SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 1-15, January 1991

AN APPROACH FOR CONSTRUCTING FAMILIES OF HOMOGENIZED EQUATIONS FOR PERIODIC MEDIA. I: AN INTEGRAL REPRESENTATION AND ITS CONSEQUENCES*

R. C. MORGAN† AND I. BABUŠKA†

Abstract. The paper, which is the first in a series of two, presents an approach by which it is possible to derive a family of homogenization approaches and assess the accuracy of any homogenization in the relation of given input data.

Key words. homogenization, composite materials

AMS(MOS) subject classifications. 35J05, 35J99, 73K20

1. Introduction. The study of periodic media is one application of partial differential equations that have highly oscillatory, periodic coefficients. Essentially, the problem is to solve the elliptic differential equation

$$-\sum_{p,q=1}^{n} \frac{\partial}{\partial x_p}\left(a_{pq}\left(\frac{x}{h}\right)\frac{\partial u^h}{\partial x_q}(x)\right)+a_0\left(\frac{x}{h}\right)u^h(x)=f(x) \tag{1}$$

on $\Omega \subset \mathbb{R}^n$ with prescribed boundary conditions or $\Omega=\mathbb{R}^n$, in which a_{pq} and a_0 are real-valued 2π-periodic functions and h is a positive number that is small in comparison to the diameter of the domain Ω.

The problem is to get the solution of (1) for relatively (to what?) small h. There is a large available mathematical literature that addresses the behavior of the solution of (1) as $h \to 0$. We mention here, for example, [3], [9], [10], and the survey [21].

One of the main applications for differential equations of type (1) is in the field of composite materials. Here the aim is to replace the composite by homogeneous materials with the bulk material properties. For various aspects we refer to [1], [2], [12], [13], [19]. A brief history is given in [2]. The accuracy of such replacement depends, of course, on the goals of the analysis. *Hence many approaches are used in applications.* The most obvious approach, namely, to use asymptotic analysis for $h \to 0$, is not always applicable because *h is given* and cannot be changed and because of particular aims of the analysis. When numerically solving problem (1) directly, we face essential difficulties of how to represent the microstructure of the composite materials. This difficulty falls into the class of solution of elliptic equations with rough coefficients. For various aspects of this problem, we refer to [6]-[8].

As was said above, various approaches can be and are used for solving (1). These approaches often give very different results (see, e.g., [10]). In addition, some of them are formulated in an abstract or in more or less analytic form. These approaches give theoretical insight, but are not well suited to a numerical treatment of the problem.

This paper presents and thoroughly analyzes an approach directed to overcoming the various major difficulties mentioned above:

(a) It allows the design of an entire class of "homogenization" formulations and judges the accuracy and reliability of any homogenization approach. It also allows the

* Received by the editors May 23, 1989; accepted for publication (in revised form) February 6, 1990. This research was partially supported by Office of Naval Research contract N00014-85-K-0169.

† Institute for Physical Science and Technology and Mathematics Department, University of Maryland, College Park, Maryland 20742.

specification of the class of problems (e.g., loads) for which a homogenization approach is applicable. In addition, it leads to a hierarchal construction of the homogenization formulations.

(b) The implementation is completely numerical, and allows adaptive modeling (selection of the equations).

We will address here only the problem with $\Omega = \mathbb{R}^n$, although very important features of the solution occur near the boundary when Ω is a bounded domain. These problems have a special character and will not be addressed here. Some brief comments will be made in § 5.

We will assume that

(i) $\Omega = \mathbb{R}^n$,

(ii) $a_0(x) \geqq \gamma_0 > 0$,

(iii) $f \in L_2(\mathbb{R}^n)$,

and that the problem is elliptic and self-adjoint.

The restriction of our analysis to a single differential equation is of a technical character only, as the ideas are also applicable to a system of equations, which would arise in elasticity problems, for example. The main idea of the approach, under the assumptions stated above, is based on the result that the solution u^h of (1) can be written in the form

$$u^h(x) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \hat{f}(t) \phi\left(\frac{x}{h}, h, t\right) e^{it \cdot x} \, dt, \tag{2}$$

in which $\hat{f}$ is the Fourier transform of f and $\phi(y, h, t)$ is a function that is 2π-periodic in y and analytic in h and t, and that solves the differential equation

$$-e^{iht \cdot y} \sum_{p,q=1}^{n} \frac{\partial}{\partial y_p} \left(a_{pq}(y) \frac{\partial}{\partial y_p} (\phi(y) \, e^{iht \cdot y}) \right) + h^2 a_0(y) \phi(y) = h^2 \tag{3}$$

on $\{y \in \mathbb{R}^n : |y_p| < \pi\}$. There are some other representations of u^h that are related to (2), but are developed in a different context. For example, in [9] (§§ 3.1 and 3.2 of Chapter 4) and [17], a spectral decomposition of the operator (3) defined on the cell (respectively, shifted cell) is used and $u^h(x)$ is expressed through Bloch expansion.

By taking various approximations of Φ, we can express approximately the solution u^h in terms of solutions of auxiliary partial differential equations with constant coefficients, or even pseudodifferential equations. We can also use Φ for the construction of the basis functions of the finite-element method. Considering the error analysis associated with the approximations of ϕ, we can design an adaptive method of selecting a "model" that would yield an approximate solution, whose accuracy meets a prescribed tolerance. These ideas are more fully discussed in [5], where we have introduced a method to systematically derive numerical, computer-oriented methods for an approximation of u^h. In § 5 we will elaborate on these ideas.

In this paper, we concentrate our attention on representation (2), whereas in [16], we make a thorough analysis of $\phi(y, h, t)$. Consequently, the properties of ϕ that are used in this paper will be stated without proof. The integral in (2) is defined as a Bochner integral of an $H^1_{-\nu}(\mathbb{R}^n)$-valued function ($H^1_{-\nu}(\mathbb{R}^n)$ is defined in the next section). As a simple application of (2), in § 5 we will give an alternate proof of the classical homogenization result (the limit of u^h as $h \to 0$).

This paper and [16] are based on the first author's Ph.D. thesis [15], in which additional details and references can be found.

2. Notation and statement of the problem. For $j=0,1$, and for any $\nu\in\mathbb{R}$, define the weighted Sobolev space $H^j_\nu(\mathbb{R}^n)$ to be the completion of $C_0^\infty(\mathbb{R}^n)$ (the complex-valued C^∞-functions that have compact support on $\mathbb{R}^n$), with respect to $\|\cdot\|_{j,\nu}$, where

$$\|u\|^2_{j,\nu}=\int_{\mathbb{R}^n}\sum_{|\alpha|\leqq j}|D^\alpha u(x)|^2\exp(2\nu|x|)\,dx.$$

(For $x\in\mathbb{R}^n$, $|x|\equiv|x_1|+\cdots+|x_n|$.) We will use $H^j(\mathbb{R}^n)$ and $\|\cdot\|_j$ to denote the standard Sobolev space and norm on $\mathbb{R}^n$ (i.e., when $\nu=0$). Next, we introduce the Sobolev spaces of periodic functions for which

$$S\equiv\{y\equiv(y_1,\cdots,y_n)\in\mathbb{R}^n\colon |y_k|<\pi \text{ for } k=1,\cdots,n\}$$

is the fundamental period. For $j=0,1$, we denote the standard Sobolev norm on S by $\|\cdot\|_{j,S}$, and we define $H^j_{\text{per}}(S)$ to be the completion, with respect to the norm $\|\cdot\|_{j,S}$, of the complex-valued C^∞-functions on $\mathbb{R}^n$ that are 2π-periodic in each coordinate variable.

Let a_{pq} $(p,q=1,\cdots,n)$ and a_0 be real-valued, 2π-periodic, L_∞-functions defined on $\mathbb{R}^n$. Furthermore, assume $a_{qp}=a_{pq}$ and assume that there exist positive constants γ_0 and γ_1 such that

$$\text{(4)}\qquad \begin{cases} a_0(x)\geqq\gamma_0 & \text{for all } \zeta_p\in\mathbb{C},\\ \displaystyle\sum_{p,q=1}^n a_{pq}(x)\zeta_q\bar\zeta_p\geqq\gamma_1\sum_{p=1}^n|\zeta_p|^2 & \end{cases}$$

almost everywhere on $\mathbb{R}^n$. For each $h>0$, define

$$\Psi(h)[u,v]=\int_{\mathbb{R}^n}\left\{\sum_{p,q=1}^n a_{pq}\left(\frac{x}{h}\right)\frac{\partial u}{\partial x_q}(x)\overline{\frac{\partial v}{\partial x_p}(x)}+a_0\left(\frac{x}{h}\right)u(x)\overline{v(x)}\right\}dx.$$

An immediate consequence of the conditions imposed on the coefficients a_0 and a_{pq} is that there exists a constant C, independent of $h>0$, such that

$$\text{(5)}\qquad \begin{aligned} &|\Psi(h)[u,v]|\leqq C\|u\|_1\|v\|_1,\\ &|\Psi(h)[v,v]|\geqq\min\{\gamma_0,\gamma_1\}\|v\|_1^2 \end{aligned}$$

for all u and v in $H^1(\mathbb{R}^n)$. Then, according to the Lax-Milgram theorem, for each $h>0$ and each $f\in L_2$, there exists a unique function $u^h\in H^1(\mathbb{R}^n)$ that satisfies

$$\text{(6)}\qquad \Psi(h)[u^h,v]=\int_{\mathbb{R}^n}f(x)\overline{v(x)}\,dx\quad\text{for all } v\in H^1(\mathbb{R}^n),$$

because

$$\text{(7)}\qquad v\mapsto\int_{\mathbb{R}^n}f(x)\overline{v(x)}\,dx$$

is a bounded linear functional on $H^1(\mathbb{R}^n)$.

Next, for each $h\in\mathbb{C}$ and $t\in\mathbb{C}^n$, define the sesquilinear form $\Phi(h,t)\colon H^1_{\text{per}}(s)\times H^1_{\text{per}}(s)\to\mathbb{C}$ by

$$\Phi(h,t)[\phi,v]\equiv\int_S\left\{\sum_{p,q=1}^n a_{pq}(y)\frac{\partial}{\partial y_p}(\phi(y)\,e^{iht\cdot y})\cdot\frac{\partial}{\partial y_p}(\overline{v(y)}\,e^{-iht\cdot y})+h^2a_0(y)\phi(y)\overline{v(y)}\right\}dy.$$

LEMMA 1. *A neighborhood $\hat{G} \subset \mathbb{C}^{n+1}$ of $\mathbb{R}^{n+1}$ can be found such that for each $(h, t) \in \hat{G}$, there exists a unique function $\phi(\cdot, h, t) \in H^1_{\text{per}}(S)$ that satisfies*

$$\Phi(h, t)[\phi(\cdot, h, t), v] = h^2 \int_S \overline{v(y)}\, dy \quad \text{for all } v \in H^1_{\text{per}}(S).$$

Furthermore, the mapping $(h, t) \in \hat{G} \mapsto \phi(\cdot, h, t) \in H^1_{\text{per}}(S)$ is holomorphic, by which we mean that about each point in $\hat{G}$, the function $(h, t) \mapsto \phi(\cdot, h, t)$ can be expanded in a power series, convergent in $H^1_{\text{per}}(S)$ and in which each coefficient is an element in $H^1_{\text{per}}(S)$.

For the most part, the proofs of statements concerning $\phi(\cdot, h, t)$ are omitted in this paper since we give a fairly comprehensive analysis of $\phi(\cdot, h, t)$ in [16].

In § 4, we show that u^h admits the representation

$$u^h(x) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \hat{f}(t) \phi\left(\frac{x}{h}, h, t\right) e^{it \cdot x}\, dt, \tag{8}$$

in which $\hat{f}(t) = (1/(2\pi)^{n/2}) \int_{\mathbb{R}^n} f(x)\, e^{-it \cdot x}\, dx$ and in which the integral is a Bochner integral of $H^1_{-\nu}(\mathbb{R}^n)$-valued functions. Our proof of (8) has as its first step the claim that for each $h > 0$ and $t \in \mathbb{R}^n$,

$$x \mapsto \phi\left(\frac{x}{h}, h, t\right) e^{it \cdot x} \tag{9}$$

solves (6) when $f(x) = e^{it \cdot x}$. However, (9) is not an element of $H^1(\mathbb{R}^n)$, and for this choice of f, (7) is not a bounded linear functional on $H^1(\mathbb{R}^n)$. Consequently, we consider $\Psi(h)$ as a sesquilinear form on $H^1_{-\nu}(\mathbb{R}^n) \times H^1_{\nu}(\mathbb{R}^n)$ for (sufficiently small) positive numbers ν.

The main tool for analyzing $\Psi(h)$ is Theorem 2 below.

THEOREM 2. *Let H_1 and H_2 be two complex Hilbert spaces with respective norms $\|\cdot\|_k$ and associated inner products $(\cdot, \cdot)_k$ for $k = 1, 2$. Let $B[\cdot, \cdot]$ be a sesquilinear form defined on $H_1 \times H_2$ for which there exist positive constants M and γ such that*

(a) $|B[u, v]| \leqq M \|u\|_1 \|v\|_2$ *for all* $u \in H_1$ *and* $v \in H_2$,

(b) $\inf_{\substack{u \in H_1 \\ \|u\|_1 = 1}} \sup_{\substack{v \in H_2 \\ \|u\|_2 \leqq 1}} |B[u, v]| \geqq \gamma > 0$,

(c) $\sup_{u \in H_1} |B[u, v]| > 0$ *for each* $v \in H_2$, $v \neq 0$.

If $f \in H_2^*$, the space of bounded conjugate-linear functionals on H_2, then there exists a unique $u_0 \in H_1$ such that

(d) $B[u_0, v] = f(v)$ *for all* $v \in H_2$,

(e) $\|u_0\|_1 \leqq \frac{1}{\gamma} \|f\|_{H_2^*}$.

A proof of Theorem 2 in the case of real Hilbert spaces can be found in [4] (as Theorem 5.2.1). The method of proof in the complex case is essentially unchanged and thus will be omitted.

We now prove the following lemma.

LEMMA 3. *There exist positive constants ν_0, C, and γ such that for all $\nu \in (0, \nu_0)$ and all $h > 0$,*

(i) $|\Psi(h)[u, v]| \leqq C \|u\|_{1,-\nu} \|v\|_{1,\nu}$,

(ii) $\inf_{\|u\|_{1,-\nu}=1} \sup_{\|v\|_{1,\nu}=1} |\Psi(h)[u, v]| \geqq \gamma > 0$,

(iii) $\sup_{u \in H^1_{-\nu}(\mathbb{R}^n)} |\Psi(h)[u, v]| > 0$ *for all* $v \in H^1_{\nu}(\mathbb{R}^n)$ *and* $v \neq 0$.

The constants ν_0, C, and γ are independent of $h > 0$; however, γ depends on ν_0.

Proof. Statement (i) follows because $\Psi(h)$ has L_∞-coefficients. To prove statement (ii), define $(Tu)(x) \equiv u(x)e^{-2\nu|x|}$ for $u \in H^1_{-\nu}(\mathbb{R}^n)$. Then for $\nu > 0$,

$$\|Tu\|^2_{1,\nu} = \int_{\mathbb{R}^n} \left(\sum_{j=1}^n \left| \frac{\partial u}{\partial x_j}(x) - 2\nu \operatorname{sgn}(x_j) u(x) \right|^2 + |u(x)|^2 \right) e^{-2\nu|x|}\, dx \tag{10}$$
$$\leqq 2(1+4n\nu^2)\|u\|^2_{1,-\nu}.$$

Now,

$$\begin{aligned}
\Psi(h)[u, Tu] &= \int_{\mathbb{R}^n} \left\{ \sum_{p,q=1}^n a_{pq}\left(\frac{x}{h}\right) \frac{\partial u}{\partial x_q}(x) \overline{\frac{\partial (Tu)}{\partial x_p}(x)} + a_0\left(\frac{x}{h}\right) u(x) \overline{T(u)} \right\} dx \\
&= \int_{\mathbb{R}^n} \left\{ \sum_{p,q=1}^n a_{pq}\left(\frac{x}{h}\right) \frac{\partial u}{\partial x_q}(x) \overline{\frac{\partial u}{\partial x_p}(x)} + a_0\left(\frac{x}{h}\right) u(x)\overline{u(x)} \right\} \exp(-2\nu|x|)\, dx \\
&\quad -2\nu \int_{\mathbb{R}^n} \left\{ \sum_{p,q=1}^n a_{pq}\left(\frac{x}{h}\right) \frac{\partial u}{\partial x_q}(x) \operatorname{sgn}(x_p) \overline{u(x)} \right\} \exp(-2\nu|x|)\, dx \\
&\equiv \Psi_1(h)[u] - 2\nu\Psi_2(h)[u].
\end{aligned} \tag{11}$$

A simple consequence of (4) is $\Psi_1(h)[u] \geqq \min\{\gamma_0, \gamma_1\}\|u\|^2_{1,-\nu}$. We also have $|\Psi_2(h)[u]| \leqq c\|u\|^2_{1,-\nu}$ for some constant c, independent of h. Combining these two inequalities with (10) and (11) yields

$$\begin{aligned}
|\Psi(h)[u, Tu]| &\geqq (\min\{\gamma_0, \gamma_1\} - 2c\nu)\|u\|^2_{1,-\nu} \\
&\geqq \frac{\min\{\gamma_0, \gamma_1\} - 2c\nu}{\sqrt{2(1+4n\nu^2)}} \|u\|_{1,-\nu}\|Tu\|_{1,\nu}.
\end{aligned}$$

Consequently, there exist positive constants ν_0 and γ, which are independent of $h > 0$, for which $\Psi(h)[u, Tu] \geqq \gamma\|u\|_{1,-\nu}\|Tu\|_{1,\nu}$ for all $\nu \in (0, \nu_0)$. This proves (ii).

Statement (iii) is proven in a similar manner. □

Throughout the remainder of this paper, we implicitly assume $\nu \in (0, \nu_0)$. Let $f \in L_2(\mathbb{R}^n)$; then

$$\left| \int_{\mathbb{R}^n} f(x)\overline{v(x)}\, dx \right| \leqq \|f\|_0\|v\|_{1,\nu} \quad \text{for all } v \in H^1_\nu(\mathbb{R}^n).$$

Lemma 3, in conjunction with Theorem 2, now yields the next theorem.

THEOREM 4. *For each $h > 0$ and $f \in L_2(\mathbb{R}^n)$ there exists a unique function $u^h \in H^1_{-\nu}(\mathbb{R}^n)$ for which*

$$\Psi(h)[u^h, v] = \int_{\mathbb{R}^n} f(x)\overline{v(x)}\, dx \quad \textit{for all } v \in H^1_\nu(\mathbb{R}^n). \tag{12}$$

Furthermore, $\|u^h\|_{1,-\nu} \leqq (1/\gamma)\|f\|_0$.

There is no ambiguity in denoting the unique solutions of (6) and (12) by u^h because $H^1_\nu(\mathbb{R}^n) \subset H^1(\mathbb{R}^n) \subset H^1_{-\nu}(\mathbb{R}^n)$ implies they are the same function. However, (12) can be solved when f belongs to a broader class of functions than the class of L_2-functions, namely, when f belongs to the dual space of $H^1_\nu(\mathbb{R}^n)$. One such function is defined by $f(x) = e^{it\cdot x}$.

3. Preliminaries.

THEOREM 5. *There exists a constant C_0 independent of t, such that*

$$\|\chi\, e^{it\cdot x}\|_{1,S} \leqq C_0(1+\|t\|)\|\chi\|_{1,S}$$

for all $\chi \in H^1(S)$ and for all $t \in \mathbb{R}^n$, where $\|t\|^2 \equiv t_1^2 + \cdots + t_n^2$.

For each $h>0$ and each $\omega \in \mathbb{Z}^n$, define

$$S(h, \omega) \equiv \{x \in \mathbb{R}^n : (\omega_j - 1)\pi h < x_j < (\omega_j + 1)\pi h, j = 1, \cdots, n\}.$$

Note that $S(1, (0, \cdots, 0)) = S$ is the fundamental domain for the periodic spaces.

LEMMA 6. *For each $h>0$ and $\nu>0$, there exists a constant $C_1(h, \nu)$ that remains bounded as $h \to 0$, such that for all $t \in \mathbb{R}^n$, and for any $\chi \in H^0_{\text{per}}(S)$,*

$$\text{(i)} \qquad \left\| \chi\left(\frac{x}{h}\right) e^{it\cdot x} \right\|_{0,-\nu} \leqq C_1(h, \nu) \|\chi\|_{0,S}$$

and for any $\chi \in H^1_{\text{per}}(S)$,

$$\text{(ii)} \qquad \left\| \chi\left(\frac{x}{h}\right) e^{it\cdot x} \right\|_{1,-\nu} \leqq (1+h^{-1}) C_1(h, \nu) \|\chi \exp(iht \cdot y)\|_{1,S}.$$

Proof.

$$\left\| \chi\left(\frac{x}{h}\right) \exp(it \cdot x) \right\|^2_{1,-\nu}$$

$$\leqq \sum_{\omega \in \mathbb{Z}^n} \int_{S(h,2\omega)} \left(\sum_{p=1}^{n} \left| \frac{\partial}{\partial x_p} \left(\chi\left(\frac{x}{h}\right) \exp(it \cdot x) \right) \right|^2 + \left| \chi\left(\frac{x}{h}\right) \right|^2 \right) \exp(-2\nu|x|)\, dx,$$

where $2\omega \equiv (2\omega_1, \cdots, 2\omega_n)$. Making the substitution $(x/h) - y + 2\pi\omega$ in the integral over $S(h, 2\omega)$ and using the periodicity of χ yields a constant $\tilde{c}$, independent of h, such that

$$\left\| \chi\left(\frac{x}{h}\right) \exp(it \cdot x) \right\|^2_{1,-\nu}$$

$$= \sum_{\omega \in \mathbb{Z}^n} \left\{ \int_S \left(h^{-1} \sum_{p=1}^{n} \left| \frac{\partial}{\partial y_p} (\chi(y) \exp(iht \cdot y)) \exp(i2\pi h\omega \cdot t) \right|^2 + |\chi(y)|^2 \right) \right.$$

$$\left. \cdot \exp(-2\nu h|y + 2\pi\omega|) h^n\, dy \right\}$$

$$\leqq \tilde{c}(1+h^{-2}) \left(\sum_{\omega \in \mathbb{Z}^n} \exp(-4\pi\nu h|\omega|) \right) h^n \|\chi \exp(iht \cdot y)\|^2_{1,S}.$$

It is not difficult to prove that there is a constant c that is independent of h such that

$$h^n \sum_{\omega \in \mathbb{Z}^n} \exp(-4\pi\nu h|\omega|) \leqq c h^n \left(1 + \int_{\mathbb{R}^n} \exp(-r\pi\nu h|x|)\, dx \right).$$

Upon setting $C_1(h, \nu) \equiv \sqrt{\tilde{c}c(h^n + (2\pi\nu)^{-n})}$, (ii) follows.

A similar argument, in which the contribution from the first-order derivatives is ignored, produces (i). □

LEMMA 7. *For all $h>0$ and $t \in \mathbb{R}^n$, we have*

$$\|\phi(\cdot, h, t)\|_{0,S} \leqq \frac{(2\pi)^{n/2}}{\gamma_0},$$

$$\|\phi(\cdot, h, t) \exp(iht \cdot y)\|_{1,S} \leqq C_2(h)$$

for some positive number $C_2(h)$.

Proof. It follows from (4) and Lemma 1 that

$$\gamma_1 \int_S \sum_{p=1}^{n} \left| \frac{\partial}{\partial y_p} (\phi(y, h, t) \exp(iht \cdot y)) \right|^2 dy + \gamma_0 h^2 \|\phi(\cdot, h, t)\|_{0,S}^2$$
$$\leqq \Phi(h, t)[\phi(\cdot, h, t), \phi(\cdot, h, t)]$$
$$\leqq (2\pi)^{n/2} h^2 \|\phi(\cdot, h, t)\|_{0,S}.$$

The lemma now follows with

$$C_2(h) \equiv \frac{(2\pi)^{n/2}}{\min\{\gamma_0, \gamma_1 h^{-2}\}}. \qquad \square$$

4. The representation of u^h. We begin with Theorem 8 below.

THEOREM 8. *For each $h > 0$ and $t \in \mathbb{R}^n$,*

(i) *$x \mapsto \phi((x/h), h, t)\, e^{it\cdot x}$ is in $H^1_{-\nu}(\mathbb{R}^n)$;*

(ii) *$\Psi(h)[\phi((x/h), h, t)\, e^{it\cdot x}, v] = \int_{\mathbb{R}^n} \exp(it\cdot x)\overline{v(x)}\, dx$ for all $v \in H^1_\nu(\mathbb{R}^n)$.*

Furthermore,

(iii) $$\left\| \phi\left(\frac{x}{h}, h, t\right) e^{it\cdot x} \right\|_{1,-\nu} \leqq \frac{1}{\gamma \nu^{n/2}}.$$

Proof. Statement (i) follows from Lemmas 6 and 7, which imply $\|\phi((x/h), h, t)\, e^{it\cdot x}\|_{1,-\nu} \leqq (1 + h^{-1}) C_1(h, \nu) C_2(h)$. Assuming that (ii) is true, (iii) is a consequence of Theorem 2, Lemma 3, and the fact $|\int_{\mathbb{R}^n} e^{it\cdot x}\overline{v(x)}\, dx| \leqq (1/\nu^{n/2})\|v\|_{1,\nu}$. The proof of (ii) is based upon determining the relationship between $\Psi(h)$ and $\Phi(h, t)$, and then using Lemma 1.

For each $h > 0$, there exists a locally finite, C^∞-partition of unity $\{\sigma_\omega(\cdot, h) : \omega \in \mathbb{Z}^n\}$ subordinated to $\{S(h, \omega) : \omega \in \mathbb{Z}^n\}$ such that $\sum_{\omega\in\mathbb{Z}^n} \sigma_\omega(\cdot, h)v$ converges to v in $H^1_\nu(\mathbb{R}^n)$ whenever $v \in H^1_\nu(\mathbb{R}^n)$. The basic requirement of the partition of unity is that $|\sigma_\omega(x, h)|$ and $|(\partial\sigma_\omega/\partial x_p)(x, h)|$ for $p = 1, \cdots, n$ are uniformly bounded for $x \in \mathbb{R}^n$ and $\omega \in \mathbb{Z}^n$. Then $v_\omega(\cdot, h) \equiv \sigma_\omega(\cdot, h)v$ has compact support in $S(h, \omega)$, and for any $\chi \in H^1_{\text{per}}(S)$,

$$\Psi(h)\left[\chi\left(\frac{x}{h}\right) e^{it\cdot x}, v\right] = \sum_{\omega\in\mathbb{Z}^n} \Psi(h)\left[\chi\left(\frac{x}{h}\right) \exp(it\cdot x), v_\omega(x, h)\right]$$

$$\leqq \sum_{\omega\in\mathbb{Z}^n} \int_{S(h,\omega)} \left(\sum_{p,q=1}^{n} a_{pq}\left(\frac{x}{h}\right) \frac{\partial}{\partial x_q}\left(\chi\left(\frac{x}{h}\right) \exp(it\cdot x)\right) \overline{\frac{\partial v_\omega}{\partial x_p}(x, h)} \right. \tag{13}$$
$$\left. + a_0\left(\frac{x}{h}\right)\chi\left(\frac{x}{h}\right)\left(\exp(it\cdot x)\overline{v_\omega(x, h)}\right) \right) dx$$

because $\Psi(h)$ is a continuous sesquilinear form on $H^1_{-\nu}(\mathbb{R}^n) \times H^1_\nu(\mathbb{R}^n)$, according to Lemma 3 (i). In an effort to transform the region of integration $S(h, \omega)$ into $S = S(1, (0, \cdots, 0))$ in each integral, make the substitution $x/h = y + \pi\tilde{\omega}$ for $x \in S(h, \omega)$, in which $\tilde{\omega}$ is the n-tuple of even integers, that is, derived from ω according to

$$\tilde{\omega}_p = \begin{cases} \omega_p, & \omega_p \text{ even}, \\ \omega_p - 1, & \omega_p \text{ odd}. \end{cases}$$

Using the periodicity of a_0, $\{a_{pq} : p, q = 1, \cdots, n\}$, and χ, each integral in (13) becomes

$$\int_{S(1,\omega-\tilde{\omega})} \left\{ h^{-2} \sum_{p,q=1}^{n} a_{pq}(y) \frac{\partial}{\partial y_q}(\chi(y) \exp(iht\cdot(y + \pi\tilde{\omega}))) \overline{\frac{\partial v_\omega}{\partial x_p}(h(y + \pi\tilde{\omega}), h)} \right. \tag{14}$$
$$\left. + a_0(y)\chi(y) \exp(iht\cdot(y + \pi\tilde{\omega}))\overline{v_\omega(h(y + \pi\tilde{\omega}), h)} \right\} h^n\, dy.$$

Next, for each $\omega \in \mathbb{Z}^n$, define

$$v^0_\omega(y, h, t) \equiv v_\omega(h(y + \pi\tilde{\omega}), h) \exp(-iht\cdot(y + \pi\tilde{\omega})) \tag{15}$$

for $y \in S(1, \omega - \tilde{\omega})$ and extend $v_\omega^0(\cdot, h, t)$ to all of $\mathbb{R}^n$ by 2π-periodicity. Since the support of $y \mapsto v_\omega(h(y + \pi\tilde{\omega}), h)$ is contained in $S(1, \omega - \tilde{\omega})$, it follows that $v_\omega^0(\cdot, h, t) \in H^1_{\text{per}}(S)$. Using (15) to substitute for $v_\omega(h(y + \pi\tilde{\omega}), h)$ in (14) yields

$$\int_{S(1,\omega-\tilde{\omega})} \left\{ h^{-2} \sum_{p,q=1}^{n} a_{pq}(y) \frac{\partial}{\partial y_q}(\chi(y) \exp(iht \cdot y)) \frac{\partial}{\partial y_q} \overline{(v_\omega^0(y, h, t) \exp(-iht \cdot y))} \right.$$
$$\left. + \sigma a_0(y) \chi(y) \overline{v_\omega^0(y, h, t)} \right\} h^n \, dy.$$

Now, the domain of integration $S(1, \omega - \tilde{\omega})$ can be replaced with $S(1, (0, \cdots, 0)) = S$, and consequently

$$\Psi(h)\left[\chi\left(\frac{x}{h}\right) \exp(it \cdot x), v\right] = \sum_{\omega \in \mathbb{Z}^n} h^{n-2} \Phi(h, t)[\chi, v_\omega^0(\cdot, h, t)] \tag{16}$$

for all $v \in H^1_\nu(\mathbb{R}^n)$.

Noting Lemma 1, it is now a simple matter to prove (ii):

$$\Psi(h)\left[\phi\left(\frac{x}{h}, h, t\right) \exp(it \cdot x), v\right] = \sum_{\omega \in \mathbb{Z}^n} h^{n-2} \Phi(h, t)[\phi(\cdot, h, t), v_\omega^0(\cdot, h, t)]$$
$$= \sum_{\omega \in \mathbb{Z}^n} h^n \int_S \overline{v_\omega^0(y, h, t)} \, dy$$
$$= \int_{\mathbb{R}^n} \exp(it \cdot x) v(x) \, dx)$$

for all $v \in H^1_\nu(\mathbb{R}^n)$, since $v_\omega^0(\cdot, h, t) \in H^1_{\text{per}}(S)$. □

LEMMA 9. *For each $h > 0$, $t \in \mathbb{R}^n \mapsto \phi((x/h), h, t) \exp(it \cdot x) \in H^1_{-\nu}(\mathbb{R}^n)$ is a continuous mapping.*

Proof. The continuity of $t \mapsto e^{itx} \in H^1_{-\nu}(\mathbb{R}^n)$ follows in a straightforward manner. Upon setting $t = 0$ and $\chi = \phi(\cdot, h, t) - \phi(\cdot, h, \tau)$ in Lemma 6,

$$\lim_{t \to \tau} \left\| \phi\left(\frac{x}{h}, h, t\right) - \phi\left(\frac{x}{h}, h, \tau\right) \right\|_{1,-\nu}$$
$$\leqq (1 + h^{-1}) C_1(h, \nu) \lim_{t \to \tau} \|\phi(\cdot, h, t) - \phi(\cdot, h, \tau)\|_{1,S}$$
$$= 0$$

follows from Lemma 1. □

For each $f \in L_2(\mathbb{R}^n)$, the Fourier transform of f is defined by

$$\hat{f}(t) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} f(x) \exp(-it \cdot x) \, dx.$$

The following notation will be used with the function f only (as it appears in (12)). For any $f \in L_2(\mathbb{R}^n)$ and $N > 0$, define $f_N \in L_2(\mathbb{R}^n)$ as the inverse Fourier transform of $\hat{f}|_{\{t: \|t\| \leqq N\}}$, i.e.,

$$f_N(x) = \frac{1}{(2\pi)^{n/2}} \int_{\|t\| \leqq N} \hat{f}(t) \exp(it \cdot x) \, dt, \qquad \hat{f}_N(t) = \begin{cases} \hat{f}(t), & \|t\| \leqq N, \\ 0, & \|t\| > N. \end{cases}$$

Parseval's inequality implies

$$\lim_{N \to \infty} \|f - f_N\|_0 = \lim_{N \to \infty} \|\hat{f} - \hat{f}_N\|_0 = 0.$$

Next, for each $N>0$, define $u^h(\cdot; N)\in H^1_{-\nu}(\mathbb{R}^n)$ by

$$u^h(x; N)=\frac{1}{(2\pi)^{n/2}}\int_{\|t\|\leqq N}\hat{f}(t)\phi\left(\frac{x}{h}, h, t\right)\exp(it\cdot x)\,dt$$

$$\left(=\frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n}\hat{f}_N(t)\phi\left(\frac{x}{h}, h, t\right)\exp(it\cdot x)\,dt\right),$$

in which the integral is to be interpreted as a Bochner integral of $H^1_{-\nu}(\mathbb{R}^n)$-valued function (cf. [18]). In order to show that the integral has such a meaning, we need to show that the integrand is strongly measurable and that $t\mapsto\|\hat{f}(t)\phi(x/h, h, t)\,e^{it\cdot x}\|_{1,-\nu}$ is Lebesgue integrable over $\{t:\|t\|\leqq N\}$. This integrability condition is satisfied as a result of Theorem 8 (iii) and because $\hat{f}$ is an L_1-function on $\{t:\|t\|\leqq N\}$. By strong measurability of the integrand, we mean that there exists a sequence of simple functions $t\mapsto w_k(\cdot, t)\in H^1_{-\nu}(\mathbb{R}^n)$ such that

$$\lim_{k\to\infty}\left\|\hat{f}(t)\phi\left(\frac{x}{h}, h, t\right)\exp(it\cdot x)-w_k(x, t)\right\|_{1,-\nu}=0. \tag{17}$$

It is easy to construct such a sequence (and we do so in the proof of Theorem 10 below) using Lemma 9 and the measurability of f. Furthermore, if the sequence of simple functions satisfies

$$\lim_{K\to\infty}\int_{\|t\|\leqq N}\left\|\hat{f}(t)\phi\left(\frac{x}{h}, h, t\right)\exp(it\cdot x)-w_k(x, t)\right\|_{1,-\nu}=0 \tag{18}$$

then

$$\int_{\|t\|\leqq n}\hat{f}(t)\phi\left(\frac{x}{h}, h, t\right)e^{it\cdot x}\,dt=\lim_{k\to\infty}\int_{\|t\|\leqq N}w_k(\cdot, t)\,dt\quad\text{in } H^1_{-\nu}(\mathbb{R}^n),$$

in which the integral of a simple function is defined in the standard manner.

THEOREM 10. *For each $N>0$,*

$$\psi(h)[u^h(\cdot; N), v]=\int_{\mathbb{R}^n}f_N(x)\overline{v(x)}\,dx\quad\textit{for all } v\in H^1_\nu(\mathbb{R}^n).$$

Proof. The essence of this proof is that a sequence of simple functions satisfying (17) and (18) can be chosen so that each value of each simple function is of the form $c\phi((x/h), h, \tau)\,e^{i\tau\cdot x}$ for some $\tau\in\{t:\|t\|\leqq N\}$ and for some complex number c. With this being the case, Theorem 8 (ii) can be used to evaluate $\Psi(h)[u, v]$ whenever u is the integral of one of these simple functions.

To begin, note that $H^1_\nu(\mathbb{R}^n)\subset L_1(\mathbb{R}^n)$. Consequently, $\hat{v}$ is continuous (cf. [15]) when $v\in H^1_\nu(\mathbb{R}^n)$, and it follows from Theorem 8 that

$$\Psi(h)\left[\phi\left(\frac{x}{h}, h, \tau\right)e^{i\tau\cdot x}, v\right]=(2\pi)^{n/2}\widehat{\bar{v}(\tau)}$$

is well defined for each $\tau\in\mathbb{R}^n$.

Set $\Omega\equiv\{t\in\mathbb{R}^n:\|t\|\leqq N\}$. Now select a sequence of simple functions $s_k:\Omega\to\mathbb{C}$ and a sequence of collections $\{\Omega_{k,j}:j=0,\cdots, M_k\}$ of measurable subsets of Ω for $k=1, 2,\cdots$ and for which the following is true. For each k:

(a) $\Omega=\cup_{j=0}^{M_k}\Omega_{k,j}$, $\Omega_{k,j}\cap\Omega_{k,l}=\phi$ for $j\neq l$;

(b) for $j=1,\cdots,M_k$, if t and τ are in $\Omega_{k,j}$, then

$$\left\| \phi\left(\frac{x}{h}, h, t\right) \exp(it\cdot x) - \phi\left(\frac{x}{h}, h, t\right) \exp(it\cdot x) \right\|_{1,-\nu} < 2^{-k},$$

and $|\hat{v}(t)-\hat{v}(\tau)|<2^{-k}$;

(c) s_k is constant on $\Omega_{k,j}$ for $j=0,\cdots,M_k$, with $s_k(t)=0$ for $t\in\Omega_{k,0}$, $s_k\to\hat{f}_N$ pointwise almost everywhere on Ω, and $|s_k(t)|\leqq|\hat{f}_N(t)|$ almost everywhere on Ω. Next, for each k, pick a $\tau^{k,j}\in\Omega_{k,j}$ for each $j=1,\cdots,M_k$, and define

$$w_k(x,t)=\begin{cases}0, & t\in\Omega_{k,0},\\ \phi(x/h,h,\tau^{k,j})\exp(i\tau^{k,j}\cdot x), & t\in\Omega_{k,j},\end{cases}$$

$$r_k(t)=\begin{cases}0, & t\in\Omega_{k,0},\\ \hat{v}(\tau^{k,j}), & t\in\Omega_{k,j}.\end{cases}$$

A consequence of the preceding construction and of Theorem 8 (iii) is

$$\left\| \hat{f}_N(t)\phi\left(\frac{x}{h}, h, t\right) e^{it\cdot x} - s_k(t)w_k(x,t) \right\|_{1,-\nu} \leqq |\hat{f}_N(t)-s_k(t)|\frac{1}{\gamma\nu^{n/2}} + |\hat{f}_N(t)|2^{-k}$$

$$\leqq \left(\frac{1}{\gamma\nu^{n/2}}+2^{-k}\right)|\hat{f}_N(t)|.$$

It follows from the first inequality that

$$\lim_{k\to\infty} \left\| \hat{f}_N(t)\phi\left(\frac{x}{h}, h, t\right) \exp(it\cdot x) - s_k(t)w_k(\cdot,t) \right\|_{1,-\nu} = 0 \quad \text{a.e.}$$

Furthermore, the Lebesgue dominated convergence theorem and the second inequality imply

$$\lim_{k\to\infty} \int_\Omega \left\| \hat{f}_N(t)\phi\left(\frac{x}{h}, h, t\right) \exp(it\cdot x) - s_k(t)w_k(\cdot,t) \right\|_{1,-\nu} dt = 0.$$

Consequently,

$$u^h(\cdot,N) = \lim_{k\to\infty} (2\pi)^{-n/2} \int_{\|t\|\leqq N} s_k(t)w_k(\cdot,t)\,dt \quad \text{in } H^1_{-\nu}(\mathbb{R}^n).$$

Finally, the continuity of $u\mapsto\Psi(h)[u,v]$ (Lemma 3); the definitions of s_k, w_k, and r_k; and Parseval's equality imply

$$\Psi(h)[u^h(\cdot;N),v] = \lim_{k\to\infty} (2\pi)^{-n/2}\left[\int_{\|t\|\leqq N} s_k(t)w_k(\cdot,t)\,dt, v\right]$$

$$= \lim_{k\to\infty} \int_{\|t\|\leqq N} s_k(t)r_k(t)\,dt$$

$$= \int_{\|t\|\leqq N} \hat{f}_N(t)\overline{\hat{v}(t)}\,dt$$

$$= \int_{\mathbb{R}^n} f_N(x)\overline{v(x)}\,dx$$

for all $v\in H^1_\nu(\mathbb{R}^n)$, because $\hat{f}_N(t)=0$ for $\|t\|>N$. □

It is now a simple matter to prove the main result of this paper.

THEOREM 11. *Suppose $h>0$ and $f\in L_2(\mathbb{R}^n)$. Let $u^h\in H^1_{-\nu}(\mathbb{R}^n)$ be the solution of* (12). *Then*

$$u^h(x)=\lim_{N\to\infty}(2\pi)^{-n/2}\int_{\|t\|\leqq N}\hat{f}(t)\phi\left(\frac{x}{h},h,t\right)\exp(it\cdot x)\,dt \quad \text{in } H^1_{-\nu}(\mathbb{R}^n), \tag{19}$$

where, for each N, the integral is defined as a Bochner integral of $H^1_{-\nu}(\mathbb{R}^n)$-valued functions.

Proof. A consequence of Theorem 10 is that

$$\Psi(h)[u^h-u^h(\cdot;N),v]=\int_{\mathbb{R}^n}(f(x)-f_N(x))\overline{v(x)}\,dx \tag{20}$$

for all $v\in H^1_{\nu}(\mathbb{R})$, from which the inequality

$$\|u^h-u^h(\cdot;N)\|_{1,-\nu}\leqq\frac{1}{\gamma}\|f-f_N\|_0 \tag{21}$$

is easily derived (cf. Theorem 2 and Lemma 3). Then (19) follows. □

Actually, (19) converges in $H^1(\mathbb{R}^n)$ even though each integral is defined only as a function in $H^1_{-\nu}(\mathbb{R}^n)$. The reasoning that allows us to identify the unique solutions of (6) and (12) also yields $u^h(\cdot;N)\in H^1(\mathbb{R}^n)$ and the fact that (20) is valid for all $v\in H^1(\mathbb{R}^n)$. Now, the Lax-Milgram theorem and (5) imply

$$\|u^h-u^h(\cdot;N)\|_1\leqq\frac{1}{\min\{\gamma_0,\gamma_1\}}\|f-f_N\|_0.$$

5. Homogenization. In this section we derive first the classical result of homogenization, which states that u^h converges, as h tends to zero, to a function that is the solution of a constant coefficient partial differential equation. This is an example of analyzing u^h through (19) and an analysis of $\phi(\cdot,h,t)$.

According to Lemma 1, it is possible to expand $\phi(\cdot,h,t)$ in powers of h, for each $t\in\mathbb{R}^n$. Consequently, we can write

$$\phi(\cdot,h,t)=\phi_0(\cdot,t)+\phi_1(\cdot,t)h+\cdots. \tag{22}$$

The functions $\{\phi_j(\cdot,t):j=0,1,\cdots\}$ can be determined by expanding (3) in powers of h and substituting (22). Here we are interested in only the constant term, and solving for it yields

$$\phi_0(\cdot,t)=g_0(t)=\frac{1}{\sum_{p,q=t}^{n}A_{pq}t_qt_p+A_0},$$

where A_0 and $\{A_{pq}: p,q=1,\cdots,n\}$ are derived from the periodic coefficients a_0 and $\{a_{pq}: p,q=1,\cdots,n\}$ and certain auxiliary functions. Complete details are given in [16], where proofs of the properties of $\phi(\cdot,h,t)$ that are stated in the following lemma can also be found.

LEMMA 12. *There exist positive constants θ and G_0, and continuous functions $g_0:\mathbb{R}^n\to(0,\infty)$ and $G_1:\{(h,t)\in\mathbb{R}^{n+1}: 0\leqq\theta h(1+\|t\|)<1\}\to(0,\infty)$ such that*

(i) $g_0(t)\leqq G_0/1+\|t\|^2$,

(ii) $\|\phi(\cdot,h,t)-g_0(t)\|_{1,S}\leqq G_1(h,t)h$ *for each $h\geqq 0$ and $t\in\mathbb{R}^n$ that satisfy $\theta h(1+\|t\|)<1$.*

A consequence of (i) is that we can define functions u_0 and $u_0(\cdot\,; N)$, for $N>0$, in $H^2(\mathbb{R}^n)$ by

$$u_0(x) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} \hat{f}(t) g_0(t) \exp(it \cdot x)\, dt,$$
$$u_0(x; N) = (2\pi)^{-n/2} \int_{\|t\|<N} \hat{f}(t) g_0(t) \exp(it \cdot x)\, dt \tag{23}$$

whenever $\hat{f} \in L_2(\mathbb{R}^n)$ (cf. [18]). Note that $u_0(\cdot\,; N) = u_0$ with $f = f_N$. We have $\hat{u}_0 = \hat{f} g_0$ and $(u_0(\cdot\,; N)) = \hat{f}_N g_0$, and then Parseval's equality implies

$$\|u_0 - u_0(\cdot\,; N)\|_0 \leqq G_0 \|f - f_N\|_0. \tag{24}$$

The integrals that define u_0 and $u_0(\cdot\,; N)$ are to be interpreted as Lebesgue integrals of numerical-valued functions. However, we want to interpret $u_0(\cdot\,; N)$ as an integral of $H^1_{-\nu}(\mathbb{R}^n)$-valued functions. In the Appendix we show that the integral in (23) can be interpreted in both ways, in an unambiguous and consistent manner. Furthermore, note that u_0 is the solution of the constant coefficient differential equation

$$-\sum_{p,q=1}^{n} A_{pq} \frac{\partial^2 u_0}{\partial x_p \partial x_q}(x) + A_0 u_0(x) = f(x). \tag{25}$$

We can now prove the classical result in homogenization.

THEOREM 13. *Let $f \in L_2(\mathbb{R}^n)$. Let $u^h \in H^1_{-\nu}(\mathbb{R}^n$ be the solution of* (12) *for $h>0$, and let $u_0 \in H^1(\mathbb{R}^n) \subset H^1_{-\nu}(\mathbb{R}^n)$ be the solution of* (25). *Then* $\lim_{h\to 0} \|u^h - u_0\|_{0,-\nu} = 0$.

Proof. For each $N>0$, we have

$$\begin{aligned} \|u^h - u_0\|_{0,-\nu} &\leqq \|u^h - u^h(\cdot\,; N)\|_{0,-\nu} \\ &\quad + \|u^h(\cdot\,; N) - u_0(\cdot\,; N)\|_{0,-\nu} + \|u_0(\cdot\,; N) - u_0\|_{0,-\nu}. \end{aligned} \tag{26}$$

Let $\varepsilon>0$ be given. It follows from (21) and (24) that we can choose an $N>0$ that makes each of the first and third terms on the right-hand side of (26) smaller than ε, uniformly in h. Next, a consequence of Lemmas 6 and 12 is that there exists $h_0 \in (0, 1/\theta(1+N))$ such that for all $h \in (0, h_0)$,

$$\begin{aligned} &\|u^h(\cdot\,; N) - u_0(\cdot\,; N)\|_{0,-\nu} \\ &\quad \leqq (2\pi)^{-n/2} \int_{\|t\|\leqq N} |\hat{f}(t)| \left\| \left(\phi\left(\frac{x}{h}, h, t\right) - g_0(t) \right) \exp(it \cdot x) \right\|_{0,-\nu} dt \\ &\quad \leqq (2\pi)^{-n/2} C_1(h, \nu) h \int_{\|t\|\leqq N} |\hat{f}(t)| G_1(h, t)\, dt \\ &\quad < \varepsilon. \end{aligned}$$

□

We have shown that an expansion of the function $\Phi(\cdot, h, t)$ leads directly to the classical homogenization. A truncated series (22) is a special approximation of Φ. Comparing it with Φ, we get precise information whether the results are for given h and f in the range of admissible accuracy.

We, of course, have many other possibilities to employ function Φ for a derivation of homogenized equation and numerical treatment of the (original) problem, so that the admissible accuracy is achieved. We mention some of them here.

(a) For given h (which in engineering applications cannot be made "sufficiently" small), we can approximate the function $\Phi(\cdot, h, t)$ in the form

$$\Phi(\cdot, h, t) \cong \sum_{k=1}^{n} \varphi_k(\cdot)\psi_k(t), \tag{27}$$

where $\psi_k(t)$ are rational functions. The approximation (27) is such that the error

$$\varepsilon = \Phi(\cdot, h, t) - \sum_{k=1} \varphi_k(\cdot)\psi_k(t)$$

is small for $t_0 \leqq t \leqq t_1$, $0 \leqq x/h \leqq 2\pi$, where $\hat{f}(t)$ is not negligible. We can also (by a smoothing technique) decompose f,

$$f = \sum_{j=1}^{n} f_j$$

so that $\hat{f}_j$ is small outside of $t_j < t < t_{j+1}$, and solve m homogenization problems separately. By this approach any prescribed accuracy can be achieved. Every $\psi_k(t) = \bar{\psi}_k(t)/\rho_k(t)$ is the symbol of the homogenized problem

$$L_1 u^h = L_2 f.$$

Here L_1, respectively, L_2, are operators associated to ψ_k, respectively, ρ_k.

(b) We can approximate the function Φ separately for every fixed value x and derive homogenized equations associated with this point. This is important in application because usually we are interested in the solutions in some specific (dangerous) places only.

(c) Often we are interested not in the solution but in the fluxes (stresses), stress intensity factors, etc., in some specific points. Writing expressions analogous to (2) for these points we can construct other adequate homogenizations.

(d) Given a concrete homogenization we can compare its symbol with Φ and get the information about the admissible range of f and accuracy of such an approach. For example, we can judge whether h is "sufficiently" small. By this we can characterize the approximation of various approaches in much more effective ways than have been used in [11].

The function $\Phi(\cdot, h, t)$ is the solution of the elliptic problem (3) with parameters h and t. The function $\Phi(\cdot, h, t)$ can be found numerically without any difficulties by the finite-element method (in the range of a priori given tolerance). The computation for various values t and h can be done in parallel. The approximation we mentioned above is then simply post-processing of the computed data.

(e) The function Φ can be directly employed for the construction of a special element for solving the original problem analogously as has been discussed in [8]. Here we will use the finite-element basis function of the form

$$\varphi_{i,j} = \rho_j(x)\Phi\left(\frac{x}{h}, h, t_i\right), \qquad i = 1, \cdots, s, \tag{28}$$

where t_i are properly chosen points and ρ_j is the classical "hat" function associated to the nodal point x_j in the finite element method. Here, of course, denoting by H the meshsize of the elements, we have $H \gg h$.

There are many other opportunities to employ the function Φ for computational purposes and to create a hierarchical family of homogenizations. The selection of the adequate homogenized equation then could be made adaptively. Nevertheless, it is out of the scope of this paper to elaborate more and we refer to [5] for additional aspects and numerical examples.

So far we assumed that $\Omega = \mathbb{R}^n$. If Ω has a bounded boundary (e.g., crack problem) the solution is much more complicated. Here always a boundary layer or singular behavior is present. A practical way to deal with this problem is to use refined meshes and basis functions of the form (24) with sufficiently large s (which also could be determined in an adaptive way).

6. Appendix. Define $w(x, t) = (w\pi)^{n/2}\hat{f}(t)g_0(t)\, e^{it\cdot x}$, where $f \in L_2(\mathbb{R}^n)$ and g_0 is defined as in Lemma 13. According to (22), $u_0(\cdot\,; N) \in H^2(\mathbb{R}^n)$ is defined by

$$u_0(x; N) = \int_{\|t\| \leqq N} w(x, t)\, dt$$

as an integral of numerical-valued functions. Since

$$t \mapsto w(\cdot, t) \text{ is strongly measurable in } H^1_{-\nu}(\mathbb{R}^n),$$

$$\int_{\|t\| \leqq N} \|w(\cdot, t)\|_{1,-\nu}\, dt \leqq (2\pi)^{n/2} G_0 \nu^{-n/2} \int_{\|t\| \leqq N} |\hat{f}(t)|\, dt < \infty,$$

it follows that

$$W = \int_{\|t\| \leqq N} w(\cdot, t)\, dt$$

can be defined as an integral of $H^1_{-\nu}(\mathbb{R}^n)$-valued functions. We want to prove the following lemma.

LEMMA 14. *$u_0(\cdot\,; N) = W$ as a function in $H^1_{-\nu}(\mathbb{R}^n)$.*

Proof. It suffices to show that $x \mapsto u_0(x; N) \exp(-\nu|x|)$ and $x \mapsto W(x) \exp(-\nu|x|)$ generate the same generalized function. It follows from (27) that a sequence $t \mapsto w_k(\cdot, t) \in H^1_{-\nu}(\mathbb{R}^n)$ of simple functions can be chosen so that

$$\lim_{k\to\infty} \|w_k(\cdot, t) - w(\cdot, t)\|_{1,-\nu} = 0 \quad \text{a.e.},$$

$$\lim_{k\to\infty} \int_{\|t\| \leqq N} \|w_k(\cdot, t) - w(\cdot, t)\|_{1,-\nu}\, dt = 0,$$

$$\|w_k(\cdot, t)\|_{1,-\nu} \leqq \frac{3}{2} \|w(\cdot, t)\|_{1,-\nu} \quad \text{a.e.}$$

Then by definition, $W = \lim_{k\to\infty} \int_{\|t\| \leqq N} w_k(\cdot, t)\, dt$ in $H^1_{-\nu}(\mathbb{R}^n)$.

Let $\psi \in C_0^\infty(\mathbb{R}^n)$. Then Fubini's theorem, the Lebesgue dominated convergence theorem, and the definitions of w_k and W imply

$$\begin{aligned}
\int_{\mathbb{R}^n} u_0(x; N) \exp(-\nu|x|)\overline{\psi(x)}\, dx &= \int_{\|t\| \leqq N} \int_{\mathbb{R}^n} w(x, t) \exp(-\nu|x|)\overline{\psi(x)}\, dx\, dt \\
&= \lim_{k\to\infty} \int_{\|t\| \leqq N} \int_{\mathbb{R}^n} w_k(x, t) \exp(-\nu|x|)\overline{\psi(x)}\, dx \\
&= \int_{\mathbb{R}^n} W(x) \exp(-\nu|x|)\overline{\psi(x)}\, dx. \qquad \square
\end{aligned}$$

REFERENCES

[1] J. D. ARCHENBACH, *A Theory of Elasticity with Microstructure for Directionally Reinforced Composites*, Springer-Verlag, Berlin, New York, 1975.

[2] I. BABUŠKA, *Homogenization and its applications: Mathematical and computational problems*, in Numerical Solution of Partial Differential Equations III, SYNSPADE 1975, B. Hubbard, ed., Academic Press, New York, pp. 89–116.

[3a] ———, *Solution of interface problems by homogenization* I, SIAM J. Math. Anal., 7 (1976), pp. 603–634.

[3b] ———, *Solution of interface problems by homogenization* II, SIAM J. Math. Anal., 7 (1976), pp. 635–645.

[3c] ———, *Solution of interface problems by homogenization* III, SIAM J. Math. Anal., 8 (1977), pp. 923–937.

[4] I. BABUŠKA AND A. K. AZIZ, *Survey lectures on the mathematical foundation of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1973, pp. 5–359.

[5] I. BABUŠKA AND R. C. MORGAN, *Composites with a periodic structure: mathematical analysis and numerical treatment*, Comput. Math. Appl., 11 (1985), pp. 995–1005.

[6] I. BABUŠKA AND J. E. OSBORN, *Finite element methods for the solution of problems with rough input data*, in Lecture Notes in Math. 1121, A. Dold and B. Echman, eds., Springer-Verlag, Berlin, New York, 1983, pp. 1–18.

[7] ———, *Generalized finite element methods, their performance and their relation to mixed methods*, SIAM J. Numer. Anal., 20 (1983), pp. 510–536.

[8] I. BABUŠKA, G. CALOZ, AND J. E. OSBORN, *Special finite elements for a class of second order elliptic problems with rough coefficients*, to appear.

[9] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, New York, 1978.

[10] J. M. BURGERS, *On a problem of homogenization*, Quart. Appl. Math., 35 (1978), pp. 421–438.

[11] C. C. CHAMS AND G. P. SENDECKYJ, *Critique on theories predicting thermoelastic properties of fibrous composites*, J. Composite Materials, 2 (1968), pp. 332–358.

[12] S. K. GARG, V. SVALBONAS, AND G. A. GURTMAN, *Analysis of Structural Composite Materials*, Marcel Dekker, New York, 1973.

[13] Z. HASHIN, *Theory of Fiber Reinforced Materials*, NASA Report NASA CR-1974, Washington, D.C., 1972.

[14] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vol. I, Dunod, Paris, 1968.

[15] R. C. MORGAN, *Mathematical aspects and computational considerations in the theory of homogenization*, Ph.D. thesis, University of Maryland, College Park, MD, 1982.

[16] R. C. MORGAN AND I. BABUŠKA, *An approach for constructing families of homogenized equations for periodic media.* II: *Properties of the kernel*, SIAM J. Math. Anal., this issue (1991), pp. 16–33.

[17] F. ODEH AND J. B. KELLER, *Partial differential equations with periodic coefficients and Bloch norms in crystals*, J. Math. Phys., 5 (1964), pp. 1499–1504.

[18] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[19] G. P. SENDECKYJ, ED., *Composite Materials*, Vol. 2, Mechanics of Composite Materials, L. J. Broutman and R. H. Krock, eds., Academic Press, New York, London, 1974.

[20] K. YOSIDA, *Functional Analysis*, Fourth edition, Springer-Verlag, Berlin, New York, 1974.

[21] V. V. ŽHIKOV, S. M. KOZLOV, O. A. OLEINIK, AND KHA T'EN NGOAN, *Averaging and G-convergence of differential operators*, Russ. Math. Surveys, 34 (1979), pp. 69–147.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 16-33, January 1991

002

AN APPROACH FOR CONSTRUCTING FAMILIES OF HOMOGENIZED EQUATIONS FOR PERIODIC MEDIA. II: PROPERTIES OF THE KERNEL*

R. C. MORGAN† AND I. BABUŠKA†

Abstract. This paper is the second in the series devoted to the study of constructions of families of homogenizations. The first paper [*SIAM J. Math. Anal.*, 22 (1991), pp. 1-15] used the properties of the kernel $\Phi(\cdot, h, t)$. In this paper these properties are established.

Key words. homogenization, composite materials

AMS(MOS) subject classifications. 35J05, 35JPP, 73K20

1. Introduction. In [8] we developed an integral representation of the solution to a differential equation that models the equations that arise in the study of periodic media (e.g., composite materials). The elliptic differential equation studied in [8] is

$$-\sum_{p,q=1}^{n} \frac{\partial}{\partial x_p}\left(a_{pq}\left(\frac{x}{h}\right)\frac{\partial u^h}{\partial x_q}(x)\right)+a_0\left(\frac{x}{h}\right)u^h(x)=f(x)$$

on $\mathbb{R}^n$, in which a_{pq} and a_0 are real-valued 2π-periodic functions and h is a given positive number. An alternate proof of the classical homogenization result (the limit of u^h as $h\to 0$) is given in [8], based on the integral formula for u^h that was developed there.

The integral representation of u^h depends on the 2π-periodic function $\phi(\cdot, h, t)$ that satisfies

$$-e^{-iht\cdot y}\sum_{p,q=1}^{n}\frac{\partial}{\partial y_p}\left(a_{pq}(y)\frac{\partial}{\partial y_q}(\phi(y,h,t)\,e^{iht\cdot y})\right)+h^2a_0(y)\phi(y,h,t)=h^2$$

on $\{y\in\mathbb{R}^n: |y_p|<\pi\}$, in which $t\in\mathbb{R}^n$. The main emphasis of [8] is placed on the function u^h; the properties of ϕ that are needed there are stated without proof. This paper presents an analysis of ϕ in order to prove these claims, namely, Lemmas 1 and 12 in [8]. Theorem 1 in this paper is equivalent to Lemma 1 in [8], whereas the content of Lemma 12 and the discussion preceding it in [8] are contained in Theorems 9 and 11 here.

In § 2, the notation used here and the equation that $\phi(\cdot, h, t)$ satisfies are given along with the statement of Theorem 1. The proof of Theorem 1 is presented in § 3. The expansion of $\phi(\cdot, h, t)$ in powers of h and properties of this expansion are developed in § 4. Section 5 is devoted to developing several analyticity results associated with families of sesquilinear forms. The results of § 5 are used extensively in § 3. Additional details and references can be found in [7].

A method for systematically developing classes of differential equations, or even pseudodifferential operators, that describe the behavior of composite materials with a periodic structure has been introduced in [3] and is based on the results of [8] and this paper.

* Received by the editors May 23, 1989; accepted for publication (in revised form) February 6, 1990. This research was partially supported by Office of Naval Research contract N00014-85-K-0169.

† Institute for Physical Science and Technology and Mathematics Department, University of Maryland, College Park, Maryland 20742.

2. Notation and statement of the problem. Let $S \equiv \{y \equiv (y_1, \cdots, y_n) \in \mathbb{R}^n : |y_k| < \pi$ for $k = 1, \cdots, n\}$, and for $j = 0, 1$, denote the standard Sobolev norm on S by $\|\cdot\|_j$. In addition, define $|\cdot|_1$ by

$$|v_1| \equiv \left(\int_S \sum_{p=1}^{n} \left| \frac{\partial v}{\partial y_p}(y) \right|^2 dy \right)^{1/2}.$$

The Sobolev spaces of periodic functions for which S is the fundamental period, is denoted by $H^j_{\text{per}}(S)$, and is defined to be the completion with respect to $\|\cdot\|_j$, of the complex-valued, C^∞-functions on $\mathbb{R}^n$ that are 2π-periodic in each coordinate variable.

Let a_{pq}, for $p, q = 1, \cdots, n$, and a_0 be real valued, 2π-periodic, with L_∞-functions defined on $\mathbb{R}^n$. Furthermore, assume $a_{qp} = a_{pq}$ and that there exist positive constants γ_0 and γ_1 such that

$$\begin{aligned} &a_0(x) \geqq \gamma_0, \\ &\sum_{p,q=1}^{n} a_{qp}(x)\zeta_q\bar{\zeta}_p \geqq \gamma_1 \sum_{p=1}^{n} |\zeta_p|^2 \quad \text{for all } \zeta_p \in \mathbb{C}, \end{aligned} \tag{1}$$

almost everywhere on $\mathbb{R}^n$. For each $h \in \mathbb{C}$ and $t \in \mathbb{C}^n$, define the sesquilinear form $\Phi(h, t) : H^1_{\text{per}}(S) \times H^1_{\text{per}}(S) \to \mathbb{C}$ by

$$\begin{aligned} &\Phi(h, t)[\phi, v] \\ &= \int_S \left\{ \sum_{p,q=1}^{n} a_{pq}(y) \frac{\partial}{\partial y_q}(\phi(y)\, e^{iht\cdot y}) \frac{\partial}{\partial y_q}(\overline{v(y)}\, e^{-iht\cdot y}) + h^2 a_0(y)\phi(y)\overline{v(y)} \right\} dy. \end{aligned} \tag{2}$$

In § 3 we will prove Theorem 1 below.

THEOREM 1. *There exists a neighborhood $\hat{G}$ of $\mathbb{R}^{n+1}$ (contained in $\mathbb{C}^{n+1}$), such that a unique function $\phi(\cdot, h, t) \in H^1_{\text{per}}(S)$ exists for each $(h, t) \in \hat{G}$ and satisfies*

$$\Phi(h, t)[\phi(\cdot, h, t), v] = h^2 \int_S \overline{v(y)}\, dy \quad \text{for all } v \in H^1_{\text{per}}(S). \tag{3}$$

Furthermore, the mapping

$$(h, t) \mapsto \phi(\cdot, h, t) \in H^1_{\text{per}}(S) \tag{4}$$

is holomorphic on $\hat{G}$ (see Definition 12 in § 5).

In the proof of Theorem 1, the following eigenvalue problem will be considered: Seek $\lambda(h, t) \in \mathbb{C}$ and a nonzero function $\psi(\cdot, h, t) \in H^1_{\text{per}}(S)$ such that

$$\Phi(h, t)[\psi(\cdot, h, t), v] = \lambda(h, t) \int_S \psi(y, h, t)\overline{v(y)}\, dy \quad \text{for all } v \in H^1_{\text{per}}(S). \tag{5}$$

Before proceeding to the proof of Theorem 1, we give two lemmas that we will use repeatedly.

LEMMA 2. *A constant C_0 exists such that*

$$\frac{1}{C_0(1 + \|t\|)} \|v\|_1 \leqq \|v\, e^{it\cdot y}\|_1 \leqq C_0(1 + \|t\|)\|v\|_1$$

for all $v \in H^1(S)$, the standard Sobolev space, and for all $t \in \mathbb{R}^n$, where $\|t\|^2 \equiv t_1^2 + \cdots + t_n^2$.

Proof. The proof of the right-hand inequality is straightforward. The inequality on the left is proved by applying the right-hand inequality to the function $w = v\, e^{it\cdot y}$:

$$\|v\|_1 = \|w\, e^{-it\cdot y}\|_1 \leqq C_0(1 + \|t\|)\|w\|_1. \qquad \square$$

LEMMA 3. *Let H be a complex Hilbert space with norm $\|\cdot\|_H$ and inner product $(\cdot,\cdot)_H$, and let $\Phi: H\times H\to\mathbb{C}$ be a sesquilinear form (i.e., $\Phi[\phi, v]$ is linear in ϕ and conjugate-linear in v). If there exist constants M and γ such that*

$$|\Phi[\phi, v]|\leqq M\|\phi\|_H\|v\|_H,$$

$$\gamma\|v\|_H^2\leqq|\phi[v, v]|$$

for all ϕ and v in H, then for each $f\in H^*$, the space of bounded conjugate-linear functionals on H, there is a unique $\phi\in H$ such that

$$\Phi[\phi, v]=f(v)\quad \text{for all } v\in H.$$

Moreover, $\|\phi\|_H\leqq(1/\gamma)\|f\|_{H^*}$.

Lemma 3 is known as the Lax–Milgram theorem (see [2]). The essence of the proof of Lemma 3 is the existence of a bounded operator A that maps H isomorphically onto H such that $\Phi[\phi, v]=(A\phi, v)_H$ for all ϕ and v in H. This fact will be used in the proof of Theorem 15 in § 5.

3. Proof of Theorem 1. The ideas and results of § 5 will be used extensively in this section. Note that $H^0_{\text{per}}(S)$ and $H^1_{\text{per}}(S)$ satisfy the conditions imposed on H and V, respectively, in § 5, i.e., $H^1_{\text{per}}(S)$ is a continuously, densely, and compactly embedded subspace of $H^0_{\text{per}}(S)$. (A discussion of spaces of periodic functions is contained in [1].) Clearly, $(h, t)\in\mathbb{C}^{n+1}\mapsto\Phi(h, t)[\phi, v]\in\mathbb{C}$ is an analytic function for each ϕ and v in $H^1_{\text{per}}(S)$.

We start here by determining an open set $G\subset\mathbb{C}^{n+1}$ such that $\mathbb{R}^{n+1}\subset G$ and such that $\Phi(h, t)$ satisfies inequalities similar to (34) and (35) for each $(h, t)\in G$. Then we will show that, in the sense of (5), zero is not an eigenvalue of $\Phi(h, t)$ when $h\neq 0$ and $(h, t)\in\mathbb{R}^{n+1}$, but that zero is a simple eigenvalue of $\Phi(0, t)$. The conclusions of Theorems 19 and 20 in § 5 get us part of the way through the proof of Theorem 1; we must investigate further the eigenvalue problem associated to $\Phi(0, t)$.

LEMMA 4. *There exist an open set $G\subset\mathbb{C}^{n+1}$ and real-valued functions M, γ, and μ such that μ is continuous on G and for each $(h, t)\in G$, $M(h, t)>0$, $\gamma(h, t)>0$*

(i) $|\Phi(h, t)[\phi, v]|\leqq M(h, t)\|\phi\|_1\|v\|_1$ *for all ϕ and v in $H^1_{\text{per}}(S)$,*

(ii) $\gamma(h, t)\|v\|_1^2\leqq\text{Re}\,(\Phi(h, t)[v, v])+\mu(h, t)\|v\|_0^2$ *for all $v\in H^1_{\text{per}}(S)$.*

Moreover, G can be chosen so that $\mathbb{R}^{n+1}\subset G$ and so that $(\bar h, \bar t)\in G$ whenever $(h, t)\in G$.

Proof. For each $z\in\mathbb{C}^n$ and ϕ and v in $H^1_{\text{per}}(S)$, define the sesquilinear form

$$B(z)[\phi, v]=\int_S\sum_{p,q=1}^n a_{pq}(y)\frac{\partial}{\partial y_q}(\phi(y)\,e^{iz\cdot y})\frac{\partial}{\partial y_q}(\overline{v(y)}\,e^{-iz\cdot y})\,dy;$$

thus, $\Phi(h, t)[\phi, v]=B(ht)[\phi, v]+h^2\int_S a_0(y)\phi(y)\overline{v(y)}\,dy$. Whenever it is convenient in this proof, we will use $ht=\rho+i\sigma$, where ρ and σ are real n-vectors. Defining

$$\begin{aligned}F(\rho, \sigma)[\phi, v]&=B(\rho+i\sigma)[\phi, v]-B(\rho)[\phi, v]\\&=\int_S\sum_{p,q=1}^n a_{pq}(y)\Big\{\sigma_p\frac{\partial}{\partial y_q}(\phi(y)\,e^{i\rho\cdot y})\overline{v(y)}\,e^{-i\rho\cdot y}\\&\qquad-\sigma_q\phi(y)\,e^{i\rho\cdot y}\frac{\partial}{\partial y_q}(\overline{v(y)}\,e^{-i\rho\cdot y})-\sigma_q\sigma_p\phi(y)\overline{v(y)}\Big\}\,dy\end{aligned}$$

it follows, since each a_{pq} is an L_∞-function, that there exists a constant K such that

$$|F(\rho, \sigma)[\phi, v]|\leqq K\|\sigma\|(1+\|\sigma\|)\|\phi\,e^{i\rho\cdot y}\|_1\|v\,e^{i\rho\cdot y}\|_1.\tag{6}$$

In addition, there is a constant K' such that

$$|B(\rho)[\phi, v]| \leqq K'|\phi\, e^{i\rho\cdot y}|_1 |v\, e^{i\rho\cdot y}|_1.$$

Using Lemma 2 and the fact that a_0 is bounded, it follows that there exists a positive number $M(h, t)$ for each (h, t) $\mathbb{C}^{n+1}$ such that (i) is true.

Next, a consequence of (1) is that

$$B(\rho)[v, v] \geqq \gamma_1 |v\, e^{i\rho\cdot y}|_1^2$$

and therefore

$$\text{(7)} \qquad \begin{aligned} \operatorname{Re}(B(\rho+i\sigma)[v, v]) + \gamma_1\|v\|_0^2 &= B(\rho)[v, v] + \operatorname{Re}(F(\rho, \sigma)[v, v]) + \gamma_1\|v\|_0^2 \\ &\geqq (\gamma_1 - K\|\sigma\|(1+\|\sigma\|))\|v\, e^{i\rho\cdot y}\|_1^2. \end{aligned}$$

We now define G by

$$G \equiv \left\{(h, t) \in \mathbb{C}^{n+1}\colon K\|\operatorname{Im}(ht)\|(1+\|\operatorname{Im}(ht)\|) < \frac{\gamma_1}{2}\right\}$$

and set

$$\gamma(h, t) = \frac{\gamma_1}{2C_0^2(1+\|\operatorname{Re}(ht)\|)^2}$$

in which C_0 is defined in Lemma 2. Finally, setting $K'' = \|a_0\|_{L_\infty(S)}$ and using (7) and Lemma 2, we conclude that

$$\begin{aligned} \operatorname{Re}(\Phi(h, t)[v, v]) + \gamma_1\|v\|_0^2 &= \operatorname{Re}(B(ht)[v, v]) + \operatorname{Re}\left(h^2\int_S a_0(y)|v(y)|^2\, dy\right) + \gamma_1\|v\|_0^2 \\ &\leqq \gamma(h, t)\|v\|_1^2 - K''|h|^2\|v\|_0^2 \end{aligned}$$

for $(h, t) \in G$, which yields (ii) with $\mu(h, t) = K''|h|^2 + \gamma_1$. □

For each $(h, t) \in G$, we can associate to $\Phi(h, t)$ a closed operator $T(h, t)$ as in Theorem 15:

$$\Phi(h, t)[\phi, v] = (T(h, t)\phi, v)_{H^0_{\text{per}}(S)} \quad \text{for all } v \in H^1_{\text{per}}(S)$$

and for all ϕ in the domain of $T(h, t)$, which is a dense subspace of $H^1_{\text{per}}(S)$. Reference is made to $T(h, t)$ in the next few paragraphs in order to draw upon the results of § 5.

For each $(h, t) \in \mathbb{R}^{n+1}$, a direct consequence of (1) and Lemma 2 is

$$\text{(8)} \qquad \Phi(h, t)[v] \geqq \gamma_1 |v\, e^{iht\cdot y}|_1^2 + \gamma_0 h^2\|v\|_0^2 \geqq \frac{\min\{\gamma_0 h^2, \gamma_1\}}{2C_0^2(1+\|(ht\|)^2}\|v\|_1^2$$

for all $v \in H^1_{\text{per}}(S)$. If $(h, t) \in \mathbb{R}^{n+1}$ and $h \neq 0$, then the hypotheses of Lemma 3 are satisfied; hence $\Phi(h, t)[\phi, v] = \int_S w(y)\overline{v(y)}\, dy$ for all $v \in H^1_{\text{per}}(S)$, is uniquely solvable for each $w \in H^0_{\text{per}}(S)$. Consequently, zero is in the resolvent set of $T(h, t)$, and by Theorem 20, $\{(h, t) \in \mathbb{R}^{n+1}\colon h \neq 0\}$ is contained in an open set in $\mathbb{C}^{n+1}$ on which (4) is holomorphic.

Now let $h = 0$. For any $\tau \in \mathbb{R}^n$ (or $\mathbb{C}^n$), we have from the first inequality in (8) that

$$\Phi(0, \tau)[v, v] \geqq \gamma_1 |v|_1^2.$$

If zero is an eigenvalue of $\Phi(0, \tau)$ with $\psi_0 \in H^1_{\text{per}}(S)$ being an associated eigenfunction, then

$$0 = \Phi(0, \tau)[\psi_0, \psi_0] \geqq \gamma_1 |\psi_0|_1,$$

and, consequently, ψ_0 is a constant function, which depends on τ. Therefore, zero is an eigenvalue of (the associated closed operator) $T(0, \tau)$, and ψ_0 is the only eigenfunction associated to zero. There are no generalized eigenfunctions when $\tau \in \mathbb{R}^n$ because the identity $\Phi(0, \tau)[v, \phi] = \overline{\Phi(0, \tau)[\varphi, v]}$ implies, by Corollary 18, that $T(0, \tau)$ is self-adjoint.

The following conclusions can now be drawn from Theorem 20. For each $\tau \in \mathbb{R}^n$, there exists a neighborhood $G_\tau \subset G$ of $(0, \tau)$, and there exists a complex-valued function λ, analytic on G_τ, such that $\lambda(0, t) = 0$ for $(0, t) \in G_\lambda$ and such that $\lambda(h, t)$ is a simple eigenvalue of $\Phi(h, t)$ when $(h, t) \in G$. Furthermore, there exist two holomorphic functions $(h, t) \in G_\tau \mapsto P(h, t)$ and $(h, t) \in G_\tau \mapsto R_2(0, h, t)$, with values in the space of bounded linear operators that map $H^0_{\text{per}}(S)$ into $H^{(1)}_{\text{per}}(S)$, such that $P(h, t)$ projects $H^0_{\text{per}}(S)$ onto the one-dimensional eigenspace spanned by the eigenvector associated to $\lambda(h, t)$, and such that

$$\phi(\cdot, h, t) = \frac{h^2}{\lambda(h, t)} P(h, t)\mathbf{1} + h^2 R_2(0, h, t)\mathbf{1} \tag{9}$$

for each $(h, t) \in G_\tau$ for which $\lambda(h, t) \neq 0$. Here $\mathbf{1}$ is the constant function that takes on the value 1 on S. Note that $\lambda(h, t)$ is not identically zero on G_τ since zero is not an eigenvalue of $\Phi(h, \tau)$ when h is a nonzero real number; therefore, this representation of $\phi(\cdot, h, t)$ is meaningful.

Clearly, the holomorphy of $(h, t) \mapsto \phi(\cdot, h, t)$ on G_τ is completely determined by that of $h^2/\lambda(h, t)$. We will show that $h^2/\lambda(h, t)$ has an analytic continuation to $h = 0$, and that there is a suitable restriction G'_τ of G_τ, on which $\lambda(h, t) = 0$ if and only if $h = 0$. This will be done by determining part of the Taylor expansion of $\lambda(h, t)$ in powers of h about $h = 0$, where it will be seen that the coefficient of h^k is zero for $k = 0, 1$.

For the moment, we assume the existence of a function ω, analytic on G_τ, such that

$$\lambda(h, t) = h^2 \omega(h, t) \quad \text{for } (h, t) \in G_\tau. \tag{10}$$

Letting $\psi(\cdot, h, t) \in H^1_{\text{per}}(S)$ be the eigenfunction associated to $\lambda(h, t)$ as in (5), the first inequality in (8) yields

$$\gamma_0 h^2 \|(\cdot, h, t)\|_0^2 \leqq \Phi(h, t)[\psi(\cdot, h, t), \psi(\cdot, h, t)] = \lambda(h, t)\|\psi(\cdot, h, t)\|_0^2$$

for $(h, t) \in G_\tau \cap \mathbb{R}^{n+1}$. Consequently, $\omega(h, t) \geqq \gamma_0$ for $(h, t) \in G_\tau \cap \mathbb{R}^{n+1}$, and by continuity, there is an open set $G'_\tau \subset G_\tau$ such that $(G_\tau \cap \mathbb{R}^{n+1}) \subset G'_\tau$ and $\operatorname{Re}(\omega(h, t)) > 0$ for $(h, t) \in G'_\tau$. Thus $(h, t) \mapsto h^2/\lambda(h, t) = 1/\omega(h, t)$ is analytic on G'_τ, from which it follows that (9) is holomorphic on G'_τ as well. Assuming that (10) is valid, Theorem 1 is proven.

In the process of determining the Taylor expansion of $\lambda(h, t)$ in powers of h about $h = 0$, we develop a similar expansion of the eigenfunction $\psi(\cdot, h, t)$. In § 5 (see (45)), we show that $\psi(\cdot, h, t)$ can be chosen so that it depends holomorphically on $(h, t) \in G_\tau$. The Taylor expansion of $\lambda(h, t)$ and $\psi(\cdot, h, t)$ in powers of h will be obtained next, by expanding the eigenvalue equation (5) in powers of h and equating the coefficients of like powers.

Upon setting

$$\Phi_0[\phi, v] \equiv \int_S \sum_{p,q=1}^{n} a_{pq}(y) \frac{\partial \phi}{\partial y_q}(y) \overline{\frac{\partial \phi}{\partial y_p}(y)}\, dy,$$

$$\Phi_1(t)[\phi, v] \equiv i \int_S \sum_{p,q=1}^{n} a_{pq}(y) \left(\phi(y) \overline{\frac{\partial v}{\partial y_p}(y)} t_q - \frac{\partial \phi}{\partial y_q}(y) v(y) t_p \right) dy, \tag{11}$$

$$\Phi_2(t)[\phi, v] \equiv \int_S \left(\sum_{p,q=1}^{n} a_{pq}(y) t_q t_p + a_0(y) \right) \phi(y) \overline{v(y)}\, dy$$

for all ϕ and v in $H^1_{\text{per}}(S)$ and for all $t \in \mathbb{C}^n$, it follows from the definition of $\Phi(h, t)$ that

$$\Phi(h, t) = \Phi_0 + \Phi_1(t)h + \Phi_2(t)h^2. \tag{12}$$

Substituting the expansions $\lambda(h, t) = \sum_{k=0}^{n} \lambda_k(t) h^k$, $\psi(\cdot, h, t) = \sum_{k=0}^{n} \psi(\cdot, t) h^k$, and (12) into (5), equating like powers of h, and noting that $\lambda_0(t) = \lambda(0, t) = 0$, we derive the following system of equations:

$$\Phi_0[\psi_k(\cdot, t), v] = \begin{cases} 0, & k = 0, \\ \lambda_1(t) \displaystyle\int_S \psi_0(y, t) \overline{v(y)}\, dy - \Phi_1(t)[\psi_0(\cdot, t), v], & k = 1, \\ \displaystyle\sum_{l=1}^{k} \lambda_l(t) \int_S \Psi_{k-l}(y, t) \overline{v(y)}\, dy & \\ \quad - \displaystyle\sum_{l=1}^{z} \Phi_l(t)[\psi_{k-l}(\cdot, t), v], & k \geqq 2, \end{cases} \tag{13}$$

for all $v \in H^1_{\text{per}}(S)$. Successively solving the equations in (13) will yield formulas for the $\lambda_k(t)$ and $\psi_k(\cdot, t)$.

The method of solving (13) will be based on Lemma 6 below. First note that

$$\Phi_0[\phi, 1] = \Phi_0[1, \phi] = \Phi_1(t)[1, 1] = 0 \tag{14}$$

for all $\phi \in H^1_{\text{per}}(S)$ and for each $t \in \mathbb{C}^n$.

DEFINITION 5.

$$W \equiv \left\{ v \in H^1_{\text{per}}(S) : \int_S v(y)\, dy = 0 \right\}.$$

Each function $H^1_{\text{per}}(S)$ can be represented uniquely as the sum of a constant function and a function in W.

LEMMA 6. *$|\cdot|_1$ is a norm on W, and there exists a constant K such that*

$$|\Phi_0[\phi, v]| \leqq K |\phi|_1 |v|_1, \qquad \Phi_0[v, v] \geqq \gamma_1 |v|_1^2$$

for all ϕ and v in W.

Each equation in (13) is of the form

$$\Phi_0[\psi_k, v] = F_k(v) \quad \text{for all } v \in H^1_{\text{per}}(S),$$

in which F_k is a conjugate-linear form on $H^1_{\text{per}}(S)$ which depends on t; $\lambda_1(t), \cdots, \lambda_k(t)$; and $\psi_0(\cdot, t), \cdots, \psi_{k-1}(\cdot, t)$. Furthermore, F_k is bounded on $H^1_{\text{per}}(S)$. It is also bounded on W, as a result of the closed graph theorem (see (31)). A consequence of (14) is that we must ensure that $F_k(1) = 0$, which will determine $\lambda_k(t)$ uniquely. Then Lemmas 3 and 6 imply that $\psi_k(\cdot, t)$ is determined uniquely as an element in W, that is, up to an additive constant. However, the arbitrary constants in $\psi_k(\cdot, t)$ will remain essentially

arbitrary because, being an eigenfunction, $\psi(\cdot, h, t)$ is uniquely determined up to a multiplicative constant only.

THEOREM 7. *For $k \geqq 1$, define functions $\tilde{\chi}_k(\cdot, t) \in W$ according to*

$$\Phi_0[\tilde{\chi}_k(\cdot, t), v] = \begin{cases} -\Phi_1(t)[1, v], & k = 1, \\ -\Phi_1(t)[\tilde{\chi}_1(\cdot, t), v] - \Phi_2(t)[1, v], & k = 2, \\ \sum_{j=2}^{k-1} \lambda_j(t) \int_S \tilde{\chi}_{k-j}(y, t)\overline{v(y)}\, dy & \\ \quad - \sum_{j=1}^{2} \Phi_j(t)[\tilde{\chi}_{k-j}(\cdot, t), v], & k \geqq 3, \end{cases} \tag{15}$$

for all $v \in W$, where

$$\lambda_l(t) = \begin{cases} 0, & l = 1, \\ \dfrac{1}{(2\pi)^n}(\Phi_1(t)[\tilde{\chi}_1(\cdot, t), 1] + \Phi_2(t)[1, 1]), & l = 2, \\ \dfrac{1}{(2\pi)^n}(\Phi_1(t)[\tilde{\chi}_{l-1}(\cdot, t), 1] + \Phi_2(t)[\tilde{\chi}_{l-2}(\cdot, t), 1]), & l \geqq 3. \end{cases} \tag{16}$$

Next, define $\psi_l(\cdot, t) \in H^1_{\text{per}}(S)$ for $l \geqq 0$ by

$$\psi_l(\cdot, t) = \begin{cases} f_0(t), & l = 0, \\ \sum_{j=0}^{l-1} f_j(t)\tilde{\chi}_{l-j}(\cdot, t) + f_l(t), & l \geqq 1, \end{cases} \tag{17}$$

in which each f_l is a holomorphic function (in a neighborhood of $t = \tau$) and $f_0(t) \neq 0$ for all t. Then (16) *and* (17) *solve* (13).

Proof. Upon solving (13) with $k = 0$, we have that $\psi_0(\cdot, t)$ must be a constant function, which we denote by $f_0(t)$ and which we assume is nonzero for each t because $f_0(t) - \psi_0(\cdot, t) = \psi(\cdot, 0, t)$ is an eigenfunction.

In order to simplify notation, we will drop the dependence on t in the remainder of the proof.

For $k = 1$, (13) becomes

$$\Phi_0[\psi_1, v] = \lambda_1 f_0 \int_S v(y)\, dy - f_0\Phi_1[1, v] \quad \text{for all } v \in H^1_{\text{per}}(S).$$

Setting $v = 1$ and using (14) yields $0 = \lambda_1 f_0 (2\pi)^n$, which implies $\lambda_1 = 0$ because $f_0 \neq 0$. Now (13) with $k = 1$ can be reduced to

$$\Phi_0[\psi_1, v] = -f_0\Phi_1[1, v] \quad \text{for all } v \in W,$$

the solution of which is given by (17) with $l = 1$.

Substituting (16) and (17) into (13) with $k = 2$ yields

$$\Phi_0[\psi_2, v] = \lambda_2 f_0 \int_S \overline{v(y)}\, dy - (f_0\Phi_1[\tilde{\chi}_1, 1] + f_1\Phi_1[1, v]) - f_0\Phi_2[1, v]$$

for all $v \in H^1_{\text{per}}(S)$. Again, set $v = 1$ and use (14) to obtain $0 = \lambda_2 f_0 (2\pi)^n - f_0(\Phi_1[\tilde{\chi}_1, 1] + \Phi_1[1, 1])$. Since $f_0 \neq 0$, the solution to this equation is given by (16). Now (13) with $k = 2$ can be reduced to

$$\Phi_0[\psi_2, v] = -f_0(\Phi_1[\tilde{\chi}_1, v] + \Phi_1[1, v]) - f_1\Phi_1[1, v] \quad \text{for all } v \in W,$$

and the solution to this equation is given in (17).

An induction argument, similar to the one we will use in the proof of Theorem 8, establishes the remaining formulas in (16) and (17). □

As we noted earlier, the arbitrary nature of the constants $f_l(t)$ can be traced to the fact that the eigenfunction $\psi(\cdot, h, t)$ is determined uniquely, up to a multiplicative constant, only. To see this, let $c(h, t) = \sum_{k=0}^{\infty} c_k(t)h^k$ be analytic in h and t, with $c_0(t) \neq 0$, and define

$$f_k^*(t) = \sum_{l=0}^{k} p^{-l} c_l(t) f_{k-l}(t), \qquad k \geq 0,$$

$$\psi_k^*(\cdot, t) = \begin{cases} f_0^*(t), & k = 0, \\ \sum_{j=0}^{k-1} f_j^*(t)\tilde{\chi}_{k-j}(\cdot, t) + f_k^*(t), & k \geq 1. \end{cases}$$

Then $\sum_{k=0}^{\infty} \psi_k^*(\cdot, t)h^k$ is the power series expansion of $c(h, t)\psi(\cdot, h, t)$; clearly the form of ψ_k^* is the same as that of ψ_k.

We have shown that (10) is correct, and consequently, the proof of Theorem 1 is complete.

4. Expansion in powers of h. The expansion of $\phi(\cdot, h, t)$ in powers of h can be determined formally by expanding (3) in powers of h and equating like powers. This formal process is valid because $h \mapsto \phi(\cdot, h, t) \in H^1_{\text{per}}(S)$ is holomorphic at $h = 0$ for each $t \in \mathbb{C}^n$ such that $(0, t) \in \hat{G}$ (see Theorem 1), and because $\Phi(h, t)[\phi, v]$ is a polynomial in h and t for fixed ϕ and v in $H^1_{\text{per}}(S)$. By substituting

$$\phi(\cdot, h, t) = \sum_{k=0}^{\infty} \phi_k(\cdot, t)h^k \tag{18}$$

and (12) into (3), we obtain the following system of equations for the coefficients $\phi_k(\cdot, t)$.

$$\Phi_0[\phi_k(\cdot, t), v] = \begin{cases} 0, & k = 0, \\ \Phi_1(t)[\phi_0(\cdot, t), v], & k = 1, \\ \int_S v(y)\, dy - \Phi_1(t)[\phi_{k-1}(\cdot, t), v] - \Phi_2(t)[\phi_{k-2}(\cdot, t), v], & k = 2, \\ -\Phi_1(t)[\phi_{k-1}(\cdot, t), v] - \Phi_2(t)[\phi_{k-2}(\cdot, t), v], & k \geq 3, \end{cases} \tag{19}$$

for all $v \in H^1_{\text{per}}(S)$. The radius of convergence of (18) depends on t, and each coefficient $\phi_k(\cdot, t)$ is in $H^1_{\text{per}}(S)$ and depends holomorphically on t.

The method of determining each ϕ_k is similar to that used in the proof of Theorem 7 to determine the Taylor expansions of $h \mapsto \lambda(h, t)$ and $h \mapsto \psi(\cdot, h, t)$. Recall that W is the subspace of $H^1_{\text{per}}(S)$ of functions that have an average value of zero. A consequence of (14) is that the right-hand side of (19) must be equal to zero when $v = 1$. On the other hand, restricting v to be in W, Lemmas 3 and 6, applied to (19), uniquely determine $\phi_k(\cdot, t)$ as an element in W (i.e., up to an additive constant), in terms of $\phi_{k-1}(\cdot, t)$ and $\phi_{k-2}(\cdot, t)$. Then $\phi_k(\cdot, t)$ becomes uniquely defined as an element in $H^1_{\text{per}}(S)$ by requiring the right-hand side of the equation for $\phi_{k+2}(\cdot, t)$ in (19) to be zero when $v = 1$.

THEOREM 8. *For each $k \geq 1$ define $\chi_k(\cdot, t) \in W$ to be the solution of*

$$\Phi_0[\chi_k(\cdot, t), v] = \begin{cases} -\Phi_1(t)[1, v], & k = 1, \\ -\Phi_1(t)[\chi_1(\cdot, t), v] - \Phi_2(t)[1, v], & k = 2, \\ -\Phi_1(t)[\chi_{k-1}(\cdot, t), v] - \Phi_2(t)[\chi_{k-2}(\cdot, t), v], & k \geq 3 \end{cases} \tag{20}$$

for all $v \in W$, and for each $k \geqq 0$ define $g_k(t) \in \mathbb{C}$ by

(21)
$$g_k(t) = \begin{cases} \dfrac{(2\pi)^n}{\Phi_1(t)[\chi_1(\cdot, t), 1] + \Phi_2(t)[1, 1]}, & k = 0, \\ -\dfrac{g_0(t)}{(2\pi)^n} \sum_{j=0}^{k-1} g_j(t)(\Phi_1(t)[\chi_{k+1-j}(\cdot, t), 1] + \Phi_2(t)[\chi_{k-j}(\cdot, t), 1]), & k \geqq 1. \end{cases}$$

Then the coefficient of h^k in (18) *is given by*

(22)
$$\phi_k(\cdot, t) = \begin{cases} g_0(t), & k = 0, \\ \sum_{j=0}^{k-1} g_j(t)\chi_{k-j}(\cdot, t) + g_k(t), & k \geqq 1. \end{cases}$$

Proof. Throughout this proof we will suppress all dependence upon t. First, each χ_k is well defined in W by (20) because of Lemmas 3 and 6 and because each right-hand side in (20) is a bounded conjugate-linear form on W (see (30) and (31)).

It follows immediately from (19) with $k = 0$ that ϕ_0 is a constant function, which we denote by g_0. For $k = 1$ in (19), we now have

$$\Phi_0[\phi_1, v] = -g_0\phi_1[1, v] \quad \text{for all } v \in H^1_{\text{per}}(S).$$

Since $\Phi_1[1, 1] = 0$ (see (14)), the solution ϕ_1 has the form of (22); χ_1 is defined by (20), but g_0 and g_1 are arbitrary constants.

We next consider $k = 2$ in (19), and substituting (22), we obtain

(23)
$$\Phi_0[\phi_2, v] = \int_S \overline{v(y)}\, dy - (g_0\Phi_1[\chi_1, v] + g_1\Phi_1[1, v]) - g_0\Phi_2[1, v] \quad \text{for all } v \in H^1_{\text{per}}(S).$$

Setting $v = 1$ and using (14) yields

$$0 = \Phi_0[\Phi_2, 1] = (2\pi)^n - g_0(\Phi_1[\chi_1, 1] + \Phi_2[1, 1]).$$

Solving for g_0 yields (21) with $k = 0$. On the other hand, requiring v to be in W, and using (20), gives

$$\begin{aligned} \Phi_0[\phi_2, v] &= -g_0(\Phi_1[\chi_1, v] + \Phi_2[1, v]) - g_1\Phi_1[1, v] \\ &= g_0\Phi_0[\chi_2, v] + g_1\Phi_0[\chi_1, v], \end{aligned}$$

which can be solved easily for ϕ_2 as a function in W. Thus, ϕ_0 is completely determined as in (22), whereas ϕ_1 and ϕ_2 have the form of (22); we have yet to show that g_1 and g_2 have been correctly defined.

Now let $k \geqq 3$, and assume that $\phi_0, \cdots, \phi_{k-3}$ are given by (22), and that ϕ_{k-2} and ϕ_{k-1} have the form of (22). That is, we are certain that $g_0, \cdots, g_{k-3}$ are correctly defined in (21), but are not sure about g_{k-2} and g_{k-1}. We wish to show that g_{k-2} is correctly defined in (21) and that ϕ_k has the form of (22). By our assumptions, (19) becomes

$$\begin{aligned} \Phi_0[\phi_k, v] = -\sum_{j=0}^{k-3} g_j(\Phi_1[\chi_{k-1-j}, v] + \Phi_2[\chi_{k-2-j}, v]) \\ - g_{k-2}(\Phi_1[\chi_1, v] + \Phi_2[1, v]) - g_{k-1}\Phi_1[1, v]. \end{aligned}$$

Setting $v = 1$ yields

$$0 = \Phi_0[\phi_k, 1] = -\sum_{j=0}^{k-3} g_j(\Phi_1[\chi_{(k-2)+1-j}, v] + \Phi_2[\chi_{(k-2)-j}, 1]) - \frac{(2\pi)^n}{g_0} g_{k-2},$$

which can be solved for g_{k-2}, and thus obtaining (21). Finally, upon requiring v to be in W, it follows from (20) that ϕ_k has the form given in (22). □

Next, we sufficiently investigate the properties of the expansion of $\phi(\cdot, h, t)$ in powers of h to be able to prove Lemma 12 in [8]. The results are stated here in Theorems 9 and 11.

We begin by determining the dependence on t of $\chi_1(\cdot, t)$. Expanding the right-hand side of (20) with $k=1$, according to (11) yields

$$\Phi_0[\chi_1(\cdot, t), v] = -i \sum_{q=1}^{n} \left(\int_S \sum_{p=1}^{n} a_{pq}(y) \overline{\frac{\partial v}{\partial y_q}(y)}\, dy \right) t_q$$

for all $v \in W$. Now define $\chi_{1;q} \in W$, for each $q = 1, \cdots, n$, to be the unique solution (cf. Lemmas 3 and 6) of

$$\Phi_0[\chi_{1;q}, v] = \int_S \sum_{p=1}^{n} a_{pq}(y) \overline{\frac{\partial v}{\partial y_q}(y)}\, dy \quad \text{for all } v \in W. \tag{24}$$

Consequently,

$$-\chi_1(\cdot, t) = i \sum_{q=1}^{n} \chi_{1;q} t_q. \tag{25}$$

Note that $\chi_{1;q}$ is a real-valued function because the same is true for each a_{pq}. Furthermore, it follows from (24) and the definition of Φ_0 that

$$\Phi_0[\chi_{1;q} + y_q, v] = 0 \quad \text{for all } v \in W. \tag{26}$$

The following formula for $g_0(t)$ can be easily obtained by substituting (11) and (25) into (21):

$$g_0(t) = \frac{1}{\sum_{p,q=1}^{n} A_{pq} t_q t_p + A_0}, \tag{27}$$

where

$$\begin{aligned} A_0 &\equiv \frac{1}{(2\pi)^n} \int_S a_0(y)\, dy, \\ A_{pq} &\equiv \frac{1}{(2\pi)^n} \int_S \left(a_{pq}(y) + \sum_{r=1}^{n} a_{pr}(y) \frac{\partial \chi_{1;q}}{\partial y_r}(y) \right) dy. \end{aligned} \tag{28}$$

THEOREM 9. *$A_0 \in \mathbb{R}$, $A_{pq} \in \mathbb{R}$, $A_{pq} = A_{qp}$, and $0 < g_0(t) \leqq 1/(\gamma_1 \|t\|^2 + \gamma_0)$ for $t \in \mathbb{R}^n$.*

Proof. A_0 and $\{A_{pq} : p, q = 1, \cdots, n\}$ are real numbers because each integrand in (28) is a real-valued function. It follows from the definition of Φ_0 and from the formula for A_{pq} that

$$A_{pq} = \frac{1}{(2\pi)^n} \Phi_0[\chi_{1;q} + y_q, y_p];$$

and upon using (26) with $v = \chi_{1;p}$ we obtain

$$A_{pq} = \frac{1}{(2\pi)^n} \Phi_0[\chi_{1;q} + y_q, \chi_{1;p} + y_p] \quad \text{for all } p, q = 1, \cdots, n. \tag{29}$$

The symmetry of A_{pq} follows from (29) (and (11)), since symmetry conditions are imposed on the coefficients a_{pq}, and since each function involved in (29) is real valued.

An immediate consequence of (1) and (28) is $A_0 \geqq \gamma_0$. Next, define $\xi(y, t) = \sum_{q=1}^{n} t_q y_q$. Then (25), (26), Lemma 6, and the fact that $\chi_1(\cdot, t)$ is a periodic function imply

$$\begin{aligned}
\sum_{p,q=1}^{n} A_{pq} t_q t_p &= \frac{1}{(2\pi)^n} \Phi_0[-i\chi_1(\cdot, t) + \xi(\cdot, t), -i\chi_1(\cdot, t) + \xi(\cdot, t)] \\
&\geqq \frac{\gamma_1}{(2\pi)^n} |-i\chi_1(\cdot, t) + \xi(\cdot, t)|_1^2 \\
&= \frac{\gamma_1}{(2\pi)^n} \int_S \sum_{p=1}^{n} \left| -i \frac{\partial \chi_1}{\partial y_p}(y, t) + t_p \right|^2 dy \\
&= \frac{\gamma_1}{(2\pi)^n} \left\{ |\chi_1(\cdot, t)|_1^2 + 2 \sum_{p=1}^{n} t_p \int_S \operatorname{Im}\left(\frac{\partial \chi_1}{\partial y_p}(y, t) \right) dy + (2\pi)^n \|t\|^2 \right\} \\
&\geqq \gamma_1 \|t\|^2.
\end{aligned}$$

Then $1/(g_0(t)) \geqq \gamma_1 \|t\|^2 + \gamma_0$ follows from (27). □

Next we prove Lemma 10 below.

LEMMA 10. *There are positive constants η and θ, which are independent of $t \in \mathbb{R}^n$ such that*

$$\|\phi_k(\cdot, t)\|_1 \leqq \eta g_0(t) \theta^k (1 + \|t\|)^k \quad \textit{for each } k \geqq 0.$$

Proof. After stating a few preliminary results, the proof is presented in three steps. First, an upper bound for $|\chi_k(\cdot, t)|_1$ is derived from (20). This result is then used with (21) in order to obtain an upper bound on $|g_k(t)|$. Finally these two bounds and (22) will give an upper bound on $\|\phi_k(\cdot, t)\|_1$.

It follows from (11) that there is a constant c_1 such that

$$|\Phi_k(t)[\phi, v]| \leqq c_1 (1 + \|t\|)^k \|\phi\|_1 \|v\|_1 \quad \text{for } k = 1, 2, \tag{30}$$

because the coefficients a_0 and a_{pq} are L_∞-functions. The closed graph theorem implies that there is a constant c_2, which we take to be larger than $(2\pi)^{n/2}$, such that

$$\|v\|_1 \leqq c_2 |v|_1 \quad \text{for all } v \in W. \tag{31}$$

We now have from (20) that

$$|\Phi_0[\chi_k(\cdot, t), v]| \leqq \begin{cases} c_1 c_2^2 (1 + \|t\|) |v|_1, & k = 1, \\ c_1 c_2^2 \{(1 + \|t\|) |\chi_1(\cdot, t)|_1 + (1 + \|t\|)^2\} |v|_1, & k = 2, \\ c_1 c_2^2 \sum_{j=1}^{2} (1 + \|t\|)^j |\xi_{k-j}(\cdot, t)|_1 |v|_2, & k \geqq 3. \end{cases}$$

Lemmas 3 and 6 imply

$$|\chi_k(\cdot, t)|_1 \leqq \begin{cases} c_e (1 + \|t\|), & k = 1, \\ c_3 (|\chi_1(\cdot, t)|_1 + (1 + \|t\|))(1 + \|t\|), & k = 2, \\ c_3 (|\chi_{k-1}(\cdot, t)|_1 + (1 + \|t\|) |\chi_{k-2}(\cdot, t)|_1)(1 + \|t\|), & k \geqq 3, \end{cases}$$

in which $c_3 \equiv c_1 c_2^2 / \gamma_1$. An induction argument proves

$$|\chi_k(\cdot, t)|_1 \leqq (c_3 + 1)^k (1 + \|t\|)^k \quad \text{for } k = 1, 2, \cdots. \tag{32}$$

Next, using (30)–(32) in (21) yields

$$|g_k(t)| \leqq \frac{g_0(t)}{(2\pi)^n} \sum_{j=0}^{k-1} |g_j(t)| c_1 c_2^2 (1+\|t\|) \cdot (|\chi_{k+1-j}(\cdot, t)|_1 + (1+\|t\|)|\chi_{k-j}(\cdot, t)|_1)$$

$$\leqq \frac{c_1 c_2^2 (c_3+2)}{(2\pi)^n} g_0(t)(1+\|t\|)^2 \sum_{j=0}^{k-1} |g_j(t)|(c_3+1)^{k-j}(1+\|t\|)^{k-j}$$

for $k \geqq 1$. A consequence of Theorem 9 is that this last inequality can be rewritten as

$$|g_k(t)| \leqq c_4 \sum_{j=0}^{k-1} |g_j(t)|(c_3+1)^{k-j}(1+\|t\|)^{k-j} \quad \text{for } k \geqq 1,$$

where

$$c_4 = \frac{2c_1 c_2^2 (c_3+2)}{(2\pi)^n \min\{\gamma_0, \gamma_1\}}.$$

Another induction argument then proves

$$|g_k(t)| \leqq g_0(t)(c_3+1)^k (c_4+1)^k (1+\|t\|)^k \quad \text{for } k \geqq 0. \tag{33}$$

Finally, substitute (32) and (33) into (22) to obtain

$$\|\phi_k(\cdot, t)\|_1 \leqq \sum_{j=0}^{k-1} g_0(t) c_2 (c_3+1)^k (c_4+1)^j (1+\|t\|)^k + g_0(t) c_2 (c_3+1)^k (c_4+1)^j (1+\|t\|)^k$$

$$\leqq c_2 g_0(t)(c_3+1)^k (c_4+2)^k (1+\|t\|)^k,$$

which finishes the proof. □

By computing a majorizing series for (18), the next theorem is a consequence of Lemma 10.

THEOREM 11. *Let η and θ be given as in Lemma* 17, *and suppose $h>0$ and $t \in \mathbb{R}^n$ satisfy $\theta(1+\|t\|)h < 1$. Then*

$$\|\phi(\cdot, h, t)\|_1 \leqq \frac{\eta g_0(t)}{1-\theta(1+\|t\|)h},$$

$$\left\| \phi(\cdot, h, t) - \sum_{j=0}^{k} \phi_j(\cdot, t) h^j \right\|_1 \leqq \frac{\eta \theta^{k+1} g_0(t)}{1-\theta(1+\|t\|)h} (1+\|t\|)^{k+1} h^{k+1}$$

for $k \geqq 0$.

5. Appendix. In this section we develop some of the theory that was used in § 3 when making some of our analyticity claims. The main goals here are Theorems 19 and 20. Throughout this section we make the following assumptions. Let V and H be separable, complex Hilbert spaces in which V is a compactly and continuously embedded dense subspace of H. We denote the associated inner products by $(\cdot, \cdot)_V$ and $(\cdot, \cdot)_H$, and the associated norms by $\|\cdot\|_V$ and $\|\cdot\|_H$. Let G be an open set in $\mathbb{C}^n$, and consider a family of sesquilinear forms $\Phi(z): V \times V \to \mathbb{C}$, defined for each $z \in G$. Suppose that there are real-valued functions M, γ, and μ defined on G such that $M(z) > 0$ and $\gamma(z) > 0$ for $z \in G$, μ is continuous on G, and for each $z \in G$

$$|\Phi(z)[\phi, v]| \leqq M(z)\|\phi\|_V \|v\|_V \quad \text{for all } \phi \text{ and } v \text{ in } V, \tag{34}$$

$$\gamma(z)\|v\|_V^2 \leqq \operatorname{Re}(\Phi(z)[v, v]) + \mu(z)\|v\|_H^2 \quad \text{for all } v \in V. \tag{35}$$

Furthermore, suppose that $z \mapsto \Phi(z)[\phi, v]$ is analytic on G, for each ϕ and v in V.

For a given $w \in H$, we want to determine the dependence on z, in particular situations, of $\phi(z) \in V$, which satisfies

$$\Phi(z)[\phi(z), v] = (w, v)_H \quad \text{for all } v \in V.$$

In so doing, we will consider the eigenvalue problem: Seek $\lambda(z) \in \mathbb{C}$ and $\psi(z) \in V$ such that

$$\Phi(z)[\psi(z), v] = \lambda(z)(\psi(z), v)_H \quad \text{for all } v \in V.$$

Let $a \in G$. If zero is not an eigenvalue of $\Phi(a)$, then we will show that $\Phi(z)$ exists for, and depends analytically on, z in a neighborhood of a. If zero is a "simple" eigenvalue of $\Phi(a)$, then we will show that λ, with $\lambda(a) = 0$, is analytic in a neighborhood of a, and we will derive an expression for $\phi(z)$, exhibiting its dependence on $\lambda(z)$.

When z is one complex variable, many of the results of this section can be found in [5]. An important difference is that we have imposed alternate conditions ((34) and (35)) on $\Phi(z)$. This allows us to conclude that ϕ is analytic with values in V, rather than in H, which is the conclusion in [5].

At this point we want to give a definition of analyticity, or holomorphy, for Banach space-valued functions of several complex variables. Several definitions are possible. In a setting more general than Banach spaces, three definitions are stated and proven to be equivalent, in Chapter III of [4]. First an open polydisc $\Delta(a, \rho)$ in $\mathbb{C}^n$ with center a and multiradius $\rho \equiv (\rho_1, \cdots, \rho_n)$, where $0 < \rho_j < \infty$, is defined by

$$\Delta(a, \rho) \equiv \{z \in \mathbb{C}^n : |z_j - a_j| < \rho_j \text{ for } j = 1, \cdots, n\}.$$

DEFINITION 12. Let W be a Banach space, and recall that G is an open set in $\mathbb{C}^n$. A function $w : G \to W$ is analytic, or holomorphic, if for each $a \in G$ there is a polydisc $\Delta(a, \rho) \subset G$ and a set of coefficients $\{w_\alpha(a) : \alpha \text{ is a multi-index}\} \subset W$ such that $\sum_{0 \leq |\alpha|} w_\alpha(a)(z-a)^\alpha$ converges in W to $w(z)$ for each $z \in \Delta(a, \rho)$.

We will have several occasions in which the next two lemmas will be used. When $n = 1$, proofs can be found in [5], and for the general case they are proved in [7]. Let H_1 and H_2 be two separable, complex Hilbert spaces and denote the inner product on H_2 by $(\cdot, \cdot)_2$. Denote the space of bounded linear operators mapping H_1 into H_2 by $B(H_1, H_2)$.

LEMMA 13. *Let $T(z) \in B(H_1, H_2)$ for each $z \in G$. The following statements are equivalent:*

(i) *$T : G \to B(H_1, H_2)$ is holomorphic;*

(ii) *$T(\cdot)\phi : G \to H_2$ is holomorphic for each $\phi \in H_1$;*

(iii) *$(T(\cdot)\phi, v)_2 : G \to \mathbb{C}$ is holomorphic for each $\phi \in H_1$ and $v \in H_2$.*

LEMMA 14. *Suppose $T : G \to B(H_1, H_2)$ is holomorphic, and let $a \in G$. If $T(a)^{-1} \in B(H_2, H_1)$, then there exists a neighborhood $G_a \subset G$ of a such that $T(z)^{-1}$ exists for $z \in G_1$ and $T(\cdot)^{-1} : G_1 \to B(H_2, H_1)$ is holomorphic.*

We are now ready to state and prove the results on which the main theorems of this section are based. We begin by showing that there exists a closed operator $T(z) : D(T(z)) \subset V \to H$ such that $\Phi(z)[\phi, v] = (T(z)\phi, v)_H$ for all $\phi \in D(T(z))$ and $v \in V$. The next theorem gives one way of constructing such an operator, which will be convenient for us in what follows. Other forms of this representation theorem can be found in [5] and [6]. (See also [1] and [9].)

Throughout this section, we will denote the domain and range of an operator T by $D(T)$ and $R(T)$. Also, in the next theorem only, the dependence on z, as we have stated so far, is inconsequential, and so we drop it.

THEOREM 15. *Let $\Phi: V \times V \to \mathbb{C}$ be a sesquilinear form for which there exist real constants $M > 0$, $\gamma > 0$, and μ such that*

$$|\Phi[\phi, v]| \leqq M \|\phi\|_V |v\|_V \quad \text{for all } \phi \text{ and } v \text{ in } V, \tag{34'}$$

$$\phi \|v\|_V^2 \leqq R(\Phi[v, v]) + \mu \|v\|_H^2 \quad \text{for all } v \in V. \tag{35'}$$

Then there is a unique closed operator $T: V \to H$ such that

(i) *$D(T)$ is dense in V;*

(ii) *$\Phi[\phi, v] = (T\phi, v)_H$ for all $\phi \in D(T)$ and $v \in V$;*

(iii) *given $\phi \in V$ and $w \in H$, if $\Phi[\phi, v] = (w, v)_H$ for all v in a dense subspace of V, then $\phi \in D(T)$ and $T\phi = w$.*

Proof. Uniqueness follows from (iii); let there be another closed operator S such that $\Phi[\phi, v] = (S\phi, v)_H$ for all $\phi \in D(S)$ and $v \in V$. Then $\phi \in D(T)$ and $T\phi = S\phi$.

Since the embedding of V in H is continuous, it follows from the Riesz representation theorem that there is a linear operator $F \in B(H, V)$ such that

$$(w, v)_H = (Fw, v)_V \quad \text{for all } w \in H \text{ and } v \in V.$$

Furthermore, F is a 1-1 map, and $R(F)$ is dense in V.

Next, it follows from (34') and (35') that $\Phi[\cdot, \cdot] + \mu(\cdot, \cdot)_H$, as a sesquilinear form on $V \times V$, satisfies the hypotheses of Lemma 3. From the discussion that follows Lemma 3, it follows that there exists an operator $A_\mu \in B(V, V)$ such that $A_\mu^{-1} \in B(V, V)$ and $\Phi[\phi, v] + \mu(\Phi, v)_H = (A_\mu \phi, v)_V$ for all ϕ and v in V. Now define

$$D(T) \equiv \{\phi \in V: A_\mu \phi \in R(F)\}$$

and set

$$T = (F^{-1} A_\mu - \mu I)_{D(T)},$$

in which I is the identity operator on H.

Clearly the choice of μ in (35') is not unique. That the definitions of $D(T)$ and T are independent of μ can be seen, as follows. Let $\mu' \neq \mu$ be a real number for which (35') remains valid when μ is replaced by μ'. (The value of $\gamma > 0$ makes no difference.) The definitions of A_μ and $A_{\mu'}$ imply

$$(A_{\mu'} \phi, v)_V - \mu'(F\phi, v)_V = (A_\mu \phi, v)_V - \mu(F\phi, v)_V$$

for all ϕ and v in V. Thus

$$(A_{\mu'} - \mu' F)\phi = (A_\mu - \mu F)\phi \quad \text{for all } \phi \in V,$$

from which it follows that $A_{\mu'} \phi \in R(F)$ if and only if $A_\mu \phi \in R(F)$, and that

$$T = (F^{-1} A_\mu - \mu I)|_{D(T)} = (F^{-1} A_{\mu'} - \mu' I)|_{D(T)}.$$

Since A_μ is an isomorphism on V and $R(F)$ is dense in V, it follows that $D(T)$ is dense in V, which proves (i).

Statement (ii) follows from the definitions of A_μ and T.

To prove (iii), let $\phi \in V$ and $w \in H$ such that $\Phi[\phi, v] = (w, v)_H$ for all v in a dense subset of V. Then

$$(A_\mu \phi - \mu F\phi, v)_V = \Phi[\phi, v] = (W, v)_H = (Fw, v)_V$$

for all v in a dense subset of V, which implies $A_\mu \phi - \mu F\phi = Fw$ so that $\phi \in D(T)$ and $T\phi = w$.

Finally, $T: D(T) \subset V \to H$ is a closed operator because $A_\mu^{-1} F \in B(H, V)$. □

Noting (34) and (35), Theorem 15 implies the existence of a unique closed operator $T(z): V \to H$ for each $z \in G$, such that $D(T(z))$ is dense in v; $\Phi(z)[\phi, v] = (T(z)\phi, v)_H$ for all $\phi \in D(T(z))$ and $v \in V$; and for any $w \in H$,

$$T(z)\phi = w \quad \text{if and only if } \phi(z)[\phi, v] = (w, v)_H \text{ for all } v \text{ in a dense subset of } V. \tag{36}$$

Since $T(z): D(T(z)) \subset V \to H$ is closed, the resolvent operator $R(\zeta, z) \equiv (T(z) - \zeta)^{-1}$ belongs to $B(H, V)$ for each $\zeta \in \rho(T(z))$, the resolvent set of $T(z)$. A consequence of (34)-(36) and Lemma 3 is that $\rho(T(z))$ contains $\{\zeta \in \mathbb{C}: -\mathrm{Re}\, \zeta \geqq \mu(z)\}$. A standard result in the spectral theory of operators is that $\rho(T(z))$ is an open set in $\mathbb{C}$. In Theorem 17 below, we will prove that

$$\mathcal{G} \equiv \{(\zeta, z) \in \mathbb{C}^{n+1}: \zeta \in \rho(T(z)) \text{ and } z \in G\} \tag{37}$$

is an open set also, and that $R: \mathcal{G} \to B(H, V)$ is holomorphic.

First we prove a preliminary result.

LEMMA 16. *For each $a \in G$, if $\zeta \in \rho(T(a))$ then there exists a neighborhood $G_{a\zeta} \subset G$, of a such that*

(i) *$\zeta \in \rho(T(z))$ and $R(\zeta, z) \in B(H, V)$ for all $z \in G_{a\zeta}$;*

(ii) *$z \mapsto R(\zeta, z) \in B(H, V)$ is holomorphic on $G_{a\zeta}$.*

Proof. The continuity of μ allows us to choose a neighborhood $G_a \subset G$ of a, and a number $\mu_a \geqq \mu(z)$ for $z \in G_a$; that is,

$$\gamma(z)\|v\|_V^2 \leqq \mathrm{Re}\,(\Phi(z)[v, v]) + \mu_a \|v\|_H^2$$

for all $v \in V$ and $z \in G_a$. We will first prove the lemma for $\zeta = -\mu_a$ and then use the identity

$$(T(z) - \zeta) r(-\mu_a, z) = I_H - (\zeta + \mu_a) R(-\mu_a, z) \quad \text{for } z \in G_a, \tag{38}$$

to prove the lemma for arbitrary $\zeta \in \rho(T(a))$.

As in the proof of Theorem 15, associate an operator $A_a(z) \in B(V, V)$ to $\Phi(z)$ such that

$$\Phi(z)[\phi, v] + \mu_a(\phi, v)_H = (A_a(z)\phi, v)_V \tag{39}$$

for all ϕ and v in V. It was shown there that while $A_a(z)$ depends on the choice of μ_a, $A_a(z) - \mu_a F$ does not depend on μ_a, where $F \in B(H, V)$ is defined by $(Fw, v)_V = (w, v)_H$ for all $w \in H$ and $v \in V$. Furthermore, $D(T(z)) = \{\phi \in V: A_a(z)\phi \in R(F)\}$ and $T(z) = (F^{-1}A_a(z) - \mu_a)|_{D(T(z))}$ for $z \in G_a$. Consequently, $T(z) + \mu_a = F^{-1}A_a(z)|_{D(T(z))}$ is a one-to-one map of $D(T(z))$ onto H, and it follows that $-\mu_a \in \rho(T(z))$ whenever $z \in G_1$. According to Lemma 13 and the hypothesis that $z \mapsto \Phi(z)[\phi, v]$ is analytic, it follows from (39) that $A_a: G_a \to B(V, V)$ is holomorphic. Since $A_a(z)^{-1} \in B(V, V)$ for each z in G_a, Lemma 14 implies that $A_a(\cdot)^{-1}: G_1 \to B(V, V)$ is holomorphic. Therefore $z \mapsto R(-\mu_a, z) = A_a(z)^{-1}F \in B(H, V)$ is holomorphic on G_a.

Now, $\zeta \in \rho(T(a))$. When $z = a$, the left-hand side of (38) is a one-to-one map of H onto H; hence its inverse exists and belongs to $B(H, H)$. As a function of z with values in $B(H, H)$, $R(-\mu_a, z)$ is holomorphic on G_a because it is holomorphic as a function with values in $B(H, V)$ and because the embedding of V into H is bounded. Thus the right-hand side of (38), as a function of z with values in $B(H, H)$, is

holomorphic on G_a, and its inverse belongs to $B(H, H)$ when $z = a$. According to Lemma 21, there is a neighborhood $G_{a\zeta} \subset G_a$ of a on which

$$z \mapsto (I - (\zeta + \mu_a)R(-\mu_a, z))^{-1} \in B(H, H)$$

is holomorphic. Therefore, $\zeta \in \rho(T(z))$ for $z \in G_{a\zeta}$, and the holomorphy of

$$z \in G_{a\zeta} \mapsto R(\zeta, z) = R(-\mu_a, z)(I - (\zeta + \mu_a)R(-\mu_a, z))^{-1} \in B(H, V)$$

follows.

THEOREM 17. *The set $\mathcal{G}$, defined by* (37), *is open in $\mathbb{C}^{n+1}$, and $(\zeta, z) \mapsto R(\zeta, z) \in B(H, V)$ is holomorphic on $\mathcal{G}$.*

Proof. Let $(\eta, a) \in \mathcal{G}$. This proof is essentially a careful repetition of the second part of Lemma 16, with η replacing $-\mu_a$. Lemma 16 implies that a neighborhood $G_{a\eta} \subset G$ of a can be found such that $\eta \in \rho(T(z))$ for $z \in G_{a\eta}$, and $z \mapsto R(\eta, z) \in B(H, V)$ is holomorphic on $G_{a\eta}$. Analogous to (38), we have

$$(T(z) - \zeta)R(\eta, z) = I - (\zeta - \eta)R(\eta, z) \quad \text{for } z \in G_{a\eta}. \tag{40}$$

Now the right-hand side of (40), as a function of (ζ, z), with values in $B(H, H)$, is holomorphic on $\mathbb{C} \times G_{a\eta}$ and takes on the value I when $(\zeta, z) = (\eta, a)$. A consequence of Lemma 14 is that there is a neighborhood $\mathcal{G}_{\eta a} \subset \mathbb{C} \times G_{a\eta}$ of $(h\ a)$ on which $(\zeta, z) \mapsto (I - (\zeta - \eta)R(\eta, z))^{-1} \in B(H, H)$ is holomorphic. Hence, $\zeta \in \rho(T(z))$ for $(\zeta, z) \in \mathcal{G}_{\eta a}$, which implies that $\mathcal{G}$ is open. It follows that

$$(\zeta, z) \mapsto R(\zeta, z) = R(\eta, z)(I - (\zeta - \eta)R(\eta, z))^{-1} \in B(H, V)$$

is holomorphic on $\mathcal{G}_{\eta a}$. □

The next corollary states a condition on the family $\{\Phi(z): z \in G\}$ which guarantees that the operator $T(z)$ is self-adjoint if $z \in G \cap \mathbb{R}^n$.

COROLLARY 18. *Suppose that $\bar{z} = (z_1, \cdots, \bar{z}_n) \in G$ whenever $z \in G$. If $\Phi(z)[v, \phi] = \Phi(\bar{z})[\phi, v]$ for all ϕ and v in V and for $z \in G$, then $T(z)^* = T(z)$ for all $z \in G$, in which $T(z)^*$ is the adjoint of $T(z)$ as an operator on H.*

Note that $R(\zeta, z)$ is compact as an operator on H, when $\zeta \in \rho(T(z))$, because $R(\zeta, z) \in B(H, V)$ and V is compactly embedded in H by hypothesis. Consequently, the spectrum of $T(a)$ consists entirely of eigenvalues that have finite multiplicity and no finite accumulation point.

Recall that given $w \in H$, we want to determine the existence and the dependence upon z of the solution $\phi(z) \in V$ of

$$\Phi(z)[\phi(z), v] = (w, v)_H \quad \text{for all } v \in V.$$

It follows from (36) that this is equivalent to solving $T(z)\phi(z) = w$. When zero is not an eigenvalue $\phi(z) = R(0, z)w$. Lemma 16 yields the following theorem.

THEOREM 19. *$\phi: \{z \in G: 0 \in \rho(T(z))\} \to V$ is holomorphic.*

When zero is a simple eigenvalue, we have the following result.

THEOREM 20. *Let $a \in G$ and suppose zero is a simple eigenvalue of $T(a)$. Then there exists a neighborhood $G_a \subset G$ of a and two functions λ and $z \mapsto P(z) \in B(H, V)$, which are holomorphic on G_a, such that $\lambda(a) = 0$, $\lambda(z)$ is a simple eigenvalue of $T(z)$, and $P(z)$ projects H onto the one-dimensional eigenspace that corresponds to $\lambda(z)$. Furthermore, there exists another holomorphic function $z \in G_a \mapsto R_2(0, z) \in B(H, V)$ such that*

$$\phi(z) = \frac{1}{\lambda(z)} P(z)w + R_2(0, z)w \tag{41}$$

for all $z \in G_a$ for which $\lambda(z) \neq 0$.

The remainder of this section will be devoted to the proof of Theorem 20. The theory concerning the eigenvalue problem associated to $T(z)$ is well developed (cf. [5]). In fact, the form of $\phi(z)$ in (41) is a direct consequence of that theory. Here we repeat many of these ideas in the process of proving the conclusions about analyticity.

Since the spectrum of $T(a)$ is a discrete set of eigenvalues having no finite accumulation point, a Jordan curve $\mathscr{C}$ can be drawn in $\rho(T(a))$ so as to enclose an open set in $\mathbb{C}$ containing zero in its interior and the other eigenvalues in the exterior of its closure. Then $\mathscr{C}\times\{a\}\subset\mathscr{G}$, where $\mathscr{G}$ is the open set defined by (37). Hence for each $\zeta\in\mathscr{C}$ there is a disc $D(\zeta, r(\zeta))\subset\mathbb{C}$ ($r(\zeta)>0$, is the radius), and a polydisc $\Delta(a, \rho(\zeta))\subset\mathbb{C}^n$ such that

$$(\zeta, a)\in D(\zeta, r(\zeta))\times\Delta(a, \rho(\zeta))\subset\mathscr{G}.$$

However, $\mathscr{C}$ is compact, so a finite set $\{\zeta_j\in\mathscr{C}: j=1, \cdots, k\}$ can be chosen such that $\{D(\zeta_j, r(\zeta_j)): j=1, \cdots, k\}$ covers $\mathscr{C}$. Consequently, $\mathscr{C}\subset\rho(T(z))$ for $z\in G_a'\equiv\bigcap_{j=1}^{k}\Delta(a, \rho(\zeta_j))$.

Next the operator $P(z)$ is defined for $z\in G_a'$ as a Riemann integral of $B(H, V)$-valued functions by

$$P(z)\equiv-\frac{1}{2\pi i}\int_{\mathscr{C}} R(\zeta, z)\,d\zeta. \tag{42}$$

It is shown in Theorems III-6.17 and VII-1.7 of [5] that $P(z)$ is a projection operator and that $P(a)$ maps H onto $M_1(a)$, the one-dimensional eigenspace associated with the eigenvalue 0 of $T(a)$. Moreover, H can be decomposed as $H=M_1(z)+M_2(z)$ for $z\in G_a'$, in which

$$M_1(z)\equiv P(z)H \quad\text{and}\quad M_2(z)\equiv(I-P(z))H. \tag{43}$$

It is also true that $P: G_a'\to B(H, V)$ is holomorphic because, by Theorem 17, the same is true of $R: \mathscr{G}\to B(H, V)$. Since $\dim(M_1(a))=1$, and since $P: G_a'\to B(H, H)$ is continuous (recall that V is continuously embedded in H), it follows from [5, ¶¶I-§ 4.6, IV-§ 3.4] that

$$\dim(M_1(z))=1 \quad\text{for } z\in G_a'. \tag{44}$$

Now let $\psi_a\in M_1(a)$ and nonzero, and define G_a to be an open connected subset of G_a' such that $a\in G_a$ and $(P(z)\psi_a, \psi_a)\neq 0$ for $z\in G_a$. Next, define

$$\psi(z)\equiv P(z)\psi_a \quad\text{for } z\in G_a; \tag{45}$$

it follows that $\psi: G_a\to V$ is holomorphic with $\psi(a)=\psi_a$.

It follows from [5, ¶¶ III-§ 5.6, III-§ 6.1] of [5] that for $z\in G_a$,

$$P(z)v\in D(T(z)) \quad\text{for all } v\in D(T(z)),$$

$$T(z)v\in M_k(z) \quad\text{for } v\in M_k(z)\cap D(T(z)) \text{ and } k=1.$$

Thus $T_k(z): M_k(z)\to M_k(z)$ for $k=1, 2$, and for $z\in G_a$ can be defined by $T_k(z)\equiv T(z)|_{M_k(z)\cap D(T(z))}$. The eigenvalue problem has now been decomposed into two eigenvalue problems, one for each $T_k(z)$ in $M_k(z)$. Of particular interest here is the eigenvalue problem for $T_1(z)$, because 0 and ψ_a form an eigenvalue-eigenfunction pair for $T_1(a)$.

According to (44), $T_1(z)$ is a one-dimensional operator. Therefore, $\lambda(z)\equiv$ trace $(T_1(z))$ is the eigenvalue of $T_1(z)$, i.e., for each $z\in G_a$,

$$T(z)\psi_a=T_1(z)\psi_z=\lambda(z)\psi_z \quad\text{for all } \psi_z\in M_1(z)\cap D(T(z)). \tag{46}$$

In (46), setting $\psi_z = P(z)\psi_a$ and taking inner products in H with ψ_a yields

$$\lambda(z) = \frac{(T(z)P(z)\psi_a, \psi_a)_H}{(P(z)\psi_a, \psi_a)_H} \quad \text{for } z \in G_a. \tag{47}$$

Since $T(z)$ is a closed operator, it follows from (42) and the identity $T(z)R(\zeta, z) = I + \zeta R(\zeta, z)$ that $T(z)P(z) = -(1/2\pi i)\int_{\mathscr{C}} \zeta R(\zeta, z)\, d\zeta$. Consequently, $z \mapsto T(z)P(z) \in B(H, V)$ is holomorphic on G_a ($\subset G_a'$). Therefore, the analyticity of λ on G_a follows from (47) because G_a was chosen so that $(P(z)\psi_a, \psi_a) \neq 0$ when $z \in G_a$. Since $\mathscr{C} \subset \rho(T(z))$ for all $z \in G_a$, $\lambda(z)$ lies in the interior of the open set enclosed by $\mathscr{C}$, whereas the remainder of the spectrum of $T(z)$ must lie in the open set that is exterior to $\mathscr{C}$.

Let $z \in G_a$ such that $\lambda(z) \neq 0$. Then $R(0, z)$ exists and commutes with $P(z)$. Consequently, we can define $R_k(0, z) \in B(H, V)$ for $k = 1, 2$, by

$$R_1(0, z) \equiv R(0, z)P(z) = P(z)R(0, z),$$

$$R_2(0, z) \equiv R(0, z)(I - P(z)) = (I - P(z))R(0, z).$$

By passing $R(0, z)$ under the integral sign in (42), then using the resolvent equation to obtain $R(0, z) - R(\zeta, z) = -\zeta R(0, z)R(\zeta, z)$ when zero and ζ are in $\rho(T(z))$, and noting that zero lies inside the open set enclosed by $\mathscr{C}$,

$$R(0, z)P(z) = R(0, z) - \frac{1}{2\pi i}\int_{\mathscr{C}} \frac{R(\zeta, z)}{\zeta}\, d\zeta \tag{48}$$

is obtained for each $z \in G_a$ such that $\lambda(z) \neq 0$. However, $\mathscr{C} \subset \rho(T(z))$ for all $z \in G_a$, so that the second term on the right-hand side of (48) is holomorphic as a function of z on G_a with values in $B(H, V)$. Consequently, $z \mapsto R_2(0, z) \in B(H, V)$ can be continued analytically to all of G_a.

Finally, it is clear from (43) and the definition of $R_1(0, z)$ that $R_1(0, z)w \in M_1(z) \cap D(T(z))$ for all $w \in H$ and for each $z \in G$ such that $\lambda(z) \neq 0$. Then (46) implies

$$\lambda(z)R_1(0, z)w = T(z)R_1(0, z)w = T(z)R(0, z)P(z)w = P(z)w,$$

from which $R_1(0, z)w = (1/\lambda(z))P(z)w$ is obtained from all $w \in H$ and for each $z \in G$ such that $\lambda(z) \neq 0$. This finishes the proof of Theorem 20 because $\phi(z) = R(0, z)w = R_1(0, z)w + R_2(0, z)w$ when $\lambda(z) \neq 0$.

REFERENCES

[1] S. AGMON, *Lecture on Elliptic Boundary Value Problems*, Van Nostrand, Princeton, NJ, 1965.

[2] I. BABUŠKA AND A. K. AZIZ, *Survey lectures on the mathematical foundation of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1973, pp. 5–359.

[3] I. BABUŠKA AND R. C. MORGAN, *Composites with a periodic structure: mathematical analysis and numerical treatment*, Comput. Math. Appl., (1985), pp. 995–1001.

[4] M. HERVÉ, *Analytic and Plurisubharmonic Functions*, A. Dold and B. Eckmann, eds., Lecture Notes in Math. 198, Springer-Verlag, Berlin, New York, 1971.

[5] T. KATO, *Perturbation Theory for Linear Operators*, Second edition, Springer-Verlag, Berlin, New York, 1976.

[6] J. L. LIONS, *Equations differentielles opérationnnelles et problèmes aux limites*, Springer-Verlag, Berlin, New York, 1961.

[7] R. C. MORGAN, *Mathematical aspects and computational considerations in the theory of homogenization*, Ph.D. thesis, University of Maryland, College Park, MD, 1982.

[8] R. C. MORGAN AND I. BABUŠKA, *An approach for constructing families of homogenized equations for periodic media. I: An integral representation and its consequences*, SIAM J. Math. Anal, this issue (1991), pp. 1–15.

[9] M. SCHECHTER, *Spectra of Partial Differential Operators*, North-Holland, Amsterdam, New York, 1971.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 34-45, January 1991

003

ASYMPTOTIC BEHAVIOUR OF SOLUTIONS OF THE POROUS MEDIUM EQUATION WITH CHANGING SIGN*

SHOSHANA KAMIN† AND JUAN LUIS VAZQUEZ‡

Abstract. It is proved that solutions of the porous medium equation $u_t = \Delta(|u|^{m-1}u)$, $m>1$, defined in $Q=\mathbb{R}^N\times(0,\infty)$ with initial data $u(x,0)$ integrable, compactly supported, and with changing sign, become nonnegative in finite time if $\int u_0(x)\,dx>0$. Precise asymptotic convergence rates then follow. The positivity result is related to finite propagation and is false for the heat equation, $m=1$. Nevertheless, we obtain asymptotic convergence rates for $1\geqq m>(N-2)_+/N$ as $t\to\infty$.

The same analysis works for the equation $u_t=\operatorname{div}(|\nabla u|^{p-1}\nabla u)$ with $p>1$. The case of zero mass, $\int u_0(x)\,dx=0$, is also studied for the porous medium equation (in $N=1$) and the solution is shown to converge to an antisymmetric profile in that case.

Key words. nonlinear parabolic equations, flows in porous media, asymptotic behavior

AMS(MOS) subject classifications. 35K65, 35B40

1. Introduction.

In this paper we study the large-time behaviour of the solutions of the porous medium equation

$$u_t=\Delta(|u|^{m-1}u) \quad \text{for } x\in\mathbb{R}^N,\quad t>0, \tag{1.1}$$

with $m>0$, $N\geqq 1$, and initial data

$$u(x,0)=u_0(x) \quad \text{for } x\in\mathbb{R}^N. \tag{1.2}$$

We begin the paper with the assumption that $m>1$. We make the assumptions

$$u_0\in L^1(\mathbb{R}^N), \tag{1.3}$$

$$u_0 \text{ has compact support,} \tag{1.4}$$

but we make *no* assumption on the sign of u_0. Our main result is the following theorem.

THEOREM 1. *Let u be the solution of* (1.1)-(1.4) *with*

$$M\equiv\int u_0(x)\,dx>0. \tag{1.5}$$

Then u becomes nonnegative in a finite time $T=T(u_0)$. Similarly u becomes nonpositive if $M<0$.

The result is false for $m=1$, i.e., the classical heat equation $u_t=\Delta u$, as the following simple example shows. Let

$$u_0(x)=-\delta(x)+M\delta(x+ae_1),$$

where δ stands for Dirac's mass, $a>0$ and $M>1$ are constants, $x=(x_1,x_2,\cdots,x_N)\in\mathbb{R}^N$, and $e_1=(1,0,\cdots,0)$. Then the solution is given by

$$u(x,t)=-E(x,t)+ME(x+ae_1,t), \tag{1.6}$$

* Received by the editors April 5, 1989; accepted for publication (in revised form) February 1, 1990.

† School of Mathematical Sciences, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. The work of this author was supported by the United States-Israel Binational Science Foundation grant 06-0361-0682-00000.

‡ Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain. The work of this author was supported by Direcciõn General de Investigación Científica (DGICYT-Spain) grant PB86-0112-C02-01 and EEC contract SC1-0019-C, and performed during the author's visit to Tel Aviv University.

where E is the fundamental solution

$$E(x, t) = (4\pi t)^{-N/2} \exp\left(-\frac{|x|^2}{4t}\right). \tag{1.7}$$

This expression is negative precisely in the region

$$x_1 > \frac{2t}{a} \lg M - \frac{a}{2}. \tag{1.8}$$

An example with $u_0 \in L^1(\mathbb{R}^N)$ is obtained by displacing the origin of time to $t = \tau > 0$. Observe that some solutions initially having changing sign eventually become positive. Take for instance

$$u_0(x) = -\delta(x) + M\delta(x + ae_1) + M\delta(x - ae_1)$$

with $M > \frac{1}{2}$.

The main property of the porous medium ($m > 1$) which is not shared by the heat equation (nor by the fast diffusion equation $u_t = \Delta(|u|^{m-1}u)$, $0 < m < 1$) is *finite propagation*, which in particular means that a solution u such that u_0 has compact support has the same property for all fixed times $t > 0$; cf. [OKC] or the survey papers [P], [Ar]. The equivalent of the fundamental solution E of the heat equation is the self-similar solution (Barenblatt's solution [B], [ZK])

$$w(x, t) = t^{-k}\left(C - \frac{(m-1)k}{2mN}|\xi|^2\right)_+^{1/(m-1)}, \tag{1.9}$$

where $k = (m - 1 + (2/N))^{-1}$, $\xi = xt^{-k/N}$, $(s)_+$ means $\max(s, 0)$ and C is an arbitrary constant which can be determined by fixing the mass

$$M = \int w(x, t)\, dx,$$

so that $C = a(m, N)M^l$ with $l = 2(m-1)k/N$. (We will often write w_M instead of w to stress dependence on M.) Observe that w_M takes the initial data $M\delta(x)$ and that w vanishes in the region

$$|x| \geqq \rho_M(t) \equiv b(m, N)(M^{m-1}t)^{k/N}. \tag{1.10}$$

Solutions for $u_0(x) = -M\delta(x)$ are obtained by just changing the signs.

We can now see an intuitive explanation for the different behaviour regarding Theorem 1 in the cases where $m > 1$ and $m = 1$. While a fundamental solution $ME(x + ae_1, t)$ will never dominate $-E(x, t)$ over the whole space $\mathbb{R}^n$ even if t is large, this will happen with $w_M(x + ae_1, t)$ and $w_{-1}(x, t) = -w_1(x, t)$. Of course, since (1.1) with $m > 1$ is nonlinear, this is no proof of our result.

The asymptotic behaviour of nonnegative solutions to problem (1.1), (1.2) has been investigated by various authors for different assumptions on u_0; cf. [AR], [K], [FK], [Ve], $\cdots$. In particular, when $u_0 \in L^1(\mathbb{R}^N)$, convergence to the self-similar solution w with the same mass M as u_0 has been established by Kamin [K] in 1973 for $N = 1$, and by Friedman and Kamin [FK] in 1980 for any space dimensions. The theorem follows.

THEOREM 2 [FK]. *Let u be the solution to* (1.1)-(1.4) *with $u \geqq 0$, $m > 1$. Then as $t \to \infty$*

$$t^k|u(x, t) - w_M(x, t)| \to 0 \tag{1.11}$$

uniformly in $x \in \mathbb{R}^N$, with $M = \int u_0(x)\, dx$ and $k = (m - 1 + (2/N))^{-1}$.

When u_0 has compact support and one space dimension, $N=1$, very detailed information on the convergence of solutions and interfaces to the self-similar profile is known (cf. [V1], [AV], [An]). Less is known in several space variables. In particular, the evolution of the support is studied by Caffarelli, Vazquez, and Wolanski [CVW] who prove that it expands at the same rate as w_M. More precisely, if supp $(u_0) \subset B_{R_0}(0)$, $M = \int u_0(x)\,dx$ $(u_0 \geqq 0)$, and we define

$$r(t) = \min\{|x|: u(x,t) > 0\}, \qquad R(t) = \max\{|x|: u(x,t) > 0\},$$

we have

$$\begin{aligned} &R(t) \geqq \rho_M(t), \\ &R(t) - r(t) \leqq 2R_0, \\ &R(t) \leqq \lambda \rho_M(t) + R_0 \quad \text{for some } \lambda = \lambda(m) > 0. \end{aligned} \tag{1.12}$$

In fact, we can do better for large times.

THEOREM 3. *As* $t \to \infty$

$$\lim_{t\to\infty} \frac{r(t)}{\rho_M(t)} = \lim_{t\to\infty} \frac{R(t)}{\rho_M(t)} = 1. \tag{1.13}$$

The proof of the theorem is just a variant of the method used to prove Theorem 2 in [FK]. Since by Theorem 1 our solutions become nonnegative, Theorems 2 and 3 apply to *all* solutions with positive mass, even if u_0 has changing sign.

Solutions with no sign restriction have been less studied. Existence and uniqueness of weak solutions are discussed in [BC], [BCP]. Continuity with a uniform modulus is established in [dB] and [S], while Hölder continuity is obtained in [dBF]. An application to model the evolution of the interface separating fresh and salt water has been proposed by de Josselin de Jong and van Duijn [JvD]. (See also [BH].)

Theorem 1 will be proved in § 2, where we also show how to obtain Theorem 3 and establish a convergence rate when $M = 0$.

We devote § 3 to the case $1 \geqq m > (N-2)_+/N$, where we have infinite speed of propagation. In that range of exponents, Theorem 2 is also valid for nonnegative solutions with finite mass. Although we do not have $u \geqq 0$ in finite time, we have convergence to a positive Barenblatt solution as $t \to \infty$ if $\int u_0(x)\,dx = M > 0$ with the rate (1.11) for all $u_0 \in L^1(\mathbb{R}^N)$.

We recall that in the range $0 < m < (N-2)/N$ $(N > 2)$ we have extinction in finite time for initial data in $L^p(\mathbb{R}^N) \cap L^1(\mathbb{R}^N)$ with $p = N(1-m)/2$. This settles the large-time behaviour.

In § 4 we extend Theorems 1 and 2 to the p-Laplacian equation

$$u_t = \operatorname{div}(|\nabla u|^{p-2}\nabla u) \tag{1.14}$$

with $p > 2$ under assumptions (1.2)-(1.5) on the initial data. This will be a key ingredient in our study of the case $M = 0$ for (1.1), which we will perform in § 5 restricted to one space dimension. If the moment $\int x u_0(x)\,dx \neq 0$, we find convergence to an antisymmetric profile obtained as the derivative of the Barenblatt profile for (1.14). Explicit computations for $m = 1$ show that similar phenomena should occur for $N > 1$.

2. Case $m > 1$. According to [BC] and [BCP], for every $u_0(x) \in L^1(\mathbb{R}^N)$ there is a unique function, $u \in C([0,\infty): L^1(\mathbb{R}^N))$, which solves (1.1) in the sense of distributions and takes on the initial data u_0. This solution is bounded for $t \geqq \tau > 0$. In fact, since the Maximum Principle holds, a bound for u is derived from the well-known estimates

for nonnegative solutions. Thus, if u_1 is the solution with initial data $u_{01} = u_0^+$ and u_2 has $u_{20} = u_0^-$ ($\equiv \max(-u_0, 0)$), then

$$-u_2(x,t) \leqq u(x,t) \leqq u_1(x,t). \tag{2.1}$$

Using the well-known estimate ("smoothing effect"; cf. [Ve], [V2]) we get

$$-c(M_{20}^{2/N}/t)^k \leqq u(x,t) \leqq c(M_{10}^{2/N}/t)^k, \tag{2.2}$$

where $M_{10} = \int u_0^+ \, dx$, $M_{20} = \int u_0^- \, dx$, $M = M_{10} - M_{20}$, and c is the exact constant appearing in the estimate for the self-similar solutions w. Throughout §§ 2–4, k will keep the value $(m-1+(2/N))^{-1}$.

The solutions have some further regularity: by [BCP, Prop. 1.6], $\nabla(|u|^{m-1}u) \in H^1_{\mathrm{loc}}(Q)$, $Q = \mathbb{R}^N \times (0,\infty)$. We may now apply the results of [dB], [dBF], and [S] and conclude that u is Hölder continuous in Q with constants and modulus depending only on M_{10} and M_{20}. This gives equicontinuity needed later in passing to the limit as $t \to \infty$.

We also have bounds for the support derived from what is known for u_1 and u_2. Thus, by (1.12) we know that u_1 and u_2, hence also u, vanish for

$$|x| \geqq Ct^{k/N}, \qquad C = c(m,N)(\max(M_{10}, M_{20}))^{(m-1)k/N} \tag{2.3}$$

if t is large. This estimate can be strengthened replacing $\max(M_{10}, M_{20})$ by M after we prove Theorem 1, but we do not need it for the moment.

Let us now define

$$M_1(t) = \int u^+(x,t)\, dx, \qquad M_2(t) = \int u^-(x,t)\, dx. \tag{2.4}$$

It is clear that $M_1(0) = M_{10}$, $M_2(0) = M_{20}$, and $M = M_1(t) - M_2(t)$ since there is conservation of mass. We also have Lemma 1 below.

LEMMA 1. *$M_1(t)$ and $M_2(t)$ are nonincreasing in t.*

Proof. Consider the solution v_τ to (1.1) with initial data at time $t = \tau$, $v_\tau(x,\tau) = u^+(x,\tau)$. By the Maximum Principle $u(x,t) \leqq v_\tau(x,t)$ for $t \geqq \tau$, $x \in \mathbb{R}^N$. Conservation of mass proves that

$$M_1(\tau) = \int v_\tau(x,\tau)\, dx = \int v_\tau(x,t)\, dx \geqq \int u^+(x,t)\, dx$$

whenever $t \geqq \tau$, and hence $M_1(\tau) \geqq M_1(t)$. The proof is similar for M_2. □

More generally it is known that for every convex function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ with $\phi(0) = 0$, the integral $\int \phi(u)\, dx$ is nonincreasing in time. Our lemma comes from the particular case $\phi(s) = s^+$ applied to u and $-u$. (We thank the referee for this observation.)

As a consequence of the monotonicity we may define the asymptotic masses

$$M_1 = \lim_{t\to\infty} M_1(t), \qquad M_2 = \lim_{t\to\infty} M_2(t). \tag{2.5}$$

Clearly M_1, $M_2 \geqq 0$ and $M_1 - M_2 = M > 0$.

We now recall the rescaling operation, which is very useful for formulating and proving asymptotic results. Given $\lambda > 0$, we associate to any solution u the function $T_\lambda u \equiv u_\lambda$ given by

$$(T_\lambda u)(x,t) = \lambda^k u(\lambda^{k/N}x, \lambda t) \tag{2.6}$$

with $k = (m-1+(2/N))^{-1}$ as above. It is easy to check that $T_\lambda u$ is again a solution of (1.1) with initial data $T_\lambda u_0(x) = \lambda^k u_0(\lambda^{k/N}x)$, which has exactly the same mass as u_0. Moreover, w_M is invariant under T_λ, i.e., $T_\lambda w_M \equiv w_M$.

In terms of T_λ the convergence result of [FK] simply says that

$$\lim_{\lambda\to\infty} T_\lambda u(x,1) = w_M(x,1) \tag{2.7}$$

uniformly in x. This holds for $u \geqq 0$. In the case of changing sign we have the next lemma.

LEMMA 2. *As* $\lambda \to \infty$

$$(T_\lambda u^+)(x,1) \to w_{M_1}(x,1), \tag{2.8}$$

$$(T_\lambda u^-)(x,1) \to w_{M_2}(x,1) \tag{2.9}$$

uniformly in $\mathbb{R}^N$.

Proof. Consider again the solution v_τ with initial data $u^+(x,\tau)$ for $\tau > 0$ (and large). By (2.7) we have

$$T_\lambda v_\tau(x,1) \to w_{M_1(\tau)}(x,1)$$

as $\lambda \to \infty$. Therefore for τ and λ large enough we obtain

$$|T_\lambda v_\tau(x,1) - w_{M_1}(x,1)| < \varepsilon$$

since $M_1(\tau) \to M_1$. We now recall that $u \leqq v_\tau$ for $t \geqq \tau$, so that

$$(T_\lambda u^+)(x,1) \leqq w_{M_1}(x,1) + \varepsilon \tag{2.10}$$

if λ is large enough. On the other hand, we know by (2.2) that the family $T_\lambda u^+(x,1)$ is uniformly bounded and that by (2.3) its supports are also contained in a fixed ball. Moreover, the family $\{T_\lambda u\}$ is uniformly equicontinuous on compact subsets of Q as explained above. Since $T_\lambda(u^+) = (T_\lambda u)^+$ the family $T_\lambda(u^+)$ is also equicontinuous. All this implies that along a sequence $\lambda_n \to \infty$, $T_\lambda(u^+)(x,1)$ converges to a function f such that

$$0 \leqq f(x) \leqq w_{M_1}(x,1),$$

$$\int f(x)\,dx = \lim \int T_{\lambda_n} u^+(x,1)\,dx = \lim M_1(\lambda_n) = M_1.$$

Together these two facts imply that $f \equiv w_{M_1}(\cdot, 1)$. The uniqueness of the limit implies that $T_\lambda u^+$ converges to f along any sequence, thus establishing (2.8). The proof of (2.9) is similar. □

COROLLARY 3. $M_2 = 0$.

Proof. The above convergences imply that

$$t^k u^+(0,t) \to c_1(m) M_1^{2k/N}, \tag{2.11}$$

$$t^k u^-(0,t) \to c_1(m) M_2^{2k/N}. \tag{2.12}$$

Since necessarily $M_1 > 0$, (2.12) *can only hold if* $M_2 = 0$. □

End of proof of Theorem 1. Take $\varepsilon \ll M$ and t_0 such that

$$\int u^-(x,t)\,dx \leqq \varepsilon$$

for $t \geqq t_0$. According to (1.12), the support of the solution v with initial data $v(x,0) = u^-(x,t_0)$ grows as $t \to \infty$ like

$$O(\varepsilon^\alpha t^\beta), \qquad \alpha = (m-1)k/N, \quad \beta = k/N.$$

By the Maximum Principle, the same happens for $u^-(x,t)$. In terms of the rescaled solution, this means that

$$T_\lambda u(x,1) \geqq 0 \quad \text{for } |x| \geqq C\varepsilon^\alpha$$

for $\lambda \geqq \lambda_1(\varepsilon)$. Now for $|x| \leqq C\varepsilon^\alpha \ll bM^{(m-1)k/N}$ we already know that $T_\lambda u(x,1) > 0$ if λ is large enough, $\lambda \geqq \lambda_2(\varepsilon)$, because of the uniform convergence (2.8). Therefore $T_\lambda u(x,1) \geqq 0$ for $\lambda \geqq \max(\lambda_1(\varepsilon), \lambda_2(\varepsilon)) = \lambda_3$ which means that $u(x,t) \geqq 0$ for every $x \in \mathbb{R}^N$ if $t \geqq \lambda_3$. □

Though we cannot obtain the same conclusion when the mass $M = 0$, as we will see in § 5, the above applies as far as Corollary 3, thus showing that $M_1 = M_2 = 0$. Thus we obtain the following result.

THEOREM 4. *Let u be a solution of* (1.1)-(1.4) *with* $\int u_0(x)\,dx = 0$. *Then*

$$\lim_{t\to\infty} t^k u(x,t) = 0 \quad \textit{uniformly in } \mathbb{R}^N.$$

Let us end this section by mentioning the modification needed in the paper [FK] to obtain the asymptotic behaviour of the support, Theorem 3. The method of finding a family of Barenblatt solutions w_τ, with growing masses that lie below the solution for $t \geqq \tau$ and letting $\tau \to \infty$, should be changed into a family from above with decreasing masses that tend to M as $\tau \to \infty$. That this can be done is already observed in Remarks 1 and 2 at the end of that paper for the case of exponents $1 \geqq m > (N-2)/N$.

3. Case $m \leqq 1$. The considerations of the previous section apply to the solutions of problem (1.1)-(1.5) when $1 \geqq m > (N-2)_+/N$ to establish the following result.

THEOREM 5. *If u is a solution with mass $M \in \mathbb{R}$, then*

$$\lim_{t\to\infty} t^k |u(x,t) - w_M(x,t)| = 0 \quad \textit{uniformly in } \mathbb{R}^N. \tag{3.1}$$

Here $w_M(x,t)$ is given by the same formula (1.9) if $m < 1$ and by the limit as $m \to 1$, which is precisely $ME(x,t)$ if $m = 1$. Observe that these solutions are positive everywhere in $Q = \mathbb{R}^N \times \mathbb{R}^+$.

The result is known for nonnegative solutions and proved in the same paper [FK, Remark 2, p. 562]. For $m = 1$ we may of course prove (3.1) directly using the representation formula.

To prove (3.1) for $m < 1$ and general u_0 we repeat the plan of § 2. We recall that existence and uniqueness of solutions can be found in [HP], while a uniform modulus of continuity is found at the end of [dB] and in [S]. We have estimates (2.1) and (2.2) (which make sense since $k > 0$ precisely for $m > (N-2)/N$) and the same rescaling operators. Lemma 1 is true without modifications.

In order to prove Lemma 2 we face the problem of infinite propagation, i.e., the supports are just $\mathbb{R}^N$. We replace control of supports with control over decay as $|x| \to \infty$. In fact we have the following lemma.

LEMMA 4. *The family $\{T_\lambda u(x,1)\}_{\lambda>1}$ decays uniformly as $|x| \to \infty$ like $O(|x|^{-\gamma})$, $\gamma = 2/(1-m)$.*

Proof. We use the following estimate from [HP]:

$$|u(x,t)| \leqq C\left[t^{-k}\left(\int_{B_{4R}(x)} |u_0| \right)^{2k/N} + (t/R^2)^{1/(1-m)} \right]. \tag{3.2}$$

We choose $4R + R_0 = |x|$, assuming that u_0 is supported in the ball $B_{R_0}(0)$ (as will be $T_r u_0$ for $r > 1$), so that the integral in the second member vanishes. Put $t = 1$ and $T_\lambda u$ instead of u to obtain the decay. □

Since $\gamma > N$ precisely for $m > (N-2)/N$ we have uniform integrability as $|x| \to \infty$. Therefore the family converges in $L^1(\mathbb{R}^N)$ weakly. In this way we may continue through the proofs of Lemma 2 and Corollary 3 and obtain (3.1). □

We remark that in the case $M > 0$, Theorem 5 implies that u will be positive for large t in at least a *core* of the form

$$|x| \leqq C(t)t^{k/N} \tag{3.3}$$

with $C(t) \to \infty$ as $t \to \infty$. But, as shown in (1.6), u may very well continue to be negative for some x, i.e., solutions may need an infinite time to become positive everywhere, so to speak. (In the example (1.6)-(1.8) for $m = 1$ the positive region extends to a distance of order of t, so $C(t) \cong t^{1-k/N} = t^{1/2}$.)

4. The *p*-Laplacian equation. The ingredients in our treatment of (1.1) can be summarized in the following list:

(i) An existence and uniqueness theory for the Cauchy problem with integrable data having compact support;
(ii) The maximum principle;
(iii) Conservation of mass;
(iv) Boundedness for positive solutions;
(v) Equicontinuity;
(vi) Invariance of the equation under mass-preserving rescaling;
(vii) Convergence of nonnegative solutions to a selfsimilar solution, invariant under rescaling;
(viii) Finite propagation (essential for Theorem 1).

All of these properties are true in the case of the equation

$$u_t = \operatorname{div}(|\nabla u|^{p-2}\nabla u) \equiv \Delta_p(u) \tag{4.1}$$

when $p > 2$. For existence and uniqueness under general initial conditions we refer to [dBH], (ii) and (iii) are well known, and (v) is proved by [dBF]. The equation is invariant under the rescaling

$$(T_\lambda u)(x, t) = \lambda^k u(\lambda^{k/N}x, \lambda t), \tag{4.2}$$

which formally equals (2.6), but now the value of k is $(p-2+(p/N))^{-1}$, a value which will be kept throughout this section and also in § 5 for $p = m+1$. The fundamental solutions are now of the form

$$\hat{w}_M(x, t) = t^{-k}(C - q|\xi|^{p/(p-1)})_+^{(p-1)/(p-2)}, \qquad \xi = xt^{-k/N}, \tag{4.3}$$

where C is again related to the mass $M = \int \hat{w}(x, t)\,dx$; $C = c(p, N)M^\alpha$ with $\alpha = p(p-2)k/N(p-1)$, and $q = ((p-2)/p)(k/N)^{1/(p-1)}$. The support of w is given by

$$|x| \leqq \hat{\rho}_M(t) = \hat{b}(p, N)(M^{p-2}t)^{k/N}. \tag{4.4}$$

A smoothing effect of the form

$$u(x, t) \leqq \hat{w}_M(0, t) \equiv \hat{c}(p, N)(M^{(p/N)}/t)^k \tag{4.5}$$

is proved in [V3] (cf. also [Ve]). Finite propagation holds exactly for $p > 2$ (cf. [DH], [HV]). Finally, the convergence of u to $\hat{w}_M$, as in (1.11), has been proved for $N = 1$ in [EV] and is extended to $N > 1$ in our paper [KV].

In this way the proofs of § 2 hold and Theorems 1-4 hold for problem (4.1), (1.2), (1.3) (with our present values of $\hat{\omega}_M$ and k).

Remarks. The explicit solutions $\hat{w}_M$ appear in [B] for $N=1, 2, 3$, and have been used by several authors. We have taken some care in writing them down with the different coefficients and exponents attached to them because they are sometimes misprinted in the literature. Observe also that many authors write $p+1$ instead of p (so that $\Delta_p(u) \equiv \operatorname{div}(|Du|^{p-1}Du)$ and then the range is $p>1$).

5. The porous medium equation with zero mass ($N=1$). Now we consider the problem

$$u_t = (|u|^{m-1}u)_{xx} \quad \text{for } (x, t) \in Q = \mathbb{R} \times (0, \infty), \tag{5.1}$$

$$u(x, 0) = u_0(x) \quad \text{for } x \in \mathbb{R}, \tag{5.2}$$

where $m>1$ and the initial data u_0 are integrable and have compact support and zero mass

$$\int u_0(x)\, dx = 0. \tag{5.3}$$

We also assume that the first moment is nonzero:

$$\int x u_0(x)\, dx \neq 0. \tag{5.4}$$

To be specific we take $P = -\int x u_0(x)\, dx > 0$. By (formally for the moment) integrating in x from $-\infty$ we obtain for the variable

$$v(x, t) = \int_{-\infty}^{x} u(y, t)\, dy \tag{5.5}$$

the problem

$$v_t = (|v_x|^{m-1}v_x)_x \quad \text{in } Q, \tag{5.6}$$

$$v(x, 0) = v_0(x) \quad \text{in } \mathbb{R}, \tag{5.7}$$

where $v_0(x) = \int_{-\infty}^{x} u_0(y)\, dy$ is bounded, and has compact support (because of (5.3)) and mass P; if $\operatorname{supp}(u_0) \subset [a, b]$ we have

$$\int_{\mathbb{R}} v_0(x)\, dx = \int_a^b dx \int_a^x u_0(y)\, dy = b \int u_0(y)\, dy - \int y u_0(y)\, dy = P.$$

Since the solution v is continuous with v_x continuous [dBF] and satisfies (5.6) in a strong sense (cf. [EV]), it is easy to show that v_x is a weak solution of (5.1), (5.2). By uniqueness $u = v_x$.

By the version of Theorem 1 for the p-Laplacian equation (with $p = m+1 > 2$), the solution of problem (5.6), (5.7) becomes nonnegative in finite time. We can now apply the results of [EV] to conclude that v converges to $\hat{w} = \hat{w}_P$ as $t \to \infty$ at a certain rate. Therefore in some sense v_x converges to $\hat{w}_x$ too. We show next that the convergence is uniform. In fact, in our case $N=1$, $p = m+1$ we have

$$\hat{w} = t^{-k}(C - q|\xi|^{(m+1)/m})_+^{m/(m-1)} \tag{5.8}$$

with $k = 1/2m$, $\xi = xt^{-k}$, and $C = C(m, P)$. Therefore,

$$z(x, t) \equiv \hat{w}_x(x, t) = -dt^{-1/m}\xi^{1/m}(C - q|\xi|^{(m+1)/m})_+^{1/(m-1)}, \tag{5.9}$$

where $d = q(m+1)/(m-1)$. It is antisymmetric in x, positive for $-\hat{\rho}(t) < x < 0$, and negative for $0 < x < \hat{\rho}(t)$, with $\hat{\rho}(t) = \hat{b}(P^{m-1}t)^k$. It decays like $t^{-1/m}$ and expands like $t^{1/2m}$, both rates different from the positive case ($t^{-1/(m-1)}$ and $t^{1/(m+1)}$, respectively).

We have for a general solution Theorem 6 below.

THEOREM 6. *As* $t \to \infty$

$$t^{1/m}|u(x,t) - z(x,t)| \to 0 \tag{5.10}$$

uniformly in $x \in \mathbb{R}$. *Moreover, u vanishes outside a region*

$$s_1(t) < x < s_2(t), \tag{5.11}$$

where

$$s_2(t) = \hat{\rho}_P(t) + o(1), \qquad s_2(t) \geqq \hat{\rho}_P(t), \tag{5.12}$$

$$s_1(t) = -\hat{\rho}_P(t) + o(1), \qquad s_1(t) \leqq -\hat{\rho}_P(t). \tag{5.13}$$

Proof. The result is a consequence of the asymptotic formulas obtained in [EV, § 5]. In fact s_1 and s_2 are the left- and right-hand interfaces of v, and for large t, $v(x,t) > 0$ precisely if $s_1(t) < x < s_2(t)$. Established are not only (5.12) and (5.13), but also estimates for the derivatives

$$s_2'(t) = \hat{\rho}_P'(t) + o(1/t), \tag{5.14}$$

$$s_1'(t) = -\hat{\rho}_P'(t) + o(1/t). \tag{5.15}$$

To go further we need to introduce the so-called "pressure variable"

$$\sigma(x,t) = \frac{m}{m-1} v^{(m-1)/m} \tag{5.16}$$

for which [EV] proves the fundamental estimate

$$(|\sigma_x|^{m-1}\sigma_x)_x \geqq -\frac{1}{2mt}, \tag{5.17}$$

which is exact for the (pressure of the) solutions $\hat{w}$. Moreover, we have the interface equation

$$(|\sigma_x|^{m-1}\sigma_x)(s_i(t), t) = -s_i'(t). \tag{5.18}$$

Combining (5.12)-(5.18) we obtain the estimate

$$|\sigma_x|^{m-1}\sigma_x + \frac{x}{2mt} = o\left(\frac{1}{t}\right) \tag{5.19}$$

in $s_1(t) < x < s_2(t)$. Integration gives

$$\sigma - \hat{\sigma} = o(t^{-1/2m}), \tag{5.20}$$

where $\hat{\sigma}$ is the pressure corresponding to $\hat{w}_P$. Now we observe that $\sigma_x = v_x v^{-1/m}$. Hence

$$u = c(m)\sigma^{1/(m-1)}\sigma_x \tag{5.21}$$

to obtain uniform convergence of u to $z = \hat{w}_x$ with an error that decays faster than $t^{-1/m}$ (we mention the exponent $1/m$ because it is the decay rate for z and u). □

As the consequence of (5.10), u will be positive in the region $-\hat{\rho}(t)(1-\varepsilon) < x < -\varepsilon\hat{\rho}(t)$ and negative in $\varepsilon\hat{\rho}(t) < x < (1-\varepsilon)\hat{\rho}(t)$ if $0 < \varepsilon < 1$ and $t \geqq t_\varepsilon$. On the other hand, it is easy to prove that v is nondecreasing in x for $x \leqq a$ and nonincreasing for $x \geqq b$ where $[a, b] \supset \operatorname{supp}(u_0)$. Therefore the only region where u changes sign for large t is $a \leqq x \leqq b$.

Similar results hold for $P>0$, i.e., $\int xu_0(x)\,dx>0$ (just change x in $-x$). On the contrary, we cannot handle the case where $M=0$, $P=0$.

Let us mention that solution (5.9) was published by Barenblatt and Zeldovich [BZ] in 1957, with the observation that it takes a dipole as initial data, $u_0(x)=\delta'(x)$. Application of (5.6) to study the propagation of (5.1) with $M=0$ appears, e.g., in [HV].

Finally, in order to have some insight about the situation in several dimensions, we discuss briefly the linear heat equation

$$u_t=\Delta u, \qquad (x,t)\in Q \tag{5.22}$$

with $u(x,0)=u_0(x)\in L^1(\mathbb{R}^N)$, $\int u_0(x)\,dx=0$, and moments

$$P_i=\int x_i u_0(x)\,dx \tag{5.23}$$

not all zero. Let $\vec{P}=(P_1,\cdots,P_N)$. Then the asymptotic development of $u(x,t)$ has as first term

$$u(x,t)\cong \sum_{i=1}^{N} P_i\cdot\frac{\partial E}{\partial x_i}=\frac{\vec{P}\cdot\vec{x}}{4t}E(x,t), \tag{5.24}$$

which is positive on the half-space $\vec{P}\cdot x>0$ and negative on $\vec{P}\cdot x<0$. Therefore there is a unique pattern (as in $N=1$) which has as parameters the amplitude and orientation of the moment vector $\vec{P}$. It is very likely that this situation holds for $m\neq 1$. □

Appendix.

A.1. At the suggestion of one of the referees we prove the necessity of the condition of compact support for the initial data u_0 (condition (1.4)) in Theorem 1. We do this by constructing an example of a solution that does not become nonnegative in finite time even though its initial data satisfy conditions (1.3) and (1.5).

We consider the solution $u(x,t)$ to (1.1) with $m>1$ and with an initial function $u_0(x)\in C(\mathbb{R}^N)$ that is positive in the ball $\{x:|x|<1\}$, negative for $|x|>1$, and satisfies (1.3) and (1.5). Let $v(x,t)$ be the solution with $v(x,0)$ equal to the positive part of $u_0(x)$. Then $v\geqq 0$ and $v\geqq u$ in $\mathbb{R}^n\times(0,\infty)$. Moreover, by Theorem 3 there is an increasing function of t, $R(t)$ (which behaves like t^k as $t\to\infty$ according to (1.13)), such that $v(\cdot,t)$ vanishes outside a ball of radius $R(t)$. This means that $u(x,t)\leqq 0$ for every $t>0$ and $|x|>R(t)$.

We now claim that in fact $u(x,t)<0$ whenever $|x|>R(t)$. To prove this, we fix a time $T>0$ and consider the solution $u(x,t)$ in a cylinder of the form $Q=B\times(0,T)$, where $B=B_\rho(a)$ is the ball of center a, radius $\rho>0$ in $\mathbb{R}^N$, and a and ρ satisfy

$$R(T)+\rho\leqq|a|. \tag{A.1}$$

Then we compare our solution u restricted to the cylinder Q with the solution $w(x,t)$ of the mixed problem

$$\begin{aligned} &w_t=\Delta(|w|^{m-1}w) && \text{in } Q,\\ &w(x,t)=0 && \text{for } |x-a|=\rho,\quad 0<t<T,\\ &w(x,0)=w_0(x) && \text{for } |x-a|<\rho \end{aligned} \tag{A.2}$$

for some suitable w_0 with $0\geqq w_0(x)\geqq u_0(x)$ in B. Now it is known [AP] that this problem admits solutions of the separated-variables form

$$w(x,t)=f(x)/(t+\tau)^\alpha \quad \text{with } \alpha=1/(m-1),$$

where $\tau>0$ is arbitrary and f is a smooth function that is negative in B. By taking τ large enough we may get $w(x,0)\geqq u(x,0)$ in B, and since also $w\geqq u$ on the lateral

boundary of Q, by the Maximum Principle we conclude that $w(x, t) \geq u(x, t)$. Hence $u(x, t) < 0$ in Q, which proves our claim. On the other hand, different variations of this example are easy to construct using the same basic idea.

A.2. Since we are commenting on the necessity of the hypotheses of Theorem 1, let us mention that the integrability assumption on u_0, condition (1.3) together with (1.4), can be weakened into the assumptions

$$u_0^- \in L^1(\mathbb{R}^N), \tag{A.3}$$

$$u_0^- \text{ has compact support,} \tag{A.4}$$

i.e., only the negative part of the initial data, $u_0^-(x) = \max(0, -u_0(x))$, needs to be integrable and have compact support. Of course, if this happens and $u_0^+ \notin L^1(\mathbb{R}^N)$ then $M = \infty$. The proof of this version of Theorem 1 follows from the standard case thanks to the Maximum Principle (consider the solutions $\bar{u}_n$ with initial data $\min(u(x), n)$, which satisfy Theorem 1 for large n and are smaller than u).

Acknowledgments. Several improvements in the text are due to a referee's observations, for which the authors are very grateful.

REFERENCES

[An] S. ANGENENT, *Large-time asymptotics for the porous media equation*, in Nonlinear Diffusion Equations and Their Equilibrium States I, W.-M. Ni, L. A. Peletier, and J. Serrin, eds., Springer-Verlag, Berlin, New York, 1988.

[Ar] D. G. ARONSON, *The porous medium equation*, in Some Problems in Nonlinear Diffusion, A. Fasano and M. Primicerio, eds., Lecture Notes in Math. 1224, Springer-Verlag, Berlin, New York, 1986.

[AP] D. G. ARONSON AND L. A. PELETIER, *Large-time behavior of solutions of the porous media equation in bounded domains*, J. Differential Equations, 39 (1981), pp. 378–412.

[AR] A. ALIKAKOS AND R. ROSTAMIAN, *On the uniformization of solutions of the porous medium equation in* $\mathbb{R}^N$, Israel J. Math, 47 (1984), pp. 270–290.

[AV] D. G. ARONSON AND J. L. VAZQUEZ, *Eventual* C^∞*-regularity and concavity for flows in one-dimensional porous media*, Arch. Rational Mech. Anal., 99 (1987), pp. 329–348.

[B] G. I. BARENBLATT, *On self-similar motions of compressible fluids in porous media*, Prikl. Mat. Mekh., 16 (1952), pp. 679–698. (In Russian.)

[BC] PH. BENILAN AND M. G. CRANDALL, *The continuous dependence on* ϕ *of the solutions of* $u_t - \Delta\phi(u) = 0$, Indiana Univ. Math. J., 30 (1981), pp. 162–177.

[BCP] PH. BENILAN, M. G. CRANDALL, AND M. PIERRE, *Solutions of the porous medium equation in* $\mathbb{R}^n$ *under optimal conditions on initial values*, Indiana Univ. Math. J., 33 (1984), pp. 51–87.

[BH] M. BERTSCH AND D. HILHORST, *On two new applications of Gurtin's coordinate transformation*, in Differential equations (Xanthi, 1987), Lecture Notes in Pure Appl. Math., 118, Dekker, New York, 1989, pp. 71–80.

[BZ] G. I. BARENBLATT AND Y. B. ZEL'DOVICH, *On dipole-type solutions in problems of nonstationary filtration of gas under polytropic regime*, Prikl. Mat. Mekh., 21 (1957), pp. 718–720.

[CVW] L. A. CAFFARELLI, J. L. VAZQUEZ, AND N. E. WOLANSKI, *Lipshitz continuity of solutions and interfaces to the N-dimensional porous medium equation*, Indiana Univ. Math. J., 36 (1987), pp. 323–401.

[DH] J. I. DIAZ AND M. A. HERRERO, *Estimates on the support of the solutions of some nonlinear elliptic and parabolic problems*, Proc. Roy. Soc. Edinburgh Sect. A, 89 (1981), pp. 249–258.

[dB] E. DI BENEDETTO, *Continuity of weak solutions to a general porous medium equation*, Indiana Univ. Math. J., 32 (1983), pp. 83–118.

[dBF] E. DI BENEDETTO AND A. FRIEDMAN, *Hölder estimates for nonlinear degenerate parabolic systems*, J. Reine Angew. Math., 357 (1985), pp. 1–22.

[dBH] E. DI BENEDETTO AND M. A. HERRERO, *On the Cauchy problem and initial traces of a degenerate parabolic equation*, Trans. Amer. Math. Soc., 314 (1989), pp. 187–224.

[EV] J. R. ESTEBAN AND J. L. VAZQUEZ, *Homogeneous diffusion in* $\mathbb{R}$ *with power-like nonlinear diffusivity*, Arch. Rational Mech. Anal., 103 (1988), pp. 39–80.

[FK] A. FRIEDMAN AND S. KAMIN, *The asymptotic behavior of gas in an N-dimensional porous medium*, Trans. Amer. Math. Soc., 262 (1980), pp. 551-563.

[HP] M. A. HERRERO AND M. PIERRE, *The Cauchy problem for* $u_t = \Delta(u^m)$ *when* $0 < m < 1$, Trans. Amer. Math. Soc., 291 (1985), pp. 145-158.

[HV] M. A. HERRERO AND J. L. VAZQUEZ, *On propagation properties of a nonlinear degenerate parabolic equation*, Comm. Partial Differential Equations, 7 (1982), pp. 145-158.

[JvD] G. DE JOSSELIN DE JONG AND C. J. VAN DUIJN, *Transverse dispersion from an originally sharp fresh-salt water interface caused by shear flow*, J. Hydrology, 84 (1986), pp. 55-79.

[K] S. KAMIN (KAMENOMOSTKAYA), *The asymptotic behavior of the solution of the filtration equation*, Israel J. Math., 14 (1973), pp. 76-87.

[KV] S. KAMIN AND J. L. VAZQUEZ, *Fundamental solutions and asymptotic behaviour for the p-Laplacian equation*, Rev. Mat. Iberoamericana, 4 (1988), pp. 339-354.

[OKC] O. A. OLEINIK, A. S. KALASHNIKOV AND Y. L. CZHOU, *The Cauchy problem and boundary problems for equations of the type of nonstationary filtration*, Izv. Akad. Nauk SSSR Ser. Mat., 22 (1958), pp. 667-704. (In Russian.)

[P] L. PELETIER, *The porous medium equation*, in Applications of Nonlinear Analysis in the Physical Sciences, H. Amann, N. Bazley, and K. Kirchgässner, eds., Pitman, Boston, 1981.

[S] P. SACKS, *Continuity of solutions of a singular parabolic equation*, Nonlinear Anal., 7 (1983), pp. 387-409.

[V1] J. L. VAZQUEZ, *Asymptotic behaviour and propagation properties of the one-dimensional flow of gas in a porous medium*, Trans. Amer. Math. Soc., 277 (1983), pp. 507-527.

[V2] ———, *Symétrisation pour* $u_t = \Delta\phi(u)$ *et applications*, C. R. Acad. Sci. Paris, Sér. I Math., 295 (1982), pp. 71-74.

[V3] ———, *Symmetrization in nonlinear parabolic problems*, Portugaliae Math., 41 (1982), pp. 339-345.

[Ve] L. VERON, *Coercivité et propriétés regularisantes des semi-groups nonlinéaires dans les espaces de Banach*, Ann. Fac. Sci. Toulouse, 1 (1979), pp. 171-200.

[ZK] I. B. ZEL'DOVICH AND A. S. KOMPANEEZ, *On the theory of heat conduction depending on temperature. Lectures dedicated on the 70th Anniversay of A. F. Joffe*, Akad. Nauk SSSR, 1950, pp. 61-71. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 46-71, January 1991

004

UNIFORM ENERGY DECAY RATES FOR EULER-BERNOULLI EQUATIONS WITH FEEDBACK OPERATORS IN THE DIRICHLET/NEUMANN BOUNDARY CONDITIONS*

J. BARTOLOMEO† AND R. TRIGGIANI‡

Abstract. The uniform stabilization problem is studied for the Euler-Bernoulli equation (though the methods apply also to the corresponding nonconstant coefficient case) defined in a smooth, bounded domain Ω of R^n, with suitable dissipative boundary feedback operators. These either are active in both the Dirichlet and Neumann boundary conditions, or are active in only the Dirichlet and inactive in the Neumann boundary condition. The uniform stabilization results presented are fully consistent with recently established exact controllability and optimal regularity theories for the solutions, which in fact motivate the choices of functional spaces in the first place. In particular, these uniform stabilization results require no geometrical conditions on Ω in the case of active Dirichlet/Neumann feedback operators, and require some geometrical conditions on Ω in the case of an active feedback operator only in the Dirichlet boundary condition, as is the case of recent exact controllability theories [I. Lasiecka and R. Triggiani, *SIAM J. Control Optim.*, 27 (1989), pp. 330-373]. Moreover, the forms of the dissipative feedback controls are natural consequences of (i) the type of boundary conditions selected; (ii) the choice that the control in the lowest boundary condition be L_2 in time and space.

Key words. Euler-Bernoulli equations, uniform stabilization

AMS(MOS) subject classifications. 35Q20, 35B37, 35B40

1. Introduction, preliminaries, and statement of main results.

1.1. Introduction and literature. Let Ω be an open bounded domain in R^n, n typically $\geqq 2$, with sufficiently smooth boundary Γ. In Ω we consider the Euler-Bernoulli mixed problem in $w(t, x)$ on an arbitrary time interval $(0, T]$ with Dirichlet and Neumann boundary conditions:

$$w_{tt} + \Delta^2 w = 0 \qquad \text{in } (0, T] \times \Omega, \tag{1.1a}$$

$$w(0, x) = w_0(x), \quad w_t(0, x) = w_1(x) \qquad \text{in } \Omega, \tag{1.1b}$$

$$w(t, \sigma) = g_1(t, \sigma) \qquad \text{in } (0, T] \times \Gamma, \tag{1.1c}$$

$$\frac{\partial w}{\partial \nu}(t, \sigma) = g_2(t, \sigma) \qquad \text{in } (0, T] \times \Gamma \tag{1.1d}$$

with nonhomogeneous forcing terms (control functions) g_1 and g_2 in the Dirichlet and Neumann boundary conditions. There has recently been a keen resurgence of interest (e.g., [34], [10], [16], and references cited therein) in plate equation theory, of which the Euler-Bernoulli equation (1.1a) is a canonical model, presumably stimulated by two main sources: (i) renewed studies in the dynamics, feasibility, and implementation of so-called large-scale flexible structures envisioned to be employed in space; (ii) recent mathematical advances in regularity theory of second-order mixed hyperbolic

* Received by the editors July 12, 1989; accepted for publication (in revised form) January 5, 1990. This research was partially supported by National Science Foundation grants DMS-8796320 and DMS-8902811 and by Air Force Office of Scientific Research grant AFOSR-87-0321. This paper was presented by the second author at the International Workshops held in Voran, Austria, July 1988. An announcement has appeared in the Proceedings, International Series of Numerical Mathematics, 91 (1989), Birkhäuser Verlag, Basel, pp. 91-400. This work is partially based on the Ph.D. thesis of the first author.

† Department of Mathematics, University of Florida, Gainesville, Florida 32611.

‡ Department of Applied Mathematics, Thornton Hall, University of Virginia, Charlottesville, Virginia 22903.

problems (canonically, the wave equation) of both Dirichlet type [12], [21], [22], [35]) and Neumann type [21]–[24], [36] with L_2-boundary data. In either case, a prime thrust of motivation has come from dynamical control studies, at either the engineering or the theoretical level.

With reference to the specific problem (1.1), we cite [15], [18], [19], [31] for optimal regularity theory and exact controllability theory with respect to classes of interest for the initial data $\{w_0, w_1\}$ and of boundary data $\{g_1, g_2\}$, which markedly improved upon regularity of prior literature [17].

In the present paper, we focus on the problem of *boundary feedback uniform stabilization* for the dynamics (1.1) by *explicit feedback operators*, to be more properly defined below. Our results are fully consistent with the corresponding exact controllability results [18], [19] with respect not only to the function spaces for $\{g_1, g_2\}$ and $\{w, w_t\}$ as mentioned above, but also to the lack of geometrical conditions on Ω when both g_1 and g_2 are active, or to the presence of similar geometrical conditions on Ω—to be expected—when only g_1 is active while g_2 is taken $g_2 \equiv 0$. We note, in passing, that uniform stabilization of (1.1) by means of a feedback operator acting on $\{w, w_t\}$, (defined in terms of the algebraic Riccati operator that arises in the study of the optimal quadratic cost problem on an infinite horizon $0 \leqq t \leqq T = \infty$) has already been achieved in the abstract treatment of [5] (see in particular Appendix 2 of [5]) as a consequence—among others—of the optimal regularity and exact controllability results mentioned before. (Indeed, because of these results, the abstract treatment of the wave equation in [28, § 5] also covers, mutatis mutandis, the plate problem (1.1), as noted in [5]. See also [30].)

Mathematically, the present work is guided by, and partially rests upon, techniques developed in two main sources: (i) the studies of exact controllability [18], [19] for problem (1.1); (ii) the study of uniform stabilization of the wave equation with boundary feedback in the Dirichlet boundary conditions [20] and in the Neumann boundary conditions (B.C.) [39]. Of course, these studies must be seen, in turn, in the context of recent investigations including (a) uniform stabilization of the wave equation with feedback in the Neumann boundary conditions [2], [8], [9], [20, § 4]; (b) regularity theory for hyperbolic equations in [12], [21], [22], [35] as well as corresponding exact controllability theory [13], [14], [25], [7], [38]; (c) exact controllability results for Euler–Bernoulli equations with different boundary conditions [13], [14], [26], [27]; and (d) corresponding optimal quadratic cost problems [4], [28], and [5]. We stress the following point of view: we choose g_1, say, in "open loop" form to be in $L^2(0, T; L^2(\Gamma))$, $T < \infty$. This determines the corresponding solution of (1.1) with $w_0 = w_1 = g_2 = 0$ to be: $\{w, w_t\} \in C([0, T]; Z)$, where Z is the space identified in (1.6) below. This is an optimal regularity result [15], [19]. Next, we choose g_2 (respectively, $\{w_0, w_1\}$) such that the corresponding solution of (1.1) with $g_1 = w_0 = w_1 = 0$ (respectively, $g_1 = g_2 = 0$) also produces $\{w, w_t\} \in C([0, T]; Z)$, again as an optimal regularity result. This leads to $\{w_0, w_1\} \in Z$ and $g_2 \in L^2(0, T; H^{-1}(\Gamma))$ [15], [19]. In other words, only one choice is made, that $g_1 \in L^2(0, T; L^2(\Gamma))$; then, we work with other data and resulting solutions in corresponding optimal spaces. Our solution of the uniform stabilization problem below is fully consistent with these "open-loop" considerations: uniform stabilization will be achieved in the space Z with controls in feedback form $g_1 \in L^2(0, \infty; L^2(\Gamma))$ and $g_2 \in L^2(0, \infty; H^{-1}(\Gamma))$; see Theorem 1.2 below. For other uniform stabilization results for plates, we refer to the monograph [10], as well as to [11], where, however, higher-order boundary conditions are treated. It should be emphasized that all these problems are very sensitive to the particular choice of boundary conditions, which in turn determine the appropriate spaces of solutions. The lower boundary

conditions considered in the present paper, which yield low-regularity spaces for the solutions, make it necessary—unlike the case of higher-order boundary conditions—to transform the original problem (1.1) in w to a new problem ((3.16) in p below), through a suitable change of variable ((3.12) below from w to p), before applying the correct multipliers (which are different from the multiplers used for higher-order boundary conditions). The case of feedback in $L^2(0, \infty; L^2(\Gamma))$ only in the Neumann B.C. requires a different state space of optimal regularity, $L^2(\Omega) \times H^{-2}(\Omega)$, and different multipliers, and will be reported elsewhere [33]. Sharp results on the lack of uniform stabilization are in [40].

1.2. Formulation of the uniform stabilization problem and main statements. Throughout the paper, we let $A: L^2(\Omega) \supset \mathcal{D}(A) \to L^2(\Omega)$ be the positive self-adjoint operator defined by

$$Af = \Delta^2 f, \qquad \mathcal{D}(A) = H^4(\Omega) \cap H_0^2(\Omega). \tag{1.2}$$

We have [6], [19, App. C]

$$\mathcal{D}(A^{1/4}) = H_0^1(\Omega), \quad \mathcal{D}(A^{3/4}) = V, \quad V = \left\{ f \in H^3(\Omega): f|_\Gamma = \left.\frac{\partial f}{\partial \nu}\right|_\Gamma = 0 \right\} \tag{1.3}$$

with equivalent norms. Thus, for $f \in \mathcal{D}(A^{1/4}) = H_0^1(\Omega)$,

$$\begin{aligned} &\|f\|_{\mathcal{D}(A^{1/4})} = \|A^{1/4} f\|_{L^2(\Omega)}, \\ &\text{equivalent to } \|f\|_{H^1(\Omega)}, \\ &\text{in turn equivalent to } \left\{ \int_\Omega |\nabla f|^2 \, d\Omega \right\}^{1/2} \end{aligned} \tag{1.4}$$

by Poincaré inequality. Similarly, for $f \in \mathcal{D}(A^{3/4}) = V$,

$$\|f\|_{\mathcal{D}(A^{3/4})} = \|A^{3/4} f\|_{L^2(\Omega)} \qquad \text{equivalent to } \left\{ \int_\Omega |\nabla(\Delta f)|^2 \, d\Omega \right\}^{1/2} \tag{1.5}$$

Our optimal space will be

$$Z = H^{-1}(\Omega) \times V' = [\mathcal{D}(A^{1/4})]' \times [\mathcal{D}(A^{3/4})]' \tag{1.6}$$

where $'$ denotes duality with respect to the $L^2(\Omega)$-topology.

Next, let $g_1 = g_2 = 0$ in (1.1). Then, the corresponding evolution of (1.1) is governed by the operator

$$\mathcal{A}_0 = \begin{vmatrix} 0 & I \\ -A & 0 \end{vmatrix},$$

which generates a strongly continuous *unitary* group on the space $\mathcal{D}(A^{1/2}) \times L^2(\Omega)$ with domain $\mathcal{D}(\mathcal{A}_0) = \mathcal{D}(A) \times \mathcal{D}(A^{1/2})$ and hence on the space Z of our interest with domain $\mathcal{D}(\mathcal{A}_0) = \mathcal{D}(A^{1/4}) \times [\mathcal{D}(A^{1/4})]' = H_0^1(\Omega) \times H^{-1}(\Omega)$, denoted by $e^{\mathcal{A}_0 t}$. Thus the free solutions of (1.1) with $g_1 = g_2 = 0$ are norm preserving on Z:

$$\|[w(t), w_t(t)]\|_Z \equiv \|e^{\mathcal{A}_0 t}[w_0, w_1]\|_Z \equiv \|[w_0, w_1]\|_Z, \qquad t \in R.$$

With this well-known result at hand, we can state the aim of the paper. Motivated by, and consistent with, the function spaces in the optimal regularity and exact controllability theories [15], [18], [19] of (1.1), we will study the question of existence and construction of explicit boundary feedback operators $\mathcal{F}_1$ and $\mathcal{F}_2$ based on the velocity w_t

$$w_t \in [\mathcal{D}(A^{3/4})]' = V' \to \mathcal{F}_1(w_t) \in L^2(0, \infty; L^2(\Gamma)), \tag{1.7}$$

$$w_t \in [\mathcal{D}(A^{3/4})]' = V' \to \mathcal{F}_2(w_t) \in L^2(0, \infty; H^{-1}(\Gamma)), \tag{1.8}$$

such that the boundary feedback functions

$$g_1 = \mathcal{F}_1(w_t), \qquad g_2 = \mathcal{F}_2(w_t) \tag{1.9}$$

inserted in (1.1c, d) produce a *strongly continuous* (*s.c.*) (*feedback*) *semigroup on Z that is exponentially stable in the uniform operator norm $\mathcal{L}(Z)$ of the space Z in* (1.6). (The feedback $B^*P[w, w_t]$ based on the Riccati operator P, referred to in § 1.1, acts instead on the full pair $\{w, w_t\}$ [5].)

Choice of operators $\mathcal{F}_1$ and $\mathcal{F}_2$. It is justified in § 2 (see (2.2)), that the following choices of $\mathcal{F}_1$ and $\mathcal{F}_2$:

$$g_1 = \mathcal{F}_1(w_t) = -k_1^2(x) G_1^* A^{-1/2} w_t = -k_1^2(x) G_1^* A A^{-3/2} w_t = -k_1^2(x) \left.\frac{\partial \Delta(A^{-3/2} w_t)}{\partial \nu}\right|_\Sigma, \tag{1.10}$$

$$\begin{aligned} g_2 = \mathcal{F}_2(w_t) &= -k_2(x)\Lambda^2 k_2(x)[G_2^* A^{-1/2} w_t] \\ &= -k_2(x)\Lambda^2 k^2(x)[G_2^* A A^{-3/2} w_t] \\ &= k_2(x)\Lambda^2 k_2(x)[\Delta(A^{-3/2} w_t)]_\Sigma \end{aligned} \tag{1.11}$$

provide reasonable *candidates* for the uniform stabilization problem of (1.1), in the sense that the closed-loop feedback dynamics with (1.10) and (1.11) inserted in (1.1c, d), respectively, is well posed in the semigroup sense in Z, and all of its solutions originating in Z decrease as $t \to +\infty$ in the Z-norm. (This, however, does not say that such Z-norms decrease to zero as $t \to +\infty$, let alone in the uniform norm of $\mathcal{L}(Z)$. To show this conclusion will be our major task.) In (1.10), (1.11) we have that:

(a) (1.12) $k_i(x)$ = smooth functions on Γ, $k_i(x) \geqq k_0 \geqq 0$;

(b) (1.13) Λ: isomorphism $H^s(\Gamma)$ onto $H^{s-1}(\Gamma)$, self-adjoint on $L^2(\Gamma)$ so that for $s = 1$, if ∇_σ denotes the tangential gradient on Γ,

$$\|\Lambda g\|_{L^2(\Gamma)} = \|g\|_{H^1(\Gamma)} = \left\{ \int_\Gamma |\nabla_\sigma g|^2 + g^2 \, d\Gamma \right\}^{1/2}; \tag{1.14}$$

(c) The operators G_i^* are the adjoints, in the sense that

$$(G_i g, v)_{L^2(\Omega)} = (g, G_i^* v)_{L^2(\Gamma)}, \qquad g \in L^2(\Gamma), \, v \in L^2(\Omega) \tag{1.15}$$

of the operators G_i defined by $G_1 g_1 = v$, respectively, $G_2 g_2 = y$, where

$$\Delta^2 v = 0 \quad \text{in } \Omega, \qquad \Delta^2 y = 0 \quad \text{in } \Omega, \tag{1.16a}$$

$$v|_\Gamma = g_1 \quad \text{on } \Gamma, \qquad y|_\Gamma = 0 \quad \text{on } \Gamma, \tag{1.16b}$$

$$\left.\frac{\partial v}{\partial \nu}\right|_\Gamma = 0 \quad \text{on } \Gamma, \qquad \left.\frac{\partial y}{\partial \nu}\right|_\Gamma = g_2 \quad \text{on } \Gamma. \tag{1.16c}$$

Elliptic theory [17], [32] gives for any $s \in R$

$$G_1: \text{continuous } H^s(\Gamma) \to H^{s+1/2}(\Omega), \tag{1.17}$$

$$G_2: \text{continuous } H^s(\Gamma) \to H^{s+3/2}(\Omega). \tag{1.18}$$

Moreover, it is proved by Green's theorem that [19, Lemma 2.0 and Lemma 4.0, respectively]

$$G_1^* A f = \left.\frac{\partial(\Delta f)}{\partial \nu}\right|_\Gamma, \qquad f \in \mathcal{D}(A), \tag{1.19}$$

$$G_2^* A f = -(\Delta f)|_\Gamma, \qquad f \in \mathcal{D}(A). \tag{1.20}$$

Identities (1.19), (1.20) are used in the last steps of (1.10) and (1.11), respectively.

Thus the resulting candidate feedback system, whose stability properties in Z we will investigate, is

$$w_{tt}+\Delta^2 w=0 \qquad \text{in } (0,\infty)\times\Omega=Q, \tag{1.21a}$$

$$w(0,x)=w_0(x)\colon\ w_t(0,x)=w_1(x) \qquad \text{in } \Omega, \tag{1.21b}$$

$$w|_\Sigma=-k_1(x)\left.\frac{\partial\Delta(A^{-3/2}w_t)}{\partial\nu}\right|_\Sigma \qquad \text{in } (0,\infty)\times\Gamma=\Sigma, \tag{1.21c}$$

$$\left.\frac{\partial w}{\partial\nu}\right|_\Sigma=k_2(x)\Lambda^2k_2(x)[\Delta(A^{-3/2}w_t)]_\Sigma \qquad \text{in } \Sigma. \tag{1.21d}$$

Using the techniques of [21], [37], problem (1.21) can be rewritten more conveniently in abstract form as

$$\frac{d}{dt}\begin{vmatrix} w\\ w_t\end{vmatrix}=\mathscr{A}\begin{vmatrix} w\\ w_t\end{vmatrix} \quad \text{on } Z, \tag{1.22}$$

$$\begin{aligned}&\mathscr{A}=\begin{vmatrix} 0 & I\\ -A & -A[G_1k_1^2G_1^*A^{-1/2}+G_2k_2\Lambda^2k_2G_2^*A^{-1/2}]\end{vmatrix},\\ &\mathscr{D}(\mathscr{A})=\{y\in Z\colon \mathscr{A}y\in Z\}.\end{aligned} \tag{1.23}$$

A more explicit description of $\mathscr{D}(\mathscr{A})$ will be given below. Our main results are as follows.

THEOREM 1.1. (i) (Well-posedness on Z). *The operator $\mathscr{A}$ in* (1.23) *is dissipative on $Z=[\mathscr{D}(A^{1/4})]'\times[\mathscr{D}(A^{3/4})]'$, see* (1.6), *and satisfies here*: range $(\lambda I-\mathscr{A})=Z$ *for* $\lambda>0$. *Thus, by the Lumer–Phillips theorem, $\mathscr{A}$ generates a strongly continuous contraction semigroup $e^{\mathscr{A}t}$ on Z. The resolvent operator $R(\lambda,\mathscr{A})$ is given by*

$$R(\lambda,\mathscr{A})=\begin{vmatrix}(I-V^{-1}(\lambda))/\lambda & V^{-1}(\lambda)A^{-1}\\ -V^{-1}(\lambda) & \lambda V^{-1}(\lambda)A^{-1}\end{vmatrix}, \tag{1.24}$$

$$V(\lambda)=(I+\lambda G_1k_1^2G_1^*A^{-1/2}+\lambda G_2k_2\Lambda^2k_2G_2^*A^{-1/2}+\lambda^2A^{-1}) \tag{1.25}$$

and is compact on Z for $\operatorname{Re}\lambda>0$. *Moreover, $0\in\rho(\mathscr{A})$, the resolvent set of $\mathscr{A}$.*

(ii) (L_2-boundedness in time of feedback operators.) *For $\{w_0,w_1\}\in Z$, we have for problem* (1.21)

$$-w|_\Sigma=k_1^2G_1^*A^{-1/2}w_t=k_1^2\frac{\partial\Delta(A^{-3/2}w_t)}{\partial\nu}\in L^2(0,\infty;L^2(\Gamma)), \tag{1.26}$$

$$-\left.\frac{\partial w}{\partial\nu}\right|_\Sigma=k_2\Lambda^2k_2G_2^*A^{-1/2}w_t=k_2\Lambda^2k_2[\Delta(A^{-3/2}w_t)]_\Sigma\in L^2(0,\infty;H^{-1}(\Gamma)). \tag{1.27}$$

More precisely

$$\int_0^\infty\|w\|^2_{L^2(\Gamma)}\,dt=\int_0^\infty\|k_1^2G_1^*A^{-1/2}w_t\|^2_{L^2(\Gamma)}\,dt\leqq\|\{w_0,w_1\}\|^2_Z, \tag{1.28}$$

$$\int_0^\infty\left\|\frac{\partial w}{\partial\nu}\right\|^2_{L^2(\Gamma)}dt=\int_0^\infty\int_\Gamma\|\Lambda k_2G_2^*A^{-1/2}w_t\|^2_{L^2(\Gamma)}\,dt\leqq\|\{w_0,w_1\}\|^2_Z. \tag{1.29}$$

(iii) *Now let $k_2\equiv0$, i.e., $(\partial w/\partial\nu)=0$ on Σ in* (1.21d). *Then the resolvent $R(\lambda,\mathscr{A})$ is well defined and compact on Z also on the imaginary axis, and hence for all* $\operatorname{Re}\lambda\geqq0$, *provided that the following elliptic uniqueness property holds true: with $\lambda>0$, if ϕ solves*

$$\Delta^2\phi=\lambda\phi \qquad \text{in } \Omega \tag{1.30}$$

$$\phi|_\Gamma=\left.\frac{\partial\phi}{\partial\nu}\right|_\Gamma=\left.\frac{\partial(\Delta\phi)}{\partial\nu}\right|_\Gamma=0 \tag{1.31}$$

(*elliptic problem with three homogeneous boundary conditions*) *then $\phi\equiv0$.*

THEOREM 1.2. (Uniform stabilization on Z with both feedback operators in the absence of geometrical conditions on Ω.) *The following property holds for the feedback problem* (1.21), *or* (1.22), (1.23): *there are constants* $M, \delta > 0$ *such that, if* $k_0 > 0$ *in* (1.12),

$$\left\| \begin{vmatrix} w(t) \\ w_t(t) \end{vmatrix} \right\|_Z = \left\| e^{\mathcal{A}t} \begin{vmatrix} w_0 \\ w_1 \end{vmatrix} \right\|_Z \leqq M e^{-\delta t} \left\| \begin{vmatrix} w_0 \\ w_1 \end{vmatrix} \right\|_Z, \quad t \geqq 0. \tag{1.32}$$

THEOREM 1.3. (Uniform stabilization on Z with only the first feedback operator g_1 and $g_2 \equiv 0$, in the presence of geometrical conditions on Ω.) *Consider the feedback problem* (1.1) *with* g_1 *given by* (1.10) *with* $k_1(x) \geqq k_0 > 0$ *while* $g_2 \equiv 0$, *i.e.*, (1.21) *with* $(\partial w/\partial \nu) = 0$ *on* Σ. *Then there are constants* $M, \delta > 0$ *such that the uniform decay* (1.32) *holds true, provided that* Ω *satisfies the following geometrical conditions: there exists a smooth vector field* $h(x) \in [C^2(\bar{\Omega})]^n$ *such that*

(i) (1.33) $h \cdot \nu \geqq \gamma > 0$ *on* Γ, ν = *unit outward normal*, γ = *constant*;

(ii) *There exists a positive constant* $\rho > 0$ *such that*

$$\int_\Omega H(x) v(x) \cdot v(x)\, d\Omega \geqq \rho \int_\Omega |v(x)|^2_{R^n}\, d\Omega \quad \forall v(x) \in [L^2(\Omega)]^n, \tag{1.34}$$

$$H(x) = \begin{vmatrix} \dfrac{\partial h_1}{\partial x_1} & \cdots & \dfrac{\partial h_1}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial h_n}{\partial x_1} & \cdots & \dfrac{\partial h_n}{\partial x_n} \end{vmatrix}. \tag{1.35}$$

(*A checkable condition for* (ii) *to hold true is that the symmetric matrix* $H(x) + H^*(x)$ *be uniformly positive definite on* $\bar{\Omega}$).

Both conditions (i), (ii) *are satisfied automatically with* $h(x) = x - x_0$ *if*

$$(x - x_0) \cdot \nu \geqq \gamma > 0 \quad \text{on } \Gamma. \tag{1.36}$$

Remark 1.1. In the proof of Theorem 1.2, it will suffice to take a radial vector field $h(x) = x - x_0$. Then, in this case, the "const" in (3.2) below can be *explicitly* estimated (as in Theorem 1.3A below). As a consequence, the proof in [41, Thm. 4.1] provides an *explicit* estimate of the constant δ of decay rate in (1.32). The same conclusion that δ in (1.32) can be *explicitly* estimated also holds true for Theorem 1.3 in the case where the vector field $h(x)$ postulated there is radial (or linear), in which case Theorem 1.3B is not needed as $K_2 = 0$ in (1.37). The loss of explicit control of the "const" in (3.2) occurs in the proof of Theorem 1.3B, which is needed for a general vector field $h(x)$ (nonlinear) to absorb a "lower-order term." These considerations on the explicit estimate, or lack thereof, of the constant δ in (1.32) are *irrespective* of whether we use criterion (3.2) or the equivalent (*) in Remark 3.2: a radial vector field yields an explicit δ, while a more general vector field requires "absorption" of "lower-order terms," and this step loses control of the constant (in the case of criterion (*) in Remark 3.2, by virtue of the proof by contradiction in the compactness/unique argument in, say [19]. □

(By a local variation of the proof at the level of (3.54), (3.55) below, we may take $\gamma = 0$; see [14], [19, footnote 2].) The proof of Theorem 1.3 will be broken up in § 3 into two main results, Theorem 1.3A, B below.

THEOREM 1.3A. *Consider the feedback problem* (1.21) *with* $\partial w/\partial \nu = 0$ *on* Σ *in* (1.21d), *i.e., with* $k_2 \equiv 0$. *Then, under assumptions* (i)/(1.33) and (ii)/(1.34) *in Theorem*

1.3, *we have for* $\beta > 0$

$$\int_0^\infty e^{-2\beta t} E(t)\, dt \leqq K_1 E(0) + K_2 \int_0^\infty e^{-2\beta t} [\|A^{-1/2} w(t)\|^2_{L^2(\Omega)} + \|A^{-1} w_t(t)\|^2_{L^2(\Omega)}]\, dt + K_3 \int_0^\infty e^{-2\beta t} \|w(t)|_\Gamma\|^2_{L^2(\Gamma)}\, dt \tag{1.37}$$

where $E(t) \equiv \|A^{-1/4} w(t)\|^2_{L^2(\Omega)} + \|A^{-3/4} w_t(t)\|^2_{L^2(\Omega)}$, *and where the constants* K_1, K_2, K_3—*which can be explicitly estimated from the proof below—are independent of* β, $0 < \beta \leqq \beta_0 < \infty$. *Also, in particular,* $K_2 = 0$ *if the assumed vector field* $h(x)$ *is radial (or linear)* (K_2 *is proportional to* $\max |\nabla(\operatorname{div} h)|$ *over* $\bar{\Omega}$).

THEOREM 1.3B. *Consider the feedback problem* (1.21) *with* $\partial w/\partial \nu = 0$ *on* Σ *in* (1.21d). *Then, under uniqueness property assumption* (iii) *of Theorem* 1.3, *we have: for every* $\varepsilon > 0$ *there exists* C_ε *such that for every* $0 < \beta$

$$\int_0^\infty e^{-2\beta t} [\|A^{-1/2} w(t)\|^2_{L^2(\Omega)} + \|A^{-1} w_t(t)\|^2_{L^2(\Omega)}]\, dt \leqq \varepsilon \int_0^\infty e^{-2\beta t} [\|A^{-1/4} w(t)\|^2_{L^2(\Omega)} + \|A^{-3/4} w_t(t)\|^2_{L^2(\Omega)}] + C_\varepsilon E(0). \tag{1.38}$$

2. Preliminaries and proof of Theorem 1.1 (sketch).

Step 1. (Abstract model for the closed-loop problem.) We follow the conceptual approach of [21], [37], [39] for the wave equation, and of [19], [26], and [27] for the Euler–Bernoulli problem. The abstract differential version of (1.1)—which corresponds to the integral version of [19, eq. (2.3)] defining the explicit input $\{g_1, g_2\} \to$ solution $\{w, w_t\}$ map—is given in additive form by

$$w_{tt} = -Aw + AG_1 g_1 + AG_2 g_2 \tag{2.1a}$$

where the operators G_i are defined by (1.16), and where the operator A on the right of (2.1a) is the extension, with the same symbol $L^2(\Omega) \to [\mathscr{D}(A)]'$ of the original operator in (1.2). The corresponding first-order equation is

$$\frac{d}{dt} \begin{vmatrix} w \\ w_t \end{vmatrix} = \begin{vmatrix} 0 & I \\ -A & 0 \end{vmatrix} \begin{vmatrix} w \\ w_t \end{vmatrix} + \begin{vmatrix} 0 \\ AG_1 g_1 + AG_2 g_2 \end{vmatrix}. \tag{2.1b}$$

Since $\left|\begin{smallmatrix} 0 & I \\ -A & 0 \end{smallmatrix}\right|$ is skew-adjoint on Z (see (1.6)), then (2.1b) and (1.6) plainly suggest that we take $g_1 = -k_1^2 G_1^* A^{-1/2} w_t$ and $g_2 = -k_2 \Lambda^2 k_2 G_2^* A^{-1/2} w_t$ (as explicitly noted in (1.10), (1.11)) as natural candidates for feedback stabilization, as this choice then makes the corresponding feedback operator $\mathscr{A}$ defined by (1.23) dissipative on Z. Indeed, for $y = [y_1, y_2] \in \mathscr{D}(\mathscr{A})$,

$$\begin{aligned} \operatorname{Re} (\mathscr{A} y, y)_Z &= -(AG_1 k_1^2 G_1^* A^{-1/2} y_2, y_2)_{[\mathscr{D}(A^{3/4})]'} - (AG_2 k_2 \Lambda^2 k_2 G_2^* A^{-1/2} y_2, y_2)_{[\mathscr{D}(A^{3/4})]'} \\ &= -(A^{-1/2} G_1 k_1^2 G_1^* A^{-1/2} y_2, y_2)_{L^2(\Omega)} - (A^{-1/2} G_2 k_2 \Lambda^2 k_2 G_2^* A^{-1/2} y_2, y_2)_{L^2(\Omega)} \\ &= -\|k_1 G_1^* A^{-1/2} y_2\|^2_{L^2(\Gamma)} - \|\Lambda k_2 G_2^* A^{-1/2} y_2\|^2_{L^2(\Gamma)} \leqq 0 \end{aligned} \tag{2.2}$$

by (1.12), (1.13). Thus the resulting closed-loop problem, where (1.10) and (1.11) are inserted in (1.1c, d), respectively, takes the form

$$w_{tt} = -Aw - A[G_1 k_1^2 G_1^* A^{-1/2} w_t + G_2 k_2 \Lambda^2 k_2 G_2^* A^{-1/2} w_t] \tag{2.3}$$

on, say, $[\mathscr{D}(A)]'$, i.e., takes the explicit p.d.e. form (1.21). Henceforth, unless otherwise noted, we will take $k_1(x) = k_2(x) = 1$ for simplicity of notation when dealing with two feedbacks; and $k_1(x) \equiv 1$, $k_2 \equiv 0$ when dealing only with the feedback g_1, while $g_2 \equiv 0$.

Step 2. (Well-posedness.) Generation by $\mathscr{A}$ of an s.c. semigroup on Z follows via the Lumer–Phillips theorem, since $\mathscr{A}$ is actually maximal dissipative; and indeed direct computations show the explicit expression (1.24), (1.25) of the resolvent operator. To solve $(\lambda I - \mathscr{A})y = z \in Z$ with $\lambda > 0$ fixed, i.e., by (1.23)

$$\lambda y_1 - y_2 = z_1 \in [\mathscr{D}(A^{1/4})]', \tag{2.4}$$

$$A[y_1 + G_1 G_1^* A^{-1/2} y_2 + G_2 \Lambda^2 G_2^* A^{-1/2} y_2] + \lambda y_2 = z_2 \in [\mathscr{D}(A^{3/4})]' \tag{2.5}$$

for $y \in \mathscr{D}(\mathscr{A})$, we apply λA^{-1} to the second equation, subtract off the first, and obtain $V(\lambda) y_2 = \lambda A^{-1} z_2 - z_1 \in [\mathscr{D}(A^{1/4})]'$ via (1.25). But $V(\lambda)$ is boundedly invertible on $[\mathscr{D}(A^{1/4})]'$, since equivalently $A^{-1/4} V(\lambda) A^{1/4} = I + \lambda A^{-1/4} G_1 G_1^* A^{-1/4} + \lambda A^{-1/4} G_2 \Lambda^2 G_2^* A^{-1/4} + \lambda^2 A^{-1}$ is boundedly invertible on $L^2(\Omega)$ for $\lambda > 0$. This way (1.24) is obtained. Finally, from (1.24), compactness of $R(\lambda, \mathscr{A})$ on Z is readily seen to be a consequence of compactness on $L^2(\Omega)$ of the following operators:

$$\begin{aligned} &A^{-1/4} V^{-1}(\lambda) A^{1/4} A^{-1/2}, \qquad A^{-3/4} V^{-1}(\lambda) A^{1/4} A^{-1/2}, \\ &A^{-1/4}(I - V^{-1}(\lambda)) A^{1/4}, \qquad A^{-1/2} A^{-1/4} V^{-1}(\lambda) A^{1/4}. \end{aligned} \tag{2.6}$$

But the first and the fourth are compact here since, as shown above, $(A^{-1/4} V(\lambda) A^{1/4})^{-1} \in \mathscr{L}(L^2(\Omega))$ and of course A^{-1} is compact. The same is then seen to hold true for the third operator in (2.6), by

$$I - V^{-1}(\lambda) = \lambda V^{-1}(\lambda) G_1 G_1^* A^{-1/2} + \lambda V^{-1}(\lambda) G_2 \Lambda^2 G_2^* A^{-1/2} + \lambda^2 V^{-1}(\lambda) A^{-1},$$

which is obtained from (1.25) by applying $V^{-1}(\lambda)$ and using the preceding results. Finally, $\mathscr{A} y = 0$ yields $y = 0$ by (1.23) and $0 \in \rho(\mathscr{A})$.

Step 3. The L_2-boundedness in time (1.28), (1.29) follows at once from (2.2) with $y = [w(t), w_t(t)]$, since then the left of (2.2) is $\frac{1}{2}(d/dt)\|\exp(\mathscr{A}t)[w_0, w_1]\|_Z^2$: integrating in t and using contraction yields the conclusion.

Step 4. By contradiction, let $\lambda = \mathrm{ir}$, r real $\neq 0$, and let $V(\mathrm{ir})x = 0$. Taking the $[\mathscr{D}(A^{1/4})]'$-inner product, we obtain from the definition of $V(\cdot)$ in (1.25) (with $k_2 \equiv 0$ and $k_1 \equiv 1$)

$$G_1^* A^{-1/2} x = G_1^* A A^{-3/2} x = \frac{\partial \Delta}{\partial \nu} A^{-3/2} x|_\Gamma = 0 \tag{2.7}$$

using (1.19). Moreover, we obtain using (2.7) in $V(\lambda)$ that $Ax = r^2 x$. Thus, either $x = 0$ as desired, or else x is an eigenvector of A, say $x = e_n$, with eigenvalue r^2. Thus, $e_n|_\Gamma = (\partial e_n / \partial \nu)|_\Gamma = 0$. Moreover, $A^{-3/2} e_n = (1/r^2)^{3/2} e_n$ and (2.7) implies $(\partial \Delta e_n / \partial \nu)|_\Gamma = 0$. Then the uniqueness property to be proved in Steo 5 implies $x = e_n = 0$. Thus, the (point) spectrum of $\mathscr{A}$ is in $\mathrm{Re}\,\lambda < 0$.

Step 5. We prove the uniqueness property for problems (1.30), (1.31). Define the nonnegative self-adjoint operator

$$Af = \Delta^2 f; \qquad \mathscr{D}(A) = \left\{ f \in H^4(\Omega) : \frac{\partial f}{\partial \nu}\bigg|_\Gamma = \frac{\partial(\Delta f)}{\partial \nu}\bigg|_\Gamma = 0 \right\}. \tag{2.8}$$

Then plainly

$$A^{1/2} f = -\Delta f; \qquad \mathscr{D}(A^{1/2}) = \left\{ f \in H^2(\Omega) : \frac{\partial f}{\partial \nu}\bigg|_\Gamma = 0 \right\}. \tag{2.9}$$

Problems (1.30), (1.31) can be rewritten taking $\lambda = r^2$, with $r > 0$

$$A\phi = r^2 \phi \quad \text{plus } \phi|_\Gamma = 0. \tag{2.10}$$

Thus either $\phi \equiv 0$ and we are done, or else ϕ is an eigenvector of A with eigenvalue r^2. Applying $A^{-1/2}$ to (2.10) yields $A^{1/2}\phi = r\phi$ plus $\phi|_\Gamma = 0$; i.e., by (2.9)

$$\begin{cases} -\Delta\phi = r\phi, \\ \phi|_\Gamma = \dfrac{\partial\phi}{\partial\nu}\Big|_\Gamma = 0 \end{cases} \tag{2.11}$$

and (2.11) plainly implies $\phi \equiv 0$ in Ω (with sufficiently smooth Γ), as desired. □

3. Proof of Theorems 1.2 and 1.3.

3.1. Preliminaries and a change of variable. For the feedback problem (1.21) in the case of Theorem 1.2 (respectively, (1.21a-c)) and $(\partial w/\partial\nu)_\Sigma = 0$ in the case of Theorem 1.3) we define (the "energy functional") $E(t) \equiv E(w, t)$ by the squared norm of the semigroup in Theorem 1.1(i) on $Z = [\mathscr{D}(A^{1/4})]' \times [\mathscr{D}(A^{3/4})]'$:

$$\begin{aligned} E(t) &= \left\| e^{\mathscr{A}t} \begin{vmatrix} w_0 \\ w_1 \end{vmatrix} \right\|_Z^2 = \left\| \begin{vmatrix} w(t) \\ w_t(t) \end{vmatrix} \right\|_Z^2 \\ &= \|A^{-1/4} w(t)\|_{L^2(\Omega)}^2 + \|A^{-3/4} w_t(t)\|_{L^2(\Omega)}^2 \leqq E(0) \end{aligned} \tag{3.1}$$

by the contraction property of Theorem 1.1(i). Our main goal will be, as usual [8], [39], to show that

$$\int_0^\infty E(t)\, dt \leqq \text{const } E(0) \quad \forall [w_0, w_1] \in \mathscr{D}(\mathscr{A}) \tag{3.2}$$

where the constant is independent of the initial data $[w_0, w_1]$ whereby (3.2) can be extended by continuity to all $\{w_0, w_1\} \in Z$. After this, Datko's theorem [3] will yield the desired uniform bound (1.32). Thus, unless otherwise stated, we assume henceforth that $[w_0, w_1] \in \mathscr{D}(\mathscr{A})$.

Remark 3.1. Instead of (3.2), we may of course use the other well-known and equivalent criterion for uniform decay of a semigroup: that there is some $0 < T < \infty$ such that

$$(*) \qquad E(T) \leqq rE(0), \quad r < 1 \quad \text{or} \quad \|e^{\mathscr{A}T}\|_{\mathscr{L}(Z)} < 1.$$

Generally speaking, each of the two criteria, while requiring very closely related approaches and computations, offers some advantages and some disadvantages over the other. In the case of constant coefficients (canonically, with Δ^2) and with a radial vector field $h(x) = x - x_0$, as in the proof of our Theorem 1.2, the use of (*) offers some streamlining in the computations over use of (3.2): we may take $\beta = 0$ below and integrate in time over a finite interval, without needing to show that terms arising from integration by parts in time go to zero as $t \to \infty$. On the other hand, in the non-constant coefficient case (in space variable) and in working with a general vector field $h(x)$ as in the proof of Theorem 1.3 below, the need arises to absorb "lower-order terms." In using (*), we resort to some arguments of compactness type already used in the corresponding exact controllability results. These, however, ultimately rely on a Holmgren-type uniqueness property, and hence require smooth (analytic) coefficients. In contrast, in using criterion (3.2), absorption of lower-order terms requires a new result, such as our Theorem 1.3B in the case of the present paper. Such a result needs only minimal smoothness of the coefficients (in space); besides, it is of interest in itself. See, for instance, [8, § 5] for a more general wave equation with Neumann feedback, where the counterpart of (3.2) is used. As noted in Remark 1.1, achievement of an *explicit* estimate of the constant δ in (1.32) is *irrespective* of whether we use criterion

(3.2) or the equivalent condition (*) above. Indeed, it depends on whether $h(x)$ is radial (linear) or not.

Returning to (1.23), we see a more explicit description of $y=[y_1, y_2]\in\mathscr{D}(\mathscr{A})$:

(i) (3.3) $y_2\in[\mathscr{D}(A^{1/4})]'$; i.e., $A^{-1/4}y_2\in L^2(\Omega)$, i.e., $y_2\in H^{-1}(\Omega)$,

(ii) (3.4) $A[y_1+G_1G_1^*A^{-1/2}y_2+G_2\Lambda^2G_2^*A^{-1/2}y_2]\in[\mathscr{D}(A^{3/4})]'$,

$$y_1+G_1G_1^*A^{-1/2}y_2+G_2\Lambda^2G_2^*A^{-1/2}y_2\in\mathscr{D}(A^{1/4})=H_0^1(\Omega),$$

which implies a fortiori

$$y_1\in H^1(\Omega). \tag{3.5}$$

Conclusion (3.5) is a consequence of $A^{-1/2}y_2\in H_0^1(\Omega)$ by (3.3) and (1.3a), and of the following maps:

$$G_1G_1^*\text{: continuous } \mathscr{D}(A^{1/4})=H_0^1(\Omega)\to H^2(\Omega), \tag{3.6}$$

$$G_2\Lambda^2G_2^*\text{: continuous } \mathscr{D}(A^{1/4})=H_0^1(\Omega)\to H^2(\Omega). \tag{3.7}$$

Indeed, returning to (1.17) and (1.18), we see that these imply, respectively,

$$G_1^*\text{: continuous } H_0^1(\Omega)\to H^{3/2}(\Gamma) \tag{3.8}$$

by duality on (1.17) with $s=-\frac{3}{2}$, and

$$G_2^*\text{: continuous } H_0^1(\Omega)\to H^{5/2}(\Gamma) \tag{3.9}$$

by duality on (1.18) with $s=-\frac{5}{2}$. Then (3.8), followed by (1.17) with $s=\frac{3}{2}$, yields (3.6). Also, (3.9) and the definition (1.13) of Λ gives first

$$\Lambda^2G_2^*\text{: continuous } H_0^1(\Omega)\to H^{1/2}(\Gamma); \tag{3.10}$$

this followed by (1.18) with $s=\frac{1}{2}$ yields (3.7). Finally, (3.6), (3.7) used in (3.4) yield (3.5) via (3.3) as desired. Next, by Theorem 1.1(i),

$$\text{If } \{w_0, w_1\}\in\mathscr{D}(\mathscr{A}) \quad \text{then } \{w(t), w_t(t)\}\in C([0, T]; \mathscr{D}(\mathscr{A})) \tag{3.11a}$$

and thus by (3.3) and (3.5)

$$A^{-1/4}w_t\in C([0, T]; L^2(\Omega)), \quad w\in C([0, T]; H^1(\Omega)), \quad \{w_0, w_1\}\in\mathscr{D}(\mathscr{A}). \tag{3.11b}$$

Motivated by the multiplier techniques of [18] and [19], and by [20], we introduce a new variable p by setting $A^{3/4}p=A^{-3/4}w_t$; i.e.,

$$p=A^{-3/2}w_t\in\begin{cases}C([0, T]; \mathscr{D}(A^{3/4})) & \text{if } \{w_0, w_1\}\in Z,\\ C([0, T]; \mathscr{D}(A^{5/4})) & \text{if } \{w_0, w_1\}\in\mathscr{D}(\mathscr{A});\end{cases} \tag{3.12a, b}$$

see (1.6), Theorem 1.1(i) and (3.11b) respectively. Thus, by (2.3),

$$\begin{aligned}p_t=A^{-3/2}w_{tt}&=-A^{-1/2}[w+G_1G_1^*A^{-1/2}w_t+G_2\Lambda^2G_2^*A^{-1/2}w_t]\\ &\in\begin{cases}L^2(0, T; \mathscr{D}(A^{1/4})) & \text{if } \{w_0, w_1\}\in Z,\\ C([0, T]; \mathscr{D}(A^{3/4}) & \text{if } \{w_0, w_1\}\in\mathscr{D}(\mathscr{A})\end{cases}\end{aligned} \tag{3.13a, b}$$

where the regularity follows from Theorem 1.1(i), (1.28), (1.29), or (3.11a), (3.4), respectively; hence

$$p_{tt}=-A^{-1/2}[w_t+G_1G_1^*A^{-1/2}w_{tt}+G_2\Lambda^2G_2^*A^{-1/2}w_{tt}], \tag{3.14a}$$

$$p_{tt}=-Ap+F_1+F_2, \tag{3.14b}$$

$$F_1=-A^{-1/2}G_1G_1^*A^{-1/2}w_{tt}, \quad F_2=-A^{-1/2}G_2\Lambda^2G_2^*A^{-1/2}w_{tt}. \tag{3.15}$$

In terms of the scalar function $p(t,x)$, $x\in\Omega$, corresponding to the vector-valued function $p(t)=p(t,\cdot)$, the abstract equation (3.14b) can be rewritten explicitly as the following Euler-Bernoulli homogeneous problem with initial condition (I.C.) well defined at $t_0>0$ by (3.13b) and (2.3) for $w_{tt}(t_0)$

$$p_{tt}+\Delta^2 p=F_1+F_2 \qquad \text{in } (t_0,\infty)\times\Omega=Q, \tag{3.16a}$$

$$p(t_0,x)=p_0=A^{-3/2}w_t(t_0);\ p_t(t_0,x)=p_1=A^{-3/2}w_{tt}(t_0) \qquad \text{in } \Omega, \tag{3.16b}$$

$$p|_\Sigma=0 \qquad \text{in } (t_0,\infty)\times\Gamma=\Sigma, \tag{3.16c}$$

$$\left.\frac{\partial p}{\partial \nu}\right|_\Sigma=0 \qquad \text{in } \Sigma \tag{3.16d}$$

where the homogeneous boundary conditions are a consequence of $p\in\mathscr{D}(A^{3/4})$ from (3.12), and of (1.3). In our argument in the sequel, we will have to consider pointwise values $p_t(t)$. Note from (3.13b) that these make sense for initial data $\{w_0,w_1\}\in\mathscr{D}(\mathscr{A})$ *as assumed*, while from (3.13a) the pointwise meaning of p_t in $H_0^1(\Omega)=\mathscr{D}(A^{1/4})$ is lost for general initial data in Z. In the analysis below of the p-system (3.16), we will use the crucial results, from (1.3)-(1.5) via (3.12) and (3.13), that

$$\|A^{-3/4}w_t\|_{L^2(\Omega)}=\|A^{3/4}p\|_{L^2(\Omega)} \quad \text{equivalent to} \quad \left\{\int_\Omega |\nabla(\Delta p)|^2\,d\Omega\right\}^{1/2}, \tag{3.17}$$

$$\|A^{3/4}p(t)\|^2_{L^2(\Omega)}\leqq E(0) \quad \forall t\geqq 0, \tag{3.18}$$

$$A^{1/4}p_t=-A^{-1/4}w+O(\|G_1^*A^{-1/2}w_t\|_{L^2(\Gamma)}+\|\Lambda G_2^*A^{-1/2}w_t\|_{L^2(\Gamma)}), \tag{3.19}$$

the $L^2(\Gamma)$-terms being the feedback in (1.28), (1.29), since G_1 and $G_2\Lambda$ are bounded on $L^2(\Gamma)$ and

$$\|A^{1/4}p_t\|_{L^2(\Omega)} \quad \text{equivalent to} \quad \left\{\int_\Omega |\nabla p_t|^2\,d\Omega\right\}^{1/2}. \tag{3.20}$$

In (3.19) the symbol O means, as usual, bounded above by a constant. The norms on the right of (3.17) and (3.20) will arise in the multiplier approach used below (following [18], [19]); this justifies the need to introduce the variable p. Before applying the multiplier approach, we need to note what follows. Since $\lambda=0\in\rho(\mathscr{A})$, the resolvent set of $\mathscr{A}$, by Theorem 1.1(i), we have for $\{w_0,w_1\}\in\mathscr{D}(\mathscr{A})$

$$\left\|\begin{vmatrix} w(t)\\ w_t(t)\end{vmatrix}\right\|_{\mathscr{D}(\mathscr{A})}=\left\|e^{\mathscr{A}t}\begin{vmatrix} w_0\\ w_1\end{vmatrix}\right\|_{\mathscr{D}(\mathscr{A})}=\left\|e^{\mathscr{A}t}\mathscr{A}\begin{vmatrix} w_0\\ w_1\end{vmatrix}\right\|_Z\leqq\left\|\mathscr{A}\begin{vmatrix} w_0\\ w_1\end{vmatrix}\right\|_Z, \qquad t\geqq 0 \tag{3.21}$$

by the contraction property. A fortiori, (3.21) implies via (3.11b) for $\{w_0,w_1\}\in\mathscr{D}(\mathscr{A})$

$$\|w(t)\|^2_{H^1(\Omega)}+\|A^{-1/4}w_t(t)\|^2_{L^2(\Omega)}\leqq\left\|\mathscr{A}\begin{vmatrix} w_0\\ w_1\end{vmatrix}\right\|^2_Z \quad \forall t\geqq t_0. \tag{3.22}$$

By (3.12), $A^{5/4}p(t)=A^{-1/4}w_t(t)$, and by (3.13)

$$\begin{aligned} A^{1/4}p_t(t)=&-A^{-1/4}w(t)-A^{-1/4}G_1G_1^*A^{-1/4}A^{-1/4}w_t(t)\\ &-A^{-1/4}G_2\Lambda^2G_2^*A^{-1/4}A^{-1/4}w_t(t), \end{aligned} \tag{3.23}$$

so that from (3.22) we obtain

$$\|A^{5/4}p(t)\|^2_{L^2(\Omega)}+\|A^{1/4}p_t(t)\|^2_{L^2(\Omega)}\leqq\text{const}\left\{E(t_0)+\left\|\mathscr{A}\begin{vmatrix} w_0\\ w_1\end{vmatrix}\right\|^2_Z\right\}, \quad t\geqq t_0 \tag{3.24}$$

since $A^{-1/4}w(t) \in C([0, T]; L^2(\Omega))$ for all $\{w_0, w_1\} \in Z$ and $G_1 G_1^* A^{-1/4}$ are bounded on $L^2(\Omega)$ (see (3.6) and (3.7)).

3.2. An identity for the *p*-system.

PROPOSITION 3.1. *The following identity holds true for problem* (3.16), *where* $\{w_0, w_1\} \in \mathscr{D}(\mathscr{A})$; *hence* $\{p_0, p_1\} \in \mathscr{D}(A^{5/4}) \times \mathscr{D}(A^{3/4})$ *by* (3.12b), (3.13b), *where* $\beta > 0$ *is an arbitrary constant, and* $Q = (t_0, \infty) \times \Omega$, $\Sigma = (t_0, \infty) \times \Gamma$, *and* $h(x)$ *is a smooth vector field*:

$$
\begin{aligned}
&\int_\Sigma e^{-2\beta t} \frac{\partial(\Delta p)}{\partial \nu} h \cdot \nabla(\Delta p)\, d\Sigma - \frac{1}{2} \int_\Sigma e^{-2\beta t} |\nabla(\Delta p)|^2 h \cdot \nu\, d\Sigma \\
&\quad = \int_Q e^{-2\beta t} H \nabla(\Delta p) \cdot \nabla(\Delta p)\, dQ + \int_Q e^{-2\beta t} H \nabla p_t \cdot \nabla p_t\, dQ \\
&\qquad + \int_Q e^{-2\beta t} p_t \nabla p_t \cdot \nabla(\operatorname{div} h)\, dQ \\
&\qquad + \frac{1}{2} \int_Q e^{-2\beta t} \{|\nabla p_t|^2 - |\nabla(\Delta p)|^2\} \operatorname{div} h\, dQ \\
&\qquad + \int_Q e^{-2\beta t} (F_1 + F_2) h \cdot \nabla(\Delta p)\, dQ \\
&\qquad - 2\beta \int_Q e^{-2\beta t} p_t h \cdot \nabla(\Delta p)\, dQ + e^{-2\beta t_0} (p_1, h \cdot \nabla(\Delta p_0))_\Omega .
\end{aligned} \tag{3.25}
$$

Proof. Most of the proof is carried out in Appendix A and Leads to identities (A.8) and (A.10). It only remains to show that for the assumed initial data

$$
\lim_{T \to x} e^{-2\beta T} (p_t(T), h \cdot \nabla(\Delta p(T)))_\Omega = 0. \tag{3.26}
$$

Indeed, we have by using (3.17), (3.18) with $T \geqq t_0$, and (3.24)

$$
\begin{aligned}
|(p_t(T), h \cdot \nabla(\Delta p(T)))_\Omega| &\leqq c_h \{\| p_t(T) \|^2_{L^2(\Omega)} + \| A^{3/4} p(T) \|^2_{L^2(\Omega)}\} \\
&\leqq \text{const} \left\{ E(t_0) + \left\| \mathscr{A} \begin{vmatrix} w_0 \\ w_1 \end{vmatrix} \right\|^2_Z \right\}.
\end{aligned} \tag{3.27}
$$

□

It will be shown below that all the terms in (3.25) are well defined. But before doing so, we will rewrite the fourth integral over Q on the right of (3.25) in a more convenient form.

PROPOSITION 3.2. *The following identity holds true for problem* (3.16), *where* $\{w_0, w_1\} \in \mathscr{D}(\mathscr{A})$ *and* $\beta > 0$

$$
\begin{aligned}
&\int_Q e^{-2\beta t} \{|\nabla p_t|^2 - |\nabla(\Delta p)|^2\} \operatorname{div} h\, dQ \\
&\quad = -\int_\Sigma e^{-2\beta t} \frac{\partial(\Delta p)}{\partial \nu} \Delta p \operatorname{div} h\, d\Sigma + \int_Q e^{-2\beta t} (F_1 + F_2) \Delta p \operatorname{div} h\, dQ \\
&\qquad + \int_Q e^{-2\beta t} \Delta p \nabla(\Delta p) \cdot \nabla(\operatorname{div} h)\, dQ - 2\beta \int_Q e^{-2\beta t} p_t \Delta p \operatorname{div} h\, dQ \\
&\qquad - \int_Q e^{-2\beta t} p_t \nabla p_t \cdot (\operatorname{div} h)\, dQ + e^{-2\beta t_0} (p_1, \Delta p_0 \operatorname{div} h)_\Omega .
\end{aligned} \tag{3.28}
$$

Proof. Most of the proof is carried out in Appendix B and leads to identity (B.4). We must show only that

$$\lim_{T\to\infty} e^{-2\beta T}\int_\Omega \nabla p(T)\cdot\nabla(p_t(T)\operatorname{div} h)\,d\Omega = 0 \tag{3.29}$$

for the assumed initial data. Indeed, as in (3.27) we have by (3.20) and (3.24) for $T \geqq t_0$,

$$\begin{aligned}\left|\int_\Omega \nabla p(T)\cdot\nabla(p_t(T)\operatorname{div} h)\,d\Omega\right| &\leqq C_h\{\|A^{1/4}p(T)\|^2_{L^2(\Omega)}+\|A^{1/4}p_t(T)\|^2_{L^2(\Omega)}\}\\ &\leqq \text{const}\left\{E(t_0)+\left\|\mathscr{A}\begin{vmatrix}w_0\\ w_1\end{vmatrix}\right\|^2_Z\right\}.\end{aligned} \tag{3.30}$$

□

By combining Proposition 3.1 with Proposition 3.2 we readily obtain the following final identity.

PROPOSITION 3.3. *The following identity holds true for problem* (3.16), *where* $\{w_0, w_1\}\in\mathscr{D}(\mathscr{A})$, $\beta>0$ *is an arbitrary constant, and* $Q=(t_0,\infty)\times\Omega$; $\Sigma=(t_0,\infty)\times\Gamma$; *and* $h(x)$ *is a smooth vector field*

$$\begin{aligned}&\int_\Sigma e^{-2\beta t}\frac{\partial(\Delta p)}{\partial\nu}h\cdot\nabla(\Delta p)\,d\Sigma-\frac{1}{2}\int_\Sigma e^{-2\beta t}|\nabla(\Delta p)|^2 h\cdot\nu\,d\Sigma\\ &\quad+\frac{1}{2}\int_\Sigma e^{-2\beta t}\frac{\partial(\Delta p)}{\partial\nu}\Delta p\operatorname{div} h\,d\Sigma\\ &=\int_Q e^{-2\beta t}H\nabla(\Delta p)\cdot\nabla(\Delta p)\,dQ+\int_Q e^{-2\beta t}H\nabla p_t\cdot\nabla p_t\,dQ\\ &\quad+\int_Q e^{-2\beta t}(F_1+F_2)h\cdot\nabla(\Delta p)\,dQ+\frac{1}{2}\int_Q e^{-2\beta t}(F_1+F_2)\Delta p\operatorname{div} h\,dQ\\ &\quad+\frac{1}{2}\int_Q e^{-2\beta t}p_t\nabla p_t\cdot\nabla(\operatorname{div} h)\,dQ+\frac{1}{2}\int_Q e^{-2\beta t}\Delta p\nabla(\Delta p)\cdot\nabla(\operatorname{div} h)\,dQ\\ &\quad-2\beta\int_Q e^{-2\beta t}p_t h\cdot\nabla(\Delta p)\,dQ-\beta\int_Q e^{-2\beta t}p_t\Delta p\operatorname{div} h\,dQ\\ &\quad+e^{-2\beta t_0}(p_1, h\cdot\nabla(\Delta p_0))_\Omega+\frac{1}{2}e^{-2\beta t_0}(p_1,\Delta p_0\operatorname{div} h)_\Omega.\end{aligned} \tag{3.31}$$

The analysis below will show a fortiori that the terms in identity (3.31) are well defined by establishing appropriate estimates thereof.

3.3. Analysis of the terms involving F_i and the initial data. Crucial terms are those involving F_1, F_2 multiplied by $h\cdot\nabla(\Delta p)$.

PROPOSITION 3.4. *For* $\{w_0, w_1\}\in\mathscr{D}(\mathscr{A})$, *we have the following identity with* $\beta>0$:

(a)

$$\begin{aligned}&\int_Q e^{-2\beta t}(F_1+F_2)h\cdot\nabla(\Delta p)\,dQ+e^{-2\beta t_0}(p_1, h\cdot\nabla(\Delta p_0))_\Omega\\ &\quad-e^{-2\beta t_0}(A^{-1/2}w(t_0), h\cdot\nabla(\Delta p_0))_\Omega\\ &=\int_{t_0}^\infty e^{-2\beta t}(A^{-1/2}G_1G_1^*A^{-1/2}w_t+A^{-1/2}G_2\Lambda^2G_2^*A^{-1/2}w_t, h\cdot\nabla(\Delta p_t))_\Omega\,dt\\ &\quad+2\beta\int_{t_0}^\infty e^{-2\beta t}(A^{-1/2}G_1G_1^*A^{-1/2}w_t\\ &\quad+A^{-1/2}G_2\Lambda^2G_2^*A^{-1/2}w_t, h\cdot\nabla(\Delta p))_\Omega\,dt.\end{aligned} \tag{3.32}$$

(b) *For* $\{w_0, w_1\} \in \mathscr{D}(\mathscr{A})$ *and for any* $\varepsilon > 0$, *the following estimate holds for the first term of* (3.32):

$$\begin{aligned}
&\int_{t_0}^{\infty} e^{-2\beta t}(A^{-1/2}G_1G_1^*A^{-1/2}w_t + A^{-1/2}G_2\Lambda^2G_2^*A^{-1/2}w_t,\ h\cdot\nabla(\Delta p_t))_\Omega\, dt \\
&= \frac{1}{\varepsilon}\, O\left\{\int_{t_0}^{\infty} e^{-2\beta t}[\|G_1^*A^{-1/2}w_t\|^2_{L^2(\Gamma)} + \|\Lambda G_2^*A^{-1/2}w_t\|^2_{L^2(\Gamma)}]\, dt\right\} \qquad (3.33)\\
&\quad + \varepsilon\left\{\int_{t_0}^{\infty} e^{-2\beta t}E(t)\, dt\right\}
\end{aligned}$$

where O denotes upper bound with a multiplicative constant *independent* of β, and t_0 and the right-hand side of (3.33) are finite by Theorem 1.1(ii), (1.28), (1.29), and the contraction property (3.1) of $E(t)$.

(c)

$$\begin{aligned}
&\int_Q e^{-2\beta t}(F_1+F_2)h\cdot\nabla(\Delta p)\, dQ + e^{-2\beta t_0}(p_1,\ h\cdot\nabla(\Delta p_0))_\Omega \\
&= O\{E(t_0)\} + \frac{1}{\varepsilon}\, O\left\{\int_{t_0}^{\infty} e^{-2\beta t}[\|G_1^*A^{-1/2}w_t\|^2_{L^2(\Gamma)} + \|\Lambda G_2^*A^{-1/2}w_t\|^2_{L^2(\Gamma)}]\, dt\right\} \qquad (3.34)\\
&\quad + \varepsilon\left\{\int_{t_0}^{\infty} e^{-2\beta t}E(t)\, dt\right\}.
\end{aligned}$$

Proof. (a) Recalling F_1 from (3.15) and integrating by parts in t we obtain

$$\begin{aligned}
-\int_Q e^{-2\beta t}F_1h\cdot\nabla(\Delta p)\, dQ &= \int_{t_0}^{\infty}(A^{-1/2}G_1G_1^*A^{-1/2}w_{tt},\ e^{-2\beta t}h\cdot\nabla(\Delta p))_\Omega\, dt \\
&= -e^{-2\beta t_0}(A^{-1/2}G_1G_1^*A^{-1/2}w_t(t_0),\ h\cdot\nabla(\Delta p_0))_\Omega \\
&\quad + 2\beta\int_{t_0}^{\infty} e^{-2\beta t}(A^{-1/2}G_1G_1^*A^{-1/2}w_t,\ h\cdot\nabla(\Delta p))_\Omega\, dt \qquad (3.35)\\
&\quad - \int_{t_0}^{\infty} e^{-2\beta t}(A^{-1/2}G_1G_1^*A^{-1/2}w_t,\ h\cdot\nabla(\Delta p_t))_\Omega\, dt
\end{aligned}$$

since for $\{w_0, w_1\} \in \mathscr{D}(\mathscr{A})$ as assumed we have

$$\lim_{T\to\infty} e^{-2\beta T}(A^{-1/2}G_1G_1^*A^{-1/2}w_t(T),\ h\cdot\nabla(\Delta p(T))_\Omega = 0. \qquad (3.36)$$

Since $A^{-1/2}G_1G_1^*$ is a bounded operator on $L^2(\Omega)$, we arrive at (3.36) readily by applying the Schwarz inequality to the inner product in Ω, using $\|A^{-1/2}w_t(T)\| = \|Ap(T)\|$ by (3.12) and $\||\nabla(\Delta p(T))|\| \sim \|A^{3/4}p(T)\|$ by (3.17), and recalling (3.24). Similarly, recalling F_2 from (3.15) we obtain, integrating by parts in t,

$$\begin{aligned}
-\int_Q e^{-2\beta t}F_2h\cdot\nabla(\Delta p)\, dQ &= \int_{t_0}^{\infty}(A^{-1/2}G_2\Lambda^2G_2^*A^{-1/2}w_{tt},\ e^{-2\beta t}h\cdot\nabla(\Delta p))_\Omega\, dt \\
&= -e^{-2\beta t_0}(A^{-1/2}G_2\Lambda^2G_2^*A^{-1/2}w_t(t_0),\ h\cdot\nabla(\Delta p_0))_\Omega \qquad (3.37)\\
&\quad + 2\beta\int_{t_0}^{\infty} e^{-2\beta t}(A^{-1/2}G_2\Lambda^2G_2^*A^{-1/2}w_t,\ h\cdot\nabla(\Delta p))_\Omega\, dt \\
&\quad - \int_{t_0}^{\infty} e^{-2\beta t}(A^{-1/2}G_2\Lambda^2G_2^*A^{-1/2}w_t,\ h\cdot\nabla(\Delta p_t))_\Omega\, dt \qquad (3.38)
\end{aligned}$$

since for $\{w_0, w_1\} \in \mathscr{D}(\mathscr{A})$ as assumed we have

$$\lim_{T\to\infty} e^{-2\beta T}(A^{-1/2}G_2\Lambda^2 G_2^* A^{-1/2} w_t(T), h\cdot\nabla(\Delta p(T)))_\Omega = 0. \tag{3.39}$$

An argument similar to the one below (3.36) establishes (3.39), this time with $A^{-1/2}G_2\Lambda^2 G_2^* A^{-1/4}$ a bounded operator on $L^2(\Omega)$ by (3.7), and with $\|A^{-1/4}w_t(T)\| = \|A^{5/4}p(T)\|$ by (3.12), so that we can once again invoke (3.24). Finally, to obtain (3.32), we sum up (3.35) with (3.37) and use (3.13) in combining the two $(\ ,\ )_\Omega$-terms for $t = t_0$. Part (a) is proved.

(b) We treat each term on the left of (3.33) separately. We will use the following lemma.

LEMMA 3.5. *For $h = [h_1(x), \cdots, h_n(x)] \in [C^2(\bar\Omega)]^n$, we have*

$$\|A^{-1/2}(h\cdot\nabla(\Delta p_t))\|_{L^2(\Omega)} \leqq C_h \|A^{1/4}p_t\|_{L^2(\Omega)}. \tag{3.40}$$

Proof of Lemma 3.5. We have that $\mathscr{D}(A^{1/2}) = H_0^2(\Omega)$ (see [19, App. C]) so that $[\mathscr{D}(A^{1/2})]' = H^{-2}(\Omega)$ (with equivalent norms). Then, since $p_t|_\Gamma = 0$ so that (3.20) applies,

$$\|A^{-1/2}(h\cdot\nabla(\Delta p_t))\|_{L^2(\Omega)} = \|h\cdot\nabla(\Delta p_t)\|_{[\mathscr{D}(A^{1/2})]'} \leqq c_1\|h\cdot\nabla(\Delta p_t)\|_{H^{-2}(\Omega)} \tag{3.41}$$

$$\leqq C_h \||\nabla(\Delta p_t)|\|_{H^{-2}(\Omega)} \leqq C_h\|p_t\|_{H^1(\Omega)} = C_h\|A^{1/4}p_t\|_{L^2(\Omega)}. \tag{3.42}$$

In going from (3.41) to (3.42) we have used (see also [17, p. 31]) the fact that if z is a scalar function in $H^{-2}(\Omega)$, and $h_i \in C^2(\bar\Omega)$, then we have $h_i z \in H^{-2}(\Omega)$ continuously, as it follows directly by duality. If $y \in H_0^2(\Omega)$ and $h_i \in C^2(\bar\Omega)$, then $h_i y \in H_0^2(\Omega)$, since $(h_i y)|_\Gamma = 0$ and $\partial(h_i y)/\partial\nu|_\Gamma = 0$ by using the same properties for y. □

Then (3.33) follows at once from (3.40) (after moving the self-adjoint $A^{-1/2}$ across the $(\ ,\ )_\Omega$-inner product) and from (3.19) and (3.1).

(c) We use (1.28), (1.29) and (3.17), (3.18) to obtain readily

$$\begin{aligned} &2\beta\int_{t_0}^\infty e^{-2\beta t}(A^{-1/2}G_1G_1^*A^{-1/2}w_t + A^{-1/2}G_2\Lambda^2G_2^*A^{-1/2}w_t, h\cdot\nabla(\Delta p))_\Omega\, dt \\ &\leqq 2\beta[E(t_0) + \int_{t_0}^\infty e^{-2\beta t}E(t)\, dt] = (2\beta+1)O(E(t_0)) \end{aligned} \tag{3.43}$$

by the contraction property of $E(t)$. Then (3.33) and (3.43) used in (3.32) readily yield (3.34). □

The integral terms in (3.31) involving the lower-order terms $F_i\Delta p$ div h are a fortiori handled by the analysis in Proposition 3.4 that deals with the integral terms involving the higher-order terms $F_i h\cdot\nabla(\Delta p)$. Thus, we have the following easy counterpart of (3.34) of Proposition 3.4c.

PROPOSITION 3.6. *For $\{w_0, w_1\} \in \mathscr{D}(\mathscr{A})$ we have the following estimate with $\beta > 0$:*

$$\begin{aligned} &\int_Q e^{-2\beta t}(F_1+F_2)\Delta p \text{ div } h\, dQ + e^{-2\beta t_0}(p_1, \Delta p_0 \text{ div } h)_\Omega \\ &= \frac{1}{\varepsilon} O\left\{\int_{t_0}^\infty e^{-2\beta t}[\|G_1^*A^{-1/2}w_t\|^2_{L^2(\Gamma)} + \|\Lambda G_2^*A^{-1/2}w_t\|^2_{L^2(\Gamma)}]\, dt\right\} \\ &\quad + \varepsilon\left(\int_{t_0}^\infty e^{-2\beta t}E(t)\, dt\right) + O\{E(t_0)\}. \end{aligned} \tag{3.44}$$

3.4. Completion of the proof of Theorem 1.2: two feedback controls g_1 and g_2, in the absence of geometrical conditions on Ω.

Right-hand side of (3.31). We will specialize the vector field $h(x)$ to a radial field $h(x)=x-x_0\in R^n$. Thus, $H(x)\equiv$ Identity, div $h\equiv \dim\Omega=n$, and the two integral terms in identity (3.31) involving ∇ (div h) vanish. As to the integral terms in (3.31) premultiplied by β, we readily see by (3.17), (3.19), (3.1), and (1.28), (1.29) that

$$2\beta\int_Q e^{-2\beta t}p_t h\cdot\nabla(\Delta p)\,dQ-\beta\int_Q e^{-2\beta t}p_t\Delta p\ \text{div}\ h\,dQ=(\beta+1)O(E(t_0)) \tag{3.45}$$

Hence, using (3.34), (3.44), and (3.45) on the right-hand side (R.H.S.) of identity (3.31), we obtain

$$\begin{aligned}\text{R.H.S. of (3.31)}=&\int_Q e^{-2\beta t}[|\nabla(\Delta p)|^2+|\nabla p_t|^2\,dQ+\varepsilon\int_{t_0}^{\infty}e^{-2\beta t}E(t)\,dt\\ &+\frac{1}{\varepsilon}O\left\{\int_{t_0}^{\infty}e^{-2\beta t}[\|G_1^*A^{-1/2}w_t\|^2_{L^2(\Gamma)}+\|\Lambda G_2^*A^{-1/2}w_t\|^2_{L^2(\Gamma)}]\,dt\right\}\\ &+O(E(t_0)).\end{aligned} \tag{3.46}$$

Thus, recalling again (3.17), (3.19), (3.20), and (3.1), we finally obtain from (3.46) with $c_{i\varepsilon}$ positive constants:

$$\text{R.H.S. of (3.31)}\geqq c_{1\varepsilon}\int_{t_0}^{\infty}e^{-2\beta t}E(t)\,dt-c_{2\varepsilon}E(t_0), \tag{3.47}$$

which is the desired estimate.

Left-hand side of (3.31). We collect the three pieces of information that are needed. First, by (1.10), (1.19), (3.12), and (1.28), we obtain

$$-w|_\Sigma=G_1^*A^{-1/2}w_t=G_1^*AA^{-3/2}w_t=\frac{\partial(\Delta p)}{\partial\nu}\in L^2(0,\infty;L^2(\Gamma)) \tag{3.48}$$

continuously in $E(0)$. Next, in a similar way, by (1.20) and (1.29) we find

$$\Lambda G_2^*A^{-1/2}w_t=\Lambda G_2^*AA^{-3/2}w_t=-\Lambda(\Delta p)\in L^2(0,\infty;L^2(\Gamma)) \tag{3.49}$$

continuously in $E(0)$, i.e., via the definition of Λ in (1.13):

$$\int_{t_0}^{\infty}\|\Lambda(\Delta p)\|^2_{L^2(\Gamma)}\,dt=\int_{t_0}^{\infty}\int_\Gamma(\Delta p)^2+|\nabla_\sigma(\Delta p)|^2\,d\Gamma\,dt\leqq E(t_0). \tag{3.50}$$

We now return to the left-hand side (L.H.S.) of identity (3.31) and see that by (3.48)–(3.50), along with $|\nabla(\Delta p)|^2=|\partial(\Delta p)/\partial\nu|^2+|\nabla_\sigma(\Delta p)|^2$, (1.28), (1.29), and (3.17) we obtain for any $\varepsilon>0$

$$\begin{aligned}\text{L.H.S. of (3.31)}&\leqq C_{h,\varepsilon}\int_\Sigma e^{-2\beta t}\left[\left(\frac{\partial(\Delta p)}{\partial\nu}\right)^2+|\nabla_\sigma(\Delta p)|^2+(\Delta p)^2\right]d\Sigma\\ &\leqq C_{h,\varepsilon}E(t_0),\end{aligned} \tag{3.51}$$

which is the desired estimate. Finally, we combine (3.47) and (3.51) and we conclude that if $\{w_0,w_1\}\in\mathcal{D}(\mathcal{A})$ and $\beta>0$ then

$$\int_{t_0}^{\infty}e^{-2\beta t}E(t)\,dt\leqq\text{const}\,E(t_0) \tag{3.52}$$

with const independent of β_0, $0<\beta\leqq\beta_0$, and of $\{w_0,w_1\}$. Letting $\beta\downarrow 0$ in (3.52) we easily obtain (3.2). The proof of Theorem 1.2 is complete. □

3.5. Completion of the proof of Theorem 1.3A: feedback control for g_1 while $g_2 \equiv 0$, under geometrical conditions for Ω. Now let $h(x)$ be a general smooth vector field as assumed in (1.33), (1.34) of Theorem 1.3. Then we estimate the two integral terms involving ∇ (div h) by

$$\int_Q e^{-2\beta t} p_t \nabla p_t \cdot \nabla(\operatorname{div} h)\, dQ + \int_Q e^{-2\beta t} \Delta p \nabla(\Delta p) \cdot \nabla(\operatorname{div} h)\, dQ$$
$$= \varepsilon O\left(\int_Q e^{-2\beta t} [|\nabla p_t|^2 + |\nabla(\Delta p)|^2]\, dQ \right)$$
$$+ \frac{1}{\varepsilon} O\left(\int_Q e^{-2\beta t} [p_t^2 + (\Delta p)^2]\, dQ \right).$$

Thus, the new contribution in the R.H.S. of (3.31) is given by the above terms when $h(x)$ is not linear, where $\|\Delta p\| = \|A^{1/2} p\| = \|A^{-1} w_t\|$ in the $L^2(\Omega)$-norm by (3.12); moreover, $\|p_t\|$ is estimated by (3.13) (with $k_2 = 0$ as we are taking $g_2 \equiv 0$ now), subject to the feedback bound (1.28). Thus the counterpart of (3.51) is

$$\text{R.H.S. of (3.31)} \geqq C_{1\varepsilon} \int_{t_0}^{\infty} e^{-2\beta t} E(t)\, dt - C_{2\varepsilon} E(t_0)$$
$$- C_{3h\varepsilon} \int_{t_0}^{\infty} e^{-2\beta t} [\|A^{-1/2} w\|^2_{L^2(\Omega)} + \|A^{-1} w_t\|^2_{L^2(\Omega)}]\, dt \tag{3.53}$$

after we use (1.34) on the matrix $H(x)$, with positive constants independent of β and t_0, where $C_{3h\varepsilon} = 0$ if $h(x)$ is linear in x ($C_{3h\varepsilon}$ is proportional to $\max |\nabla(\operatorname{div} h)|$ over $\bar{\Omega}$). In the absence of a feedback control on g_2, i.e., with $g_2 \equiv 0$, we estimate the L.H.S. of (3.31) differently. We use the assumption (1.33) on h and obtain, since $\|\Delta p\|_{L^2(\Gamma)} \leqq C\|\Delta p\|_{H^1(\Omega)} \leqq C\|p\|_{H^3(\Omega)} \leqq C\|A^{3/4} p\|_{L^2(\Omega)}$ because $p|_\Gamma = (\partial p/\partial \nu)|_\Gamma = 0$,

$$\text{L.H.S. of (3.31)} \leqq \frac{c_h}{\varepsilon} \int_\Sigma e^{-2\beta t} \left(\frac{\partial(\Delta p)}{\partial \nu} \right)^2 d\Sigma + \left(\varepsilon - \frac{\gamma}{2} \right) \int_\Sigma e^{-2\beta t} |\nabla(\Delta p)|^2\, dQ$$
$$- \varepsilon \int_{t_0}^{\infty} e^{-2\beta t} \|A^{3/4} p\|^2_{L^2(\Omega)}\, dt \tag{3.54}$$

$$\leqq O(E(t_0)) - \varepsilon \int_{t_0}^{\infty} e^{-2\beta t} E(t)\, dt \tag{3.55}$$

after selecting $\varepsilon < \gamma/2$, dropping the $|\nabla(\Delta p)|^2$-term, and recalling (3.48) and (3.17). Combining (3.53) with (3.55) yields (1.37) in Theorem 1.3A. (A local variation of the argument in (3.54), (3.55) also allows us to take $\gamma = 0$; see [14], [19, footnote 2].) □

3.6. Proof of Theorem 1.3B.

Step 1.[1] From $\|A^{-1/2} w\|^2 = (A^{-1/4} w, A^{-3/4} w)$ and $\|A^{-1} w_t\|^2 = (A^{-3/4} w_t, A^{-5/4} w_t)$ we obtain for any $\varepsilon > 0$

$$2 \int_0^{\infty} e^{-2\beta t} \|A^{-1/2} w(t)\|^2\, dt \leqq \varepsilon \int_0^{\infty} e^{-2\beta t} \|A^{-1/4} w(t)\|^2\, dt$$
$$+ \frac{1}{\varepsilon} \int_0^{\infty} e^{-2\beta t} \|A^{-3/4} w(t)\|^2\, dt, \tag{3.56}$$

[1] All norms and inner products in this subsection are in $L^2(\Omega)$, unless otherwise noted explicitly. Note that the absorption within ε of Theorem 1.3B is in the space variable, not in the time variable as in [8], [39].

$$2\int_0^\infty e^{-2\beta t}\|A^{-1}w_t(t)\|^2\,dt \leqq \varepsilon\int_0^\infty e^{-2\beta t}\|A^{-3/4}w_t(t)\|^2\,dt + \frac{1}{\varepsilon}\int_0^\infty e^{-2\beta t}\|A^{-5/4}w_t(t)\|^2\,dt. \tag{3.57}$$

Thus, in view of (3.56), (3.57), to prove (1.38) in Theorem 1.3B, all we need is the following result.

PROPOSITION 3.7. *With reference to the closed-loop problem* (1.21) *with* $k_2 \equiv 0$, *i.e.*, $\partial w/\partial \nu \equiv 0$ *on* Σ, *for every* ε_1 *there is a* $C_{\varepsilon_1} > 0$ *such that for every* $\beta > 0$

(a)

$$\int_0^\infty e^{-2\beta t}\|A^{-3/4}w(t)\|^2\,dt \leqq \varepsilon_1\int_0^\infty e^{-2\beta t}\|A^{-3/4}w_t(t)\|^2\,dt + C_{\varepsilon_1}E(0), \tag{3.58}$$

(b)

$$\int_0^\infty e^{-2\beta t}\|A^{-5/4}w_t(t)\|^2\,dt \leqq \varepsilon_1\int_0^\infty e^{-2\beta t}\|A^{-5/4}w_{tt}(t)\|^2\,dt + C_{\varepsilon_1}E(0) \tag{3.59}$$

where by (2.3)

$$\|A^{-5/4}w_{tt}\|^2 \leqq 2\|A^{-1/4}w\|^2 + \text{const}\,\|G_1^*A^{-1/2}w_t\|^2_{L^2(\Gamma)} \quad \textit{a.e. in } t, \tag{3.60}$$

so that (3.59) *implies by* (3.60) *and by the feedback bound* (1.28)

$$\int_0^\infty e^{-2\beta t}\|A^{-5/4}w_t(t)\|^2\,dt \leqq 2\varepsilon_1\int_0^\infty e^{-2\beta t}\|A^{-1/4}w(t)\|^2\,dt + C'_{\varepsilon_1}E(0). \tag{3.61}$$

Step 3. Using (3.58) and (3.61) in (3.56) and (3.57), respectively, with $\varepsilon_1 < \varepsilon^2$ we obtain (1.35) in Theorem 1.3B.

Step 4. *Proof of Proposition* 3.7. Patterned after the proof in [39, § 3.2] of Theorem 2 in [8] for the case of wave equation problem considered there. We introduce a new variable

$$u(t,x) = \phi(t)w(t,x), \quad \phi \in C^\infty(R), \quad \phi(0) = \phi'(0) = \phi''(0) = 0, \quad \phi(t) \equiv 1, \quad \text{for } t \geqq 1. \tag{3.62}$$

Then, in the new variable u, (1.21) with $\partial w/\partial \nu \equiv 0$ on Σ becomes

$$u_{tt} + \Delta^2 u = b \qquad \text{in } (0,\infty)\times\Omega, \tag{3.63a}$$

$$u|_{t=0} = 0;\ u_t|_{t=0} = 0 \qquad \text{in } \Omega, \tag{3.63b}$$

$$u|_\Sigma = -G_1^*A^{1/2}u_t + G_1^*A^{-1/2}\phi' w \qquad \text{in } (0,\infty)\times\Gamma, \tag{3.63c}$$

$$\left.\frac{\partial u}{\nabla \nu}\right|_\Sigma \equiv 0 \qquad \text{in } (0,\infty)\times\Gamma, \tag{3.63d}$$

$$b = \phi'' w + 2\phi' w_t, \tag{3.64}$$

whose explicit solution, according to the operator model in e.g., [18], [19] is

$$u(t) = A\int_0^t S(t-\tau)G_t[-G_1^*A^{-1/2}u_t(\tau) + G_1^*A^{-1/2}\phi'(\tau)w(\tau)]\,d\tau + \int_0^t S(t-\tau)b(\tau)\,d\tau \tag{3.65}$$

where $S(t)=\int_0^t C(\tau)\,d\tau$, $C(\cdot)$ being the cosine operator on $L^2(\Omega)$ generated by the negative self-adjoint operator $-A$. Taking the Laplace transform of (3.65) with $\hat{u}_t(\lambda)=\lambda\hat{u}(\lambda)$ by (3.63b) and $\widehat{S(t)}=R(\lambda^2,-A)$, $\lambda=\beta+i\alpha$, we obtain

$$\begin{aligned}[I+\lambda AR(\lambda^2,-A)G_1G_1^*A^{-1/2}]\hat{u}(\lambda)\\ =AR(\lambda^2,-A)G_1G_1^*A^{-1/2}[\widehat{\phi' w}](\lambda)+R(\lambda^2,-A)\hat{b}(\lambda).\end{aligned}\tag{3.66}$$

But, if we recall the definition of $V(\lambda)$ in (1.25) (with $k_1\equiv 1$, $k_2\equiv 0$), we see that

$$\begin{aligned}I+\lambda AR(\lambda^2,-A)G_1G_1^*A^{-1/2}&=R(\lambda^2,-A)A[I+\lambda G_1G_1^*A^{-1/2}+\lambda^2A^{-1}]\\ &=AR(\lambda^2,-A)V(\lambda)\end{aligned}\tag{3.67}$$

inserted in (3.66) yields

$$\begin{aligned}AR(\lambda^2,-A)V(\lambda)\hat{u}(\lambda)&=AR(\lambda^2,-A)G_1G_1^*A^{-1/2}[\widehat{\phi' w}](\lambda)\\ &\quad+AR(\lambda^2,-A)A^{-1}\hat{b}(\lambda),\end{aligned}\tag{3.68}$$

valid at least for all $\lambda=\beta+i\alpha$, $\beta\geqq 0$, with $\lambda^2=\beta^2-\alpha^2+2i\alpha\beta\neq\{-\mu_n,\ n=1,2\cdots\}$, $\mu_n>0$, the eigenvalues of A; i.e., except $\beta=0$ and $\alpha^2=\mu_n$, or $\lambda=\lambda_n=\pm i\sqrt{\mu_n}$, where $R(\lambda^2,-A)$ is *not* defined. Then (3.68) yields after a crucial cancellation

$$\hat{u}(\lambda)=V^{-1}(\lambda)[G_1G_1^*A^{-1/2}(\widehat{\phi' w})+A^{-1}\hat{b}](\lambda).\tag{3.69}$$

Moreover, since $\phi' w$ and b both vanish at $t=0$ by (3.64) and (3.62), we have

$$\begin{aligned}\lambda[\widehat{\phi' w}](\lambda)&=[(\widehat{\phi' w})_t](\lambda)=[\widehat{\phi'' w}+\widehat{\phi' w_t}](\lambda),\\ \lambda\hat{b}(\lambda)&=[\hat{b}_t](\lambda)=[\widehat{\phi''' w}+3\widehat{\phi'' w_1}+2\widehat{\phi' w_{tt}}](\lambda)\end{aligned}\tag{3.70}$$

and thus by (3.69), (3.70)

$$[\hat{u}_t](\lambda)=\lambda\hat{u}(\lambda)=V^{-1}(\lambda)[G_1G_1^*A^{-1/2}[(\widehat{\phi' w})_t]+A^{-1}\hat{b}_t](\lambda).\tag{3.71}$$

Then, by Theorem 1.1(iii), the resolvent $R(\lambda,\mathcal{A})$ is well defined also on the imaginary axis. Hence, we have that $V^{-1}(\lambda)\in\mathcal{L}(L^2(\Omega))$ in the closed right half plane $\operatorname{Re}\lambda\geqq 0$, including the imaginary axis $\beta=0$, and is holomorphic in $\operatorname{Re}\lambda>0$. Moreover, for any λ in the *closed* rectangle $\mathcal{R}_{\alpha_0}$: $0\leqq\operatorname{Re}\lambda\leqq 1$, $|\operatorname{Im}\lambda|\leqq\alpha_0$, with $\alpha_0>0$ arbitrary, we have

$$\|V^{-1}(\lambda)\|_{\mathcal{L}(L^2(\Omega))}\leqq C_{\alpha_0},\qquad \lambda\in\mathcal{R}_{\alpha_0}.\tag{3.72}$$

Proof of part (a), (3.58). With $\lambda=\beta+i\alpha$, β fixed, $0<\beta\leqq 1$, we obtain from (3.69) and (3.72) by use of Parseval equality and obvious majorizations on bounded operators since $\phi'=\phi''\equiv 0$ for $t\geqq 1$

$$\begin{aligned}&\int_{|\alpha|\leqq\alpha_0}\|A^{-3/4}\hat{u}(\lambda)\|^2\,d\alpha\\ &\leqq C_{\alpha_0}\left\{\int_{|\alpha|\leqq\alpha_0}\|[A^{-1/4}\widehat{\phi' w}](\lambda)\|^2\,d\alpha+\int_{|\alpha|\leqq\alpha_0}\|[A^{-3/4}\hat{b}](\lambda)\|^2\,d\alpha\right\}\end{aligned}$$

$$\leqq C_{\alpha_0}\left\{\int_0^\infty e^{-2\beta t}\|A^{-1/4}\phi' w\|^2\,dt+\int_0^\infty e^{-2\beta t}\|A^{-3/4}(\phi'' w+2\phi' w_t)\|^2\,dt\right\} \tag{3.73}$$

(by (3.64b))

$$\leqq C_\phi C_{\alpha_0}\left\{\int_0^1 \|A^{-1/4}w\|^2+\|A^{-3/4}w_t\|^2\,dt\right\}\leqq C_\phi C_{\alpha_0}E(0)$$

by the contraction property of $E(t)$. Next, for $|\alpha|>\alpha_0>0$, where $1/|\lambda|^2\leqq(\alpha^2/\alpha_0^2)(1/|\lambda|^2)\leqq 1/\alpha_0^2$, the Parseval equality gives

$$\begin{aligned}\int_{|\alpha|>\alpha_0}|A^{-3/4}\hat{u}(\lambda)\|^2\,d\alpha &= \int_{|\alpha|>\alpha_0}\frac{1}{|\lambda|^2}\|A^{-3/4}\lambda\hat{u}(\lambda)\|^2\,d\alpha\\ &\leqq\frac{1}{\alpha_0^2}\int_{|\alpha|>\alpha_0}\|A^{-3/4}\hat{u}_t(\lambda)\|^2\,d\alpha\\ &\leqq\frac{2\pi}{\alpha_0^2}\int_0^\infty e^{-2\beta t}\|A^{-3/4}u_t\|^2\,dt.\end{aligned} \tag{3.74}$$

Choosing $1/\alpha_0^2=\varepsilon_1$, we obtain from (3.73), (3.74) and the Parseval equality

$$2\pi\int_0^\infty e^{-2\beta t}\|A^{-3/4}u(t)\|^2\,dt\leqq\varepsilon_1\int_0^\infty e^{-2\beta t}\|A^{-3/4}u_t(t)\|^2\,dt+C_{\varepsilon_1}E(0), \tag{3.75}$$

which is an inequality of the type desired, but for u, not w. We return from u to w: since $u\equiv w$ for $t\geqq 1$ by (3.62), and $u_t=\phi' w+\phi w_t$, so that

$$\|A^{-3/4}u_t\|^2=O(\|A^{-1/4}w\|^2+\|A^{-3/4}w_t\|^2), \tag{3.76}$$

$$\begin{aligned}\int_0^\infty e^{-2\beta t}\|A^{-3/4}w(t)\|^2\,dt &= \int_0^1 e^{-2\beta t}\|A^{-3/4}w(t)\|^2\,dt+\int_1^\infty e^{-2\beta t}\|A^{-3/4}u(t)\|^2\,dt\\ &\leqq CE(0)+\varepsilon_1\int_0^1 e^{-2\beta t}\|A^{-3/4}u_t(t)\|^2\,dt\\ &\quad+\varepsilon_1\int_1^\infty e^{-2\beta t}\|A^{-3/4}w_t(t)\|^2\,dt+C_{\varepsilon_1}E(0)\\ &\leqq C_{\varepsilon_1}E(0)+\varepsilon_1\int_0^1 e^{-2\beta t}\|A^{-3/4}w_t(t)\|^2\,dt\\ &\quad+\varepsilon_1\int_1^\infty e^{-2\beta t}\|A^{-3/4}w_t(t)\|^2\,dt\end{aligned} \tag{3.77}$$

(by (3.76) in $0\leqq t\leqq 1$)

and part (a), (3.58) is proved.

Proof of part (b), (3.59). The proof is conceptually similar. We start from (3.71). We have that the resolvent $R(\lambda,\mathscr{A})$ is well defined in all of $\operatorname{Re}\lambda\geqq 0$ and hence (see the proof of Theorem 1.1 following (2.5))

$$\|A^{-1/4}V^{-1}(\lambda)A^{1/4}\|_{\mathscr{L}(L^2(\Omega))}\leqq C_{\alpha_0},\qquad \lambda\in\mathscr{R}_{\alpha_0}. \tag{3.78}$$

Then, from (3.71), writing $A^{-5/4}V^{-1}(\lambda)A^{1/4}A^{-5/4}\hat{b}$ and using (3.78) and (3.70), since $\phi' = \phi'' = \phi''' = 0$ for $t \geqq 1$ we obtain

$$\begin{aligned}
&\int_{|\alpha|\leqq\alpha_0} \|A^{-5/4}\hat{u}_t(\lambda)\|^2\, d\alpha \\
&\quad \leqq C_{\alpha_0}\left\{\int_{|\alpha|\leqq\alpha_0} \|G_1^* A^{-1/2}[(\phi' w)_t](\lambda)\|^2\, d\alpha + \int_{|\alpha|\leqq\alpha_0} \|A^{-5/4}[\hat{b}_t](\lambda)\|^2\, d\alpha\right\} \\
&\quad \leqq C_{\alpha_0}\left\{\int_0^\infty e^{-2\beta t}\|G_1^* A^{-1/2}(\phi'' w + \phi' w_t)\|^2\, dt\right. \\
&\qquad \left. + \int_0^\infty e^{-2\beta t}\|A^{-5/4}(\phi''' w + 3\phi'' w_t + 2\phi' w_{tt}))\|^2\, dt\right\} \quad \text{(by (3.70))} \qquad (3.79) \\
&\quad \leqq C_\phi C_{\alpha_0}\left\{\int_0^1 [\|A^{-1/4}w\|^2 + \|G_1^* A^{-1/2} w_t\|^2_{L^2(\Gamma)}]\, dt\right. \\
&\qquad \left. + \int_0^1 [\|A^{-1/4}w\|^2 + \|A^{-3/4}w_t\|^2 + \|A^{-5/4}w_{tt}\|^2]\, dt\right\} \\
&\quad \leqq C_{\alpha_0} E(0),
\end{aligned}$$

since from (2.3) $A^{-5/4}w_{tt} = -A^{-1/4}w - A^{-1/4}G_1 G_1^* A^{-1/2} w_t$, and then (3.1), the contraction of $E(t)$, and (1.28) apply. The rest of the proof proceeds as before. We obtain as in (3.74)

$$\int_{|\alpha|>\alpha_0} \|A^{-5/4}\hat{u}_t(\lambda)\|^2\, d\alpha \leqq \frac{2\pi}{\alpha_0^2}\int_0^\infty e^{-2\beta t}\|A^{-5/4}u_{tt}\|^2\, dt \qquad (3.80)$$

and have as in (3.75) with $\varepsilon_1 = 1/\alpha_0^2$

$$2\pi\int_0^\infty e^{-2\beta t}\|A^{-5/4}u_t(t)\|^2\, dt \leqq \varepsilon_1\int_0^\infty e^{-2\beta t}\|A^{-5/4}u_{tt}(t)\|^2\, dt + C_{\varepsilon_1}E(0), \qquad (3.81)$$

from which the passage from u satisfying (3.81) to w satisfying (3.59) takes place as before. The proof of Theorem 1.3B is complete. □

Appendix A. Proof of Proposition 3.1. Adapting the multiplier technique of [18], [19] to present circumstances, where $t_0 \leqq t \leqq \infty$, we multiply (3.16a) by $e^{-2\beta t}h \cdot \nabla(\Delta p)$, where $\beta > 0$ is an arbitrary constant and $h(x) \in C^2(\bar{\Omega})$ is a vector field on $\bar{\Omega}$. For future reference to uniform stabilization problems for (1.1a) with boundary conditions of possibly different type from (1.1c, d), we will first derive a general identity for p which solves only (3.16a) with no use of boundary conditions (3.16c, d) (see (A.8) below). Only subsequently will we specialize such identity (A.8) to p which also satisfies the boundary conditions (3.16c, d).

Identity for p which satisfies (3.16a). With $Q = (t_0, \infty)\times\Omega$, $\Sigma = (t_0, \infty)\times\Gamma$, we multiply (3.16a) by $e^{-2\beta t}h \cdot \nabla(\Delta p)$ and integrate by parts. We will use the identity, obtained via the divergence theorem

$$\int_\Omega fh \cdot \nabla\psi\, d\Omega = \int_\Gamma f\psi h \cdot \nu\, d\Gamma - \int_\Omega \psi h \cdot \nabla f\, d\Omega - \int_\Omega f\psi \operatorname{div} h\, d\Omega \qquad (A.1)$$

with f, ψ two $H^1(\Omega)$-functions. In addition we will use the identity

$$\int_Q e^{-2\beta t}\Delta\phi h\cdot\nabla\phi\,dQ=\int_\Sigma e^{-2\beta t}\frac{\partial\phi}{\partial\nu}h\cdot\nabla\phi\,d\Sigma-\frac{1}{2}\int_\Sigma e^{-2\beta t}|\nabla\phi|^2h\cdot\nu\,d\Sigma$$
$$-\int_Q e^{-2\beta t}H\nabla\phi\cdot\nabla\phi\,dQ+\frac{1}{2}\int_Q e^{-2\beta t}|\nabla\phi|^2\operatorname{div}h\,dQ \tag{A.2}$$

already proved in, e.g., [39, eq. (A.3), App. A] (with similar multiplier $e^{-2\beta t}h\cdot\nabla\phi$), where $H(x)$ is the matrix defined in (1.33) (transpose of the Jacobian of $h(x)$).

Term $p_{tt}\,e^{-2\beta t}h\cdot\nabla(\Delta p)$. Integrating by parts in t

$$\int_Q e^{-2\beta t}p_{tt}h\cdot\nabla(\Delta p)\,dQ=\lim_{T\to\infty}e^{-2\beta T}(p_t(T),h\cdot\nabla(\Delta p(T)))_\Omega$$
$$-e^{-2\beta t_0}(p_1,h\cdot\nabla(\Delta p_0))_\Omega \tag{A.3}$$
$$+2\beta\int_Q e^{-2\beta t}p_th\cdot\nabla(\Delta p)\,dQ-\int_Q e^{-2\beta t}p_th\cdot\nabla(\Delta p_t)\,dQ$$

(using (A.1) in the last integral above with $f=p_t$ and $\psi=\Delta p_t$)

$$=\lim_{T\to\infty}e^{-2\beta T}(p_t(T),h\cdot\nabla(\Delta p(T))_\Omega-e^{-2\beta t_0}(p_1,h\cdot\nabla(\Delta p_0))_\Omega$$
$$+2\beta\int_Q e^{-2\beta t}p_th\cdot\nabla(\Delta p)\,dQ-\int_\Sigma e^{-2\beta t}p_t\Delta p_th\cdot\nu\,d\Sigma \tag{A.4}$$
$$+\int_Q e^{-2\beta t}p_t\Delta p_t\operatorname{div}h\,dQ+\int_Q e^{-2\beta t}\Delta p_th\cdot\nabla p_t\,dQ.$$

Invoking (A.2) with $\phi=p_t$ for the last integral in (A.5) we obtain

$$\int_Q e^{-2\beta t}p_{tt}h\cdot\nabla(\Delta p)\,dQ=\lim_{T\to\infty}e^{-2\beta T}(p_t(T),h\cdot\nabla(\Delta p)(T)))_\Omega-e^{-2\beta t_0}(p_1,h\cdot\nabla(\Delta p_0))_\Omega$$
$$-\int_\Sigma e^{-2\beta t}p_t\Delta p_th\cdot\nu\,d\Sigma+\int_\Sigma e^{-2\beta t}\frac{\partial p_t}{\partial\nu}h\nabla p_t\,d\Sigma$$
$$-\frac{1}{2}\int_\Sigma e^{-2\beta t}|\nabla p_t|^2h\cdot\nu\,d\Sigma \tag{A.5}$$
$$-\int_Q e^{-2\beta t}H\nabla p_t\cdot\nabla p_t\,dQ+\frac{1}{2}\int_Q e^{-2\beta t}|\nabla p_t|^2\operatorname{div}h\,dQ$$
$$+2\beta\int_Q e^{-2\beta t}p_th\cdot\nabla(\Delta p)\,dQ+\int_Q e^{-2\beta t}\Delta p_tp_t\operatorname{div}h\,dQ.$$

Now, using Green's first theorem on the last integral at the right of (A.5) along with the identity

$$\nabla p_t\cdot\nabla(p_t\operatorname{div}h)=p_t\nabla(\operatorname{div}h)\cdot\nabla p_t+|\nabla p_t|^2\operatorname{div}h$$

we finally obtain from (A.5)

$$\int_Q e^{-2\beta t}p_{tt}h\cdot\nabla(\Delta p)\,dQ=\lim_{T\to\infty}[e^{-2\beta t}(p_t,h\cdot\nabla(\Delta p))_\Omega]_{t_0}^T-\int_\Sigma e^{-2\beta t}p_t\Delta p_th\cdot\nu\,d\Sigma$$
$$-\frac{1}{2}\int_\Sigma e^{-2\beta t}|\nabla p_t|^2h\cdot\nu\,d\Sigma+\int_\Sigma e^{-2\beta t}\frac{\partial p_t}{\partial\nu}h\cdot\nabla p_t\,d\Sigma$$

$$
\begin{aligned}
&+\int_\Sigma e^{-2\beta t}\frac{\partial p_t}{\partial \nu}p_t \operatorname{div} h\, dQ-\int_Q e^{-2\beta t}H\nabla p_t\cdot\nabla p_t\, dQ\\
&-\frac{1}{2}\int_Q e^{-2\beta t}|\nabla p_t|^2 \operatorname{div} h\, dQ\\
&-\int_Q e^{-2\beta t}p_t\nabla(\operatorname{div} h)\cdot\nabla p_t\, dQ\\
&+2\beta\int_Q e^{-2\beta t}p_t h\cdot\nabla(\Delta p)\, dQ.
\end{aligned}
\tag{A.6}
$$

Term $e^{-2\beta t}\Delta^2 ph\cdot\nabla(\Delta p)$. Using identity (A.2), this time with $\phi=\Delta p$, we obtain

$$
\begin{aligned}
&\int_Q e^{-2\beta t}\Delta^2 ph\cdot\nabla(\Delta p)\, dQ\\
&\quad=\int_\Sigma e^{-2\beta t}\frac{\partial(\Delta p)}{\partial\nu}h\cdot\nabla(\Delta p)\, d\Sigma-\frac{1}{2}\int_\Sigma e^{-2\beta t}|\nabla(\Delta p)|^2 h\cdot\nu\, d\Sigma\\
&\quad\quad-\int_Q e^{-2\beta t}H\nabla(\Delta p)\cdot\nabla(\Delta p)\, dQ+\frac{1}{2}\int_Q e^{-2\beta t}|\nabla(\Delta p)|^2\operatorname{div} h\, dQ.
\end{aligned}
\tag{A.7}
$$

Summing up (A.6) and (A.7) and recalling (3.16a) we obtain

$$
\begin{aligned}
&\int_\Sigma e^{-2\beta t}\frac{\partial p_t}{\partial\nu}h\cdot\nabla p_t\, d\Sigma+\int_\Sigma e^{-2\beta t}\frac{\partial p_t}{\partial\nu}p_t\operatorname{div} h\, d\Sigma-\int_\Sigma e^{-2\beta t}p_t\Delta p_t h\cdot\nu\, d\Sigma\\
&\quad-\frac{1}{2}\int_\Sigma e^{-2\beta t}|\nabla p_t|^2 h\cdot\nu\, d\Sigma+\int_\Sigma e^{-2\beta t}\frac{\partial(\Delta p)}{\partial\nu}h\cdot\nabla(\Delta p)\, d\Sigma\\
&\quad-\frac{1}{2}\int_\Sigma e^{-2\beta t}|\nabla(\Delta p)|^2 h\cdot\nu\, d\Sigma\\
&=\int_Q e^{-2\beta t}[H\nabla(\Delta p)\cdot\nabla(\Delta p)+H\nabla p_t\cdot\nabla p_t]\, dQ\\
&\quad+\frac{1}{2}\int_Q e^{-2\beta t}\{|\nabla p_t|^2-|\nabla(\Delta p)|^2\}\operatorname{div} h\, dQ+\int_Q e^{-2\beta t}p_t\nabla p_t\cdot\nabla(\operatorname{div} h)\, dQ\\
&\quad-2\beta\int_Q e^{-2\beta t}p_t h\cdot\nabla(\Delta p)\, dQ+\int_Q e^{-2\beta t}(F_1+F_2)h\cdot\nabla(\Delta p)\, dQ\\
&\quad-\lim_{T\to\infty}[e^{-2\beta t}(p_t, h\cdot\nabla(\Delta p))_\Omega]_{t_0}^T
\end{aligned}
\tag{A.8}
$$

which is the desired identity for p which satisfies (3.16a).

Specialization of the left-hand side of (A.8) *to p which satisfies also the boundary conditions* (3.16c, d). Recalling (3.16c, d) we have

$$
p_t|_\Sigma\equiv 0,\quad \nabla p\perp\Gamma\quad\text{and}\quad|\nabla p|=\left|\frac{\partial p}{\partial\nu}\right|\equiv 0\quad\text{on }\Sigma\quad\text{(by (3.16d))}
\tag{A.9a}
$$

$$
\left.\frac{\partial p_t}{\partial\nu}\right|_\Sigma\equiv 0,\quad \nabla p_t\perp\Gamma\quad\text{and}\quad|\nabla p_t|=\left|\frac{\partial p_t}{\partial\nu}\right|\equiv 0\quad\text{on }\Sigma.
\tag{A.9b}
$$

Thus, using (3.16c, d) and (A.9a, b) on the L.H.S. of (A.8), we find that this simplifies to

$$\text{(A.10)}\quad \text{L.H.S. of (A.8)} = \int_\Sigma e^{-2\beta t}\frac{\partial(\Delta p)}{\partial \nu} h\cdot\nabla(\Delta p)\,d\Sigma - \frac{1}{2}\int_\Sigma e^{-2\beta t}|\nabla(\Delta p)|^2 h\cdot \nu\,d\Sigma.$$

Appendix B. Proof of Proposition 3.2. Again, we will first obtain an identity, (B.3) below, for p which solves only (3.16a) and for an arbitrary smooth vector field $h\in C^2(\bar\Omega)$. Next, we shall specialize this identity (B.3) to the case where p also satisfies the boundary conditions (3.16c, d). We multiply (3.16a) by $e^{-2\beta t}\Delta p$ div h and integrate over Q by parts in t and by Green's first theorem:

$$\begin{aligned}\int_Q e^{-2\beta t}p_{tt}\Delta p \text{ div } h\,dQ &= \lim_{T\to\infty}\left[e^{-2\beta t}\int_\Omega \Delta p p_t \text{ div } h\,d\Omega\right]_{t_0}^{T} + 2\beta\int_Q e^{-2\beta t}p_t\Delta p \text{ div } h\,dQ\\ &\quad -\int_Q e^{-2\beta t}p_t\Delta p_t \text{ div } h\,dQ\\ &= \lim_{T\to\infty}\left[e^{-2\beta t}\int_\Omega \Delta p p_t \text{ div } h\,d\Omega\right]_{t_0}^{T} \qquad \text{(B.1)}\\ &\quad +2\beta\int_Q e^{-2\beta t}p_t\Delta p \text{ div } h\,dQ - \int_\Sigma e^{-2\beta t}\frac{\partial p_t}{\partial \nu}p_t \text{ div } h\,d\Sigma\\ &\quad +\int_Q e^{-2\beta t}|\nabla p_t|^2 \text{ div } h\,dQ + \int_Q e^{-2\beta t}p_t\nabla p_t\cdot\nabla(\text{div } h)\,dQ.\end{aligned}$$

Also, again by Green's first theorem

$$\begin{aligned}\int_Q e^{-2\beta t}\Delta^2 p\Delta p \text{ div } h\,dQ &= \int_\Sigma e^{-2\beta t}\frac{\partial(\Delta p)}{\partial\nu}\Delta p \text{ div } h\,d\Sigma\\ &\quad -\int_Q e^{-2\beta t}|\nabla(\Delta p)|^2 \text{ div } h\,dQ \qquad \text{(B.2)}\\ &\quad -\int_Q e^{-2\beta t}\Delta p\nabla(\Delta p)\cdot\nabla(\text{div } h)\,dQ.\end{aligned}$$

Summing up (B.1) and (B.2) and recalling (3.16a) we find the identity

$$\begin{aligned}&\int_Q e^{-2\beta t}\{|\nabla p_t|^2 - |\nabla(\Delta p)|^2\}\text{ div } h\,dQ\\ &\quad = \int_\Sigma e^{-2\beta t}\frac{\partial p_t}{\partial\nu}p_t \text{ div } h\,d\Sigma - \int_\Sigma e^{-2\beta t}\frac{\partial(\Delta p)}{\partial\nu}\Delta p \text{ div } h\,d\Sigma\\ &\quad\quad +\int_Q e^{-2\beta t}\Delta p\nabla(\Delta p)\cdot\nabla(\text{div } h)\,dQ \qquad \text{(B.3)}\\ &\quad\quad -2\beta\int_Q e^{-2\beta t}p_t\Delta p \text{ div } h\,dQ - \int_Q e^{-2\beta t}p_t\nabla p_t\cdot\nabla(\text{div } h)\,dQ\\ &\quad\quad -\lim_{T\to\infty}\left[e^{-2\beta t}\int_\Omega \Delta p p_t \text{ div } h\,d\Omega\right]_{t_0}^{T} + \int_Q e^{-2\beta t}(F_1+F_2)\Delta p \text{ div } h\,dQ\end{aligned}$$

for p satisfying (3.16a). If p now satisfies also the boundary conditions (3.16c, d), then (B.3) specializes to

$$\begin{aligned}
&\int_Q e^{-2\beta t}\left\{|\nabla p_t|^2-|\nabla(\Delta p)|^2\right\}\operatorname{div} h\,dQ\\
&\quad=-\int_\Sigma e^{-2\beta t}\frac{\partial(\Delta p)}{\partial\nu}\Delta p\operatorname{div} h\,d\Sigma+\int_Q e^{-2\beta t}\Delta p\nabla(\Delta p)\cdot\nabla(\operatorname{div} h)\,dQ\\
&\quad-2\beta\int_Q e^{-2\beta t}p_t\Delta p\operatorname{div} h\,dQ-\int_Q e^{-2\beta t}p_t\nabla p_t\cdot\nabla(\operatorname{div} h)\,dQ\\
&\quad+\int_Q e^{-2\beta t}(F_1+F_2)\Delta p\operatorname{div} h\,dQ\\
&\quad+\lim_{T\to\infty}e^{-2\beta T}\int_\Omega\nabla p(T)\cdot\nabla(p_t(T)\operatorname{div} h)\,d\Omega\\
&\quad-e^{-2\beta t_0}\int_\Omega\nabla p_0\cdot\nabla(p_1\operatorname{div} h)\,d\Omega.
\end{aligned}\tag{B.4}$$

REFERENCES

[1] J. BARTOLOMEO, *Uniform stabilization of the Euler-Bernoulli equation with active Dirichlet and non-active Neumann boundary feedback controls*, Ph.D. thesis, Dept. of Mathematics, University of Florida, Gainesville, FL, 1988.

[2] G. CHEN, *Energy decay estimates and exact controllability of the wave equation in a bounded domain*, J. Math. Pures Appl., 9 (1979), pp. 249-274.

[3] R. DATKO, *Extending a theorem of Liapunov to Hilbert space*, J. Math. Anal. Appl., 32 (1970), pp. 610-616.

[4] G. DA PRATO, I. LASIECKA, AND R. TRIGGIANI, *A direct study of Riccati equations arising in boundary control problems for hyperbolic equations*, J. Differential Equations, 64 (1986), pp. 26-47.

[5] F. FLANDOLI, I. LASIECKA, AND R. TRIGGIANI, *Algebraic Riccati equations with non-smoothing observation arising in hyperbolic and Euler-Bernouli equations*, Ann. Mat. Pura Appl. (IV) (1989), pp. 307-382.

[6] P. GRISVARD, *Caracterization de quelques espaces d'interpolation*, Arch. Rational Mech. Anal., 25 (1967), pp. 40-63.

[7] L. F. HO, *Observabilité frontière de l'équation des ondes*, C. R. Acad. Sci. Sér I Math., 302 (1986), pp. 443-446.

[8] J. LAGNESE, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163-182.

[9] ———, *A note on boundary stabilization of wave equations*, SIAM J. Control Optim., 26 (1988), pp. 1250-1256.

[10] ———, *Boundary Stabilization of Thin Plates*, SIAM Studies in Applied Mathematics, 10, Society for Industrial and Applied Mathematics, Philadelphia, 1989.

[11] ———, *Uniform boundary stabilization of homogeneous isotropic plates*, in Lecture Notes in Comput. Sci. 102 (1987), Springer-Verlag, Berlin, New York, pp. 204-215.

[12] J. L. LIONS, *Contrôle des systèmes distribués singuliers*, Gauthier-Villars, Paris, 1983.

[13] ———, *Exact controllability, stabilization and perturbations*, SIAM Rev., 30 (1988), pp. 1-68.

[14] ———, *Controllabilité exacte, perturbations et stabilization de systèmes distribués*, Vols. 1 and 2, Masson, Paris, to appear.

[15] ———, *Un résultat de regularité pour l'opérator* $(\partial^2/\partial t^2)+\Delta^2$, in Current Topics in Partial Differential Equations, Y. Ohya et al., eds, Kinokuniya, Tokyo, 1986.

[16] J. LAGNESE AND J. L. LIONS, *Modeling Analysis and Control of Thin Plates*, Masson, Paris, 1988.

[17] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems*, Vols. I and II, Springer-Verlag, Berlin, New York, 1972.

[18] I. LASIECKA AND R. TRIGGIANI, *Exact controllability of the Euler-Bernoulli equation with $L^2(\Sigma)$-control only in the Dirichlet boundary conditions*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 8 (1988), pp. 35-42.

[19] ———, *Exact controllability of the Euler-Bernoulli equation with controls in the Dirichlet and Neumann boundary conditions: a nonconservative case*, SIAM J. Control Optim., 27 (1989), pp. 330-373.

[20] ———, *Uniform exponential energy decay of wave equations in a bounded region with $L_2(0, \infty; L_2(\Gamma))$-feedback control in the Dirichlet boundary conditions*, J. Differential Equations, 66 (1987), pp. 340-390.

[21] ———, *A cosine operator approach to modeling $L_2(0, T; L_2(\Gamma))$-boundary input hyperbolic equations*, Appl. Math. Optim., 7 (1981), pp. 35-83.

[22] ———, *Regularity of hyperbolic equations under $L_2(0, T; L_2(\Gamma))$-boundary terms*, Appl. Math. Optim., 10 (1983), pp. 275-286.

[23] ———, *Sharp regularity theory for second-order hyperbolic equations of Neumann type. Part* I: *L_2 nonhomogeneous data*, Ann. Mat. Pura Appl., to appear.

[24] ———, *Trace regularity of the solutions of the wave equation with homogeneous Neumann boundary conditions and compactly supported data*, J. Math. Anal. Appl., 141 (1989), pp. 49-71.

[25] ———, *Exact controllability of the wave equation with Neumann boundary control*, Appl. Math. Optim., 19 (1989), pp. 243-290.

[26] ———, *Regularity theory for a class of nonhomogeneous Euler-Bernoulli equations: a cosine operator approach*, Boll. Un. Mat. Ital. B(7), 3 (1989), pp. 199-228.

[27] ———, *Exact controllability of the Euler-Bernoulli equation with boundary controls for displacement and moment*, J. Math. Anal. Appl., 146 (1990), pp. 1-33.

[28] ———, *Riccati equations for hyperbolic partial differential equations with $L_2(0, T; L_2(\Gamma))$-Dirichlet boundary terms*, SIAM J. Control Optim., 24 (1986), pp. 884-926.

[29] ———, *A lifting theorem for the time regularity of solutions to abstract equations with unbounded operators and applications to hyperbolic equations*, Proc. Amer. Math. Soc., 104 (1988), pp. 745-755.

[30] ———, *Infinite horizon quadratic cost problems for boundary control problems*, Proc. 26th Conference on Decision and Control, Los Angeles, CA, 1987, pp. 1005-1010.

[31] ———, *Further results on Exact controllability for an Euler-Bernoulli problem*, 1989, Lecture Notes Control and Information Sci., Springer-Verlag, Berlin; Proceedings of Workshop held at Montpelier, France, Jan. 1989, J. P. Zolésio, ed., to appear.

[32] J. NECAS, *Les méthodes directes en théorie des équations élliptique*, Masson, Paris, 1967.

[33] N. OURADA AND R. TRIGGIANI, *Uniform stabilization for Euler-Bernoulli equations with feedback operator only in the Neumann boundary condition*, preprint, 1989, Differential and Integral Equations, to appear.

[34] D. RUSSELL, *Mathematical models for the elastic beam and their control-theoretic implications*, in Semigroups, Theory and Applications, Vol. II, Res. Notes Math. 152, Pitman, Boston, pp. 177-216.

[35] I. LASIECKA, J. L. LIONS, AND R. TRIGGIANI, *Non-homogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., 65 (1986), pp. 149-192.

[36] W. SYMES, *A trace theorem for solutions of the wave equation and the remote determination of acoustic sources*, Math. Methods Appl. Sci., 5 (1983), pp. 35-93.

[37] R. TRIGGIANI, *A cosine operator approach to modeling $L_2(0, T; L_2(\Gamma))$-boundary input problems for hyperbolic systems*, in Proc. Eighth IFIP Conference, University of Wurzburg, Wurzburg, Federal Republic of Germany, July 1977, Lecture Notes in Control and Information Sci., Springer-Verlag, Berlin, New York, 1978, pp. 380-390.

[38] ———, *Exact boundary controllability on $L^2(\Omega) \times H^{-1}(\Omega)$ for the wave equation with Dirichlet control acting on a portion of the boundary, and related problems*, Appl. Math. Optim., 18 (1988), pp. 241-277.

[39] ———, *Wave equation on a bounded domain with boundary dissipation: an operator approach*, J. Math. Anal. Appl., 137 (1989), pp. 438-461.

[40] ———, *Finite rank relatively bounded perturbations of semi-group generators. Part* III: *a sharp result on the lack of uniform stabilization*, Differential Integral Equations, 3 (1990), pp. 503-522.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 72-106, January 1991

005

GLOBAL EXISTENCE AND ASYMPTOTIC STABILITY FOR A NONLINEAR INTEGRODIFFERENTIAL EQUATION MODELING HEAT FLOW*

DEBORAH BRANDON†

Abstract. This paper studies initial value problems that arise from models for one-dimensional heat flow (with finite wave speeds) in materials with memory. Under assumptions that ensure compatibility of the constitutive relations with the second law of thermodynamics, the resulting integrodifferential equation is hyperbolic near equilibrium. The existence of unique, globally (in time) defined, classical solutions to the problems under consideration is established, provided the data are smooth and sufficiently close to equilibrium. Both Dirichlet and Neumann boundary conditions are treated, as well as the problem on the entire real line.

Local existence is proved using a contraction-mapping argument which involves estimates for linear hyperbolic partial differential equations with variable coefficients. Global existence is obtained by deriving a priori energy estimates. These estimates are based on inequalities for strongly positive Volterra kernels (including a new inequality that is needed due to the form of the constitutive relations). Furthermore, compatibility with the second law plays an essential role in the proof in order to obtain an existence result under less restrictive assumptions on the data.

Key words. integrodifferential equation, second sound, heat flow, hyperbolic equation

AMS(MOS) subject classifications. 45K05, 35L60, 80A20

Introduction. In this paper we establish global existence and asymptotic stability of solutions to initial value problems arising from integral models for heat flow that were introduced in [2]. These models are based on Gurtin and Pipkin's theory of heat conduction [6]. The situations we are concerned with are such that the heat flux depends on the temporal history of the temperature gradient (and possibly on the present value and the history of the temperature), but is independent of the present value of the temperature gradient.

As in [2], we restrict our attention to one-dimensional rigid heat conductors in which the only nonzero component of the heat flux is its x-component, q. Here q and the absolute temperature $\theta > 0$ are functions of x and time t. Moreover, we assume that the material under consideration is homogeneous and has unit density. The first two laws of thermodynamics then take the form

$$e_t + q_x = r, \tag{0.1}$$

$$\eta_t \geqq -\left(\frac{q}{\theta}\right)_x + \frac{r}{\theta}, \tag{0.2}$$

where $e = e(x, t)$ is the (specific) internal energy, $r = r(x, t)$ is the external heat supply, and $\eta = \eta(x, t)$ is the (specific) entropy. Subscripts t and x indicate partial derivatives. If we define the (specific) free energy $\psi = \psi(x, t)$ through

$$\psi := e - \theta\eta, \tag{0.3}$$

* Received by the editors June 28, 1989; accepted for publication (in revised form) February 2, 1990. This work was supported by the U.S. Air Force under grants AFOSR-85-0307 and AFOSR-87-0191. The work was concluded at the Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

then the law of balance of energy (0.1) and the entropy inequality (0.2) can be combined to give the Clausius-Duhem inequality

$$\psi_t + \eta\theta_t + \frac{q\theta_x}{\theta} \leqq 0. \tag{0.4}$$

Gurtin and Pipkin consider materials characterized by constitutive equations that express $\psi(x, t)$, $\eta(x, t)$, and $q(x, t)$ as functionals of $(\theta(x, t), \bar{\theta}^t(x, \cdot), \bar{\theta}^t_x(x, \cdot))$. Here $\bar{\theta}^t$ and $\bar{\theta}^t_x$ denote the summed histories up to time t of the temperature and the temperature gradient. The summed history up to time t of θ is defined by

$$\bar{\theta}^t(x, s) := \int_{t-s}^{t} \theta(x, z)\, dz, \qquad x \in B, \quad s \geqq 0, \tag{0.5}$$

where $B \subset \mathbb{R}$ denotes the interval occupied by the body. Gurtin and Pipkin require that their constitutive relations be compatible with thermodynamics in the sense that the Clausius-Duhem inequality (0.4) is satisfied for all smooth processes consistent with the constitutive relations. They derive conditions that are both necessary and sufficient for compatibility with thermodynamics. These conditions can be summarized roughly as follows:

(i) The entropy is minus the derivative of the free energy with respect to the present value of the temperature;

(ii) The heat flux is determined from the free energy through a differential equation called the heat flux relation;

(iii) A functional differential inequality called the dissipation inequality, holds for all smooth processes.

We note that by virtue of (0.3), condition (ii) implies a relation between q and e and hence, e will generally depend on $\bar{\theta}^t_x$.

MacCamy considered a model motivated by Gurtin and Pipkin's linearized constitutive equations [10]. He replaced the linear equation for the heat flux with

$$q(x, t) = -\int_0^\infty a(s) f(\theta_x(x, t-s))\, ds, \tag{0.6}$$

but retained the linear equation for the internal energy

$$\begin{aligned} e(x, t) &= b + c\theta(x, t) - \int_0^\infty \beta'(s)\bar{\theta}^t(x, s)\, ds \\ &= b + c\theta(x, t) + \int_0^\infty \beta(s)\theta(x, t-s)\, ds. \end{aligned} \tag{0.7}$$

Here b and c are constants, a and β are smooth kernels that decay sufficiently rapidly at infinity, and f is a smooth function. MacCamy proved global existence and asymptotic stability for a corresponding initial boundary value problem. Similar existence theorems for MacCamy's model were established by Dafermos and Nohel [5] and Staffans [13].

MacCamy does not address the issue of compatibility with thermodynamics. However, we can show that there are smooth processes consistent with (0.6), (0.7) but for which an inequality implied by (0.4) is violated; within the context of [5], [10], and [13] this probably is not a serious difficulty since the solutions discussed there remain close to equilibrium (i.e., close to a state where θ is a constant and $\theta_x \equiv 0$), and under reasonable assumptions on a, β, and f, the aforementioned inequality is

satisfied by a suitable class of processes that are close to equilibrium. (See [2, § 1] for further details.)

Here we consider the constitutive relations[1]

$$\psi(x,t)=\hat{\psi}(\theta(x,t))+\int_0^\infty \hat{\Psi}(s,\theta(x,t),\bar{\theta}^t(x,s),\bar{\theta}_x^t(x,s))\,ds,$$

$$\eta(x,t)=-\hat{\psi}'(\theta(x,t))-\int_0^\infty \hat{\Psi}_{,2}(s,\theta(x,t),\bar{\theta}^t(x,s),\bar{\theta}_x^t(x,s))\,ds, \tag{0.8}$$

$$q(x,t)=-\theta(x,t)\int_0^\infty \hat{\Psi}_{,4}(s,\theta(x,t),\bar{\theta}^t(x,s),\bar{\theta}_x^t(x,s))\,ds,$$

and hence by (0.3) we have

$$e(x,t)=\hat{e}(\theta(x,t))+\int_0^\infty \hat{E}(s,\theta(x,t),\bar{\theta}^t(x,s),\bar{\theta}_x^t(x,s))\,ds \tag{0.9}$$

with

$$\hat{e}(\nu):=\hat{\psi}(\nu)-\nu\hat{\psi}'(\nu),\qquad \hat{E}(s,\nu,\alpha,\gamma):=\hat{\Psi}(s,\nu,\alpha,\gamma)-\nu\hat{\Psi}_{,2}(s,\nu,\alpha,\gamma), \tag{0.10}$$
$$s,\nu>0,\quad \alpha\geqq 0,\quad \gamma\in\mathbb{R}.$$

Here $\hat{\Psi}$ is normalized so that

$$\hat{\Psi}(s,\nu,\nu s,0)=0\quad \forall s,\nu>0 \tag{0.11}$$

and $\hat{\Psi}$ satisfies hypotheses which ensure that the integrals in (0.8) will be well behaved for a reasonable class of functions θ.

We assume that $\hat{\Psi}$ satisfies

$$\hat{\Psi}_{,1}(s,\nu,\alpha,\gamma)+\nu\hat{\Psi}_{,3}(s,\nu,\alpha,\gamma)\leqq 0\qquad \forall s,\nu>0,\quad \alpha\geqq 0,\quad \gamma\in\mathbb{R} \tag{0.12}$$

and thus by the main result obtained in [2] the constitutive relations (0.8) are compatible with thermodynamics and

$$\hat{\Psi}_{,j}(s,\nu,\nu s,0)=0,\qquad j=1,2,3,4\quad \forall s,\nu>0. \tag{0.13}$$

Substitution of $(0.8)_3$ and (0.9) into the law of balance of energy (0.1) yields

$$\begin{aligned}&\hat{c}_I(\theta(x,t),\bar{\theta}^t(x,\cdot),\bar{\theta}_x^t(x,\cdot))\theta_t(x,t)\\&\quad+\frac{\partial}{\partial x}\left\{\int_0^\infty \hat{Q}(s,\theta(x,t),\bar{\theta}^t(x,s),\bar{\theta}_x^t(x,s))\,ds\right\}\\&\quad+\int_0^\infty \hat{E}_{,3}(s,\theta(x,t),\bar{\theta}^t(x,s),\bar{\theta}_x^t(x,s))[\theta(x,t)-\theta(x,t-s)]\,ds\\&\quad+\int_0^\infty \hat{E}_{,4}(s,\theta(x,t),\bar{\theta}^t(x,s),\bar{\theta}_x^t(x,s))[\theta_x(x,t)-\theta_x(x,t-s)]\,ds=r(x,t),\end{aligned} \tag{0.14}$$
$$x\in B,\quad t\geqq 0.$$

Here $\hat{Q}$ is given by

$$\hat{Q}(s,\nu,\alpha,\gamma):=-\nu\hat{\Psi}_{,4}(s,\nu,\alpha,\gamma),\qquad s,\nu>0,\quad \alpha\geqq 0,\quad \gamma\in\mathbb{R}, \tag{0.15}$$

[1] We use $F_{,j}$ to denote the partial derivative of a function F with respect to its jth argument.

and

$$\hat{c}_I(\theta(s,t), \bar{\theta}^t(x,\cdot), \bar{\theta}^t_x(x,\cdot)) \tag{0.16}$$
$$:= \hat{e}'(\theta(x,t)) + \int_0^\infty \hat{E}_{,2}(s, \theta(x,t), \bar{\theta}^t(x,s), \bar{\theta}^t_x(x,s))\, ds$$

is the instantaneous heat capacity at $(\theta(x,t), \bar{\theta}^t(x,\cdot), \bar{\theta}^t_x(x,\cdot))$; the equilibrium heat capacity $\hat{c}_E(\nu)$ at the temperature ν is given by

$$\hat{c}_E(\nu) := \hat{e}'(\nu). \tag{0.17}$$

It is generally assumed in practice that the heat capacities are positive.

We seek a smooth solution to (0.14) subject to the initial conditions

$$\begin{aligned} &\theta(x,t) = \varphi(x,t), && x \in B, \quad t < 0, \\ &\theta(x,0) = \theta_0(x), && x \in B, \end{aligned} \tag{0.18}$$

and appropriate boundary conditions if $B \neq \mathbb{R}$. Here $\varphi > 0$ and $\theta_0 > 0$ are prescribed smooth functions. Observe that (0.18) permits a temporal jump discontinuity in θ at $t = 0$. Even if such a discontinuity is present in the data, we can obtain a solution that is smooth for $t \geqq 0$ provided that θ_0 and $r(\cdot, 0)$ satisfy certain compatibility conditions at the endpoints of B.

It follows from the arguments of Gurtin and Pipkin [6] that compatibility with thermodynamics, strict positivity of the equilibrium heat capacity, and some assumptions of nondegeneracy imply that (0.14) is of hyperbolic type near equilibrium. (See [12, Chap. II] for determination of type for equations with memory terms.) The characteristic speeds for (0.14) are not constant, and it is therefore possible that weak waves will be amplified and shocks will develop. On the other hand, (0.14) includes a natural damping mechanism induced by memory. It is not clear which effect is dominant. A great deal of insight into this question is given by Chen [3], who assumed the existence of solutions containing singularities called temperature rate waves and obtained a formula for the amplitude of these waves. He found that an amplitude of small initial value decays as $t \to \infty$, and if the initial amplitude is large then blowup may occur in finite time. This suggests that when the data are close to equilibrium, (0.14) has a global solution, whereas if the data are sufficiently far away from equilibrium the solution may develop singularities in finite time.

To keep the analysis relatively clean, while retaining the important features (from the point of view of the analysis) of (0.14) we treat the following special case in detail:

$$\begin{aligned} \psi(x,t) &= \hat{\psi}(\theta(x,t)) - \frac{1}{\theta(x,t)} \int_0^\infty a'(s) F(\bar{\theta}^t_x(x,s))\, ds, \\ \eta(x,t) &= -\hat{\psi}'(\theta(x,t)) - \frac{1}{\theta(x,t)^2} \int_0^\infty a'(s) F(\bar{\theta}^t_x(x,s))\, ds, \\ q(x,t) &= \int_0^\infty a'(s) F'(\bar{\theta}^t_x(x,s))\, ds. \end{aligned} \tag{0.19}$$

Here $\hat{\psi}: (0,\infty) \to \mathbb{R}$, $a: [0,\infty) \to \mathbb{R}$, and $F: \mathbb{R} \to \mathbb{R}$ are smooth functions with $a \in W^{3,1}(0,\infty)$ and $F(0) = 0$. Observe that unlike (0.8), in (0.19) there is no dependence on the summed history of the temperature; moreover, here the kernel (a') factors out. We assume that

$$a \text{ is convex}, \qquad F \geqq 0; \tag{0.20}$$

the arguments used in [2] can be applied in the present setting to show that (0.20) implies that the constitutive equations (0.19) are compatible with thermodynamics. We note that by (0.20) we have

$$a' \leqq 0, \quad a \geqq 0, \quad F'(0) = 0, \quad F''(0) \geqq 0. \tag{0.21}$$

The corresponding equation for e is

$$e(x, t) = \hat{e}(\theta(x, t)) - \frac{2}{\theta(x, t)} \int_0^\infty a'(s) F(\bar{\theta}_x^t(x, s))\, ds, \tag{0.22}$$

where $\hat{e}$ is as in $(0.10)_1$. Thus (0.1) yields

$$\begin{aligned} &\left(\hat{e}'(\theta(x, t)) + \frac{2}{\theta(x, t)^2} \int_0^\infty a'(s) F(\bar{\theta}_x^t(x, s))\, ds\right) \theta_t(x, t) \\ &\quad + \int_0^\infty a'(s) F''(\bar{\theta}_x^t(x, s)) \bar{\theta}_{xx}^t(x, s)\, ds \\ &\quad - \frac{2}{\theta(x, t)} \int_0^\infty a'(s) F'(\bar{\theta}_x^t(x, s))[\theta_x(x, t) - \theta_x(x, t-s)]\, ds = r(x, t), \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad x \in B, \quad t \geqq 0. \end{aligned} \tag{0.23}$$

We establish global existence and asymptotic stability of smooth solutions to the initial value problem (0.23), (0.18) for smooth data (r, φ, θ_0) that are close to equilibrium. We treat Dirichlet and Neumann boundary conditions as well as the problem with $B = \mathbb{R}$. We also make some remarks concerning the extension of our work to the initial value problem (0.14), (0.18).

To indicate the nature of our results let us consider the case where $B = [0, 1]$, $\varphi \equiv \theta_0 \equiv \theta^*$, with Dirichlet boundary conditions

$$\theta(0, t) = \theta(1, t) = \theta^*, \qquad t \geqq 0, \tag{0.24}$$

where $\theta^* > 0$ is a given constant.

In order to prove global existence of solutions to (0.23), (0.24), (0.18) we need to make additional assumptions on the constitutive relations and on the data. Concerning the constitutive equations we require that

$$a \not\equiv 0, \tag{0.25}$$

and we strengthen the inequality $(0.21)_4$ to the strict inequality

$$F''(0) > 0. \tag{0.26}$$

These two conditions imply that the linearized relation for the heat flux is nontrivial. We also assume that the equilibrium heat capacity is strictly positive, i.e.,[2]

$$\hat{e}' > 0. \tag{0.27}$$

Assumptions (0.25)–(0.27) imply that (0.23) is hyperbolic near equilibrium. Since we have dependence on the summed history of the temperature gradient (for which we do not obtain a pointwise bound), we need to make a growth restriction on F that is related to the decay rate of a. In addition, we assume that the heat supply r is smooth, decays with time, and is small in a sense that will be stated more precisely later.

[2] For our purposes it suffices to assume that $\hat{e}'(\theta^*) > 0$; however, assumption (0.27) is in accord with experience and leads to certain simplifications in the proofs of Theorems 1.2 and 1.3.

Moreover, to ensure the existence of smooth classical solutions the heat supply r must satisfy the condition

$$r(0,0)=r(1,0)=r_t(0,0)=r_t(1,0)=0, \tag{0.28}$$

which guarantees compatibility of the data with the boundary conditions at $t=0$. We note that our assumptions imply that a is a strongly positive-definite kernel in the sense of [11]. Inequalities for such kernels play an essential role in the proof of global existence.

Observe that if $r\equiv 0$, then $\theta\equiv\theta^*$ is a solution. We look for classical solutions to (0.23), (0.24), (0.18) near the prescribed equilibrium temperature θ^* for $t\geqq 0$. We show that (0.23), (0.24), (0.18) has a unique solution $\theta>0$ with θ, θ_x, θ_t, θ_{xx}, θ_{xt}, θ_{tt}, θ_{xxx}, θ_{xxt}, θ_{xtt}, $\theta_{ttt}\in C([0,\infty);L^2(0,1))$ and θ, θ_x, θ_t, θ_{xx}, θ_{xt}, $\theta_{tt}\in L^2((0,\infty);L^2(0,1))\cap L^\infty((0,\infty);L^2(0,1))$. Moreover, as $t\to\infty$, $\theta(\cdot,t)\to\theta^*$ and $\theta_x(\cdot,t)$, $\theta_t(\cdot,t)\to 0$ uniformly on $[0,1]$. An analogous result can be obtained for Neumann boundary conditions as well as for the problem with $B=\mathbb{R}$.

The arguments used to prove global existence in [5], [10], and [13] for MacCamy's model are similar in spirit to the arguments used here. The primary differences between our existence proof and those for MacCamy's model arise from the dependence of e on the summed history of θ_x. This dependence complicates the analysis and necessitates the use of a new inequality for strongly positive-definite kernels. Global existence is obtained by deriving a priori estimates; in these derivations we exploit the compatibility of our constitutive relations with thermodynamics, i.e., we make use of the entropy inequality (0.2). It is interesting to note that we can obtain an existence result for (0.23), (0.24), (0.18) without utilizing the thermodynamical restrictions, provided the linearized equation has the appropriate features. However, the compatibility conditions imposed on our constitutive relations by the thermodynamical restrictions allow us to establish a global existence result under less restrictive assumptions on the data.

The paper is organized as follows. Precise statements of global existence results are given in § 1. Section 2 is concerned with appropriate local existence results and with properties of strongly positive-definite kernels relevant to our needs. Section 3 is devoted to the proof of the theorems stated in § 1; the proof for the problem with Dirichlet boundary conditions is discussed in detail and remarks are made concerning other boundary conditions.

1. Statement of results. We first consider the problem

$$\begin{aligned}&\left(\hat{e}'(\theta(x,t))+\frac{2}{\theta(x,t)^2}\int_0^\infty a'(s)F(\bar{\theta}_x^t(x,s))\,ds\right)\theta_t(x,t)\\&\quad+\int_0^\infty a'(s)F''(\bar{\theta}_x^t(x,s))\bar{\theta}_{xx}^t(x,s)\,ds\\&\quad-\frac{2}{\theta(x,t)}\int_0^\infty a'(s)F'(\bar{\theta}_x^t(x,s))[\theta_x(x,t)-\theta_x(x,t-s)]\,ds=r(x,t),\\&\qquad\qquad x\in[0,1],\quad t\geqq 0,\end{aligned} \tag{1.1}$$

$$\theta(x,t)=\theta^*,\qquad x\in[0,1],\quad t<0, \tag{1.2}$$

$$\theta(x,0)=\theta_0(x),\qquad x\in[0,1], \tag{1.3}$$

$$\theta(0,t)=\theta(1,t)=\theta^*,\qquad t\geqq 0. \tag{1.4}$$

Here, $\theta^*>0$ is a given constant and $\theta_0:[0,1]\to(0,\infty)$ is a prescribed smooth function.

Concerning $\hat{e}$, F, and a we require that

$$\hat{e} \in C^4(0, \infty), \tag{1.5}$$

$$\hat{e}' > 0; \tag{1.6}$$

$$F \in C^5(\mathbb{R}), \tag{1.7}$$

$$F(0) = 0, \quad F''(0) > 0, \quad F(\gamma) \geqq 0 \quad \forall \gamma \in \mathbb{R}, \tag{1.8}$$

and there are constants $K > 0$, $k > 1$ such that

$$|F^{(j)}(\xi) - F^{(j)}(0)| \leqq K(|\xi| + |\xi|^k), \qquad j = 0, 1, 2, 3, 4, 5, \quad \xi \in \mathbb{R}; \tag{1.9}$$

$$a \in W^{3,1}(0, \infty), \quad a \text{ is strongly positive definite}, \quad a'' \geqq 0, \tag{1.10}$$

and

$$\int_0^\infty |a'(z)| z^{k+1}\, dz, \int_0^\infty |a''(z)| z^{k+1}\, dz, \int_0^\infty |a'''(z)| z^k\, dz < \infty. \tag{1.11}$$

We note that there is some redundancy in (1.11) due to assumption (1.10); however, we feel that the present form (which is not as compact as possible) is clearer from an expository point of view. The definition of a strongly positive-definite kernel is given in the next section. For now, it suffices to know that

(i) $(1.10)_{1,2}$ implies $a(0) > 0$;

(ii) If $a \in W^{3,1}(0, \infty)$, $a \not\equiv 0$, and $a'' \geqq 0$, then a is strongly positive definite.

The data are assumed to have the following regularity:

$$\theta_0 \in H^3(0, 1), \tag{1.12}$$

$$r, r_x, r_t, r_{xt}, r_{tt} \in C([0, \infty); L^2(0, 1)) \cap L^2((0, \infty); L^2(0, 1)) \cap L^\infty((0, \infty); L^2(0, 1)), \tag{1.13}$$

$$r(\cdot, 0) \in H^2(0, 1), \qquad r_{ttt} \in L^2((0, \infty); L^2(0, 1)). \tag{1.14}$$

We also assume that the following compatibility conditions hold on the boundary:

$$\theta_0(0) = \theta_0(1) = \theta^*, \tag{1.15}$$

$$r(0, 0) = r(1, 0) = 0, \tag{1.16}$$

$$r_t(0, 0) = a(0)F''(0)\left(-\theta_0''(0) + \frac{2}{\theta^*}\theta_0'(0)^2\right), \tag{1.17}$$

$$r_t(1, 0) = a(0)F''(0)\left(-\theta_0''(1) + \frac{2}{\theta^*}\theta_0'(1)^2\right). \tag{1.18}$$

The interpretation of (1.15) is clear; conditions (1.16)–(1.18) ensure that $\theta_t(\cdot, 0)$ and $\theta_{tt}(\cdot, 0)$ vanish on the boundary. To state our results, it is convenient to define

$$\Theta_0 := \int_0^1 ([\theta_0(x) - \theta^*]^2 + \theta_0'(x)^2 + \theta_0''(x)^2)\, dx \tag{1.19}$$

and

$$\mathbf{R}_0 := \sup_{t \geqq 0} \int_0^1 (r^2 + r_t^2)(x, t)\, dx + \int_0^1 r_x^2(x, 0)\, dx + \left(\sup_{\substack{x \in [0,1] \\ t \geqq 0}} |r(x, t)|\right) + \int_0^\infty \int_0^1 (r^2 + r_t^2 + r_{tt}^2)(x, t)\, dx\, dt. \tag{1.20}$$

We establish the following result.

THEOREM 1.1. *Assume that* (1.5)-(1.11) *hold. Then there is a constant* $\delta > 0$ *such that for all* θ_0 *and* r *satisfying* (1.12)-(1.18) *and*

$$\Theta_0 + R_0 \leqq \delta^2, \tag{1.21}$$

the initial value problem (1.1)-(1.4) *has a unique solution* $\theta > 0$ *with*

$$\theta, \theta_x, \theta_t, \theta_{xx}, \theta_{xt}, \theta_{tt}, \theta_{xxx}, \theta_{xxt}, \theta_{xtt}, \theta_{ttt} \in C([0, \infty); L^2(0, 1)) \tag{1.22}$$

and

$$\theta - \theta^*, \theta_x, \theta_t, \theta_{xx}, \theta_{xt}, \theta_{tt} \in L^\infty((0, \infty); L^2(0, 1)) \cap L^2((0, \infty); L^2(0, 1)). \tag{1.23}$$

Moreover, as $t \to \infty$

$$\theta(\cdot, t) \to \theta^* \tag{1.24}$$

and

$$\theta_x(\cdot, t), \theta_t(\cdot, t) \to 0 \quad \textit{uniformly on } [0, 1]. \tag{1.25}$$

Remark 1.1. The constant δ in Theorem 1.1 depends on θ^* and on properties of the functions appearing in the constitutive relations.

Remark 1.2. By the Sobolev embedding theorem, (1.22) implies that $\theta \in C^2([0, 1] \times [0, \infty))$.

A result analogous to Theorem 1.1 can be established if we replace the Dirichlet boundary conditions (1.4) with Neumann boundary conditions

$$\theta_x(0, t) = \theta_x(1, t) = 0, \qquad t \geqq 0. \tag{1.26}$$

Remark 1.3. Under the assumptions of Theorem 1.2 below (1.26) holds if and only if

$$q(0, t) = q(1, t) = 0, \qquad t \geqq 0. \tag{1.27}$$

(Recall that the heat flux q is given by $(0.19)_3$.) It is obvious that (1.26) implies (1.27). In order to show that (1.27) implies (1.26) we first differentiate the relation for the heat flux $(0.19)_3$ with respect to t on the boundary, making use of (1.2). We then add and subtract terms to obtain the identity

$$\begin{aligned} a(0)\theta_x(\xi, t) + \int_0^t a'(t-s)\theta_x(\xi, s)\, ds \\ = \frac{1}{F''(0)} \frac{\partial}{\partial t} \left\{ \int_0^t \theta_x(\xi, s) \int_{t-s}^\infty a'(y) \right. \\ \left. \cdot \int_0^1 [F''(z\bar{\theta}_x^t(\xi, y)) - F''(0)]\, dz\, dy\, ds \right\}, \\ \xi = 0, 1, \quad t \geqq 0. \end{aligned} \tag{1.28}$$

We can now solve (1.28) for θ_x and make use of Lemma 2.3 below to show that (1.26) is the only continuous solution of (1.28) that vanishes at $t = 0$.

We now require that r satisfy (1.13), (1.14), and

$$r \in L^1((0, \infty); L^2(0, 1)); \tag{1.29}$$

in addition, we assume that the following compatibility conditions hold:

$$\theta_0'(0) = \theta_0'(1) = 0, \tag{1.30}$$

$$r(0,0) = r(1,0) = 0, \tag{1.31}$$

$$r_t(0,0) = -a(0)F''(0)\theta_0''(0), \tag{1.32}$$

$$r_t(1,0) = -a(0)F''(0)\theta_0''(1). \tag{1.33}$$

THEOREM 1.2. *Assume that* (1.5)-(1.11) *hold. Then there is a constant* $\delta > 0$ *such that for every* θ_0 *and* r *that satisfy* (1.12)-(1.14), (1.29)-(1.33) *and*

$$\Theta_0 + R_0 + \left(\int_0^\infty \left(\int_0^1 r(x,t)^2 \, dx \right)^{1/2} dt \right)^2 \leqq \delta^2 \tag{1.34}$$

the initial value problem (1.1)-(1.3), (1.26) *has a unique solution* $\theta > 0$ *with*

$$\theta, \theta_x, \theta_t, \theta_{xx}, \theta_{xt}, \theta_{tt}, \theta_{xxx}, \theta_{xxt}, \theta_{xtt}, \theta_{ttt} \in C([0,\infty); L^2(0,1)), \tag{1.35}$$

$$\theta_x, \theta_t, \theta_{xx}, \theta_{xt}, \theta_{tt} \in L^\infty((0,\infty); L^2(0,1)) \cap L^2((0,\infty); L^2(0,1)), \tag{1.36}$$

and

$$\theta \in L^\infty((0,\infty); L^2(0,1)). \tag{1.37}$$

Furthermore, as $t \to \infty$, $\theta(\cdot, t)$ *converges to a constant* $\theta^{**} > 0$ *uniformly on* $[0,1]$ *and*

$$\theta_x(\cdot, t), \theta_t(\cdot, t) \to 0 \quad \textit{uniformly on } [0,1]. \tag{1.38}$$

Remark 1.4. The value of θ^{**} can be determined from (1.1) as follows. If the assumptions of Theorem 1.2 hold and θ satisfies (1.1)-(1.3), (1.26) then integrating (1.1) over $[0,1]\times[0,t]$, $t > 0$, and passing to the limit as $t \to \infty$ yields

$$\hat{e}(\theta^{**}) = \int_0^1 \hat{e}(\theta_0(x)) \, dx + \int_0^\infty \int_0^1 r(x,t) \, dx \, dt. \tag{1.39}$$

By (1.6) $\hat{e}$ is strictly monotone and hence there is a unique solution θ^{**} of (1.39).

Let us now consider the problem stated below in which the heat conductor occupies the entire real line:

$$\begin{aligned} &\left(\hat{e}'(\theta(x,t)) + \frac{2}{\theta(x,t)^2} \int_0^\infty a'(s) F(\bar{\theta}_x^t(x,s)) \, ds \right) \theta_t(x,t) \\ &\quad + \int_0^\infty a'(s) F''(\bar{\theta}_x^t(x,s)) \bar{\theta}_{xx}^t(x,s) \, ds \\ &\quad - \frac{2}{\theta(x,t)} \int_0^\infty a'(s) F'(\bar{\theta}_x^t(x,s)) [\theta_x(x,t) - \theta_x(x,t-s)] \, ds = r(x,t), \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad x \in \mathbb{R}, \quad t \geqq 0, \end{aligned} \tag{1.40}$$

$$\theta(x,t) = \theta^*, \qquad x \in \mathbb{R}, \quad t < 0, \tag{1.41}$$

$$\theta(x,0) = \theta_0(x), \qquad x \in \mathbb{R}. \tag{1.42}$$

We assume that

$$\theta_0 - \theta^* \in H^3(\mathbb{R}), \tag{1.43}$$

$$\begin{aligned} r_x, r_t, r_{xt}, r_{tt} \in C([0,\infty); L^2(\mathbb{R})) \cap L^2((0,\infty); L^2(\mathbb{R})) \\ \cap L^\infty((0,\infty); L^2(\mathbb{R})), \end{aligned} \tag{1.44}$$

$$r \in C([0,\infty); L^2(\mathbb{R})) \cap L^1((0,\infty); L^2(\mathbb{R})) \cap L^\infty((0,\infty); L^2(\mathbb{R})), \tag{1.45}$$

$$r(\cdot,0) \in H^2(\mathbb{R}), \qquad r_{ttt} \in L^2((0,\infty); L^2(\mathbb{R})). \tag{1.46}$$

Note that (1.45) implies $r \in L^2((0,\infty); L^2(\mathbb{R}))$. We define

$$\Theta_1 := \int_{-\infty}^{\infty} ([\theta_0(x) - \theta^*]^2 + \theta_0'(x)^2 + \theta_0''(x)^2)\, dx \tag{1.47}$$

and

$$\begin{aligned} R_1 := \sup_{t \geqq 0} \int_{-\infty}^{\infty} (r^2 + r_t^2)(x,t)\, dx + \int_{-\infty}^{\infty} r_x^2(x,0)\, dx + \left(\sup_{\substack{x \in \mathbb{R} \\ t \geqq 0}} |r(x,t)| \right)^2 \\ + \int_0^{\infty} \int_{-\infty}^{\infty} (r^2 + r_t^2 + r_{tt}^2)(x,t)\, dx\, dt \\ + \left(\int_0^{\infty} \left(\int_{-\infty}^{\infty} r(x,t)^2\, dx \right)^{1/2} dt \right)^2. \end{aligned} \tag{1.48}$$

THEOREM 1.3. *If* (1.5)-(1.11) *hold, then there is a constant* $\delta > 0$ *such that when* θ_0 *and* r *satisfy* (1.43)-(1.46) *and*

$$\Theta_1 + R_1 \leqq \delta^2, \tag{1.49}$$

the initial value problem (1.40)-(1.42) *has a unique solution* $\theta > 0$ *with*

$$\theta - \theta^*,\ \theta_x,\ \theta_t,\ \theta_{xx},\ \theta_{xt},\ \theta_{tt},\ \theta_{xxx},\ \theta_{xxt},\ \theta_{xtt},\ \theta_{ttt} \in C([0,\infty); L^2(\mathbb{R})), \tag{1.50}$$

$$\theta_x,\ \theta_t,\ \theta_{xx},\ \theta_{xt},\ \theta_{tt} \in L^\infty((0,\infty); L^2(\mathbb{R})) \cap L^2((0,\infty); L^2(\mathbb{R})), \tag{1.51}$$

and

$$\theta - \theta^* \in L^\infty((0,\infty); L^2(\mathbb{R})). \tag{1.52}$$

In addition, as $t \to \infty$,

$$\theta(\cdot,t) \to \theta^* \quad \textit{uniformly on } \mathbb{R} \tag{1.53}$$

and

$$\theta_x(\cdot,t),\ \theta_t(\cdot,t) \to 0 \quad \textit{uniformly on } \mathbb{R} \textit{ and in } L^2(\mathbb{R}). \tag{1.54}$$

Remark 1.5. A detailed proof of Theorem 1.1 is given in § 3. With some minor modifications the argument used to establish Theorem 1.1 can be applied to prove Theorems 1.2 and 1.3; these modifications are discussed in § 3.

Remark 1.6. Assumption (1.11) is not the weakest possible to obtain the global estimates of § 3. However, in order to establish Lemma 2.2 below, the replacement of (1.11) with a weaker assumption would necessitate a more complicated argument than the one used in this paper.

The results established here can be modified and extended, as is illustrated below.

(i) *Solutions with less regularity.* Using a density argument we can show that under weaker assumptions on the data, our initial value problems have a unique, globally defined solution with less regularity than the solutions discussed above. More precisely, for instance in Theorem 1.1, if we replace (1.12)-(1.18) with

$$\theta_0 \in H^2(0,1), \tag{1.55}$$

$$\begin{aligned} r,\ r_t \in C([0,\infty); L^2(0,1)) \cap L^2((0,\infty); L^2(0,1)) \\ \cap L^\infty((0,\infty); L^2(0,1)), \end{aligned} \tag{1.56}$$

$$r \in L^\infty((0,1)\times(0,\infty)), \quad r(\cdot,0) \in H^1(0,1), \quad r_{tt} \in L^2((0,\infty); L^2(0,1)), \tag{1.57}$$

and

$$\theta_0(0) = \theta_0(1) = \theta^*, \qquad r(0,0) = r(1,0) = 0, \tag{1.58}$$

then the result of Theorem 1.1 is true with (1.22) replaced with

$$\theta, \theta_x, \theta_t, \theta_{xx}, \theta_{xt}, \theta_{tt} \in C([0,\infty); L^2(0,1)). \tag{1.59}$$

(ii) *Nonequilibrium history.* Results analogous to Theorems 1.1–1.3 can be obtained if a more general history is prescribed. For example, a result similar to Theorem 1.1 can be established if (1.2) is replaced by

$$\theta(x,t) = \varphi(x,t), \qquad x \in [0,1], \quad t < 0, \tag{1.60}$$

where $\varphi : [0,1] \times (-\infty, 0] \to (0,\infty)$ satisfies

$$\begin{aligned} &\varphi_x, \varphi_{xx}, \varphi_{xt}, \varphi_{xxx}, \varphi_{xxt}, \varphi_{xtt} \in C((-\infty,0]; L^2(0,1)) \\ &\quad \cap L^2((-\infty,0); L^2(0,1)) \cap L^\infty((-\infty,0); L^2(0,1)), \end{aligned} \tag{1.61}$$

$$\int_0^\infty a'(s) F'(\bar{\varphi}_x^0(\cdot, s))\, ds \in H^3(0,1), \tag{1.62}$$

and the compatibility conditions (1.16)–(1.18) are modified accordingly. In addition, the quantity

$$\begin{aligned} \Phi := &\sup_{t \in (-\infty,0)} \int_0^1 (\varphi_x^2 + \varphi_{xx}^2 + \varphi_{xt}^2)(x,t)\, dx \\ &+ \int_{-\infty}^0 \int_0^1 (\varphi_x^2 + \varphi_{xx}^2 + \varphi_{xt}^2)(x,t)\, dx\, dt \\ &+ \int_0^1 \left(\int_0^\infty a'(s) \frac{\partial^2}{\partial x^2} \{F'(\bar{\varphi}_x^0(x,s))\}\, ds \right)^2 dx \end{aligned} \tag{1.63}$$

must be sufficiently small, i.e., condition (1.21) is to be replaced with

$$\Theta_0 + \Phi + R_0 \leqq \delta^2. \tag{1.64}$$

In § 3 we discuss modifications needed in order to adapt the proof of Theorem 1.1 to this case. We note that for the analogue of Theorem 1.2, if we assume that

$$\varphi_x(0,t) = \varphi_x(1,t) = 0, \qquad t \leqq 0 \tag{1.65}$$

then, following the procedure discussed in Remark 1.3, we can show that (1.26) is equivalent to (1.27).

(iii) *General integral models.* These results can be extended to the case where the constitutive equations (0.8) are considered. In these equations the dependence on the summed history of θ is nontrivial; hence a term involving $\theta(x,t)$ appears in the analogue of (1.1). In the corresponding linearized equation the coefficient of $\theta(x,t)$ is

$$E^* := \int_0^\infty \hat{E}_{,3}(s, \theta^*, \theta^* s, 0)\, ds; \tag{1.66}$$

we can show that compatibility with thermodynamics implies that E^* is nonnegative and hence the methods we use here can be adopted to produce analogous results to those stated in Theorems 1.1–1.3. The precise statement of the technical assumptions required would be very complicated and not very illuminating, e.g., the mapping

$$s \mapsto Q_{,4}(s, \theta^*, \theta^* s, 0) \tag{1.67}$$

would have to be such that our assumptions on

$$s \mapsto a'(s)F''(0) \tag{1.68}$$

would hold. We will not discuss this case in further detail.

2. Preliminaries. We begin by stating a local existence result for (1.1)-(1.4). We first note that (1.1) is hyperbolic near equilibrium, but may lose its evolutionary character at states sufficiently far from equilibrium. To ensure that (1.1)-(1.4) is well posed we assume that θ_0 is close to equilibrium in the sense described below. We choose $\varepsilon \in (0, \theta^*)$ sufficiently small so that there are constants e^*, $q^* > 0$ with the following property:

$$\hat{e}'(w(x,t)) + \frac{2}{w(x,t)^2}\int_0^\infty a'(s)F(\bar{w}_x^t(x,s))\,ds \geqq e^* \qquad \forall x \in [0,1], \quad t \in [0,T], \tag{2.1}$$

and

$$-\int_0^\infty a'(s)F''(\bar{w}_x^t(x,s))\,ds \geqq q^* \quad \forall x \in [0,1], \quad t \in [0,T], \tag{2.2}$$

for every $T > 0$ and every $w \in L^\infty((-\infty, T); H^1(0,1))$ satisfying

$$|w(x,t) - \theta^*|, |w_x(x,t)| \leqq \varepsilon \quad \forall x \in [0,1], \quad t \in (-\infty, T]. \tag{2.3}$$

Such a choice is possible by virtue of our assumptions on a and F. (Indeed, the left-hand sides of (2.1) and (2.2) are strictly positive when $w(x,t) \equiv \theta^*$. A simple perturbation about $w = \theta^*$ guarantees the existence of a suitable ε. In fact, we may take $e^* = \frac{1}{2}\hat{e}'(\theta^*)$ and $q^* = \frac{1}{2}a(0)F''(0)$.) We assume that θ_0 satisfies

$$|\theta_0(x) - \theta^*|, |\theta_0'(x)| \leqq \eta \quad \forall x \in [0,1], \tag{2.4}$$

for some $\eta \in (0, \varepsilon)$.

We can now state the following lemma.

LEMMA 2.1. *Assume that* (1.5)-(1.18) *and* (2.4) *are satisfied. Then the initial value problem* (1.1)-(1.4) *has a unique solution* $\theta > 0$, *defined on a maximal time interval* $[0, T_0)$, $T_0 > 0$, *with*

$$\theta, \theta_x, \theta_t, \theta_{xx}, \theta_{xt}, \theta_{tt}, \theta_{xxx}, \theta_{xxt}, \theta_{xtt}, \theta_{ttt} \in C([0, T_0); L^2(0,1)) \tag{2.5}$$

and

$$|\theta(x,t) - \theta^*|, |\theta_x(x,t)| < \varepsilon \quad \forall x \in [0,1], \quad t \in [0, T_0). \tag{2.6}$$

Moreover, if

$$\sup_{\substack{x \in [0,1] \\ t \in [0,T_0)}} |\theta(x,t) - \theta^*|, \sup_{\substack{x \in [0,1] \\ t \in [0,T_0)}} |\theta_x(x,t)| < \varepsilon \tag{2.7}$$

and

$$\sup_{t \in [0,T_0)} \int_0^1 (\theta^2 + \theta_x^2 + \theta_t^2 + \theta_{xx}^2 + \theta_{xt}^2 + \theta_{tt}^2 + \theta_{xxx}^2 + \theta_{xxt}^2 + \theta_{xtt}^2 + \theta_{ttt}^2)(x,t)\,dx < \infty, \tag{2.8}$$

then $T_0 = \infty$.

A result analogous to Lemma 2.1 can be established if we replace (1.4) by (1.26) (i.e., if instead of Dirichlet boundary conditions we consider Neumann boundary

conditions) and (1.15)-(1.18) with (1.29)-(1.33). Similarly to (1.1)-(1.4), the initial value problem (1.1)-(1.3), (1.26) has a unique solution θ defined on a maximal time interval $[0, T_0)$, $T_0 > 0$ satisfying (2.5) and (2.6). We can also obtain a corresponding result for the case when the heat conductor occupies the entire real line; the assumptions required in this case would be the analogues on $\mathbb{R}$ of the assumptions stated above.

The proof of Lemma 2.1 is given in Chapter III of [1][3]; this proof is very technical but standard. Proofs similar in spirit have been used by several authors to obtain local existence results (cf., e.g., [12, Chap. III]) and hence we omit the proof of the lemma. It is interesting to note that although compatibility of the constitutive relations (0.19), (0.22) with thermodynamics determines the form of (1.1) it plays no further role in the proof of Lemma 2.1. However, a bootstrapping argument, in which the thermodynamical restrictions play an essential part, can be applied to strengthen the result described in Lemma 2.1. More precisely, we can show that under the assumptions of Lemma 2.1, if θ satisfies (1.1)-(1.4) on a maximal time interval $[0, T_0)$, $T_0 > 0$ (and hence θ satisfies the entropy inequality (0.2)), then a bound on the $L^\infty([0, T_0); L^2(0, 1))$ norms of θ and its derivatives through order 2 implies that there is a bound on the aforementioned norms of third-order derivatives of θ. Hence, we can establish the following lemma.

LEMMA 2.2. *Suppose that the assumptions of Lemma* 2.1 *hold and that* θ *is a solution of* (1.1)-(1.4) *on a maximal time interval* $[0, T_0)$, $T_0 > 0$. *If* θ *satisfies* (2.7) *and*

$$\sup_{t \in [0, T_0)} \int_0^1 (\theta^2 + \theta_x^2 + \theta_t^2 + \theta_{xx}^2 + \theta_{xt}^2 + \theta_{tt}^2)(x, t)\, dx < \infty, \tag{2.9}$$

then $T_0 = \infty$.

Remark 2.1. If θ is a solution of (1.1)-(1.4), then θ satisfies the entropy inequality (0.2), where the entropy and the heat flux are given by $(0.19)_2$ and $(0.19)_3$, i.e.,

$$\begin{aligned} &-\frac{\partial}{\partial x}\left\{\frac{1}{\theta(x, t)} \int_0^\infty a'(s) F'(\bar{\theta}_x^t(x, s))\, ds\right\} \\ &\leqq \frac{\partial}{\partial t}\left\{-\hat{\psi}'(\theta(x, t)) - \frac{1}{\theta(x, t)^2} \int_0^\infty a'(s) F(\bar{\theta}_x^t(x, s))\, ds\right\} - \frac{r(x, t)}{\theta(x, t)}. \end{aligned} \tag{2.10}$$

Recall that

$$\hat{e}(\nu) := \hat{\psi}(\nu) - \nu \hat{\psi}'(\nu), \qquad \nu > 0, \tag{2.11}$$

hence (1.5) implies

$$\hat{\psi}'' \in C(0, \infty). \tag{2.12}$$

Before proving Lemma 2.2 we introduce the following definition. For $T > 0$ and $0 < h < T$, we define the forward difference operator Δ_h (with respect to the time

[3] Assumption $(1.15)_2$ of Chapter III of [1] does not suffice to ensure that $\theta_t(\cdot, 0+) \in H^2(0, 1)$. We need to make the additional assumption that

$$\int_0^\infty a'(s) F'(\bar{\varphi}_x^0(\cdot, s))\, ds \in H^3(0, 1),$$

where $\varphi : [0, 1] \times [-\infty, 0] \to (0, \infty)$ is a prescribed general history, i.e.,

$$\theta(x, t) = \varphi(x, t), \qquad x \in [0, 1], \quad t < 0.$$

However, the arguments used to prove Theorem 1.1 of [1, Chap. III] remain valid.

variable) by

$$(\Delta_h w)(x,t) := w(x,t+h) - w(x,t), \qquad x \in [0,1], \quad t \in [0, T-h] \tag{2.13}$$

for every $w \in C([0,T]; L^2(0,1))$.

Proof of Lemma 2.2. Let θ be a solution of (1.1)–(1.4) on a maximal time interval $[0, T_0)$, $T_0 > 0$, such that (2.7) holds. Our aim is to show that if $T_0 < \infty$, then

$$\sup_{t \in [0,T_0)} \int_0^1 (\theta^2 + \theta_x^2 + \theta_t^2 + \theta_{xx}^2 + \theta_{xt}^2 + \theta_{tt}^2)(x,t)\, dx = \infty. \tag{2.14}$$

For this purpose it is convenient to introduce the quantities

$$\gamma_2(t) := \sup_{s \in [0,t]} \int_0^1 (\theta^2 + \theta_x^2 + \theta_t^2 + \theta_{xx}^2 + \theta_{xt}^2 + \theta_{tt}^2)(x,s)\, dx, \qquad t \in [0, T_0), \tag{2.15}$$

$$\gamma_3(t) := \sup_{s \in [0,t]} \int_0^1 (\theta^2 + \theta_x^2 + \theta_t^2 + \theta_{xx}^2 + \theta_{xt}^2 + \theta_{tt}^2 + \theta_{xxx}^2 + \theta_{xxt}^2 + \theta_{xtt}^2 + \theta_{ttt}^2)(x,s)\, dx, \qquad t \in [0, T_0), \tag{2.16}$$

$$\Theta := \int_0^1 (\theta_0(x)^2 + \theta_0'(x)^2 + \theta_0''(x)^2 + \theta_0'''(x)^2)\, dx, \tag{2.17}$$

and

$$\begin{aligned} R := {} & \sup_{t \in (0,\infty)} \int_0^1 (r^2 + r_x^2 + r_t^2 + r_{xt}^2 + r_{tt}^2)(x,t)\, dx + \int_0^1 r_{xx}^2(x,0)\, dx \\ & + \int_0^\infty \int_0^1 (r^2 + r_x^2 + r_t^2 + r_{xt}^2 + r_{tt}^2 + r_{ttt}^2)(x,t)\, dx\, dt. \end{aligned} \tag{2.18}$$

In the following calculations we make use of the inequalities

$$\left(\sum_{i=1}^N A_i \right)^2 \leqq N \sum_{i=1}^N A_i^2, \qquad A_1, \cdots, A_N \in \mathbb{R}, \tag{2.19}$$

$$|AB| \leqq \frac{A^2}{4\lambda} + \lambda B^2, \qquad A, B \in \mathbb{R}, \quad \lambda > 0, \tag{2.20}$$

and

$$\| A * B \|_{L^p((0,T);L^2(0,1))} \leqq \| A \|_{L^1(0,\infty)} \| B \|_{L^p((0,T);L^2(0,1))} \tag{2.21}$$

for every $T > 0$, $A \in L^1(0,\infty)$, and $B \in L^p((0,T); L^2(0,1))$, where $1 \leqq p \leqq \infty$ and $A * B$ denotes the convolution of A with B. We use Γ to denote a (possible large) positive generic constant which is independent of θ_0, r, and T_0.

We first differentiate (1.1) twice with respect to t and then apply the forward difference operator Δ_h to the resulting expression. We multiply the new equation by $\Delta_h \theta_{tt}$ and integrate over $[0,1] \times [0,t]$, $t \in (0, T_0)$. After several integrations by parts, we divide both sides by h^2 and let $h \downarrow 0$ to obtain the identity

(2.22)

$$\begin{aligned} & \frac{1}{2} \int_0^1 \left(\hat{e}'(\theta(x,t)) + \frac{2}{\theta(x,t)^2} \int_0^\infty a'(s) F(\bar{\theta}_x^t(x,s))\, ds \right) \theta_{ttt}^2(x,t)\, dx \\ & \quad - \frac{1}{2} \int_0^1 \int_0^\infty a'(s) F''(\bar{\theta}_x^t(x,s))\, ds\, \theta_{xtt}^2(x,t)\, dx \end{aligned}$$

$$
\begin{aligned}
&= -\int_0^t \int_0^1 \frac{\partial}{\partial x}\left\{\frac{1}{\theta(x,s)}\int_0^\infty a'(z)F'(\bar\theta_x^s(x,z))\,dz\right\}\theta_{ttt}^2(x,s)\,dx\,ds\\
&\quad+\frac{1}{2}\int_0^1 \hat e'(\theta_0(x))\theta_{ttt}^2(x,0)\,dx+\frac{1}{2}a(0)F''(0)\int_0^1\theta_{xtt}^2(x,0)\,dx\\
&\quad+\frac{1}{2}\int_0^t\int_0^1\frac{\partial}{\partial s}\left\{\hat e'(\theta(x,s))+\frac{2}{\theta(x,s)^2}\int_0^\infty a'(z)F(\bar\theta_x^s(x,z))\,dz\right\}\theta_{ttt}^2(x,s)\,dx\,ds\\
&\quad-\frac{1}{2}\int_0^t\int_0^1\frac{\partial}{\partial s}\left\{\int_0^\infty a'(z)F''(\bar\theta_x^s(x,z))\,dz\right\}\theta_{xtt}^2(x,s)\,dx\,ds\\
&\quad+\int_0^t\int_0^1\frac{\partial^3}{\partial s^3}\left\{r(x,s)-\frac{2}{\theta(x,s)}\int_0^s a'(z)F'(\bar\theta_x^s(x,z))\theta_x(x,s-z)\,dz\right\}\theta_{ttt}(x,s)\,dx\,ds\\
&\quad-\int_0^t\int_0^1\theta_t(x,s)\frac{\partial^3}{\partial s^3}\left\{\hat e'(\theta(x,s))+\frac{2}{\theta(x,s)^2}\right.\\
&\qquad\qquad\left.\cdot\int_0^\infty a'(z)F(\bar\theta_x^s(x,z))\,dz\right\}\theta_{ttt}(x,s)\,dx\,ds\\
&\quad-3\int_0^t\int_0^1\frac{\partial}{\partial s}\left\{\theta_{tt}(x,s)\frac{\partial}{\partial s}\left\{\hat e'(\theta(x,s))+\frac{2}{\theta(x,s)^2}\right.\right.\\
&\qquad\qquad\left.\left.\cdot\int_0^\infty a'(z)F(\bar\theta_x^s(x,z))\,dz\right\}\right\}\theta_{ttt}(x,s)\,dx\,ds\\
&\quad+\int_0^t\int_0^1\frac{\partial^2}{\partial s^2}\left\{\int_0^s a'(z)F''(\bar\theta_x^s(x,z))\theta_{xx}(x,s-z)\,dz\right\}\theta_{ttt}(x,s)\,dx\,ds\\
&\quad+\int_0^t\int_0^1\frac{\partial^2}{\partial s^2}\left\{\int_0^s a'(z)F'''(\bar\theta_x^s(x,z))\bar\theta_{xx}^s(x,z)\theta_x(x,s-z)\,dz\right\}\theta_{ttt}(x,s)\,dx\,ds\\
&\quad-\int_0^t\int_0^1\theta_x(x,s)\frac{\partial^2}{\partial s^2}\left\{\int_0^s a'(z)F'''(\bar\theta_x^s(x,z))\bar\theta_{xx}^s(x,z)\,dz\right\}\theta_{ttt}(x,s)\,dx\,ds\\
&\quad-2\int_0^t\int_0^1\theta_{xt}(x,s)\frac{\partial}{\partial s}\left\{\int_0^s a'(z)F'''(\bar\theta_x^s(x,z))\bar\theta_{xx}^s(x,z)\,dz\right\}\theta_{ttt}(x,s)\,dx\,ds\\
&\quad+\int_0^t\int_0^1\theta_x(x,s)\frac{\partial^3}{\partial s^3}\left\{\frac{2}{\theta(x,s)}\int_0^\infty a'(z)F'(\bar\theta_x^s(x,z))\,dz\right\}\theta_{ttt}(x,s)\,dx\,ds\\
&\quad+3\int_0^t\int_0^1\frac{\partial}{\partial s}\left\{\theta_{xt}(x,s)\frac{\partial}{\partial s}\left\{\frac{2}{\theta(x,s)}\int_0^\infty a'(z)F'(\bar\theta_x^s(x,z))\,dz\right\}\right\}\theta_{ttt}(x,s)\,dx\,ds\\
&\quad-\int_0^t\int_0^1\frac{\partial^2}{\partial s^2}\left\{\int_0^\infty a'(z)F''(\bar\theta_x^s(x,z))\,dz\right\}\theta_{xx}(x,s)\theta_{ttt}(x,s)\,dx\,ds\\
&\quad-2\int_0^t\int_0^1\frac{\partial}{\partial s}\left\{\int_0^\infty a'(z)F''(\bar\theta_x^s(x,z))\,dz\right\}\theta_{xxt}(x,s)\theta_{ttt}(x,s)\,dx\,ds,\qquad t\in[0,T_0).
\end{aligned}
$$

Here $\theta_{ttt}(\cdot,0)$ and $\theta_{xtt}(\cdot,0)$ are determined from (1.1). Only a partial description of the estimates involved in obtaining (2.23), (2.25), and (2.29) (from (2.22)) will be given since the calculations are similar in spirit to those described in § 3 below. Making use of (2.1), (2.2), (2.10) (to estimate the first term on the right-hand side of (2.22)),

(2.19)-(2.21), and the Sobolev embedding theorem we can show that

$$\begin{aligned}&\int_0^1 (\theta_{xtt}^2+\theta_{ttt}^2)(x,t)\,dx\\ &\leqq \Gamma\Big\{\Theta+R+\gamma_2^2(t)+\gamma_2^3(t)+[1+R^{1/2}+(1+t)(\gamma_2^{1/2}(t)+\gamma_2^{(k+1)/2}(t))]\\ &\qquad\cdot\int_0^t\int_0^1 (\theta_{xxx}^2+\theta_{xxt}^2+\theta_{xtt}^2+\theta_{ttt}^2)(x,s)\,dx\,ds\Big\}\quad \forall t\in[0,T_0).\end{aligned}\tag{2.23}$$

We differentiate (1.1) twice with respect to t, square the resulting expression, and then integrate over $[0,1]$ to obtain the inequality

$$\begin{aligned}&\int_0^1\left(\int_0^\infty a'(s)F''(\bar\theta_x^t(x,s))\,ds\right)^2\theta_{xxt}^2(x,t)\,dx\\ &\leqq 5\int_0^1\left(\frac{\partial^3}{\partial t^3}\left\{\hat e(\theta(x,t))-\frac{2}{\theta(x,t)}\int_0^\infty a'(s)F(\bar\theta_x^t(x,s))\,ds\right\}\right)^2dx\\ &\quad+5\int_0^1\left(\theta_{xx}(x,t)\frac{\partial}{\partial t}\left\{\int_0^\infty a'(s)F''(\bar\theta_x^t(x,s))\,ds\right\}\right)^2dx\\ &\quad+5\int_0^1\left(\frac{\partial}{\partial t}\left\{\int_0^t a'(s)F''(\bar\theta_x^t(x,s))\theta_{xx}(x,t-s)\,ds\right\}\right)^2dx\\ &\quad+5\int_0^1\left(\frac{\partial}{\partial t}\left\{\int_0^t a'(s)\frac{\partial}{\partial t}(F''(\bar\theta_x^t(x,s)))\bar\theta_{xx}^t(x,s)\,ds\right\}\right)^2dx\\ &\quad+5\int_0^1 r_{tt}^2(x,t)\,dx\quad\forall t\in[0,T_0).\end{aligned}\tag{2.24}$$

Using (2.2) to get a lower bound on the left-hand side of (2.24), and estimating the right-hand side of (2.24) we can show that

$$\begin{aligned}&\int_0^1\theta_{xxt}^2(x,t)\,dx\leqq\Gamma\Big\{R+\gamma_2(t)+\gamma_2^{k+3}(t)\\ &\qquad+(1+\gamma_2(t)+\gamma_2^k(t))\\ &\qquad\cdot\int_0^1(\theta_{xtt}^2+\theta_{ttt}^2)(x,t)\,dx\Big\}\quad\forall t\in[0,T_0).\end{aligned}\tag{2.25}$$

We now differentiate (1.1) once with respect to t and then once with respect to x. We square the result and integrate over $[0,1]$ to get

$$\begin{aligned}&\int_0^1\left(\int_0^\infty a'(s)F''(\bar\theta_x^t(x,s))\,ds\right)^2\theta_{xxx}^2(x,t)\,dx\\ &\leqq 5\int_0^1\left(\frac{\partial^3}{\partial x\,\partial t^2}\left\{\hat e(\theta(x,t))-\frac{2}{\theta(x,t)}\int_0^\infty a'(s)F(\bar\theta_x^t(x,s))\,ds\right\}\right)^2dx\\ &\quad+5\int_0^1\left(\theta_{xx}(x,t)\frac{\partial}{\partial x}\left\{\int_0^\infty a'(s)F''(\bar\theta_x^t(x,s))\,ds\right\}\right)^2dx\\ &\quad+5\int_0^1\left(\frac{\partial}{\partial x}\left\{\int_0^t a'(s)F''(\bar\theta_x^t(x,s))\theta_{xx}(x,t-s)\,ds\right\}\right)^2dx\\ &\quad+5\int_0^1\left(\frac{\partial}{\partial x}\left\{\int_0^t a'(s)\frac{\partial}{\partial t}\{F''(\bar\theta_x^t(x,s))\}\bar\theta_{xx}^t(x,s)\,ds\right\}\right)^2dx\\ &\quad+5\int_0^1 r_{xt}^2(x,t)\,dx\quad\forall t\in[0,T_0).\end{aligned}\tag{2.26}$$

To obtain bounds on the third and fourth terms on the right-hand side of (2.26) we make use of the following observation. We have

$$g(x,t)=g(x,0)+\int_0^t g_t(x,s)\,ds, \qquad x\in[0,1], \quad t\in[0,T] \tag{2.27}$$

and hence

$$\int_0^1 g^2(x,t)\,dx \leqq 2\int_0^1 g^2(x,0)\,dx+2t\int_0^t\int_0^1 g_t^2(x,s)\,dx\,ds \quad \forall t\in[0,T] \tag{2.28}$$

for every $T>0$ and every smooth function $g:[0,1]\times[0,T]\to\mathbb{R}$. Thus we arrive at the inequality

$$\begin{aligned}\int_0^1 \theta_{xxx}^2(x,t)\,dx \leqq \Gamma\Big\{ & R+\gamma_2(t)+\gamma_2^{k+3}(t)\\ & +(1+\gamma_2(t)+\gamma_2^k(t))\int_0^1(\theta_{xxt}^2+\theta_{xtt}^2)(x,t)\,dx\\ & +t(1+\gamma_2(t)+\gamma_2^{k+2}(t))\int_0^t\int_0^1\theta_{xxx}^2(x,s)\,dx\,ds\Big\} \quad \forall t\in[0,T_0).\end{aligned} \tag{2.29}$$

As remarked earlier, we will not give further details of the calculations involved to obtain the above estimates.

Combining (2.23), (2.25), and (2.29) it is easy to show that there is a constant $N=N(k)>2$ such that

$$\begin{aligned}&\int_0^1(\theta_{xxx}^2+\theta_{xxt}^2+\theta_{xtt}^2+\theta_{ttt}^2)(x,t)\,dx\\ &\quad\leqq\bar\Gamma\Big\{\Theta+\Theta^2+R+R^2+\gamma_2(T_0)+\gamma_2^N(T_0)\\ &\qquad+[1+R^{1/2}+R+(1+T_0)(\gamma_2^{1/2}(T_0)+\gamma_2^N(T_0))]\\ &\qquad\cdot\int_0^t\int_0^1(\theta_{xxx}^2+\theta_{xxt}^2+\theta_{xtt}^2+\theta_{ttt}^2)(x,s)\,dx\,ds\Big\} \quad \forall t\in[0,T_0),\end{aligned} \tag{2.30}$$

where $\bar\Gamma$ is a fixed positive constant independent of r, θ_0, and T_0. Thus Gronwall's inequality implies

$$\begin{aligned}\gamma_3(T_0)\leqq\bar\Gamma[&\Theta+\Theta^2+R+R^2+\gamma_2(T_0)+\gamma_2^N(T_0)]\\ &\cdot\exp\{\bar\Gamma T_0[1+R^{1/2}+R+(1+T_0)(\gamma_2^{1/2}(T_0)+\gamma_2^N(T_0))]\}.\end{aligned} \tag{2.31}$$

According to Lemma 2.1, if $T_0<\infty$ then $\gamma_3(T_0)=\infty$ and hence (2.31) leads to the desired conclusion. □

In the analysis of (1.1)–(1.4) we make essential use of several properties of strongly positive-definite kernels. A function $b\in L^1_{\mathrm{loc}}[0,\infty)$ is said to be *positive definite* if

$$\int_0^t w(s)\int_0^s b(s-z)w(z)\,dz\,ds\geqq 0 \quad \forall t\geqq 0, \tag{2.32}$$

for every $w\in C[0,\infty)$. The kernel b is said to be *strongly positive definite* if there is a constant $c>0$ such that the mapping $t\mapsto b(t)-c\,e^{-t}$ is positive definite.

This definition is generally not easy to check directly. We can show that if $b \in L^1(0, \infty)$, then b is strongly positive definite if and only if there is a constant $c > 0$ such that

$$\text{Re } \mathscr{L}[b](i\omega) \geqq \frac{c}{\omega^2 + 1} \quad \forall \omega \in \mathbb{R}, \tag{2.33}$$

where $\mathscr{L}[\cdot]$ denotes the Laplace transform. It is useful to know that if $b \in C^2[0, \infty)$ and

$$(-1)^j b^{(j)}(t) \geqq 0 \quad \forall t \geqq 0, \quad j = 0, 1, 2, \quad b' \not\equiv 0, \tag{2.34}$$

then b is strongly positive definite. With sufficient regularity, we can obtain information concerning the pointwise behaviour near zero of strongly positive-definite functions. In particular, $(1.10)_{1,2}$ imply that

$$a(0) > 0, \qquad a'(0) < 0. \tag{2.35}$$

This follows easily by expressing $a(0)$ and $a'(0)$ in terms of the Laplace transform of a (cf., e.g., [7, § 2]). Condition (2.35) plays an important role in the analysis. See, for example, [11] for more information on strongly positive-definite kernels.

To obtain certain estimates, we need to solve (1.1) for θ_{xx}. For this purpose we recall that for each $y \in L^1_{\text{loc}}[0, \infty)$, the equation

$$a(0)w(t) + \int_0^t a'(t-s)w(s)\, ds = y(t), \qquad t \geqq 0, \tag{2.36}$$

has a unique solution $w \in L^1_{\text{loc}}[0, \infty)$; this solution is given by

$$w(t) = \frac{1}{a(0)}\left(y(t) + \int_0^t m(t-s)y(s)\, ds\right), \qquad t \geqq 0, \tag{2.37}$$

where m, the *resolvent kernel of* a', is defined to be the unique solution of the resolvent equation

$$a(0)m(t) + \int_0^t m(t-s)a'(s)\, ds = -a'(t), \qquad t \geqq 0. \tag{2.38}$$

Using a Paley-Wiener type argument, $(1.10)_{1,2}$, and properties of strongly positive kernels, we establish the following lemma.

LEMMA 2.3. *Assume that* $(1.10)_{1,2}$ *is satisfied. Then the solution* m *to* (2.38) *satisfies* $m' \in L^1(0, \infty)$.

Remark 2.2. Under assumptions $(1.10)_{1,2}$ and $(1.11)_1$ we can also show that

$$m(t) = \frac{a(0)}{\mathscr{L}[a](0)} + M(t), \qquad t \geqq 0, \tag{2.39}$$

where $M \in L^1(0, \infty)$.

Proof of Lemma 2.3. Define $\Pi := \{\xi \in \mathbb{C}: \text{Re } \xi \geqq 0\}$. Formally taking Laplace transforms in (2.38) we find that

$$\mathscr{L}[m](\xi) = \frac{-\mathscr{L}[a'](\xi)}{a(0) + \mathscr{L}[a'](\xi)}, \qquad \xi \in \Pi. \tag{2.40}$$

Recall that $(1.10)_{1,2}$ imply (2.35). Thus, by (2.38) we have

$$\mathscr{L}[m'](\xi) = \frac{a'(0)}{a(0)} - \frac{\xi \mathscr{L}[a'](\xi)}{a(0) + \mathscr{L}[a'](\xi)}, \qquad \xi \in \Pi. \tag{2.41}$$

After a simple computation we obtain

$$\mathscr{L}[m'](\xi)=\frac{a'(0)}{a(0)}+\frac{a(0)}{\mathscr{L}[a](\xi)}-\xi, \qquad \xi\in\Pi. \tag{2.42}$$

By (2.33) and the maximum principle for analytic functions $\mathscr{L}[a]$ does not vanish on Π. Hence, by $(2.35)_1$ and (2.33), $\mathscr{L}[m']$ is locally analytic on Π in the sense of Definition 2.1 of [9]. Observe that for ξ near infinity we have

$$\begin{aligned}\mathscr{L}[m'](\xi)&=\frac{-a(0)(\xi\mathscr{L}[a'](\xi)-a'(0))+a'(0)\mathscr{L}[a'](\xi)}{a(0)(a(0)+\mathscr{L}[a'](\xi))}\\&=\frac{-a(0)\mathscr{L}[a''](\xi)+a'(0)\mathscr{L}[a'](\xi)}{a(0)(a(0)+\mathscr{L}[a'](\xi))}.\end{aligned} \tag{2.43}$$

Thus $\mathscr{L}[m']$ is locally analytic at infinity and $\mathscr{L}[m'](\infty)=0$. Therefore, by Proposition 2.3 of [9] $m'\in L^1(0,\infty)$, and the proof is complete. □

Before describing our next result we introduce the following notation (which is also used in the next section). For $b\in L^1_{\mathrm{loc}}[0,\infty)$ we define

$$Q(w,t,b):=\int_0^t\int_0^1 w(x,s)\int_0^s b(s-z)w(x,z)\,dz\,dx\,ds, \qquad t\in[0,T], \tag{2.44}$$

for every $T>0$ and every $w\in C([0,T];L^2(0,1))$. The result below was motivated by Lemma 2 of [8].

LEMMA 2.4. *Assume that* $(1.10)_{1,2}$ *hold. Then there exists a constant* $L>0$ *such that*

$$\begin{aligned}\int_0^1 w^2(x,t)\,dx\leqq L\int_0^1 w^2(x,0)\,dx+L\int_0^t\int_0^1 w^2(x,s)\,dx\,ds\\+L\liminf_{h\downarrow 0}\frac{1}{h^2}Q(\Delta_h w,t,a)\quad\forall t\in[0,T],\end{aligned} \tag{2.45}$$

for every $T>0$ *and every* $w\in C([0,T];L^2(0,1))$ *and consequently, by Lemma* 2.5 *of* [7], *there is a constant* $L^*>0$ *such that*

$$\begin{aligned}\int_0^1 w^2(x,t)\,dx+\int_0^t\int_0^1 w^2(x,s)\,dx\,ds\leqq L^*\int_0^1 w^2(x,0)\,dx+L^*Q(w,t,a)\\+L^*\liminf_{h\downarrow 0}\frac{1}{h^2}Q(\Delta_h w,t,a)\quad\forall t\in[0,T],\end{aligned} \tag{2.46}$$

for every $T>0$ *and every* $w\in C([0,T];L^2(0,1))$.

For the proof of Lemma 2.4 it is convenient to introduce the following notation:

$$e(t):=e^{-t}, \qquad t\in[0,\infty). \tag{2.47}$$

In addition, for $T>0$ and $0<h<T$, we define the quantity

$$(D_h w)(x,t):=\int_0^t \Delta_h w(x,s)\,ds, \qquad t\in[0,T-h], \tag{2.48}$$

for every $w\in C([0,T];L^2(0,1))$. We note that

$$(D_h w)(x,t)=\int_t^{t+h} w(x,s)\,ds-\int_0^h w(x,s)\,ds, \qquad t\in[0,T-h]. \tag{2.49}$$

Proof of Lemma 2.4. We first observe that by $(1.10)_{1,2}$ there exists a constant $c>0$ such that

$$0 \leqq Q(v, t, e) \leqq cQ(v, t, a) \quad \forall t \in [0, T], \tag{2.50}$$

for every $T>0$ and every $v \in C([0, T]; L^2(0, 1))$. Let $T>0$, $h \in (0, T)$, and $w \in C([0, T]; L^2(0, 1))$ be given. Integration by parts (twice) leads to the following identity:

$$\begin{aligned} Q(\Delta_h w, t, e) = &\frac{1}{2}\int_0^1 (D_h w)(x, t)^2\, dx + \int_0^t \int_0^1 (D_h w)(x, s)^2\, dx\, ds \\ &- \int_0^1 (D_h w)(x, t) \int_0^t e^{-(t-s)} (D_h w)(x, s)\, ds\, dx \\ &- \int_0^t \int_0^1 (D_h w)(x, s) \int_0^s e^{-(s-z)} (D_h w)(x, z)\, dz\, dx\, ds. \end{aligned} \tag{2.51}$$

Dividing both sides of (2.51) by h^2 and letting $h \downarrow 0$ we can show that $\lim_{h \downarrow 0} (1/h^2) Q(\Delta_h w, t, e)$ exists and is given by

$$\begin{aligned} \lim_{h \downarrow 0} \frac{1}{h^2} Q(\Delta_h w, t, e) = &\frac{1}{2}\int_0^1 [w(x, t) - w(x, 0)]^2\, dx + \int_0^t \int_0^1 [w(x, s) - w(x, 0)]^2\, dx\, ds \\ &- \int_0^1 [w(x, t) - w(x, 0)] \int_0^t e^{-(t-s)} [w(x, s) - w(x, 0)]\, ds\, dx \\ &- \int_0^t \int_0^1 [w(x, s) - w(x, 0)] \int_0^s e^{-(s-z)} [w(x, z) - w(x, 0)]\, dz\, dx\, ds. \end{aligned} \tag{2.52}$$

After some simple computations we obtain the following expression for the last two terms on the right-hand side of (2.52):

$$\begin{aligned} &- \int_0^1 [w(x, t) - w(x, 0)] \int_0^t e^{-(t-s)} [w(x, s) - w(x, 0)]\, ds\, dx \\ &\quad = -\int_0^1 w(x, t) \int_0^t e^{-(t-s)} w(x, s)\, ds\, dx \\ &\qquad + \int_0^1 w(x, 0) \int_0^t e^{-(t-s)} w(x, s)\, ds\, dx \\ &\qquad - \int_0^1 w^2(x, 0)[1 - e^{-t}]\, dx + \int_0^1 w(x, t) w(x, 0)[1 - e^{-t}]\, dx, \end{aligned} \tag{2.53}$$

$$\begin{aligned} &- \int_0^t \int_0^1 [w(x, s) - w(x, 0)] \int_0^s e^{-(s-z)} [w(x, z) - w(x, 0)]\, dz\, dx\, ds \\ &\quad = -Q(w, t, e) - \int_0^t \int_0^1 w^2(x, 0)\, dx\, ds + \int_0^1 w^2(x, 0)[1 - e^{-t}]\, dx \\ &\qquad + 2\int_0^t \int_0^1 w(x, s) w(x, 0)\, dx\, ds - \int_0^t \int_0^1 w(x, s) w(x, 0)\, e^{-s}\, dx\, ds \\ &\qquad - \int_0^t \int_0^1 w(x, s) w(x, 0)\, e^{-(t-s)}\, dx\, ds. \end{aligned} \tag{2.54}$$

Hence, (2.52) implies

$$\begin{aligned}\frac{1}{2}\int_0^1 w^2(x,t)\,dx &= \lim_{h\downarrow 0}\frac{1}{h^2}Q(\Delta_h w,t,e)+Q(w,t,e)-\int_0^t\int_0^1 w^2(x,s)\,dx\,ds\\ &\quad -\frac{1}{2}\int_0^1 w^2(x,0)\,dx+\int_0^1 w(x,t)\int_0^t e^{-(t-s)}w(x,s)\,ds\,dx\\ &\quad +\int_0^1 w(x,t)w(x,0)\,e^{-t}\,dx\\ &\quad +\int_0^t\int_0^1 w(x,s)w(x,0)\,e^{-s}\,dx\,ds.\end{aligned}\tag{2.55}$$

To complete the proof we use the inequality (2.20); for $\lambda>0$

$$\begin{aligned}&\left|\int_0^1 w(x,t)\int_0^t e^{-(t-s)}w(x,s)\,ds\,dx\right|\\ &\leqq \lambda\int_0^1 w^2(x,t)\,dx+\frac{1}{4\lambda}\int_0^1\left(\int_0^t e^{-(t-s)}w(x,s)\,ds\right)^2 dx\\ &\leqq \lambda\int_0^1 w^2(x,t)\,dx+\frac{1}{8\lambda}\int_0^t\int_0^1 w^2(x,s)\,dx\,ds,\end{aligned}\tag{2.56}$$

$$\left|\int_0^1 w(x,t)w(x,0)\,e^{-t}\,dx\right|\leqq\lambda\int_0^1 w^2(x,t)\,dx+\frac{1}{4\lambda}\int_0^1 w^2(x,0)\,dx,\tag{2.57}$$

and similarly

$$\begin{aligned}&\left|\int_0^t\int_0^1 w(x,s)w(x,0)\,e^{-s}\,dx\,ds\right|\\ &\leqq\int_0^t\int_0^1 w^2(x,s)\,dx\,ds+\frac{1}{4}\int_0^t e^{-2s}\,ds\int_0^1 w^2(x,0)\,dx\\ &\leqq\int_0^t\int_0^1 w^2(x,s)\,dx\,ds+\frac{1}{8}\int_0^1 w^2(x,0)\,dx.\end{aligned}\tag{2.58}$$

Hence, if $\lambda>0$ is chosen to be sufficiently small the desired conclusion follows from (2.50). □

3. Proof of Theorem 1.1. We choose $\varepsilon\in(0,\theta^*)$ as in the first paragraph of § 2. If (1.21) holds with $\delta<\eta/2$, for some $\eta\in(0,\varepsilon)$, then the Sobolev embedding theorem implies

$$|\theta_0(x)-\theta^*|,\ |\theta_0'(x)|\leqq\sqrt{2\Theta_0}<\eta\quad\forall x\in[0,1].\tag{3.1}$$

Therefore, by Lemmas 2.1 and 2.2, the initial value problem (1.1)-(1.4) has a unique solution $\theta>0$ that satisfies

$$\theta,\ \theta_x,\ \theta_t,\ \theta_{xx},\ \theta_{xt},\ \theta_{tt},\ \theta_{xxx},\ \theta_{xxt},\ \theta_{xtt},\ \theta_{ttt}\in C([0,T_0);L^2(0,1))\tag{3.2}$$

and

$$|\theta(x,t)-\theta^*|,\ |\theta_x(x,t)|<\varepsilon\quad\forall x\in[0,1],\quad t\in[0,T_0)\tag{3.3}$$

on a maximal time interval $[0, T_0)$, $T_0 > 0$. Our aim is to show that if (1.21) holds for $\delta > 0$ sufficiently small, then

$$\sup_{t\in[0,T_0)} \int_0^1 ([\theta(x,t)-\theta^*]^2+\theta_x^2(x,t)+\theta_t^2(x,t)+\theta_{xx}^2(x,t) + \theta_{xt}^2(x,t)+\theta_{tt}^2(x,t))\,dx < \infty \tag{3.4}$$

and

$$\sup_{\substack{x\in[0,1]\\ t\in[0,T_0)}} |\theta(x,t)-\theta^*|, \qquad \sup_{\substack{x\in[0,1]\\ t\in[0,T_0)}} |\theta_x(x,t)| < \varepsilon \tag{3.5}$$

and hence $T_0 = \infty$ (by Lemma 2.2). For this purpose it is convenient to introduce the quantities

$$\begin{aligned}\mathscr{E}(t) := \sup_{s\in[0,t]} \int_0^1 &([\theta(x,s)-\theta^*]^2+\theta_x^2(x,s)+\theta_t^2(x,s)+\theta_{xx}^2(x,s) \\ &+\theta_{xt}^2(x,s)+\theta_{tt}^2(x,s))\,dx \\ +\int_0^t \int_0^1 &([\theta(x,s)-\theta^*]^2+\theta_x^2(x,s)+\theta_t^2(x,s)+\theta_{xx}^2(x,s) \\ &+\theta_{xt}^2(x,s)+\theta_{tt}^2(x,s))\,dx\,ds, \qquad t\in[0,T_0)\end{aligned} \tag{3.6}$$

and

$$\begin{aligned}\nu(t) := \sup_{\substack{x\in[0,1]\\ s\in[0,t]}} &([\theta(x,s)-\theta^*]^2+\theta_x^2(x,s)+\theta_t^2(x,s))^{1/2} \\ &+\left(\int_0^t \left(\sup_{x\in[0,1]} |\theta_x(x,s)|\right)^2 ds\right)^{1/2}, \qquad t\in[0,T_0).\end{aligned} \tag{3.7}$$

Equation (1.1) can be rewritten as follows:

$$\begin{aligned}&\hat{e}'(\theta^*)\theta_t(x,t) - F''(0)\int_0^t a(t-s)\theta_{xx}(x,s)\,ds \\ &= -[\hat{e}'(\theta(x,t))-\hat{e}'(\theta^*)]\theta_t(x,t) - \int_0^t \theta_{xx}(x,s) \\ &\quad \cdot \int_{t-s}^\infty a'(z)[F''(\bar{\theta}_x^t(x,z))-F''(0)]\,dz\,ds \\ &\quad -\frac{2}{\theta(x,t)^2}\theta_t(x,t)\int_0^\infty a'(s)F(\bar{\theta}_x^t(x,s))\,ds \\ &\quad +\frac{2}{\theta(x,t)}\theta_x(x,t)\int_0^\infty a'(s)F'(\bar{\theta}_x^t(x,s))\,ds \\ &\quad -\frac{2}{\theta(x,t)}\int_0^t a'(s)F'(\bar{\theta}_x^t(x,s))\theta_x(x,t-s)\,ds + r(x,t), \\ &\qquad\qquad x\in[0,1], \quad t\in[0,T_0).\end{aligned} \tag{3.8}$$

In the derivation of this equation from (1.1) we make use of (1.2) and (1.8). The second terms on both sides of (3.8) are obtained through the following computation:

$$\int_0^\infty a'(s)F''(\bar\theta_x^t(x,s))\bar\theta_{xx}^t(x,s)\,ds = \int_0^\infty \int_{t-s}^t a'(s)F''(\bar\theta_x^t(x,s))\theta_{xx}(x,z)\,dz\,ds$$

$$= \int_0^t \theta_{xx}(x,z)\int_{t-z}^\infty a'(s)F''(\bar\theta_x^t(x,s))\,ds\,dz. \tag{3.9}$$

The aim of the computations that follow is to establish the inequality (3.40) below; to do so we employ energy methods. We use two main types of estimates in this argument:

(i) Estimates derived directly from energy integrals;

(ii) Additional estimates obtained from equation (3.8) through the use of inverse Volterra operators.

In our energy integrals, the left-hand side of (3.8) will lead to positive-definite contributions and the right-hand side will lead to terms that are small provided the solution is near equilibrium. We make essential use of Lemma 2.4 in the estimates of type (i); in addition, in order to estimate the energy integral of highest order we must exploit compatibility of our constitutive relations (0.19), (0.22) with thermodynamics; i.e., we make use of the fact that a solution of (1.1)–(1.4) satisfies the entropy inequality (0.2). (See Remark 2.1 for further details.) Lemma 2.3 plays an important role in the estimates of type (ii). A reader who is unfamiliar with energy methods and seeks further motivation for our computations may wish to look at the argument following (3.40) before reading the derivation of (3.40).

In the numerous estimations that follow we make frequent use of the inequalities (2.19)–(2.21). We use Γ to denote a (possibly large) positive generic constant which is independent of θ_0, r, and T_0.

To obtain our first energy integral we multiply equation (3.8) by $(\theta-\theta^*)$ and integrate over $[0,1]\times[0,t]$, $t\in[0,T_0)$. After integration by parts we find that

$$\frac{1}{2}\hat e'(\theta^*)\int_0^1[\theta(x,t)-\theta^*]^2\,dx + F''(0)Q(\theta_x,t,a)$$

$$= \frac{1}{2}\hat e'(\theta^*)\int_0^1[\theta_0(x)-\theta^*]^2\,dx + \int_0^t\int_0^1[\theta(x,s)-\theta^*]$$

$$\cdot\Bigg(-[\hat e'(\theta(x,s))-\hat e'(\theta^*)]\theta_t(x,s)$$

$$-\int_0^s \theta_{xx}(x,y)\int_{s-y}^\infty a'(z)[F''(\bar\theta_x^s(x,z))-F''(0)]\,dz\,dy$$

$$-\frac{2}{\theta(x,s)^2}\theta_t(x,s)\int_0^\infty a'(z)F(\bar\theta_x^s(x,z))\,dz$$

$$+\frac{2}{\theta(x,s)}\theta_x(x,s)\int_0^\infty a'(z)F'(\bar\theta_x^s(x,z))\,dz - \frac{2}{\theta(x,s)}$$

$$\cdot\int_0^s a'(z)F'(\bar\theta_x^s(x,z))\theta_x(x,s-z)\,dz + r(x,s)\Bigg)\,dx\,ds, \qquad t\in[0,T_0). \tag{3.10}$$

We next differentiate (3.8) with respect to t:

$$\begin{aligned}
&\hat{e}'(\theta^*)\theta_{tt}(x,t) - F''(0)a(0)\theta_{xx}(x,t) - F''(0)\int_0^t a'(t-s)\theta_{xx}(x,s)\,ds \\
&\quad = \frac{\partial}{\partial t}\Big\{ -[\hat{e}'(\theta(x,t)) - \hat{e}'(\theta^*)]\theta_t(x,t) \\
&\qquad - \int_0^t \theta_{xx}(x,s)\int_{t-s}^{\infty} a'(z)[F''(\bar{\theta}_x^t(x,z)) - F''(0)]\,dz\,ds \\
&\qquad - \frac{2}{\theta(x,t)^2}\theta_t(x,t)\int_0^{\infty} a'(s)F(\bar{\theta}_x^t(x,s))\,ds \\
&\qquad + \frac{2}{\theta(x,t)}\theta_x(x,t)\int_0^{\infty} a'(s)F'(\bar{\theta}_x^t(x,s))\,ds \\
&\qquad - \frac{2}{\theta(x,t)}\int_0^t a'(s)F'(\bar{\theta}_x^t(x,s))\theta_x(x,t-s)\,ds + r(x,t)\Big\}, \\
&\qquad\qquad x\in[0,1], \quad t\in[0,T_0).
\end{aligned} \tag{3.11}$$

Multiplying this equation by θ_t and integrating over $[0,1]\times[0,t]$, $t\in[0,T_0)$ we obtain the following expression:

$$\begin{aligned}
&\frac{1}{2}\hat{e}'(\theta^*)\int_0^1 \theta_t^2(x,t)\,dx + F''(0)Q(\theta_{xt},t,a) \\
&\quad = F''(0)\int_0^t\int_0^1 a(s)\theta_0''(x)\theta_t(x,s)\,dx\,ds + \frac{1}{2}\hat{e}'(\theta^*)\int_0^1 \theta_t^2(x,0)\,dx + \int_0^t\int_0^1 \theta_t(x,s)\frac{\partial}{\partial s} \\
&\qquad \cdot\Big\{ -[\hat{e}'(\theta(x,s)) - \hat{e}'(\theta^*)]\theta_t(x,s) \\
&\qquad - \int_0^s \theta_{xx}(x,y)\int_{s-y}^{\infty} a'(z)[F''(\bar{\theta}_x^s(x,z)) - F''(0)]\,dz\,dy \\
&\qquad - \frac{2}{\theta(x,s)^2}\theta_t(x,s)\int_0^{\infty} a'(z)F(\bar{\theta}_x^s(x,z))\,dz \\
&\qquad + \frac{2}{\theta(x,s)}\theta_x(x,s)\int_0^{\infty} a'(z)F'(\bar{\theta}_x^s(x,z))\,dz \\
&\qquad - \frac{2}{\theta(x,s)}\int_0^s a'(z)F'(\bar{\theta}_x^s(x,z))\theta_x(x,s-z)\,dz + r(x,s)\Big\}\,dx\,ds, \qquad t\in[0,T_0).
\end{aligned} \tag{3.12}$$

We note that according to (3.8) we have

$$\theta_t(x,0) = \frac{1}{\hat{e}'(\theta_0(x))}r(x,0), \qquad x\in[0,1]. \tag{3.13}$$

Differentiation of (3.11) with respect to t yields (after integrating several terms by parts)

(3.14)

$$\begin{aligned}
&\hat{e}'(\theta^*)\theta_{ttt}(x,t) - F''(0)a(0)\theta_{xxt}(x,t) - F''(0)\int_0^t a'(t-s)\theta_{xxt}(x,s)\,ds \\
&\quad = F''(0)a'(t)\theta_0''(x)
\end{aligned}$$

$$+\frac{\partial^2}{\partial t^2}\left\{-[\hat{e}'(\theta(x,t))-\hat{e}'(\theta^*)]\theta_t(x,t)-\frac{2}{\theta(x,t)}\right.$$
$$\left.\cdot\int_0^t a'(s)F'(\bar{\theta}_x^t(x,s))\theta_x(x,t-s)\,ds+r(x,t)\right\}$$
$$+\frac{\partial}{\partial t}\left\{\int_0^t \theta_{xx}(x,t-s)a'(s)[F''(\bar{\theta}_x^t(x,s))-F''(0)]\,ds\right.$$
$$\left.+\int_0^t \theta_{xx}(x,s)\int_{t-s}^t a'(z)F'''(\bar{\theta}_x^t(x,z))\theta_x(x,t-z)\,dz\,ds\right\}$$
$$-\theta_{xx}(x,t)\frac{\partial}{\partial t}\left\{\int_0^\infty a'(s)[F''(\bar{\theta}_x^t(x,s))-F''(0)]\,ds\right\}$$
$$-\theta_x(x,t)\frac{\partial}{\partial t}\left\{\int_0^t \theta_{xx}(x,s)\int_{t-s}^\infty a'(z)F'''(\bar{\theta}_x^t(x,z))\,dz\,ds\right\}$$
$$-\frac{\partial}{\partial x}\left\{\theta_{xt}(x,t)\int_0^\infty a'(s)[F''(\bar{\theta}_x^t(x,s))-F''(0)]\,ds\right\}$$
$$-\theta_t(x,t)\frac{\partial^2}{\partial t^2}\left\{\frac{2}{\theta(x,t)^2}\int_0^\infty a'(s)F(\bar{\theta}_x^t(x,s))\,ds\right\}$$
$$-\theta_{tt}(x,t)\frac{\partial}{\partial t}\left\{\frac{4}{\theta(x,t)^2}\int_0^\infty a'(s)F(\bar{\theta}_x^t(x,s))\,ds\right\}$$
$$-\theta_{ttt}(x,s)\frac{2}{\theta(x,t)^2}\int_0^\infty a'(s)F(\bar{\theta}_x^t(x,s))\,ds$$
$$+\theta_x(x,t)\frac{\partial^2}{\partial t^2}\left\{\frac{2}{\theta(x,t)}\int_0^\infty a'(s)F'(\bar{\theta}_x^t(x,s))\,ds\right\}$$
$$+\theta_{xt}(x,t)\frac{\partial}{\partial t}\left\{\frac{4}{\theta(x,t)}\int_0^\infty a'(s)F'(\bar{\theta}_x^t(x,s))\,ds\right\}$$
$$+\theta_{xtt}(x,t)\frac{2}{\theta(x,t)}\int_0^\infty a'(s)F'(\bar{\theta}_x^t(x,s))\,ds,\qquad x\in[0,1],\quad t\in[0,T_0).$$

In analogy with the previous calculation, we multiply (3.14) by θ_{tt} and integrate over $[0,1]\times[0,t]$, $t\in[0,T_0)$. The resulting relation is

(3.15)

$$\frac{1}{2}\hat{e}'(\theta^*)\int_0^1 \theta_{tt}^2(x,t)\,dx+F''(0)Q(\theta_{xtt},t,a)$$
$$=F''(0)\int_0^t\int_0^1 a'(s)\theta_0''(x)\theta_{tt}(x,s)\,dx\,ds-F''(0)\int_0^1 a(t)\theta_{xt}(x,0)\theta_{xt}(x,t)\,dx$$
$$+F''(0)a(0)\int_0^1 \theta_{xt}^2(x,0)\,dx+F''(0)\int_0^t\int_0^1 a'(s)\theta_{xt}(x,0)\theta_{xt}(x,s)\,dx\,ds$$
$$+\frac{1}{2}\hat{e}'(\theta^*)\int_0^1 \theta_{tt}^2(x,0)\,dx+\int_0^t\int_0^1 \theta_{tt}(x,s)\frac{\partial^2}{\partial s^2}$$
$$\cdot\left\{-[\hat{e}'(\theta(x,s))-\hat{e}'(\theta^*)]\theta_t(x,s)\right.$$
$$\left.-\frac{2}{\theta(x,s)}\int_0^s a'(z)F'(\bar{\theta}_x^s(x,z))\theta_x(x,s-z)\,dz+r(x,s)\right\}dx\,ds$$

$$+\int_0^t\int_0^1 \theta_{tt}(x,s)\frac{\partial}{\partial s}\Big\{\int_0^s \theta_{xx}(x,s-z)a'(z)[F''(\bar\theta_x^s(x,z))-F''(0)]\,dz + \int_0^s \theta_{xx}(x,y) \cdot \int_{s-y}^s a'(z)F'''(\bar\theta_x^s(x,z))\theta_x(x,s-z)\,dz\,dy\Big\}\,dx\,ds$$

$$-\int_0^t\int_0^1 \theta_{tt}(x,s)\theta_{xx}(x,s)\frac{\partial}{\partial s}\Big\{\int_0^\infty a'(z)[F''(\bar\theta_x^s(x,z))-F''(0)]\,dz\Big\}\,dx\,ds$$

$$-\int_0^t\int_0^1 \theta_{tt}(x,s)\theta_x(x,s)\frac{\partial}{\partial s}\Big\{\int_0^s \theta_{xx}(x,y)\int_{s-y}^\infty a'(z)F'''(\bar\theta_x^s(x,z))\,dz\,dy\Big\}\,dx\,ds$$

$$+\frac{1}{2}\int_0^1 \theta_{xt}^2(x,t)\int_0^\infty a'(s)[F''(\bar\theta_x^t(x,s))-F''(0)]\,ds\,dx$$

$$-\frac{1}{2}\int_0^t\int_0^1 \theta_{xt}^2(x,s)\frac{\partial}{\partial s}\Big\{\int_0^\infty a'(z)[F''(\bar\theta_x^s(x,z))-F''(0)]\,dz\Big\}\,dx\,ds$$

$$-\int_0^t\int_0^1 \theta_{tt}(x,s)\theta_t(x,s)\frac{\partial^2}{\partial s^2}\Big\{\frac{2}{\theta(x,s)^2}\int_0^\infty a'(z)F(\bar\theta_x^s(x,z))\,dz\Big\}\,dx\,ds$$

$$-\int_0^t\int_0^1 \theta_{tt}^2(x,s)\frac{\partial}{\partial s}\Big\{\frac{4}{\theta(x,s)^2}\int_0^\infty a'(z)F(\bar\theta_x^s(x,z))\,dz\Big\}\,dx\,ds$$

$$-\int_0^1 \frac{1}{\theta(x,t)^2}\theta_{tt}^2(x,t)\int_0^\infty a'(s)F(\bar\theta_x^t(x,s))\,ds\,dx$$

$$+\int_0^t\int_0^1 \theta_{tt}^2(x,s)\frac{\partial}{\partial s}\Big\{\frac{1}{\theta(x,s)^2}\int_0^\infty a'(z)F(\bar\theta_x^s(x,z))\,dz\Big\}\,dx\,ds$$

$$+\int_0^t\int_0^1 \theta_{tt}(x,s)\theta_x(x,s)\frac{\partial^2}{\partial s^2}\Big\{\frac{2}{\theta(x,s)}\int_0^\infty a'(z)F'(\bar\theta_x^s(x,z))\,dz\Big\}\,dx\,ds$$

$$+\int_0^t\int_0^1 \theta_{tt}(x,s)\theta_{xt}(x,s)\frac{\partial}{\partial s}\Big\{\frac{4}{\theta(x,s)}\int_0^\infty a'(z)F'(\bar\theta_x^s(x,z))\,dz\Big\}\,dx\,ds$$

$$-\int_0^t\int_0^1 \theta_{tt}^2(x,s)\frac{\partial}{\partial x}\Big\{\frac{1}{\theta(x,s)}\int_0^\infty a'(z)F'(\bar\theta_x^s(x,z))\,dz\Big\}\,dx\,ds, \qquad t\in[0,T_0).$$

We note that (3.13) implies

$$\theta_{xt}(x,0)=-\frac{\hat e''(\theta_0(x))}{\hat e'(\theta_0(x))}\theta_0'(x)\theta_t(x,0)+\frac{1}{\hat e'(\theta_0(x))}r_x(x,0), \qquad x\in[0,1] \tag{3.16}$$

and from (3.11) we have

$$\theta_{tt}(x,0)=\frac{F''(0)a(0)}{\hat e'(\theta_0(x))}\theta_0''(x)-\frac{\hat e''(\theta_0(x))}{\hat e'(\theta_0(x))}\theta_t^2(x,0) -\frac{2F''(0)a(0)}{\theta_0(x)\hat e'(\theta_0(x))}\theta_0'(x)^2+\frac{1}{\hat e'(\theta_0(x))}r_t(x,0), \qquad x\in[0,1]. \tag{3.17}$$

We add (3.10), (3.12), and (3.15) and make use of Lemma 2.4 to obtain a lower bound on the left-hand side of the resulting identity. We then make some routine estimations to derive the inequality

$$\begin{aligned}&\int_0^1 ([\theta(x,t)-\theta^*]^2+\theta_x^2(x,t)+\theta_t^2(x,t)+\theta_{xt}^2(x,t)+\theta_{tt}^2(x,t))\,dx\\&\quad+\int_0^t\int_0^1 (\theta_x^2(x,s)+\theta_{xt}^2(x,s))\,dx\,ds\\&\leqq\Gamma\{\Theta_0+R_0\}+\Gamma\{\sqrt{\Theta_0}+\sqrt{R_0}\}\sqrt{\mathscr{E}(t)}\\&\quad+\Gamma\sqrt{R_0}\,\mathscr{E}(t)+\Gamma\{\nu(t)+\nu^{k+2}(t)\}\mathscr{E}(t)\quad\forall t\in[0,T_0).\end{aligned}\tag{3.18}$$

To indicate how (3.18) was derived we show detailed estimations of certain typical terms of (3.10), (3.12), and (3.15), as follows. Many of the terms can be estimated in a simple way, for instance,

$$\begin{aligned}&\left|\int_0^t\int_0^1 \hat{e}'''(\theta(x,s))\theta_t^2(x,s)\theta_{tt}(x,s)\,dx\,ds\right|\\&\leqq\sup_{\substack{x\in[0,1]\\s\in[0,t]}}|\hat{e}'''(\theta(x,s))\theta_t(x,s)|\int_0^t\int_0^1|\theta_t(x,s)\theta_{tt}(x,s)|\,dx\,ds\\&\leqq\sup_{|\theta-\theta^*|<\varepsilon}|\hat{e}'''(\theta)|\sup_{\substack{x\in[0,1]\\s\in[0,t]}}|\theta_t(x,s)|\int_0^t\int_0^1|\theta_{tt}(x,s)\theta_t(x,s)|\,dx\,ds\\&\leqq\Gamma\nu(t)\int_0^t\int_0^1(\theta_{tt}^2(x,s)+\theta_t^2(x,s))\,dx\,ds\\&\leqq\Gamma\nu(t)\mathscr{E}(t)\quad\forall t\in[0,T_0)\end{aligned}\tag{3.19}$$

or

$$\begin{aligned}&\left|F''(0)\int_0^t\int_0^1 a'(s)\theta_0''(x)\theta_{tt}(x,s)\,dx\,ds\right|\\&\leqq F''(0)\left(\int_0^t\int_0^1 a'(s)^2\theta_0''(x)^2\,dx\,ds\right)^{1/2}\left(\int_0^t\int_0^1\theta_{tt}^2(x,s)\,dx\,ds\right)^{1/2}\\&\leqq F''(0)\left(\int_0^\infty a'(s)^2\,ds\right)^{1/2}\left(\int_0^1\theta_0''(x)^2\,dx\right)^{1/2}\sqrt{\mathscr{E}(t)}\\&\leqq\Gamma\sqrt{\Theta_0}\sqrt{\mathscr{E}(t)}\quad\forall t\in[0,T_0).\end{aligned}\tag{3.20}$$

Some of the terms must be rewritten carefully before they are estimated: e.g., the term estimated in (3.25) below arises from

$$\int_0^t\int_0^1\theta_{tt}(x,s)\frac{\partial^2}{\partial s^2}\left\{\frac{2}{\theta(x,s)}\int_0^s a'(z)F'(\bar{\theta}_x^s(x,z))\theta_x(x,s-z)\,dz\right\}dx\,ds,\tag{3.21}$$

which appears on the right-hand side of (3.15). We first differentiate the integral appearing in the integrand of (3.21) once with respect to s and then make the following change of variable:

$$\begin{aligned}&\int_0^s a'(z)F'(\bar{\theta}_x^s(x,z))\theta_{xt}(x,s-z)\,dz\\&\quad=\int_0^s a'(s-\zeta)F'(\bar{\theta}_x^s(x,s-\zeta))\theta_{xt}(x,\zeta)\,d\zeta.\end{aligned}\tag{3.22}$$

We next differentiate the right-hand side of (3.22) with respect to s and then repeat the same change of variable to obtain the integral estimated in (3.25). We note that a similar procedure is used when differentiating terms of the form

$$\int_{s-y}^{\infty} a'(z)F'''(\bar{\theta}_x^s(x,z))\,dz \tag{3.23}$$

with respect to s; the change of variable in this case takes the form

$$\int_{s-y}^{\infty} a'(z)F'''(\bar{\theta}_x^s(x,z))\,dz = \int_{-\infty}^{y} a'(s-\zeta)F'''(\bar{\theta}_x^s(x,s-\zeta))\,d\zeta. \tag{3.24}$$

We now continue to show some typical calculations. The computations below are more involved than those used in (3.19) and (3.20): we obtain a bound on a term appearing on the right-hand side of (3.15):

(3.25)

$$\begin{aligned}
&\left|\int_0^t\int_0^1 \theta_{tt}(x,s)\frac{2}{\theta(x,s)}\int_0^s a'(z)F''(\bar{\theta}_x^s(x,z))\theta_x(x,s-z)\theta_{xt}(x,s-z)\,dz\,dx\,ds\right| \\
&\quad\leqq \Gamma \sup_{\substack{x\in[0,1]\\ s\in[0,t]}}|\theta_x(x,s)|\int_0^t\int_0^1|\theta_{tt}(x,s)|\int_0^s|a'(z)|(F''(0) \\
&\qquad +|F''(\bar{\theta}_x^s(x,z))-F''(0)|)|\theta_{xt}(x,s-z)|\,dz\,dx\,ds \\
&\quad\leqq \Gamma\nu(t)\int_0^t\int_0^1|\theta_{tt}(x,s)| \\
&\qquad\cdot\int_0^s|a'(z)|(F''(0)+K(|\bar{\theta}_x^s(x,z)|+|\bar{\theta}_x^s(x,z)|^k))|\theta_{xt}(x,s-z)|\,dz\,dx\,ds \\
&\quad\leqq \Gamma\nu(t)\int_0^t\int_0^1|\theta_{tt}(x,s)|\int_0^s|a'(z)| \\
&\qquad\cdot\left(F''(0)+K\left[\sqrt{z}\left(\int_{s-z}^s\theta_x^2(x,\xi)\,d\xi\right)^{1/2}\right.\right. \\
&\qquad\quad\left.\left.+(\sqrt{z})^k\left(\int_{s-z}^s\theta_x^2(x,\xi)\,d\xi\right)^{k/2}\right]\right)|\theta_{xt}(x,s-z)|\,dz\,dx\,ds \\
&\quad\leqq \Gamma\nu(t)\left(\int_0^t\int_0^1\theta_{tt}^2(x,s)\,dx\,ds\right)^{1/2} \\
&\qquad\cdot\left(\int_0^t\int_0^1\left(\int_0^s|a'(z)|[1+\nu(t)\sqrt{z}\right.\right. \\
&\qquad\quad\left.\left.+\nu^k(t)(\sqrt{z})^k]|\theta_{xt}(x,s-z)|\,dz\right)^2 dx\,ds\right)^{1/2} \\
&\quad\leqq \Gamma\nu(t)\sqrt{\mathscr{E}(t)}\left(\int_0^t\int_0^1\theta_{xt}^2(x,s)\,dx\,ds\right)^{1/2}
\end{aligned}$$

$$\cdot\left(\int_0^\infty |a'(z)|\,dz+\nu(t)\int_0^\infty |a'(z)|\sqrt{z}\,dz\right.$$

$$\left.+\nu^k(t)\int_0^\infty |a'(z)|(\sqrt{z})^k\,dz\right)$$

$$\leqq\Gamma\{\nu(t)+\nu^2(t)+\nu^{k+1}(t)\}\mathscr{E}(t)\leqq\Gamma\{\nu(t)+\nu^{k+1}(t)\}\mathscr{E}(t)\quad\forall t\in[0,T_0)$$

and from (3.10) we estimate the following term:

(3.26)

$$\left|\int_0^t\int_0^1[\theta(x,s)-\theta^*]\int_0^s\theta_{xx}(x,y)\int_{s-y}^\infty a'(z)[F''(\bar\theta_x^s(x,z))-F''(0)]\,dz\,dy\,dx\,ds\right|$$

$$\leqq K(\nu(t)+\nu^k(t))\int_0^t\int_0^1|\theta(x,s)-\theta^*|\int_0^s|\theta_{xx}(x,y)|$$

$$\cdot\int_{s-y}^\infty|a'(z)|(\sqrt{z}+(\sqrt{z})^k)\,dz\,dy\,dx\,ds$$

$$\leqq K(\nu(t)+\nu^k(t))\sqrt{\mathscr{E}(t)}\left(\int_0^t\int_0^1\left(\int_0^s|\theta_{xx}(x,y)|\int_{s-y}^\infty|a'(z)|(\sqrt{z}\right.\right.$$

$$\left.\left.+(\sqrt{z})^k)\,dz\,dy\right)^2dx\,ds\right)^{1/2}$$

$$\leqq K(\nu(t)+\nu^k(t))\mathscr{E}(t)\int_0^\infty\int_s^\infty|a'(z)|(\sqrt{z}+(\sqrt{z})^k)\,dz\,ds$$

$$\leqq\Gamma\{\nu(t)+\nu^k(t)\}\mathscr{E}(t)\quad\forall t\in[0,T_0).$$

The rest of the terms on the right-hand side of (3.10), (3.12), and (3.15), except for the last term in (3.15), are handled in a similar fashion to (3.19), (3.20), (3.25), and (3.26). (Recall that (1.8) implies $F(0)=F'(0)=0$.) The last term on the right-hand side of (3.15) is first estimated from above, making use of compatibility with thermodynamics, i.e., we utilize the entropy inequality (0.2) in the estimation below: by (2.10) we have

$$-\int_0^t\int_0^1\theta_{tt}^2(x,s)\frac{\partial}{\partial x}\left\{\frac{1}{\theta(x,s)}\int_0^\infty a'(z)F'(\bar\theta_x^s(x,z))\,dz\right\}dx\,ds$$

$$\leqq\int_0^t\int_0^1\theta_{tt}^2(x,s)\left(\frac{\partial}{\partial s}\left\{-\hat\psi'(\theta(x,s))-\frac{1}{\theta(x,s)^2}\right.\right.$$

(3.27)

$$\left.\left.\cdot\int_0^\infty a'(z)F(\bar\theta_x^s(x,z))\,dz\right\}-\frac{r(x,s)}{\theta(x,s)}\right)dx\,ds$$

$$\forall t\in[0,T_0),$$

where $\hat\psi''\in C(0,\infty)$ (see Remark 2.1). Thus, it can be shown that

(3.28)

$$-\int_0^t\int_0^1\theta_{tt}^2(x,s)\frac{\partial}{\partial x}\left\{\frac{1}{\theta(x,s)}\int_0^\infty a'(z)F'(\bar\theta_x^s(x,z))\,dz\right\}dx\,ds$$

$$\leqq\Gamma\{\nu(t)+\nu^{k+1}(t)\}\mathscr{E}(t)+\Gamma\sqrt{R_0}\,\mathscr{E}(t)\quad\forall t\in[0,T_0).$$

Additional estimates are derived directly from equation (3.8) in the following manner. In order to obtain a temporal-L^2 estimate for θ_t we first multiply (3.8) by θ_{xx} and integrate the resulting identity over $[0, 1]\times[0, t]$, $t\in[0, T_0)$. We arrive at the relation

$$\begin{aligned}
&\frac{1}{2}\hat{e}'(\theta^*)\int_0^1 \theta_x^2(x, t)\,dx + F''(0)Q(\theta_{xx}, t, a)\\
&\quad=\frac{1}{2}\hat{e}'(\theta^*)\int_0^1 \theta_0'(x)^2\,dx\\
&\qquad+\int_0^t\int_0^1 \theta_{xx}(x, s)\Big([\hat{e}'(\theta(x, s))-\hat{e}'(\theta^*)]\theta_t(x, s)\\
&\qquad\qquad+\int_0^s \theta_{xx}(x, y)\int_{s-y}^{\infty} a'(z)[F''(\bar{\theta}_x^s(x, z))-F''(0)]\,dz\,dy\\
&\qquad\qquad+\frac{2}{\theta(x, s)^2}\theta_t(x, s)\int_0^{\infty} a'(z)F(\bar{\theta}_x^s(x, z))\,dz\\
&\qquad\qquad-\frac{2}{\theta(x, s)}\theta_x(x, s)\int_0^{\infty} a'(z)F'(\bar{\theta}_x^s(x, z))\,dz\\
&\qquad\qquad+\frac{2}{\theta(x, s)}\int_0^s a'(z)F'(\bar{\theta}_x^s(x, z))\theta_x(x, s-z)\,dz - r(x, s)\Big)\,dx\,ds,
\end{aligned}\tag{3.29}$$

$$t\in[0, T_0).$$

This relation leads to the inequality

$$Q(\theta_{xx}, t, a)\leqq\Gamma\Theta_0+\Gamma\sqrt{R_0}\sqrt{\mathscr{E}(t)}+\Gamma\{\nu(t)+\nu^k(t)\}\mathscr{E}(t)\quad\forall t\in[0, T_0).\tag{3.30}$$

We now square (3.8) and integrate over $[0, 1]\times[0, t]$, $t\in[0, T_0)$. The squares of the terms on the right-hand side of (3.8) are under control, and the square of the convolution term on the left-hand side of (3.8) can be estimated using (1.10), Lemma 4.2 of [13], and (3.30). Hence, we arrive at the estimate

$$\begin{aligned}
\int_0^t\int_0^1 \theta_t^2(x, s)\,dx\,ds &\leqq\Gamma\{\Theta_0+R_0\}+\Gamma\sqrt{R_0}\sqrt{\mathscr{E}(t)}\\
&\quad+\Gamma\{\nu(t)+\nu^{2k}(t)\}\mathscr{E}(t)\quad\forall t\in[0, T_0).
\end{aligned}\tag{3.31}$$

Equation (3.11) can be written as

$$\hat{e}'(\theta^*)\theta_{tt}(x, t)-F''(0)a(0)\theta_{xx}(x, t)-F''(0)\int_0^t a'(t-s)\theta_{xx}(x, s)\,ds = G_t(x, t),\tag{3.32}$$

$$x\in[0, 1],\quad t\in[0, T_0),$$

where $G(x, t)$ denotes the right-hand side of (3.8).

Solving for θ_{xx} in terms of θ_{tt} and G_t (see (2.36) and (2.37)) we get

$$\begin{aligned}
-F''(0)a(0)\theta_{xx}(x, t) &= G_t(x, t)-\hat{e}'(\theta^*)\theta_{tt}(x, t)\\
&\quad+\int_0^t m(t-s)[G_t(x, s)-\hat{e}'(\theta^*)\theta_{tt}(x, s)]\,ds
\end{aligned}\tag{3.33}$$

$$x\in[0, 1],\quad t\in[0, T_0)$$

where m is the resolvent of a' (see (2.38)). We note that by (3.8) $G(x,0)-\hat{e}'(\theta^*)\theta_t(x,0)=0$ for all $x\in[0,1]$, and by (2.38) $m(0)=-a'(0)/a(0)$. Thus, after integrating the last term on the right-hand side of (3.33) by parts we arrive at the expression

$$\begin{aligned}-F''(0)a(0)\theta_{xx}(x,t)&\\ =G_t(x,t)-\hat{e}'(\theta^*)\theta_{tt}(x,t)-\frac{a'(0)}{a(0)}[G(x,t)-\hat{e}'(\theta^*)\theta_t(x,t)]&\\ +\int_0^t m'(t-s)[G(x,s)-\hat{e}'(\theta^*)\theta_t(x,s)]\,ds, \qquad x\in[0,1],\quad t\in[0,T_0).&\end{aligned}\tag{3.34}$$

We now square (3.34) and integrate over $[0,1]$. By (3.18) and Lemma 2.3 we have

$$\begin{aligned}\int_0^1\theta_{xx}^2(x,t)\,dx\leqq\Gamma\{\Theta_0+R_0\}+\Gamma\{\sqrt{\Theta_0}+\sqrt{R_0}\}\sqrt{\mathscr{E}(t)}&\\ +\Gamma\sqrt{R_0}\,\mathscr{E}(t)+\Gamma\{\nu(t)+\nu^{2k+2}(t)\}\mathscr{E}(t)\quad\forall t\in[0,T_0).&\end{aligned}\tag{3.35}$$

To obtain a temporal-L^2 bound on θ_{tt} we multiply (3.34) by θ_{tt} and integrate over $[0,1]\times[0,t]$, $t\in[0,T_0)$. We note that

$$\begin{aligned}\int_0^t\int_0^1\theta_{tt}(x,s)\theta_{xx}(x,s)\,dx\,ds&=-\int_0^t\int_0^1\theta_x(x,s)\theta_{xtt}(x,s)\,dx\,ds\\ &=-\int_0^1\theta_x(x,t)\theta_{xt}(x,t)\,dx+\int_0^1\theta_0'(x)\theta_{xt}(x,0)\,dx\\ &\quad+\int_0^t\int_0^1\theta_{xt}^2(x,s)\,dx\,ds,\qquad t\in[0,T_0).\end{aligned}\tag{3.36}$$

Thus we have

$$\begin{aligned}\int_0^t\int_0^1\theta_{tt}^2(x,s)\,dx\,ds\leqq\Gamma\{\Theta_0+R_0\}+\Gamma\{\sqrt{\Theta_0}+\sqrt{R_0}\}\sqrt{\mathscr{E}(t)}&\\ +\Gamma\sqrt{R_0}\,\mathscr{E}(t)+\Gamma\{\nu(t)+\nu^{k+2}(t)\}\mathscr{E}(t)\quad\forall t\in[0,T_0).&\end{aligned}\tag{3.37}$$

(Here, we make crucial use of (2.20).) We now square (3.34) and integrate over $[0,1]\times[0,t]$, $t\in[0,T_0)$ using (3.18), (3.31), (3.37), and Lemma 2.3 to obtain the following estimate:

$$\begin{aligned}\int_0^t\int_0^1\theta_{xx}^2(x,s)\,dx\,ds\leqq\Gamma\{\Theta_0+R_0\}+\Gamma\{\sqrt{\Theta_0}+\sqrt{R_0}\}\sqrt{\mathscr{E}(t)}&\\ +\Gamma\sqrt{R_0}\,\mathscr{E}(t)+\Gamma\{\nu(t)+\nu^{2k+2}(t)\}\mathscr{E}(t)\quad\forall t\in[0,T_0).&\end{aligned}\tag{3.38}$$

Observe that by Poincaré's inequality there is a constant $c>0$ such that

$$\int_0^t\int_0^1[\theta(x,s)-\theta^*]^2\,dx\,ds\leqq c\int_0^t\int_0^1\theta_x^2(x,s)\,dx\,ds\quad\forall t\in[0,T_0).\tag{3.39}$$

It follows from (3.18), (3.31), (3.35), and (3.37)–(3.39) that

$$\begin{aligned}\mathscr{E}(t)\leqq\Gamma\{\Theta_0+R_0\}+\Gamma\{\sqrt{\Theta_0}+\sqrt{R_0}\}\sqrt{\mathscr{E}(t)}+\Gamma\sqrt{R_0}\,\mathscr{E}(t)&\\ +\Gamma\{\nu(t)+\nu^{2k+2}(t)\}\mathscr{E}(t)\quad\forall t\in[0,T_0).&\end{aligned}\tag{3.40}$$

Using (2.20), (3.40) yields

$$\mathscr{E}(t)\leqq\bar{\Gamma}\{\Theta_0+R_0\}+\bar{\Gamma}\sqrt{R_0}\,\mathscr{E}(t)+\bar{\Gamma}\{\nu(t)+\nu^{2k+2}(t)\}\mathscr{E}(t)\quad\forall t\in[0,T_0),\tag{3.41}$$

where $\bar{\Gamma}$ denotes a fixed positive constant which is independent of θ_0, r, and T_0. We choose $\bar{\mathscr{E}}$, $\delta > 0$ such that

$$\bar{\mathscr{E}} < \varepsilon^2/2, \quad \bar{\Gamma}\{\sqrt{2\bar{\mathscr{E}}} + (\sqrt{2\bar{\mathscr{E}}})^{2k+2}\} \leqq \tfrac{1}{6}, \quad \bar{\Gamma}\delta^2 \leqq \tfrac{1}{6}\bar{\mathscr{E}}, \quad \bar{\Gamma}\delta \leqq \tfrac{1}{6}, \tag{3.42}$$

and

$$\delta < \tfrac{1}{2}\eta \tag{3.43}$$

for some $\eta \in (0, \varepsilon)$.

Suppose now that (1.21) holds for the above choice of δ. By the Sobolev embedding theorem

$$\nu(t) \leqq \sqrt{2\mathscr{E}(t)} \quad \forall t \in [0, T_0). \tag{3.44}$$

Thus, it follows from (3.41) that for any $t \in [0, T_0)$ with $\mathscr{E}(t) \leqq \bar{\mathscr{E}}$, we actually have $\mathscr{E}(t) \leqq \frac{1}{2}\bar{\mathscr{E}}$. Hence by continuity, if $\mathscr{E}(0) \leqq \frac{1}{2}\bar{\mathscr{E}}$ then

$$\mathscr{E}(t) \leqq \tfrac{1}{2}\bar{\mathscr{E}} \quad \forall t \in [0, T_0). \tag{3.45}$$

It is possible to choose a smaller $\delta > 0$ (if necessary) so that (1.21) implies $\mathscr{E}(0) \leqq \frac{1}{2}\bar{\mathscr{E}}$. Consequently, for $\delta > 0$ small enough, (3.45) holds; moreover, by the Sobolev embedding theorem

$$\sup_{\substack{x\in[0,1]\\ t\in[0,T_0)}} |\theta(x,t) - \theta^*|, \quad \sup_{\substack{x\in[0,1]\\ t\in[0,T_0)}} |\theta_x(x,t)| \leqq \sqrt{2\bar{\mathscr{E}}} < \varepsilon. \tag{3.46}$$

Therefore, by Lemma 2.2 we have $T_0 = \infty$. In addition, (1.23) is an immediate consequence of (3.45). Moreover, (1.24) and (1.25) follow from (1.23) by standard embedding inequalities, e.g., from (1.23) we have

$$\theta - \theta^* \in L^\infty((0,\infty); L^2(0,1)) \tag{3.47}$$

and

$$\theta_x, \theta_{xt} \in L^2((0,\infty); L^2(0,1)). \tag{3.48}$$

We note that (3.48) implies

$$\theta_x(\cdot, t) \to 0 \quad \text{in } L^2(0,1) \quad \text{as } t \to \infty. \tag{3.49}$$

Observe that

$$\begin{aligned} [\theta(x,t) - \theta^*]^2 &= 2\int_0^x [\theta(\xi,t) - \theta^*]\theta_x(\xi,t)\, d\xi \\ &\leqq 2\int_0^1 |\theta(\xi,t) - \theta^*||\theta_x(\xi,t)|\, d\xi \\ &\leqq 2\left(\int_0^1 [\theta(\xi,t) - \theta^*]^2\, d\xi\right)^{1/2}\left(\int_0^1 \theta_x^2(\xi,t)\, d\xi\right)^{1/2} \quad \forall x \in [0,1], \quad t \geqq 0. \end{aligned} \tag{3.50}$$

Hence, by (3.47) and (3.49)

$$\theta(\cdot, t) \to \theta^* \quad \text{uniformly on } [0,1] \quad \text{as } t \to \infty. \tag{3.51}$$

This completes the proof of Theorem 1.1. □

The proofs of Theorems 1.2 and 1.3 are very similar to the proof above. In both cases, however, since we cannot use Poincaré's inequality, we do not obtain a temporal-L^2 estimate for $\theta-\theta^*$, and hence before we proceed with the calculations we divide equations (1.1) and (1.40) by $\hat{e}'(\theta(x,t))$. For the same reason, in Theorem 1.2, for example, we require that (1.29) hold in order to obtain the following estimate:

$$
\begin{aligned}
&\left|\int_0^t\int_0^1 \frac{1}{\hat{e}'(\theta(x,s))}[\theta(x,s)-\theta^*]r(x,s)\,dx\,ds\right| \\
&\quad\leqq \Gamma\int_0^t\left(\int_0^1[\theta(x,s)-\theta^*]^2\,dx\right)^{1/2}\left(\int_0^1 r^2(x,s)\,dx\right)^{1/2}ds \\
&\quad\leqq \Gamma\sup_{s\in[0,t]}\left(\int_0^1[\theta(x,s)-\theta^*]^2\,dx\right)^{1/2}\int_0^t\left(\int_0^1 r^2(x,s)\,dx\right)^{1/2}ds \\
&\quad\leqq \Gamma\sqrt{\mathscr{E}(t)}\int_0^\infty\left(\int_0^1 r^2(x,t)\,dx\right)^{1/2}dt \quad \forall t\in[0,T_0);
\end{aligned}
\tag{3.52}
$$

the other terms with which we must be careful can be handled by integration by parts, e.g.,

$$
\begin{aligned}
&\left|\int_0^t\int_0^1 \frac{1}{\hat{e}'(\theta(x,s))}[\theta(x,s)-\theta^*]\int_0^\infty a'(z)[F''(\bar{\theta}_x^s(x,z))-F''(0)]\bar{\theta}_{xx}^s(x,z)\,dz\,dx\,ds\right| \\
&\quad=\left|-\int_0^t\int_0^1 \frac{1}{\hat{e}'(\theta(x,s))}[\theta(x,s)-\theta^*]\frac{\partial}{\partial x}\right. \\
&\qquad\left.\cdot\left\{\int_0^\infty a(z)[F''(\bar{\theta}_x^s(x,z))-F''(0)]\theta_x(x,s-z)\,dz\right\}dx\,ds\right| \\
&\quad=\left|\int_0^t\int_0^1\left(-\frac{\hat{e}''(\theta(x,s))}{\hat{e}'(\theta(x,s))^2}[\theta(x,s)-\theta^*]+\frac{1}{\hat{e}'(\theta(x,s))}\right)\theta_x(x,s)\right. \\
&\qquad\left.\cdot\int_0^s a(z)[F''(\bar{\theta}_x^s(x,z))-F''(0)]\theta_x(x,s-z)\,dz\,dx\,ds\right| \\
&\quad\leqq\Gamma(\nu(t)+\nu^{k+1}(t))\mathscr{E}(t) \quad \forall t\in[0,T_0).
\end{aligned}
\tag{3.53}
$$

The argument to show that $\theta(\cdot,t)\to\theta^{**}$ uniformly on $[0,1]$ as $t\to\infty$ for Theorem 1.2 is essentially the same as the argument used to establish an analogous result in § 3 of [4]: We first observe that standard embedding inequalities yield (1.38) as well as boundedness of θ on $[0,1]\times[0,\infty)$. Hence, every sequence of times tending to infinity has a subsequence on which θ converges uniformly to a constant, namely, θ^{**}.

Theorems 1.2 and 1.3 can be proved using an argument in the same spirit as in [13], i.e., instead of taking temporal derivatives of the equation and multiplying by corresponding time derivatives of θ, we can take spatial derivatives of the equation and multiply by appropriate x derivatives of θ. This cannot be done for Theorem 1.1 since we have a term involving $\theta_x(x,t)$ on the right-hand side of (1.1) which would lead to uncontrollable boundary terms.

In the case of nonequilibrium history the argument is essentially the same. The main modification needed arises when we want to make use of an inequality of the form (2.21); we then extend a to $\mathbb{R}$ by zero. To give an indication of where such a

modification is needed we consider the analogue of the term treated in (3.26):

$$\left|\int_0^t\int_0^1 [\theta(x,s)-\theta^*]\int_{-\infty}^s \theta_{xx}(x,y)\int_{s-y}^{\infty} a'(z)[F''(\bar\theta_x^s(x,z))-F''(0)]\,dz\,dy\,dx\,ds\right|$$

$$\leqq \left|\int_0^t\int_0^1 [\theta(x,s)-\theta^*]\int_{-\infty}^0 \theta_{xx}(x,y)\right.$$

$$\left.\cdot\int_{s-y}^{\infty} a'(z)[F''(\bar\theta_x^s(x,z))-F''(0)]\,dz\,dy\,dx\,ds\right| \tag{3.54}$$

$$+\left|\int_0^t\int_0^1 [\theta(x,s)-\theta^*]\int_0^s \theta_{xx}(x,y)\right.$$

$$\left.\cdot\int_{s-y}^{\infty} a'(z)[F''(\bar\theta_x^s(x,z))-F''(0)]\,dz\,dy\,dx\,ds\right|.$$

The second term on the right-hand side of (3.54) can clearly be handled in the same manner as in (3.26) and once a is extended by zero, the first term on the right-hand side of (3.54) can also be treated in the same way.

Remark 3.1. We note that in order to obtain a priori bounds in the above proof it suffices to assume that the data satisfy (1.55)-(1.58). It is in the proof of local existence that we need the original assumptions on the data (1.12)-(1.18).

Remark 3.2. If, for example, in the case of Theorem 1.1 assumption (1.21) is replaced with

$$\Theta_0+\int_0^1 \theta_0'''(x)^2\,dx+R_0+\sup_{t\geqq 0}\int_0^1 (r_x^2+r_{xt}^2+r_{tt}^2)(x,t)\,dx$$

$$+\int_0^1 r_{xx}^2(x,0)\,dx+\int_0^{\infty}\int_0^1 (r_x^2+r_{xt}^2+r_{ttt}^2)(x,t)\,dx\leqq\delta^2, \tag{3.55}$$

then we can establish the existence of a unique solution $\theta>0$ satisfying (1.22)-(1.25); moreover,

$$\theta_{xxx},\ \theta_{xxt},\ \theta_{xtt},\ \theta_{ttt}\in L^{\infty}((0,\infty);\ L^2(0,1))\cap L^2((0,\infty);\ L^2(0,1)), \tag{3.56}$$

and

$$\theta_{xx}(\cdot,t),\ \theta_{xt}(\cdot,t),\ \theta_{tt}(\cdot,t)\to 0 \quad \text{uniformly on } [0,1] \tag{3.57}$$

as $t\to\infty$. The arguments used to establish such a result are similar in spirit to the arguments used to prove Theorem 1.1 except that here there is no need to make use of the entropy inequality (0.2) or any other consequence of the thermodynamical restrictions.

Acknowledgments. I thank my thesis advisor W. J. Hrusa for numerous valuable discussions. Many of the results described in this paper are due to his suggestions.

REFERENCES

[1] D. BRANDON, *On a class of models for heat flow in materials with memory*, Ph.D. thesis, Department of Mathematics, Carnegie-Mellon University, Pittsburgh, PA, 1988.

[2] D. BRANDON AND W. J. HRUSA, *Construction of a class of integral models for heat flow in materials with memory*, J. Integral Equations Appl., 1 (1988), pp. 175-201.

[3] P. J. CHEN, *On the growth and decay of one-dimensional temperature rate waves*, Arch. Rational Mech. Anal., 35 (1969), pp. 1-15.

[4] B. D. COLEMAN, W. J. HRUSA, AND D. R. OWEN, *Stability of equilibrium for a nonlinear hyperbolic system describing heat propagation by second sound in solids*, Arch. Rational Mech. Anal., 94 (1986), pp. 267–289.

[5] C. M. DAFERMOS AND J. A. NOHEL, *Energy methods for nonlinear hyperbolic Volterra integrodifferential equations*, Comm. Partial Differential Equations, 4 (1979), pp. 219–278.

[6] M. E. GURTIN AND A. C. PIPKIN, *A general theory of heat conduction with finite wave speeds*, Arch. Rational Mech. Anal., 31 (1968), pp. 113–126.

[7] W. J. HRUSA AND J. A. NOHEL, *The Cauchy problem in one-dimensional nonlinear viscoelasticity*, J. Differential Equations, 59 (1985), pp. 388–412.

[8] W. J. HRUSA AND M. RENARDY, *A model equation for viscoelasticity with a strongly singular kernel*, SIAM J. Math. Anal., 19 (1988), pp. 257–269.

[9] G. S. JORDAN, O. J. STAFFANS, AND R. L. WHEELER, *Local analyticity in weighted L^1-spaces and applications to stability problems for Volterra equations*, Trans. Amer. Math. Soc., 274 (1982), pp. 749–782.

[10] R. C. MACCAMY, *An integro-differential equation with application in heat flow*, Quart. J. Appl. Math., 35 (1977), pp. 1–19.

[11] J. A. NOHEL AND D. F. SHEA, *Frequency domain methods for Volterra equations*, Adv. in Math., 22 (1976), pp. 278–304.

[12] M. RENARDY, W. J. HRUSA, AND J. A. NOHEL, *Mathematical Problems in Viscoelasticity*, Longmans Press, London, 1987.

[13] O. J. STAFFANS, *On a nonlinear hyperbolic Volterra equation*, SIAM J. Math. Anal., 11 (1980), pp. 793–812.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 107-128, January 1991

006

RAPIDLY STRETCHING PLASTIC JETS: THE LINEARIZED PROBLEM*

FERNANDO REITICH†

Abstract. A linearized version of the Levy-von Mises equations modeling the evolution of rapidly stretching plastic jets is studied. Under the assumptions of axial symmetry and stress-free surface, existence and uniqueness of a solution is shown for the resulting (nonstandard) linear initial-boundary value problem. Some growth and periodicity properties of the solution are also established.

Key words. jets, Levy-von Mises equations, linearization

AMS(MOS) subject classifications. 35Q20, 76D25

Introduction. A typical example of a rapidly stretching jet is furnished by the jet produced by a shaped-charge [1], [2], [9], [12]. A shaped-charge consists of an explosive with a conical cavity lined with a thin metal sheet; the explosion will cause the metal to collapse toward the axis where an extremely high velocity jet will instantly be formed. The velocity of the particles in these jets increases linearly with the distance from the rear end, so that the jet experiences a very significant stretching.

In a recent paper, Romero [14] analyzes the stability of these jets using the Levy-von Mises equations for an incompressible perfectly plastic material (see § 1 below). He finds that a rapidly stretching plastic jet can be initially stable due to inertial effects, a result that had been anticipated by Frankel and Weihs [4] based on their study of the stability of a capillary jet of an ideal fluid (see also, e.g., Rayleigh [13], Weber [15], Levich [11], Goldin et al. [8], Bogy [3]).

Assuming that the jet is axially symmetric and that its surface is stress-free, Romero finds a particular solution ("the undisturbed flow"), for which the axial velocity is linearly increasing. He then linearizes the equations about it, introducing scaled space variables (r, z) $(0<r<1, -\infty<z<\infty)$ and a scaled time variable t $(0\leqq t\leqq T)$, to reduce this linear system to a system of three equations, essentially of the form

$$\frac{\partial\phi}{\partial t}=\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\phi}{\partial r}\right)-\frac{\partial^2\phi}{\partial z^2}-\frac{\partial p}{\partial z}, \tag{0.1}$$

$$\frac{\partial\psi}{\partial t}=\frac{\partial^2\psi}{\partial z^2}-\frac{\partial p}{\partial r}, \tag{0.2}$$

$$\operatorname{div}(\phi,\psi)=0, \tag{0.3}$$

where ϕ, ψ, p correspond to scaled versions of the perturbations in the axial velocity, radial velocity, and pressure, respectively.

Here r represents the radial variable scaled according to the radius of the free boundary (i.e., the moving boundary of the jet) and z is an appropriately scaled axial variable. The stress-free and compatibility conditions on the surface of the jet yield nonstandard boundary conditions for (0.1)-(0.3) of the form

$$p=B_1(\phi,\psi,\Omega) \quad \text{at } r=1, \tag{0.4}$$

* Received by the editors May 22, 1989; accepted for publication (in revised form) February 8, 1990.
† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

$$\frac{\partial \phi}{\partial r} = B_2(\psi, \Omega) \quad \text{at } r = 1, \tag{0.5}$$

$$\frac{\partial \Omega}{\partial t} = \psi \quad \text{at } r = 1, \tag{0.6}$$

where B_1 and B_2 are certain first-order differential operators and $\Omega = \Omega(t, z)$ denotes the perturbation in the radius of the jet scaled according to the actual radius.

To complete the setting of the problem we need to impose initial conditions, namely,

$$\phi(0, r, z) = \phi_0(r, z), \tag{0.7}$$

$$\Omega(0, 0) = \Omega_0. \tag{0.8}$$

The system (0.1)–(0.8) is a nonstandard linear problem, due both to the "mixed type" differential equations and the boundary conditions.

While the main purpose of [14] is to study the stability of (0.1)–(0.8), the main result of the present paper is an existence and uniqueness theorem for this system. Romero's analysis of the stability of short wavelength solutions

$$\begin{aligned} &\phi(t, r, z) = \phi^*(t, r, \mu)\, e^{i\mu z}, \qquad \psi(t, r, z) = \psi^*(t, r, \mu)\, e^{i\mu z}, \\ &p(t, r, z) = p^*(t, r, \mu)\, e^{i\mu z}, \qquad \Omega(t, z) = \Omega^*(t, \mu)\, e^{i\mu z}, \end{aligned} \tag{0.9}$$

with $|\mu| \gg 1$

leads him to the conclusion that such disturbances do not undergo significant growth. However, his argument is based on the assumption that, in (0.9),

$$\begin{aligned} &\phi^*(t, r, \mu) = \tilde{\phi}(r, \mu)\, e^{\gamma_1 t}, \qquad \psi^*(t, r, \mu) = \tilde{\psi}(r, \mu)\, e^{\gamma_2 t}, \\ &p^*(t, r, \mu) = \tilde{p}(r, \mu)\, e^{\gamma_3 t}, \qquad \Omega^*(t, \mu) = \tilde{\Omega}(\mu)\, e^{\gamma_4 t}, \end{aligned} \tag{0.10}$$

which translates into quite restrictive conditions for ϕ_0. In this paper we consider general initial conditions $\phi_0(r, z)$. Since the jet is expected to eventually become unstable (cf. [14]), the system may develop ill-posedness (as in the backward heat equation): for this reason we assume that ϕ_0 is entire analytic in the variable z. More precisely, we show that if $\phi_0(x, y, z) = \phi_0((x^2+y^2)^{1/2}, z)$ is $C^{2+\alpha}$ in (x, y) (for some $\alpha \in (0, 1)$) and if $\phi_0(x, y, \cdot)$ extends to an entire function in z of order 1 and sufficiently small type, then there exists a unique solution of (0.1)–(0.8).

In § 1 we briefly describe the linearization procedure which leads to (0.1)–(0.6). In § 2 we reduce the system into another system which is much more convenient to work with; in particular, ψ and Ω do not appear in the new system. In § 3 we state the main existence and uniqueness results (Theorem 3.1) and in § 4 we prove a preliminary lemma (Lemma 4.1), which immediately implies Theorem 3.1 in the special case of initial conditions of the form $\phi_0(x, y, z) = u_0(x, y)z^n (n \geqq 0)$. Section 5 is devoted to the proof of Theorem 3.1 and, finally, in § 6 we establish periodicity and growth properties of the solution.

1. The linearized system. We shall assume that the flow obeys the laws for an incompressible perfectly plastic material satisfying the Levy–von Mises equations and that the surface of the jet is stress-free. The Levy–von Mises equations are

$$\rho\left(\frac{\partial \bar{u}}{\partial t} + \bar{u} \cdot \nabla \bar{u}\right) = -\nabla p + \operatorname{div} \bar{\bar{T}}, \tag{1.1}$$

$$\nabla \cdot \bar{u} = 0, \tag{1.2}$$

where ρ is the constant material density, p is the pressure, and $\bar{\bar{T}}$ is the deviatoric stress. In Cartesian coordinates $\bar{\bar{T}}$ is given by

$$T_{ij} = 2\mu\dot{\varepsilon}_{ij}, \tag{1.3}$$

where $\dot{\varepsilon}_{ij}$ is the rate of strain tensor

$$\dot{\varepsilon}_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \tag{1.4}$$

μ is the effective viscosity

$$\mu = Y(2\dot{\varepsilon}_{kl}\dot{\varepsilon}_{kl})^{-1/2}, \tag{1.5}$$

and Y is the yield stress of the material.

Assuming axial symmetry, let (z, r) be the cylindrical coordinates and let (u, v) be their corresponding velocities. Then, a special solution, satisfying stress-free boundary conditions and the compatibility relation of the time rate of change of the free surface to the velocity field at the boundary, is given by (see [14])

$$u_0(t, r, z) = \frac{\beta(0)z}{q(t)}, \tag{1.6}$$

$$v_0(t, r, z) = \frac{-\beta(0)r}{2q(t)}, \tag{1.7}$$

$$p_0(t, r, z) = \frac{3}{8}\rho\beta(t)^2(r^2 - R_0(t)^2) - \frac{Y}{3^{1/2}}, \tag{1.8}$$

$$R_0(t, z) = \frac{a_0(0)}{q(t)^{1/2}}, \tag{1.9}$$

where $a_0(0)$ is the initial radius of the jet, $\beta(0)$ is the initial strain rate, $\beta(t) = \beta(0)/q(t)$, and $q(t) = \beta(0)t + 1$.

Linearizing about this solution and introducing scaled variables

$$\hat{t} = \ln(q(t)), \tag{1.10}$$

$$\hat{r} = \frac{r}{R_0(t)}, \tag{1.11}$$

$$\hat{z} = \frac{z}{L(t)}, \tag{1.12}$$

where $L(t) = q(t)L(0)$ and the parameter $L(0)$ is the initial length scale, the linearized system takes the form (after dropping the ^'s)

$$\frac{\partial\phi}{\partial t} = \Gamma^{-2}\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\phi}{\partial r}\right) - \alpha^2\frac{\partial^2\phi}{\partial z^2}\right) - \phi - \alpha^2\frac{\partial p}{\partial z}, \tag{1.13}$$

$$\frac{\partial\psi}{\partial t} - 2\psi = \alpha^2\Gamma^{-2}\frac{\partial^2\psi}{\partial z^2} - \frac{\partial p}{\partial r}, \tag{1.14}$$

$$\frac{\partial\phi}{\partial z} + \frac{1}{r}\frac{\partial}{\partial r}(r\psi) = 0 \tag{1.15}$$

for $(t, r, z) \in [0, T] \times (0, 1] \times \mathbb{R}$ $(T > 0)$, and

$$\Gamma^2 p = 2\left(\frac{\partial \psi}{\partial r} + \frac{1}{2}\frac{\partial \phi}{\partial z}\right) + \frac{3}{4}\Omega\Gamma^2 \equiv \Gamma^2 B_1(\phi, \psi, \Omega) \quad \text{at } r = 1, \tag{1.16}$$

$$\frac{\partial \phi}{\partial r} = 3\alpha^2 \frac{\partial \Omega}{\partial z} - \alpha^2 \frac{\partial \psi}{\partial z} \equiv B_2(\psi, \Omega) \quad \text{at } r = 1, \tag{1.17}$$

$$\frac{\partial \Omega}{\partial t} = \psi \quad \text{at } r = 1, \tag{1.18}$$

(1.19) ψ is finite as r approaches zero,

where ϕ, ψ, p, and Ω are scaled versions of the perturbations in the axial velocity, radial velocity, pressure, and jet radius, respectively. Also, the functions α and Γ depend only on t and are given by

$$\alpha(t)^2 = \alpha(0)^2 e^{-3t}, \qquad \Gamma(t)^2 = \Gamma(0)^2 e^{-3t},$$

where

$$\alpha(0)^2 = (a_0(0)/L(0))^2, \qquad \Gamma(0)^2 = 3^{1/2}\rho a_0(0)^2 \beta(0)^2 / Y.$$

We supplement the system by adding the initial conditions

$$\phi(0, r, z) = \phi_0(r, z), \tag{1.20}$$

$$\Omega(0, 0) = \Omega_0. \tag{1.21}$$

2. An equivalent linear system. The purpose of this section is to reduce the nonstandard system (1.13)-(1.21) into a form which will be more recognizable and easier to work with. The new system will not involve Ω and ψ and it will be amenable to the theory of elliptic and parabolic differential equations.

The first step is to note that if (ϕ, ψ, p, Ω) is a solution of (1.13)-(1.20) with $\Omega(0, 0) = \Omega_0$, then

$$(\phi, \psi, p + \tfrac{3}{4}(\omega_0 - \Omega_0), \Omega + \omega_0 - \Omega_0)$$

is a solution of (1.13)-(1.20) satisfying $\Omega(0, 0) = \omega_0$. Thus, we may assume that

$$\Omega(0, 0) = -\frac{1}{3}\int_0^1 r \frac{\partial \phi_0}{\partial z}(r, 0)\, dr. \tag{2.1}$$

LEMMA 2.1. *Assume that*

$$\frac{\partial \phi_0}{\partial r}(1, z) = 0, \qquad z \in \mathbb{R}. \tag{2.2}$$

Then, (ϕ, ψ, p, Ω) *is a solution of* (1.13)-(1.20) *satisfying* (2.1) *if and only if*

(i) (ϕ, p) *satisfy*

$$\frac{\partial \phi}{\partial t} + \phi = -\alpha^2 \frac{\partial p}{\partial z} + \Gamma^{-2}\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \phi}{\partial r}\right) - \alpha^2 \frac{\partial^2 \phi}{\partial z^2}\right), \tag{2.3a}$$

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial p}{\partial r}\right) = \frac{\partial}{\partial t}\left(\frac{\partial \phi}{\partial z}\right) - 2\frac{\partial \phi}{\partial z} - \alpha^2 \Gamma^{-2}\frac{\partial^3 \phi}{\partial z^3} \tag{2.3b}$$

for $(t, r, z) \in [0, T] \times (0, 1] \times \mathbb{R}$, *with initial-boundary conditions given by*

$$\frac{\partial \phi}{\partial r} = -3\alpha^2 \int_0^t \int_0^1 \rho \frac{\partial^2 \phi}{\partial z^2}(s, \rho, z)\, d\rho\, ds - \alpha^2 \int_0^1 \rho \frac{\partial^2 \phi_0}{\partial z^2}(\rho, z)\, d\rho + \alpha^2 \int_0^1 \rho \frac{\partial^2 \phi}{\partial z^2}(t, \rho, z)\, d\rho \quad \text{at } r = 1, \tag{2.4}$$

$$p = -\frac{3}{4} \int_0^t \int_0^1 \rho \frac{\partial \phi}{\partial z}(s, \rho, z)\, d\rho\, ds - \Gamma^{-2} \frac{\partial \phi}{\partial z} - \frac{1}{4} \int_0^1 \rho \frac{\partial \phi_0}{\partial z}(\rho, z)\, d\rho + 2\Gamma^{-2} \int_0^1 \rho \frac{\partial \phi}{\partial z}(t, \rho, z)\, d\rho \quad \text{at } r = 1, \tag{2.5}$$

$$\phi(0, r, z) = \phi_0(r, z); \tag{2.6}$$

and

(ii) ψ *and* Ω *are given by*

$$\psi(t, r, z) = -\frac{1}{r} \int_0^r \rho \frac{\partial \phi}{\partial z}(t, \rho, z)\, d\rho, \tag{2.7}$$

$$\Omega(t, z) = -\int_0^t \int_0^1 \rho \frac{\partial \phi}{\partial z}(s, \rho, z)\, d\rho\, ds - \frac{1}{3} \int_0^1 \rho \frac{\partial \phi_0}{\partial z}(\rho, z)\, d\rho. \tag{2.8}$$

Proof. Suppose that (ϕ, ψ, p, Ω) is a solution of (1.13)-(1.20) satisfying (2.1). We want to show that (2.3)-(2.8) hold.

First note that (2.3a), (2.6) are exactly (1.13), (1.20). To prove (2.3b) multiply (1.14) by r and differentiate with respect to r to obtain (using (1.15))

$$-r \frac{\partial}{\partial t}\left(\frac{\partial \phi}{\partial z}\right) + 2r \frac{\partial \phi}{\partial z} = -\frac{\partial}{\partial r}\left(r \frac{\partial p}{\partial r}\right) - cr \frac{\partial^3 \phi}{\partial z^3},$$

where $c = \alpha(t)^2 \Gamma(t)^{-2} = a_0(0)^2 (L(0)\Gamma(0))^{-2} \geqq 0$. Dividing by r, we get

$$-\frac{\partial}{\partial t}\left(\frac{\partial \phi}{\partial z}\right) + 2 \frac{\partial \phi}{\partial z} = -\frac{1}{r} \frac{\partial}{\partial r}\left(r \frac{\partial p}{\partial r}\right) - c \frac{\partial^3 \phi}{\partial z^3},$$

which gives (2.3b).

In order to prove (2.4), (2.5) we must first establish (2.7), (2.8). Multiplying (1.15) by r and integrating with respect to r we get

$$\psi(t, 1, z) - r\psi(t, r, z) = -\int_r^1 \rho \frac{\partial \phi}{\partial z}(t, \rho, z)\, d\rho.$$

Using (1.19) we conclude that

$$\psi(t, 1, z) = -\int_0^1 \rho \frac{\partial \phi}{\partial z}(t, \rho, z)\, d\rho,$$

so that

$$-r\psi(t, r, z) = -\int_r^1 \rho \frac{\partial \phi}{\partial z}(t, \rho, z)\, d\rho + \int_0^1 \rho \frac{\partial \phi}{\partial z}(t, \rho, z)\, d\rho = \int_0^r \rho \frac{\partial \phi}{\partial z}(t, \rho, z)\, d\rho,$$

and (2.7) follows. Now we can use (1.18) and (2.7) to obtain

$$\begin{aligned}\Omega(t,z) &= \int_0^t \psi(s,1,z)\,ds + \Omega(0,z) \\ &= -\int_0^t \int_0^1 \rho \frac{\partial \phi}{\partial z}(s,\rho,z)\,d\rho\,ds + \Omega(0,z).\end{aligned} \tag{2.9}$$

On the other hand, specializing (1.17) at $t=0$ and using (2.7) once again

$$\frac{\partial \phi_0}{\partial r}(1,z) - \alpha(0)^2\left(\int_0^1 \rho \frac{\partial^2 \phi_0}{\partial z^2}(\rho,z)\,d\rho\right) = 3\alpha(0)^2 \frac{\partial \Omega}{\partial z}(0,z),$$

which, due to assumption (2.2), is

$$\frac{\partial \Omega}{\partial z}(0,z) = -\frac{1}{3}\int_0^1 \rho \frac{\partial^2 \phi_0}{\partial z^2}(\rho,z)\,d\rho. \tag{2.10}$$

But (2.1) and (2.10) imply

$$\Omega(0,z) = -\frac{1}{3}\int_0^1 \rho \frac{\partial \phi_0}{\partial z}(\rho,z)\,d\rho, \tag{2.11}$$

which, together with (2.9), gives (2.8).

Thus, we may now use (2.7), (2.8) to rewrite the boundary condition (1.17) as

$$\begin{aligned}\frac{\partial \phi}{\partial r} - \alpha^2 \int_0^1 \rho \frac{\partial^2 \phi}{\partial z^2}(t,\rho,z)\,d\rho = -3\alpha^2 \int_0^t \int_0^1 \rho \frac{\partial^2 \phi}{\partial z^2}(s,\rho,z)\,d\rho\,ds \\ -\alpha^2 \int_0^1 \rho \frac{\partial^2 \phi_0}{\partial z^2}(\rho,z)\,d\rho \quad \text{at } r=1,\end{aligned}$$

which is (2.4).

Finally, using (1.15) we can rewrite (1.16) as

$$2\left(-\psi - \frac{\partial \phi}{\partial z} + \frac{1}{2}\frac{\partial \phi}{\partial z}\right) - \Gamma^2 p + \frac{3}{4}\Omega\Gamma^2 = 0 \quad \text{at } r=1,$$

or

$$-2\psi - \frac{\partial \phi}{\partial z} - \Gamma^2 p + \frac{3}{4}\Omega\Gamma^2 = 0 \quad \text{at } r=1, \tag{2.12}$$

and (2.5) follows from (2.7), (2.8), and (2.12).

From the above computations it is clear how to proceed in proving the converse statement, that is, if ϕ, p satisfy (2.3)-(2.6) and ψ, Ω are given by (2.7) and (2.8), then (ϕ, ψ, p, Ω) is a solution of (1.13)-(1.20) satisfying (2.1).

This completes the proof of Lemma 2.1. □

It will be convenient to rewrite the system (2.3)-(2.6) in Cartesian coordinates. Taking r to be the radial variable, $r=(x^2+y^2)^{1/2}$, and setting $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$, B_1 = unit disc in $\mathbb{R}^2$, $G = [0,T] \times B_1 \times \mathbb{R}$, $\Gamma_0 = [0,T] \times \partial B_1 \times \mathbb{R}$, we can write

$$\frac{\partial \phi}{\partial t} - \Gamma(t)^{-2}\Delta\phi + c\frac{\partial^2 \phi}{\partial z^2} + \phi = -\alpha(t)^2 \frac{\partial p}{\partial z} \quad \text{in } G, \tag{2.13a}$$

$$\Delta p = \frac{\partial}{\partial t}\left(\frac{\partial \phi}{\partial z}\right) - 2\frac{\partial \phi}{\partial z} - c\frac{\partial^3 \phi}{\partial z^3} \quad \text{in } G, \tag{2.13b}$$

$$\frac{\partial\phi}{\partial\nu} = -3\frac{\alpha(t)^2}{2\pi}\int_0^t\int_{B_1}\frac{\partial^2\phi}{\partial z^2}(s,x,y,z)\,dx\,dy\,ds$$

$$+\frac{\alpha(t)^2}{2\pi}\int_{B_1}\frac{\partial^2\phi}{\partial z^2}(t,x,y,z)\,dx\,dy \tag{2.13c}$$

$$-\frac{\alpha(t)^2}{2\pi}\int_{B_1}\frac{\partial^2\phi_0}{\partial z^2}(x,y,z)\,dx\,dy \quad \text{on } \Gamma_0,$$

$$p = -\frac{3}{8\pi}\int_0^t\int_{B_1}\frac{\partial\phi}{\partial z}(s,x,y,z)\,dx\,dy\,ds - \Gamma(t)^{-2}\frac{\partial\phi}{\partial z}$$

$$-\frac{1}{8\pi}\int_{B_1}\frac{\partial\phi_0}{\partial z}(x,y,z)\,dx\,dy \tag{2.13d}$$

$$+\frac{\Gamma(t)^{-2}}{\pi}\int_{B_1}\frac{\partial\phi}{\partial z}(t,x,y,z)\,dx\,dy \quad \text{on } \Gamma_0,$$

$$\phi(0,x,y,z) = \phi_0(x,y,z), \tag{2.13e}$$

where ν = exterior unit normal to ∂B_1 and $c = \alpha(t)^2\Gamma(t)^{-2} = (a_0(0)/L(0)\Gamma(0))^2$ $(c \geqq 0)$.

3. Statement of the existence and uniqueness results. Before stating our main result, which will be proved in § 5, we need to define the class of "admissible" initial data.

Let $C^{2+\alpha}(\overline{B_1})$ denote the classical space of functions defined on $\overline{B_1}$ whose second derivatives are Hölder continuous with exponent α $(\alpha \in (0,1))$, and let $\bar{C}_\alpha$, $\bar{C}_{2+\alpha}$ denote the parabolic Hölder spaces on $[0,T]\times\overline{B_1}$ (see, e.g., Friedman [6, pp. 61–63]). For $f \in C^{2+\alpha}(\overline{B_1})$, $g \in \bar{C}_\alpha$, $h \in \bar{C}_{2+\alpha}$ we introduce the notation

$$\|f\|_{2+\alpha}^0 = \|f\|_{2+\alpha,\overline{B_1}},$$

$$\|g\|_\alpha = \|g\|_{\alpha,[0,T]\times\overline{B_1}},$$

$$\|h\|_{2+\alpha} = \|h\|_{2+\alpha,[0,T]\times\overline{B_1}}.$$

Finally, if k is a function satisfying $k \in \bar{C}_\alpha$, $k(t,\cdot) \in C^{2+\alpha}(\overline{B_1})$ $(t \in [0,T])$, we shall write

$$|||k||| = \|k\|_{\alpha,[0,T]\times\overline{B_1}} + \sup_{t\in[0,T]}\|k(t,\cdot)\|_{2+\alpha,\overline{B_1}}.$$

DEFINITION. A function $f: \overline{B_1}\times\mathbb{R}\to\mathbb{R}$ is said to belong to the class A_η^α $(0<\alpha<1, 0<\eta)$ if f satisfies the following conditions:

(i) For each $(x,y) \in \overline{B_1}$, $f(x,y,\cdot)$ extends to an entire function in $\mathbb{C}$.

(ii) For each $\zeta \in \mathbb{C}$, $f(\cdot,\zeta) \in C^{2+\alpha}(\overline{B_1})$.

(iii) For each $\varepsilon > 0$, there exists a constant $c_\varepsilon \geqq 0$ such that

$$\|f(\cdot,\zeta)\|_{2+\alpha,\overline{B_1}} \leqq c_\varepsilon\, e^{(\eta+\varepsilon)|\zeta|} \quad \forall\zeta\in\mathbb{C}. \tag{3.1}$$

(iv) $\partial_\nu f(x,y,\zeta) = 0$ for $(x,y) \in \partial B_1$, $\zeta \in \mathbb{C}$.

(Note that condition (iv) coincides with (2.2).) Examples of functions in A_η^α can be constructed as follows:

(1) Let $h \in C^{2+\alpha}(\overline{B_1})$ with $\partial_\nu h = 0$ on ∂B_1, and let $g(\zeta)$ be an entire function of order 1 and type η. Then $f(x,y,\zeta) = h(x,y)g(\zeta) \in A_\eta^\alpha$.

(2) Let $\varphi: \overline{B_1}\times\mathbb{R}\to\mathbb{R}$ satisfy:

(a) $\partial_\nu\varphi = 0$, $(x,y) \in \partial B_1$.

(b) $\|\varphi(\cdot,\mu)\|_{2+\alpha} \leqq M$ for $|\mu| \leqq \eta/(2\pi)$, for some constant $M \geqq 0$.

(c) $\varphi(x, y, \mu) = 0$ for $|\mu| > \eta/(2\pi)$.

Then, the function

$$f(x, y, \zeta) = \varphi(x, y, \cdot)\hat{}(\zeta) = \int_{\mathbb{R}} \varphi(x, y, \mu)\, e^{-2\pi i \mu \zeta}\, d\mu$$

belongs to A_η^α.

THEOREM 3.1. *Let $T > 0$, $0 < \alpha < 1$. Then, there exists a number $\eta_0 < 1$ such that, for every $\phi_0 \in A_{\eta_0}^\alpha$, there exists a unique solution (ϕ, p) of (2.13) satisfying:*

(i) $\phi \equiv \phi_0$ *at* $t = 0$.

(ii) *For each* $(x, y) \in \overline{B_1}$, $t \in [0, T]$, $\phi(t, x, y, \cdot)$, *and* $p(t, x, y, \cdot)$ *extend to entire functions in* $\mathbb{C}$.

(iii)

$$\phi(\cdot, \zeta) \in \bar{C}_{2+\alpha}([0, T] \times \overline{B_1}), \qquad \zeta \in \mathbb{C},$$

$$p(\cdot, \zeta) \in \bar{C}_\alpha([0, T] \times \overline{B_1}), \qquad \zeta \in \mathbb{C},$$

$$p(t, \cdot, \zeta) \in C^{2+\alpha}(\overline{B_1}), \qquad \zeta \in \mathbb{C}, \quad t \in [0, T].$$

(iv) *For every $\varepsilon > 0$, there exists $c_\varepsilon > 0$ such that*

$$\|\phi(\cdot, \zeta)\|_{2+\alpha} + |||p(\cdot, \zeta)||| \leqq c_\varepsilon\, e^{(\eta_0 + \varepsilon)|\zeta|} \quad \forall \zeta \in \mathbb{C}. \tag{3.2}$$

Remark. It will be evident from the proof of Theorem 3.1 that, for each α fixed, $\eta_0 \to 0$ as $T \to \infty$.

Note that $\phi_0(x, y, \cdot)$ is not necessarily a tempered distribution for $\phi_0 \in A_\eta^\alpha$. This prevents us from using the Paley-Wiener theorem (see § 6) in the proof of Theorem 3.1. The underlying approach for proving this theorem is by superposition: we expand the initial data into a power series in z, $\phi_0(x, y, z) = \sum_{n=0}^\infty a_n(x, y) z^n$, solve the problem with initial data $\phi_0^n(x, y, z) = a_n(x, y) z^n$, and then sum over n. In order to solve for ϕ_0^n we take the Fourier transform in the z variable, solve the resulting problem (using a fixed-point argument) and then take the inverse transform. Finally, using the growth condition for ϕ_0, we shall show that the resulting series actually converges to a solution of (2.13).

By Fourier transforming (2.13) in z and setting $\phi(t, x, y, \cdot)\hat{}(\mu) = u(t, x, y, \mu)$, $p(t, x, y, \cdot)\hat{}(\mu) = v(t, x, y, \mu)$, $\phi_0(x, y, \cdot)\hat{}(\mu) = u_0(x, y, \mu)$, we are led to the following system:

$$\frac{\partial u}{\partial t} - \Gamma^{-2} \Delta u + c(2\pi i \mu)^2 u + u = -\alpha^2 2\pi i \mu v \quad \text{in } G, \tag{3.3a}$$

$$\Delta v = 2\pi i \mu \frac{\partial u}{\partial t} - 2(2\pi i \mu) u - c(2\pi i \mu)^3 u \quad \text{in } G, \tag{3.3b}$$

$$\begin{aligned} \frac{\partial u}{\partial \nu} = &-\frac{3\alpha^2}{2\pi} (2\pi i \mu)^2 \int_0^t \int_{B_1} u(s, x, y, \mu)\, dx\, dy\, ds \\ &+ \frac{\alpha^2}{2\pi} (2\pi i \mu)^2 \int_{B_1} u(t, x, y, \mu)\, dx\, dy \\ &- \frac{\alpha^2}{2\pi} (2\pi i \mu)^2 \int_{B_1} u_0(x, y, \mu)\, dx\, dy \quad \text{on } \Gamma_0, \end{aligned} \tag{3.3c}$$

$$v=-\frac{3}{8\pi}(2\pi i\mu)\int_0^t\int_{B_1}u(s,x,y,\mu)\,dx\,dy\,ds-\Gamma^{-2}(2\pi i\mu)u$$

$$-\frac{(2\pi i\mu)}{8\pi}\int_{B_1}u_0(x,y,\mu)\,dx\,dy \tag{3.3d}$$

$$+\frac{\Gamma^{-2}}{\pi}(2\pi i\mu)\int_{B_1}u(t,x,y,\mu)\,dx\,dy\quad\text{on }\Gamma_0,$$

$$u(0,x,y,\mu)=u_0(x,y,\mu). \tag{3.3e}$$

Replacing $-iv$ by v we may rewrite (3.3) as

$$\frac{\partial u}{\partial t}-\Gamma^{-2}\Delta u+(1-4\pi^2c\mu^2)u=\alpha^2 2\pi\mu v\quad\text{in } G, \tag{3.4a}$$

$$\Delta v=2\pi\mu\frac{\partial u}{\partial t}+(8\pi^3c\mu^3-4\pi\mu)u\quad\text{in } G, \tag{3.4b}$$

$$\frac{\partial u}{\partial\nu}=6\pi\alpha^2\mu^2\int_0^t\int_{B_1}u-2\pi\alpha^2\mu^2\int_{B_1}u+2\pi\alpha^2\mu^2\int_{B_1}u_0\quad\text{on }\Gamma_0, \tag{3.4c}$$

$$v=-\frac{3}{4}\mu\int_0^t\int_{B_1}u-\Gamma^{-2}2\pi\mu u-\frac{\mu}{4}\int_{B_1}u_0+2\Gamma^{-2}\mu\int_{B_1}u\quad\text{on }\Gamma_0, \tag{3.4d}$$

$$u(0,x,y,\mu)=u_0(x,y,\mu). \tag{3.4e}$$

Note that, for fixed μ, (3.4) is an "elliptic-parabolic" system. In the next section we prove existence and uniqueness of a solution for the special case where u_0 is independent of μ (which constitutes the main step towards proving Theorem 3.1).

4. Proof of Theorem 3.1 in a special case.

LEMMA 4.1. *Let* $u_0:\overline{B_1}\to\mathbb{R}$ *satisfy*

$$u_0\in C^{2+\alpha}(\overline{B_1}), \tag{4.1}$$

$$\frac{\partial u_0}{\partial\nu}=0\quad\text{on }\partial B_1. \tag{4.2}$$

Then, there exists a unique solution of (3.4) *satisfying*: $u(t,x,y,\cdot)$, $v(t,x,y,\cdot)\in C^\infty(\mathbb{R})$, $\partial_\mu^l u(\cdot,\mu)\in\bar{C}_{2+\alpha}([0,T]\times\overline{B_1})$, $\partial_\mu^l v(\cdot,\mu)\in\bar{C}_\alpha([0,T]\times\overline{B_1})$, $\partial_\mu^l v(t,\cdot,\mu)\in C^{2+\alpha}(\overline{B_1})$, *and* $\partial_t\partial_\mu^l v(t,\cdot,\mu)\in C^{2+\alpha}(B_1)\cap C^0(\overline{B_1})$ $(l\geqq 0)$.

Furthermore, for each $l\geqq 0$, *we have*

$$\left\|\frac{\partial^l u}{\partial\mu^l}(\cdot,0)\right\|_{2+\alpha}\leqq c_1c_2^l l!\,\|u_0\|^0_{2+\alpha}, \tag{4.3}$$

$$\left|\!\left|\!\left|\frac{\partial^l v}{\partial\mu^l}(\cdot,0)\right|\!\right|\!\right|\leqq c_1c_2^l l!\,\|u_0\|^0_{2+\alpha} \tag{4.4}$$

for some constants c_1, $c_2>0$.

Assume for a moment that the conclusions of the lemma hold, and let (u,v) be the solution of (3.4) for u_0 independent of μ. Then, if we set

$$\phi(t,x,y,z)=\langle(-2\pi i)^{-n}\partial_\mu^n\delta,\,u(t,x,y,\mu)\,e^{2\pi i\mu z}\rangle$$

$$=\frac{1}{(2\pi i)^n}\sum_{k=0}^n\binom{n}{k}(2\pi iz)^k\partial_\mu^{n-k}u(t,x,y,0)$$

and

$$p(t,x,y,z)=\langle(-2\pi i)^{-n}\partial_\mu^n\delta,\ iv(t,x,y,\mu)\,e^{2\pi i\mu z}\rangle$$
$$=\frac{i}{(2\pi i)^n}\sum_{k=0}^{n}\binom{n}{k}(2\pi iz)^k\partial_\mu^{n-k}v(t,x,y,0),$$

where δ is the Dirac delta acting on the variable μ, it is easy to check that (ϕ, p) is a solution of (2.13) satisfying (i)-(iv) of Theorem 3.1 with $\phi_0(x,y,z)=u_0(x,y)z^n$.

Proof of Lemma 4.1. We shall use c_μ to denote a generic constant (not always the same), depending only on μ. We divide the proof into five steps.

Step I (Transformation of the system). First we transform (3.4) into a more convenient form. Using (3.4a), we may replace (3.4b) by

$$\Delta v-4\pi^2\mu^2\alpha^2v=2\pi\mu\Gamma^{-2}\Delta u+2\pi\mu(8\pi^2c\mu^2-3)u$$

and setting $w=v-2\pi\mu\Gamma^{-2}u$, we are led to

$$\frac{\partial u}{\partial t}-\Gamma^{-2}\Delta u+(1-8\pi^2c\mu^2)u=\alpha^2 2\pi\mu w \quad \text{in } G, \tag{4.5a}$$

$$\Delta w-4\pi^2\alpha^2\mu^2w=(24\pi^3c\mu^3-6\pi\mu)u \quad \text{in } G, \tag{4.5b}$$

with initial-boundary conditions given by

$$\frac{\partial u}{\partial\nu}=6\pi\alpha^2\mu^2\int_0^t\int_{B_1}u-2\pi\alpha^2\mu^2\int_{B_1}u+2\pi\alpha^2\mu^2\int_{B_1}u_0 \quad \text{on } [0,T]\times\partial B_1\times\mathbb{R}, \tag{4.5c}$$

$$w=-\frac{3}{4}\mu\int_0^t\int_{B_1}u-\Gamma^{-2}4\pi\mu u-\frac{\mu}{4}\int_{B_1}u_0+2\Gamma^{-2}\mu\int_{B_1}u \quad \text{on } [0,T]\times\partial B_1\times\mathbb{R}, \tag{4.5d}$$

$$u(0,x,y,\mu)=u_0(x,y) \quad \text{on } B_1. \tag{4.5e}$$

Finally, we want to write the right-hand side of (4.5c) in terms of u_0 and w.

Integrating (4.5a) over B_1 we get

$$f'(t)+(1-8\pi^2c\mu^2)f(t)-\Gamma^{-2}2\pi\frac{\partial u}{\partial\nu}=g(t), \tag{4.6}$$

where

$$f(t)=\int_{B_1}u\,dx\,dy, \qquad g(t)=2\pi\alpha^2\mu\int_{B_1}w\,dx\,dy,$$

and using (4.5c) we can replace $\partial_\nu u$ in (4.6), to get

$$f'(t)+(1-4\pi^2c\mu^2)f(t)-12\pi^2c\mu^2\int_0^t f(s)\,ds-4\pi^2c\mu^2\int_{B_1}u_0=g(t). \tag{4.7}$$

Since $f(0)=\int_{B_1}u_0(x,y)\,dx\,dy$, we may solve (4.7) for f in terms of u_0 and w. Replacing in (4.5c), we see that, on ∂B_1,

$$\frac{\partial u}{\partial\nu}=E_2(u_0,w),$$

where $E_2(u_0,w)$ is an expression in (u_0,w) (which can be explicitly written) depending only on t and with the property that, if $w\in\bar{C}_\alpha([0,T]\times\overline{B_1})$ and $u_0\in C^{2+\alpha}(\overline{B_1})$, we have $\partial_tE_2(u_0,w)\in\bar{C}_\alpha([0,T]\times\overline{B_1})$.

Thus, if we set

$$E_1(u_0, u) = -\frac{3}{4}\mu \int_0^t \int_{B_1} u - \Gamma^{-2} 4\pi\mu u - \frac{\mu}{4} \int_{B_1} u_0 + 2\Gamma^{-2}\mu \int_{B_1} u,$$

then we need to study system (4.5a), (4.5b) subject to the initial-boundary conditions

$$\frac{\partial u}{\partial \nu} = E_2(u_0, w) \quad \text{on } \partial B_1, \tag{4.8a}$$

$$w = E_1(u_0, u) \quad \text{on } \partial B_1, \tag{4.8b}$$

$$u(0, x, y, \mu) = u_0(x, y) \quad \text{on } B_1. \tag{4.8c}$$

Step II (Local existence). We now show that there exists a solution of (4.5a), (4.5b), (4.8) if T is small enough.

Fix $\mu \in \mathbb{R}$. Given $F \in \bar{C}_{2+\alpha}$ such that $F(0, x, y) = u_0(x, y)$, let $\mathcal{R}F = L$ be the unique solution of

$$\Delta L - 4\pi^2\alpha^2\mu^2 L = (24\pi^3 c\mu^3 - 6\pi\mu)F \quad \text{on } [0, T] \times B_1, \tag{4.9a}$$

$$L = E_1(u_0, F) \quad \text{on } [0, T] \times \partial B_1. \tag{4.9b}$$

It is then clear that $L = \mathcal{R}F$ satisfies:

$$L(t, \cdot) \in C^{2+\alpha}(\overline{B_1}), \quad \partial_t L \in C^{2+\alpha}(B_1), \quad \text{and} \quad L \in \bar{C}_\alpha.$$

Furthermore, if $F_1, F_2 \in \bar{C}_{2+\alpha}$, $F_1(0, x, y) = F_2(0, x, y) = u_0(x, y)$, and $L_i = \mathcal{R}F_i$ $(i = 1, 2)$, then, for some constant c_μ we have

$$\|L_1 - L_2\|_\alpha \leqq c_\mu \|F_1 - F_2\|_E, \tag{4.10}$$

where

$$\|h\|_E = \|h\|_{\infty, D} + \left\|\frac{\partial h}{\partial x}\right\|_{\infty, D} + \left\|\frac{\partial h}{\partial y}\right\|_{\infty, D} + \left\|\frac{\partial^2 h}{\partial x^2}\right\|_{\infty, D} + \left\|\frac{\partial^2 h}{\partial y^2}\right\|_{\infty, D} + \left\|\frac{\partial^2 h}{\partial x\, \partial y}\right\|_{\infty, D} + \left\|\frac{\partial h}{\partial t}\right\|_{\infty, D}$$

and $D = [0, T] \times \overline{B_1}$.

Indeed, this is a straightforward consequence of the maximum principle and the following result.

LEMMA 4.2. *Let* $U, W \in C^{2+\alpha}(\overline{B_1})$, $V \in C^{1+\alpha}(\overline{B_1})$ *satisfy*

$$\Delta U - \lambda^2 U = V \quad \text{on } B_1 \quad (B_1 \subset \mathbb{R}^2),$$

$$U = W \quad \text{on } \partial B_1.$$

Then

$$\|\nabla U\|_0 \leqq 6[\|V\|_1 + \lambda^2 \|W\|_0 + \|W\|_2],$$

where $\|h\|_k = \sum_{|\alpha| \leqq k} \|\partial^\alpha h / \partial x^{\alpha_1}\, \partial y^{\alpha_2}\|_{\infty, B_1}$.

(See, e.g., Gilbarg and Trudinger [7, p. 48]).

Now, given $L \in \bar{C}_\alpha$ let $\mathcal{S}L = F$ be the unique solution of

$$\frac{\partial F}{\partial t} - \Gamma^{-2}\Delta F + (1 - 8\pi^2 c\mu^2)F = \alpha^2 2\pi\mu L \quad \text{on } [0, T] \times B_1, \tag{4.11a}$$

$$\frac{\partial F}{\partial \nu} = E_2(u_0, L) \quad \text{on } [0, T] \times \partial B_1, \tag{4.11b}$$

$$F(0, x, y) = u_0(x, y) \quad \text{on } B_1. \tag{4.11c}$$

Note that, from the hypotheses on u_0 ((4.1) and (4.2)), the compatibility condition

$$\frac{\partial u_0}{\partial \nu} = E_2(u_0, L) \quad \text{on } \partial B_1 \quad \text{at } t = 0 \tag{4.12}$$

is satisfied. Thus, $F = \mathcal{S}L \in \bar{C}_{2+\alpha}$ and the parabolic Schauder estimates (see, e.g., Ladyženskaja, Solonnikov, and Ural'ceva [10, pp. 320-321]) imply that:

If L_1, $L_2 \in \bar{C}_\alpha$, $F_i = \mathcal{S}L_i$ $(i = 1, 2)$, then

$$\|F_1 - F_2\|_{2+\alpha} \leqq c_\mu \|L_1 - L_2\|_\alpha. \tag{4.13}$$

Let $\mathcal{Q} = \mathcal{S}\mathcal{R}$; then, combining the above results, we obtain

$$\|\mathcal{Q}(F_1) - \mathcal{Q}(F_2)\|_{2+\alpha} \leqq c_\mu T^{\alpha/2} \|F_1 - F_2\|_{2+\alpha} \tag{4.14}$$

for every F_1, $F_2 \in \bar{C}_{2+\alpha}$ such that $F_i(0, x, y) = u_0(x, y)$ $(i = 1, 2)$. (Note that $(\mathcal{R}F_1 - \mathcal{R}F_2)(0, x, y) \equiv 0$.)

Let $X = \{F \in \bar{C}_{2+\alpha} \mid F(0, x, y) = u_0(x, y)\} \subset \bar{C}_{2+\alpha}$. Then, choosing $T = T_0 = T_0(\mu)$ small enough, we may apply the Contraction Mapping Theorem to conclude that there exists a unique $F_1 \in X$ such that $\mathcal{Q}(F_1) = F_1$. Hence, if we set $L_1 = \mathcal{R}F_1$, then (F_1, L_1) satisfy:

$$\frac{\partial F_1}{\partial t} - \Gamma^{-2}\Delta F_1 + (1 - 8\pi^2 c\mu^2)F_1 = \alpha^2 2\pi\mu L_1 \quad \text{on } [0, T_0] \times B_1, \tag{4.15a}$$

$$\Delta L_1 - 4\pi^2\alpha^2\mu^2 L_1 = (24\pi^3 c\mu^3 - 6\pi\mu)F_1 \quad \text{on } [0, T_0] \times B_1 \tag{4.15b}$$

and

$$\frac{\partial F_1}{\partial \nu} = E_2(u_0, L_1) \quad \text{on } [0, T_0] \times \partial B_1, \tag{4.15c}$$

$$L_1 = E_1(u_0, F_1) \quad \text{on } [0, T_0] \times \partial B_1, \tag{4.15d}$$

$$F_1(0, x, y) = u_0(x, y) \quad \text{on } B_1, \tag{4.15e}$$

with $F_1 \in \bar{C}_{2+\alpha}$, $L_1 \in \bar{C}_\alpha$, $L_1(t, \cdot) \in C^{2+\alpha}(\overline{B_1})$, and $\partial_t L_1(t, \cdot) \in C^{2+\alpha}(B_1) \cap C^0(\overline{B_1})$.

We now want to show that

$$\|F_1\|_{2+\alpha} \leqq c_\mu \|u_0\|^0_{2+\alpha}, \tag{4.16}$$

$$\sup_{t \in [0, T_0]} \|L_1(t, \cdot)\|^0_{2+\alpha} \leqq c_\mu \|u_0\|^0_{2+\alpha}, \tag{4.17}$$

and

$$\left\| \frac{\partial L_1}{\partial t} \right\|_0 = \left\| \frac{\partial L_1}{\partial t} \right\|_{\infty, [0, T_0] \times B_1} \leqq c_\mu \|u_0\|^0_{2+\alpha}. \tag{4.18}$$

Using the elliptic Schauder estimates and the maximum principle (in (4.9)), (4.17) and (4.18) easily follow from (4.16). Finally, (4.16) (for $T_0(\mu)$ small enough) follows from (4.15), Lemma 4.2, and the parabolic Schauder estimates.

Step III (Global existence and uniqueness). In order to prove global existence of a solution on $[0, T] \times B_1$ we just need to write

$$u_1(x, y) = u_0(x, y) + 3 \int_0^{T_0} F_1(s, x, y)\, ds$$

and repeat the above argument for the system

$$\frac{\partial F}{\partial t} - \Gamma^{-2}\Delta F + (1 - 8\pi^2 c\mu^2)F = \alpha^2 2\pi\mu L \quad \text{on } [T_0, 2T_0] \times B_1, \tag{4.19a}$$

$$\Delta L - 4\pi^2\alpha^2\mu^2 L = (24\pi^3 c\mu^3 - 6\pi\mu)F \quad \text{on } [T_0, 2T_0] \times B_1 \tag{4.19b}$$

with initial boundary conditions

$$(4.19c)\quad \frac{\partial F}{\partial \nu}=6\pi\alpha^2\mu^2\int_{T_0}^{t}\int_{B_1}F-2\pi\alpha^2\mu^2\int_{B_1}F+2\pi\alpha^2\mu^2\int_{B_1}u_1 \quad \text{for } (x,y)\in\partial B_1,$$

$$(4.19d)\quad L=-\frac{3}{4}\mu\int_{T_0}^{t}\int_{B_1}F-\Gamma^{-2}4\pi\mu F-\frac{\mu}{4}\int_{B_1}u_1+2\Gamma^{-2}\mu\int_{B_1}F \quad \text{for } (x,y)\in\partial B_1,$$

$$(4.19e)\quad F(T_0,x,y)=F_1(T_0,x,y) \quad \text{on } B_1.$$

(Here we use the fact that the size of the time interval for local existence depends only on μ, not on the initial data.)

Thus, we have proved that there exists a solution (u, v) of (3.4) satisfying:

$$u(\cdot,\mu)\in\bar{C}_{2+\alpha}([0,T]\times\overline{B_1}),\qquad v(\cdot,\mu)\in\bar{C}_{\alpha}([0,T]\times\overline{B_1}),$$
$$v(t,\cdot,\mu)\in C^{2+\alpha}(\overline{B_1}),\qquad \partial_t v(t,\cdot,\mu)\in C^{2+\alpha}(B_1)\cap C^0(\overline{B_1}).$$

Also,

$$(4.20)\quad \|u(\cdot,\mu)\|_{2+\alpha}+|||v(\cdot,\mu)|||\leqq c_\mu\|u_0\|^0_{2+\alpha}.$$

The uniqueness of the solution (u, v) easily follows from (4.5), (4.8) and the uniqueness statement in the Contraction Mapping Theorem.

Step IV (Regularity of the solution). We want to show:

(a) $u(t,x,y,\cdot)$, $v(t,x,y,\cdot)\in C^\infty(\mathbb{R})$.

(b) $\partial^l_\mu u(\cdot,\mu)\in\bar{C}_{2+\alpha}([0,T]\times\overline{B_1})$, $\partial^l_\mu v(\cdot,\mu)\in\bar{C}_\alpha([0,T]\times\overline{B_1})$, $\partial^l_\mu v(t,\cdot,\mu)\in C^{2+\alpha}(\overline{B_1})$, $\partial_t\partial^l_\mu v(t,\cdot,\mu)\in C^{2+\alpha}(B_1)\cap C^0(\overline{B_1})$ for each $l\geqq 0$.

Fix $\mu_0\in\mathbb{R}$ and let, for $h\in\mathbb{R}$,

$$U_0(x,y,t)=u(t,x,y,\mu_0),\qquad U_h(x,y,t)=u(t,x,y,\mu_0+h),$$
$$V_0(x,y,t)=v(t,x,y,\mu_0),\qquad V_h(x,y,t)=v(t,x,y,\mu_0+h).$$

Then, using (3.4) we can write a system of equations for

$$\left(\frac{U_h-U_0}{h}\right)\quad\text{and}\quad\left(\frac{V_h-V_0}{h}\right),$$

and, proceeding as before, we can prove, with the aid of (4.20), that

$$\left\|\left(\frac{U_h-U_0}{h}\right)\right\|_{2+\alpha}+\left|\left|\left|\left(\frac{V_h-V_0}{h}\right)\right|\right|\right|\leqq c_{\mu_0}(\|U_h\|_{2+\alpha}+|||V_h|||+\|u_0\|^0_{2+\alpha})$$
$$\leqq c_{\mu_0}\|u_0\|^0_{2+\alpha}$$

for small h. Thus we may conclude that $u(t,x,y,\cdot)$ and $v(t,x,y,\cdot)$ are differentiable at μ_0 and that $\partial^l_\mu u(\cdot,\mu_0)$, $\partial^l_\mu v(\cdot,\mu_0)$ satisfy (b) for $l=1$.

Furthermore, we have

$$(4.21a)\quad \frac{\partial}{\partial t}\frac{\partial u}{\partial\mu}-\Gamma^{-2}\Delta\frac{\partial u}{\partial\mu}+(1-4\pi^2c\mu^2)\frac{\partial u}{\partial\mu}$$
$$=\alpha^2 2\pi\mu\frac{\partial v}{\partial\mu}+(8\pi^2c\mu u+2\pi\alpha^2 v)\quad\text{on } [0,T]\times B_1\times\mathbb{R},$$

$$(4.21b)\quad \Delta\frac{\partial v}{\partial\mu}=2\pi\mu\frac{\partial}{\partial t}\frac{\partial u}{\partial\mu}+(8\pi^3c\mu^3-4\pi\mu)\frac{\partial u}{\partial\mu}$$
$$+\left[(24\pi^3c\mu^2-4\pi)u+2\pi\frac{\partial u}{\partial t}\right]\quad\text{on } [0,T]\times B_1\times\mathbb{R},$$

$$\frac{\partial}{\partial\nu}\frac{\partial u}{\partial\mu}=6\pi\alpha^2\mu^2\int_0^t\int_{B_1}\frac{\partial u}{\partial\mu}-2\pi\alpha^2\mu^2\int_{B_1}\frac{\partial u}{\partial\mu}$$

$$+\left\{12\pi\alpha^2\mu\int_0^t\int_{B_1}u-4\pi\alpha^2\mu\int_{B_1}u+4\pi\alpha^2\mu\int_{B_1}u_0\right\} \tag{4.21c}$$

on $[0, T]\times\partial B_1\times\mathbb{R}$,

$$\frac{\partial v}{\partial\mu}=-\frac{3}{4}\mu\int_0^t\int_{B_1}\frac{\partial u}{\partial\mu}-\Gamma^{-2}2\pi\mu\frac{\partial u}{\partial\mu}+2\Gamma^{-2}\mu\int_{B_1}\frac{\partial u}{\partial\mu}$$

$$+\left\{-\frac{1}{4}\int_{B_1}u_0-\frac{3}{4}\int_0^t\int_{B_1}u-2\pi\Gamma^{-2}u+2\Gamma^{-2}\int_{B_1}u\right\} \tag{4.21d}$$

on $[0, T]\times\partial B_1\times\mathbb{R}$,

$$\frac{\partial u}{\partial\mu}(0, x, y, \mu)=0. \tag{4.21e}$$

By repeating this procedure we may conclude that (a) and (b) hold.

Step V (Bounds on the μ-derivatives at $\mu=0$). Consider the system (4.5) and let (for $l\geqq 0$)

$$F_l(t, x, y)=\frac{1}{l!}\frac{\partial^l u}{\partial\mu^l}(t, x, y, 0),$$

$$L_l(t, x, y)=\frac{1}{l!}\frac{\partial^l w}{\partial\mu^l}(t, x, y, 0),$$

and $F_l\equiv L_l\equiv 0$ if $l<0$. It is then clear that, for $l\geqq 0$,

$$\frac{\partial F_l}{\partial t}-\Gamma^{-2}\Delta F_l+F_l=\alpha^2 2\pi L_{l-1}+8\pi^2 cF_{l-2} \quad \text{on } [0, T]\times B_1, \tag{4.22a}$$

$$\Delta L_l=4\pi^2\alpha^2 L_{l-2}+24\pi^3 cF_{l-3}-6\pi F_{l-1} \quad \text{on } [0, T]\times B_1, \tag{4.22b}$$

$$\frac{\partial F_l}{\partial\nu}=6\pi\alpha^2\int_0^t\int_{B_1}F_{l-2}-2\pi\alpha^2\int_{B_1}F_{l-2}+2\pi\alpha^2\int_{B_1}F_{l-2}^0 \quad \text{on } [0, T]\times\partial B_1, \tag{4.22c}$$

$$L_l=-\frac{3}{4}\int_0^t\int_{B_1}F_{l-1}-\Gamma^{-2}4\pi F_{l-1}-\frac{1}{4}\int_{B_1}F_{l-1}^0+2\Gamma^{-2}\int_{B_1}F_{l-1} \quad \text{on } [0, T]\times\partial B_1, \tag{4.22d}$$

$$F_l(0, x, y)=F_l^0(x, y) \quad \text{on } B_1, \tag{4.22e}$$

where

$$F_l^0(x, y)=\begin{cases}u_0(x, y) & \text{if } l=0,\\ 0 & \text{if } l\neq 0.\end{cases} \tag{4.23}$$

First we note that (4.22) implies that

$$F_{2k+1}\equiv L_{2k}\equiv 0 \qquad (k\geqq 0).$$

On the other hand, for $l\geqq 0$,

$$\|F_l\|_{2+\alpha}\leqq\gamma_1(\|L_{l-1}\|_\alpha+\|F_{l-2}\|_{2+\alpha}+\|F_l^0\|_{2+\alpha}^0+\|F_{l-2}^0\|_{2+\alpha}^0), \tag{4.24a}$$

$$|||L_{l+1}|||\leqq\gamma_2(|||L_{l-1}|||+\|F_{l-2}\|_{2+\alpha}+\|F_l\|_{2+\alpha}+\|F_l^0\|_{2+\alpha}^0) \tag{4.24b}$$

for some constants γ_1, γ_2.

But then, from (4.24) we may conclude that

$$|||L_{l+1}||| \leqq \gamma_2(|||L_{l-1}||| + \|F_{l-2}\|_{2+\alpha} + \|F_l^0\|_{2+\alpha}^0) + \gamma_1\gamma_2(|||L_{l-1}||| + \|F_{l-2}\|_{2+\alpha} + \|F_l^0\|_{2+\alpha}^0 + \|F_{l-2}\|_{2+\alpha}^0)$$

so that

$$|||L_{l+1}||| \leqq \gamma_3(|||L_{l-1}||| + \|F_{l-2}\|_{2+\alpha} + \|F_l^0\|_{2+\alpha}^0 + \|F_{l-2}\|_{2+\alpha}^0) \tag{4.25}$$

for some constant γ_3.

Let $a_l = \|F_l\|_{2+\alpha} + |||L_{l+1}|||$; then, from (4.24a) and (4.25) we get

$$a_l \leqq \gamma(a_{l-2} + \|F_l^0\|_{2+\alpha}^0 + \|F_{l-2}^0\|_{2+\alpha}^0)$$

for some constant γ.

In particular, for $k \geqq 2$

$$a_{2k} \leqq \gamma a_{2(k-1)}$$

so that

$$a_{2k} \leqq \gamma^{k-1} a_2. \tag{4.26}$$

But,

$$\begin{aligned} a_2 &\leqq \gamma(a_0 + \|F_0^0\|_{2+\alpha}^0) \\ &\leqq \gamma(\gamma\|F_0^0\|_{2+\alpha}^0 + \|F_0^0\|_{2+\alpha}^0) \\ &\leqq (\gamma+\gamma^2)\|u_0\|_{2+\alpha}^0. \end{aligned}$$

Hence, there exists a constant γ_0 such that

$$a_l \leqq \gamma_0^l \|u_0\|_{2+\alpha}^0. \tag{4.27}$$

Finally, since $v = w + \Gamma^{-2} 2\pi\mu u$, we conclude from (4.27) that (4.3) and (4.4) hold.

Thus, the proof of Lemma 4.1 is complete. □

5. Proof of Theorem 3.1.

5.1. Existence. Write $\phi_0(x, y, \zeta) = \sum_{n=0}^{\infty} a_n(x, y)\zeta^n$. Then, using the fact that $\phi_0 \in A_{\eta_0}^{\alpha}$ and Cauchy's formula, we get that $a_n \in \bar{C}_{2+\alpha}(\overline{B_1})$ and, for every $\varepsilon > 0$

$$\|a_n\|_{2+\alpha}^0 \leqq c_\varepsilon \frac{(\eta_0+\varepsilon)^n e^n}{n^n} \quad \text{for } n \geqq 1. \tag{5.1}$$

Now let $(A_n(t, x, y, \mu), B_n(t, x, y, \mu))$ be the solution of (3.4) given by Lemma 4.1 when u_0 is replaced by a_n.

Finally, set

$$\phi(t, x, y, \zeta) = \sum_{n=0}^{\infty} \frac{1}{(2\pi i)^n} \sum_{k=0}^{n} \binom{n}{k} (2\pi i\zeta)^k \frac{\partial^{n-k} A_n}{\partial \mu^{n-k}}(t, x, y, 0), \tag{5.2a}$$

$$p(t, x, y, \zeta) = \sum_{n=0}^{\infty} \frac{i}{(2\pi i)^n} \sum_{k=0}^{n} \binom{n}{k} (2\pi i\zeta)^k \frac{\partial^{n-k} B_n}{\partial \mu^{n-k}}(t, x, y, 0). \tag{5.2b}$$

We claim:

If η_0 is small enough, then (ϕ, p) defined by (5.2) is a solution of (2.13) satisfying (i)-(iv) of Theorem 3.1.

Proof. We have

$$\|\phi(\cdot, \zeta)\|_{2+\alpha} \leqq \sum_{n=0}^{\infty} \frac{1}{(2\pi)^n} \sum_{k=0}^{n} \binom{n}{k} (2\pi|\zeta|)^k \left\| \frac{\partial^{n-k} A_n}{\partial \mu^{n-k}}(\cdot, 0) \right\|_{2+\alpha}.$$

But from (4.3) we know that

$$\left\| \frac{\partial^{n-k} A_n}{\partial \mu^{n-k}}(\cdot, 0) \right\|_{2+\alpha} \leqq c_1 c_2^{n-k} (n-k)! \|a_n\|^0_{2+\alpha}$$

so that

$$\|\phi(\cdot, \zeta)\|_{2+\alpha} \leqq c_1 \sum_{n=0}^{\infty} n! \sum_{k=0}^{n} \frac{|\zeta|^k}{k!} c_2^{n-k} \|a_n\|^0_{2+\alpha}. \tag{5.3}$$

Thus, using (5.1),

$$\|\phi(\cdot, \zeta)\|_{2+\alpha} \leqq c_1 c_\varepsilon \left(\sum_{n=1}^{\infty} n! \sum_{k=1}^{n} \frac{|\zeta|^k}{k!} c_2^{n-k} \frac{(\eta_0+\varepsilon)^n e^n}{n^n} + \sum_{n=1}^{\infty} n! c_2^n \frac{(\eta_0+\varepsilon)^n e^n}{n^n} + 1 \right),$$

and changing the order of summation,

$$\|\phi(\cdot, \zeta)\|_{2+\alpha} \leqq c_1 c_\varepsilon \left(\sum_{k=1}^{\infty} \frac{[(\eta_0+\varepsilon)|\zeta|]^k}{k!} \sum_{n \geqq k} (c_2(\eta_0+\varepsilon))^{n-k} \frac{n!\, e^n}{n^n} + \sum_{n=1}^{\infty} (c_2(\eta_0+\varepsilon))^n \frac{n!\, e^n}{n^n} + 1 \right). \tag{5.4}$$

By using the estimate

$$\frac{n!\, e^n}{n^n} \leqq c_3 n^{1/2}, \tag{5.5}$$

we get

$$\|\phi(\cdot, \zeta)\|_{2+\alpha} \leqq c_1 c_3 c_\varepsilon \left(\sum_{k=1}^{\infty} \frac{[(\eta_0+\varepsilon)|\zeta|]^k}{k!} \sum_{n \geqq k} (c_2(\eta_0+\varepsilon))^{n-k} n^{1/2} + \sum_{n=1}^{\infty} (c_2(\eta_0+\varepsilon))^n n^{1/2} + 1 \right),$$

that is,

$$\|\phi(\cdot, \zeta)\|_{2+\alpha} \leqq \tilde{c}_\varepsilon \left(\sum_{k=1}^{\infty} \frac{[(\eta_0+\varepsilon)|\zeta|]^k}{k!} \sum_{n \geqq 0} (c_2(\eta_0+\varepsilon))^n (n+k)^{1/2} + \sum_{n=1}^{\infty} (c_2(\eta_0+\varepsilon))^n n^{1/2} + 1 \right) \qquad (\tilde{c}_\varepsilon = c_1 c_3 c_\varepsilon). \tag{5.6}$$

Therefore, if $\eta_0 < 1/c_2$ and $0 < \varepsilon < 1/c_2 - \eta_0$, then (5.6) implies

$$\|\phi(\cdot, \zeta)\|_{2+\alpha} \leqq \hat{c}_\varepsilon (1+|\zeta|)\, e^{(\eta_0+\varepsilon)|\zeta|}, \tag{5.7a}$$

for some constant $\hat{c}_\varepsilon$.

In a similar way, but using (4.4) instead of (4.3), we can prove that

$$|||p(\cdot, \zeta)||| \leqq \tilde{\tilde{c}}_\varepsilon (1+|\zeta|)\, e^{(\eta_0+\varepsilon)|\zeta|}. \tag{5.7b}$$

Since (5.7) clearly implies (3.2) we see that (ϕ, p) satisfies (ii)-(iv) of Theorem 3.1, and it only remains to show that (ϕ, p) is a solution of (2.13).

First we check (2.13e); we have

$$\phi(0,x,y,z)=\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}\sum_{k=0}^{n}\binom{n}{k}(2\pi iz)^k\frac{\partial^{n-k}a_n}{\mu^{n-k}}(x,y)$$

$$=\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}(2\pi iz)^n a_n(x,y)=\phi_0(x,y,z).$$

Now set

$$A_{n,k}=\partial_\mu^{n-k}A_n(\cdot,0),\qquad B_{n,k}=\partial_\mu^{n-k}B_n(\cdot,0).$$

Then from (3.4) we have

$$\frac{\partial A_{n,k}}{\partial t}-\Gamma^{-2}\Delta A_{n,k}+A_{n,k}=4\pi^2c(n-k)(n-k-1)A_{n,k+2}$$
$$+2\pi\alpha^2(n-k)B_{n,k+1}\quad\text{on }[0,T]\times B_1, \tag{5.8a}$$

$$\Delta B_{n,k}=2\pi(n-k)\frac{\partial A_{n,k+1}}{\partial t}+8\pi^3c(n-k)(n-k-1)(n-k-2)A_{n,k+3}$$
$$-4\pi(n-k)A_{n,k+1}\quad\text{on }[0,T]\times B_1 \tag{5.8b}$$

and, on $[0,T]\times\partial B_1$,

$$\frac{\partial A_{n,k}}{\partial\nu}=6\pi\alpha^2(n-k)(n-k-1)\int_0^t\int_{B_1}A_{n,k+2}$$
$$-2\pi\alpha^2(n-k)(n-k-1)\int_{B_1}A_{n,k+2}+2\pi\alpha^2(n-k)(n-k-1)\delta_{k,n-2}\int_{B_1}a_n, \tag{5.8c}$$

$$B_{n,k}=-\frac{3}{4}(n-k)\int_0^t\int_{B_1}A_{n,k+1}-\frac{1}{4}\delta_{k,n-1}\int_{B_1}a_n$$
$$-\Gamma^{-2}2\pi(n-k)A_{n,k+1}+2\Gamma^{-2}(n-k)\int_{B_1}A_{n,k+1}. \tag{5.8d}$$

Using (5.8a),

$$\frac{\partial\phi}{\partial t}-\Gamma^{-2}\Delta\phi+\phi$$

$$=\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}\sum_{k=0}^{n}\binom{n}{k}(2\pi iz)^k\left[\frac{\partial A_{n,k}}{\partial t}-\Gamma^{-2}\Delta A_{n,k}+A_{n,k}\right]$$

$$=\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}\sum_{k=0}^{n}\binom{n}{k}(2\pi iz)^k[4\pi^2c(n-k)(n-k-1)A_{n,k+2}+2\pi\alpha^2(n-k)B_{n,k+1}]$$

$$=-c\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}\sum_{k=0}^{n-2}\binom{n}{k}(2\pi i)^{k+2}(n-k)(n-k-1)A_{n,k+2}z^k$$
$$-\sum_{n=0}^{\infty}\frac{i}{(2\pi i)^n}\sum_{k=0}^{n-1}\binom{n}{k}(2\pi i)^{k+1}\alpha^2(n-k)B_{n,k+1}z^k$$

$$=-c\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}\sum_{j=2}^{n}\binom{n}{j-2}(2\pi i)^j(n-j+2)(n-j+1)A_{n,j}z^{j-2}$$
$$-\alpha^2\sum_{n=0}^{\infty}\frac{i}{(2\pi i)^n}\sum_{j=1}^{n}\binom{n}{j-1}(2\pi i)^j(n-j+1)B_{n,j}z^{j-1}$$

so that

$$\frac{\partial\phi}{\partial t}-\Gamma^{-2}\Delta\phi+\phi=-c\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}\sum_{j=2}^{n}\binom{n}{j}(2\pi i)^j j(j-1)z^{j-2}A_{n,j}$$
$$-\alpha^2\sum_{n=0}^{\infty}\frac{i}{(2\pi i)^n}\sum_{j=1}^{n}\binom{n}{j}(2\pi i)^j jz^{j-1}B_{n,j}$$
$$=-c\frac{\partial^2\phi}{\partial z^2}-\alpha^2\frac{\partial p}{\partial z}$$

and therefore (2.13a) is satisfied.

Similarly, using (5.8d) we see that, on Γ_0,

$$p=\sum_{n=0}^{\infty}\frac{i}{(2\pi i)^n}\sum_{k=0}^{n}\binom{n}{k}(2\pi iz)^k B_{n,k}$$
$$=\sum_{n=0}^{\infty}\frac{i}{(2\pi i)^n}\sum_{k=0}^{n}\binom{n}{k}(2\pi iz)^k\left[-\frac{3}{4}(n-k)\int_0^t\int_{B_1}A_{n,k+1}-\frac{1}{4}\delta_{k,n-1}\right.$$
$$\left.\cdot\int_{B_1}a_n-\Gamma^{-2}2\pi(n-k)A_{n,k+1}+2\Gamma^{-2}(n-k)\int_{B_1}A_{n,k+1}\right].$$

Hence,

$$p=-\frac{3}{8\pi}\int_0^t\int_{B_1}\left(\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}\sum_{k=0}^{n-1}\binom{n}{k}(2\pi i)^{k+1}z^k(n-k)A_{n,k+1}\right)$$
$$-\frac{1}{8\pi}\int_{B_1}\left(\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}\sum_{k=0}^{n}\binom{n}{k}(2\pi i)^{k+1}z^k\delta_{k,n-1}a_n\right)$$
$$-\Gamma^{-2}\left(\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}\sum_{k=0}^{n-1}\binom{n}{k}(2\pi i)^{k+1}z^k(n-k)A_{n,k+1}\right)$$
$$+\frac{\Gamma^{-2}}{\pi}\int_{B_1}\left(\sum_{n=0}^{\infty}\frac{1}{(2\pi i)^n}\sum_{k=0}^{n-1}\binom{n}{k}(2\pi i)^{k+1}z^k(n-k)A_{n,k+1}\right)$$
$$=-\frac{3}{8\pi}\int_0^t\int_{B_1}\frac{\partial\phi}{\partial z}-\frac{1}{8\pi}\int_{B_1}\frac{\partial\phi_0}{\partial z}-\Gamma^{-2}\frac{\partial\phi}{\partial z}+\frac{\Gamma^{-2}}{\pi}\int_{B_1}\frac{\partial\phi}{\partial z},$$

so that (2.13d) also holds.

In a similar way we can show that (2.13b) and (2.13c) are satisfied, thus completing the existence proof for Theorem 3.1. □

5.2. Uniqueness. Let (ϕ, p) be a solution of (2.13) satisfying (i)-(iv) of Theorem 3.1 for $\phi_0\equiv 0$. We wish to show that $\phi\equiv p\equiv 0$. Write

$$\phi(t,x,y,z)=\sum_{n=0}^{\infty}\phi^n(t,x,y)z^n/n!,\qquad p(t,x,y,z)=\sum_{n=0}^{\infty}p^n(t,x,y)z^n/n!.$$

Then, for each $n\geqq 0$,

$$\frac{\partial\phi^n}{\partial t}-\Gamma^{-2}\Delta\phi^n+\phi^n=-\alpha^2p^{n+1}-c\phi^{n+2}\quad\text{on } [0,T]\times B_1, \tag{5.9a}$$

$$\Delta p^n=\frac{\partial\phi^{n+1}}{\partial t}-2\phi^{n+1}-c\phi^{n+3}\quad\text{on } [0,T]\times B_1, \tag{5.9b}$$

(5.9c) $$\frac{\partial \phi^n}{\partial \nu} = -\frac{3\alpha^2}{2\pi}\int_0^t \int_{B_1} \phi^{n+2} + \frac{\alpha^2}{2\pi}\int_{B_1} \phi^{n+2} \quad \text{on } [0, T]\times \partial B_1,$$

(5.9d) $$p^n = -\frac{3}{8\pi}\int_0^t \int_{B_1} \phi^{n+1} - \Gamma^{-2}\phi^{n+1} + \frac{\Gamma^{-2}}{\pi}\int_{B_1} \phi^{n+1} \quad \text{on } [0, T]\times \partial B_1,$$

(5.9e) $$\phi^n(0, x, y) = 0.$$

Using (5.9a), (5.9c), (5.9e) together with the parabolic Schauder estimates we get

(5.10) $$\|\phi^n\|_{2+\alpha} \leqq k_1(|||p^{n+1}||| + \|\phi^{n+2}\|_{2+\alpha})$$

for some constant k_1.

Also, using (5.9b), (5.9d) and the maximum principle

$$|||p^{n-1}||| \leqq k_2(\|\phi^n\|_{2+\alpha} + \|\phi^{n+2}\|_{2+\alpha})$$

and using (5.10)

(5.11) $$|||p^{n-1}||| \leqq k_3(|||p^{n+1}||| + \|\phi^{n+2}\|_{2+\alpha})$$

for some constant k_3.

For $n \geqq 0$ set

$$\gamma_n = \|\phi^n\|_{2+\alpha} + |||p^{n-1}|||,$$

where, by definition, $p^{-1} \equiv 0$. Then, (5.10) and (5.11) give

(5.12) $$\gamma_n \leqq k_0\gamma_{n+2}, \qquad n \geqq 0$$

for some constant k_0.

On the other hand, using Cauchy's formula and part (iv) of Theorem 3.1 we see that, for $m \geqq 2$,

$$\gamma_m \leqq c_\varepsilon\left[m!\frac{(\eta_0+\varepsilon)^m e^m}{m^m} + (m-1)!\frac{(\eta_0+\varepsilon)^{m-1} e^{m-1}}{(m-1)^{m-1}}\right]$$

and from (5.5)

(5.13) $$\gamma_m \leqq \tilde{c}_\varepsilon(\eta_0+\varepsilon)^m m^{1/2} \quad \text{for } \varepsilon > 0, \quad m \geqq 2.$$

Fix $n \geqq 0$; then from (5.12)

(5.14) $$\gamma_n \leqq k_0^l\gamma_{n+2l} \quad \forall l \geqq 0$$

so that from (5.13), (5.14) we obtain

(5.15) $$\gamma_n \leqq \tilde{c}_\varepsilon k_0^l(\eta_0+\varepsilon)^{n+2l}(n+2l)^{1/2}, \qquad l \geqq 1.$$

Thus, for every $l \geqq 1$

(5.16) $$\gamma_n \leqq \tilde{c}_\varepsilon(\eta_0+\varepsilon)^n[((\eta_0+\varepsilon)^2 k_0)^l(n+2l)^{1/2}].$$

Hence, letting $l \to \infty$ we get

$$\gamma_n = 0 \quad \forall n \geqq 0,$$

provided $\eta_0 < (1/k_0)^{1/2}$ and $0 < \varepsilon < (1/k_0)^{1/2} - \eta_0$ and, consequently,

$$\phi \equiv p \equiv 0.$$

This completes the proof of Theorem 3.1 provided we choose $\eta_0 < \min(1/c_2, (1/k_0)^{1/2})$. (Recall that c_2 is the constant in (4.3), (4.4).) □

6. Properties of the solution.

THEOREM 6.1. *Let $\phi_0 \in A^{\alpha}_{\eta_0}$ where $0<\alpha<1$ and η_0 is as in Theorem* 3.1. *Let (ϕ, p) be the solution of* (2.13) *given in Theorem* 3.1. *Then*:

(a) *If ϕ_0 is periodic in z (i.e., $\phi_0(\cdot, z+Z)=\phi_0(\cdot, z)$ for some $Z\in\mathbb{R}$ and all $z\in\mathbb{R}$) then also ϕ and p are periodic in z, with the same period Z.*

(b) *Let $m\geqq 2$ and assume that*

$$\|\phi_0(\cdot, z)\|^0_{2+\alpha} \leqq \frac{c_0}{(1+|z|)^m} \tag{6.1}$$

for some constant $c_0>0$ $(z\in\mathbb{R})$. Then

$$\|\phi(\cdot, z)\|_{2+\alpha} + |||p(\cdot, z)||| = o((1+|z|)^{2-m}) \quad \text{as } z\to\infty. \tag{6.2}$$

THEOREM 6.2. *Let ϕ_0, ϕ and p be as in Theorem* 6.1. *If we assume that $\phi_0(x, y, z) = \phi_0(x, y, -z)$ and*

$$\int_{\mathbb{R}} \left\| \frac{\partial \phi_0}{\partial z} \right\|^0_{2+\alpha} dz < \infty, \tag{6.3}$$

then

$$\phi(t, x, y, z) = \phi(t, x, y, -z), \qquad p(t, x, y, z) = -p(t, x, y, -z) \tag{6.4}$$

and

$$\left\| \frac{\partial \phi}{\partial z}(\cdot, z) \right\|_{2+\alpha} + \left|\left| \frac{\partial p}{\partial z}(\cdot, z) \right|\right| = o(1) \quad \text{as } z\to\infty. \tag{6.5}$$

In order to prove Theorems 6.1 and 6.2 we shall use the following theorem.

THEOREM (Paley-Wiener). *Let K be a compact, convex, and balanced subset of $\mathbb{R}^n$, and let $I_K(\eta) = \sup_{\mu\in K} |\langle \eta, \mu\rangle|$ (where $\langle \eta, \mu\rangle = \eta\cdot\mu = \sum_{i=1}^n \eta_i\mu_i$).*

Let $T\in\mathcal{S}'$ (i.e., T is a tempered distribution). Then the following are equivalent:

(i) supp $T\subset K$.

(ii) *$\hat{T}$ has an entire holomorphic extension to $\mathbb{C}^n$ so that, for some $m\geqq 0$,*

$$|\hat{T}(\zeta)| \leqq c(1+|\zeta|)^m \exp I_K(\operatorname{Im}(\zeta)).$$

(*Actually m can be taken to be the order of T.*)

(For a proof see, e.g., [5, pp. 145-146].)

Proof of Theorem 6.1. Note that (a) follows immediately from the uniqueness of solutions: in fact, if we set $\tilde{\phi}(t, x, y, z) = \phi(t, x, y, z+Z)$, $\tilde{p}(t, x, y, z) = p(t, x, y, z+Z)$ it is clear that $(\tilde{\phi}, \tilde{p})$ is a solution of (2.13) satisfying (i)-(iv) of Theorem 3.1, so that, by uniqueness, we must have $\tilde{\phi}\equiv\phi$, $\tilde{p}\equiv p$.

We now turn to the proof of (b). Clearly, $u_0(x, y, \mu) = \phi_0(x, y, \cdot)\hat{}(\mu)$ satisfies

$$u_0(\cdot, \mu) \in \bar{C}_{2+\alpha}(\overline{B_1}), \tag{6.6a}$$

$$u_0(x, y, \cdot) \in C^{m-2}(\mathbb{R}). \tag{6.6b}$$

Also, an application of the Paley-Wiener theorem shows that $u_0(x, y, \cdot)$ is compactly supported and that, for each $(x, y)\in\overline{B_1}$

$$\text{supp } u_0(x, y, \cdot) \subset I_0 = \{\mu \,|\, |\mu| \leqq \eta_0\}. \tag{6.6c}$$

We may then consider the Fourier transformed system (3.3) with initial data $u_0(x, y, \mu)$. Following the proof of Lemma 4.1 we see that there exists a solution (u, v) of (3.3) satisfying:

$$u(t, x, y, \cdot), v(t, x, y, \cdot) \in C_0^{m-2}(\mathbb{R}),$$

$$\frac{\partial^l u}{\partial \mu^l}(\cdot, \mu) \in \bar{C}_{2+\alpha}([0, T] \times \overline{B_1}) \qquad (0 \leqq l \leqq m-2),$$

$$\frac{\partial^l v}{\partial \mu^l}(\cdot, \mu) \in \bar{C}_{\alpha}([0, T] \times \overline{B_1}) \qquad (0 \leqq l \leqq m-2),$$

$$\frac{\partial^l v}{\partial \mu^l}(t, \cdot, \mu) \in C^{2+\alpha}(\overline{B_1}) \qquad (0 \leqq l \leqq m-2),$$

$$\frac{\partial}{\partial t}\frac{\partial^l v}{\partial \mu^l}(t, \cdot, \mu) \in C^{2+\alpha}(B_1) \cap C^0(\overline{B_1}) \qquad (0 \leqq l \leqq m-2).$$

Furthermore, for each $(x, y) \in \overline{B_1}$, $t \in [0, T]$

$$\text{supp } u(t, x, y, \cdot) \subset I_0 \quad \text{and} \quad \text{supp } v(t, x, y, \cdot) \subset I_0. \tag{6.7}$$

Then, if we set

$$\tilde{\phi}(t, x, y, z) = u(t, x, y, \cdot)^{\vee}(z) \quad \text{and} \quad \tilde{p}(t, x, y, z) = v(t, x, y, \cdot)^{\vee}(z)$$

we see that $(\tilde{\phi}, \tilde{p})$ is a solution of (2.13) satisfying (6.2). But yet another application of the Paley-Wiener theorem (using (6.7)) implies that $(\tilde{\phi}, \tilde{p})$ satisfy (i)-(iv) of Theorem 3.1. Thus, by uniqueness, $(\tilde{\phi}, \tilde{p}) = (\phi, p)$, and this completes the proof of (b). □

Proof of Theorem 6.2. If $\phi_0(x, y, z) = \phi_0(x, y, -z)$, set $\tilde{\phi}(t, x, y, z) = \phi(t, x, y, -z)$ and $\tilde{p}(t, x, y, z) = -p(t, x, y, -z)$. It is easily checked then that $(\tilde{\phi}, \tilde{p})$ is a solution of (2.13) satisfying (i)-(iv) of Theorem 3.1 and hence, by uniqueness, we obtain $\tilde{\phi} \equiv \phi$ and $\tilde{p} \equiv p$. This proves (6.4).

Finally, (6.5) may be proved using an argument similar to that used for proving (6.2). □

Acknowledgments. The author thanks L. A. Romero for presenting the problem that is the subject of this paper and Professor A. Friedman for his help and encouragement. He also thanks Professor E. Fabes for several enlightening discussions.

REFERENCES

[1] J. R. ASAY AND L. D. BIRTHOFF, Report SAND 78 Sandia National Laboratories, Albuquerque, NM, 1978.

[2] G. BIRKHOFF, D. P. MAC DOUGALL, E. M. PUGH, AND G. I. TAYLOR, *Explosives with lined cavities*, J. Appl. Phys., 19 (1948), pp. 563-582.

[3] D. B. BOGY, *Drop formation in a circular liquid jet*, Ann. Rev. Fluid Mech., 11 (1979), pp. 207-228.

[4] I. FRANKEL AND D. WEIHS, *Stability of a capillary jet with linearly increasing axial velocity (with application to shaped charges)*, J. Fluid Mech., 155 (1985), pp. 289-307.

[5] A. FRIEDMAN, *Generalized Functions and Partial Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[6] ———, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[7] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Second edition, Springer-Verlag, Berlin, New York, 1983.

[8] M. GOLDIN, J. YERUSHALMI, R. PFEFFER, AND R. SHINNAR, *Breakup of a laminar capillary jet of a viscoelastic fluid*, J. Fluid Mech., 38 (1969), pp. 689–711.

[9] D. E. GRADY, *Fragmentation of rapidly expanding jets and sheets*, in Hypervelocity Impact: Proceedings of the 1986 Symposium, Pergamon Press, Oxford, 1987.

[10] O. A. LADYŽENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasi-Linear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[11] V. G. LEVICH, *Physiochemical Hydrodynamics*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[12] E. M. PUGH, R. J. EICHELBERGER, AND N. ROSTOKER, *Theory of jet formation by charges with lined conical cavities*, J. Appl. Phys., 23 (1952), pp. 532–536.

[13] J. W. RAYLEIGH, *The Theory of Sound*, 2nd ed., Macmillan, London, 1926.

[14] L. A. ROMERO, *The instability of rapidly stretching plastic jets*, J. Appl. Phys., to appear.

[15] C. WEBER, *Zum Zerfall eines Flüssigkeitsstrahles*, Z. Angew. Math. Mech., 11 (1931), pp. 136–154.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 129-137, January 1991

007

NONLINEAR AGE-DEPENDENT POPULATION DYNAMICS WITH CONSTANT SIZE*

TANYA KOSTOVA† AND FABIO MILNER‡

Abstract. The existence of populations of constant size is studied for the Gurtin-MacCamy model. Necessary and sufficient conditions are given and some examples of their implications are considered.

Key words. population dynamics, age-structured, constant size population

AMS(MOS) subject classification. 92A15

1. Introduction. In their classic paper Gurtin and MacCamy [2] considered the following equations describing nonlinear age-dependent population dynamics with an age distribution $u = u(a, t)$ satisfying

$$\begin{aligned} &u_a + u_t = -d(a, P(t))u, \qquad a, t > 0, \\ &u(0, t) = \int_0^\infty b(a, P(t))u(a, t)\, da, \qquad t \geqq 0, \\ &u(a, 0) = \phi(a), \qquad a \geqq 0, \end{aligned} \tag{1.1}$$

where a is the age, t is the time, d is the deathrate, b is the birthrate, and $P(t) = \int_0^\infty u(a, t)\, da$ is the total population size.

In [2] a necessary and sufficient condition for the existence of an equilibrium age distribution $u(a)$ of (1.1) with a prescribed population size $P_0 = \int_0^\infty u(a)\, da$ was derived. More precisely, given a constant $P_0 > 0$ and assuming that $b(\cdot, P_0) \exp(-\int_0^\cdot d(s, P_0)\, ds) \in L^1(\mathbb{R}^+)$, a necessary and sufficient condition for the equations

$$\begin{aligned} &u_a = -d(a, P_0)u, \qquad a > 0, \\ &u(0) = \int_0^\infty b(a, P_0)u(a)\, da, \\ &P_0 = \int_0^\infty u(a)\, da \end{aligned}$$

to have a solution is that

$$\int_0^\infty b(a, P_0) \exp\left(-\int_0^a d(\tau, P_0)\, d\tau\right) da = 1. \tag{1.2}$$

In this case $u(a)$ is determined explicitly as

$$u(a) = \frac{P_0 \exp(-\int_0^a d(\tau, P_0)\, d\tau)}{\int_0^\infty \exp(-\int_0^a d(\tau, P_0)\, d\tau)\, da}. \tag{1.3}$$

* Received by the editors June 19, 1989; accepted for publication December 11, 1989.

† Institute of Mathematics, Bulgarian Academy of Sciences, Acad. G. Bonchev Str. BG. 8, 1113 Sofia, Bulgaria.

‡ Dipartimento di Mathematica, Universita degli Studi di Roma, Via Fontanile di Carcaricola, 00133 Roma, Italy and Department of Mathematics, Purdue University, West Lafayette, Indiana 47907. The work of this author was partially supported by National Science Foundation Research grant DMS-8813258.

Equilibrium solutions are of major importance, as they may be asymptotically stable provided some conditions on the birthrate $b(a, P)$, death rate $d(a, P)$, and the initial age distribution $\phi(a)$ are fulfilled; thus they may determine the asymptotic behavior of the solutions of (1.1) having suitable initial age distributions. That is why much attention has been paid to the investigation of the properties of the steady state (we refer to [1]-[5], just to mention some of this work). Still, in real situations the available data on population dynamics is more often given in terms of the total population size than in terms of the age distribution. There are contemporary examples of human populations whose size does not change in time, though they may have a time-dependent age distribution. An interesting question for which we find the answer here is what the relation between the birth and death rates is in such occasions.

This paper is organized as follows. In § 2 we derive some necessary and sufficient conditions for the existence of a population of constant size in terms of its initial distribution and its fertility and mortality. In § 3 we give some examples of consequences of these results and discuss their implications.

2. Populations of constant size. Let P_0 be a prescribed number. We shall assume that b, d, and ϕ in (1.1) are piecewise continuous functions, $d \not\equiv 0$, and ϕ is integrable. We shall say $u(a, t)$ is a *distribution of constant size* P_0 if $P(t) \equiv P_0$. Consider equations (1.1) and also assume the following compatibility conditions on initial and boundary data:

$$\int_0^\infty \phi(a)\, da = P_0, \qquad \phi(0) = \int_0^\infty b(a, P_0)\phi(a)\, da. \tag{2.1}$$

Suppose that (1.1) has a solution $u(a, t)$ such that $\int_0^\infty u(a, t)\, da = P_0$ for all t. Integrating the first equation of (1.1) in the age variable, we obtain the following equation for $P(t)$:

$$P'(t) = \int_0^\infty [b(a, P_0) - d(a, P_0)]u(a, t)\, da, \qquad t > 0,$$

and, therefore, in our case, $P(t) \equiv P_0$ is equivalent to the condition:

$$\int_0^\infty b(a, P_0)u(a, t)\, da = \int_0^\infty d(a, P_0)u(a, t)\, da, \qquad t \geqq 0. \tag{2.2}$$

This last equality is a necessary and sufficient condition for $u(a, t)$ to be of constant size but it still does not determine an explicit relation between $b(a, P_0)$ and $d(a, P_0)$. It obviously leads to the conclusion that $u(a, t)$ is the solution of the system

$$\begin{aligned} &u_a + u_t = -d(a, P_0)u, && a, t > 0, \\ &u(0, t) = \int_0^\infty d(a, P_0)u(a, t)\, da, && t \geqq 0, \\ &u(a, 0) = \phi(a), && a \geqq 0. \end{aligned} \tag{2.3}$$

Remark 2.1. The condition $b(a, P_0) = d(a, P_0)$ in (1.1) is sufficient for the validity of (2.2) and, therefore, any solution of (2.3) is of constant size. This system is one which also admits equilibrium age distributions in the case where $\int_0^\infty d(a, P_0)\, da = \infty$, since then, $\int_0^\infty d(a, P_0) \exp\left(-\int_0^a d(s, P_0)\, ds\right) da = 1$ (see (1.2)). Equilibrium distributions exist only when $\phi(a)$ is given by (1.3). Hence, for any other function ϕ, the solution of (2.3) is of constant size but not an equilibrium one.

Several questions arise, which we shall answer here.

(a) When is there a nonequilibrium solution $u(a, t)$ of (1.1) with $P(t) \equiv P_0$?

(b) Given an initial age distribution $\phi(a)$ with $\int_0^\infty \phi(a)\, da = P_0$, does a solution with constant size exist?

(c) If the answer to (b) is yes, what condition must $b(a, P_0)$ and $d(a, P_0)$ satisfy?

It is well known (see [2]) that the solution of (1.1) can be represented in the following implicit way:

$$u(a, t) = \begin{cases} B(t-a, P_0) \exp\left(-\int_0^a d(\tau, P_0)\, d\tau\right), & t \geqq a, \\ \phi(a-t) \exp\left(-\int_0^t d(a-t+\tau, P_0)\, d\tau\right), & t < a, \end{cases} \tag{2.4}$$

where

$$B(s, P_0) = \int_0^\infty b(a, P_0) u(a, s)\, da \tag{2.5}$$

is the birth function.

Combining (2.4) and (2.5) we obtain the following formula:

$$\begin{aligned} B(t, P_0) = &\int_0^t b(a, P_0) B(t-a, P_0) \exp\left(-\int_0^a d(\tau, P_0)\, d\tau\right) da \\ &+ \int_t^\infty b(a, P_0) \phi(a-t) \exp\left(-\int_0^t d(a-t+\tau, P_0)\, d\tau\right) da. \end{aligned} \tag{2.6}$$

Recall now that we have assumed that $P(t) \equiv P_0$ and, therefore, (2.2) is valid. This means that $B(t, P_0) = \int_0^\infty d(a, P_0) u(a, t)\, da$, and thus we shall also have

$$\begin{aligned} B(t, P_0) = &\int_0^t d(a, P_0) B(t-a, P_0) \exp\left(-\int_0^a d(\tau, P_0)\, d\tau\right) da \\ &+ \int_t^\infty d(a, P_0) \phi(a-t) \exp\left(-\int_0^t d(a-t+\tau, P_0)\, d\tau\right) da. \end{aligned} \tag{2.7}$$

Conversely, if (2.7) is valid, then (2.2) will hold (because (2.6) holds) and $u(a, t)$ will be of constant size. Equation (2.7) is another necessary and sufficient condition for the existence of a population of constant size P_0; but it is still implicit, as we do not know $B(t, P_0)$. We can use Laplace transforms to obtain an equation involving only b, d, and ϕ. For this purpose, assume that $d(a, P_0)$ and $b(a, P_0)$ are bounded. As $u(\cdot, t) \in L^1(\mathbb{R}^+)$ for each t (since $\int_0^\infty u(a, t)\, da \equiv P_0$), then $B(t, P_0) = \int_0^\infty b(a, P_0) u(a, t)\, da \leqq \|b\|_{L^\infty} P_0$, that is, $B(t, P_0)$ is a bounded function of t. On the other hand, $\int_t^\infty b(a, P_0) \phi(a-t) \exp\left(-\int_0^t d(a-t+\tau, P_0)\, d\tau\right) da \leqq \|b\|_{L^\infty} \int_0^\infty \phi(s)\, ds = \|b\|_{L^\infty} P_0$ and, similarly, $\int_t^\infty d(a, P_0) \phi(a-t) \exp\left(-\int_0^t d(a-t+\tau, P_0)\, d\tau\right) da$ is bounded. Therefore, for $s \geqq 0$, we can take the Laplace transform $\mathscr{L}$ of both sides of (2.6) and (2.7) and, making use of the special convolution form of the first integrals in the right-hand sides, we arrive at

$$\mathscr{L}(B)(s) = \mathscr{L}(B)(s) \mathscr{L}(\Pi)(s) + \mathscr{L}(B_0)(s), \tag{2.8}$$

$$\mathscr{L}(B)(s) = \mathscr{L}(B)(s) \mathscr{L}(\theta)(s) + \mathscr{L}(D_0)(s), \tag{2.9}$$

where

$$\mathcal{L}(f)(s) = \int_0^\infty e^{-st} f(t)\, dt,$$

$$\Pi(t) = b(t, P_0) \exp\left(-\int_0^t d(\tau, P_0)\, d\tau\right),$$

$$\theta(t) = d(t, P_0) \exp\left(-\int_0^t d(\tau, P_0)\, d\tau\right), \tag{2.10}$$

$$B_0(t) = \int_0^\infty \phi(a) b(a+t, P_0) \exp\left(-\int_0^t d(a+\tau, P_0)\, d\tau\right) da,$$

$$D_0(t) = \int_0^\infty \phi(a) d(a+t, P_0) \exp\left(-\int_0^t d(a+\tau, P_0)\, d\tau\right) da.$$

Therefore, if $u(a, t)$ is a solution of (1.1) of constant size P_0, then (2.9) holds for all $s > 0$ and, of course, (2.8) does too.

Conversely, if (2.9) is valid then, as (2.8) holds, (2.6) and (2.7) will also hold and the solution of (1.1) will be of constant size.

Now consider the quantity

$$1 - \mathcal{L}(\theta)(s) = 1 - \int_0^\infty e^{-st}\, d(t, P_0) \exp\left(-\int_0^t d(\tau, P_0)\, d\tau\right) dt.$$

For $s = 0$,

$$0 \leqq \mathcal{L}(\theta)(0) = 1 - \exp\left(-\int_0^\infty d(\tau, P_0)\, d\tau\right) \leqq 1.$$

If we note that for $s > 0$, we have $0 \leqq \mathcal{L}(\theta)(s) < \mathcal{L}(\theta)(0)$, then we see that

$$1 - \mathcal{L}(\theta)(s) > 0 \quad \text{for } s > 0.$$

Therefore, we can express $\mathcal{L}(B)$ from (2.9) as

$$\mathcal{L}(B)(s) = \mathcal{L}(D_0)(s)/(1 - \mathcal{L}(\theta)(s)),$$

which, substituted in (2.8), yields the relation

$$\frac{1 - \mathcal{L}(\Pi)(s)}{1 - \mathcal{L}(\theta)(s)} \mathcal{L}(D_0)(s) = \mathcal{L}(B_0)(s);$$

that is,

$$[1 - \mathcal{L}(\Pi)(s)]\mathcal{L}(D_0)(s) = \mathcal{L}(B_0)(s)[1 - \mathcal{L}(\theta)(s)], \qquad s > 0. \tag{2.11}$$

This last equality can also be rewritten in another way, as we show below. Having in mind that for $s > 0$ and for $X(\cdot, t) \in L^1(\mathbb{R}^+)$, $t > 0$, the following relation holds:

$$\mathcal{L}\left(\int_0^\infty X(a, t)\, da\right) = \frac{1}{s}\int_0^\infty X(a, 0)\, da + \frac{1}{s}\mathcal{L}\left(\int_0^\infty \frac{\partial X}{\partial t}(a, t)\, da\right),$$

we obtain, after some sample computations, the relations

$$\mathcal{L}(B_0)(s) = \frac{1}{s}\phi(0)(1 - \mathcal{L}(\Pi)(s))$$

$$-\frac{1}{s}\mathcal{L}\left(\int_0^\infty (\phi'(a) + d(a, P_0)\phi(a)) b(a+t, P_0)\right.$$

$$\left. \cdot \exp\left(-\int_0^t d(a+\tau, P_0)\, d\tau\right) da\right)(s),$$

and

$$\mathscr{L}(D_0)(s)=\frac{1}{s}\int_0^\infty d(a)\phi(a)\,da(1-\mathscr{L}(\theta)(s))$$
$$-\frac{1}{s}\mathscr{L}\Big(\int_0^\infty[\phi'(a)+d(a,P_0)\phi(a)]d(a+t,P_0)$$
$$\cdot\exp\Big(-\int_0^t d(a+\tau,P_0)\,d\tau\Big)\,da\Big)(s).$$

It follows that (2.11) is equivalent to the following two relations:

$$\phi(0)=\int_0^\infty d(a,P_0)\phi(a)\,da,$$
$$[1-\mathscr{L}(\Pi)]\mathscr{L}\Big(\int_0^\infty[\phi'(s)+d(s,P_0)\phi(s)]d(s+t,P_0)$$
$$\cdot\exp\Big(-\int_0^t d(s+\tau,P_0)\,d\tau\Big)\,ds\Big) \tag{2.12}$$
$$=[1-\mathscr{L}(\theta)]\mathscr{L}\Big(\int_0^\infty[\phi'(s)+d(s,P_0)\phi(s)]b(s+t,P_0)$$
$$\cdot\exp\Big(-\int_0^t d(s+\tau,P_0)\,d\tau\Big)\,ds\Big),$$

as seen by substituting $\phi(0)$, D_0, and B_0 in (2.11). We have shown that (2.11) and (2.12) are other necessary and sufficient conditions for the solution (1.1) to be of constant size.

Let $Q(t)$ and $R(t)$ be defined as

$$Q(t)=\int_0^\infty[\phi'(s)+d(s,P_0)\phi(s)]d(s+t,P_0)\exp\Big(-\int_0^t d(s+\tau,P_0)\,d\tau\Big)\,ds,$$
$$R(t)=\int_0^\infty[\phi'(s)+d(s,P_0)\phi(s)]b(s+t,P_0)\exp\Big(-\int_0^t d(s+\tau,P_0)\,d\tau\Big)\,ds.$$

Taking inverse Laplace transforms, we obtain the following result.

THEOREM 2.1. *Suppose that ϕ, b, d are bounded, nonnegative, piecewise continuous functions, satisfying the compatibility conditions* (2.1). *The necessary and sufficient condition for the solution of* (1.1) *to be of constant size P_0 is that either of the following relations holds*:

(i) $$D_0(t)-\int_0^t D_0(s)\Pi(t-s)\,ds=B_0(t)-\int_0^t B_0(s)\theta(t-s)\,ds,$$

where D_0, B_0, Π, and θ are defined by (2.10),

(ii) $$Q(t)-\int_0^t Q(s)\Pi(t-s)\,ds=R(t)-\int_0^t R(s)\theta(t-s)\,ds,$$
$$\phi(0)=\int_0^\infty d(a,P_0)\phi(a)\,da.$$

Next we shall derive another necessary and sufficient condition for the solution of (1.1) to be of constant size, which proves useful for some special cases and does not impose boundedness restrictions on b, d, and ϕ. Let us assume again that (1.1) has a solution of constant size P_0. As $P_0 \equiv \int_0^\infty u(a, t)\, da$, then integrating (2.4) in age and differentiating in time, we get

$$0 = \int_0^t B_t(t-a, P_0) \exp\left(-\int_0^a d(\tau, P_0)\, d\tau\right) da + \int_t^\infty \frac{d}{dt}\left[\phi(a-t) \exp\left(-\int_0^t d(a-t+\tau, P_0)\, dt\right)\right] da. \tag{2.13}$$

The second term in the right-hand side of (2.13) depends only on the functions ϕ and d and (2.13) may then be considered as a Volterra equation for $B_t(t-a, P_0)$. Let us denote

$$F(t) = -\int_t^\infty \frac{d}{dt}\left[\phi(a-t) \exp\left(-\int_0^t d(a-t+\tau, P_0)\, d\tau\right)\right] da$$
$$= \int_t^\infty [\phi'(a-t) + \phi(a-t) d(a-t, P_0)] \exp\left(-\int_0^t d(a-t+\tau, P_0)\, d\tau\right) da.$$

Consequently,

$$F(t) = \int_0^\infty [\phi'(s) + \phi(s) d(s, P_0)] \exp\left(-\int_0^t d(s+\tau, P_0)\, d\tau\right) ds. \tag{2.14}$$

Let $H(s) = B_t(s, P_0)$. Then, $H(t)$ satisfies the following Volterra equation of the first kind

$$\int_0^t H(s) \exp\left(-\int_0^{t-s} d(\tau, P_0)\, d\tau\right) ds = F(t). \tag{2.15}$$

On the other hand, since (2.2) is valid for all t, including $t = 0$, it follows from (2.2) and (2.5) that

$$\phi(0) = B(0, P_0) = \int_0^\infty b(a, P_0)\phi(a)\, da = \int_0^\infty d(a, P_0)\phi(a)\, da.$$

Then, (2.14) implies that

$$F(0) = -\phi(0) + \int_0^\infty \phi(s) d(s, P_0)\, ds = 0.$$

Therefore, we can use the following well-known result. Let $I = [0, T]$ and let $S = \{(t, s): 0 \leqq s \leqq t \leqq T\}$. Let $F \in C^1(I)$ with $F(0) = 0$ and let $K(t, s)$, $\partial K/\partial t \in C^0(S)$. Assume that $K(t, t) \neq 0$ for all $t \in I$. Under these conditions $\int_0^t K(t, s) y(s)\, ds = F(t)$ has a unique solution $y \in C^0(I)$.

Applying this theorem, we see that, if $d \in C^0(I)$ and

$$[\phi'(\cdot) + \phi(\cdot) d(\cdot, P_0)] \exp\left(-\int_0^t d(\cdot + \tau, P_0)\, d\tau\right) \in L^1(\mathbb{R}^+),$$

then (2.15) has a unique solution H:

$$H = \mathscr{L}^{-1}\left(\mathscr{L}(F)/\mathscr{L}\left(\exp\left(-\int d(\cdot, P_0)\right)\right)\right),$$

where $\mathscr{L}^{-1}(f)$ is the inverse Laplace transform of f. Therefore, $B(t, P_0)$ can be determined explicitly in terms of H by integrating

$$B(t, P_0) = \phi(0) + \int_0^t H(s)\, ds. \tag{2.16}$$

If B is then inserted in the renewal equation (2.6), we arrive at a necessary condition which $b(a, P_0)$ and $d(a, P_0)$ must satisfy.

On the other hand, if u is a solution of (1.1), then it satisfies (2.6) with $B(t, P_0) = \int_0^\infty b(a, P_0)u(a, t)\, da$. Assume that (2.14)-(2.16) are fulfilled. It follows that $P'(t) = 0$ and therefore $P(t) \equiv P_0$. Thus, we have proved the following result.

THEOREM 2.2. *Let $d \in C^0(I)$ and let $\phi \in L^1(\mathbb{R}^+)$ and b satisfy the compatibility conditions* (2.1) *and* $[\phi'(\cdot) + \phi(\cdot)d(\cdot, P_0)] \exp(-\int_0^t d(\cdot + \tau, P_0)\, d\tau) \in L^1(\mathbb{R}^+)$ *for each $t \in I$. Then,* (1.1) *has a solution with constant size P_0 in the time interval $I = [0, T]$ if and only if $b(a, P_0)$, $d(a, P_0)$, $\phi(a)$, and $B(t)$ satisfy* (2.6) *and* (2.14)-(2.16). *In such case, the solution $u(a, t)$ is given explicitly by* (2.4), *where $B(t)$ is determined from* (2.15), *and* (2.16).

Remark 2.2. Note that in this theorem we do not impose boundedness conditions on b and ϕ as we did in Theorem 2.1.

Remark 2.3. Let $P_0 > 0$ and let the assumptions of Theorem 2.2 hold. Let $d(a, P_0)$ and $\phi(a)$ be such that $F(t) \not\equiv 0$ (which is always possible, in view of (2.14)). Then, $B(t)$ can be determined uniquely and $u(a, t)$ subsequently from it. We should note that such populations of constant size are not equilibrium distributions. Indeed, suppose that $u(a, t) \equiv u(a)$. Then, $B(t) = B(0) = \phi(0)$ and, therefore, $H(u) = 0$ and $F(t) = 0$. This is a contradiction. Therefore $u(a, t)$ is not an equilibrium distribution of (1.1).

The only case when there does not exist a nonequilibrium solution with constant size is when $\phi(s) = \phi(0) \exp(-\int_0^s d(\tau, P_0)\, d\tau)$. In this case $F \equiv 0$ (since $\phi' + d\phi \equiv 0$) and so, $H \equiv 0$, which gives $B(t) \equiv B(0) = \phi(0)$. Therefore, the relation following (2.15) transforms into $1 = \int_0^\infty b(a, P_0) \exp(-\int_0^a d(\tau, P_0)\, d\tau)\, da$ and $u(a, t) = \phi(0) \exp(-\int_0^a d(\tau, P_0)\, d\tau) = u(a)$. In other words, from an initial equilibrium distribution the only possible solution of (1.1) with constant size is that equilibrium distribution (1.3). We have proved the following result.

THEOREM 2.3. *Let $P_0 > 0$ be given and let b and ϕ satisfy the compatibility conditions* (2.1). *Then, for any continuous function $d(a, P_0)$ such that $F(t) \not\equiv 0$ and such that* $(\phi'(s) + \phi(s)d(s, P_0)) \exp(-\int_0^t d(s + \tau, P_0)\, d\tau) \in L^1$, *there exists a nonequilibrium solution $u(a, t)$ of* (1.1) *such that $\int_0^\infty u(a, t)\, da \equiv P_0$. Furthermore, if $\phi(a) = \phi(0) \exp(-\int_0^a d(\tau, P_0)\, d\tau)$, then, the solution is the equilibrium one, $u(a, t) \equiv \phi(a)$.*

Theorem 2.3 answers the questions (a), (b), and (c) at the beginning of this section.

3. Some examples. We shall treat some special cases of (1.1) and discuss its solutions of constant size applying the above theory.

(I) Let $d(a, P) \equiv d$. Then $u(a, t)$ is a solution of constant size P_0 if and only if (2.12) holds. Since $1 - \mathscr{L}(\theta)$ can vanish at one point, at most, and all the functions of which we take transforms are piecewise continuous, it follows that (2.12), for the case of a constant death rate, is equivalent to the following relations:

$$\int_0^\infty [\phi'(a) + d\phi(a)]\, da = 0, \tag{3.1}$$

$$\int_0^\infty [\phi'(a) + d\phi(a)]b(a + t, P_0)\, da = 0, \qquad t \geqq 0. \tag{3.2}$$

The last relation follows by observing that, if (3.1) holds and d is constant, then, the left-hand side of the second equation of (2.12) vanishes. Conversely, (3.1) and (3.2) imply that u has constant size invoking Theorem 2.1. In summary, if the death rate is constant and b and ϕ are piecewise continuous bounded functions such that the compatibility conditions (2.1) hold, then u is of constant size if and only if (3.1) and (3.2) hold. In such case, the solution of (1.1) of size P_0 is

$$u(a,t)=\begin{cases} dP_0\, e^{-ad}, & t \geqq a, \\ \phi(a-t)\, e^{-td}, & t < a. \end{cases}$$

(II) Let

$$\phi(s)=\begin{cases} K>0, & 0 \leqq s \leqq c, \\ 0 \quad \text{otherwise}, \end{cases}$$

be a uniform distribution on the interval $(0, c)$ so that $P_0 = Kc$. Assume that all individuals die before reaching age c. Then, $\int_0^c d(a, P_0)\, da = +\infty$. Now note that (1.1) and (2.2) give in this case

$$\phi(0) = B(0) = \int_0^\infty b(a, P_0)u(a, 0)\, da = \int_0^c b(a, P_0)\phi(0)\, da$$
$$= \int_0^c d(a, P_0)\phi(0)\, da,$$

which implies

$$\int_0^c d(a, P_0)\, da = \int_0^c b(a, P_0)\, da = 1,$$

a contradiction. Therefore, given a uniform initial age-distribution on the finite age interval $(0, c)$, if the initial and boundary date are compatible, then, for any death rate $d(a)$ such that the life span of any individual is at most c, the population cannot have constant size, independently of the birthrate b.

(III) Let $b(a, P) \equiv b$. Then, u is of constant size P_0 if, and only if,

$$B(t) = \int_0^\infty b(a, P_0)u(a, t)\, da = b \int_0^\infty u(a, t)\, da = bP(t) \equiv bP_0,$$

which is equivalent to the following relations:

$$B_t(t) = 0, \qquad B(0) = bP_0.$$

Using Theorem 2.2 we see that these relations are equivalent to the following necessary and sufficient conditions for a population with constant birthrate b to have constant size P_0:

$$\int_0^\infty [\phi'(a) + d(a, P_0)\phi(a)] \exp\left(-\int_0^t d(a+\tau, P_0)\, d\tau\right) da = 0, \qquad t \geqq 0,$$

$$\phi(0) = bP_0.$$

In such a case, u is explicitly given by

$$u(a,t)=\begin{cases} bP_0 \exp\left(-\int_0^a d(\tau, P_0)\, d\tau\right), & t \geqq a, \\ \phi(a-t) \exp\left(-\int_0^t d(a-t+\tau, P_0)\, d\tau\right), & t < a. \end{cases}$$

REFERENCES

[1] R. H. ELDERKIN, *Population models with globally age-dependent dynamics: on computing the steady state*, Comput. Math. Appl., 9 (1983), pp. 371–376.

[2] M. E. GURTIN AND R. C. MACCAMY, *Non-linear age-dependent population dynamics*, Arch. Rational Mech. Anal., 54 (1974), pp. 281–300.

[3] V. G. MATSENKO AND V. N. RUBANOVSKII, *Application of the direct Lyapunov method to analysis of the dynamics of an age-structured biopopulation*, Zh. Vychisl. Mat. i Mat. Fiz., 23 (1983), pp. 326–332. (In Russian.)

[4] G. F. WEBB, *Theory of Nonlinear Age-Dependent Population Dynamics*, Marcel Dekker, New York, 1985.

[5] E. WEISTOCK AND C. RORRES, *Local stability of an age-structured population with density-dependent fertility and mortality*, SIAM J. Appl. Math., 47 (1987), pp. 589–604.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 138-145, January 1991

008

L^{∞}_{loc}-ESTIMATES FOR LOCAL SOLUTIONS OF DEGENERATE PARABOLIC EQUATIONS*

DANIELE ANDREUCCI†

Abstract. A local sup-estimate for local (sub)solutions of degenerate parabolic equations is proved. Such *local* estimates do not involve any dependence on the initial and boundary data, but rather, provide a bound for the solution in a given domain, only in terms of some integral norm of the solution in a larger domain. Estimates of this kind for solutions of linear *nondegenerate* parabolic equations are due to Moser (*Comm. Pure Appl. Math.*, 17 (1964), pp. 101-134). Sharp estimates of this kind have not been available in the literature in the degenerate case. The new input here is an interpolation process that permits one to deal with the degeneracy of the equation. The estimates shown here are sharp and reduce to the classical ones of Moser in the linear nondegenerate case.

Key words. degenerate parabolic equations, local solutions, local sup-estimates

AMS(MOS) subject classifications. 35K65, 35B45

1. Introduction. Local sup-estimates for solutions of the porous media equation (or its quasilinear variations) are employed in several papers to study the behavior (local or at infinity) of such solutions (e.g., [1], [3], [4]).

We are interested here in local sup-estimates that are completely independent of initial or boundary data. Such a dependence typically appears in estimates relying on the maximum principle. To our knowledge, quantitative estimates of this kind are not present in the literature, in a form that is sharp or comparable with analogous results for nondegenerate equations. In this paper we prove a sharp estimate (see (1.6) below), coinciding with known results in the linear case [6]. The estimate will take the form of an interpolation inequality with a free parameter.

We remark that sup-estimates for solutions defined in the whole space $\mathbb{R}^N$ have been obtained in [2], where, in particular, a connection is traced between regularizing effect and sup-estimates.

We will consider nonnegative local subsolutions of the equation

$$u_t - \operatorname{div} \vec{a}(x, t, u, \nabla\phi(u)) = 0, \tag{1.1}$$

where $\phi : [0, \infty) \to [0, \infty)$ is a locally AC (absolutely continuous) function satisfying, for some $\Lambda > 1$,

$$1 < \frac{\phi'(s)s}{\phi(s)} \leqq \Lambda \quad \text{a.e. } s > 0. \tag{1.2}$$

The first inequality in (1.2) can be allowed to be nonstrict; see Remarks 3.1 and 3.3 below. For the sake of brevity we carry out the proofs only in case (1.2).

We define $\Phi(s) = \phi(s)s^{-1}$, $s > 0$, and assume in the following that the function $(x, t) \to \vec{a}(x, t, u(x, t), \nabla\phi(u(x, t))$ is measurable and satisfies

$$\vec{a}(x, t, u, \nabla\phi(u)) \cdot \nabla u \geqq \Lambda^{-1}\Phi(u)|\nabla u|^2, \tag{1.3}$$

$$|\vec{a}(x, t, u, \nabla\phi(u))| \leqq \Lambda\Phi(u)|\nabla u| \tag{1.4}$$

(of course the constants in (1.3), (1.4) are named Λ just for notational simplicity).

* Received by the editors April 21, 1989; accepted for publication (in revised form) January 12, 1990.
† Istituto Matematico "Ulisse Dini," Università di Firenze, viale Morgagni 67/a, 50134 Florence, Italy.

In order to state our main result we introduce the following notation:

$$B_\rho = \{x \in \mathbb{R}^N \,|\, |x - x_0| < \rho\},$$

$$Q_\infty = B_\rho \times \left(\frac{t}{2}, t\right),$$

$$Q_0 = B_{(1+\sigma)\rho} \times \left(\frac{t}{2} - \frac{\sigma}{2} t, t\right),$$

where ρ, $t > 0$, $\sigma \in (0, 1)$, $x_0 \in \mathbb{R}^N$ are given. We also use the notation

$$\fint\!\!\int_{Q_0} f\,dx\,dt = (\text{meas}\,(Q_0))^{-1} \int\!\!\int_{Q_0} f\,dx\,dt.$$

Let us define a local subsolution u to (1.1) in Q_0 as a function

$$u \in C(t_0, t; L^1(B_{(1+\sigma)\rho})), \qquad t_0 \equiv \frac{t}{2} - \frac{\sigma}{2} t,$$

$$\nabla \phi(u) \in L^2(Q_0),$$

satisfying

$$u_t - \operatorname{div} \vec{a}(x, t, u, \nabla \phi(u)) \leqq 0, \tag{1.5}$$

in the usual weak sense in Q_0 (see, for example, [5]). Then we have the following theorem.

THEOREM. *Let u be a bounded nonnegative local subsolution of* (1.1) *in Q_0. Then for all η, $\varepsilon > 0$*

$$\begin{aligned} \phi(\|u\|_{\infty, Q_\infty}) &\leqq \eta \frac{\rho^2}{t} \Phi^{-1}\left(\eta \frac{\rho^2}{t}\right) \\ &+ \gamma\left(1 + \frac{1}{\eta}\right)^{(N+2)/2\varepsilon} \left(\frac{t}{\rho^2}\right)^{1/\varepsilon} \left(\fint\!\!\int_{Q_0} \Phi(u) \phi(u)^\varepsilon \,dx\,d\tau\right)^{1/\varepsilon}, \end{aligned} \tag{1.6}$$

where $\gamma = \gamma(\varepsilon, \sigma, \Lambda, N)$: γ becomes unbounded as ε (or σ, or Λ^{-1}) tends to zero.

We note that (1.6) is also valid for any nonnegative subsolution which can be approximated locally by bounded subsolutions.

In § 2 we prove the theorem, and in § 3 we collect some comments and generalizations.

2. Proof of (1.6). We will need the following elementary inequalities:

$$h^\Lambda \phi(s) \leqq \phi(sh) \leqq h \phi(s), \qquad s > 0, \quad h \in (0, 1), \tag{2.1}$$

$$h^{\Lambda - 1} \Phi(s) \leqq \Phi(sh) \leqq \Phi(s), \qquad s > 0, \quad h \in (0, 1), \tag{2.2}$$

(2.1) is proved integrating on (hs, s):

$$\frac{1}{s} \leqq \frac{\phi'(s)}{\phi(s)} \leqq \frac{\Lambda}{s},$$

and (2.2) follows from (2.1).

We notice that (2.1), (2.2) imply that $\phi(s)$ and $\phi(s)^\alpha s^{-2}$ $(\alpha > 2)$ are increasing functions of s, and that Φ is nondecreasing.

LEMMA. *Let us define for* $z,\ z_0 \geqq 0$, *and* $\alpha > 2$

$$F(z, z_0) = \left(\int_{z_0}^{z} \left(\int_{z_0}^{r} \frac{\phi(s)^\alpha}{s^2}\, ds \right)_+ dr \right)^{1/2}.$$

Then we have for all $z > 0$, $z \neq z_0$,

$$\frac{\phi(z)^\alpha}{z^2} \geqq c \left| \frac{\partial F}{\partial z}(z, z_0) \right|^2, \tag{2.3}$$

where $c = c(\alpha, \Lambda)$.

Proof. If $z \in (0, z_0)$, $F(z, z_0) \equiv 0$ and there is nothing to prove.

If $z > z_0$,

$$\left| \frac{\partial F}{\partial z}(z, z_0) \right|^2 = \frac{\left(\int_{z_0}^{z} \frac{\phi(s)^\alpha}{s^2}\, ds \right)^2}{4 \left(\int_{z_0}^{z} dr \int_{z_0}^{r} \frac{\phi(s)^\alpha}{s^2}\, ds \right)}. \tag{2.4}$$

Hence

$$4 \left| \frac{\partial F}{\partial z} \right|^2 \leqq \frac{\phi(z)^{2\alpha}}{z^4} (z - z_0)^2 \left(z^{-2} \int_{z_0}^{z} dr \int_{z_0}^{r} \phi(s)^\alpha\, ds \right)^{-1}.$$

Now we consider two cases: (i) $z_0 \geqq z/2$ and (ii) $z_0 < z/2$. In case (i)

$$\int_{z_0}^{z} dr \int_{z_0}^{r} \phi(s)^\alpha\, ds \geqq \int_{z_0}^{z} dr \int_{z_0}^{r} \phi\left(\frac{z}{2}\right)^\alpha ds \geqq 2^{-\Lambda\alpha - 1} \phi(z)^\alpha (z - z_0)^2.$$

In case (ii)

$$\int_{z_0}^{z} dr \int_{z_0}^{r} \phi(s)^\alpha\, ds \geqq \int_{z/2}^{z} dr \int_{z/2}^{r} \phi\left(\frac{z}{2}\right)^\alpha ds \geqq 2^{-\Lambda\alpha - 3} \phi(z)^\alpha z^2.$$

Therefore (2.3) is proved with $c = 2^{-\Lambda\alpha - 1}$.

Remark 1.1. Notice that (2.3) implies $F(\cdot, z_0) \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^+)$ and that $F(z, \cdot)$ is nonincreasing on $\mathbb{R}^+$.

Proof of the theorem. We use the approach of [1] and define for all $n \geqq 0$

$$t_n = \frac{t}{2} - \frac{\sigma t}{2^{n+1}}, \quad \rho_n = \rho + \frac{\sigma}{2^n}\rho, \quad k_n = k - \frac{k}{2^{n+1}},$$

$$B_n = B_{\rho_n}, \quad Q_n = B_{\rho_n} \times (t_n, t), \quad B_n(\tau) = B_n \times \{\tau\};$$

here $k > 0$ is to be chosen later.

We consider also a cutoff function ζ_n such that

$$\zeta_n(x, \tau) \equiv 0, \quad (x, \tau) \notin Q_n, \qquad \zeta_n(x, \tau) \equiv 1, \quad (x, \tau) \in Q_{n+1},$$

$$|\nabla \zeta_n| \leqq \frac{2^{n+1}}{\sigma\rho}, \qquad 0 \leqq \zeta_{n\tau} \leqq \frac{2^{n+2}}{\sigma t},$$

and we use as testing functions in the weak formulation of (1.5)

$$f_n(x, \tau) = \left(\int_{k_{n+1}}^{u(x,\tau)} \frac{\phi(s)^\alpha}{s^2}\, ds \right)_+ \zeta_n(x, \tau)^2,$$

with $\alpha > 2$.

The time part of (1.5) yields for all $\bar{t} \in (t_n, t)$ (the calculations below can be made rigorous by means of a Steklov averaging process)

$$\begin{aligned}
&\iint_{Q_n \cap \{\tau < \bar{t}\}} u_\tau \left(\int_{k_{n+1}}^{u} \frac{\phi(s)^\alpha}{s^2}\, ds \right)_+ \zeta_n^2 \, dx\, d\tau \\
&\quad = \iint_{Q_n \cap \{\tau < \bar{t}\}} \zeta_n^2 \frac{\partial}{\partial \tau} F(u, k_{n+1})^2 \, dx\, d\tau \\
&\quad \geqq -\frac{2^{n+3}}{\sigma t} \iint_{Q_n} F(u, k_n)^2 \, dx\, d\tau + \int_{B_n(\bar{t})} F(u, k_{n+1})^2 \zeta_n^2 \, dx.
\end{aligned} \tag{2.5}$$

We use the lemma in treating the space part:

$$\begin{aligned}
-\iint_{Q_n} (\operatorname{div} \vec{a}) f_n \, dx\, d\tau &= \iint_{Q_n} \vec{a} \cdot \nabla \left(\left(\int_{k_{n+1}}^{u} \frac{\phi(s)^\alpha}{s^2}\, ds \right)_+ \zeta_n^2 \right) dx\, d\tau \\
&\geqq \frac{1}{\Lambda} \iint_{Q_n \cap \{u > k_{n+1}\}} \Phi(u) \frac{\phi(u)^\alpha}{u^2} |\nabla u|^2 \zeta_n^2 \, dx\, d\tau \\
&\quad - 2\Lambda \iint_{Q_n} \Phi(u) \left(\int_{k_{n+1}}^{u} \frac{\phi(s)^\alpha}{s^2}\, ds \right)_+ \zeta_n |\nabla \zeta_n| |\nabla u| \, dx\, d\tau \\
&\geqq 2^{-3\alpha\Lambda} \iint_{Q_n} \Phi(u) |\nabla F(u, k_{n+1})|^2 \zeta_n^2 \, dx\, d\tau \\
&\quad - \frac{2\Lambda^2}{\theta} \iint_{Q_n} \Phi(u) |\nabla \zeta_n|^2 F(u, k_{n+1})^2 \, dx\, d\tau \\
&\quad - \frac{\theta}{2} \iint_{Q_n} \Phi(u) \zeta_n^2 |\nabla u|^2 F(u, k_{n+1})^{-2} \left(\int_{k_{n+1}}^{u} \frac{\phi(s)^\alpha}{s^2}\, ds \right)_+^2 dx\, d\tau.
\end{aligned}$$

But

$$F(z, z_0)^{-2} \left(\int_{z_0}^{z} \frac{\phi(s)^\alpha}{s^2}\, ds \right)^2 = 4 \left| \frac{\partial F}{\partial z}(z, z_0) \right|^2;$$

therefore, for a $\theta > 0$ suitably chosen, we arrive at the following inequality:

$$\begin{aligned}
-\iint_{Q_n} (\operatorname{div} \vec{a}) f_n \, dx\, d\tau &\geqq 2^{-4\alpha\Lambda} \Phi(k) \iint_{Q_n} |\nabla F(u, k_{n+1}) \zeta_n|^2 \, dx\, d\tau \\
&\quad - 2^{8\alpha\Lambda} \frac{2^{2n}}{\sigma^2 \rho^2} \iint_{Q_n} \Phi(u) F(u, k_n)^2 \, dx\, d\tau,
\end{aligned} \tag{2.6}$$

where we have also used $u \geqq k_{n+1} > k_0 = (k/2)$, and (2.2).

Adding (2.5) to (2.6) we get

$$\begin{aligned}
&\sup_{t_n \leqq \tau \leqq t} \int_{B_n(\tau)} F(u, k_{n+1})^2 \zeta_n^2 \, dx + \Phi(k) \iint_{Q_n} |\nabla F(u, k_{n+1}) \zeta_n|^2 \, dx\, d\tau \\
&\qquad \leqq \gamma_1 \frac{2^{2n}}{\sigma^2 t} \left(1 + \frac{t}{\rho^2} \Phi(\|u\|_{\infty, Q_0}) \right) \iint_{Q_n} F(u, k_n)^2 \, dx\, d\tau, \qquad \gamma_1 = 2^{20\alpha\Lambda}.
\end{aligned} \tag{2.7}$$

Now fix $\eta > 0$. First we consider the case

$$\mu \equiv \frac{t}{\rho^2} \Phi(\|u\|_{\infty, Q_0}) \geqq \eta. \tag{2.8}$$

Then the right-hand side (r.h.s.) of (2.7) is majorized by

$$\gamma_1 \frac{2^{2n}}{\sigma^2 t}\left(1+\frac{1}{\eta}\right)\mu \int\int_{Q_n} F(u, k_n)^2 \, dx \, d\tau.$$

We also need the following estimate for

$$|A_{n+1}| \equiv \text{meas}\,\{(x, \tau) \in Q_n | u(x, \tau) > k_{n+1}\}.$$

Using (2.1) again we have

$$\begin{aligned} \int\int_{Q_n} F(u, k_n)^2 \, dx \, d\tau &\geqq |A_{n+1}| \int_{k_n}^{k_{n+1}} dr \int_{k_n}^{r} \frac{\phi(s)^\alpha}{s^2} \, ds \\ &\geqq |A_{n+1}| \frac{\phi(k_n)^\alpha}{k_n^2} \frac{(k_{n+1}-k_n)^2}{2} \\ &\geqq 2^{-\alpha\Lambda(n+2)} |A_{n+1}| \phi(k)^\alpha. \end{aligned} \tag{2.9}$$

From (2.7)-(2.9) and the embedding of [5, pp. 74-75], there follows

$$\begin{aligned} &\int\int_{Q_{n+1}} F(u, k_{n+1})^2 \, dx \, d\tau \\ &\quad \leqq \int\int_{Q_n} F(u, k_{n+1})^2 \zeta_n^2 \, dx \, d\tau \\ &\quad \leqq |A_{n+1}|^{2/(N+2)} \left(\int\int_{Q_n} [F(u, k_{n+1})^2 \zeta_n^2]^{(N+2)/N} \, dx \, d\tau \right)^{N/(N+2)} \\ &\quad \leqq C(N) |A_{n+1}|^{2/(N+2)} \left(\int\int_{Q_n} |\nabla F(u, k_{n+1}) \zeta_n|^2 \, dx \, d\tau \right)^{N/(N+2)} \\ &\qquad \cdot \left(\sup_{t_n \leqq \tau \leqq t} \int_{B_n(\tau)} F(u, k_{n+1})^2 \zeta_n^2 \, dx \right)^{2/(N+2)} \\ &\quad \leqq DB^n \left(1+\frac{1}{\eta}\right) \mu (\sigma^2 t)^{-1} \phi(k)^{-\alpha(2/(N+2))} \Phi(k)^{-N/(N+2)} \\ &\qquad \cdot \left(\int\int_{Q_0} F(u, k_n)^2 \, dx \, d\tau \right)^{1+(2/N+2)}, \end{aligned} \tag{2.10}$$

with $D = C(N) 2^{25\alpha\Lambda}$, $B = 2^{2\alpha\Lambda}$; here and below $C(N)$ is a constant depending only on N.

If k is such that

$$\begin{aligned} \int\int_{Q_0} F(u, k_0)^2 \, dx \, d\tau &\leqq \int\int_{Q_0} F(u, 0)^2 \, dx \, d\tau \\ &= D^{-(N+2)/2} B^{-((N+2)/2)^2} \left[(\sigma^2 t)^{-1} \left(1+\frac{1}{\eta}\right) \mu \right]^{-(N+2)/2} \cdot \phi(k)^\alpha \Phi(k)^{N/2}, \end{aligned}$$

Lemma 5.6 on page 95 of [5] implies that $\iint_{Q_n} F(u, k_n)^2 \, dx \, d\tau \to 0$ as $n \to \infty$, i.e., $\|u\|_{\infty, Q_\infty} \leqq k$.

Thus, taking into account $F(u,0)^2 \leqq \phi(u)^\alpha$ and recalling the definitions of Φ and μ,

$$\phi(\|u\|_{\infty,Q_\infty})^{\alpha+(N/2)}\|u\|_{\infty,Q_\infty}^{-(N/2)} \leqq \gamma_2(\sigma\rho)^{-(2+N)}\left(1+\frac{1}{\eta}\right)^{(N+2)/2}\phi(\|u\|_{\infty,Q_0})^{\alpha+(N/2)-\varepsilon}\|u\|_{\infty,Q_0}^{-1-(N/2)}\iint_{Q_0}\phi(u)^{1+\varepsilon}\,dx\,d\tau \tag{2.11}$$

($\gamma_2 = C(N)2^{60\alpha\Lambda N^2}$) if we assume also $\alpha > 1+\varepsilon$, e.g.,

$$\alpha = 2+\varepsilon. \tag{2.12}$$

Since obviously

$$\|u\|_{\infty,Q_\infty} \leqq \|u\|_{\infty,Q_0}, \qquad \phi(u)\|u\|_{\infty,Q_0}^{-1} \leqq \phi(u)u^{-1} \quad \text{in } Q_0,$$

we get from (2.11)

$$\phi(\|u\|_{\infty,Q_\infty})^{\alpha+(N/2)} \leqq \gamma_2(\sigma\rho)^{-(2+N)}\left(1+\frac{1}{\eta}\right)^{(N+2)/2}\phi(\|u\|_{\infty,Q_0})^{\alpha+(N/2)-\varepsilon} \cdot \iint_{Q_0}\Phi(u)\phi(u)^\varepsilon\,dx\,d\tau. \tag{2.13}$$

Let us define

$$\beta = 1-\frac{\varepsilon}{\alpha+(N/2)},$$

$$Q(s,\tau) = \{(x,\theta)\,|\,|x-x_0|<s,\ \tau<\theta<t\},$$

$$U(s,\tau) = \phi(\|u\|_{\infty,Q(s,\tau)})^{\alpha+(N/2)}.$$

With this notation, on applying Young's inequality to (2.13) we have

$$U\left(\rho,\frac{t}{2}\right) \leqq \delta U(\rho_0,t_0) + \gamma_3(\rho_0-\rho)^{-(2+N)/(1-\beta)}M_0, \tag{2.14}$$

for any $\delta \in (0,1)$; here

$$M_0 = \delta^{-\beta/(1-\beta)}\left(1+\frac{1}{\eta}\right)^{(N+2)/(2(1-\beta))}\left(\iint_{Q_0}\Phi(u)\phi(u)^\varepsilon\,dx\,d\tau\right)^{1/(1-\beta)}$$

and $\gamma_3 = \gamma_2^{1/(1-\beta)}$.

Define

$$s_0 = \rho, \qquad s_{i+1}-s_i = (1-\sigma)\sigma^i(\sigma\rho),$$

$$\tau_0 = \frac{t}{2}, \qquad \tau_i - \tau_{i+1} = (1-\sigma)\sigma^i\left(\sigma\frac{t}{2}\right),$$

for $i = 0, 1, 2, \cdots$ and note that (2.14) holds in the form

$$U(s_i,\tau_i) \leqq \delta U(s_{i+1},\tau_{i+1}) + \gamma_3((1-\sigma)\sigma\rho)^{-(N+2)/(1-\beta)}\sigma^{-(N+2)/(1-\beta)i}M_0.$$

By iteration

$$U(s_0,\tau_0) \leqq \delta^n U(s_n,\tau_n) + \gamma_3((1-\sigma)\sigma\rho)^{-(N+2)/(1-\beta)}M_0\sum_{i=0}^{n}[\delta\sigma^{-(N+2)/(1-\beta)}]^i.$$

Now we choose $\delta = \frac{1}{2}\sigma^{(N+2)/(1-\beta)}$ and let $n \to \infty$. Then, taking the $(\alpha + (N/2))$th root of both sides of the inequality so obtained, we arrive at

$$\phi(\|u\|_{\infty,Q_\infty}) \leqq \gamma_4 \left(\frac{t}{\rho^2}\right)^{1/\varepsilon} (1-\sigma)^{-(N+2)/\varepsilon} \sigma^{-(N+\alpha)^2/\varepsilon^2} \left(1+\frac{1}{\eta}\right)^{(N+2)/2\varepsilon} \cdot \left(\fint\!\!\!\int_{Q_0} \Phi(u)\phi(u)^\varepsilon \, dx\, d\tau\right)^{1/\varepsilon}, \tag{2.15}$$

$\gamma_4 = (2\gamma_2)^{1/\varepsilon}$.

Finally, if (2.8) is not true, we have

$$\|u\|_{\infty,Q_\infty} \leqq \Phi^{-1}\left(\eta \frac{\rho^2}{t}\right). \tag{2.16}$$

Estimate (1.6) follows when we combine (2.15) with (2.16) and $\phi(\Phi^{-1}(x)) = x\Phi^{-1}(x)$ (note that Φ^{-1} exists because of the strict inequality in (1.2)).

3. Comments and generalizations.

Remark 3.1. If $\phi(u) = u^m$, $m > 1$, (1.6) reads (take $\varepsilon = \lambda/m$, $\lambda > 0$)

$$\|u\|_{\infty,Q_\infty} \leqq \left(\eta \frac{\rho^2}{t}\right)^{1/(m-1)} + \gamma \left(\frac{t}{\rho^2}\right)^{1/\lambda} \left(1+\frac{1}{\eta}\right)^{(N+2)/2\lambda} \left(\fint\!\!\!\int_{Q_0} u^{m-1+\lambda} \, dx\, d\tau\right)^{1/\lambda},$$

for all $\eta, \lambda > 0$.

In the linear case $\phi(u) = u$, $\Phi \equiv 1$, we can take $\eta = t/\rho^2$ in (2.8), so that our estimate takes the form (2.15), which in turn reduces to Moser's sup-estimate [6].

Remark 3.2. A solution of $u_t - \Delta u^m = 0$, $m > 1$, is given by

$$V(x,t) = \alpha |x|^{2/(m-1)} \left(1 - \frac{t}{T^*}\right)^{-1/(m-1)}, \qquad t \in (0, T^*), \quad x \in \mathbb{R}^N,$$

$$T^* = (m-1)(2m(N(m-1)+2)\alpha^{m-1})^{-1}, \quad \alpha \text{ fixed.}$$

Applying to V the estimate of Remark 3.1, we can see that the functional dependence on u of the second term on the r.h.s. of (1.6) is sharp.

We also note that, in (1.6), a corrective term balancing the second term on the r.h.s. is needed (just consider solutions $u \equiv \text{const}$, and let $\rho \to \infty$). Moreover, in our estimate, the behavior as $\rho \to \infty$ of the corrective term $(\eta\rho^2/t)\Phi^{-1}(\eta\rho^2/t)$, is the one predicted by sup-estimates at infinity for solutions of (1.1) ([1], [3], [4]).

The interpolation form assumed by (1.6) allows us to give our estimate for the subsolution u itself, rather than for $w = \max(u, c)$ ($c > 0$ fixed arbitrarily: see Remark 3.3 for an estimate of w).

Remark 3.3. Assume ϕ satisfies (1.2) (where we now allow the first inequality to be nonstrict) only for $s > \Lambda \geqq 1$, and u satisfies in Q_0

$$u_t - \operatorname{div} \vec{a}(x,t,u,\nabla\phi(u)) \leqq b(x,t,u,\nabla\phi(u)),$$

where

$$\vec{a}(x,t,u,\nabla\phi(u)) \cdot \nabla u \geqq \Lambda^{-1}\Phi(u)|\nabla u|^2 - \Lambda(1+u\phi(u)),$$
$$|\vec{a}(x,t,u,\nabla\phi(u))| \leqq \Lambda(\Phi(u)|\nabla u| + \phi(u) + 1),$$
$$|b(x,t,u,\nabla\phi(u))| \leqq \Lambda(\Phi(u)|\nabla u| + \phi(u) + 1).$$

Then, using the techniques of the lemma and theorem above, we can prove for $w = \max(u, 2\Lambda)$

$$\phi(\|w\|_{\infty,Q_\infty}) \leqq \gamma \left\{(1+t)\left(\frac{t}{\rho^2} + \frac{\rho^2}{t}\right)^2 (1 + \Phi(\Lambda)^{-1})\right\}^{(N+2)/2\varepsilon} \left(\fint\!\!\!\int_{Q_0} \Phi(w)\phi(w)^\varepsilon \, dx\, d\tau\right)^{1/\varepsilon}. \tag{3.1}$$

Indeed, in the proof we can take $k \geqq 2\Lambda$, so that (2.1), (2.2) still hold in $[k/2, \infty)$. Moreover, (2.8) certainly holds if we take

$$\eta = \Phi(\Lambda)\frac{t}{\rho^2};$$

this choice accounts for the appearance of $\Phi(\Lambda)$ in (3.1).

Remark 3.4. Solutions of variable sign. Assume $\phi : \mathbb{R} \to \mathbb{R}$ is an increasing AC function satisfying (1.2) almost everywhere $s \in \mathbb{R}$, and u is a bounded local solution of (1.1) in Q_0 (in the class defined above); u is not required to have constant sign. Then the previous arguments, with minor changes, prove estimates similar to (1.6) for the positive and negative parts of u. Analogous extensions hold if ϕ merely satisfies the assumptions of Remark 3.3 for large values of $|s|$.

REFERENCES

[1] D. ANDREUCCI AND E. DIBENEDETTO, *A new approach to initial traces in nonlinear filtration*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[2] PH. BÉNILAN AND J. BERGER, *Estimation uniforme de la solution de $u_t = \Delta\varphi(u)$ et caractérisation de l'effet régularisant*, C.R. Acad. Sci. Paris, Sér. I Math., 300 (1985), pp. 573-576.

[3] PH. BÉNILAN, M. G. CRANDALL, AND M. PIERRE, *Solutions of the porous medium equation in $\mathbb{R}^n$ under optimal conditions on initial values*, Indiana Univ. Math. J., 33 (1984), pp. 51-87.

[4] B. E. J. DAHLBERG AND C. E. KENIG, *Non-negative solutions of generalized porous medium equations*, Rev. Mat. Iberomericana, 2 (1986), pp. 267-305.

[5] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'TZEVA, *Linear and quasilinear equations of parabolic type*, Trans. Math. Monographs, 23, American Mathematical Society, Providence, RI, 1968.

[6] J. MOSER, *A Harnack inequality for parabolic differential equations*, Comm. Pure Appl. Math., 17 (1964), pp. 101-134.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 146-156, January 1991

009

AN INVERSE PROBLEM FOR A CLASS OF QUASILINEAR PARABOLIC EQUATIONS*

YANPING LIN†

Abstract. The identification of the source control $q = q(t)$ of one-dimensional quasilinear parabolic equations is considered via additional information on the solution of integral type. Existence, uniqueness, and continuous dependence of the solution upon the data are demonstrated by employing some a priori estimates, compactness arguments, and the strong maximum principle.

Key words. inverse problem, parabolic, a priori estimates, maximum principle

AMS(MOS) subject classifications. 35R25, 35R30

1. Introduction. We study the identification of the unknown source control $q = q(t)$ in the following quasilinear parabolic equation:

$$\begin{aligned} &u_t = a(x, t, u, u_x)_x + q(t)u + F(x, t, u, u_x, q(t)) \quad \text{in } Q_T, \\ &u(x, 0) = \phi(x), \qquad 0 < x < 1, \\ &u(0, t) = f(t), \quad u(1, t) = g(t), \quad 0 < t < T, \end{aligned} \tag{1.1}$$

subject to the integral identity

$$\int_0^{s(t)} \Phi(x, t)u(x, t)\, dx = E(t), \qquad 0 < t < T, \quad 0 < s(t) \leqq 1, \tag{1.2}$$

where $Q_T = \{(x, t) | 0 < x < 1, 0 < t < T\}$ with $T > 0$, the functions $a = a(x, t, u, p)$, $F = F(x, t, u, p, q)$, ϕ, f, g, s, E, and Φ are known and

$$a_p = \frac{\partial a}{\partial p}(x, t, u, p) \geqq a_0 > 0. \tag{1.3}$$

Problem (1.1)-(1.2) and other similar inverse problems, or parameters identification problems, have recently been studied by several authors both in one- and n-dimensional spaces [3], [4], [5], [15], [16]. In [3]-[5] Cannon and Lin solve these problems classically, while abstract semigroup methods were employed in [15] and [16]. When $a = p$ and (1.1) is subject to the second boundary conditions $u_x(0, t) = f(t)$ and $u_x(1, t) = g(t)$, problem (1.1)-(1.2) has been treated by Cannon and Lin [5] with $s(t) \equiv 1$. There existence, uniqueness, and stability of the solution are derived via potential theoretic representation techniques. It is obvious that the method used in [5] will no longer be good for problem (1.1)-(1.2) due to the nonlinearity of the leading term, which suggests the need for another alternative.

To interpret the integral condition (1.2), we consider the following. First, if u is a temperature, then (1.1)-(1.2) can be regarded as a control problem with source control. Here we investigate the identification of the source control $q(t)$ necessary to produce the specified or desired energy $E(t)$ on a portion of the domain. Second, let us consider the example of certain chemicals absorbing light at various frequencies given by Cannon, Esteva, and van der Hoek in [2]. The intensity of such light on a

* Received by the editors April 19, 1989; accepted for publication (in revised form) December 12, 1989.
† Department of Mathematics and Statistics, McGill University, Montreal, Québec, Canada, H3A 2K6. Present address, Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

photoelectric cell gives an electric signal which is proportional to the total amount of chemical present in the volume through which the light passes. If u denotes the concentration of such a chemical which is diffusing in a straight glass tube with x measured in the direction of the axis of the tube, then the electric signal produced by a light beam passing through the tube at the right angles to the tube between $x=0$ and $x=s(t)$ is proportional to the integral in (1.2) (with $\Phi=1$) which is the total mass of the chemical in $0\leqq x\leqq s(t)$ at the time t. For such diffusion processes, the integral condition (1.2) arises naturally and can be used as supplementary information in the determination of the unknown concentration u and the source $q(t)$. We can also find other examples in which the integral condition (1.2) arises, for instance, heat transmission in a thin rod in particle diffusion in turbulent plasma [1], [7], [11], [12].

Following [5] we want to eliminate the term $q(t)u$ in (1.1) by introducing the following transformations:

$$v(x,t)=u(x,t)\exp\left\{\int_0^t -q(\xi)\,d\xi\right\},\qquad r(t)=\exp\left\{-\int_0^t q(\xi)\,d\xi\right\}, \tag{1.4}$$

$$u(x,t)=v(x,t)\exp\left\{\int_0^t q(\xi)\,d\xi\right\},\quad q(t)=\frac{-\dot r(t)}{r(t)},\quad \dot r=\frac{dr}{dt}. \tag{1.5}$$

Thus, (1.1)-(1.2) can be written equivalently, by $(u,q)\to(v,r)$, as

$$\begin{aligned} &v_t=r(t)a\left(x,t,\frac{v}{r},\frac{v_x}{r}\right)_x+r(t)F\left(x,t,\frac{v}{r},\frac{v_x}{r},\frac{-\dot r}{r}\right)\quad \text{in } Q_T,\\ &v(x,0)=\phi(x),\qquad 0<x<1,\\ &v(0,t)=r(t)f(t),\quad v(1,t)=r(t)g(t),\quad 0<t<T, \end{aligned} \tag{1.6}$$

and

$$r(t)=\frac{1}{E(t)}\int_0^{s(t)}\Phi(x,t)v(x,t)\,dx,\qquad 0<t<T,\quad 0<s(t)\leqq 1. \tag{1.7}$$

Problem (1.6)-(1.7) now can be viewed as a quasilinear parabolic equation with nonlocal boundary condition and nonlinear functional of the solution if we substitute (1.7) into (1.6). Although (1.6)-(1.7) looks more complicated than does problem (1.1)-(1.2), it is easier to handle as we shall see below. We now define the solution pair (u,q) and (v,r).

DEFINITION 1.1. A pair (u,q) is called a solution for problem (1.1)-(1.2) if there exists an α, $0<\alpha<1$, such that

$$u\in C^{1+\alpha}(\bar Q_T)\cap C^{2+\alpha}(Q_T)\quad\text{and}\quad q\in C^{\alpha/2}([0,T]),$$

and that the pair (u,q) satisfy equations (1.1)-(1.2).

DEFINITION 1.2. A pair (v,r) is called a solution for problem (1.6)-(1.7) if $r(t)\neq 0$ and there exists an α, $0<\alpha<1$, such that

$$v\in C^{1+\alpha}(\bar Q_T)\cap C^{2+\alpha}(Q_T)\quad\text{and}\quad r\in C^{1+\alpha/2}([0,T]),$$

and that the pair (v,r) satisfy equations (1.6)-(1.7).

Here and throughout this paper we shall use the standard notation for Hölder spaces, C^{α}, $C^{1+\alpha}$, etc. defined in [9]. We shall also use the following notation:

$$\|v(\cdot,t)\|_{L^p(0,1)}=\left(\int_0^1|v(x,t)|^p\,dx\right)^{1/p},\qquad 1\leqq p<\infty,$$

$$\|v(\cdot,t)\|_{L^\infty(0,1)}=\operatorname{ess}\sup_{x\in(0,1)}|v(x,t)|,$$

for the functions $v = v(x, t)$ defined on Q_T, and $\|f\|_{L^p(0,T)}$, $1 \leqq p \leqq \infty$, for the functions defined on $(0, T)$.

Under these definitions we have the following lemma.

LEMMA 1.1. *If (u, q) is a unique solution pair for* (1.1)-(1.2), *then the pair (v, r) defined by* (1.4)-(1.5) *is a unique solution pair for* (1.6)-(1.7) *and vice versa provided that $r(t) > 0$ for $t \in [0, T)$.*

Proof. It is an elementary argument which we omit. □

Our approach as in [5] is that we shall show problem (1.6)-(1.7) has a unique solution pair (v, r) with $r(t) > 0$, and then by Lemma 1.1 we will conclude that the inverse problem (1.1)-(1.2) possesses a unique solution pair (u, q) via the transformations (1.4)-(1.5).

In order to have a direct relation between $\dot{r}$ and r, v, v_x, we differentiate (1.7) and use (1.6) to obtain

$$\begin{aligned}\dot{r} = &-\frac{\dot{E}}{E^2}\int_0^{s(t)} \Phi v\,dx + \frac{1}{E}\left\{\Phi(s(t), t)v(s(t), t)\dot{s}(t) + \int_0^{s(t)} \Phi_t v\,dx\right\} \\ &+\frac{1}{E}r(t)\left\{\Phi(s(t), t)a\left(s(t), t, \frac{v(s(t), t)}{r(t)}, \frac{v_x(s(t), t)}{r(t)}\right)\right. \\ &\left.-\Phi(0, t)a\left(0, t, \frac{v(0, t)}{r(t)}, \frac{v_x(0, t)}{r(t)}\right)\right\} \qquad (1.8) \\ &-\frac{1}{E}\int_0^{s(t)} \Phi_x(x, t)r(t)a\left(x, t, \frac{v}{r}, \frac{v_x}{r}\right)dx \\ &+\frac{1}{E}\int_0^{s(t)} \Phi(x, t)r(t)F\left(x, t, \frac{v}{r}, \frac{v_x}{r}, \frac{-\dot{r}}{r}\right)dx.\end{aligned}$$

We now state our assumptions on the data.

Assumption (H1). Assume that $a = a(x, t, u, p)$ is a smooth function with respect to all of its variables and satisfies the following growth conditions:

$$|a(x, t, u, p)| \leqq C(1 + |u| + |p|), \qquad |a_u(x, t, u, p)| + |a_x(x, t, u, p)| \leqq C. \qquad (1.9)$$

Here and in what follows, we denote by C a generic constant which depends only upon known quantities.

Assumption (H2). $f \geqq 0$, $g \geqq 0$, $\phi \geqq 0$ (ϕ is not identically zero), $0 < s(t) \leqq 1$, $\Phi(x, t) > 0$, $E(t) > 0$, and there exists α, $0 < \alpha < 1$, such that

$$\phi \in C^{2+\alpha}([0, 1]), \quad \Phi \in C^{2+\alpha}(\bar{Q}_T), \quad E, s, f, g \in C^{1+\alpha/2}([0, T]).$$

Assumption (H3). $F = F(x, t, u, p, q)$ is a smooth function with respect to all of its variables, $F \geqq 0$ and

$$|F(x, t, u, p, q)| \leqq \delta|q| + C(|u| + |p| + 1), \qquad (1.10)$$

where $\delta > 0$ is such that

$$0 \leqq \delta^* = \delta \max_{0 \leqq t \leqq T}\left\{E(t)^{-1}\int_0^{s(t)} \Phi(x, t)\,dx\right\} < 1. \qquad (1.11)$$

Remark 1.1. A consequence of (H3) is that equation (1.8) is uniquely solvable for $\dot{r}$, given r, v, v_x.

Assumption (H4). The data satisfies the following compatibility conditions:

$$\phi(0)=f(0), \quad \phi(1)=g(0), \quad E(0)=\int_0^{s(0)} \Phi(x,0)\phi(x)\,dx,$$

$$\dot{f}(0)=a(x,0,\phi(x),\phi_x(x))_x|_{x=0}+F(0,0,\phi(0),\phi_x(0),q(0)),$$

$$\dot{g}(0)=a(x,0,\phi(x),\phi_x(x))_x|_{x=1}+F(q,0,\phi(1),\phi_x(1),q(0)),$$

where $q(0)$ is uniquely determined from known data and satisfies

$$\dot{E}(0)=\Phi(s(0),0)\phi(s(0))\dot{s}(0)+\int_0^{s(0)} \Phi_t(x,0)\phi(x)\,dx+q(0)E(0)+a(x,0,\phi,\phi_x)|_0^{s(0)}$$

$$+\int_0^{s(0)} \Phi(x,0)F(x,0,\phi,\phi_x,q(0))\,dx-\int_0^{s(0)} \Phi_x(x,0)a(x,0,\phi,\phi_x)\,dx.$$

This paper is organized in the following way. In § 2, we derive by using integral inequalities, Young's inequality, and interpolation inequalities [6], [8], some a priori bounds on the solution pair (v, r). In § 3, we approximate our solution pair (v, r) by a sequence of smooth functions $\{v^\theta, r^\theta\}$, and we employ the a priori estimates obtained in § 2, the strong maximum principle and compactness arguments to conclude that there is a subsequence of $\{v^\theta, r^\theta\}$ which will converge to the real solution of problem (1.6)-(1.7). Also, we demonstrate the uniqueness and continuous dependence of the solution upon the data, and finally, we state our main results for problem (1.1)-(1.2).

2. A priori bounds on (v, r). We shall in this section derive various a priori bounds for the solution pair (v, r) for problem (1.6)-(1.7) under the assumption that the solution pair (v, r) exists and is such that $r=r(t)\neq 0$ for $t\in[0, T)$.

First, we shall show the following result.

LEMMA 2.1. *There exists a $C>0$ such that*

$$\iint_{Q_t} v_{xx}^2\,dx\,d\tau+\int_0^1 v_x^2\,dx+\int_0^t (r^2+\dot{r}^2)\,d\tau\leqq C, \qquad t\in(0, T), \tag{2.1}$$

and

$$\|v\|_{C(\bar{Q}_T)}+\|r\|_{C([0,T])}\leqq C. \tag{2.2}$$

Proof. We multiply equation (1.6) by v_{xx} and integrate over Q_t to obtain

$$\begin{aligned}\iint_{Q_t} v_t v_{xx}\,dx\,d\tau &= \iint_{Q_t} r(t)a\left(x,t,\frac{v}{r},\frac{v_x}{r}\right)_x v_{xx}\,dx\,d\tau \\ &\quad+\iint_{Q_t} r(t)F\left(x,t,\frac{v}{r},\frac{v_x}{r},\frac{-\dot{r}}{r}\right)v_{xx}\,dx\,d\tau \\ &= I_1+I_2.\end{aligned} \tag{2.3}$$

We know from integration by parts that

$$\begin{aligned}\iint_{Q_t} v_t v_{xx}\,dx\,d\tau &= \int_0^t v_t v_x|_0^1\,d\tau-\iint_{Q_t} v_x v_{xt}\,dx\,d\tau \\ &= \int_0^t v_t v_x|_0^1\,d\tau-\frac{1}{2}\int_0^1 v_x^2\,dx+\frac{1}{2}\int_0^1 \phi_x^2\,dx.\end{aligned}$$

Since

$$r(t)a\left(x, t, \frac{v}{r}, \frac{v_x}{r}\right)_x = a_p v_{xx} + a_u v_x + r(t)a_x,$$

we have from assumption (H1),

$$\begin{aligned} I_1 &\geqq a_0 \iint_{Q_t} v_{xx}^2 \, dx\, d\tau - C \iint_{Q_t} |v_{xx}|(|r|+|v|+|v_x|) \, dx\, d\tau \\ &\geqq (a_0-\varepsilon) \iint_{Q_t} v_{xx}^2 \, dx\, d\tau - C(\varepsilon) \iint_{Q_t} (|r|^2+|v|^2+|v_x|^2) \, dx\, d\tau. \end{aligned} \tag{2.4}$$

Similarly, we see from assumption (H3) that

$$|I_2| \leqq \varepsilon \iint_{Q_t} v_{xx}^2 \, dx\, d\tau + C(\varepsilon)\left\{\iint_{Q_t} (|v|^2+|v_x|^2) \, dx\, d\tau + \int_0^t (|r|^2+|\dot{r}|^2) \, d\tau\right\}. \tag{2.5}$$

It is easy to see from (1.8) and assumptions (H1)-(H3) that

$$|\dot{r}| \leqq C\{\|v(\cdot, t)\|_{L^\infty(0,1)} + \|v_x(\cdot, t)\|_{L^\infty(0,1)}\} + \delta^*|\dot{r}|$$

from which it follows that

$$|\dot{r}| \leqq \frac{C}{1-\delta^*}\{\|v(\cdot, t)\|_{L^\infty(0,1)} + \|v_x(\cdot, t)\|_{L^\infty(0,1)}\}. \tag{2.6}$$

But, from $v(0, t) = r(t)f(t)$ and $v(1, t) = r(t)g(t)$, it follows that

$$|v_t(0, t)| \leqq C|\dot{r}f + r\dot{f}| \leqq C(|r|+|\dot{r}|), \qquad |v_t(1, t)| \leqq C|\dot{r}g + r\dot{g}| \leqq C(|r|+|\dot{r}|). \tag{2.7}$$

Since

$$\int_0^t v_t v_x|_0^1 \, d\tau = \int_0^t (v_t(1, \tau)v_x(1, \tau) - v_t(0, \tau)v_x(0, \tau)) \, d\tau, \tag{2.8}$$

we obtain from (2.6)-(2.8) and the trace inequalities [14] that

$$\begin{aligned} \left|\int_0^t v_t v_x|_0^1 \, d\tau\right| &\leqq C \int_0^t \|v_x(\cdot, \tau)\|_{L^\infty(0,1)}(|r|+|\dot{r}|) \, d\tau \\ &\leqq C \int_0^t (\|v_x(\cdot, \tau)\|_{L^\infty(0,1)} + \|v(\cdot, \tau)\|_{L^\infty(0,1)}) \, d\tau \\ &\leqq \varepsilon C \iint_{Q_t} v_{xx}^2 \, dx\, d\tau + C(\varepsilon) \iint_{Q_t} (|v_x|^2+|v|^2) \, dx\, d\tau. \end{aligned} \tag{2.9}$$

Hence, by taking ε small and fixed, we have from (2.3)-(2.9) that

$$\iint_{Q_t} v_{xx}^2 \, dx\, d\tau + \int_0^1 v_x^2 \, dx \leqq C\left\{1 + \iint_{Q_t} (|v|^2+|v_x|^2) \, dx\, d\tau + \int_0^t (|r|^2+|\dot{r}|^2) \, d\tau\right\}. \tag{2.10}$$

Since

$$\iint_{Q_t} v^2 \, dx\, d\tau \leqq C \int_0^t |f|^2 \, d\tau + C \iint_{Q_t} v_x^2 \, dx\, d\tau \tag{2.11}$$

we see via (2.6) and the trace inequalities that

$$\begin{aligned} \int_0^t (|r|^2+|\dot{r}|^2) \, d\tau &\leqq C \int_0^t (\|v(\cdot, \tau)\|_{L^\infty(0,1)} + \|v_x(\cdot, \tau)\|_{L^\infty(0,1)}) \, d\tau \\ &\leqq \varepsilon C \iint_{Q_t} v_{xx}^2 \, dx\, d\tau + C(\varepsilon) \iint_{Q_t} (|v|^2+|v_x|^2) \, dx\, d\tau, \end{aligned} \tag{2.12}$$

after substituting (2.11)-(2.12) with ε sufficiently small (fixed) into (2.10) and using Gronwall's lemma, we have

$$\int\int_{Q_t} v_{xx}^2\,dx\,d\tau+\int_0^1 v_x^2\,dx\leqq C$$

and then by (2.12)

$$\int_0^t (|r|^2+|\dot r|^2)\,d\tau\leqq C.$$

Finally, (2.2) is a direct application of (2.1). □

LEMMA 2.2. *There exists $C>0$ such that*

$$\|v_x\|_{L^\infty(\bar Q_T)}+\|\dot r\|_{L^\infty(0,T)}\leqq C. \tag{2.13}$$

We need the following two well-known inequalities to prove our Lemma 2.2. The interpolation inequality:

$$\|u\|_{L^\infty(0,1)}\leqq C\|u\|_{H^1(0,1)}^{2/3}\|u\|_{L^1(0,1)}^{1/3},\qquad u\in H^1(0,1), \tag{2.14}$$

and Young's inequality:

$$AB\leqq \eta\frac{A^\sigma}{\sigma}+\eta^{-h/\sigma}\frac{B^h}{h},\quad A,B\geqq 0,\quad \eta>0,\quad \frac{1}{\sigma}+\frac{1}{h}=1,\quad \sigma,h>1. \tag{2.15}$$

Proof of Lemma 2.2. Let $P>2$; then we have

$$\begin{aligned}\int_0^1 v_x^P\,dx-\int_0^1 \phi_x^P\,dx&=\int_0^t\frac{d}{dt}\int_0^1 v_x^P\,dx\,d\tau\\&=\int_0^t Pv_x^{P-1}v_t|_0^1\,d\tau-\int\int_{Q_t}P(P-1)v_x^{P-2}v_{xx}v_t\,dx\,d\tau\\&=J_1+J_2.\end{aligned}$$

From assumptions (H1)-(H3) and (1.6), we see that there holds for a small $\varepsilon>0$

$$\begin{aligned}J_2&=-\int\int_{Q_t}P(P-1)v_x^{P-2}v_{xx}\left\{a_pv_{xx}+a_uv_x+r(t)a_x+r(t)F\left(x,t,\frac{v}{r},\frac{v_x}{r},\frac{-\dot r}{r}\right)\right\}dx\,d\tau\\&\leqq -a_0\int\int_{Q_t}P(P-1)v_x^{P-2}v_{xx}^2\,dx\,d\tau\\&\quad+\int\int_{Q_t}P(P-1)|v_x|^{p-2}|v_{xx}|(|v_x|+|v|+|r|+|\dot r|)\,dx\,d\tau\\&\leqq(\varepsilon-a_0)\int\int_{Q_t}P(P-1)v_x^{P-2}v_{xx}^2\,dx\,d\tau\\&\quad+C(\varepsilon)\int\int_{Q_t}P(P-1)|v_x|^{P-2}(|v_x|^2+|v|^2+|r|^2+|\dot r|^2)\,dx\,d\tau\end{aligned}$$

and by (2.12) and Lemma 2.1,

$$\begin{aligned}J_2&\leqq(\varepsilon-a_0)\int\int_{Q_t}P(P-1)v_x^{P-2}v_{xx}^2\,dx\,d\tau\\&\quad+C(\varepsilon)\int\int_{Q_t}P(P-1)(\|v_x(\cdot,\tau)\|_{L^\infty(0,1)}^{P}+\|v_x(\cdot,\tau)\|_{L^\infty(0,1)}^{P-2})\,dx\,d\tau.\end{aligned}$$

We know from (2.6) and (2.8) that

$$|J_1| \leqq C \int_0^t P(\|v(\cdot, \tau)\|_{L^\infty(0,1)}^P + \|v_x(\cdot, \tau)\|_{L^\infty(0,1)}^P)\, d\tau,$$

and then, we obtain from the estimates for $J_i's$ that

$$\begin{aligned} &\int_0^1 v_x^P\, dx + \iint_{Q_t} P(P-1) v_x^{P-2} v_{xx}^2\, dx\, d\tau \\ &\leqq CP^2 + C\int_0^1 \phi_x^P\, dx + C\int_0^T P^2 \|v_x(\cdot, \tau)\|_{L^\infty(0,1)}^P\, d\tau. \end{aligned} \tag{2.16}$$

Here we have used (2.2) to bound $\|v\|_{L^\infty(Q_T)}$ and $\|r\|_{L^\infty(0,T)}$.

If $\|v_x(\cdot, t)\|_{L^\infty(0,1)} \leqq \max\{1, \|\phi_x\|_{L^\infty(0,1)}\}$ for all $t \in (0, T)$, then we are done. Otherwise, it follows from the continuity of v_x that the interval $(0, T)$ can be decomposed into two parts: $(0, T) = G_T \cup L_T$, where

$$G_T = \{t \mid \|v_x(\cdot, t)\|_{L^\infty(0,1)} \geqq \max\{1, \|\phi_x\|_{L^\infty(0,1)}\},\ t \in (0, T)\},$$

$$L_T = \{t \mid \|v_x(\cdot, t)\|_{L^\infty(0,1)} < \max\{1, \|\phi_x\|_{L^\infty(0,1)}\},\ t \in (0, T)\},$$

and $\text{Meas}(G_T) > 0$. Thus,

$$1 + \int_0^1 \phi_x^P\, dx \leqq \frac{2}{\text{Meas}\,(G_T)} \int_{G_T} \|v_x(\cdot, \tau)\|_{L^\infty(0,1)}^P\, d\tau \leqq C\int_0^T \|v_x(\cdot, \tau)\|_{L^\infty(0,1)}^P\, d\tau,$$

and hence, (2.16) will become

$$\int_0^1 v_x^P\, dx + \iint_{Q_T} P(P-1) v_x^{P-2} v_{xx}^2\, dx\, d\tau \leqq C\int_0^T P^2 \|v_x(\cdot, \tau)\|_{L^\infty(0,1)}^P\, d\tau. \tag{2.17}$$

Here the constant C is dependent upon Meas (G_T).

Now we see from the interpolation inequality (2.14) that

$$\|v_x^{P/2}\|_{L^\infty(0,1)} \leqq C\|v_x^{P/2}\|_{H^1(0,1)}^{2/3} \|v_x^{P/2}\|_{L^1(0,1)}^{1/3},$$

and then, by Young's inequality (2.15) with $\sigma = \frac{3}{2}$ and $h = 3$,

$$\begin{aligned} \|v_x\|_{L^\infty(0,1)}^P &\leqq C\|v_x^{P/2}\|_{H^1(0,1)}^{4/3} \|v_x^{P/2}\|_{L^1(0,1)}^{2/3}, \\ &\leqq C\eta \|v_x^{P/2}\|_{H^1(0,1)}^2 + C\eta^{-2} \|v_v^{P/2}\|_{L^1(0,1)}^2. \end{aligned}$$

Since,

$$\|v_x^{P/2}\|_{H^1(0,1)}^2 = \frac{P^2}{4} \int_0^1 v_x^{P-2} v_{xx}^2\, dx + \int_0^1 v_x^P\, dx$$

and

$$\|v_x^{P/2}\|_{L^1(0,1)}^2 = \|v_x\|_{L^{P/2}(0,1)}^P,$$

we see that for $t \in [0, T]$,

$$\begin{aligned} &\int_0^1 v_x^P\, dx + P(P-1) \int_0^T \int_0^1 v_x^{P-2} v_{xx}^2\, dx\, dt \\ &\leqq CP(P-1)\left\{ P^2\eta \int_0^T \int_0^1 v_x^{P-2} v_{xx}^2\, dx\, dt + \eta \int_0^T \int_0^1 v_x^P\, dx\, dt \right\} \\ &\quad + CP^2\eta^{-2} \int_0^T \|v_x(\cdot, t)\|_{L^{P/2}(0,1)}^P\, dt. \end{aligned}$$

If we let $\eta = \min\{1/2CP^2, 1/2Tp(p-1)\}$, we obtain from the above inequality

$$\sup_{0<t<T} \int_0^1 v_x^P \, dx \leqq CP^4 \int_0^T \|v_x(\cdot, t)\|^P_{L^{P/2}(0,1)} \, dt \leqq CP^4 \sup_{0<t<T} \|v_x(\cdot, t)\|^P_{L^{P/2}(0,1)}. \tag{2.18}$$

Let $P = P_k = 2^k$ and $A_k = \sup_{0<t<T} (\int_0^1 v_x^{P_k} \, dx)^{1/P_k}$. We take the P_kth root of (2.18) and, by induction, obtain

$$A_k \leqq d_k A_{k-1} \leqq \cdots \leqq \left\{ \prod_{m=1}^{k} d_m \right\} A_1,$$

where $d_m = (CP_m^4)^{1/P_m}$. But, we know that

$$\prod_{m=1}^{k} d_m \leqq \exp\left\{ (\ln C + \ln 2) \sum_{m=1}^{\infty} \frac{4m+1}{2^m} \right\} < \infty,$$

so that $\prod_{m=1}^{\infty} d_m \leqq C$. Consequently, we can conclude from $\|v_x\|_{L^\infty(\bar{Q}_T)} = \lim_{k\to\infty} A_k$ and $A_1 \leqq C$ (by Lemma 2.1) that

$$\|v_x\|_{L^\infty(\bar{Q}_T)} \leqq C. \tag{2.19}$$

Then, by (2.6) it follows that $\|\dot{r}\|_{L^\infty(0,1)} \leqq C$. □

LEMMA 2.3. *There exist $C > 0$ and $\beta \in (0, \alpha)$ such that*

$$\|v\|_{C^{1+\beta}(\bar{Q}_T)} + \|r\|_{C^{1+\beta/2}([0,T])} \leqq C. \tag{2.20}$$

Proof. If we write equation (1.6) as the linear equation

$$v_t = a_p v_{xx} + B,$$

where

$$B(x, t) = a_u v_x + r(t) a_x + r(t) F\left(x, t, \frac{v}{r}, \frac{v_x}{r}, \frac{-\dot{r}}{r}\right)$$

and a_p are bounded functions by Lemma 2.2 and assumptions (H1)-(H3). Thus, we see from Theorem 6 of [13, p. 363] that there exist $\beta \in (0, \alpha)$ such that (2.20) holds. □

As a corollary of this lemma we have from Schauder's estimates [9], [14] that

$$\|v\|_{C^{2+\beta}(Q_T)} \leqq C(r_*), \tag{2.21}$$

where $r_* = \inf\{|r(t)| | t \in (0, T)\}$.

3. Existence and uniqueness of the solutions. In this section we shall use standard approximation techniques to prove the existence of our solution pair (v, r).

Let $\theta > 0$ be a small parameter. We define a sequence $\{v^\theta, r^\theta, \rho^\theta\}$ by

$$v_t^\theta = r^\theta(t) a\left(x, t, \frac{v^\theta}{r^\theta}, \frac{v_x^\theta}{r^\theta}\right)_x + r^\theta(t) F\left(x, t, \frac{v^\theta}{r^\theta}, \frac{v_x^\theta}{r^\theta}, \frac{-\rho^\theta}{r^\theta}\right) \quad \text{in } Q_T,$$

$$v^\theta(x, 0) = \phi(x), \qquad 0 < x < 1, \tag{3.1}$$

$$v^\theta(0, t) = r^\theta(t) f(t), \quad v^\theta(1, t) = r^\theta(t) g(t), \quad 0 < t < T,$$

and

$$r^\theta(t)=\frac{1}{E(t)}\int_0^{s(t)}\Phi(x,t)\bar v^\theta(x,t)\,dx,\qquad 0<t<T,\quad 0<s(t)\leqq 1, \tag{3.2}$$

$$\begin{aligned}\rho^\theta(t)=&-\frac{\dot E}{E^2}\int_0^{s(t)}\Phi\bar v^\theta\,dx+\frac{1}{E}\left\{\Phi(s(t),t)\bar v^\theta(s(t),t)\dot s(t)+\int_0^{s(t)}\Phi_t\bar v^\theta\,dx\right\}\\&+\frac{1}{E}r^\theta(t)\left\{\Phi(s(t),t)a\left(s(t),t,\frac{\bar v^\theta(s(t),t)}{r^\theta(t)},\frac{\bar v_x^\theta(s(t),t)}{r^\theta(t)}\right)\right.\\&\left.-\Phi(0,t)a\left(0,t,\frac{\bar v^\theta(0,t)}{r^\theta(t)},\frac{\bar v_x^\theta(0,t)}{r^\theta(t)}\right)\right\}\\&-\frac{1}{E}\int_0^{s(t)}\Phi_x(x,t)r^\theta(t)a\left(x,t,\frac{\bar v^\theta}{r^\theta},\frac{\bar v_x^\theta}{r^\theta}\right)dx\\&+\frac{1}{E}\int_0^{s(t)}\Phi(x,t)r^\theta(t)F\left(x,t,\frac{\bar v^\theta}{r^\theta},\frac{\bar v_x^\theta}{r^\theta},\frac{-\rho^\theta}{r^\theta}\right)dx,\end{aligned} \tag{3.3}$$

where

$$\bar v^\theta(x,t)=\begin{cases}\phi(x), & 0\leqq t\leqq\theta,\\ v^\theta(x,t-\theta), & \theta\leqq t<T.\end{cases} \tag{3.4}$$

In fact, ρ^θ is an approximation of the derivative $\dot r$ of r. We know from (3.4) that $\bar v^\theta$ is determined by initial data ϕ in $0\leqq t\leqq\theta$ and $r^\theta(0)=1$ by the compatibility assumption (H4). We see from (3.2)-(3.3) and assumption (H3) that $\{r^\theta,\rho^\theta\}$ is uniquely determined in $0\leqq t\leqq\theta$ by Lemma 1.2 of [5]. Then, we can solve (3.1) uniquely in the usual classical sense [9], [14] in $0\leqq t\leqq\theta$ and $v^\theta>0$ by the strong maximum principle [10, pp. 74-75]. This will determine $\bar v^\theta$ in $\theta\leqq t\leqq 2\theta$ by (3.4), and then the pair $\{r^\theta,\rho^\theta\}$ is determined in $\theta\leqq t\leqq 2\theta$ with $r^\theta>0,\cdots$. By induction, we have that problem (3.1)-(3.4) possesses a unique triple $\{v^\theta,r^\theta,\rho^\theta\}$ with $r^\theta>0$.

We now apply the arguments for Lemmas 2.1 and 2.2 to (3.1)-(3.3) and see that there exist $\beta\in(0,\alpha]$ and $C>0$ independent of θ,

$$\|v^\theta\|_{C^{1+\beta}(\bar Q_T)}+\|r^\theta\|_{C^{\beta/2}([0,T])}+\|\rho^\theta\|_{C^{\beta/2}([0,T])}\leqq C \tag{3.5}$$

and $\|v^\theta\|_{C^{2+\beta}(Q_T)}\leqq C(r_*^\theta)$, where $r_*^\theta=\inf\{r^\theta(t)|t\in(0,T)\}$. Thus, we see from compactness arguments that there exist $v\in C^{1+\beta}(\bar Q_T)$, $r\in C^{\beta/2}([0,T])$, and $\rho\in C^{\beta/2}([0,T])$, and a subsequence of $\{v^\theta,r^\theta,\rho^\theta\}$, also denoted by itself, such that

$$v^\theta\to v,\quad r^\theta\to r,\quad \rho^\theta\to\rho\quad\text{as }\theta\to 0, \tag{3.6}$$

and the convergence is uniform in $C^{1+\lambda}(\bar Q_T)\times C^{\lambda/2}([0,T])\times C^{\lambda/2}([0,T])$ if $\lambda<\beta$.

Unfortunately, we cannot take the limit in (3.1)-(3.3) as $\theta\to 0$ since $r^\theta>0$ will not guarantee that the limit function $r(t)>0$ for all $t\in[0,T)$. We must first show that $r(t)>0$ on $[0,T)$. Since $r^\theta(0)=1$, we see that $r(0)=1$. By continuity there exists at least a small time interval such that $r(t)>0$. Let $T^*\in(0,T)$ be defined by

$$T^*=\inf\{t|r(t)=0,\ t\in(0,T)\}>0. \tag{3.7}$$

Thus, by letting $\theta\to 0$ in (3.1)-(3.4) in $\Omega\times(0,T^*)$, we see that (v,r,ρ) will be a solution of (3.1)-(3.3) in Q_{T^*} with θ and the bar on v removed. From the strong maximum principle [10, pp. 74-75] and continuity of v, we see that $v(x,T^*)>0$ in $(0,1)$. Since

$$r(T^*)=\frac{1}{E(T^*)}\int_0^{s(T^*)}\Phi(x,T^*)v(x,T^*)\,dx>0$$

and $r(T^*)=0$ by (3.7), this contradiction implies that $r(t)>0$ in $[0,T]$.

Hence, the limit in (3.1)-(3.3) as $\theta\to 0$ can be taken and the triple (v,r,ρ) is actually a global solution for problem (3.1)-(3.3) without θ and the bar on v. This is

because $\|v^\theta\|_{C^{2+\beta}(Q_T)} \leqq C$, where $C>0$ is independent of θ, but depends on $r_* = \inf\{r(t) | t \in (0, t)\} > 0$ which depends upon the data. The identity $\dot{r} \equiv \rho$ in $[0, T]$ follows from differentiating (3.2) (with θ and the bar on u removed) and comparing $\dot{r}$ and ρ together with Lemma 2.1 of [4, p. 598].

We shall now summarize the above as the following theorem.

THEOREM 3.1. *Under assumptions* (H1)-(H4), *there exists a unique solution pair* (v, r) *with* $r>0$ *for problem* (1.6)-(1.7), *which is continuously dependent upon the data.*

Proof. The existence follows from the above argument. If we let (v^k, r^k) be the two solutions corresponding to the data $\{\phi^k, f^k, g^k, s^k, E^k, \Phi^k, F^k\}$ $(k=1, 2)$, then there exists an $M>0$ such that

$$\|v^k\|_{C^{2,1}(\bar{Q}_T)} + \|r^k\|_{C^{1+\lambda}([0,T])} \leqq M, \qquad k=1, 2,$$

where M depends only upon the data. Hence, it follows from an argument similar to that of [4] and [5] that there exists a $C>0$ dependent upon the data such that the following stability estimate holds:

$$\begin{aligned} &\|v^1 - v^2\|_{C^{1+\lambda}(\bar{Q}_T)} + \|r^1 - r^2\|_{C^{1+\lambda/2}([0,T])} \\ &\quad \leqq C\{\|f^1 - f^2\|_{C^{1+\lambda/2}([0,T])} + \|g^1 - g^2\|_{C^{1+\lambda/2}([0,T])} + \|s^1 - s^2\|_{C^{1+\lambda/2}([0,T])} \\ &\qquad + \|E^1 - E^2\|_{C^{1+\lambda/2}([0,T])} + \|\phi^1 - \phi^2\|_{C^{1+\lambda/2}([0,1])} \\ &\qquad + \|F^1 - F^2\|_{L^\infty(\bar{Q}_T \times [-N,N]^3)}\}, \end{aligned} \tag{3.8}$$

for some $0<\lambda \leqq \alpha<1$, where n is such that $|v/r|, |v_x/r|, |\dot{r}/r| \leqq N$ for all x and t. □

THEOREM 3.2. *Under assumptions* (H1)-(H4), *there exists a unique solution pair* (u, p) *for problem* (1.6)-(1.7), *which is continuously dependent upon the data.*

Proof. The proof follows from the transformation (1.4)-(1.5), Lemma 1.1, and Theorem 3.1. □

Remark 3.1. We see from the method developed in this paper that if we replace (1.1) by the more general form

$$u_t = a(x, t, u, u_x, q(t))_x + q(t)u + F(x, t, u, u_x, q(t)),$$

our results still hold if the following additional growth conditions are satisfied:

$$|a(x, t, u, p, q)| \leqq \delta_1 |q| + C(1 + |u| + |p|),$$

where $\delta_1 > 0$ is such that

$$0 \leqq \delta_1^* = \delta_1 \max_{0 \leqq t \leqq T} \left\{ E(t)^{-1} \left(2\|\Phi(\cdot, t)\|_{L^\infty(0,1)} + \int_0^{s(t)} |\Phi_x(x, t)|\, dx \right) \right\} < 1$$

and

$$0 \leqq \delta^* + \delta_1^* < 1.$$

Actually, these conditions are required to guarantee the unique solvability for $\dot{r}$ in the following equation:

$$\begin{aligned} \dot{r} = &-\frac{\dot{E}}{E^2} \int_0^{s(t)} \Phi v\, dx + \frac{1}{E} \left\{ \Phi(s(t), t) v(s(t), t) \dot{s}(t) + \int_0^{s(t)} \Phi_t v\, dx \right\} \\ &+ \frac{1}{E} r(t) \left\{ \Phi(s(t), t) a\left(s(t), t, \frac{v(s(t), t)}{r(t)}, \frac{v_x(s(t), t)}{r(t)}, \frac{-\dot{r}(t)}{r(t)} \right) \right. \\ &\left. - \Phi(0, t) a\left(0, t, \frac{v(0, t)}{r(t)}, \frac{v_x(0, t)}{r(t)}, \frac{-\dot{r}(t)}{r(t)} \right) \right\} \end{aligned}$$

$$-\frac{1}{E}\int_0^{s(t)} \Phi_x(x,t)r(t)a\left(x,t,\frac{v}{r},\frac{v_x}{r},\frac{-\dot{r}}{r}\right)dx$$

$$+\frac{1}{E}\int_0^{s(t)} \Phi(x,t)r(t)F\left(x,t,\frac{v}{r},\frac{v_x}{r},\frac{-\dot{r}}{r}\right)dx,$$

if r, v, and v_x are given.

Acknowledgment. The author thanks Professor J. R. Cannon for many comments, suggestions, and criticisms during the preparation of this paper.

REFERENCES

[1] J. R. CANNON, *The One-Dimensional Heat Equation*, Encyclopedia of Mathematics and Its Applications, Vol. 23, Addison-Wesley, Reading, MA, 1984.

[2] J. R. CANNON, S. P. ESTEVA, AND J. VAN DER HOEK, *A Galerkin procedure for the diffusion equation subject to the specification of mass*, SIAM J. Numer. Anal., 24 (1987), pp. 499–515.

[3] J. R. CANNON AND Y. LIN, *Determination of a parameter $p(t)$ in some quasi-linear parabolic differential equations*, Inverse Problems, 4 (1988), pp. 35–45.

[4] ———, *Determination of a parameter $p(t)$ in a Hölder class for some semi-linear parabolic equations*, Inverse Problems, 4 (1988), pp. 596–605.

[5] ———, *An inverse problem of finding a parameter in a semi-linear heat equation*, J. Math. Anal. Appl., 145 (1990), pp. 470–484.

[6] J. R. CANNON AND H. YIN, *On a class of non-classical parabolic problems*, J. Differential Equations, 79 (1989), pp. 266–288.

[7] K. L. DECKERT AND C. G. MAPLE, *Solution for diffusion with integral type boundary conditions*, Proc. Iowa Acad. Sci., 70 (1963), pp. 354–361.

[8a] L. C. EVANS, *A free boundary value problem: the flow of two immiscible fluids in a one-dimensional porous medium*, I, Indiana Univ. Math. J., 26 (1977), pp. 915–932.

[8b] ———, *A free boundary value problem: the flow of two immiscible fluids in a one-dimensional porous medium*, II, Indiana Univ. Math. J., 27 (1987), pp. 93–111.

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

[10] ———, *Variational Principle and Free Boundary Problems*, John Wiley, New York, 1982.

[11] N. I. IONKIN, *Solution of a boundary-value problem in heat conduction with a nonclassical boundary condition*, J. Differential Equations, 13 (1977), pp. 204–211.

[12] L. I. KAMYNIN, *A boundary value problem in the theory of heat conduction with a non-classical boundary condition*, U.S.S.R. Comput. Math. and Math. Phys., 4 (1964), pp. 33–59.

[13] S. N. KRUZKOV, *Nonlinear parabolic problems in two independent variables*, Trans. Moscow Math. Soc., 16 (1967), pp. 355–375.

[14] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URALECVA, *Linear and Quasi-linear Equations of Parabolic Type*, Amer. Math. Soc. Transl., 23, Providence, RI, 1968.

[15a] A. I. PRILEPKO AND D. G. ORLOVSKII, *Determination of the evolution parameter of an equation and inverse problems of mathematical physics*, I, J. Differential Equations, 21 (1985), pp. 119–129.

[15b] ———, *Determination of the evolution parameter of an equation and inverse problems of mathematical physics*, II, J. Differential Equations, 21 (1985), pp. 694–701.

[16] A. I. PRILEPKO AND V. V. SOLO'EV, *Solvability of the inverse boundary value problem of finding a coefficient of a lower order term in a parabolic equation*, J. Differential Equations, 23 (1987), pp. 136–143.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 157-168, January 1991

010

ENTIRE SOLUTIONS OF REAL AND COMPLEX MONGE-AMPÈRE EQUATIONS*

TAKAŜI KUSANO† AND CHARLES A. SWANSON‡

Abstract. Real and complex Monge-Ampère equations

$$\det\left(\frac{\partial^2 u}{\partial x_i \partial x_j}\right) = f(|x|, u, |\nabla u|), \qquad x \in \mathbb{R}^N, \quad N \geqq 2, \tag{A}$$

$$\det\left(\frac{\partial^2 u}{\partial z_i \partial \bar{z}_j}\right) = f(|z|, u, |\nabla u|), \qquad z \in \mathbb{C}^N, \quad N \geqq 2 \tag{B}$$

are considered in the entire spaces $\mathbb{R}^N$ and $\mathbb{C}^N$, respectively, $N \geqq 2$. A unified fixed-point approach is used to generate various conditions for (A) to have radial, strictly convex solutions $u(x)$ in $\mathbb{R}^N$ that are asymptotic to positive constant multiples of $|x|$ as $|x| \to \infty$, and for (B) to have radial, strictly plurisubharmonic solutions $u(z)$ in $\mathbb{C}^N$ that are asymptotic to positive constant multiples of $\log|z|$ as $|z| \to \infty$.

Key words. Monge-Ampère equation, entire solution, convex, plurisubharmonic, Gaussian curvature

AMS(MOS) subject classifications. 35J60, 35Q99

1. Introduction.

Our primary objective is to present a unified approach for establishing the existence and structure of radial entire solutions of real and complex Monge-Ampère equations

$$\det\left(\frac{\partial^2 u}{\partial x_i \partial x_j}\right) = f(|x|, u, |\nabla u|), \qquad x \in \mathbb{R}^N, \quad N \geqq 2, \tag{A}$$

$$\det\left(\frac{\partial^2 u}{\partial z_i \partial \bar{z}_j}\right) = f(|z|, u, |\nabla u|), \qquad z \in \mathbb{C}^N, \quad N \geqq 2, \tag{B}$$

respectively, $i, j = 1, \cdots, N$, under various hypotheses in §2 on the function $f \in C(\bar{\mathbb{R}}_+ \times \mathbb{R}_+ \times \bar{\mathbb{R}}_+, \mathbb{R}_+)$ or $f \in C(\bar{\mathbb{R}}_+ \times \mathbb{R} \times \bar{\mathbb{R}}_+, \mathbb{R}_+)$, where $\mathbb{R}_+ = (0, \infty)$, $\bar{\mathbb{R}}_+ = [0, \infty)$. As usual, $|x|$ and $|z|$ denote the Euclidean and Hermitian norms of points $x = (x_1, \cdots, x_N)$ and $z = (z_1, \cdots, z_N)$ in real and complex N-space $\mathbb{R}^N$ and $\mathbb{C}^N$, respectively, and ∇ denotes the gradient with respect to the coordinates in either space. An *entire solution* of (A) is defined to be a real-valued function $u \in C^2(\mathbb{R}^N)$ satisfying (A) at every point of $\mathbb{R}^N$. An entire solution of (B) is a real-valued function $u \in C^2(\mathbb{C}^N)$ satisfying (B) at every point of $\mathbb{C}^N$. Our attention will be directed toward *radial solutions* of (A) or (B), i.e., solutions that are functions of $t = |x|$ or $t = |z|$, respectively.

The main theorems in §2 contain various sufficient conditions for (A) to have infinitely many radial entire solutions $u(x)$, which are strictly convex and asymptotic to positive constant multiples of $|x|$ as $|x| \to \infty$, and for (B) to have infinitely many radial entire solutions $u(z)$, which are strictly plurisubharmonic and asymptotic to positive constant multiples of $\log|z|$ as $|z| \to \infty$. Such solutions are positive throughout $\mathbb{R}^N$ [or $\mathbb{C}^N$] under the hypotheses of Theorems 2.1A and 2.3A (or Theorems 2.1B and 2.3B, respectively). The proofs are given in §3 on the basis of the Schauder-Tychonov fixed-point theorem. Some variants of these results are included in §4.

* Received by the editors May 29, 1989; accepted for publication (in revised form) February 1, 1990.

† Department of Mathematics, Faculty of Science, Hiroshima University, Hiroshima 730, Japan. This author's research was partly supported by Japan's Ministry of Education Grant-in-Aid for Scientific Research 62302004.

‡ Department of Mathematics, University of British Columbia, Vancouver, British Columbia V6T 1Y4, Canada. This author's research was supported by National Sciences and Engineering Research Council of Canada grant A-3105.

Since their origin in geometry two centuries ago, Monge-Ampère equations have retained their central theoretical role in various geometric problems, including the problem of constructing manifolds with prescribed Gaussian curvature; see, for example, Bakel'man [3], Kazdan [9], and Pogorelov [17]-[19]. Among the numerous investigations of Monge-Ampère equations, we mention only the recent studies [1]-[10], [12]-[23]. Most of the literature, however, has been concerned primarily with solvability and regularity questions for boundary value problems in *bounded* domains, whereas information on Monge-Ampère equations in unbounded domains seems to be limited to results of Popivanov and Kutev [20] concerning the Neumann problem for (A) in exterior domains, and the present authors [11] for (A) in $\mathbb{R}^2$.

Standard calculations [7], [8], [16] show that the existence of a positive radial entire solution $u(x)=y(t)(t=|x|)$ of (A) is equivalent to the existence of a positive solution $y\in C^2(\bar{\mathbb{R}}_+)$ of the ordinary differential equation

$$([y'(t)]^N)'=Nt^{N-1}f(t,y(t),|y'(t)|),\qquad t>0 \tag{$1.1)_A$}$$

satisfying initial conditions $y(0)=c>0$ and $y'(0)=0$. A parallel statement for a positive radial entire solution $u(z)=y(t)(t=|z|)$ of (B) holds if $(1.1)_A$ is replaced by

$$([ty'(t)]^N)'=N2^{N+1}t^{2N-1}f(t,y(t),\tfrac{1}{2}|y'(t)|),\qquad t>0. \tag{$1.1)_B$}$$

In §3 we consider the more general problem of existence of solutions $y\in C^2(\bar{\mathbb{R}}_+)$ of the equation

$$([t^\alpha y'(t)]^N)'=t^{N(\alpha+1)-1}g(t,y(t),y'(t)),\qquad t>0 \tag{1.2}$$

satisfying $y(0)=c$, $y'(0)=0$, and $y'(t)>0$ for $t>0$, where $\alpha\in[0,1]$ is a constant. Evidently (1.2) is of type $(1.1)_A$ or $(1.1)_B$ when $\alpha=0$ or $\alpha=1$, respectively. Such a problem for (1.2) will be solved by obtaining an appropriate solution of the associated integrodifferential equation

$$y(t)=c+\int_0^t s^{-\alpha}\left[\int_0^s r^{N(\alpha+1)-1}g(r,y(r),y'(r))\,dr\right]^{1/N}ds,\qquad t\geqq 0 \tag{1.3}$$

by a fixed-point analysis.

2. Statement of theorems and examples. The existence of radial entire solutions of (A) and (B) will be proved under conditions on the function f selected from the list below. For convenience the domain of f is taken to be $\mathbb{D}=\bar{\mathbb{R}}_+\times\mathbb{R}\times\bar{\mathbb{R}}_+$; a restriction to $\bar{\mathbb{R}}_+\times\mathbb{R}_+\times\bar{\mathbb{R}}_+$ is understood if only positive solutions of (A) or (B) are being considered.

(f_1) $f(t,y,z)$ is positive and continuous in $\mathbb{D}$, and nondecreasing with respect to y and z.

(f_2) $k^{-N}f(t,ky,kz)$ is a nondecreasing function of k in some interval $(0,k_0]$ and $\lim_{k\to 0+}k^{-N}f(t,ky,kz)=0$ for each $(t,y,z)\in\mathbb{D}$.

(f_2') $k^{-N}f(t,ky,kz)$ is a nonincreasing function of k in some interval $[k_0,\infty)$ and $\lim_{k\to\infty}k^{-N}f(t,ky,kz)=0$ for each $(t,y,z)\in\mathbb{D}$.

(f_3) $\lim_{y\to-\infty}f(t,y,z)=0$ for each fixed (t,z).

(f_4) $\displaystyle\int_0^\infty t^{N-1}f(t,at,a)\,dt<\infty$ for some constant $a>0$.

(f_5) $\displaystyle\int_1^\infty t^{2N-1}f\left(t,a\log t,\frac{a}{t}\right)dt<\infty$ for some constant $a>0$.

THEOREM 2.1A. *If* (f_1), (f_4), *and one of* (f_2), (f_2') *hold, then* (A) *has an infinitude of positive radial strictly convex entire solutions* $u(x)$ *in* $\mathbb{R}^N$ *such that* $|x|^{-1}u(x)$ *has a positive finite limit as* $|x|\to\infty$.

THEOREM 2.1B. *If* (f_1), (f_5), *and one of* (f_2), (f_2') *hold, then* (B) *has an infinitude of positive radial strictly plurisubharmonic entire solutions* $u(z)$ *in* $\mathbb{C}^N$ *such that* $(\log|z|)^{-1}u(z)$ *has a positive finite limit as* $|z|\to\infty$.

It can be verified easily that the conclusion of Theorem 2.1A holds for the example

$$\det(\partial^2 u/\partial x_i\,\partial x_j) = p(|x|)u^\gamma + q(|x|)|\nabla u|^\delta, \qquad x\in\mathbb{R}^N, \tag{2.1}$$

where either $\gamma>1$ and $\delta>1$ or $\gamma<1$ and $\delta<1$, and p, q are positive continuous functions in $\bar{\mathbb{R}}_+$ such that

$$\int_0^\infty t^{N+\gamma-1}p(t)\,dt<\infty, \qquad \int_0^\infty t^{N-1}q(t)\,dt<\infty. \tag{2.2}$$

Likewise Theorem 2.1B applies to the complex Monge-Ampère equation

$$\det(\partial^2 u/\partial z_i\,\partial\bar z_j) = p(|z|)u^\gamma + q(|z|)|\nabla u|^\delta, \qquad z\in\mathbb{C}^N,$$

where γ, δ are as in (2.1), but (2.2) is replaced by

$$\int_1^\infty t^{2N-1}(\log t)^\gamma p(t)\,dt<\infty, \qquad \int_1^\infty t^{2N-\delta-1}q(t)\,dt<\infty. \tag{2.3}$$

THEOREM 2.2A. *Conditions* (f_1), (f_3), *and* (f_4) *imply that* (A) *has an infinitude of radial strictly convex entire solutions* $u(x)$ *in* $\mathbb{R}^N$ *such that* $|x|^{-1}u(x)$ *has a positive finite limit as* $|x|\to\infty$.

THEOREM 2.2B. *Conditions* (f_1), (f_3), *and* (f_5) *imply that* (B) *has an infinitude of radial strictly plurisubharmonic entire solutions* $u(z)$ *in* $\mathbb{C}^N$ *such that* $(\log|z|)^{-1}u(z)$ *has a positive finite limit as* $|z|\to\infty$.

Theorem 2.2A is applicable to the example

$$\det(\partial^2 u/\partial x_i\,\partial x_j) = p(|x|)\,e^u, \qquad x\in\mathbb{R}^N$$

for any positive continuous function p in $\bar{\mathbb{R}}_+$ satisfying

$$\int^\infty t^{N-1}e^{at}p(t)\,dt<\infty \quad \text{for some } a>0.$$

Similarly, Theorem 2.2B applies to the equation

$$\det(\partial^2 u/\partial z_i\,\partial\bar z_j) = p(|z|)e^u, \qquad z\in\mathbb{C}^N,$$

where p is a positive continuous function in $\bar{\mathbb{R}}_+$ satisfying

$$\int^\infty t^{2N-1+a}p(t)\,dt<\infty \quad \text{for some } a>0.$$

The final theorems in this section assert, in effect, that hypotheses (f_2), (f_2') can be deleted from Theorems 2.1A and 2.1B provided the integrals in (f_4) or (f_5) are sufficiently small. It is convenient to prove these results for the equations:

$$\det(\partial^2 u/\partial x_i\,\partial x_j) = \lambda f(|x|, u, |\nabla u|), \qquad x\in\mathbb{R}^N, \tag{A_λ}$$

$$\det(\partial^2 u/\partial z_i\,\partial\bar z_j) = \lambda f(|z|, u, |\nabla u|), \qquad z\in\mathbb{C}^N, \tag{B_λ}$$

where λ denotes a small positive parameter.

THEOREM 2.3A. *For every* λ *in some interval* $(0, \lambda_0]$, *conditions* (f_1) *and* (f_4) *are sufficient for* (A_λ) *to have an infinitude of positive radial entire solutions with the properties stated in Theorem* 2.1A.

THEOREM 2.3B. *For every* λ *in some interval* $(0, \lambda_0]$, *conditions* (f_1) *and* (f_5) *are sufficient for* (B_λ) *to have an infinitude of positive radial entire solutions with the properties stated in Theorem* 2.1B.

As an illustration, consider an equation of prescribed Gaussian curvature:

$$\det(\partial^2 u/\partial x_i\,\partial x_j) = \lambda p(|x|)(1+|\nabla u|^2)^{(N+2)/2}, \qquad x \in \mathbb{R}^N, \tag{2.4}$$

where p is a positive continuous function in $\bar{\mathbb{R}}_+$ satisfying

$$\int_0^\infty t^{N-1} p(t)\,dt < \infty. \tag{2.5}$$

If λ is sufficiently small, Theorem 2.3A implies that (2.4) has positive radial entire solutions $u(x)$ in $\mathbb{R}^N$ that are asymptotic to constant multiples of $|x|$ as $|x| \to \infty$. If $u(x)$ is any such solution of (2.4), it is obvious that $u(x)+K$ also is an entire solution of (2.4) for any constant K.

We note that if the exponent $(N+2)/2$ is replaced by $\nu \in (0, N/2)$, then (f_2') is satisfied, and hence Theorem 2.1A implies that (2.4) has, if (2.5) holds, positive radial entire solutions $u(x) \sim (\text{constant})|x|$ at ∞ for arbitrary $\lambda > 0$.

3. Proofs of the theorems. The proofs of parts A and B of the theorems in §2 will be given essentially together by considering the generalized equation (1.2). We employ the following notation throughout this section without further comment:

$$\phi_\alpha(t) = \begin{cases} 1/(1-\alpha)\max\{1, t^{1-\alpha}\} & \text{if } 0 \leqq \alpha < 1, \\ \max\{1, \log(et)\} & \text{if } \alpha = 1, \end{cases}$$

$$\psi_\alpha(t) = \min\{1, t^{-\alpha}\}, \qquad 0 \leqq \alpha \leqq 1.$$

Hypotheses on the function g in (1.2) will be selected from (g_1), (g_2), (g_2'), (g_3), defined to be identical to (f_1), (f_2), (f_2'), (f_3), respectively, with g replacing f; and we also adjoin the condition

$$\int_0^\infty t^{N(\alpha+1)-1} g(t, a\phi_\alpha(t), a\psi_\alpha(t))\,dt < \infty \tag{$g_{4+\alpha}$}$$

for some constant $a > 0$.

THEOREM 3.1. *If* (g_1), $(g_{4+\alpha})$ *and one of* (g_2), (g_2') *hold, then* (1.2) *has infinitely many positive solutions* $y \in C^2(\bar{\mathbb{R}}_+)$ *such that* $y'(0) = 0$ *and* $\lim_{t\to\infty} y(t)/\phi_\alpha(t)$ *is positive and finite.*

THEOREM 3.2. *If* (g_1), (g_3), *and* $(g_{4+\alpha})$ *hold, then* (1.2) *has infinitely many solutions* $y \in C^2(\bar{\mathbb{R}}_+)$ *such that* $y'(0) = 0$ *and* $\lim_{t\to\infty} y(t)/\phi_\alpha(t)$ *is positive and finite.*

Proof of Theorem 3.1. In view of $(g_{4+\alpha})$ and (g_2) or (g_2'), the dominated convergence theorem implies that

$$\lim_{c\to 0+} c^{-N} \int_0^\infty t^{N(\alpha+1)-1} g(t, c\phi_\alpha(t), c\psi_\alpha(t))\,dt = 0,$$

or

$$\lim_{c\to\infty} c^{-N} \int_0^\infty t^{N(\alpha+1)-1} g(t, c\phi_\alpha(t), c\psi_\alpha(t))\,dt = 0,$$

respectively. Thus, under either hypothesis (g_2) or (g_2'), there exists a positive constant c such that

$$\int_0^\infty t^{N(\alpha+1)-1} g(t, 2c\phi_\alpha(t), 2c\psi_\alpha(t))\, dt \leqq c^N. \tag{3.1}$$

Since $(g_{4+\alpha})$ implies a fortiori

$$\int_0^\infty t^{N-1} g(t, a\phi_\alpha(t), a\psi_\alpha(t))\, dt < \infty$$

for the same $a > 0$, virtually the same argument as for (3.1) ensures the existence of a constant $c > 0$ such that

$$\int_0^\infty t^{N-1} g(t, 2c\phi_\alpha(t), 2c\psi_\alpha(t))\, dt \leqq c^N. \tag{3.2}$$

Obviously c can be chosen such that both (3.1) and (3.2) are satisfied; in fact, there exists a continuum of such positive constants c.

Let $\mathscr{C}^1(\bar{\mathbb{R}}_+)$ denote the Fréchet space of all $\mathscr{C}^1$-functions in $\bar{\mathbb{R}}_+$ with the topology of uniform convergence of functions and their first derivatives on compact intervals of $\bar{\mathbb{R}}_+$. For a fixed constant c satisfying (3.1) and (3.2), we define a closed convex subset $\mathscr{Y}_\alpha$ of $\mathscr{C}^1(\bar{\mathbb{R}}_+)$ and a mapping $\mathscr{F}_\alpha : \mathscr{Y}_\alpha \to \mathscr{C}^1(\bar{\mathbb{R}}_+)$ by

$$\mathscr{Y}_\alpha = \{y \in \mathscr{C}^1(\bar{\mathbb{R}}_+) : c \leqq y(t) \leqq 2c\phi_\alpha(t),\ 0 \leqq y'(t) \leqq 2c\psi_\alpha(t),\ t \geqq 0\}, \tag{3.3}$$

$$\mathscr{F}_\alpha y(t) = c + \int_0^t s^{-\alpha} \left[\int_0^s r^{N(\alpha+1)-1} g(r, y(r), y'(r))\, dr \right]^{1/N} ds, \qquad t \geqq 0. \tag{3.4}$$

In order to conclude from the Schauder–Tychonov fixed-point theorem that there exists $y \in \mathscr{Y}_\alpha$ such that $\mathscr{F}_\alpha y = y$, we will now verify that $\mathscr{F}_\alpha$ maps $\mathscr{Y}_\alpha$ continuously into a relatively compact subset of $\mathscr{Y}_\alpha$.

If $y \in \mathscr{Y}_\alpha$ and $0 \leqq \alpha < 1$, then (3.1) and (g_1) imply that

$$c \leqq \mathscr{F}_\alpha y(t) \leqq c + \int_0^t s^{-\alpha} \left[\int_0^\infty r^{N(\alpha+1)-1} g(r, 2c\phi_\alpha(r), 2c\psi_\alpha(r))\, dr \right]^{1/N} ds$$

$$\leqq c + \frac{c}{1-\alpha} t^{1-\alpha} \leqq 2c\phi_\alpha(t), \qquad t \geqq 0.$$

If $y \in \mathscr{Y}_\alpha$ and $\alpha = 1$, we use (3.1), (3.2), and (g_1) to obtain

$$c \leqq \mathscr{F}_1 y(t) \leqq c + \int_0^t s^{-1} \left[\int_0^s r^{2N-1} g(r, 2c\phi_1(r), 2c\psi_1(r))\, dr \right]^{1/N} ds$$

$$\leqq c + \int_0^1 \left[\int_0^s r^{N-1} g(r, 2c\phi_1(r), 2c\psi_1(r))\, dr \right]^{1/N} ds$$

$$\leqq c + c = 2c\phi_1(t), \qquad 0 \leqq t \leqq 1,$$

$$c \leqq \mathscr{F}_1 y(t) \leqq c + \int_0^1 s^{-1} \left[\int_0^s r^{2N-1} g(r, 2c\phi_1(r), 2c\psi_1(r))\, dr \right]^{1/N} ds$$

$$+ \int_1^t s^{-1} \left[\int_0^s r^{2N-1} g(r, 2c\phi_1(r), 2c\psi_1(r))\, dr \right]^{1/N} ds$$

$$\leqq 2c + c \log t = c + c \log et \leqq 2c\phi_1(t), \qquad t \geqq 1.$$

Thus $c \leqq \mathscr{F}_\alpha y(t) \leqq 2c\phi_\alpha(t)$ for all $t \geqq 0$, $\alpha \in [0, 1]$. Moreover, if $y \in \mathscr{Y}_\alpha$, $0 \leqq \alpha \leqq 1$, also

$$\begin{aligned} 0 \leqq (\mathscr{F}_\alpha y)'(t) &= t^{-\alpha}\left[\int_0^t r^{N(\alpha+1)-1} g(r, y(r), y'(r))\, dr\right]^{1/N} \\ &\leqq \left[\int_0^t r^{N-1} g(r, 2c\phi_\alpha(r), 2c\psi_\alpha(r))\, dr\right]^{1/N} \\ &\leqq c \quad \text{for } 0 \leqq t \leqq 1, \\ 0 \leqq (\mathscr{F}_\alpha y)'(t) &\leqq t^{-\alpha}\left[\int_0^\infty r^{N(\alpha+1)-1} g(r, 2c\phi_\alpha(r), 2c\psi_\alpha(r))\, dr\right]^{1/N} \\ &\leqq ct^{-\alpha} \quad \text{for } t \geqq 1, \end{aligned}$$

implying that $0 \leqq (\mathscr{F}_\alpha y)'(t) \leqq 2c\psi_\alpha(t)$ for all $t \geqq 0$. Hence $\mathscr{F}_\alpha y \in \mathscr{Y}_\alpha$, and accordingly $\mathscr{F}_\alpha$ maps $\mathscr{Y}_\alpha$ into itself.

To prove the continuity of $\mathscr{F}_\alpha$, let $\{y_n\}$ be a sequence in $\mathscr{Y}_\alpha$ with $\lim_{n\to\infty} y_n = y \in \mathscr{C}^1(\bar{\mathbb{R}}_+)$ in the $\mathscr{C}^1(\bar{\mathbb{R}}_+)$-topology. We use the abbreviations $\nu = N(\alpha+1)-1$ and

$$G_n(t) = g(t, y_n(t), y_n'(t)), \qquad G(t) = g(t, y(t), y'(t)).$$

Then, for $0 \leqq \alpha \leqq 1$,

$$\begin{aligned} |(\mathscr{F}_\alpha y_n)(t) - (\mathscr{F}_\alpha y)(t)| &\leqq \int_0^t s^{-\alpha}\left|\left[\int_0^s r^\nu G_n(r)\, dr\right]^{1/N} - \left[\int_0^s r^\nu G(r)\, dr\right]^{1/N}\right| ds \\ &\leqq \int_0^t s^{-\alpha}\left[\int_0^s r^\nu |G_n(r) - G(r)|\, dr\right]^{1/N} ds \end{aligned}$$

for all $t \geqq 0$, and

$$\begin{aligned} |(\mathscr{F}_\alpha y_n)'(t) - (\mathscr{F}_\alpha y)'(t)| &\leqq t^{-\alpha}\left|\left[\int_0^t r^\nu G_n(r)\, dr\right]^{1/N} - \left[\int_0^t r^\nu G(r)\, dr\right]^{1/N}\right| \\ &\leqq t^{-\alpha}\left[\int_0^t r^\nu |G_n(r) - G(r)|\, dr\right]^{1/N} \\ &\leqq \left[\int_0^t r^{N-1} |G_n(r) - G(r)|\, dr\right]^{1/N}, \qquad t \geqq 0. \end{aligned}$$

In view of (g_1) and (3.1)-(3.3), the dominated convergence theorem implies that $(\mathscr{F}_\alpha y_n)(t) \to (\mathscr{F}_\alpha y)(t)$ and $(\mathscr{F}_\alpha y_n)'(t) \to (\mathscr{F}_\alpha y)'(t)$ as $n \to \infty$ uniformly on compact subintervals of $\bar{\mathbb{R}}_+$, establishing the continuity of $\mathscr{F}_\alpha$.

Local equicontinuity of the set $(\mathscr{F}_0\mathscr{Y}_0)' = \{(\mathscr{F}_0 y)' : y \in \mathscr{Y}_0\}$ is a consequence of the following inequality, holding for all $y \in \mathscr{Y}_0$, $0 \leqq t_1 < t_2 \leqq T < \infty$:

$$\begin{aligned} |(\mathscr{F}_0 y)'(t_2) - (\mathscr{F}_0 y)'(t_1)| &= \left|\left[\int_0^{t_2} r^\nu G(r)\, dr\right]^{1/N} - \left[\int_0^{t_1} r^\nu G(r)\, dr\right]^{1/N}\right| \\ &\leqq \left[\int_{t_1}^{t_2} r^\nu g(r, 2c\phi_0(r), 2c\psi_0(r))\, dr\right]^{1/N}. \end{aligned}$$

Then, for any compact interval $I = [0, T]$ and arbitrary $\varepsilon > 0$, there is a corresponding $\delta > 0$, independent of t_1, $t_2 \in I$ and $y \in \mathscr{Y}_0$, such that $|(\mathscr{F}_0 y)'(t_2) - (\mathscr{F}_0 y)'(t_1)| < \varepsilon$ for all t_1, $t_2 \in I$ with $|t_2 - t_1| < \delta$.

The proof of local equicontinuity of $(\mathscr{F}_\alpha \mathscr{Y}_\alpha)'$ for $\alpha \in (0, 1]$ requires the modification indicated below:

$$
\begin{aligned}
&|(\mathscr{F}_\alpha y)'(t_2) - (\mathscr{F}_\alpha y)'(t_1)| \\
&\quad = \left| t_2^{-\alpha} \left[\int_0^{t_2} r^\nu G(r)\, dr \right]^{1/N} - t_1^{-\alpha} \left[\int_0^{t_1} r^\nu G(r)\, dr \right]^{1/N} \right| \\
&\quad \leqq t_2^{-\alpha} \left| \left[\int_0^{t_2} r^\nu G(r)\, dr \right]^{1/N} - \left[\int_0^{t_1} r^\nu G(r)\, dr \right]^{1/N} \right| \\
&\qquad + |t_2^{-\alpha} - t_1^{-\alpha}| \left[\int_0^{t_1} r^\nu G(r)\, dr \right]^{1/N} \\
&\quad \leqq t_2^{-\alpha} \left[\int_{t_1}^{t_2} r^\nu G_\alpha(r)\, dr \right]^{1/N} + |t_2^{-\alpha} - t_1^{-\alpha}| \left[\int_0^{t_1} r^\nu G_\alpha(r)\, dr \right]^{1/N},
\end{aligned} \tag{3.5}
$$

where $G_\alpha(r) = g(r, 2c\phi_\alpha(r), 2c\psi_\alpha(r))$. Define

$$
K_\alpha = \left[\frac{1}{\nu+1} \max_{0 \leqq r \leqq T} G_\alpha(r) \right]^{1/N},
$$

and note that $\nu+1 = N(\alpha+1)$ and

$$
t_2^{\nu+1} - t_1^{\nu+1} \leqq (\nu+1) t_2^\nu (t_2 - t_1).
$$

Then (3.5) yields

$$
\begin{aligned}
|(\mathscr{F}_\alpha y)'(t_2) - (\mathscr{F}_\alpha y)'(t_1)| &\leqq K_\alpha \left[t_2^{-\alpha} (t_2^{\nu+1} - t_1^{\nu+1})^{1/N} + \frac{t_1^{\alpha+1}(t_2^\alpha - t_1^\alpha)}{t_1^{2\alpha}} \right] \\
&\leqq K_\alpha [(\nu+1)^{1/N} t_2^{1-1/N} (t_2 - t_1)^{1/N} + t_1^{1-\alpha} (t_2 - t_1)^\alpha],
\end{aligned}
$$

implying the equicontinuity of $(\mathscr{F}_\alpha \mathscr{Y}_\alpha)'$ on any compact interval I.

The local equicontinuity of $\mathscr{F}_\alpha \mathscr{Y}_\alpha$ can be verified more easily, and the local uniform boundedness of $\mathscr{F}_\alpha \mathscr{Y}_\alpha$ and $(\mathscr{F}_\alpha \mathscr{Y}_\alpha)'$ is clear. It then follows from Ascoli's theorem that $\mathscr{F}_\alpha \mathscr{Y}_\alpha$ has compact closure in $\mathscr{Y}_\alpha$.

Therefore we can apply the Schauder-Tychonov theorem to conclude that $\mathscr{F}_\alpha$ has a fixed point $y \in \mathscr{Y}_\alpha$. Clearly $y(t)$ satisfies (1.3), and hence also $y(t)$ is a solution of the original differential equation (1.2) such that $y(0) = c$, $y(t) > 0$ for $t \geqq 0$, $y'(0) = 0$, and

$$
y'(t) = t^{-\alpha} \left[\int_0^t s^{N(\alpha+1)-1} g(s, y(s), y'(s))\, ds \right]^{1/N} > 0 \quad \text{for } t > 0.
$$

It follows from this formula for $y'(t)$ and L'Hôpital's rule that $y \in C^2(\bar{\mathbb{R}}_+)$; in particular

$$
y''(0) = \lim_{t \to 0+} \frac{y'(t)}{t} = \left[\frac{g(0, c, 0)}{(\alpha+1)N} \right]^{1/N}. \tag{3.6}
$$

The proof can be given readily from the equations

$$
\frac{y'(t)}{t} = \left[t^{-N(\alpha+1)} \int_0^t s^{N(\alpha+1)-1} g(s, y(s), y'(s))\, ds \right]^{1/N}, \qquad t > 0,
$$

$$
\begin{aligned}
y''(t) = &-\alpha \left[t^{-N(\alpha+1)} \int_0^t s^{N(\alpha+1)-1} g(s, y(s), y'(s))\, ds \right]^{1/N} \\
&+ \frac{1}{N} g(t, y(t), y'(t)) \left[t^{-N(\alpha+1)} \int_0^t s^{N(\alpha+1)-1} g(s, y(s), y'(s))\, ds \right]^{(1-N)/N},
\end{aligned}
$$

$$
t > 0,
$$

in the limit $t \to 0+$. For these conclusions it is important that $g(0, c, 0) > 0$, as guaranteed by hypothesis (g_1) (see also (f_1)). We remark that (f_2) or (f_2') is easily possible under the positivity condition in (f_1); see example (2.1) as an illustration.

Using the above formula for $y'(t)$ (or (1.3)), we conclude that $t^\alpha y'(t)$ is nondecreasing for $t \geqq 1$ and bounded above on account of (3.1), (3.3), and the nondecreasing hypothesis in (f_1). Therefore the limit

$$\lim_{t\to\infty} \frac{y(t)}{\phi_\alpha(t)} = \lim_{t\to\infty} t^\alpha y'(t) = \left[\int_0^\infty s^{N(\alpha+1)-1} g(s, y(s), y'(s))\, ds\right]^{1/N}$$

is positive and finite. (For the left equality we used the now well-known general form of L'Hôpital's rule, not requiring that $\lim_{t\to\infty} |y(t)| = \infty$). This proves the asymptotic property in Theorem 3.1. An infinitude of such solutions of (1.2) exists corresponding to a continuum of allowable initial values c satisfying (3.1) and (3.2).

Proof of Theorem 3.2. On account of $(g_{4+\alpha})$ and the alternative hypothesis (g_3), the dominated convergence theorem yields

$$\lim_{b\to-\infty} \int_0^\infty t^{N(\alpha+1)-1} g(t, b + a\phi_\alpha(t), a\psi_\alpha(t))\, dt = 0,$$

and hence, as in (3.1) and (3.2), there exists a constant b_0 such that both

$$\int_0^\infty t^{N(\alpha+1)-1} g(t, b + a\phi_\alpha(t), a\psi_\alpha(t))\, dt \leqq a^N, \tag{3.7}$$

$$\int_0^\infty t^{N-1} g(t, b + a\phi_\alpha(t), a\psi_\alpha(t))\, dt \leqq a^N \tag{3.8}$$

for all $b \leqq b_0$. With (3.7) and (3.8) replacing (3.1) and (3.2), respectively, almost identical procedure to that used for Theorem 3.1 shows that the mapping (3.4) has a fixed point in the modified set

$$\mathcal{Y}_\alpha = \{y \in \mathcal{C}^1(\bar{\mathbb{R}}_+) : b \leqq y(t) \leqq b + a\phi_\alpha(t),\ 0 \leqq y'(t) \leqq a\psi_\alpha(t),\ t \geqq 0\}$$

for any $b \leqq b_0$. The proof is then completed as in Theorem 3.1.

Proof of Theorems 2.1(A, B) *and* 2.2(A, B). Equation (1.2) specializes to $(1.1)_A$ and $(1.1)_B$ in the respective cases

$$\alpha = 0, \qquad g(t, y, z) = Nf(t, y, z),$$

$$\alpha = 1, \qquad g(t, y, z) = N2^{N+1} f\left(t, y, \frac{z}{2}\right).$$

It is easy to see that the hypotheses $(f_1) - (f_5)$ in Theorems 2.1A, B and 2.2A, B imply the corresponding hypotheses in Theorems 3.1 and 3.2. In particular, (f_5) implies that

$$\int_e^\infty t^{2N-1} f\left(t, a \log t, \frac{a}{t}\right) dt < \infty$$

for some constant $a > 0$. Let $\tilde{a} = a/2$. Then for $t \geqq e$, since $g(t, y, z)$ is nondecreasing in y and z,

$$\begin{aligned} g(t, \tilde{a}\phi_1(t), \tilde{a}\psi_1(t)) &= g\left(t, \frac{a}{2}(1 + \log t), \frac{a}{2t}\right) \\ &\leqq g\left(t, a \log t, \frac{2a}{t}\right) \\ &= N2^{N+1} f\left(t, a \log t, \frac{a}{t}\right), \end{aligned}$$

and hence (g_5) holds. It follows that the conclusions of Theorems 3.1 and 3.2 apply to $(1.1)_A$ and $(1.1)_B$, implying the existence of radial entire solutions $u(x)=y(|x|)$ and $u(z)=y(|z|)$ of (A) and (B), respectively, with the asymptotic behavior described in Theorems 2.1A, B and 2.2A, B. The strict convexity of $u(x)=y(|x|)$ in part (A) of these theorems is a consequence of

$$\det(\partial^2 u/\partial x_i\,\partial x_j)=\left[\frac{y'(t)}{t}\right]^{N-1}y''(t)>0$$

for $t=|x|\geqq 0$. (For $t=0$, this is obtained from a limit as $t\to 0+$, as in (3.6)). The proof given by Delanoë [7, p. 339] shows that the solutions $u(z)$ in Theorems 2.1B and 2.2B are strictly plurisubharmonic in $\mathbb{C}^N$.

In order to prove Theorems 2.3A, B, we obtain positive radial entire solutions $u(x)=y(t)(t=|x|)$ of (A_λ) and $u(z)=y(t)(t=|z|)$ of (B_λ) as positive solutions $y\in C^2(\bar{\mathbb{R}}_+)$ of the ordinary differential equations

$$([y'(t)]^N)'=\lambda N t^{N-1}f(t,y(t),|y'(t)|),\qquad t>0,$$
$$([ty'(t)]^N)'=\lambda N 2^{N+1}t^{2N-1}f(t,y(t),\tfrac{1}{2}|y'(t)|),\qquad t>0,$$

respectively, containing a positive parameter λ. These equations are both of the type

$$([t^\alpha y'(t)]^N)'=\lambda t^{N(\alpha+1)-1}g(t,y(t),y'(t)),\qquad t>0 \tag{3.9}$$

for $\alpha=0$ or $\alpha=1$, respectively. Accordingly, the following result implies the truth of both Theorems 2.3A, B.

THEOREM 3.3. *If* (g_1) *and* $(g_{4+\alpha})$ *hold,* $0\leqq\alpha\leqq 1$, *then* (3.9), *for every* λ *in some interval* $(0,\lambda_0]$, *has an infinitude of positive solutions* $y\in C^2(\bar{\mathbb{R}}_+)$ *such that* $y'(0)=0$ *and* $\lim_{t\to\infty}y(t)/\phi_\alpha(t)$ *is positive and finite.*

Proof. In Theorem 3.1, condition (g_2) or (g_2') was needed only to obtain (3.1) and (3.2). However, since (3.9) contains a parameter λ, conditions (g_1) and $(g_{4+\alpha})$ are sufficient for the existence of a number $\lambda_0>0$ such that

$$\lambda_0\int_0^\infty t^{N(\alpha+1)-1}g(t,2c\phi_\alpha(t),2c\psi_\alpha(t))\,dt\leqq c^N,$$
$$\lambda_0\int_0^\infty t^{N-1}g(t,2c\phi_\alpha(t),2c\psi_\alpha(t))\,dt\leqq c^N$$

for all c in some compact subinterval of $(0,a)$, where a is the number in $(g_{4+\alpha})$ (preceding Theorem 3.1). The proof then proceeds as in Theorem 3.1 with λg replacing g.

4. Variations and extensions. The following variant of Theorem 3.1 applies to the equation

$$([t^\alpha y'(t)]^N)'=t^{N(\alpha+1)-1}g(t,y(t)),\qquad t>0 \tag{4.1}$$

for $0\leqq\alpha\leqq 1$, where now $g(t,y)$ is *nonincreasing* in y and satisfies

$$\int_0^\infty t^{N(\alpha+1)-1}g(t,a)\,dt<\infty \tag{4.2}$$

for some positive constant a.

THEOREM 4.1. *Suppose that* $g(t,y)$ *is positive and continuous in* $\bar{\mathbb{R}}_+\times\mathbb{R}_+$ *and nonincreasing with respect to* y *for fixed* t. *If* (4.2) *holds, then* (4.1) *has an infinitude of positive solutions* $y\in C^2(\bar{\mathbb{R}}_+)$ *such that* $y'(0)=0$ *and* $\lim_{t\to\infty}y(t)/\phi_\alpha(t)$ *is positive and finite.*

Proof. By the nonincreasing assumption for $g(t, y)$ and by (4.2), a constant $c > a$ can be selected large enough that both

$$\int_0^\infty t^{N(\alpha+1)-1} g(t, c)\, dt \leqq c^N \quad \text{and} \quad \int_0^\infty t^{N-1} g(t, c)\, dt \leqq c^N.$$

For such a number c, consider the mapping $\mathscr{F}_\alpha$ defined by

$$(\mathscr{F}_\alpha y)(t) = c + \int_0^t s^{-\alpha} \left[\int_0^s r^{N(\alpha+1)-1} g(r, y(r))\, dr \right]^{1/N} ds, \qquad t \geqq 0.$$

The procedure in Theorem 3.1 can then be used to verify that $\mathscr{F}_\alpha$ has a fixed point $y \in \mathscr{Y}_\alpha$, where $\mathscr{Y}_\alpha$ is given by (3.3), generating the stated solution in Theorem 4.1.

This theorem implies the following results for the real Monge–Ampère equation

$$\det (\partial^2 u / \partial x_i\, \partial x_j) = f(|x|, u), \qquad x \in \mathbb{R}^N \tag{4.3}$$

and the complex Monge–Ampère equation

$$\det (\partial^2 u / \partial z_i\, \partial \bar{z}_j) = f(|z|, u), \qquad z \in \mathbb{C}^N, \tag{4.4}$$

where f in (4.3) or (4.4) satisfies the respective conditions

$$\int_0^\infty t^{N-1} f(t, a)\, dt < \infty, \tag{4.5}$$

$$\int_0^\infty t^{2N-1} f(t, a)\, dt < \infty \tag{4.6}$$

for some positive constant a.

THEOREM 4.2A. *Suppose $f(t, u)$ is positive and continuous in $\bar{\mathbb{R}}_+ \times \mathbb{R}_+$ and nonincreasing with respect to u for fixed t. If* (4.5) *holds, then* (4.3) *has an infinitude of positive radial strictly convex entire solutions $u(x)$ in $\mathbb{R}^N$ such that $|x|^{-1} u(x)$ has a positive finite limit as $|x| \to \infty$.*

THEOREM 4.2B. *If $f(t, u)$ is as in Theorem* 4.2A *and* (4.6) *holds, then* (4.4) *has an infinitude of positive radial entire solutions $u(z)$ in $\mathbb{C}^N$ such that $(\log |z|)^{-1} u(z)$ has a positive finite limit as $|z| \to \infty$.*

Our results in §3 also can be applied to another class of complex Monge–Ampère equations

$$\det (\partial^2 u / \partial z_i\, \partial \bar{z}_j) = f(|z|, u, \Phi u), \qquad z \in \mathbb{C}^N, \tag{4.7}$$

where Φ denotes the operator defined by

$$\Phi = \sum_{j=1}^N \left(x_j \frac{\partial}{\partial x_j} + y_j \frac{\partial}{\partial y_j} \right), \qquad x_j = \operatorname{Re} z_j, \quad y_j = \operatorname{Im} z_j.$$

This type of equation has been studied by Derridj [8] and Popivanov and Kutev [21]. The ordinary differential equation for radial solutions $u(z) = y(t)$ $(t = |z|)$ of (4.7) is, instead of $(1.1)_B$,

$$([t y'(t)]^N)' = N 2^{N+1} t^{2N-1} f(t, y(t), t y'(t)), \qquad t > 0. \tag{4.8}$$

Application of Theorems 3.1 and 3.2 to (4.8) yields the following results under the hypothesis

$$\int_1^\infty t^{2N-1} f(t, a \log t, a)\, dt < \infty \quad \text{for some constant } a > 0. \tag{4.9}$$

THEOREM 4.3. *If* (4.9), (f_1), *and one of* (f_2), (f_2') *hold, then* (4.8) *has an infinitude of positive radial entire solutions* $u(z)$ *in* $\mathbb{C}^N$ *such that* $(\log|z|)^{-1}u(z)$ *has a positive finite limit as* $|z|\to\infty$.

THEOREM 4.4. *If* (4.9), (f_1), *and* (f_3) *hold, then* (4.8) *has an infinitude of radial entire solutions* $u(z)$ *in* $\mathbb{C}^N$ *such that* $(\log|z|)^{-1}u(z)$ *has a positive finite limit as* $|z|\to\infty$.

Our methods for (A) and (B) can be extended without essential change to establish the existence of radial entire solutions for *systems* of Monge-Ampère equations of any of the three types

(i) $$\det(\partial^2 u_k/\partial x_i\,\partial x_j)=f_k(|x|,u_1,\cdots,u_M,|\nabla u_1|,\cdots,|\nabla u_M|),$$
$$k=1,\cdots,M,\quad x\in\mathbb{R}^N,\quad N\geqq 2,$$

(ii) $$\det(\partial^2 u_k/\partial z_i\,\partial\bar z_j)=f_k(|z|,u_1,\cdots,u_M,|\nabla u_1|,\cdots,|\nabla u_M|),$$
$$k=1,\cdots,M,\quad z\in\mathbb{C}^N,\quad N\geqq 2,$$

(iii) $$\det(\partial^2 u_k/\partial z_i\,\partial\bar z_j)=f_k(|z|,u_1,\cdots,u_M,\Phi u_1,\cdots,\Phi u_M),$$
$$k=1,\cdots,M,\quad z\in\mathbb{C}^N,\quad N\geqq 2.$$

An example of the real Monge-Ampère equation (A) is

$$\det(\partial^2 u/\partial x_i\,\partial x_j)=2^N(2|x|^2+1)\exp[(N-\gamma)|x|^2]u^\gamma,\qquad x\in\mathbb{R}^N, \tag{4.10}$$

where γ is a positive constant. It is easily checked that $u(x)=\exp(|x|^2)$ is a positive radial entire solution of (4.10). If $\gamma>N$, then (f_1), (f_2), and (f_4) hold, and hence Theorem 2.1(A) shows that (4.10) has positive radial entire solutions $u(x)$ with the asymptotic behavior $u(x)\sim\omega|x|$ as $|x|\to\infty$ for some positive constant ω. Equation (4.10) then has positive entire solutions having different types of asymptotic behavior at infinity. A similar example is easy to construct for (B).

REFERENCES

[1] T. AUBIN, *Equations de Monge-Ampère réelles*, J. Funct. Anal., 41 (1981), pp. 345-377.

[2] ———, *Nonlinear Analysis on Manifolds, Monge-Ampère Equations*, Grundlehren Math. Wiss. 252, Springer-Verlag, Berlin, New York, 1982.

[3] I. YA. BAKEL'MAN, *Geometric Methods of Solutions of Elliptic Equations*, Nauka, Moscow, 1965. (In Russian.)

[4] E. BEDFORD AND B. A. TAYLOR, *The Dirichlet problem for a complex Monge-Ampère equation*, Invent. Math., 37 (1976), pp. 1-44.

[5] L. CAFFARELLI, L. NIRENBERG, AND J. SPRUCK, *The Dirichlet problem for nonlinear second-order elliptic equations* I. *Monge-Ampère equation*, Comm. Pure Appl. Math., 37 (1984), pp. 369-402.

[6] L. CAFFARELLI, J. J. KOHN, L. NIRENBERG, AND J. SPRUCK, *The Dirichlet problem for nonlinear second-order elliptic equations* II. *Complex Monge-Ampère and uniformly elliptic equations*, Comm. Pure Appl. Math., 38 (1985), pp. 318-344.

[7] Ph. DELANOË, *Radially symmetric boundary value problems for real and complex elliptic Monge-Ampère equations*, J. Differential Equations, 58 (1985), pp. 318-344.

[8] M. DERRIDJ, *Sur L'existence et la régularité de solutions radiales pour des équations de type Monge-Ampère complexe*, Math. Ann., 280 (1988), pp. 33-43.

[9] J. L. KAZDAN, *Prescribing the Curvature of a Riemannian Manifold*, CBMS-NSF Regional Conferences Series in Mathematics 57, American Mathematical Society, Providence, RI, 1985.

[10] N. V. KRYLOV, *Boundedly nonhomogeneous elliptic and parabolic equations in a domain*, Izv. Akad. Nauk. SSSR Ser. Mat., 47 (1983), pp. 75-108. (In Russian.)

[11] T. KUSANO AND C. A. SWANSON, *Existence theorems for elliptic Monge-Ampère equations in the plane*, Differential Integral Equations, 3 (1990), pp. 487-493.

[12] P. L. LIONS, *Sur les équations de Monge-Ampère* I, Manuscripta Math., 41 (1983), pp. 1-43.

[13] ———, *Sur les équations de Monge-Ampère* II, Arch. Rational Mech. Anal., 89 (1985), pp. 93-122.

[14] P. L. LIONS, *Two remarks on Monge-Ampère equations*, Ann. Mat. Pura Appl., 142 (1985), pp. 263-275.
[15] P. L. LIONS, N. S. TRUDINGER, AND J. I. E. URBAS, *The Dirichlet problem for the equation of prescribed Gauss curvature*, Bull. Austral. Math. Soc., 28 (1983), pp. 217-231.
[16] D. MONN, *Regularity of the complex Monge-Ampère equation for radially symmetric functions of the unit ball*, Math. Ann., 275 (1986), pp. 501-511.
[17] A. V. POGORELOV, *Monge-Ampère Equations of Elliptic Type*, Khar'kov University Press, Khar'kov, U.S.S.R. 1960. (Translated from Russian by L. F. Boron, P. Noordhoff, Groningen, the Netherlands, 1964.)
[18] ———, *Extrinsic Geometry of Convex Surfaces*, Nauka, Moscow, 1969. (Translated from Russian by Israel Program for Scientific Translations, American Mathematical Society, Providence, RI, 1973.)
[19] ———, *The Multidimensional Minkowski Problem*, John Wiley, New York, 1978.
[20] P. R. POPIVANOV AND N. D. KUTEV, *On solvability of the degenerate complex Monge-Ampère equation*, Dokl. Akad. Nauk SSSR, 301 (1988), pp. 1317-1320.
[21] ———, *Interior and exterior boundary value problems for the degenerate Monge-Ampère operator*, Hiroshima Math. J., 19 (1989), pp. 167-179.
[22] N. S. TRUDINGER AND J. I. E. URBAS, *The Neumann problem for equations of Monge-Ampère type*, Comm. Pure Appl. Math., 39 (1986), pp. 539-563.
[23] J. I. E. URBAS, *Regularity of generalized solutions of Monge-Ampère equations*, Math. Z., 197 (1988), pp. 365-393.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 169-180, January 1991

011

SYMMETRY AND BIFURCATION TO 2π-PERIODIC SOLUTIONS OF NONLINEAR SECOND-ORDER EQUATIONS WITH $2\pi/m$-PERIODIC FORCINGS*

M. FÜRKOTTER† AND H. M. RODRIGUES‡

Abstract. Consider the equation $\ddot{u}+u=g(u,p)+\mu f(t)$ where p, μ are small parameters, f is an even continuous $2\pi/m$-periodic function, $m\geqq 2$ is an integer, and g is an odd smooth nonlinear function of u. The main result is that, under certain conditions, the small 2π-periodic solutions maintain some symmetry properties of the forcing function $f(t)$, when $\mu\neq 0$. Other interesting results describe the changes of the number of such solutions as p and μ vary in a small neighborhood of the origin. The main tool used in this work is the Lyapunov-Schmidt method.

Key words. periodic solutions, symmetry, bifurcation, nonlinear equations, small solutions

AMS(MOS) subject classifications. primary 34A34, 34C15, 34C25

1. Introduction.

We consider the equation

$$\ddot{u}+u=g(u,p)+\mu f(t) \tag{1.1}$$

where p, μ are small parameters, f is an even continuous $2\pi/m$-periodic function, g is an odd function of u, sufficiently smooth, and $m\geqq 2$ is an integer.

Our main results are, under certain conditions on g and f, that the small 2π-periodic solutions of (1.1) maintain some symmetry properties of the forcing term $f(t)$, when $\mu\neq 0$. We also find the bifurcation curves and describe the changes of the number of such solutions, as (p,μ) varies in a small neighborhood of the origin. A conjecture which was stated in Fürkotter and Rodrigues [2] is proved.

Hale and Rodrigues [1], [5] by studying Duffing's equation $\ddot{u}+u=pu-u^3+\mu\cos t$, showed that the only small 2π-periodic solutions are even functions of t, if $\mu\neq 0$. They also stated the same result for a general even forcing function with minimal period 2π under the condition $\int_0^{2\pi} f(s)\cos s\,ds\neq 0$.

Rodrigues and Vanderbauwhede [6] generalized this result for equations such as (1.1) where f satisfies the former hypothesis and $g(u,p)=O(|pu|+u^2)$ as $(u,p)\to(0,0)$. They also presented an abstract version for equations in Banach spaces. Vanderbauwhede [7] also considers problems related to those above in an abstract form.

Fürkotter and Rodrigues [2] considered the case in which f is π-periodic, that is, $m=2$.

In this work we emphasize the case where f is $2\pi/m$-periodic for $m\geqq 3$, but we also make some comments about the case $m=2$ because, in some respects, the techniques presented here are different from the ones of the work above. The main features of these papers are to find a set of small 2π-periodic solutions of (1.1) and to prove that these are the only feasible solutions.

In § 2, using the Lyapunov-Schmidt method, we show that symmetries in (1.1) imply symmetries in the solutions of the auxiliary equation. We call special attention to Theorem 2.1 which plays a central role in this work.

* Received by the editors August 30, 1988; accepted for publication (in revised form) December 12, 1989. This research was partially supported by the Conselho Nacional de Desenvolvimento Cientifico e Tecnológico, Brazil (CNPq) under processo 301994/85-MA and the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

† Faculdade de Ciências e Tecnologia, UNESP, Presidente Prudente, SP, Brasil.

‡ Instituto de Ciências Matemáticas de São Carlos, USP, São Carlos, SP, Brasil.

In § 3, under certain conditions on $g(u, p)$ and on f, we prove that if $u(t)$ is a small 2π-periodic solution of (1.1) then there exists k, $-m/2 < k \leqq m/2$, such that $u(t + k\pi/m)$ is even in t, for (p, μ) small and $\mu \neq 0$. This is stated in Theorem 3.3, where it is required that a certain coefficient, $\rho = \rho(g, f)$, is nonzero.

Our results indicate that the bifurcation equation are more degenerate when more symmetries are present in (1.1).

At the end of § 3 we give some examples.

In § 4 we prove that the condition $\rho \neq 0$ is generic. It is also proved that ρ only depends on the coefficients of the Taylor expansion of $g(u, 0)$, around $u = 0$, up to the order $m+1$ if m is even and up to the order m if m is odd.

An application which can be reduced to (1.1) is the equation $\ddot{v} + \omega_0^2 v + G(v) = \sigma f(\omega t)$ where $G(v)$ is $O(v^2)$ as v tends to zero, f is an even $2\pi/m$-periodic function and we look for $2\pi/\omega$-periodic solutions for ω close to ω_0. This includes the pendulum equation and many other mechanical and electrical oscillators. If we let $u(t) \overset{\text{def}}{=} v(t/\omega)$ and $\omega_0^2/\omega^2 = 1 - p$, we get an equation like (1.1).

We thank a referee for pointing out that this problem could also be treated by using a complexification technique and representation of group $\boldsymbol{Z}_m$, such as appears in Golubitsky, Stewart, and Schaeffer [3].

A similar approach is used by Vanderbauwhede [8], where he considers a related problem.

We point out that, if we suppose that $f(t + \pi/m) = -f(t)$ for m even, the conditions of this work no longer hold and the bifurcation equations become more degenerate. This is a harder problem which will be presented in a future work. The case $m = 2$ is treated by Fürkotter and Rodrigues in [2].

2. The auxiliary equation. Consider equation (1.1)

$$\ddot{u} + u = g(u, p) + \mu f(t),$$

where (p, μ) varies in a small neighborhood of the origin, and the following hypothesis:

(A) f is a real $2\pi/m$-periodic, even function, continuous on R, and $m \geqq 2$ is an integer.

(B) g is a C^∞ real function defined in a neighborhood of $(u, p) = (0, 0)$, odd in u, and $g(u, p) = pu + \alpha u^3 + \beta u^5 + O(|pu^3| + |u|^7)$, as (u, p) goes to $(0, 0)$.

Let $\mathcal{P}$ be the space of all 2π-periodic real functions, continuous on R, with the norm $\|w\| = \sup_{0 \leqq t \leqq 2\pi} |w(t)|$, and let $\mathcal{P}^{(2)}$ be the space of all 2π-periodic real functions, with second derivative continuous on R, with the norm $\|w\| = \sup \{|w^{(j)}(t)|,\ 0 \leqq t \leqq 2\pi,\ j = 0, 1, 2\}$.

On these spaces we consider the projection

$$(Pw)(t) \overset{\text{def}}{=} \frac{\cos t}{\pi} \int_0^{2\pi} w(s) \cos s \, ds + \frac{\sin t}{\pi} \int_0^{2\pi} w(s) \sin s \, ds. \tag{2.1}$$

The Fredholm alternative implies that the equation $\ddot{u} + u = h(t)$, with h in $\mathcal{P}$, has a solution in $\mathcal{P}^{(2)}$ if and only if $Ph = 0$. Moreover, if $Ph = 0$ then there exists a unique solution $u(t)$ in $\mathcal{P}^{(2)}$ such that $Pu = 0$. We indicate this solution by $\mathcal{K}h$. From the variation of constants formula, we obtain

$$\mathcal{K}h = (I - P)\left[-\cos(\cdot) \int_0^{(\cdot)} h(s) \sin s \, ds + \sin(\cdot) \int_0^{(\cdot)} h(s) \cos s \, ds \right]. \tag{2.2}$$

Following the usual procedure of the Lyapunov-Schmidt method, the problem of finding a 2π-periodic solution $u(t)$ of (1.1) is reduced to that of finding a solution w in $\mathcal{P}^{(2)}$ of the following equations:

$$w = \mathcal{K}(I-P)[g(r\cos(\cdot-\phi)+w, p)+\mu f(\cdot)], \tag{2.3a}$$

$$P[g(r\cos(\cdot-\phi)+w, p)+\mu f(\cdot)] = 0 \tag{2.3b}$$

where $u(t) = r\cos(t-\phi)+w(t)$, $r\in R$ and $\phi\in(-\pi/2, \pi/2]$.

Equations (2.3a) and (2.3b) are called the auxiliary and the bifurcation equation, respectively. It follows from the implicit function theorem that (2.3a) has a unique small solution, for (p, μ) in a small neighborhood of the origin. We denote this solution by $w^*(r, \phi, p, \mu)(t)$. If we substitute in (2.3b) we obtain the following equivalent system of equations:

$$F(r, \phi, p, \mu) \overset{\text{def}}{=} \frac{1}{\pi}\int_0^{2\pi} g(r\cos s + w^*(r, \phi, p, \mu)(s+\phi), p)\cos s\, ds = 0, \tag{2.4a}$$

$$G(r, \phi, p, \mu) \overset{\text{def}}{=} \frac{1}{\pi}\int_0^{2\pi} g(r\cos s + w^*(r, \phi, p, \mu)(s+\phi), p)\sin s\, ds = 0. \tag{2.4b}$$

The following lemma gives information about some symmetries and estimates of w^*.

LEMMA 2.1. *If hypotheses* (A) *and* (B) *are satisfied, then the solution* w^* *of* (2.3a) *has the following properties*:

(2.5) $w^*(0, \phi, p, \mu)(t)$ *is an even* $2\pi/m$*-periodic function of* t *and is independent of* ϕ;

(2.6) $w^*(r, k\pi/m, p, \mu)(t+k\pi/m)$ *is even in* t *for* $-m/2 < k \leqq m/2$;

(2.7) $w^*(r, \phi, p, 0)(t+\phi)$ *is even in* t;

(2.8) $w^*(0, \phi, p, \mu) = \mu\mathcal{K}f + O(|p\mu|+|\mu^3|)$ *as* $(p, \mu)\to(0, 0)$;

(2.9) $w^*(r, \phi, p, \mu) = w^*(0, \phi, p, \mu) + rS(r, \phi, p, \mu)$
where $S(r, \phi, p, \mu) = O(r^2+|\mu|)$, *as* $(r, p, \mu)\to(0, 0, 0)$.

If m *is even then the following properties hold*:

$$w^*(r, \phi, p, \mu)(t) = w^*(-r, \phi, p, \mu)(t-\pi), \tag{2.10}$$

$$w^*(r, \phi, p, \mu)(t) = -w^*(r, \phi, p, -\mu)(t-\pi). \tag{2.11}$$

If $m = 3$, *the following hold*:

$$w^*(r, -\pi/3, p, \mu)(t-\pi/3) = w^*(r, \pi/3, p, \mu)(t+\pi/3), \tag{2.12}$$

$$w^*(-r, \pi/3, p, \mu)(t-2\pi/3) = w^*(r, 0, p, \mu)(t). \tag{2.13}$$

Proof. Properties (2.5)-(2.7), and (2.10)-(2.13) follow essentially from the fact that the auxiliary equation is invariant under certain transformations. Properties (2.8) and (2.9) can be proved in a natural way. □

Let $\mathcal{P}_{2\pi\times 2\pi} = \{f: R\times R\to R: f(t+2\pi, \phi) = f(t, \phi) = f(t, \phi+2\pi)$ for every $(t, \phi)\in R\times R$, f continuous$\}$ with the sup norm.

Let m and n be positive integers, with $m\geqq 2$ and $n\leqq m-1$. If n is even we define $\mathcal{F}_n$ as the space of the functions $y\in\mathcal{P}_{2\pi\times 2\pi}$, such that $y(t, \phi)$ can be written in the form:

$$\sum_{j=0}^{n/2}[a_{n,2j}(t)\cos 2j(t-\phi) + b_{n,2j}(t)\sin 2j(t-\phi)] \tag{2.14}$$

where $a_{n,2j}$ is an even $2\pi/m$-periodic function and $b_{n,2j}$ is an odd $2\pi/m$-periodic function, for $j = 0, \cdots, n/2$.

If n is an odd integer, we define $\mathcal{F}_n$ as the space of the functions $y \in \mathcal{P}_{2\pi\times 2\pi}$, such that $y(t, \phi)$ can be written in the form:

$$\sum_{j=0}^{(n-1)/2} [a_{n,2j+1}(t) \cos (2j+1)(t-\phi) + b_{n,2j+1}(t) \sin (2j+1)(t-\phi)] \tag{2.15}$$

where $a_{n,2j+1}$ is an even $2\pi/m$-periodic function and $b_{n,2j+1}$ is an odd $2\pi/m$-periodic function for $j = 0, 1, \cdots, (n-1)/2$.

Remark 2.1. In Lemmas 2.2–2.4 and Theorem 2.1 we allow ϕ to vary in R. To avoid picking up the same solution twice, in § 3 we restrict ϕ to $(-\pi/2, \pi/2]$.

LEMMA 2.2. *$\mathcal{F}_n$ is closed in $\mathcal{P}_{2\pi\times 2\pi}$.*

Proof. Let us assume first that n is even and that y is in $\mathcal{F}_n$. Then it has the form (2.14).

It is possible to prove that

$$a_{n,2j}(t) = \frac{1}{\pi} \int_0^{2\pi} y(t, \phi) \cos 2j(t-\phi)\, d\phi$$

and

$$b_{n,2j}(t) = \frac{1}{\pi} \int_0^{2\pi} y(t, \phi) \sin 2j(t-\phi)\, d\phi.$$

From this fact it follows that if a sequence $y_k \in \mathcal{F}_n$ converges in $\mathcal{P}_{2\pi\times 2\pi}$, then its limit is in $\mathcal{F}_n$.

If n is odd the proof is similar. □

LEMMA 2.3. *If q_i, n_i, β are positive integers, $y_i \in \mathcal{F}_{n_i}$, $i = 1, \cdots, \beta$ and $\alpha = \sum_{i=1}^{\beta} q_i n_i \leqq m-1$, then $\prod_{i=1}^{\beta} y_i^{q_i} \in \mathcal{F}_\alpha$.*

The next lemma plays an important role in this work.

LEMMA 2.4. *If m and n are positive integers, with $m \geqq 2$, $n \leqq m-1$, and $f \in (I-P)\mathcal{F}_n$ then $\mathcal{K}f$, that is, the function $(t, \phi) \mapsto \mathcal{K}f(\cdot, \phi)(t)$, belongs to $(I-P)\mathcal{F}_n$.*

Proof. Let us suppose first that n is even.

Our purpose is to prove that there exist coefficients $a_{n,2j}$, $b_{n,2j}$ in such a way that a function $x(t)$ given by (2.14) is a solution of $\ddot{x} + x = f(t, \phi)$ and $x \in (I-P)\mathcal{F}_n$. Since $\mathcal{K}f(\cdot, \phi)$ is the unique 2π-periodic solution which belongs to the range of $I-P$, it will follow that $x = \mathcal{K}f(\cdot, \phi)$.

If we substitute (2.14) into $\ddot{x} + x = f(t, \phi)$, where $f(t, \phi) \overset{\text{def}}{=} \sum_{j=0}^{n/2} [A_{n,2j}(t) \cos 2j(t-\phi) + B_{n,2j}(t) \sin 2j(t-\phi)]$ and equate coefficients, we obtain the equivalent system:

$$\ddot{a}_{n,2j} - (4j^2-1)a_{n,2j} + 4j\dot{b}_{n,2j} = A_{n,2j},$$
$$\ddot{b}_{n,2j} - (4j^2-1)b_{n,2j} + 4j\dot{a}_{n,2j} = B_{n,2j}$$

for $j = 0, \cdots, n/2$.

Now, if we let $y_1 = a_{n,2j}$, $\dot{y}_1 = y_2$, $y_3 = 3b_{n,2j}$, $\dot{y}_3 = y_4$, $y = \text{col}\,(y_1, y_2, y_3, y_4)$, we obtain the equivalent equation:

$$\dot{y} = C_j y + F_j$$

where $F_j = \text{col}\,(0, A_{n,2j}, 0, B_{n,2j})$ and

$$C_j = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 4j^2-1 & 0 & 0 & -4j \\ 0 & 0 & 0 & 1 \\ 0 & 4j & 4j^2-1 & 0 \end{bmatrix}$$

for $j = 0, \cdots, n/2$.

The eigenvalues of C_j are $\pm(2j+1)i$ and $\pm(2j-1)i$.

If m is even then the system $\dot{y} = C_j y$ is noncritical with respect to $\mathcal{P}_{2\pi/m}$, the space of $2\pi/m$-periodic continuous functions. The same holds if m is odd, $0 \leqq j \leqq n/2 \leqq (m-1)/2$, or $0 \leqq j \leqq n/2 < (m-1)/2$. If $j = n/2 = (m-1)/2$, then the system $\dot{y} = C_j y$ is critical with respect to $\mathcal{P}_{2\pi/m}$.

For the noncritical cases the equation $\dot{y} = C_j y + F_j$ has a unique $2\pi/m$-periodic solution $y(t) = \operatorname{col}(y_1(t), y_2(t), y_3(t), y_4(t))$. Since $z(t) \overset{\text{def}}{=} \operatorname{col}(y_1(-t), -y_2(-t), -y_3(-t), y_4(-t))$ is also a $2\pi/m$-periodic solution, it follows that $y_1(t)$ must be even and $y_3(t)$ must be odd functions of t.

For $j = n/2 = (m-1)/2$ we have a critical case. Following Hale [4, p. 275] we have that

$$\phi(t) \overset{\text{def}}{=} \begin{bmatrix} \cos mt & \sin mt \\ -m \sin mt & m \cos mt \\ \sin mt & -\cos mt \\ m \cos mt & m \sin mt \end{bmatrix}$$

is a matrix whose columns form a basis of the space of $2\pi/m$-periodic solutions of $\dot{y} = C_j(y)$, and

$$\Psi(t) = \begin{bmatrix} -(m-2)\sin mt & \cos mt & (m-2)\cos mt & \sin mt \\ (m-2)\cos mt & \sin mt & (m-2)\sin mt & -\cos mt \end{bmatrix}$$

is a matrix whose rows form a basis of the space of $2\pi/m$-periodic solutions of the adjoint equation, $\dot{z} = -zC_j$.

If we define projections $\bar{P}$, $\bar{Q}$ as in Hale [4, (2.5), p. 276], we obtain

$$\bar{P}f = \Phi(\cdot)\left[\int_0^{2\pi/m} \Phi'(t)\Phi(t)\,dt\right]^{-1} \int_0^{2\pi/m} \Phi'(t)f(t)\,dt,$$

$$\bar{Q}f = \psi'(\cdot)\left[\int_0^{2\pi/m} \Psi(t)\Psi'(t)\,dt\right]^{-1} \int_0^{2\pi/m} \Psi(t)f(t)\,dt.$$

$Pf(\cdot, \phi) = 0$ implies that $\bar{Q}F_j = 0$, for $j = (m-1)/2$.

Then $\dot{y} = C_j y + F_j$ has a unique $2\pi/m$-periodic solution $y(t)$ such that $(\bar{P}y)(t) \equiv 0$.

If $z(t) \overset{\text{def}}{=} \operatorname{col}(y_1(-t), -y_2(-t), -y_3(-t), y_4(-t))$, then $z(t)$ is also a solution of the same equation with $(\bar{P}z)(t) \equiv 0$. This implies that y_1 is even and $y_3(t)$ is odd.

The above information will provide a solution $x(t)$ of $\ddot{x} + x = f(t, \phi)$, and the condition $\bar{P}y = 0$ implies that $Px = 0$. This shows that $x \in (I-P)\mathcal{F}_n$ and that $x(t) = \mathcal{K}f(\cdot, \phi)(t)$.

The case n odd is similar.

LEMMA 2.5. *Let X be a Banach space, and let $I \subset R$ be an interval. Let $\xi: I \to X$ and $g: X \to X$ be functions with continuous derivatives up to the order n. Let $H = g \circ \xi$. Then for $n \geqq 1$, $\partial^n H/\partial r^n(r)$ can be written as a sum of terms of the form*

$$\gamma_i \frac{\partial^i g}{\partial u^i}(\xi(r))\left(\frac{d^{\alpha_1^i}\xi}{dr^{\alpha_1^i}}\right)^{\beta_1^i} \cdots \left(\frac{d^{\alpha_{k_i}^i}\xi}{dr^{\alpha_{k_i}^i}}\right)^{\beta_{k_i}^i}$$

where $\alpha_1^i\beta_1^i + \cdots + \alpha_{k_i}^i\beta_{k_i}^i = n$, $\beta_1^i + \cdots + \beta_{k_i}^i = i$ and γ_i are constants, for $i = 1, \cdots, n$. Moreover, if $i > 1$, then $\alpha_j^i < n$ and $\partial g/\partial u(\xi(r))d^n\xi/dr^n$ is the only term containing $d^n\xi/dr^n$.

The next theorem will be very important in the proof of our main results.

THEOREM 2.1. *Suppose hypotheses* (A) *and* (B) *are satisfied. If $1 \leqq n \leqq m-1$, then $\partial^n w^*/\partial r^n(0, \cdot, p, \mu)(\cdot)$ belongs to $(I-P)\mathcal{F}_n$. Moreover, it has the form* (2.14) *or* (2.15),

for m even or odd, respectively, with the coefficients $a_{ij}(t)=a_{ij}(p,\mu)(t)$ *and* $b_{ij}(t)=b_{ij}(p,\mu)(t)$.

Proof. We will do the proof by induction. If $n=1$ then $\partial w^*/\partial r(0,\cdot,p,\mu)(\cdot)$ is the unique solution of $\mathcal{H}y=y$, where $(\mathcal{H}y)(t,\phi)\overset{\text{def}}{=}\mathcal{K}(I-P)\times[\partial g/\partial u(w^*(0,\phi,p,\mu)(\cdot),p)(y+\cos(\cdot-\phi))](t)$ is a uniform contraction with respect to p, μ for (p,μ) in a small neighborhood of $(0,0)$.

From Lemmas 2.1 and 2.4, after some calculations we prove that $(I-P)\mathcal{F}_1$ is invariant under $\mathcal{H}$. Since, by Lemma 2.2, $(I-P)\mathcal{F}_1$ is closed in $\mathcal{P}_{2\pi\times 2\pi}$ it follows that the fixed point of $\mathcal{H}$ is in $(I-P)\mathcal{F}_1$.

Now let us assume that the result is true up to order $n-1$. We will prove that it is true for n.

$y=\partial^n w^*/\partial r^n(0,\cdot,p,\mu)$ is the unique solution of $y=\mathcal{H}y$ where

$$(\mathcal{H}y)(t,\phi)\overset{\text{def}}{=}\mathcal{K}(I-P)\left[\frac{\partial g}{\partial u}(w^*(0,\phi,p,\mu),p)y+T(\phi,p,\mu)\right](t),$$

and $T(\phi,p,\mu)$, by Lemma 2.5, can be written as a sum of terms of the form

$$\gamma_i\frac{\partial^i g}{\partial u^i}(w^*,p)\left(\cos(\cdot-\phi)+\frac{\partial w^*}{\partial r}\right)^{\beta_1^i}\left(\frac{\partial^{\alpha_2^i}w^*}{\partial r^{\alpha_2^i}}\right)^{\beta_2^i}\cdots\left(\frac{\partial^{\alpha_{k_i}^i}w^*}{\partial r^{\alpha_{k_i}^i}}\right)^{\beta_{k_i}^i}$$

where $i>1$ and $\partial^l w^*/\partial r^l$, above, means $\partial^l w^*/\partial r^l(0,\phi,p,\mu)$ for $l=0,1,\cdots,\alpha_{k_i}^i$. Moreover, $\alpha_j^i<n$, $\alpha_1^i\beta_1^i+\cdots+\alpha_{k_i}^i\beta_{k_i}^i=n$ and $\beta_1^i+\cdots+\beta_{k_i}^i=i$, for $j=1,\cdots,k_i$, $i=1,\cdots,n$.

As before $\mathcal{H}$ is a uniform contraction in $\mathcal{P}_{2\pi\times 2\pi}$ for (p,μ) in a small neighborhood of $(0,0)$.

From Lemmas 2.1, 2.3, and 2.4, after some calculations it follows that $(I-P)\mathcal{F}_n$ is closed in $\mathcal{P}_{2\pi\times 2\pi}$. Then the fixed point of $\mathcal{H}$ belongs to $(I-P)\mathcal{F}_n$. □

3. The bifurcation equations. Since in § 2 we obtained much information about the solution of the auxiliary equation, we now are able to analyze the bifurcation equations given in (2.4).

LEMMA 3.1. *Under hypotheses* (A) *and* (B) *the following hold*:

(i) $G(r,\phi,p,0)\equiv 0$.

(ii) *If m is even, then F and G are odd functions of r and even functions of* μ.

Proof. The first part follows from Lemma 2.1, (2.7) and the second part follows from Lemma 2.1, (2.10) and (2.11). □

THEOREM 3.1. *Suppose hypotheses* (A) *and* (B) *are satisfied. Then for* (r,p,μ) *in a small neighborhood of the origin,* $G(r,\phi,p,\mu)=r^{m-1}\mu\sin m\phi(\rho+\cdots)$, *if m is odd and* $G(r,\phi,p,\mu)=r^{m-1}\mu^2\sin m\phi(\rho+\cdots)$, *if m is even, where* ρ *is independent of* ϕ, p, μ *and* $(\cdots)$ *indicates terms of order* $O(|p|+|\mu|+|r|)$ *uniformly on* ϕ, *as* $(r,p,\mu)\to(0,0,0)$.

Proof. We will first prove that

$$\frac{\partial^l G}{\partial r^l}(0,\phi,p,\mu)=0,\qquad l=1,2,\cdots,m-2.$$

If we let $H(r,s)\overset{\text{def}}{=}g(r\cos(s-\phi)+w^*(r,\phi,p,\mu)(s),p)$, then

$$\frac{\partial^l G}{\partial r^l}(0,\phi,p,\mu)=\frac{1}{\pi}\int_0^{2\pi}\frac{\partial^l H}{\partial r^l}(0,s)\sin(s-\phi)\,ds.$$

From Lemma 2.5 it follows that $\partial^l H/\partial r^l(0,s)$ is a sum of terms of the form

$$\gamma_i\frac{\partial^i g}{\partial u^i}(w^*,p)\left(\cos(\cdot-\phi)+\frac{\partial w^*}{\partial r}\right)^{\beta_1^i}\left(\frac{\partial^{\alpha_2^i}w^*}{\partial r^{\alpha_2^i}}\right)^{\beta_2^i}\cdots\left(\frac{\partial^{\alpha_{\beta_{k_i}}^i}w^*}{\partial r}\right)^{\alpha_{\beta_{k_i}}^i}$$

where $\partial^q w^*/\partial r^q = \partial^q w^*/\partial r^q(0, \phi, p, \mu)(s)$, $q = 0$, $\alpha_1^i, \cdots, \alpha_{k_i}^i$, $\alpha_1^i\beta_1^i + \cdots + \alpha_{k_i}^i\beta_{k_i}^i = l$ and $\beta_1^i + \cdots + \beta_{k_i}^i = i$.

Let us assume first that l is even. From Theorem 2.1 and Corollary 2.1 it follows that $\partial^l G/\partial r^l(0, \phi, p, \mu)$ is a sum of integrals of the form

$$\int_0^{2\pi} [a(s)\cos 2j(s-\phi) + b(s)\sin 2j(s-\phi)]\sin(s-\phi)\,ds$$

where $0 \leqq j \leqq l/2 < (m-1)/2$ and $a(s)$, $b(s)$ are $2\pi/m$-periodic functions.

That integral can be written as

$$\frac{1}{2}\left\{-\sin(2j+1)\phi \int_0^{2\pi} [a(s)\cos(2j+1)s + b(s)\sin(2j+1)s]\,ds \right.$$
$$\left. + \sin(2j-1)\phi \int_0^{2\pi} [a(s)\cos(2j-1)s + b(s)\sin(2j-1)s]\,ds\right\}.$$

Since $j < (m-1)/2$ implies that $(2j+1) < m$ and since $a(s)$, $b(s)$ are $2\pi/m$-periodic it follows that the above integrals vanish.

The case l odd, $l < m-1$ is similar.

The same idea shows that $(\partial^{m-1}/\partial r^{m-1})G(0, \phi, p, \mu)$ is a sum of integrals of the form

$$-\frac{\sin m\phi}{2}\int_0^{2\pi} [a(s)\cos ms + b(s)\sin ms]\,ds$$

where $a(s) = a(p, \mu)(s)$, $b(s) = b(p, \mu)(s)$ are $2\pi/m$-periodic functions of s.

From Lemma 2.1, (2.6) it follows that $G(r, k\pi/m, p, \mu) \equiv 0$.

The proof can be completed by using Lemma 3.1 and the above results. □

Remark 3.1. To evaluate ρ it is sufficient to compute

$$\left(\frac{\partial^{m+1}G}{\partial r^{m-1}\partial\mu^2}(0, \phi, 0, 0)\right)\Big/\sin m\phi \quad \text{or} \quad \left(\frac{\partial^m G}{\partial r^{m-1}\partial\mu}(0, \phi, 0, 0)\right)\Big/\sin m\phi,$$

for m even or odd, respectively.

The case $m = 2$ is considered by Fürkotter and Rodrigues in [2]. The cases where $m = 3, 4, 5$ are considered with details in § 3 in the section of examples.

THEOREM 3.2. *Suppose* (A) *and* (B) *are satisfied and* $\rho \neq 0$. *Then the only small* 2π*-periodic solutions of* (1.1) *are such that* $u(t + k\pi/m)$ *is even in* t, *for some* k, $-m/2 < k \leqq m/2$, *for* (p, μ) *small and* $\mu \neq 0$.

Proof. Since $G(r, \phi, p, \mu) = r^{m-1}\mu \sin m\phi(\rho + \cdots)$ if m is odd, $G = 0$, $\mu \neq 0$ implies $r = 0$ or $\sin m\phi = 0$. Since $u(t) = r\cos(t-\phi) + w^*(r, \phi, p, \mu)(t)$, from Lemma 2.1 it follows that $u(t)$ is even if $r = 0$. If $\sin m\phi = 0$ then $\phi = k\pi/m$, $-m/2 < k \leqq m/2$. Still from Lemma 2.1 it follows that $u(t + k\pi/m) = r\cos t + w^*(r, k\pi/m, p, \mu)(t + k\pi/m)$ is even in t. □

In what follows we will assume that $\rho \neq 0$.

Now let us analyze the first bifurcation equation (2.4a),

$$F(r, \phi, p, \mu) = \frac{1}{\pi}\int_0^{2\pi} g(r\cos s + w^*(r, \phi, p, \mu)(s + \phi), p)\cos s\,ds = 0.$$

If we let $g(u, p) = pu + \alpha u^3 + O(|pu^3| + |u^5|)$ and since $F(0, \phi, p, \mu) \equiv 0$, after some calculations we obtain for $m > 3$,

$$F(r, \phi, p, \mu) = r(p + \tfrac{3}{4}\alpha r^2 + 3\alpha\eta r\mu + 3\alpha\lambda\mu^2 + \cdots) = 0$$

where

$$\lambda \overset{\text{def}}{=} \frac{1}{2\pi}\int_0^{2\pi}[(\mathcal{K}f)(t)]^2\,dt, \qquad \eta \overset{\text{def}}{=} \frac{\cos 3\phi}{4\pi}\int_0^{2\pi}(\mathcal{K}f)(t)\cos 3t\,dt$$

and $\cdots$ indicates higher-order terms.

We know that $r=0$ solves $F=0$ and $u(t)=w^*(0,\phi,p,\mu)(t)$ is $2\pi/m$-periodic in t.

In order to find other solutions, we consider $J \overset{\text{def}}{=} F/r$. To find multiple roots of $J=0$ we consider the system:

$$J(r,\phi,p,\mu)=p+\tfrac{3}{4}\alpha r^2+3\alpha\eta r\mu+3\alpha\lambda\mu^2+\cdots=0,$$

$$J_r(r,\phi,p,\mu)=\tfrac{3}{2}\alpha r+3\alpha\eta\mu+\cdots=0.$$

Since $\det(\partial(J,J_r)/\partial(p,r))=\frac{3}{2}\alpha$, for $r=p=\mu=0$, if $\alpha\neq 0$, from the implicit function theorem it follows that p and r can be found as functions of μ in a small neighborhood of the origin, for each fixed ϕ.

In what follows, the case $m=3$ requires a different treatment from the case $m>3$.

If $m=3$ the admissible values of ϕ are $\phi=0$ and $\phi=\pm\pi/3$. From Lemma 2.1, (2.12) it follows that $F(r,-\pi/3,p,\mu)=F(r,\pi/3,p,\mu)$, which implies that the bifurcation equations are the same for $\phi=-\pi/3$ and $\phi=\pi/3$.

Also from Lemma 2.1, (2.13) it follows that $F(-r,\pi/3,p,\mu)=-F(r,0,p,\mu)$, which implies $J(-r,\pi/3,p,\mu)=J(r,0,p,\mu)$. Therefore the bifurcation curve $p=p(\mu)$ for $\phi=0$ is the same as for $\phi=\pi/3$, while $r=r(\mu)$ changes sign.

The bifurcation curve for $\phi=0$ is given by $p=3\alpha(\eta^2-\lambda)\mu^2+O(|\mu|^3)$. The value of r where the bifurcation occurs is given by $r=-2\eta\mu+O(\mu^2)$.

If $m>3$ we assure that $J_r(0,\phi,p,\mu)\equiv 0$. This follows from Theorem 2.1 and Lemma 2.1.

It also follows from Theorem 2.1 that $J(0,\phi,p,\mu)$ is independent of ϕ.

Since $J(0,\phi,p,\mu)$ does not depend on ϕ and $J_r(0,\phi,p,\mu)\equiv 0$, it follows that the solution $p=p(\mu)$ which we obtain by solving $J(0,\phi,p,\mu)=0$ and $r=0$ is the unique solution of $J(0,\phi,p,\mu)=0$, $J_r(0,\phi,p,\mu)=0$ of r and p as functions of μ. Therefore for $m>3$ we have a unique bifurcation curve, which is given by

$$p=-3\alpha\lambda\mu^2+O(|\mu|^3).$$

The next theorem is very interesting and it describes the changes of the number of the small $2\pi/m$-periodic solutions of (1.1) as (p,μ) crosses the bifurcation curve.

THEOREM 3.3. *Suppose* (A) *and* (B) *are satisfied,* $m\geqq 3$, $\alpha\neq 0$ *and* ρ, *given in Remark* 3.1, *is nonzero. Then there exists a unique bifurcation curve* Γ, *which is given by* $p=-3\alpha\lambda\mu^2+O(\mu^3)$ *where* $\lambda=(1/2\pi)\int_0^{2\pi}[(\mathcal{K}f)(s)]^2\,ds$ *and* $\mathcal{K}f$ *is the* $2\pi/m$*-periodic solution of* $\ddot{u}+u=f(t)$, *if* $m>3$ *and by* $p=3\alpha(\eta^2-\lambda)\mu^2+O(\mu^3)$ *if* $m=3$, *where* $\eta^2=(\cos 3\phi/4\pi)\int_0^{2\pi}(\mathcal{K}f)(t)\cos 3t\,dt$.

The curve Γ *divides a neighborhood of the origin into regions as is shown in Figs.* 3.1 *and* 3.2 *for* $m>3$ *and* $m=3$, *respectively, for* α, $\lambda>0$, $\mu\neq 0$. *The number of* 2π*-periodic solutions of* (1.1) *is indicated in the figures.*

Examples. In what follows we will analyze the cases $m=3$, 4, 5. We recall that the case $m=2$ was studied in Fürkotter and Rodrigues [2] where we considered the example $f(t)=1+\cos 2t$. From the calculation that will be presented below, if $g(u,0)=\alpha u^3+\beta u^5+O(|u^7|)$, to compute the value of ρ for $m=3$, we only use α, while for $m=4$, 5 both α and β have contribution on the value of ρ.

For $m=3$, we have $\rho=-(3/4\pi)\alpha\int_0^{2\pi}(\mathcal{K}f)(t)\cos 3t\,dt$.

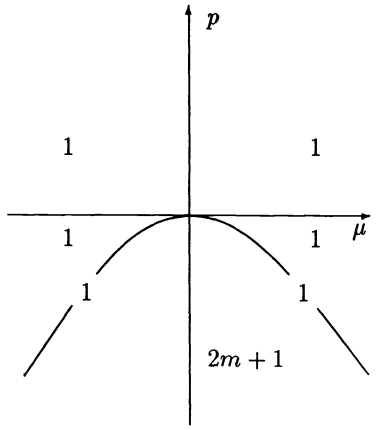


FIG. 3.1

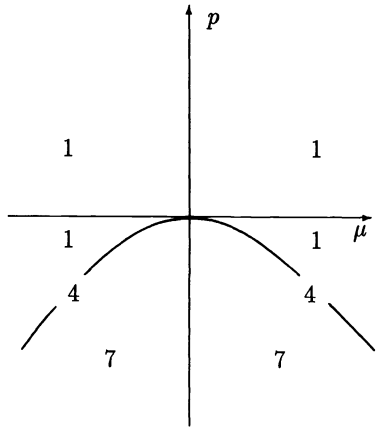


FIG. 3.2

For $m = 4$,

$$\rho = -\frac{9}{4\pi}\alpha^2 \int_0^{2\pi} \mathcal{K}(I-P)[\cos(\cdot)(\mathcal{K}f)^2](s)\cos 3s\, ds$$
$$-\frac{9}{4\pi}\alpha^2\left(\int_0^{2\pi} (\mathcal{K}f)(s)\mathcal{K}[\cos 2(\cdot)\mathcal{K}f](s)\cos 2s\, ds\right.$$
$$\left.-\int_0^{2\pi} (\mathcal{K}f)(s)\mathcal{K}[\sin 2(\cdot)\mathcal{K}f](s)\sin 2s\, ds\right)$$
$$+\frac{3}{64\pi}\alpha^2 \int_0^{2\pi} [(\mathcal{K}f)(s)]^2 \cos 4s\, ds - \frac{5\beta}{4\pi}\int_0^{2\pi} [(\mathcal{K}f)(s)]^2 \cos 4s\, ds.$$

For $m = 5$,

$$\rho = \frac{45}{64\pi}\alpha^2 \int_0^{2\pi} (\mathcal{K}f)(s)\cos 5s\, ds + \frac{3}{64\pi}\alpha^2 \int_0^{2\pi} (\mathcal{K}f)(s)\cos 5s\, ds$$
$$-\frac{5}{16\pi}\beta \int_0^{2\pi} (\mathcal{K}f)(s)\cos 5s\, ds.$$

Let us now consider the equation $\ddot{u}+u=g(u,p)+\mu f(t)$, where $g(u,p)=pu+\alpha u^3+\beta u^5+\cdots$, for some specific examples of $f(t)$:

(1) If $f(t)=\cos 3t$, then $\rho=3\alpha/32$, $\lambda=1/128$, and $\eta=-(1/32)\cos 3\phi$.

(2) If $f(t)=1+\cos 4t$, then $\rho=-\alpha^2/8+\beta/6$, $\lambda=1/450$, and $\eta=0$.

(3) If $f(t)=\cos 5t$, then $\rho=-(5/1024)\alpha^2+(5/384)\beta$, $\lambda=1/1152$, and $\eta=0$.

4. The genericity of the condition $\rho\neq 0$. In this chapter we will suppose that $g\in C^{m+2}$

$$g(u,0)=\sum_{i=1}^{(m-1)/2}\alpha_{2i+1}u^{2i+1}+O(|u|^{m+2}),$$

if m is odd and $g\in C^{m+3}$,

$$g(u,0)=\sum_{i=1}^{m/2}\alpha_{2i+1}u^{2i+1}+O(|u|^{m+3}),$$

if m is even. In both cases we suppose that $g(\cdot,p)$ is odd.

LEMMA 4.1. *Under the above assumptions and* (A), *if* $k\geqq 3$ *is an integer,* $k\leqq(m-1)/2$ *if* m *is odd* $k\leqq m/2$ *if* m *is even, then there exist continuous functions,* $W_1=W_1(\alpha_3,\cdots,\alpha_{2k-1},r,\phi,\mu,f)=O((|r|+|\mu|)^3)$, $W_2=W_2(r,\phi,\mu,f)=O((|r|+|\mu|)^{2k+1})$, *such that*

$$w^*(r,\phi,0,\mu)=\mu(\mathcal{K}f)+W_1+\alpha_{2k+1}W_2+O((|r|+|\mu|)^{2k+3}).$$

The proof of the above lemma can be done by induction on k.

LEMMA 4.2. *Under the assumptions of Lemma* 4.1, *the following hold*:

(1) *If* $m\geqq 4$ *is even, then there exists a continuous function* $K(\alpha_3,\cdots,\alpha_{m-1},f)$, *such that*

$$\rho=-\frac{m(m+1)}{2^m\pi}\alpha_{m+1}\int_0^{2\pi}[(\mathcal{K}f)(s)]^2\cos ms\,ds+K(\alpha_3,\cdots,\alpha_{m-1},f).$$

(2) *If* $m\geqq 5$ *is odd, then there exists a continuous function* $K(\alpha_3,\cdots,\alpha_{m-2},f)$, *such that*

$$\rho=-\frac{m}{2^{m-1}\pi}\alpha_m\int_0^{2\pi}(\mathcal{K}f)(s)\cos ms\,ds+K(\alpha_3,\cdots,\alpha_{m-2},f).$$

Proof. If m is even, from Lemma 4.1 it follows that

$$w^*=\mu\mathcal{K}f+W_1+\alpha_{m+1}W_2+\cdots$$

where $\cdots\overset{\text{def}}{=}O((|r|+|\mu|)^{m+3})$.

Therefore,

$$\begin{aligned}
G(r,\phi,0,\mu)&=\frac{1}{\pi}\int_0^{2\pi}\sum_{l=1}^{m/2}\alpha_{2l+1}(r\cos s+w^*(s+\phi))^{2l+1}\sin s\,ds+\cdots\\
&=\frac{1}{\pi}\int_0^{2\pi}\alpha_{m+1}(r\cos s+w^*(s+\phi))^{m+1}\sin s\,ds\\
&\quad+\frac{1}{\pi}\int_0^{2\pi}\sum_{l=1}^{(m-2)/2}\alpha_{2l+1}(r\cos s+w^*(s+\phi))^{2l+1}\sin s\,ds+\cdots\\
&=\frac{1}{\pi}\int_0^{2\pi}\alpha_{m+1}[r\cos s+\mu(\mathcal{K}f)(s+\phi)]^{m+1}\sin s\,ds\\
&\quad+\frac{1}{\pi}\int_0^{2\pi}\sum_{l=1}^{(m-2)/2}\alpha_{2l+1}[r\cos s+\mu(\mathcal{K}f)(s+\phi)+W_1]^{2l+1}\sin s\,ds+\cdots.
\end{aligned}$$

From Theorem 3.1, it follows that

$$G(r, \phi, p, \mu) = r^{m-1}\mu^2 \sin m\phi[\rho + O(|p|+|\mu|+|r|)].$$

Therefore, to determine ρ it suffices to consider the part of G, which contains terms involving $r^{m-1}\mu^2$, for $p=0$.

Let us first consider the term

$$\frac{\alpha_{m+1}}{\pi}\int_0^{2\pi}[r\cos s + \mu(\mathcal{K}f)(s+\phi)]^{m+1}\sin s\, ds.$$

The part of this term that involves $r^{m-1}\mu^2$ is given by

$$\frac{\alpha_{m+1}}{\pi}\frac{m(m+1)}{2}r^{m-1}\mu^2\int_0^{2\pi}\cos^{m-1}s[(\mathcal{K}f)(s+\phi)]^2\sin s\, ds.$$

By induction we can prove that

$$\cos^{m-1} s \sin s = \frac{1}{2^{m-1}}\Big\{[\sin ms - \sin(m-2)s] + \sum_{j=1}^{(m-2)/2}\beta_j[\sin(m-2j)s - \sin(m-2(j+1))s]\Big\}$$

where β_j are constants.

Since $\mathcal{K}f$ is $2\pi/m$-periodic, we have that

$$\int_0^{2\pi}[(\mathcal{K}f)(s+\phi)]^2\sin ks\, ds = 0 \quad \text{if } 0 \leqq k < m.$$

Therefore,

$$\frac{\alpha_{m+1}}{\pi}\frac{m(m+1)}{2}r^{m-1}\mu^2\int_0^{2\pi}[(\mathcal{K}f)(s+\phi)]^2\cos^{m-1}s\sin s\, ds$$
$$= -\frac{\alpha_{m+1}}{2^m\pi}m(m+1)r^{m-1}\mu^2\sin m\phi\int_0^{2\pi}[(\mathcal{K}f)(s)]^2\cos ms\, ds.$$

If we define $K(\alpha_3, \cdots, \alpha_{m-1}, f)$ as the coefficient of the term $r^{m-1}\mu^2\sin m\phi$ obtained from

$$\frac{1}{\pi}\int_0^{2\pi}\sum_{l=1}^{(m-2)/2}\alpha_{2l+1}[r\cos + \mu(\mathcal{K}f)(s+\phi) + W_1]^{2l+1}\sin s\, ds,$$

then we conclude that ρ has the stated form.

The second part of our lemma is similar. □

THEOREM 4.1. *Under the assumptions on g, in the beginning of this section, the condition $\rho \neq 0$ is generic.*

Proof. Let us consider the case $m \geqq 2$, even. The case $m=2$ is simple. For $m \geqq 4$ we have, from Lemma 4.2 that

$$\rho = \alpha_{m+1}c_m\int_0^{2\pi}[(\mathcal{K}f)(s)]^2\cos ms\, ds + K(\alpha_3, \cdots, \alpha_{m-1}, f)$$

where $c_m \neq 0$.

Without loss of generality, we can assume that

$$\int_0^{2\pi}[(\mathcal{K}f)(s)]^2\cos ms\, ds \neq 0.$$

If $\rho(f, g) = 0$, then we consider $\bar{g}(u) = g(u, 0) + \varepsilon u^{m+1}$. Then $\rho(f, \bar{g}) = \varepsilon c_m \int_0^{2\pi} [(\mathcal{H}f)(s)]^2 \cos ms \, ds \neq 0$ and $\bar{g}$ is close to g in the c^{m+3}-topology, if ε is small.

The remaining part of the proof is easy.

The case m odd is similar. □

REFERENCES

[1] S. N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, Berlin, 1982.

[2] M. FÜRKOTTER AND H. M. RODRIGUES, *Periodic solutions of forced nonlinear second order equations: symmetry and bifurcations*, SIAM J. Math. Anal., 17 (1986), pp. 1319-1331.

[3] M. GOLUBITSKY, I. N. STEWART, AND D. G. SCHAEFFER, *Singularities and Groups in Bifurcation Theory*, Vol. II, Applied Mathematical Science Series 69, Springer-Verlag, New York, 1988.

[4] J. K. HALE, *Ordinary Differential Equations*, Krieger, New York, 1980.

[5] J. K. HALE AND H. M. RODRIGUES, *Bifurcation in the Duffing equation with independent parameters* II, Proc. Roy. Soc. Edinburgh Sect. A, 79 (1977), pp. 317-326.

[6] H. M. RODRIGUES AND A. VANDERBAUWHEDE, *Symmetric perturbations of nonlinear equations: symmetry of small solutions*, Nonlinear Anal. Theory Methods Appl., 2 (1978), pp. 27-46.

[7] A. VANDERBAUWHEDE, *Local Bifurcation and Symmetry*, Pitman, Boston, 1982.

[8] ———, *Bifurcation of subharmonic solutions in time-reversible systems*, I, J. Appl. Math. Phys., 37 (1986), pp. 455-477.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 181-212, January 1991

012

ANALYSIS AND COMPUTATION OF SYMMETRY-BREAKING BIFURCATION AND SCALING LAWS USING GROUP THEORETIC METHODS*

P. J. ASTON†

Abstract. Group theoretic methods are used to analyse symmetry-breaking bifurcation for nonlinear equations defined on a real Hilbert space. An important result is the decomposition of the Hilbert space into orthogonal isotypic components, since the Jacobian of the nonlinear operator can be decomposed on the isotypic components. This decomposition is exploited in the detection and computation of bifurcation points. Then scaling laws that arise in many problems are considered, and a natural context is developed for the existence of a scaling law based on the symmetry of the problem. The effect of the scaling law on the bifurcation theory is explored. This theory is applied to the gravity wave problem. Also shown is the way in which the theory can extend to boundary value problems, where the natural group equivariance of the equations is destroyed by the boundary conditions.

Key words. symmetry-breaking bifurcation, symmetry groups, isotypic components, scaling laws, gravity waves

AMS(MOS) subject classifications. 58F14, 22E70

1. Introduction. In this paper, we use group theoretic methods to analyse symmetry-breaking bifurcation for problems defined on a real Hilbert space. Sattinger (1977), (1979) pioneered work in this field, developing much of the theory in a complex Banach space. Fujii, Mimura, and Nishiura (1982) extended this theory by considering the standard, or isotypic, decomposition of a complex Hilbert space. Bossavit (1986) describes in detail how this isotypic decomposition can be employed when solving linear boundary value problems on a complex Hilbert space. However, there are some fundamental differences between group representation theory in real and complex spaces, most notably the corollary to Schur's lemma (Corollary 2.9), and so not all of these results generalise naturally to the real Hilbert space setting. Vanderbauwhede (1982) considered such problems in real Banach spaces but avoided the differences between the theory in real and complex spaces.

Golubitsky, Stewart, and Schaeffer (1988) have considered in detail the analysis of bifurcation in real finite-dimensional spaces on the premise that problems defined on an infinite-dimensional space can often be reduced to finite dimensions using the Lyapunov–Schmidt procedure. However, our aim is to develop efficient numerical methods for the detection and direct computation of symmetry-breaking bifurcation points. Thus we work with a nonlinear, parameter-dependent equation in a real Hilbert space, and investigate properties of the equation and the Jacobian of the nonlinear operator that can be exploited numerically. As bifurcation from a branch of nontrivial solutions cannot be analysed analytically, in general, we develop the theory assuming certain generic conditions.

In § 2, we present systematically the group theoretic results required later on, culminating in the main result of the section, Theorem 2.11, which gives the isotypic decomposition of a real Hilbert space. We turn our attention to bifurcation problems in § 3 and apply the results of the previous section, assuming that the problem is equivariant with respect to a representation of a compact Lie group. We show that the

* Received by the editors November 8, 1989; accepted for publication (in revised form) June 3, 1990. This work was supported by the Science and Engineering Research Council.

† Department of Mathematics, University of Surrey, Guildford GU2 5XH, United Kingdom.

Jacobian can be decomposed on the isotypic components and exploit this decomposition in the detection and direct computation of bifurcation points. Other authors (Werner (1988), Healey (1988b), Dellnitz and Werner (1989)) have considered the computational advantage of restricting to fixed-point subspaces. However, for the determination of bifurcation points, we will see that the use of isotypic components is a much more powerful tool.

In § 4, we consider scaling laws which arise in many problems. A natural context for the existence of a scaling law is developed based on the symmetry of the problem, and the effect of the scaling law on the bifurcation theory of § 3 is then explored. Often one branch of solutions is related to another by a simple scaling, and so there is no need to compute such a branch. However, we show that bifurcation points can occur on a scaled branch which do not exist on the original branch. We also show how the use of a scaling law can lead to a proof of existence of bifurcating branches at a mode interaction point in some cases.

Finally, in § 5, we apply the preceding theory to the gravity wave problem on a fluid of infinite depth. We also show how the theory can be extended to boundary value problems where the natural group equivariance of the equations is destroyed by the boundary conditions.

2. Group theoretic results. In this section we collect the group theoretic results required in later sections. We develop the theory systematically for the sake of numerical analysts who may not be familiar with it. We consider the group representation theory on a *real* Hilbert space, whereas most textbooks work with complex Hilbert spaces. Many of the results for real spaces are proved in essentially the same way as the corresponding results on complex spaces, and so we do not give these proofs, but refer the reader to Barut and Raczka (1986) or Bröcker and Tom Dieck (1985) (which we will henceforth abbreviate as [BR] and [BtD], respectively). However, there are fundamental differences in the theory of irreducible representations on real and complex spaces (see Corollary 2.9 and the following assertions), and so we then prove the relevant results culminating in our main result of the section, Theorem 2.11, which gives the decomposition of a real Hilbert space into its orthogonal isotypic components.

We will consider only *compact Lie groups.* If Γ is a compact Lie group, we say that Σ is a *subgroup* of Γ if $\sigma\delta^{-1} \in \Sigma$, for all $\sigma, \delta \in \Sigma$, and Σ is closed in Γ so that Σ is also a Lie group [BtD, p. 28]. Also, Σ is a *normal subgroup* of Γ if it is a subgroup and $\gamma\Sigma = \Sigma\gamma$ for all $\gamma \in \Gamma$. If Σ is a normal subgroup of Γ, then the quotient group $\Gamma/\Sigma = \{\gamma\Sigma\colon \gamma \in \Gamma\}$ is also a Lie group [BtD, p. 35]. The *centre* of Γ, defined by

$$Z(\Gamma) = \{\delta \in \Gamma\colon \delta\gamma = \gamma\delta \ \forall \gamma \in \Gamma\},$$

is a normal subgroup of Γ. Further, a *homomorphism* of Lie groups is a smooth (infinitely differentiable) group homomorphism and an *isomorphism* (denoted by $\cong$) is an invertible homomorphism. The kernel of a homomorphism $\beta : \Gamma \to \tilde{\Gamma}$ is the set of elements of Γ that are mapped onto the identity element of $\tilde{\Gamma}$ and is a normal subgroup of Γ. The following result plays an important role in our derivation of scaling laws in § 4.

LEMMA 2.1. *Let Γ and $\tilde{\Gamma}$ be compact Lie groups and let $\beta : \Gamma \to \tilde{\Gamma}$ be a homomorphism with kernel K. Then*

$$\beta(\Gamma) \cong \Gamma/K. \tag{2.1}$$

Proof. There is a group isomorphism between $\beta(\Gamma)$ and Γ/K (Fraleigh (1977, p. 114)) and so it remains to prove that it is smooth. Now K is a (closed) normal subgroup of Γ and so Γ/K is a Lie group. Also $\beta(\Gamma)$ is a closed subgroup of $\tilde{\Gamma}$ since

β is continuous, and is thus a Lie group. Finally, a bijective homomorphism of Lie groups is an isomorphism [BtD, p. 22] which proves the result. □

Every compact Lie group Γ has a unique normalised Haar measure $d\gamma$ [BtD, p. 46] which has the invariance property that for any continuous function $f:\Gamma\to\mathbf{R}$ and any $\delta\in\Gamma$,

$$\int_\Gamma f(\gamma)\,d\gamma=\int_\Gamma f(\delta\gamma)\,d\gamma=\int_\Gamma f(\gamma\delta)\,d\gamma=\int_\Gamma f(\gamma^{-1})\,d\gamma$$

and is normalised such that

$$\int_\Gamma d\gamma=1.$$

If X is a real Hilbert space with inner product $(,)$, then integration can be extended to continuous functions $h:\Gamma\to X$ by defining $\int_\Gamma h(\gamma)\,d\gamma\in X$ to be the unique Riesz representor of the continuous linear functional $x\to\int_\Gamma (h(\gamma),x)\,d\gamma$ [BtD, p. 48]. Thus

$$\left(\int_\Gamma h(\gamma)\,d\gamma,x\right)=\int_\Gamma (h(\gamma),x)\,d\gamma. \tag{2.2}$$

Let X be a real Hilbert space with inner product $(,)$ and let the space of linear homeomorphisms from X to itself be denoted by $GL(X)$. If Γ is a compact Lie group, a *representation* of Γ on X is a group homomorphism $T:\Gamma\to GL(X)$ such that the mapping $(\gamma,x)\to T(\gamma)x$ of $\Gamma\times X$ onto X is continuous. The dimension of a representation is defined to be the dimension of the space X. An *action* of Γ on X is a continuous mapping

$$\rho:\Gamma\times X\to X,\qquad (\gamma,x)\to\rho(\gamma,x)\equiv\gamma x$$

such that

$$\mathbf{1}x=x,\qquad (\gamma_1\gamma_2)x=\gamma_1(\gamma_2 x)$$

for all $x\in X$ and $\gamma_1,\gamma_2\in\Gamma$, where $\mathbf{1}$ is the group identity element. For any action ρ of Γ on X, we can define a representation T of Γ on X by

$$T(\gamma)x\equiv\gamma x.$$

Then T is called the representation of Γ on X induced by the action ρ.

A representation T is called *orthogonal* if $T(\gamma)$ is orthogonal for all $\gamma\in\Gamma$. Important results concerning orthogonal representations of compact groups are given in the following two lemmas (cf. [BR, pp. 166, 140]).

LEMMA 2.2. *Let T be an arbitrary representation of a compact group Γ on X. Then the inner product $\langle,\rangle$ on X defined by*

$$\langle x,y\rangle\equiv\int_\Gamma (T(\gamma)x,T(\gamma)y)\,d\gamma \tag{2.3}$$

for all $x,y\in X$, defines a norm equivalent to the original one, relative to which T is an orthogonal representation of Γ.

The inner product defined by (2.3) is called Γ-*invariant* since

$$\langle x,y\rangle=\langle T(\gamma)x,T(\gamma)y\rangle$$

for all $x,y\in X$, $\gamma\in\Gamma$. *We will henceforth assume that X is a real, separable Hilbert space with a Γ-invariant inner product denoted by $\langle,\rangle$, so that T is an orthogonal representation of Γ on X.*

We say that representations T and $\tilde{T}$ of Γ on real Hilbert spaces X and $\tilde{X}$, respectively, are *equivalent* if there exists a linear homeomorphism $A: X \to \tilde{X}$ such that

$$AT(\gamma) = \tilde{T}(\gamma)A \quad \forall \gamma \in \Gamma.$$

If A is also orthogonal, we say that T and $\tilde{T}$ are *orthogonally equivalent.* We then have the following result.

LEMMA 2.3. *Two equivalent orthogonal representations are orthogonally equivalent.*

Two structures that will be used extensively in later sections are fixed-point subspaces and isotropy subgroups. For any subgroup Σ of Γ, the *fixed-point subspace* X^{Σ} is defined by

$$X^{\Sigma} = \{x \in X: T(\sigma)x = x \ \forall \sigma \in \Sigma\},$$

which is a closed subspace of X since it is the intersection of the null spaces of the bounded linear operators $T(\sigma) - I$ for all $\sigma \in \Sigma$. Also, for any $x \in X$,

$$\Sigma_x = \{\gamma \in \Gamma: T(\gamma)x = x\}$$

is a subgroup of Γ called the *isotropy subgroup* of x. (It is closed in Γ because T is a representation and so it is also a Lie group.) We extend the notion of isotropy subgroups of elements to subspaces, and so we define the isotropy subgroup Σ_Y of a subspace Y of X by

$$\Sigma_Y = \{\gamma \in \Gamma: T(\gamma)y = y \ \forall y \in Y\}.$$

A subspace W of X is Γ-*invariant* if $T(\gamma)w \in W \ \forall w \in W$, $\gamma \in \Gamma$. A nontrivial, closed, Γ-invariant subspace W of X is Γ-*irreducible* if it has no proper, closed, Γ-invariant subspaces. Otherwise it is Γ-*reducible.* (If there is no ambiguity with regard to the group Γ, we will refer to a subspace simply as invariant, irreducible, etc.)

We now aim to show that the Hilbert space X can be decomposed as an orthogonal direct sum of finite-dimensional irreducible subspaces, that is,

$$X = \sum_i \oplus X_i$$

where each X_i is irreducible, X_i and X_j are orthogonal for $i \neq j$, and every $x \in X$ can be decomposed into the convergent series

$$x = \sum_i x_i, \qquad x_i \in X_i.$$

The first result we require is the following (cf. [BR, pp. 141–142]).

LEMMA 2.4. *Let W be a closed, Γ-invariant subspace of X and let $W^{\perp}$ be the orthogonal complement of W such that $X = W \oplus W^{\perp}$. Then*

(i) *$W^{\perp}$ is a closed, Γ-invariant subspace of X.*

(ii) *The restriction of $T(\gamma)$ to W is also a representation of Γ.*

COROLLARY 2.5. *Every finite-dimensional invariant subspace of X can be decomposed as an orthogonal direct sum of irreducible subspaces.*

The next theorem shows that the infinite-dimensional space X can be decomposed in a similar way (cf. [BR, pp. 169–170]).

THEOREM 2.6. *Let Γ be a compact Lie group and let T be an orthogonal representation of Γ on the real Hilbert space X. Then X can be decomposed as an orthogonal direct sum of finite-dimensional irreducible subspaces.*

COROLLARY 2.7. *Every irreducible subspace of X is finite dimensional.*

A representation on an irreducible subspace is itself called irreducible. We now take a closer look at those linear operators that commute with irreducible representations, since this is where the group representation theory on real and complex spaces

starts to differ. Let W be an irreducible subspace of X on which the (irreducible) representation of Γ is τ. The *endomorphism algebra of* τ (over $\mathbf{R}$) is

$$D^{\Gamma}(\tau) = \{A: W \to W: A\tau(\gamma) = \tau(\gamma)A \ \forall \gamma \in \Gamma \text{ and } A \text{ is linear}\}.$$

A fundamental result in the theory of group representations is Schur's lemma and its corollary, which can be found in Kirillov (1976).

THEOREM 2.8 (Schur's lemma). *Let τ_1 and τ_2 be irreducible representations of the compact Lie group Γ on subspaces W_1 and W_2 of X, respectively. Let $A: W_1 \to W_2$ be a linear mapping such that*

$$A\tau_1(\gamma) = \tau_2(\gamma)A$$

for all $\gamma \in \Gamma$. Then A is either zero, or it is invertible, in which case τ_1 and τ_2 are equivalent representations.

COROLLARY 2.9. *Let τ be an irreducible representation of Γ. Then $D^{\Gamma}(\tau)$ is isomorphic to either $\mathbf{R}$, $\mathbf{C}$, or $\mathbf{H}$ (the quaternions) which are one-, two-, or four-dimensional algebras over $\mathbf{R}$, respectively, and τ (and W) are said to be of real, complex, or quaternionic type, respectively.*

If τ is an irreducible representation of real type on the subspace W, then the only linear mappings that commute with τ are real multiples of the identity, and in this case τ (and W) are also called *absolutely irreducible.* (Note that for any irreducible representation on a complex space, the only commuting linear mappings are complex multiples of the identity.)

The following result plays an important role in the proof of Theorem 2.11.

THEOREM 2.10. *Let W be a Γ-irreducible subspace of X with corresponding irreducible representation τ. If $A \in D^{\Gamma}(\tau)$ is self-adjoint, then it is a (real) multiple of the identity on W.*

Proof. As A is self-adjoint, all its eigenvalues are real. Now consider the three classes of irreducible representations:

(i) If τ is of real type then $A = aI$, $a \in \mathbf{R}$ and so the result holds in this case.

(ii) If τ is of complex type, then $D^{\Gamma}(\tau) \cong \mathbf{C}$ and so

$$A = aI + bM, \qquad a, b \in \mathbf{R}$$

where $M^2 = -I$. Thus the only eigenvalues of A are $a \pm ib$, which are real if and only if $b = 0$. Then $A = aI$ as required.

(iii) If τ is of quaternionic type, then $D^{\Gamma}(\tau) \cong \mathbf{H}$ and so

$$A = aI + bM_i + cM_j + dM_k$$

where $M_i^2 = M_j^2 = M_k^2 = -I$, $M_iM_j = -M_jM_i = M_k$, $M_jM_k = -M_kM_j = M_i$, and $M_kM_i = -M_iM_k = M_j$. Using these relations, it is easily shown that the only eigenvalues of A are $a \pm i(b^2 + c^2 + d^2)^{1/2}$, which are real if and only if $b = c = d = 0$, giving $A = aI$. □

The main result of this section is the decomposition of the Hilbert space X into orthogonal isotypic components. This decomposition and the associated projections are well known for complex Hilbert spaces (see Knapp (1986)).

THEOREM 2.11. *Let T be an orthogonal representation of Γ on X and let $X = \sum_i \oplus X_i$, where each X_i is irreducible. Let $\{X_{i_k}\}$ be a maximal set of mutually inequivalent irreducible subspaces. For each k, let V_k^{Γ} be the closure of the sum of all the irreducible subspaces X_i which are equivalent to X_{i_k}. Then*

(i) *$V_k^{\Gamma} \perp V_l^{\Gamma}$, $k \neq l$ and $X = \sum_k \oplus V_k^{\Gamma}$.*

(ii) *The linear operator P_k^{Γ} defined by*

$$P_k^{\Gamma} = \frac{n_k}{d_k} \int_{\Gamma} \chi_k(\gamma) T(\gamma)\, d\gamma \tag{2.4}$$

is an orthogonal projection operator whose range is V_k^Γ, *where* n_k *is the dimension of* X_{i_k}, d_k *is the dimension of the endomorphism algebra associated with* X_{i_k} (*i.e.*, 1, 2, *or* 4), *and* $\chi_k : \Gamma \to \mathbf{R}$ *is the character of the irreducible representation* τ_{i_k} *associated with* X_{i_k}, *defined by*

$$\chi_k(\gamma) = \operatorname{Tr}(\tau_k(\gamma))$$

where Tr *denotes the trace.*

Before giving a proof of this theorem, we make some remarks. The subspaces $V_k^\Gamma = P_k^\Gamma X$ are called the Γ-*isotypic components* of X. If Γ is a finite group, the Haar integral is replaced by a finite sum and so the projections onto the isotypic components become

$$P_k^\Gamma = \frac{n_k}{d_k|\Gamma|} \sum_{\gamma\in\Gamma} \chi_k(\gamma) T(\gamma)$$

where $|\Gamma|$ is the order of the group Γ. Also, if T is a unitary representation on a complex Hilbert space then the projections onto the isotypic components are

$$P_k^\Gamma = n_k \int_\Gamma \overline{\chi_k(\gamma)} T(\gamma)\, d\gamma.$$

On an equivalence class of irreducible representations, the character, dimension, and type are all constant, and so the isotypic decomposition does not depend on the choice of the maximal set of nonequivalent irreducible subspaces. This ensures that the isotypic decomposition is unique, in contrast to the decomposition of X into irreducible subspaces (Theorem 2.6) which is not unique.

Theorem 2.6, Corollary 2.7, and Theorem 2.11 are analogous to part of the Peter–Weyl theorem for unitary representations on a complex Hilbert space (see Knapp (1986)).

Proof of Theorem 2.11. In order to prove the theorem, we need some basic results on group characters of real representations (see [BtD, pp. 80, 101]). If χ_k is the character of the irreducible representation τ_k, then for all $\gamma, \delta \in \Gamma$

(i) $\chi_k(\gamma^{-1}) = \chi_k(\gamma)$,

(ii) $\chi_k(\delta\gamma\delta^{-1}) = \chi_k(\gamma)$,

(iii) $$\int_\Gamma \chi_k(\gamma)\chi_j(\gamma)\, d\gamma = \begin{cases} 0, & \tau_j, \tau_k \text{ not equivalent,} \\ d_k, & \tau_j, \tau_k \text{ equivalent.} \end{cases}$$

For the sake of convenience, we will omit the superscript Γ on the operators P_k^Γ and the isotypic components V_k^Γ throughout the proof. The first step is to show that P_k commutes with the representation T. Let $\delta \in \Gamma$. Then

$$\begin{aligned} T(\delta^{-1})P_k T(\delta) &= \frac{n_k}{d_k} \int_\Gamma \chi_k(\gamma) T(\delta)^{-1} T(\gamma) T(\delta)\, d\gamma \\ &= \frac{n_k}{d_k} \int_\Gamma \chi_k(\gamma) T(\delta^{-1}\gamma\delta)\, d\gamma \\ &= \frac{n_k}{d_k} \int_\Gamma \chi_k(\delta\gamma\delta^{-1}) T(\gamma)\, d\gamma \\ &= \frac{n_k}{d_k} \int_\Gamma \chi_k(\gamma) T(\gamma)\, d\gamma \\ &= P_k \end{aligned}$$

using the homomorphism property of T, the invariance of the Haar measure, and result (ii) on group characters, which proves that

$$P_k T(\delta) = T(\delta) P_k.$$

In order to prove that P_k is an orthogonal projection, it is sufficient to show that it is a bounded, self-adjoint projection on X. Now P_k is bounded since Γ is compact, χ_k and T are continuous, and n_k is finite. We now prove that it is self-adjoint. If $x, y \in X$ then

$$\begin{aligned}
\langle P_k x, y\rangle &= \left\langle \frac{n_k}{d_k} \int_\Gamma \chi_k(\gamma) T(\gamma) x \, d\gamma, y \right\rangle \\
&= \left\langle x, \frac{n_k}{d_k} \int_\Gamma \chi_k(\gamma) T(\gamma)^* y \, d\gamma \right\rangle \\
&= \left\langle x, \frac{n_k}{d_k} \int_\Gamma \chi_k(\gamma) T(\gamma^{-1}) y \, d\gamma \right\rangle \\
&= \left\langle x, \frac{n_k}{d_k} \int_\Gamma \chi_k(\gamma^{-1}) T(\gamma) y \, d\gamma \right\rangle \\
&= \left\langle x, \frac{n_k}{d_k} \int_\Gamma \chi_k(\gamma) T(\gamma) y \, d\gamma \right\rangle \\
&= \langle x, P_k y\rangle
\end{aligned}$$

using (2.2), the fact that T is an orthogonal representation, the homomorphism property of T, the invariance of the Haar measure, and result (i) on group characters. Thus P_k is self-adjoint.

Now consider the restriction of P_k to an irreducible subspace W with corresponding (irreducible) representation τ_j. As $\tau_j(\gamma)$: $W \to W$, then $P_k : W \to W$ also. Since P_k commutes with T, we conclude that $P_k|_W \in D^\Gamma(\tau_j)$. As P_k is also self-adjoint, it follows from Theorem 2.10 that $P_k|_W = cI|_W$, $c \in \mathbf{R}$. Taking the trace (on W) gives

$$cn_j = \frac{n_k}{d_k} \int_\Gamma \chi_k(\gamma) \chi_j(\gamma) \, d\gamma$$

where n_j and χ_j are the dimension and character, respectively, of τ_j. If τ_j is equivalent to τ_k, then $n_j = n_k$ and so $c = 1$ using result (iii) on group characters. Similarly, if τ_j is not equivalent to τ_k, then $c = 0$. Thus P_k acts as the identity on all irreducible subspaces whose representation is equivalent to τ_k, and as zero on every other irreducible subspace so that $P_k^2 = P_k$ and $P_j P_k = 0$, $j \neq k$. Hence P_k is an orthogonal projection and the isotypic components are mutually orthogonal.

Now let $M = \sum_k \oplus V_k$ and suppose that $M \neq X$. Then M is a closed, invariant subspace and so $M^\perp$ is also a closed, invariant subspace by Lemma 2.4(i). Thus by Theorem 2.6 we conclude that $M^\perp$ can be decomposed as an orthogonal sum of irreducible subspaces, which contradicts the definition of the isotypic components V_k. Hence $M^\perp = 0$ and so $M = X$. □

3. Symmetry-breaking bifurcation theory. In this section, we apply the group theoretic results of § 2 to symmetry-breaking bifurcation theory, which we approach from a computational viewpoint. Our analysis will therefore deal with a nonlinear equation in an infinite-dimensional real Hilbert space, without first reducing the problem to finite dimensions using the Lyapunov–Schmidt procedure.

Consider the bifurcation problem

$$g(x, \lambda) = 0 \tag{3.1}$$

where $g: X \times \mathbf{R} \to X$ is a C^2-mapping and X is a real, separable Hilbert space. We suppose that g satisfies the equivariance condition

$$T(\gamma) g(x, \lambda) = g(T(\gamma)x, \lambda), \qquad \gamma \in \Gamma \tag{3.2}$$

where T is a representation of the compact Lie group Γ on X. We assume, without loss of generality (Lemma 2.2), that X has a Γ-invariant inner product so that the representation T is an orthogonal representation of Γ. It follows immediately from (3.2) that if (x, λ) is a solution of (3.1), then $(T(\gamma)x, \lambda)$ is also a solution for all $\gamma \in \Gamma$. These are called *conjugate solutions.*

Throughout this section, we assume the existence of a smooth curve of solutions of (3.1) contained in $X^\Gamma \times \mathbf{R}$, which we call the primary branch. We then consider the possibility of bifurcation from such a branch. Note that Γ can be *any* group that satisfies (3.2), not necessarily the largest such group.

The following important result uses the isotypic decomposition of Theorem 2.11.

THEOREM 3.1. *Let P_k^Γ be the projection onto the Γ-isotypic components of X. If $x \in X^\Gamma$, then $g_x(x, \lambda): V_k^\Gamma \to V_k^\Gamma$ for all k, where $V_k^\Gamma \equiv P_k^\Gamma X$.*

Proof. Taking the Fréchet derivative of (3.2) with respect to x gives

$$T(\gamma) g_x(x, \lambda)\varphi = g_x(T(\gamma)x, \lambda) T(\gamma)\varphi = g_x(x, \lambda) T(\gamma)\varphi \tag{3.3}$$

for all $\gamma \in \Gamma$, $x, \varphi \in X$, and $\lambda \in \mathbf{R}$ since $x \in X^\Gamma$. It follows from this that

$$P_k^\Gamma g_x(x, \lambda) = g_x(x, \lambda) P_k^\Gamma$$

using definition (2.4) of the projection operators P_k^Γ. The result follows. □

As a result of Theorem 3.1, if $x \in X^\Gamma$ then $g_x(x, \lambda)$ can be decomposed into "block diagonal" form as

$$g_x(x, \lambda) = \operatorname{diag}\,(g_x^k(x, \lambda))$$

where $g_x^k(x, \lambda) = g_x(x, \lambda)|_{V_k^\Gamma}$ and $g_x^k(x, \lambda): V_k^\Gamma \to V_k^\Gamma$. Thus $g_x(x, \lambda)$ has a nontrivial nullspace if and only if $g_x^k(x, \lambda)$ has a nontrivial nullspace for at least one k. This means that in general there are many different possible modes of bifurcation, although, as we will see, some of these do not arise generically.

This decomposition can be used in solving any linear equation where the linear operator $A: X \to X$ commutes with a representation of a compact Lie group Γ on the (real or complex) Hilbert space X and is the underlying principle of the paper by Bossavit (1986). In particular the linear equation $Ax = b$ can be decomposed into several subproblems as

$$A^k x_k = P_k^\Gamma b$$

where $A^k = A|_{V_k^\Gamma}$, $A^k: V_k^\Gamma \to V_k^\Gamma$ and $x_k \in V_k^\Gamma$. The solution to the original problem is then $x = \sum_k x_k$. Similarly, the eigenvalue problem $Ax = \lambda x$ can be decomposed as

$$A^k x_k = \lambda x_k.$$

When applied to boundary-value problems, this decomposition can result in a considerable savings computationally, as the subproblems can be solved on domains or "symmetry cells" smaller than the origin domain (see Bossavit (1986)). Since the subproblems are independent, computational efficiency can be increased by solving them in parallel. Saffman (1980) observed this decomposition in his study of bifurcation of gravity waves on deep water but gave no explanation of it.

We are interested in analysing singular points $(x_0, \lambda_0) \in X^\Gamma \times \mathbf{R}$ on the primary branch at which $\mathcal{N}_0 = \text{Null}\,(g_x(x_0, \lambda_0))$ is nontrivial. Now $\mathcal{N}_0$ is a closed subspace of X since $g_x(x, \lambda)$ is a bounded linear operator. Also, it follows from (3.3) that $\mathcal{N}_0$ is Γ-invariant since $x_0 \in X^\Gamma$. We now consider which isotypic components generically give rise to a nontrivial null space of $g_x(x, \lambda)$. For this result, we consider two classes of operators that map $X \times \mathbf{R} \to X$. We define C_1 to be the class of Γ-equivariant C^2-mappings and C_2 to be the class of operators $g \in C_1$ such that $g(x, \lambda) = \nabla G(x, \lambda)$ for some function $G: X \times \mathbf{R} \to \mathbf{R}$. Note that if $g \in C_2$, then $g_x(x, \lambda)$ is self-adjoint for all $(x, \lambda) \in X \times \mathbf{R}$.

THEOREM 3.2. *Suppose that $g_x(x, \lambda)$ is a Fredholm operator of index zero and let $g_x^k(x, \lambda) \equiv g_x(x, \lambda)|_{V_k^\Gamma}$, where V_k^Γ is an isotypic component of X associated with the irreducible representation τ_k of Γ. Also, let $\mathcal{GN}_0^k$ be the generalised nullspace of $g_x^k(x_0, \lambda_0)$ for some $(x_0, \lambda_0) \in X^\Gamma \times \mathbf{R}$ on the primary branch.*

(a) *Suppose that $g \in C_1$. Then*

(i) *If τ_k is absolutely irreducible and $\mathcal{GN}_0^k$ is nontrivial, then generically $\mathcal{GN}_0^k$ is irreducible (and hence absolutely irreducible).*

(ii) *If τ_k is non-absolutely irreducible, then generically $\mathcal{GN}_0^k$ is trivial.*

(b) *Suppose that $g \in C_2$. If $\mathcal{GN}_0^k$ is nontrivial, then generically $\mathcal{GN}_0^k$ is irreducible regardless of the type of τ_k.*

Proof. We base our proof on the ideas contained in the sketch proof of Golubitsky, Stewart, and Schaeffer (1988, p. 84).

(a) We first prove that if $\mathcal{GN}_0^k$ is not irreducible then a small perturbation makes it irreducible. This is done in two steps. The first step consists of introducing a perturbation that makes the generalised nullspace and the nullspace coincide. The second step then reduces the generalised nullspace to an irreducible subspace by an appropriate perturbation.

Let $g \in C_1$. As we are considering the linear operator $g_x^k(x, \lambda)$, we restrict attention to the space V_k^Γ. Since $g_x(x, \lambda)$ is a Fredholm operator of index zero, $\mathcal{GN}_0^k$ is finite dimensional and V_k^Γ can be decomposed as (Dancer (1971))

$$V_k^\Gamma = \mathcal{GN}_0^k \oplus Y.$$

Also, since $x_0 \in X^\Gamma$, $\mathcal{GN}_0^k$ is Γ-invariant and so, by Lemma 2.4, Y is also Γ-invariant.

Define $\mathcal{N}_0^k \equiv \text{Null}\,(g_x^k(x_0, \lambda_0))$ and suppose that $\mathcal{N}_0^k \neq \{0\}$ and $\mathcal{N}_0^k \neq \mathcal{GN}_0^k$. Then $\mathcal{GN}_0^k$ can be decomposed as

$$\mathcal{GN}_0^k = \mathcal{N}_0^k \oplus M. \tag{3.4}$$

For the first step, we define the linear operator $L_1: X \to X$ by

$$L_1|_{(V_k^\Gamma)^\perp} = 0, \quad L_1|_Y = 0, \quad L_1|_{\mathcal{N}_0^k} = 0, \quad L_1|_M = \text{Id}|_M$$

and we define $\tilde{g}(x, \lambda) \equiv g(x, \lambda) + \varepsilon_1 L_1 x \in C_1$. Then

$$\tilde{g}_x^k(x, \lambda) \equiv \tilde{g}_x(x, \lambda)|_{V_k^\Gamma} = g_x^k(x, \lambda) + \varepsilon_1 L_1.$$

It is then easily shown that

$$\text{Null}\,(\tilde{g}_x^k(x_0, \lambda_0)) = \mathcal{N}_0^k$$

and that the generalised nullspace of $\tilde{g}_x^k(x_0, \lambda_0)$ coincides with $\mathcal{N}_0^k$.

For the second step, we decompose $\mathcal{N}_0^k$ as

$$\mathcal{N}_0^k = V \oplus W$$

where V is Γ-irreducible and W is Γ-invariant. We then define another linear operator $L_2: X \to X$ by

$$L_2|_{(V_k^\Gamma)^\perp} = 0, \quad L_2|_Y = 0, \quad L_2|_M = 0,$$
$$L_2|_V = 0, \qquad L_2|_W = \mathrm{Id}|_W$$

and we define $\hat{g}(x, \lambda) \equiv \tilde{g}(x, \lambda) + \varepsilon_2 L_2 x \in C_1$. Then

$$\hat{g}_x^k(x, \lambda) \equiv \hat{g}_x(x, \lambda)|_{V_k^\Gamma} = \tilde{g}_x^k(x, \lambda) + \varepsilon_2 L_2.$$

In this case, the generalised nullspace of $\hat{g}_x^k(x_0, \lambda_0)$ coincides with the nullspace and is equal to V. Thus if the generalised nullspace of $\tilde{g}_x^k(x_0, \lambda_0)$ is not irreducible, then a small perturbation makes it irreducible.

We now consider whether an irreducible generalised nullspace is stable under perturbation. Suppose that $\mathcal{GN}_0^k = \mathcal{N}_0^k$ and is Γ-irreducible. Then a Lyapunov–Schmidt reduction can be used to give an equivalent finite-dimensional problem

$$\tilde{G}(X, \Lambda) = 0, \qquad \tilde{G}: V \times \mathbf{R} \to \tilde{V} \tag{3.5}$$

where $V \equiv \mathcal{N}_0^k$, $\tilde{G}$ is Γ-equivariant, and $\tilde{G}(0, 0) = 0$ (Golubitsky and Schaeffer (1985, Chap. VII)). By restricting attention to the isotypic component V_k^Γ, we conclude that the restrictions of the representation T of Γ to V and $\tilde{V}$ are equivalent and so there exists an invertible linear operator $A: \tilde{V} \to V$ which commutes with T. Thus, the modified equations

$$G(X, \Lambda) \equiv A\tilde{G}(X, \Lambda) = 0, \qquad G: V \times \mathbf{R} \to V \tag{3.6}$$

have the same solutions as (3.5) and G is also Γ-equivariant.

Now the representation of Γ on V is τ_k and $G_X(X, \Lambda)$ commutes with this representation. Thus $G_X(X, \Lambda) \in D^\Gamma(\tau_k)$, the endomorphism algebra of τ_k. By the proof of Theorem 2.10, if τ_k is of real type (i.e., absolutely irreducible), then $G_X(X, \Lambda)$ has only one (repeated) real eigenvalue $a(\Lambda)$ with

$$a(0) = 0. \tag{3.7}$$

If τ_k is of complex type, then $G_X(X, \Lambda)$ has only the complex conjugate eigenvalues $a(\Lambda) \pm ib(\Lambda)$ with

$$a(0) = b(0) = 0. \tag{3.8}$$

Finally, if τ_k is of quaternionic type, then $G_X(X, \Lambda)$ has only the complex conjugate eigenvalues $a(\Lambda) \pm i(b(\Lambda)^2 + c(\Lambda)^2 + d(\Lambda)^2)^{1/2}$ with

$$a(0) = b(0) = c(0) = d(0) = 0. \tag{3.9}$$

Clearly the solution (3.7) of one equation in one variable is stable under perturbation, whereas the solutions (3.8) and (3.9) of more than one equation in only one variable are not stable under perturbation. In these cases, perturbations exist that reduce the nullspace of $G_X(X, \Lambda)$ to being trivial in a neighbourhood of the origin.

Thus if τ_k is of real type, then $\mathcal{GN}_0^k$ is generically irreducible. If τ_k is not of real type, then a small perturbation destroys the generalised nullspace. Also, if $\mathcal{GN}_0^k$ is trivial, then for any sufficiently small perturbation, it will remain trivial and so in this case, generically $\mathcal{GN}_0^k$ is trivial.

(b) If $g \in C_2$, then $\mathcal{GN}_0^k = \mathcal{N}_0^k$ and so, in the decomposition (3.4), $M = \{0\}$. The perturbed problem

$$\tilde{g}(x, \lambda) \equiv g(x, \lambda) + \varepsilon L_2 x \in C_2$$

then has an irreducible nullspace, following the proof of (a).

Performing a Lyapunov-Schmidt reduction in this case leads to the finite-dimensional problem

$$G(X, \Lambda) = 0, \qquad G: V \times \mathbf{R} \to V$$

where $V \equiv \mathcal{N}_0^k$, G is Γ-equivariant, $G(0, 0) = 0$, and $G_X(X, \Lambda)$ is self-adjoint. Since $G_X(X, \Lambda) \in D^\Gamma(\tau_k)$ also, by Theorem 2.10 the eigenvalues of $G_X(X, \Lambda)$ are of the form $a(\Lambda)$ irrespective of the type of τ_k. In this case, the solution

$$a(0) = 0$$

is stable under perturbation and so generically $\mathcal{GN}_0^k$ is irreducible irrespective of the type of τ_k. □

We note that if the generalised nullspace of $g_x(x_0, \lambda_0)$ is irreducible, then it coincides with the nullspace and is then the "smallest" possible invariant subspace. Also, Hopf bifurcation is associated with complex type irreducible representations (see Golubitsky, Stewart, and Schaeffer (1988, Chap. XVI)).

Let $(x_0, \lambda_0) \in X^\Gamma \times \mathbf{R}$ be a point on the primary branch such that $g_x^k(x_0, \lambda_0)$ has a nontrivial nullspace contained in V_k^Γ for some k. If $\psi \in \text{Null}\,(g_x^k(x_0, \lambda_0)^*)$, where $*$ denotes the adjoint operator, then $\psi \in V_k^\Gamma$, since X is a Hilbert space and $g_x^k(x_0, \lambda_0): V_k^\Gamma \to V_k^\Gamma$, and so ψ has the property that $\langle \psi, x \rangle = 0$ for all $x \in \sum_{j \neq k} \oplus V_j^\Gamma$ due to the orthogonality of the isotypic components. Also, $g_\lambda(x_0, \lambda_0) \in X^\Gamma$ since $x_0 \in X^\Gamma$.

The isotypic component V_1^Γ say, corresponding to the trivial (irreducible) representation, is the fixed-point space X^Γ whose irreducible subspaces are one-dimensional and absolutely irreducible. If the nullspace of $g_x^1(x_0, \lambda_0)$ is nontrivial, then, from Theorem 3.2, generically it will be one-dimensional. Also, if $\psi \in \text{Null}\,(g_x^1(x_0, \lambda_0)^*)$, then $\psi \in V_1^\Gamma = X^\Gamma$. Thus, if $g_\lambda(x_0, \lambda_0) \neq 0$, then generically $\langle \psi, g_\lambda(x_0, \lambda_0) \rangle \neq 0$, in which case $g_\lambda(x_0, \lambda_0) \notin \text{Range}\,(g_x(x_0, \lambda_0))$ and so a singularity of $g_x^1(x_0, \lambda_0)$ corresponds generically to a limit point (Moore and Spence (1980)). It is easy to prove from (3.2) that if $x_0 \in X^\Gamma$ then $g_{xx}(x_0, \lambda_0): X^\Gamma \times X^\Gamma \to X^\Gamma$, and so generically $\langle \psi, g_{xx}(x_0, \lambda_0)\varphi\varphi \rangle \neq 0$, where $\varphi \in \text{Null}\,(g_x^1(x_0, \lambda_0))$. Thus generically the limit point is simple.

However, if the nullspace of $g_x^k(x_0, \lambda_0)$ is nontrivial for $k \neq 1$ and $\psi \in \text{Null}\,(g_x(x_0, \lambda_0)^*)$, then $\psi \in V_k^\Gamma$ and so $\langle \psi, g_\lambda(x_0, \lambda_0) \rangle = 0$ since ψ and $g_\lambda(x_0, \lambda_0)$ are in different (orthogonal) isotypic components. Thus $g_\lambda(x_0, \lambda_0) \in \text{Range}\,(g_x(x_0, \lambda_0))$ and so, for $k \neq 1$, we would expect a singularity of $g_x^k(x_0, \lambda_0)$ to give rise to a bifurcation point. Conditions for the existence of a bifurcating branch of solutions at such a singular point are given in Theorem 3.4. Therefore, we conclude that, generically, symmetry must be broken for bifurcation from a nontrivial solution branch to occur.

In some cases the linear operators $g_x^k(x, \lambda)$ can themselves be further decomposed as the following result shows.

THEOREM 3.3. *Let τ_k be an irreducible representation of Γ of dimension $n_k > 1$. If Σ is a subgroup of Γ with the property that $\tau_k|_\Sigma$ is not Σ-irreducible and has at least two nonequivalent Σ-irreducible components, then V_k^Γ can be decomposed as*

$$V_k^\Gamma = \sum_{j=1}^{m} \oplus V_{k,j}^{\Gamma,\Sigma}$$

where $V_{k,j}^{\Gamma,\Sigma} = V_k^\Gamma \cap V_j^\Sigma \neq \{0\}$, V_j^Σ *are the* Σ*-isotypic components of* X *that intersect* V_k^Γ *nontrivially and* m *is the number of nonequivalent* Σ*-irreducible components of* τ_k. *If* $x \in X^\Gamma$, *then* $g_x^k(x, \lambda)$: $V_{k,j}^{\Gamma,\Sigma} \to V_{k,j}^{\Gamma,\Sigma}$ *and so* $g_x^k(x, \lambda)$ *can itself be decomposed as* $g_x^k(x, \lambda) = \operatorname{diag}(g_x^{k,j}(x, \lambda))$, *where* $g_x^{k,j}(x, \lambda)$: $V_{k,j}^{\Gamma,\Sigma} \to V_{k,j}^{\Gamma,\Sigma}$.

Proof. First, note that $\Sigma \neq \Gamma$ or else τ_k would not be Γ-irreducible, and Σ cannot be the trivial group since all the irreducible components of the trivial group are equivalent, so Σ must be a proper subgroup of Γ. The number of Σ-irreducible components of τ_k cannot exceed its dimension, and so $m \leqq n_k$ and is therefore finite as τ_k is finite dimensional by Corollary 2.7.

From the hypotheses of the theorem we conclude that each Γ-irreducible subspace of the isotypic component V_k^Γ can be decomposed into m Σ-irreducible subspaces, which we can assume to be an orthogonal direct sum by Theorem 2.6. Collecting together the equivalent, Σ-irreducible subspaces in V_k^Γ, we obtain a Σ-isotypic decomposition of V_k^Γ resulting in the stated decomposition of V_k^Γ.

Now if $x \in X^\Gamma$, then $x \in X^\Sigma$ also as Σ is a subgroup of Γ and so, by Theorem 3.1, $g_x(x, \lambda)$: $V_j^\Sigma \to V_j^\Sigma$ for all j. Thus $g_x(x, \lambda)$: $V_{k,j}^{\Gamma,\Sigma} \to V_{k,j}^{\Gamma,\Sigma}$ and so $g_x^k(x, \lambda)$ can be decomposed as stated. □

Let $(x_0, \lambda_0) \in X^\Gamma \times \mathbf{R}$ be a point on the primary branch at which $\mathcal{N}_0 \equiv \operatorname{Null}(g_x(x_0, \lambda_0))$ is nontrivial and irreducible with corresponding irreducible representation τ_k. Then $\mathcal{N}_0 \subset V_k^\Gamma$. If Σ is a subgroup of Γ that satisfies the hypotheses of Theorem 3.3, then $\mathcal{N}_0$ can be decomposed into Σ-isotypic components $\mathcal{N}_0^j = \mathcal{N}_0 \cap V_j^\Sigma$, $j = 1, \cdots, m$, which are not Γ-invariant. Clearly $\operatorname{Null}(g_x^{k,j}(x_0, \lambda_0)) = \mathcal{N}_0^j$ and so the subspaces $\mathcal{N}_0^j$, $j = 1, \cdots, m$ are *all* nontrivial. Thus in order to detect a singular point, it is sufficient to consider $g_x^{k,j}(x, \lambda)$ for only one j. In particular, if there is a Σ-isotypic component of $\mathcal{N}_0$ of odd dimension with $j = 1$, say, then the sign of the determinant of $g_x^{k,1}(x, \lambda)$ can be checked numerically while following the primary branch of solutions in order to detect a singular point, and the other $g_x^{k,j}(x, \lambda)$, $j \neq 1$ need not be considered. A special case of this occurs when the Σ-irreducible decomposition of τ_k includes the trivial representation with multiplicity 1, in which case $\dim(\mathcal{N}_0 \cap X^\Sigma) = 1$ since X^Σ is the Σ-isotypic component of X corresponding to the trivial irreducible representation. This important special case is featured in the Equivariant Branching Lemma of Cicogna (1981) and Vanderbauwhede (1982).

THEOREM 3.4 (Equivariant Branching Lemma). *Let* $x_0 \in X^\Gamma$. *Suppose that* $g_x(x_0, \lambda_0)$ *is Fredholm of index zero and that* $\mathcal{N}_0 \equiv \operatorname{Null}(g_x(x_0, \lambda_0))$ *is nontrivial. If*

(i) $\mathcal{N}_0 \cap X^\Gamma = \{0\}$;

(ii) *There exists an isotropy subgroup* Σ *of* Γ *such that*

$$\dim(\mathcal{N}_0 \cap X^\Sigma) = 1;$$

(iii) *The nondegeneracy condition*

$$\langle \psi_0, g_{x\lambda}(x_0, \lambda_0)\varphi_0 + g_{xx}(x_0, \lambda_0)\varphi_0 v \rangle \neq 0 \tag{3.10}$$

is satisfied where $\varphi_0 \in \mathcal{N}_0 \cap X^\Sigma$, $\psi_0 \in \operatorname{Null}(g_x(x_0, \lambda_0)^*) \cap X^\Sigma$, *and* $v \in X^\Gamma$ *is the unique solution of*

$$g_x(x_0, \lambda_0)v + g_\lambda(x_0, \lambda_0) = 0,$$

then there exists a secondary branch of solutions tangent to φ_0 *with isotropy subgroup* Σ.

The Equivariant Branching Lemma gives sufficient conditions for the existence of a bifurcating branch of solutions that apply to many practical situations, but bifurcation can also occur when the secondary branches are contained in X^Σ with $\dim(\mathcal{N}_0 \cap X^\Sigma) > 1$ (see Lauterbach (1986)).

It can be proved in a way similar to the proof of Theorem 3.1 that if $x_0 \in X^\Gamma$, then $g_{x\lambda}(x_0, \lambda_0)$: $V_k^\Gamma \to V_k^\Gamma$ and $g_{xx}(x_0, \lambda_0)$: $V_k^\Gamma \times X^\Gamma \to V_k^\Gamma$. Also, for any subgroup Ω of Γ, $g_{x\lambda}(x_0, \lambda_0)$: $X^\Omega \to X^\Omega$ and $g_{xx}(x_0, \lambda_0)$: $X^\Omega \times X^\Gamma \to X^\Omega$. Thus if $\varphi_0 \in V_k^\Gamma \cap X^\Sigma$, then

$$g_{x\lambda}(x_0, \lambda_0)\varphi_0 + g_{xx}(x_0, \lambda_0)\varphi_0 v \in V_k^\Gamma \cap X^\Sigma$$

also and so the nondegeneracy condition (3.10) is satisfied generically.

Our next consideration is the type of bifurcation, i.e., pitchfork or transcritical, which occurs as a result of the Equivariant Branching Lemma. Since the secondary branches are contained in $X^\Sigma \times \mathbf{R}$, it is sufficient to consider the reduced problem

$$g_\Sigma(x, \lambda) = 0, \qquad g_\Sigma \equiv g|_{X^\Sigma \times \mathbf{R}} : X^\Sigma \times \mathbf{R} \to X^\Sigma. \tag{3.11}$$

Now the closed fixed-point space X^Σ is invariant with respect to the normaliser N_Γ of Σ in Γ, defined by

$$N_\Gamma(\Sigma) = \{\gamma \in \Gamma : \gamma\Sigma = \Sigma\gamma\}.$$

Thus the reduced problem (3.11) is equivariant with respect to the quotient group $N_\Gamma(\Sigma)/\Sigma$. The following result has been proved by Dellnitz and Werner (1989).

THEOREM 3.5. *Let* $\mathcal{N}_0 = \text{Null}\,(g_x(x_0, \lambda_0))$, *where* $x_0 \in X^\Gamma$. *Suppose there exists an isotropy subgroup* Σ *of* Γ *such that* $\dim(\mathcal{N}_0 \cap X^\Sigma) = 1$.

(i) *If* $\dim \mathcal{N}_0 = 1$ *then* Σ *is the isotropy subgroup of* $\mathcal{N}_0$, $N_\Gamma(\Sigma) = \Gamma$, *and* $N_\Gamma(\Sigma)/\Sigma = \Gamma/\Sigma$ *is isomorphic to* Z_2.

(ii) *If* $\dim \mathcal{N}_0 > 1$ *then* $N_\Gamma(\Sigma)/\Sigma$ *is either isomorphic to* Z_2 *or is trivial.*

Now if $N_\Gamma(\Sigma)/\Sigma$ is isomorphic to Z_2, then the bifurcation is a symmetric pitchfork (see Werner and Spence (1984)) with conjugate secondary branches where, if (x, λ) is a solution of (3.11) on a secondary branch, then $(T(\delta)x, \lambda)$ is the corresponding solution on the conjugate branch for all $\delta \in N(\Sigma)\backslash\Sigma$. Alternatively, if $N_\Gamma(\Sigma)/\Sigma$ is trivial, then the bifurcation will be nonsymmetric (either pitchfork or transcritical) and the two secondary branches have distinct (nonconjugate) solutions.

We now consider extended systems of equations for which a bifurcation point is an isolated solution. The following result is a generalisation of Theorem 3.1 of Werner and Spence (1984), which considers the case $\Gamma = Z_2$. It is proved in a similar way and is straightforward.

THEOREM 3.6. *Let* $(x_0, \lambda_0) \in X^\Gamma \times \mathbf{R}$ *and let* $\mathcal{N}_0 = \text{Null}\,(g_x(x_0, \lambda_0))$ *have finite dimension. If* $\mathcal{N}_0 \cap V_k^\Gamma \neq \{0\}$ *for some* $k \neq 1$ *and conditions* (i) *and* (ii) *of the Equivariant Branching Lemma* (*Theorem* 3.4) *hold, then the extended system*

$$\begin{aligned} &G(y) = 0, \qquad G : Y \to Y, \\ &G(y) \equiv \begin{pmatrix} g(x, \lambda) \\ g_x(x, \lambda)\varphi \\ \langle l, \varphi \rangle - 1 \end{pmatrix}, \\ &y = (x, \varphi, \lambda) \in Y \equiv X^\Gamma \times (V_k^\Gamma \cap X^\Sigma) \times \mathbf{R}, \qquad l \in V_k^\Gamma \cap X^\Sigma \end{aligned} \tag{3.12}$$

has an isolated solution $(x_0, \varphi_0, \lambda_0)$, *where* $\varphi_0 \in \mathcal{N}_0 \cap V_k^\Gamma \cap X^\Sigma$ *if and only if the nondegeneracy condition* (3.10) *is satisfied.*

Alternative forms of the extended system (3.12) may also be used, for example

$$\begin{aligned} &\tilde{G}(x, \lambda) \equiv \begin{pmatrix} g(x, \lambda) \\ \langle \psi, g_x(x, \lambda)\varphi \rangle \end{pmatrix} = 0, \\ &\tilde{G} : X^\Gamma \times \mathbf{R} \to X^\Gamma \times \mathbf{R} \end{aligned} \tag{3.13}$$

where $\varphi, \psi \in V_k^\Gamma \cap X^\Sigma$ are the singular vectors associated with the singular value $\sigma = \langle \psi, g_x(x, \lambda)\varphi \rangle$, which are related by the equations

$$\begin{aligned} g_x(x, \lambda)\varphi &= \sigma\psi, \qquad \langle \psi, \psi \rangle = 1, \\ g_x^*(x, \lambda)\psi &= \sigma\varphi, \qquad \langle \varphi, \varphi \rangle = 1. \end{aligned} \tag{3.14}$$

Finally, all the theory of this section extends to the situation when $g: X \times \mathbf{R} \to Y$ and satisfies the equivariance condition

$$\tilde{T}(\gamma) g(x, \lambda) = g(T(\gamma)x, \lambda), \qquad \gamma \in \Gamma \tag{3.15}$$

where T and $\tilde{T}$ are equivalent orthogonal representations of Γ on X and Y, respectively. In particular,

$$\tilde{P}_k^\Gamma g_x(x, \lambda) = g_x(x, \lambda) P_k^\Gamma$$

where $\tilde{P}_k^\Gamma$ is the projection onto the Γ-isotypic component $\tilde{V}_k^\Gamma$ of Y (since the dimension, character, and type of equivalent irreducible representations are constant). Thus $g_x(x, \lambda)\colon V_k^\Gamma \to \tilde{V}_k^\Gamma$. All the other results can be generalised in a similar way.

4. Scaling laws. For many problems, it is possible to take one solution and perform a "change of scale" to give another solution. Thus for bifurcation problems, distinct branches of solutions are often related by a simple scaling. This relation can be expressed as a *scaling law*, which is similar to the equivariance condition (3.2) but also involves the parameter of the problem. (See Scovel, Kevrekidis, and Nicolaenko (1988) for an example of such a scaling law.)

We first consider the problem of finding such scaling laws in a given problem and show that the symmetry of the problem defines a natural context for the existence of scaling laws. We then consider how the scaling law relates to the bifurcation theory of § 3 and show that, in some cases, the existence of bifurcating branches at a mode interaction point can be proved using the scaling.

Consider the bifurcation problem

$$g(x, \lambda) = 0, \qquad g: X \times \mathbf{R} \to Y \tag{4.1}$$

where X and Y are real Hilbert spaces and g is C^2. We suppose that g satisfies the equivariance condition

$$\tilde{T}(\gamma) g(x, \lambda) = g(T(\gamma)x, \lambda) \quad \forall \gamma \in \Gamma \tag{4.2}$$

where T and $\tilde{T}$ are equivalent representations of the compact Lie group Γ on X and Y, respectively. By Lemma 2.2, we can assume without loss of generality that T and $\tilde{T}$ are orthogonal representations. For the sake of simplicity we will take $Y = X$ and $T = \tilde{T}$ but all our results generalise to the case $Y \neq X$, $T \neq \tilde{T}$.

We now show that a natural context for scaling laws can be derived by considering the group Γ. Our aim is to define a subproblem of (4.1), where the group Γ acts on a "different scale," whose solutions are related by a scaling transformation to the solutions of (4.1). The solutions of the subproblem will, however, also be solutions of (4.1). The subproblem will be defined by use of a particular fixed-point space.

Let $\beta: \Gamma \to \Gamma$ be an epimorphism (i.e., β is a homomorphism and $\beta(\Gamma) = \Gamma$) with a nontrivial kernal K. Then the mapping

$$\Gamma/K \to \Gamma, \qquad \gamma K \to \beta(\gamma) \tag{4.3}$$

is an isomorphism by Lemma 2.1. The quotient group Γ/K has a natural representation on the fixed-point space X^K given by

$$T_K: \Gamma/K \to GL(X^K), \qquad T_K(\gamma K) \equiv T(\gamma)|_{X^K} \tag{4.4}$$

since K acts trivially on X^K. This representation is well defined since X^K is closed and Γ-invariant (as K is a normal subgroup of Γ) and so $T(\gamma)\colon X^K \to X^K$ is a homeomorphism by Lemma 2.4 (ii) for all $\gamma \in \Gamma$. It is well known and easily proved that $g\colon X^K \times \mathbf{R} \to X^K$ and so we define our subproblem to be $g_K(x, \lambda) = 0$, where $g_K \equiv g|_{X^K \times \mathbf{R}}$. Clearly g_K is equivariant with respect to the representation T_K of Γ/K.

We have established that the groups Γ/K and Γ are isomorphic. The corresponding representations on X^K and X, respectively, are equivalent if there exists a linear homeomorphism $h\colon X \to X^K$ such that

$$T_K(\gamma K)h \equiv T(\gamma)h = hT(\beta(\gamma)) \quad \forall \gamma \in \Gamma. \tag{4.5}$$

If such an h exists, we can assume without loss of generality that it is orthogonal by Lemma 2.3 since T_K and T are orthogonal representations. We are now in a position to define a scaling law:

If there exists an orthogonal linear homeomorphism $h\colon X \to X^K$ satisfying (4.5), *and constants $b, c, l \in \mathbf{R}\backslash\{0\}$, such that*

$$chg(x, \lambda) = g(bhx, l\lambda) \equiv g_K(bhx, l\lambda) \tag{4.6}$$

then (4.6) *is called a scaling law.*

If such a scaling law exists, then $h' \equiv hT(\gamma)$ is also an orthogonal linear homeomorphism from X to X^K for all $\gamma \in \Gamma$, which satisfies (4.6) and also satisfies (4.5) if $\gamma \in Z(\Gamma)$, the centre of Γ. The following result proves the converse.

THEOREM 4.1. *If h_1 and h_2 are scaling transformations that satisfy* (4.5) *and* (4.6) *for fixed constants $b, c, l \neq 0$, then $h_2 = h_1 T(\gamma)$ for some $\gamma \in Z(\Gamma)$, the centre of Γ (restricting attention to the Γ-equivariance of g).*

Proof. It is straightforward to show that if h_1 and h_2 both satisfy (4.6), then

$$h_1^{-1}h_2 g(x, \lambda) = g(h_1^{-1}h_2 x, \lambda).$$

Now $h_1^{-1}h_2 \in GL(X)$ and is orthogonal. As we have restricted attention to the Γ-equivariance of g, we conclude that $h_1^{-1}h_2 = T(\gamma)$ for some $\gamma \in \Gamma$. As h_1 and h_2 also satisfy (4.5), then $\gamma \in Z(\Gamma)$. □

Since nothing is gained by combining scaling and group transformations, we restrict attention to one scaling transformation h satisfying (4.5) and (4.6).

The existence of a scaling law imposes conditions on Γ and X. Since K is nontrivial, the requirement that Γ/K be isomorphic to Γ cannot be satisfied if Γ is a finite group or if K is not a finite group. Also, the existence of the linear homeomorphism $h\colon X \to X^K$ requires that X and X^K have the same dimension, which cannot hold if X is finite dimensional.

The scaling law defines a relation between the solutions of $g = 0$ and $g_K = 0$. Thus $(x, \lambda) \in X \times \mathbf{R}$ is a solution of $g = 0$ if and only if $(bhx, l\lambda) \in X^K \times \mathbf{R}$ is a solution of $g_K = 0$ (although, of course, solutions of $g_K = 0$ are also solutions of $g = 0$). Also, (4.5) defines a relation between the representations T of Γ on X and T_K of Γ/K on X^K with the property that, for any $x \in X$, $\gamma \in \Gamma$, there exists $y \in X$ such that

$$y = T(\beta(\gamma))x \;\Leftrightarrow\; hy = T_K(\gamma K)hx = T(\gamma)hx.$$

We now consider how a scaling law relates to the bifurcation theory of § 3. An immediate consequence of the scaling law (4.6) is that if bifurcation occurs at $(x_0, \lambda_0) \in X \times \mathbf{R}$ then it must also occur at $(bhx_0, l\lambda_0) \in X^K \times \mathbf{R}$, since both the primary and secondary branches are "rescaled" by h. We now attempt to express this relation more formally.

Let Ω be a subgroup of Γ and define

$$\Omega^\beta = \{\gamma \in \Gamma\colon \beta(\gamma) \in \Omega\}, \tag{4.7}$$

which is also a subgroup of Γ. Clearly $\beta(\Omega^\beta)=\Omega$, K is a (normal) subgroup of Ω^β, and from Lemma 2.1, Ω^β/K is isomorphic to Ω. We then have the following results.

LEMMA 4.2. *Let Ω_x and Ω_{hx} be the isotropy subgroups of $x\in X$ and $hx\in X^K$, respectively, where h satisfies* (4.5). *If Ω is a subgroup of Γ, then $\Omega_x=\Omega$ if and only if $\Omega_{hx}=\Omega^\beta$ (i.e., $(\Omega_x)^\beta=\Omega_{hx}$).*

Proof. First, suppose that $\Omega_x=\Omega$. Then for every $\gamma\in\Omega^\beta$,

$$T(\gamma)hx=hT(\beta(\gamma))x=hx$$

using (4.5) and as $\beta(\gamma)\in\Omega=\Omega_x$. Thus $\gamma\in\Omega_{hx}$ and so Ω^β is a subgroup of Ω_{hx}. Similarly, for every $\gamma\in\Omega_{hx}$,

$$hT(\beta(\gamma))x=T(\gamma)hx=hx$$

again using (4.5). Since h is a linear homeomorphism, we conclude that $T(\beta(\gamma))x=x$ and so $\beta(\gamma)\in\Omega_x=\Omega$. Hence $\gamma\in\Omega^\beta$ and so Ω_{hx} is a subgroup of Ω^β. Combining these results, we conclude that $\Omega_{hx}=\Omega^\beta$.

The converse is proved similarly. □

LEMMA 4.3. *If $h: X\to X^K$ is an orthogonal linear homeomorphism satisfying* (4.5), *then the restriction of h to X^Ω is an orthogonal linear homeomorphism onto X^{Ω^β} for every subgroup Ω of Γ.*

Proof. If $x\in X^\Omega$ and $\gamma\in\Omega^\beta$, then $\beta(\gamma)\in\Omega$ and so, by (4.5),

$$T(\gamma)hx=hT(\beta(\gamma))x=hx.$$

Thus $hx\in X^{\Omega^\beta}$. It remains to prove that the mapping $h: X^\Omega\to X^{\Omega^\beta}$ is a surjection since h is an orthogonal linear homeomorphism on X.

Let $y\in X^{\Omega^\beta}$. Then $y=hx$ for some $x\in X$. If $\tilde\gamma\in\Omega$, then there exists $\gamma\in\Omega^\beta$ such that $\beta(\gamma)=\tilde\gamma$ since β is surjective. Thus

$$hT(\tilde\gamma)x=T(\gamma)hx=T(\gamma)y=y=hx$$

using (4.5) and as $y\in X^{\Omega^\beta}$. Since h is a linear homeomorphism (on X), we conclude that $T(\tilde\gamma)x=x$ for all $\tilde\gamma\in\Omega$ and so $x\in X^\Omega$. Thus h is a surjection as required. □

LEMMA 4.4. *Let Ω be a subgroup of Γ and let h satisfy* (4.5). *Then W is a closed, Ω-invariant subspace of X if and only if $W^h\equiv hW$ is a closed, Ω^β-invariant subspace of X^K.*

Proof. Since h is a homeomorphism, W is closed if and only if W^h is closed. Let $w\in W$. If $\gamma\in\Omega^\beta$, then $\beta(\gamma)\in\Omega$, and by (4.5) we have

$$T(\gamma)hw=hT(\beta(\gamma))w.$$

If W is Ω-invariant, then for all $w\in W$, $T(\beta(\gamma))w\in W$ for all $\beta(\gamma)\in\Omega$ and so $T(\gamma)hw\in W^h$ for all $\gamma\in\Omega^\beta$. Thus W^h is Ω^β-invariant. Conversely, if W^h is Ω^β-invariant, then for all $w\in W$, $T(\gamma)hw\in W^h$ for all $\gamma\in\Omega^\beta$ and so $T(\beta(\gamma))w\in W$ for all $\gamma\in\Omega^\beta$, since h is a homeomorphism. Now $\beta:\Omega^\beta\to\Omega$ is surjective and so $T(\tilde\gamma)w\in W$ for all $\tilde\gamma\in\Omega$, $w\in W$. Thus W is Ω-invariant. □

LEMMA 4.5. *Let Ω be a subgroup of Γ and let h satisfy* (4.5). *Then W is an Ω-irreducible subspace of X if and only if $W^h\equiv hW$ is an Ω^β-irreducible subspace of X^K.*

Proof. The equivalent result that W is an Ω-reducible subspace of X if and only if W^h is an Ω^β-reducible subspace follows immediately from Lemma 4.4. □

LEMMA 4.6. *Let Ω be a subgroup of Γ and let h satisfy* (4.5). *Then W is an Ω-absolutely irreducible subspace of X if and only if $W^h\equiv hW$ is an Ω^β-absolutely irreducible subspace of X^K.*

Proof. First suppose that W is Ω-absolutely irreducible. Let $A: W^h \to W^h$ commute with the representation of Ω^β on W^h. If $x \in W$ and $\gamma \in \Omega^\beta$, then $\beta(\gamma) \in \Omega$ and we have

$$\begin{aligned} h^{-1}AhT(\beta(\gamma))x &= h^{-1}AT(\gamma)hx \\ &= h^{-1}T(\gamma)Ahx \\ &= T(\beta(\gamma))h^{-1}Ahx \end{aligned}$$

using (4.5) and since A commutes with $T(\gamma)$ for all $\gamma \in \Omega^\beta$. Thus $h^{-1}Ah$ commutes with the representation of Ω on W, and as W is Ω-absolutely irreducible, we conclude that $h^{-1}Ah = cI|_W$, $c \in \mathbf{R}$. Thus $A = cI|_{W^h}$ and so W^h is Ω^β-absolutely irreducible.

The converse is proved similarly. □

We now consider how the presence of a scaling law affects bifurcation theory. Henceforth, we will assume the existence of a smooth branch of solutions of (4.1) contained in $X^\Omega \times \mathbf{R}$ for some subgroup Ω of Γ, which we will refer to as the primary branch. As a consequence of the scaling law (4.6), by Lemma 4.3 there must also be another branch of solutions of (4.1), contained in $X^{\Omega^\beta} \times \mathbf{R}$, which we will refer to as the scaled branch. (Indeed there may be many scaled branches obtained by a repeated application of the scaling law (4.6), but for our analysis we consider only one such branch.)

In the results that follow, we use the notation

$$\mathcal{N}_0 = \text{Null}\,(g_x(x_0, \lambda_0)), \qquad \mathcal{N}_0^* = \text{Null}\,(g_x(x_0, \lambda_0)^*),$$

$$\mathcal{N}_h = \text{Null}\,(g_x(bhx_0, l\lambda_0)), \qquad \mathcal{N}_h^* = \text{Null}\,(g_x(bhx_0, l\lambda_0)^*),$$

where $(x_0, \lambda_0) \in X \times \mathbf{R}$ satisfies (4.1). We also make the assumption that $g_x(x, \lambda)$ is a Fredholm operator of index zero.

LEMMA 4.7. *If $x_0 \in X^\Omega$, then $\mathcal{N}_0$ is a closed, Ω-invariant subspace of X, $\mathcal{N}_h$ is a closed, Ω^β-invariant subspace of X, and $h\mathcal{N}_0 = \mathcal{N}_h \cap X^K$. Also, $\mathcal{N}_0$ is Ω-absolutely irreducible if and only if $h\mathcal{N}_0$ is Ω^β-absolutely irreducible.*

Proof. Since $g_x(x, \lambda)$ is bounded, both $\mathcal{N}_0$ and $\mathcal{N}_h$ are closed. It is well known and easily proved that if $x \in X^\Sigma$ for some subgroup Σ of Γ, then Null $(g_x(x, \lambda))$ is Σ-invariant. Thus since $x_0 \in X^\Omega$, $\mathcal{N}_0$ is Ω-invariant. Also, $hx_0 \in X^{\Omega^\beta}$ by Lemma 4.3 and so $\mathcal{N}_h$ is Ω^β-invariant. Taking the Fréchet derivative of (4.6) with respect to x and evaluating it at (x_0, λ_0) gives

$$chg_x(x_0, \lambda_0)\varphi = bg_x(bhx_0, l\lambda_0)h\varphi \tag{4.8}$$

for all $\varphi \in X$. Thus if $\varphi \in \mathcal{N}_0$, then $h\varphi \in \mathcal{N}_h$ and so $h\mathcal{N}_0 \subseteq \mathcal{N}_h \cap X^K$. Also, if $\tilde{\varphi} \in \mathcal{N}_h \cap X^K$, then $\tilde{\varphi} = h\varphi$ for some $\varphi \in X$ and so from (4.6), $\varphi \in \mathcal{N}_0$. Hence $h\mathcal{N}_0 = \mathcal{N}_h \cap X^K$. The final statement of the lemma regarding absolute irreducibility follows directly from Lemma 4.6. □

We now define a *bifurcation point* to be a point (x_0, λ_0) such that the Equivariant Branching Lemma (Theorem 3.4) guarantees the existence of a secondary branch of solutions bifurcating from (x_0, λ_0), and we then have the following result.

THEOREM 4.8. *If $x_0 \in X^\Omega$, then (x_0, λ_0) is a bifurcation point where the secondary branch of solutions has isotropy subgroup Σ if and only if $(bhx_0, l\lambda_0)$ is a bifurcation point where the secondary branch of solutions has isotropy subgroup Σ^β.*

Proof. We must prove that the conditions of the Equivariant Branching Lemma hold at (x_0, λ_0) if and only if they hold at $(bhx_0, l\lambda_0)$. First, if $\dim \mathcal{N}_0 > 0$ then $\dim \mathcal{N}_h > 0$, since $h\mathcal{N}_0 \subseteq \mathcal{N}_h$ by Lemma 4.7 and h is a homeomorphism. Conversely, if $\dim \mathcal{N}_h > 0$

and $\dim(\mathcal{N}_h \cap X^{\Sigma^\beta}) = 1$ then $\dim(\mathcal{N}_h \cap X^K) \geqq 1$ since K is a subgroup of Σ^β. Thus $\dim \mathcal{N}_0 > 0$ since $h\mathcal{N}_0 = \mathcal{N}_h \cap X^K$ by Lemma 4.7 and h is a homeomorphism.

Now consider the three hypotheses of the Equivariant Branching Lemma:

(i) We observe that, since h is injective,

$$\begin{aligned} h(\mathcal{N}_0 \cap X^\Omega) &= h\mathcal{N}_0 \cap X^{\Omega^\beta} \\ &= \mathcal{N}_h \cap X^K \cap X^{\Omega^\beta} \\ &= \mathcal{N}_h \cap X^{\Omega^\beta} \end{aligned}$$

using Lemma 4.3, Lemma 4.7, and the fact that $X^{\Omega^\beta} \subseteq X^K$ since K is a subgroup of Ω^β. Thus $\mathcal{N}_0 \cap X^\Omega = \{0\}$ if and only if $\mathcal{N}_h \cap X^{\Omega^\beta} = \{0\}$ since h is a homeomorphism.

(ii) By an argument similar to (i), we have $h(\mathcal{N}_0 \cap X^\Sigma) = \mathcal{N}_h \cap X^{\Sigma^\beta}$ since K is a subgroup of Σ^β. Thus $\dim(\mathcal{N}_0 \cap X^\Sigma) = 1$ if and only if $\dim(\mathcal{N}_h \cap X^{\Sigma^\beta}) = 1$ since h is a homeomorphism.

(iii) For the sake of brevity, we introduce the notation $g^0 = g(x_0, \lambda_0)$, $g^h = g(bhx_0, l\lambda_0)$, etc. The nondegeneracy condition to be satisfied at (x_0, λ_0) is

$$\langle \psi_0, g^0_{x\lambda}\varphi_0 + g^0_{xx}\varphi_0 v\rangle \neq 0 \tag{4.9}$$

where $\varphi_0 \in \mathcal{N}_0 \cap X^\Sigma$, $\psi_0 \in \mathcal{N}_0^* \cap X^\Sigma$ and $v \in X^\Omega$ is the unique solution of

$$g^0_x v + g^0_\lambda = 0. \tag{4.10}$$

Similarly, the nondegeneracy condition to be satisfied at $(bhx_0, l\lambda_0)$ is

$$\langle \psi_h, g^h_{x\lambda}\varphi_h + g^h_{xx}\varphi_h v_h\rangle \neq 0 \tag{4.11}$$

where $\varphi_h \in \mathcal{N}_h \cap X^{\Sigma^\beta}$, $\psi_h \in \mathcal{N}_h^* \cap X^{\Sigma^\beta}$, and $v_h \in X^{\Omega^\beta}$ is the unique solution of

$$g^h_x v_h + g^h_\lambda = 0. \tag{4.12}$$

By differentiating the scaling law (4.6), it is easily shown that $v \in X^\Omega$ is a solution of (4.10) if and only if $v_h = (b/l)hv \in X^{\Omega^\beta}$ is a solution of (4.12) since $b, c, l \neq 0$. Also, we saw in (ii) that $h(\mathcal{N}_0 \cap X^\Sigma) = \mathcal{N}_h \cap X^{\Sigma^\beta}$ and so we can write $\varphi_h = h\varphi_0$. As h is an orthogonal transformation, we have

$$\begin{aligned} \langle x_1, g^0_x x_2\rangle &= \langle hx_1, hg^0_x x_2\rangle \\ &= \frac{b}{c}\langle hx_1, g^h_x hx_2\rangle \end{aligned}$$

for all $x_1, x_2 \in X$ using (4.8). Thus

$$\langle (g^0_x)^* x_1, x_2\rangle = \frac{b}{c}\langle (g^h_x)^* hx_1, hx_2\rangle,$$

and so $x_1 \in \mathcal{N}_0^*$ if and only if $hx_1 \in \mathcal{N}_h^*$, that is, $h\mathcal{N}_0^* = \mathcal{N}_h^* \cap X^K$. Hence $h(\mathcal{N}_0^* \cap X^\Sigma) = \mathcal{N}_h^* \cap X^{\Sigma^\beta}$ by Lemma 4.3 and the fact that $X^{\Sigma^\beta} \subseteq X^K$, and so we can write $\psi_h = h\psi_0$. It is then a matter of calculation to verify that

$$\begin{aligned} \langle \psi_h, g^h_{x\lambda}\varphi_h + g^h_{xx}\varphi_h v_h\rangle &= \left\langle h\psi_0, g^h_{x\lambda}h\varphi_0 + \frac{b}{l}g^h_{xx}h\varphi_0 hv\right\rangle \\ &= \left\langle h\psi_0, \frac{c}{bl}h(g^0_{x\lambda}\varphi_0 + g^0_{xx}\varphi_0 v)\right\rangle \\ &= \frac{c}{bl}\langle \psi_0, g^0_{x\lambda}\varphi_0 + g^0_{xx}\varphi_0 v\rangle \end{aligned}$$

since h is orthogonal. As $b, c, l \neq 0$, we conclude that (4.9) holds if and only if (4.11) holds. □

As a consequence of Theorem 4.8, it is clear that every bifurcation point on the primary branch can be rescaled to produce a bifurcation point on the scaled branch. However, a bifurcation point on the scaled branch can be scaled back onto the primary branch if and only if the isotropy subgroup $\tilde{\Sigma}$ of the secondary branch has K as a subgroup, since both the scaled branch and the secondary branch that bifurcates from it then have solutions in the range of h. In this case, the branch that bifurcates from the primary branch has isotropy subgroup $\beta(\tilde{\Sigma})$. It is possible that the symmetry of K can be lost at a bifurcation point on the scaled branch so that the bifurcation point cannot be scaled back to a bifurcation point of the primary branch.

The situation may occur where $\Omega^\beta = \Omega$ (e.g., $\Omega = \Gamma$), in which case the primary and scaled branches could coincide. However, Theorem 4.8 still holds in this situation and the two secondary branches then bifurcate from the same branch at different values of λ (provided that $l \neq 1$).

In § 3 (Theorem 3.5 and following) we saw that the type of bifurcation can be determined by considering $N_\Omega(\Sigma)/\Sigma$. It is intuitively obvious that the type of bifurcation from the primary and scaled branches must be the same. This is formalised by the following result.

THEOREM 4.9. *The mapping $\tilde{\beta}$ defined by*

$$\tilde{\beta}: N_{\Omega^\beta}(\Sigma^\beta)/\Sigma^\beta \to N_\Omega(\Sigma)/\Sigma, \qquad \gamma\Sigma^\beta \to \beta(\gamma)\Sigma$$

is an isomorphism.

Proof. We first show that $\beta(N_{\Omega^\beta}(\Sigma^\beta)) = N_\Omega(\Sigma)$. Let $\gamma \in N_{\Omega^\beta}(\Sigma^\beta)$. Then $\gamma\Sigma^\beta = \Sigma^\beta\gamma$ by the definition of $N_{\Omega^\beta}(\Sigma^\beta)$ and so $\beta(\gamma)\Sigma = \Sigma\beta(\gamma)$ since $\beta(\Sigma^\beta) = \Sigma$. Thus $\beta(N_{\Omega^\beta}(\Sigma^\beta))$ is a subgroup of $N_\Omega(\Sigma)$.

Now let $\gamma \in N_\Omega(\Sigma)$. Then $\gamma\sigma\gamma^{-1} \in \Omega$ for all $\sigma \in \Sigma$. Now $\gamma = \beta(\tilde{\gamma})$ for some $\tilde{\gamma} \in \Gamma$ since β is surjective. Thus for any $\tilde{\sigma} \in \Sigma^\beta$, $\beta(\tilde{\gamma}\tilde{\sigma}\tilde{\gamma}^{-1}) = \gamma\sigma\gamma^{-1} \in \Sigma$, where $\sigma = \beta(\tilde{\sigma}) \in \Sigma$, and so $\tilde{\gamma}\tilde{\sigma}\tilde{\gamma}^{-1} \in \Sigma^\beta$ for all $\tilde{\sigma} \in \Sigma^\beta$ giving $\tilde{\gamma} \in N_{\Omega^\beta}(\Sigma^\beta)$. Thus $N_\Omega(\Sigma)$ is a subgroup of $\beta(N_{\Omega^\beta}(\Sigma^\beta))$ and so we conclude that $\beta(N_{\Omega^\beta}(\Sigma^\beta)) = N_\Omega(\Sigma)$.

It is straightforward to prove that $\tilde{\beta}$ is a homomorphism and so we now prove that it is bijective. Let $\gamma \in N_\Omega(\Sigma)$. Then $\gamma\Sigma \in N_\Omega(\Sigma)/\Sigma$. Since β is surjective, there exists $\tilde{\gamma} \in \Gamma$ such that $\beta(\tilde{\gamma}) = \gamma$. By the above result, we conclude that $\tilde{\gamma} \in N_{\Omega_\beta}(\Sigma^\beta)$. Thus $\tilde{\beta}(\tilde{\gamma}\Sigma^\beta) = \gamma\Sigma$ and $\tilde{\gamma}\Sigma^\beta \in N_{\Omega^\beta}(\Sigma^\beta)/\Sigma^\beta$ and so $\tilde{\beta}$ is surjective.

Let $\gamma_1, \gamma_2 \in N_{\Omega^\beta}(\Sigma^\beta)$. Then

$$\begin{aligned}
\tilde{\beta}(\gamma_1\Sigma^\beta) = \tilde{\beta}(\gamma_2\Sigma^\beta) &\Rightarrow \beta(\gamma_1)\Sigma = \beta(\gamma_2)\Sigma \\
&\Rightarrow \beta(\gamma_2^{-1}\gamma_1)\Sigma = \Sigma \\
&\Rightarrow \beta(\gamma_2^{-1}\gamma_1) \in \Sigma \\
&\Rightarrow \gamma_2^{-1}\gamma_1 \in \Sigma^\beta \\
&\Rightarrow \gamma_2^{-1}\gamma_1\Sigma^\beta = \Sigma^\beta \\
&\Rightarrow \gamma_1\Sigma^\beta = \gamma_2\Sigma^\beta
\end{aligned}$$

using the definition of $\tilde{\beta}$ and the homomorphism property of β. Thus $\tilde{\beta}$ is also injective. We therefore conclude that $\tilde{\beta}$ is an isomorphism since it is a bijective homomorphism of Lie groups [BtD, p. 22]. □

Throughout our analysis, we have not been able to say that $\mathcal{N}_h = h\mathcal{N}_0$, only that $h\mathcal{N}_0 \subseteq \mathcal{N}_h$. In the generic situation, both $\mathcal{N}_0$ and $\mathcal{N}_h$ are (absolutely) irreducible (Theorem 3.2), in which case equality holds by Lemma 4.5 (cf. Lemma 4.6 also).

However, in a two-parameter problem, it is possible that mode interaction could occur on the scaled branch so that $\mathcal{N}_h \neq h\mathcal{N}_0$ at a critical value of the second parameter. In this case, Theorem 4.8 still applies, ensuring the existence of a secondary branch bifurcating from the scaled branch, thus providing some information about the bifurcating branches at a point of mode interaction.

We note that it is not possible for a scaled branch to be the continuation of the primary branch, since the branches have different isotropy subgroups, and the isotropy subgroup of a branch of solutions is preserved globally along the branch (Healey (1988a)). However, the primary and scaled branches could intersect at a bifurcation point.

The purpose of this analysis has been to determine when branches of solutions are related by a simple scaling, since the numerical computation of two such branches is clearly unnecessary and wasteful. Thus we conclude that it is sufficient to compute only those branches of solutions whose isotropy subgroup Ω satisfies either $\Omega^\beta = \Omega$ or $\Omega^\beta \neq \Omega$ and K is not a subgroup of Ω. However, it may be necessary to scale the solution up (in the second case) in order to check for the existence of bifurcation points on the scaled branch that do not occur on the primary branch.

5. Application to the gravity wave problem. In this section, we apply the theory of the previous sections to the problem of determining the profile of periodic, two-dimensional, irrotational travelling waves on the surface of an inviscid, incompressible fluid of constant density in a channel of infinite depth in the presence of gravity g, but without surface tension. By taking a frame of reference that moves with the travelling waves, the problem can be posed as a steady-state equation involving c, the speed of the flow at infinite depth, as the bifurcation parameter. Bifurcation from the trivial solution was discussed by Nekrasov (1920), who formulated the problem as an integral equation on the free surface, and by Levi-Civita (1925). More recently, Chen and Saffman (1980) and Saffman (1980) have discovered numerically secondary bifurcations on the primary branches far away from the trivial solution. We extend their results and describe a pattern which occurs in the secondary bifurcations.

The profile of the free surface can be found by solving the following differential equation (Okamoto (1990)):

$$G(\theta, \nu) \equiv \nu e^{-2\tau(s)}\tau'(s) + e^{\tau(s)} \sin \theta(s) = 0 \tag{5.1}$$

where $\nu = c^2/g$, the 2π-periodic function $\theta(s)$ given by

$$\theta(s) \sim \sum_{k=1}^{\infty} (a_k \sin ks + b_k \cos ks) \tag{5.2}$$

is the angle between the horizontal and the free surface at the point $(x(s), y(s))$ defined by

$$(x(s), y(s)) = -\int_0^s (e^{\tau(t)} \cos \theta(t), e^{\tau(t)} \sin \theta(t))\, dt, \tag{5.3}$$

and $-\tau$ is the conjugate of θ, given by

$$\tau(s) \sim \sum_{k=1}^{\infty} (a_k \cos ks - b_k \sin ks). \tag{5.4}$$

By integrating around a contour in the $x-y$ plane, it can be shown (see Okamoto (1990)) that

$$\int_0^{2\pi} e^{\tau(s)} \sin \theta(s)\, ds = 0. \tag{5.5}$$

This condition ensures that the resulting wave profile is periodic since (5.5) is equivalent to $y(2\pi)=0$ and $y(0)=0$ from (5.3). A similar argument can be used to show that

$$\int_0^{2\pi} \theta(s)\, ds = 0, \tag{5.6}$$

which is the justification for omitting the constant term from the Fourier series in (5.2).

We pose this problem in a Hilbert space setting so that the results of the previous sections can be applied. Thus we define

$$X = \left\{ f(s) \sim \sum_{k=1}^{\infty} (a_k \sin ks + b_k \cos ks) : \sum_{k=1}^{\infty} k^2(a_k^2 + b_k^2) < \infty \right\}$$

with inner product

$$\langle f_1, f_2 \rangle_X = \frac{1}{\pi} \int_0^{2\pi} f_1'(s) f_2'(s)\, ds$$

and similarly

$$Y = \left\{ f(s) \sim \sum_{k=1}^{\infty} (a_k \sin ks + b_k \cos ks) : \sum_{k=1}^{\infty} (a_k^2 + b_k^2) < \infty \right\}$$

with inner product

$$\langle f_1, f_2 \rangle_Y = \frac{1}{\pi} \int_0^{2\pi} f_1(s) f_2(s)\, ds.$$

Clearly $X \subset H^1[0, 2\pi]$, $Y \subset L^2[0, 2\pi]$, and $X \subset Y$. Also, if $\theta \in X$, then $\tau \in X$ and so $G: X \times \mathbf{R} \to Y$ (as the constant component of G vanishes due to (5.5)).

Nekrasov's integral equation can be derived from this formulation by solving (5.1) for τ in terms of θ and ν and then substituting the resulting expression for τ into an integral equation expressing the conjugacy relation between θ and τ.

Having set the problem up, we must determine the group that satisfies the equivariance condition (3.2) and look for scaling laws of the form (4.6). We define the following action of the group $O(2)$ on Y:

$$R_\alpha f(s) = f(s+\alpha), \qquad \alpha \in [0, 2\pi), \tag{5.7a}$$

$$Sf(s) = -f(-s), \tag{5.7b}$$

$$SR_\alpha f(s) = -f(-s-\alpha), \qquad \alpha \in [0, 2\pi). \tag{5.7c}$$

Then X is an $O(2)$-invariant subspace of Y. This action defines representations of $O(2)$ on X and Y. It follows from (5.2) and (5.4) that $S\tau(s) = \tau(-s)$, and it is then straightforward to show that G is equivariant with respect to the representations of $O(2)$ on X and Y induced by the action (5.7). Also, the inner products on X and Y are $O(2)$-invariant, and so the representations of $O(2)$ on X and Y induced by the group action (5.7) are orthogonal representations. Further, X and Y are linearly homeomorphic and the representations on X and Y are equivalent as required.

When seeking scaling laws of the form (4.6), we first determine all the discrete normal subgroups of $O(2)$. These consist of the finite cyclic groups Z_n, $n \in \mathbf{Z}^+$, generated by $R_{2\pi/n}$, which are the kernels of the epimorphisms $\beta_n: O(2) \to O(2)$ defined by

$$\beta_n(R_\alpha) = R_{n\alpha}, \qquad \alpha \in [0, 2\pi),$$

$$\beta_n(S) = S.$$

The fixed-point subspaces corresponding to the finite cyclic groups are

$$X^{Z_n} = \{f \in X : f(s) = f(s + 2\pi/n)\}$$

and similarly for Y^{Z_n}, $n \in \mathbf{Z}^+$. We now define the mapping h_n by

$$h_n f(s) = f(ns), \qquad n \in \mathbf{Z}^+.$$

Clearly h_n maps X onto X^{Z_n} and Y onto Y^{Z_n}. It is easily verified that h_n is an orthogonal linear homeomorphism which satisfies

$$R_\alpha h_n = h_n R_{n\alpha},$$

$$S h_n = h_n S.$$

Thus h_n meets all the requirements for a scaling transformation and we observe that G satisfies the scaling law

$$h_n G(\theta, \nu) = G(h_n \theta, \nu/n), \qquad n \in \mathbf{Z}^+. \tag{5.8}$$

We note that by Theorem 4.1 the scaling transformations h_n are uniquely defined since the centre of $O(2)$ is trivial.

5.1. Bifurcation from the trivial solution. The gravity wave problem (5.1) has the trivial solution $\theta = 0$ for all $\nu \geqq 0$ which corresponds to a flat free surface. This is the only solution that has $O(2)$ as its isotropy subgroup. We first consider bifurcation from the trivial solution, which is of course, well known, and so we only review it briefly in the current framework. Linearising (5.1) about the trivial solution gives

$$G_\theta(0, \nu)\tilde{\theta} = \nu\tilde{\tau}' + \tilde{\theta}$$

for all $\tilde{\theta} \in X$, where $-\tilde{\tau}$ is the conjugate of $\tilde{\theta}$.

The nontrivial irreducible representations of $O(2)$ on X (and Y) are

$$R_\alpha = \begin{pmatrix} \cos n\alpha & \sin n\alpha \\ -\sin n\alpha & \cos n\alpha \end{pmatrix}, \quad S = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad n = 1, 2, 3, \cdots.$$

It is easily verified that all these irreducible representations are absolutely irreducible. The corresponding isotypic components of X are

$$V_n = \text{span}\,\{\sin ns, \cos ns\}, \qquad n = 1, 2, 3, \cdots$$

and similarly for Y. It is interesting to observe that in this case, the isotypic components are irreducible. We note that the irreducible representation $R_\alpha = I$, $S = -I$ does not arise since $\theta(s) = c$, $c \in \mathbf{R}$ does not belong to X (or Y).

When $\nu = 1/n$, $n \in \mathbf{Z}^+$, then Null $(G_\theta(0, \nu)) = V_n$ and is thus irreducible. Now the isotropy subgroup of V_n is the cyclic group Z_n generated by $R_{2\pi/n}$. However, dim $(V_n \cap X^{D_n}) = 1$, where D_n is the dihedral group generated by $R_{2\pi/n}$ and S. It is a matter of calculation to verify that the nondegeneracy condition (3.10) is satisfied with $\varphi_0 = \sin ns \in V_n \cap X^{D_n}$. Thus by the Equivariant Branching Lemma, there is a branch of solutions with isotropy subgroup D_n which bifurcates from the trivial solution at $\nu = 1/n$. On such a branch, θ has the form

$$\theta(s) \sim \sum_{k=1}^{\infty} a_k \sin nks.$$

The normaliser of D_n in $O(2)$ is D_{2n} and $D_{2n}/D_n \cong Z_2$. Thus each bifurcation is a symmetric pitchfork by Theorem 3.5 and subsequent assertions. The distinct conjugate solutions of $\theta \in X^{D_n}$ are $R_\alpha \theta(s)$ for $\alpha \in (0, 2\pi/n)$ and so there is a one-parameter

"sheet" of solutions (see Sattinger (1973)). The only conjugate solution that is symmetric (relative to S) is

$$R_{\pi/n}\theta(s) \sim \sum_{k=1}^{\infty} (-1)^k a_k \sin nks,$$

which is the conjugate branch of the pitchfork. We will refer to the branch with $a_1 > 0$ initially as primary branch n and to the branch with $a_1 < 0$ initially as primary conjugate branch n.

When θ is symmetric (relative to S), it is an odd function of s. However, the wave profile is then an even function of x.

Clearly, the scaling transformations h_n do not affect the trivial solution. If we let $\Sigma = \{I, S\}$, then $\Sigma^{\beta_n} = D_n$, $n = 2, 3, 4, \cdots$, where Σ^{β_n} is defined by (4.7). Using the scaling law (5.8), we can see from Theorem 4.8 that all the primary branches which bifurcate from the trivial solution and have isotropy subgroup D_n are a scaled copy of the first branch, and so there is essentially only one distinct primary branch.

5.2. Bifurcation from D_n-symmetric branches. In general, it is not possible to study secondary bifurcation analytically, and so we solve the problem numerically. Thus we now consider the possible modes of secondary bifurcation from D_n-symmetric branches, assuming the generic conditions that the nullspace of $G_\theta(\theta, \nu)$, $\theta \in X^{D_n}$ is D_n-irreducible and that the nondegeneracy condition (3.10) is satisfied. The nontrivial irreducible representations of the group D_n, generated by $R_{2\pi/n}$ and S, on X (and Y) are given by

(i) $R_{2\pi/n} = I, \quad S = -I,$

(ii) $R_{2\pi/n} = -I, \quad S = I \quad (n \text{ even}),$

(iii) $R_{2\pi/n} = -I, \quad S = -I \quad (n \text{ even})$

(iv) $$R_{2\pi/n} = \begin{bmatrix} \cos\dfrac{2\pi m}{n} & \sin\dfrac{2\pi m}{n} \\ -\sin\dfrac{2\pi m}{n} & \cos\dfrac{2\pi m}{n} \end{bmatrix},$$

$$S = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

$$m = 1, \cdots, \tfrac{1}{2}n - 1 \quad (n \text{ even}),$$

$$= 1, \cdots, \tfrac{1}{2}(n-1) \quad (n \text{ odd}).$$

Again, all of these irreducible representations are absolutely irreducible and the corresponding isotypic components of X are

$$\tilde{V}_1 = \left\{ f(s) \sim \sum_{k=1}^{\infty} a_k \cos nks : f \in X \right\},$$

$$\tilde{V}_2 = \left\{ f(s) \sim \sum_{k=0}^{\infty} a_k \sin\left(nk + \frac{n}{2}\right)s : f \in X \right\} \quad (n \text{ even}),$$

$$\tilde{V}_3 = \left\{ f(s) \sim \sum_{k=0}^{\infty} a_k \cos\left(nk + \frac{n}{2}\right)s : f \in X \right\} \quad (n \text{ even}),$$

$$\tilde{V}_{4,m} = \Big\{ f(s) \sim \sum_{k=0}^{\infty} [a_k \sin(nk+m)s + b_k \sin(n(k+1)-m)s + c_k \cos(nk+m)s + d_k \cos(n(k+1)-m)s] : f \in X \Big\},$$

$$m = 1, \cdots, \tfrac{1}{2}n - 1 \quad (n \text{ even}),$$
$$= 1, \cdots, \tfrac{1}{2}(n-1) \quad (n \text{ odd}),$$

and similarly for Y. Since the linear operator $G_\theta(\theta, \nu)$, $\theta \in X^{D_n}$ maps each isotypic component of X to the corresponding isotypic component of Y (by the generalisation of Theorem 3.2), we can consider the situation when $\mathcal{N} \equiv \text{Null}\,(G_\theta(\theta, \nu))$ is an irreducible subspace of each isotypic component separately. By differentiating the equivariance condition (3.15) with respect to x, it is easily verified that if $\varphi \in \text{Null}\,(G_\theta(\theta, \nu))$, then $R_{\pi/n}\varphi \in \text{Null}\,(G_\theta(R_{\pi/n}\theta, \nu))$, and so for every singular point on the primary branch, there is a corresponding singular point on the primary conjugate branch to consider. We are then interested in any conjugacy relation between the secondary branches that bifurcate from the primary and primary conjugate branches.

(i) At every solution $(\theta, \nu) \in X^{D_n} \times \mathbf{R}$ of (5.1), we have

$$G_\theta(\theta, \nu) A\theta = 0$$

where $A\theta \equiv (d/d\alpha)\theta(s+\alpha)|_{\alpha=0} = \theta'(s)$. This is easily shown by differentiating the R_α-equivariance condition with respect to α and evaluating at $\alpha = 0$. Now if $\theta \in X^{D_n}$, then $\theta'(s) \in \tilde{V}_1$; therefore the component of $G_\theta(\theta, \nu)$ associated with $\tilde{V}_1$ is singular at every (nontrivial) solution point, and so the Equivariant Branching Lemma cannot be applied directly in this case. Zufiria (1987b) has investigated this class of bifurcation for the gravity wave problem and we will not consider it further. A consideration of bifurcation of this type for more general problems is given in Aston, Spence, and Wu (1990).

(ii) If $\mathcal{N} \subset \tilde{V}_2$ is irreducible, then $\dim \mathcal{N} = 1$ and so the bifurcation will be a symmetric pitchfork by Theorem 3.5. The isotropy subgroup of $\tilde{V}_2$ is $D_{n/2}$ generated by $R_{4\pi/n}$ and S and so this case corresponds to a period-doubling bifurcation. The secondary conjugate branch is obtained by applying $R_{2\pi/n}$. Now if $\varphi \in \mathcal{N} \subset \tilde{V}_2$, then $R_{\pi/n}\varphi \in \tilde{V}_3$, and so the corresponding secondary branches which bifurcate from the primary conjugate branch are associated with a different isotypic component and are not symmetric relative to S. We note, however, that all these branches are $O(2)$-conjugate and so there is essentially only one distinct secondary branch of solutions at the bifurcation point. We will refer to this case as bifurcation of mode $(n/2)^+$.

(iii) If $\mathcal{N} \subset \tilde{V}_3$ is irreducible, then $\dim \mathcal{N} = 1$. The isotropy subgroup of $\tilde{V}_3$ is the dihedral group $\tilde{D}_{n/2}$, generated by $R_{4\pi/n}$ and $SR_{2\pi/n}$. Thus we again have a symmetric pitchfork, period-doubling bifurcation. In this case, if $\varphi \in \mathcal{N} \subset \tilde{V}_3$ then $R_{\pi/n}\varphi \in \tilde{V}_2$ and so the secondary branches that bifurcate from the primary conjugate branch are symmetric with respect to S. Again, there is essentially only one distinct secondary branch of solutions. We will refer to this case as bifurcation of mode $(n/2)^-$. The period-doubling scenario is summarised in Fig. 1.

Before moving on, we must emphasise the difference in the isotropy subgroups of the secondary branches emanating from the primary and primary conjugate branches associated with period-doubling bifurcation. In Toland and Jones (1985) and Jones and Toland (1986), bifurcation at a double eigenvalue in the capillary-gravity wave problem (i.e., when the surface tension is nonzero) is considered, together with the effect of a small perturbation in the surface tension from a critical value. The group

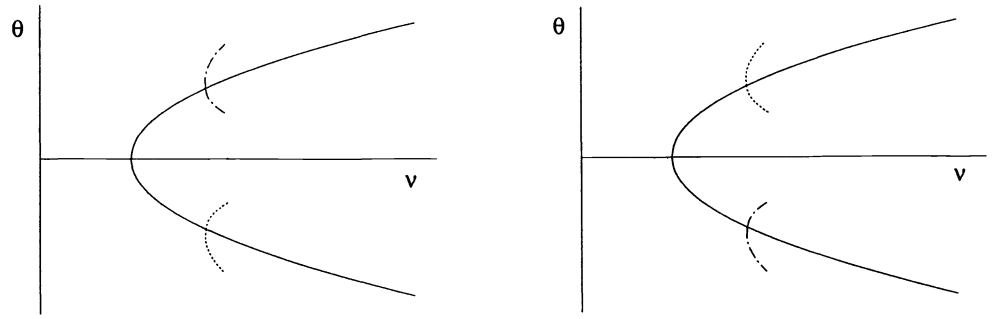


FIG. 1. *Period-doubling bifurcations.* —— D_n*-symmetric solutions,* – · – $D_{n/2}$*-symmetric solutions, and* ····· $\tilde{D}_{n/2}$*-symmetric solutions.*

that satisfies the equivariance condition for the capillary-gravity wave problem is the same group $O(2)$, and so the period-doubling scenario just described applies for this problem. Toland and Jones considered only symmetric waves (that is, solutions which satisfy $S\theta = \theta$) and found that at mode interactions between branch 2 and higher branches, a perturbation in the surface tension resulted in secondary bifurcations on branch 2 on both the primary and primary conjugate branches. They then claimed that these bifurcation points are related by symmetry, which clearly cannot be true since they are associated with the different isotypic components $\tilde{V}_2$ and $\tilde{V}_3$. Thus we must conclude that the two secondary bifurcation points are, in fact, independent of each other.

(iv) We first observe that these D_n-irreducible representations restricted to the subgroup $Z_2 = \{I, S\}$ are not Z_2-irreducible. Thus by Theorem 3.3, the isotypic components $\tilde{V}_{4,m}$ can be decomposed as

$$\tilde{V}_{4,m} = \tilde{V}^s_{4,m} \oplus \tilde{V}^a_{4,m}$$

where

$$\tilde{V}^s_{4,m} = \{f \in \tilde{V}_{4,m} : Sf = f\},$$

$$\tilde{V}^a_{4,m} = \{f \in \tilde{V}_{4,m} : Sf = -f\}.$$

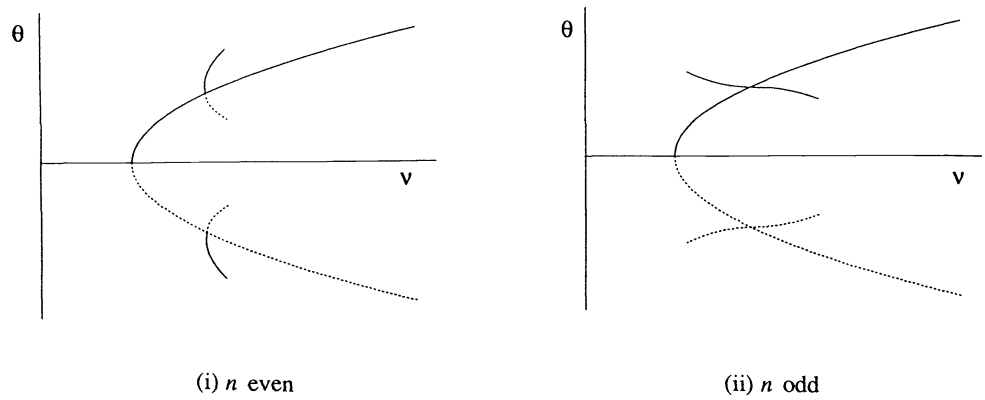
Also, the linear operator $G_\theta(\theta, \nu)$, $\theta \in X^{D_n}$ maps the subspace $\tilde{V}^s_{4,m}$ of X to the corresponding subspace of Y and similarly for $\tilde{V}^a_{4,m}$. Thus if $\mathcal{N} \subset \tilde{V}_{4,m}$ is D_n-irreducible, then $\dim \mathcal{N} = 2$ but $\dim(\mathcal{N} \cap \tilde{V}^s_{4,m}) = 1$ and $\dim(\mathcal{N} \cap \tilde{V}^a_{4,m}) = 1$. Now the isotropy subgroup of $\tilde{V}^a_{4,m}$ is the cyclic group Z_l, where $l = \gcd(m, n)$, whereas the isotropy subgroup of $\tilde{V}^s_{4,m}$ is the dihedral group D_l. Thus $\dim(\mathcal{N} \cap X^{D_l}) = 1$ and so there is a secondary branch of solutions with isotropy subgroup D_l. Clearly such a bifurcation can be detected numerically by a sign change in the determinant of the approximation to $G_\theta(\theta, \nu)|_{\tilde{V}^s_{4,m}}$ and so the linear operator $G_\theta(\theta, \nu)|_{\tilde{V}^a_{4,m}}$ need not be considered.

If n/l is odd, then the normaliser of D_l in D_n is simply D_l, and so the bifurcation is nonsymmetric by Theorem 3.5 and the solutions on the two secondary branches are distinct. However, in this case, a shift by π/l (i.e., an odd multiple of π/n) applied to $\varphi \in \mathcal{N} \cap X^{D_l}$ results in a symmetric function, and so the secondary branches bifurcating from the primary conjugate branch are $O(2)$-conjugate to those bifurcating from the primary branch.

If n/l is even, then the normaliser of D_l in D_n is D_{2l} and $D_{2l}/D_l \cong Z_2$. Thus the bifurcation is a symmetric pitchfork by Theorem 3.5, where the secondary conjugate

branch is obtained by applying the shift operator $R_{\pi/l}$. In this case, no shift by an odd multiple of π/n applied to $\varphi \in \mathcal{N} \cap X^{D_l}$ results in a symmetric function. Since dim $\mathcal{N} = 2$ on the primary conjugate branch, we can again restrict attention to the symmetric part of the isotypic component $\tilde{V}_{4,m}$, resulting in a distinct secondary branch of solutions bifurcating from the primary conjugate branch.

Thus in each case, there are two distinct branches of solutions which bifurcate. Dellnitz and Werner (1989) show that no other subgroups of D_n give rise to distinct branches of solutions via the Equivariant Branching Lemma. We will refer to this case as bifurcation of mode m (for $m < n/2$). These results are summarised in Fig. 2.



(i) n even (ii) n odd

FIG. 2. *Structure of secondary bifurcations from primary branch* $n \geqq 3$. —— *distinct solutions and* ····· *conjugate solutions.*

We note that if $l \neq 1$, then the secondary bifurcation point can be scaled down onto branch n/l and so, on any branch, the only bifurcation points that have not already occurred on lower branches are those for which m and n are coprime.

5.3. Numerical procedures for the gravity-wave problem. A pseudospectral collocation method is used to solve the gravity wave problem (5.1) numerically using the Fourier coefficients of the function θ as the degrees of freedom. This method has the advantage of spectral accuracy as θ is a smooth function. Also, the fixed-point subspaces and isotypic components of the finite-dimensional approximation spaces are easily identified. However, there is the drawback that the Jacobian matrix is full and has no banded structure.

On primary branch n, it is sufficient to solve for θ on the interval $(0, \pi/n)$ using the equally spaced meshpoints $s_i = (i\pi/(N+1)n)$, $i = 1, 2, \cdots, N$, where N is the number of sine functions in the approximation for $\theta \in X^{D_n}$. For a given vector of Fourier coefficients, both the functions θ and τ can be evaluated at the collocation points using only one Fast Fourier Transform (FFT). The FFT will be most efficient if N is chosen such that $N+1$ is a power of 2.

Since secondary bifurcation points must occur on the primary conjugate branch whenever they occur on the primary branch, it is sufficient to check only the primary branch for possible bifurcation points. When seeking bifurcation from branch n, $n \geqq 3$ of mode $m < n/2$, it is sufficient to solve for $\varphi \in \tilde{V}^S_{4,m}$ on the interval (or symmetry cell) $(0, 2\pi/n]$. We observe that in this case φ is always symmetric (relative to S). Now it is computationally convenient to work only with symmetric functions and so, when seeking bifurcation from branch 2, we can find both possible modes of bifurcation

by checking both the primary and primary conjugate branches for bifurcation with $\varphi \in \tilde{V}_2$ (so that φ is symmetric). The isotypic component $\tilde{V}_3$ does not then need to be considered. Also, it is sufficient to solve for $\varphi \in \tilde{V}_2$ on the interval $(0, \pi/2]$. Thus if N collocation points are used to solve for θ on a primary branch, all the possible bifurcation points can be found by considering only matrices of order at most $2N+2$, irrespective of the increase in the period of the solution on the secondary branch. This clearly illustrates the power of this approach.

In order to find a secondary bifurcation point, the path-following package PITCON (Rheinboldt (1986)) is used to follow a primary branch of solutions away from the trivial solution, initially using the first Fourier coefficient associated with the lowest-order sine function in the expression for θ as the continuation parameter. At each step, the determinant of the Jacobian matrix defined on an isotypic component (or the appropriate subspace of an isotypic component) of the finite-dimensional approximation space is evaluated. Clearly, many such matrices can be found and their determinants evaluated at each step and, as the matrices are dependent only on the primary branch solution, this procedure could be performed in parallel, resulting in a considerable saving in computation time. Once a sign change of a determinant has been detected, the bifurcation point can be found directly by solving the extended system (3.13) by Newton's method, using the last computed point on the primary branch as the initial approximation. The singular value σ and the singular vectors φ and ψ are found by performing inverse iteration on the eigenvalue equations

$$J^T J\varphi = \sigma^2 \varphi, \qquad JJ^T \psi = \sigma^2 \psi,$$

which can be derived from (3.14), where J is the Jacobian matrix under consideration. Only one LU-decomposition of the matrix J is required to perform the inverse iteration since forward and backward substitutions can then be used twice for each system. Generally, the vectors φ and ψ change very little at each Newton iteration of the extended system, and so one inverse iteration is usually sufficient to update φ and ψ between Newton iterates. Alternatively, the vectors φ and ψ need not be corrected after every Newton iterate, in which case an updated value of σ is obtained as

$$\sigma = \psi^T J\varphi$$

using the updated matrix J. This procedure saves the costly LU-decomposition of J at some of the Newton iterates.

The full Newton method is used to solve the extended system, as the cost of evaluating the function, which requires computing the singular value σ, is more expensive than the decomposition of the Jacobian of the extended system. Also, the same storage space can then be used for the matrix J and the Jacobian of the extended system so that storage for only one matrix is required.

The advantage of using the extended system (3.13) instead of (3.12) is that the vector of derivatives of the singular value σ with respect to the Fourier coefficients of θ can be evaluated efficiently using FFTs. Otherwise, this operation would be very costly.

Once a secondary bifurcation point has been located, branch switching is easily achieved by adding a small multiple of the singular vector φ onto the symmetric solution and using one of the (nonzero) Fourier coefficients of φ as the continuation parameter.

It is well known that the primary branch solutions develop sharp crests as the amplitude of the wave increases, resulting in a limiting wave profile that has a 120-degree corner at the crest (Amick, Fraenkel, and Toland (1982)). This produces a boundary layer in θ at the origin as the primary branches are followed away from the trivial

solution, and so the number of terms N in the Fourier expansion for θ has to be increased as the computation proceeds. Clearly, a local mesh refinement strategy would be required to follow the primary branches up towards the limiting wave, which is not possible with a spectral method (although see Tanaka (1983)). However, we will not be concerned with the near limiting waves.

5.4. Numerical results for the gravity-wave problem. Primary branches 2–10 were followed in the range $1/n \leqq \nu \leqq 1.19/n$, $n = 2, 3, \cdots, 10$, checking for all possible modes of bifurcation. On branch 2, no bifurcation of mode 1^- was found, but a bifurcation point of mode 1^+ was found. Exactly one of every other possible mode of bifurcation was found on each of the branches 3–10 in the specified range. These bifurcation points, scaled back onto the first primary branch for purposes of comparison, are presented in Table 1. The bifurcation points on branches 2 and 3 agree closely with the results of Chen and Saffman (1980), and those of mode 1 agree with the results of Saffman (1980) but are given to greater accuracy. The results in Table 1 were all computed with $N = 255$. By taking $N = 383$, we can verify that the error in the computed values of ν is of the order of 10^{-11}, whereas the error in the computed values of $\|\theta\|_\infty$ is of the order of 10^{-6}. Also, Zufiria (1987b) checked primary branch 1 for a bifurcation which breaks the reflectional symmetry (associated with the isotypic component $\tilde{V}_1$), and found none.

TABLE 1
Bifurcation points on branches 2–10.

Branch	Mode	$\|\theta\|_\infty$	ν	Chen and Saffman (1980) ν	Saffman (1980) ν
2	1^+	0.4406510	1.17535981841	1.175351	1.175
3	1	0.4402020	1.17513246000	1.175102	1.175
4	1	0.4396202	1.17483980293		1.175
5	1	0.4391432	1.17460105308		1.175
5	2	0.4404909	1.17527864507		
6	1	0.4387711	1.17441341967		
7	1	0.4384939	1.17426460109		
7	2	0.4399007	1.17498076008		
7	3	0.4405696	1.17531849738		
8	1	0.4382697	1.17414451964		
8	3	0.4404000	1.17523265296		
9	1	0.4380851	1.17404592240		
9	2	0.4393677	1.17471330800		
9	4	0.4406018	1.17533484504		
10	1	0.4379308	1.17396367239		1.174
10	3	0.4399996	1.17503053726		

From these results there appears to be an ordering of the bifurcation points. Let $\nu_{m,n}$ be the value of ν at which there is a bifurcation from branch n of mode m, scaled back to the first branch (by multiplying by n). Then these results suggest the following ordering on the bifurcation points:

$$\frac{m_1}{n_1} < \frac{m_2}{n_2} \Rightarrow \nu_{m_1,n_1} < \nu_{m_2,n_2}.$$

This ordering is the same as that found by Zufiria (1987a) for a Hamiltonian approximation to the finite-depth problem. Such an ordering has also been explored

by Scovel, Kevrekidis, and Nicolaenko (1988) in their investigation of the Kuramoto-Sivashinsky equation. It is interesting to observe that both Zufiria and Scovel et al. were solving fourth-order ordinary differential equations that are $O(2)$-equivariant and satisfy scaling laws of the form (4.6).

We seek to determine the value ν_∞ defined by

$$\nu_\infty = \lim_{n\to\infty} \nu_{1,n}.$$

To this end, we computed bifurcation points of mode 1 on higher branches; the results are given in Table 2. We make the assumption that as $n \to \infty$,

$$\nu_{1,n} \simeq \nu_\infty + \alpha \frac{1}{n}$$

for some $\alpha \in \mathbf{R}$. Using the last two values in Table 2 for $n = 500$ and $n = 1000$, we obtain an approximation to ν_∞ given by

$$\nu_\infty = 1.17307544040.$$

In the same way, we obtain an approximation to the norm of θ at $\nu = \nu_\infty$, which is given by

$$\|\theta(\nu_\infty)\|_\infty = 0.4362479.$$

Having established this pattern in the bifurcation structure, we conjecture that *there are no secondary bifurcation points on any primary branch with* $\|\theta\|_\infty < 0.4362479$.

5.5. Extension to nonperiodic problems. The procedure outlined in this section based on $O(2)$-equivariance can be applied to nonperiodic problems in certain cases. As an example, consider a second-order, ordinary differential equation of the form

$$\begin{aligned} &g(x, \lambda) \equiv x'' + f(x, x', \lambda) = 0, \\ &g : H^2[0, \pi] \times \mathbf{R} \to L^2[0, \pi], \end{aligned} \tag{5.9}$$

together with the homogeneous Neumann boundary conditions

$$x'(0) = x'(\pi) = 0. \tag{5.10}$$

We can define a corresponding periodic problem by $g(x, \lambda) = 0$, $g : H^2[-\pi, \pi] \times \mathbf{R} \to L^2[-\pi, \pi]$, together with 2π-periodic boundary conditions. The periodic problem is equivariant with respect to the translations

$$R_\alpha x(s) = x(s + \alpha), \qquad \alpha \in \mathbf{R},$$

and we assume that it is also equivariant with respect to the reflection

$$S_1 x(s) = x(-s).$$

TABLE 2
Bifurcation points of mode 1 *on higher branches.*

Branch	Mode	$\|\theta\|_\infty$	ν
20	1	0.4371592	1.17355440666
50	1	0.4366304	1.17327590512
100	1	0.4364422	1.17317717624
250	1	0.4363263	1.17311646942
500	1	0.4362871	1.17309598530
1000	1	0.4362675	1.17308571285

Thus the periodic problem is $O(2)$-equivariant. However, the only group action that respects the Neumann boundary conditions and leaves the interval $[0, \pi]$ invariant is $S_1 R_\pi$, the reflection about the midpoint of the interval.

Now any solution of the Neumann problem can be extended onto the interval $[-\pi, 0]$ using the reflection operator S_1. This extended function is C^1 at $s=0$ and it is also in $H^2[-\pi, \pi]$. Since g is equivariant with respect to S_1, the extended function satisfies (5.9) on the interval $[-\pi, \pi]$. Similarly, this function can be extended to a 2π-periodic function which satisfies (5.9) on the whole of the real line using the translation operators $R_{2\pi n}$, $n \in \mathbf{Z}$. Thus every solution of the Neumann problem can be extended to a solution of the periodic problem. Now the S_1-symmetric solutions of the periodic problem are precisely those which satisfy the Neumann boundary conditions (since 2π-periodic even functions are also even about $s=\pi$). Thus in order to find solutions of the Neumann problem, we can use the procedure described in this section to find S_1-symmetric solutions of the corresponding periodic problem. This technique has been employed previously by Fujii, Mimura, and Nishiura (1982) and by Armbruster and Dangelmayr (1987).

Similarly, if the boundary conditions associated with (5.9) are the homogeneous Dirichlet conditions

$$x(0) = x(\pi) = 0, \tag{5.11}$$

and g is equivariant with respect to R_α, $\alpha \in \mathbf{R}$ and the odd reflection S_2 defined by

$$S_2 x(s) = -x(-s),$$

then any solution of (5.9) which satisfies the boundary conditions (5.11) can be extended onto the interval $[-\pi, 0]$ using the reflection operator S_2. This extended function is C^1 at $s=0$, is in $H^2[-\pi, \pi]$, and satisfies (5.9) on the interval $[-\pi, \pi]$. As before, this function can be extended periodically over the real line. In this case, the S_2-symmetric solutions of the periodic problem are those which satisfy the boundary conditions (5.11). Thus we can find solutions of the Dirichlet problem by using the procedure described in this section to find S_2-symmetric solutions of the corresponding $O(2)$-equivariant periodic problem.

Many equations of the form (5.9) are equivariant with respect to the group generated by R_α, S_1, and S_2, which we will call $O^2(2)$. In this case, branches that bifurcate from the $O^2(2)$-symmetric solution (which is often the trivial solution) have isotropy subgroup generated either by $R_{2\pi/n}$, S_1, and $S_2 R_{\pi/n}$, in order to satisfy the Neumann boundary conditions (5.10), or by $R_{2\pi/n}$, S_2, and $S_1 R_{\pi/n}$ in order to satisfy the Dirichlet boundary conditions (5.11). These two classes of solution are conjugate solutions of the 2π-periodic problem and are related by the shift operator $R_{\pi/2n}$. In both cases, the group is the dihedral group D_{2n} generated by the "rotation" $S_1 S_2 R_{\pi/n}$ and the reflection either S_1 or S_2 (since $(S_1 S_2 R_{\pi/n})^2 = R_{2\pi/n}$). The analysis of secondary bifurcation is then very similar to that of § 5.2 but is applied to the group D_{2n}.

We now consider the case where (5.9) is defined on the interval $[0, \pi/2]$ and the associated boundary conditions are the mixed conditions

$$x(0) = x'(\pi/2) = 0. \tag{5.12}$$

If g is equivariant with respect to $O^2(2)$ generated by R_α, S_1, and S_2, then any solution of (5.9) that satisfies the boundary conditions (5.12) can be extended onto the interval $[\pi/2, \pi]$ using the reflection operator $S_1 R_\pi$, and this function can then be extended onto the interval $[-\pi, 0]$ using the reflection operator S_2. This extended function is in $H^2[-\pi, \pi]$, satisfies (5.9) on the interval $[-\pi, \pi]$, and can be extended

periodically over the real line as before. The solutions of the $O^2(2)$-equivariant, 2π-periodic problem which satisfy the boundary conditions (5.12) are those which are symmetric relative to S_2 and S_1R_π. Such solutions must have period $2\pi/n$ for some *odd* integer n.

In all of these cases, we have been able to extend the group of symmetries under consideration from the discrete group Z_2 or the trivial group to the continuous group $O(2)$ or $O^2(2)$). Thus it is possible that a scaling law could exist for the problem as described in § 4.

Finally, we consider the situation where the boundary conditions and the group reflection do not "match" as they do in the above examples. For the purposes of illustration, we will consider the case where g is equivariant with respect to R_α and S_1 only and we have the Dirichlet boundary conditions (5.11). If a solution of (5.9) that satisfies the boundary conditions (5.11) is extended onto $[-\pi, 0]$ using the reflection S_1, the resulting function is only C^0 at $s=0$ and so cannot be in $H^2[-\pi, \pi]$. Thus the procedure of selecting solutions of the 2π-periodic problem that have certain symmetries breaks down in this case. Hence we must consider only the group actions that respect the boundary conditions. This leaves the group Z_2 generated by the reflection S_1R_π about the midpoint of the interval. Bifurcation could then occur by breaking the reflectional symmetry, resulting in a symmetric pitchfork bifurcation. However, generically, secondary bifurcation will not occur since all the symmetry in the problem has been broken. Alternatively, if (5.9) has a trivial solution for all λ, then bifurcation from the trivial solution might occur without breaking the reflectional symmetry, and generically this would be a transcritical bifurcation. On such a primary branch, secondary bifurcation might occur by breaking the reflectional symmetry. Clearly, a scaling law of the type described in § 4 cannot hold in this case since the group Z_2 is discrete.

Thus we conclude that if g is just $O(2)$-equivariant, then the choice of boundary conditions makes a fundamental difference to the structure of the solution set of (5.9). All these extensions can also be applied to elliptic partial differential equations in rectangular domains of higher dimension.

Acknowledgments. I would like to acknowledge many helpful discussions with Professor J. F. Toland throughout the course of this work.

REFERENCES

C. J. AMICK, L. E. FRAENKEL, AND J. F. TOLAND (1982), *On Stokes conjecture for the wave of extreme form*, Acta Math., 148, pp. 193–214.

D. ARMBRUSTER AND G. DANGELMAYR (1987), *Coupled stationary bifurcations in nonflux boundary value problems*. Math. Proc. Cambridge Philos. Soc., 101, pp. 167–191.

P. J. ASTON, A. SPENCE, AND W. WU (1990), *Bifurcation to rotating waves in equations with $O(2)$-symmetry*, submitted to SIAM J. Appl. Math.

A. O. BARUT AND R. RACZKA (1986), *Theory of Group Representations and Applications*, Second revised edition, World Scientific, Singapore.

A. BOSSAVIT (1986), *Symmetry, groups and boundary value problems. A progressive introduction to noncommutative harmonic analysis of partial differential equations in domains with geometrical symmetry*, Comput. Math. Appl. Mech. Engrg., 56, pp. 167–215.

T. BRÖCKER AND T. TOM DIECK (1985), *Representations of Compact Lie Groups*, Springer-Verlag, Berlin, New York.

B. CHEN AND P. G. SAFFMAN (1980), *Numerical evidence for the existence of new types of gravity waves of permanent form on deep water*, Stud. Appl. Math., 62, pp. 1–21.

G. CICOGNA (1981), *Symmetry breakdown from bifurcation*, Lett. Nuovo Cimento (2), 31, pp. 600–602.

E. N. DANCER (1971), *Bifurcation theory in real Banach space*, Proc. London Math. Soc. (3), 23, pp. 699–734.

M. DELLNITZ AND B. WERNER (1989), *Computational methods for bifurcation problems with symmetries—with special attention to steady state and Hopf bifurcation points*, J. Comput. Appl. Math., 26, pp. 97–123.

J. B. FRALEIGH (1977), *A First Course in Abstract Algebra*, Second edition, Addison-Wesley, Reading, MA, London.

H. FUJII, M. MIMURA, AND Y. NISHIURA (1982), *A picture of the global bifurcation diagram in ecological interacting and diffusing systems*, Phys. D, 5, pp. 1–42.

M. GOLUBITSKY AND D. G. SCHAEFFER (1985), *Singularities and Groups in Bifurcation Theory*, Volume 1, Appl. Math. Sci., 51, Springer-Verlag, Berlin, New York.

M. GOLUBITSKY, I. STEWART, AND D. G. SCHAEFFER (1988), *Singularities and Groups in Bifurcation Theory*, Volume 2, Appl. Math. Sci., 69, Springer-Verlag, Berlin, New York.

T. J. HEALEY (1988a), *Global bifurcation and continuation in the presence of symmetry with an application to solid mechanics*, SIAM J. Math. Anal., 19, pp. 824–840.

———, (1988b), *A group theoretic approach to computational bifurcation problems with symmetry*, Comput. Methods Appl. Math. Engrg., 67, pp. 257–295.

M. C. W. JONES AND J. F. TOLAND (1986), *Symmetry and bifurcation of capillary-gravity waves*, Arch. Rational Mech. Anal., 96, pp. 29–53.

A. A. KIRILLOV (1976), *Elements of the Theory of Representations*, Springer-Verlag, Berlin, New York.

A. W. KNAPP (1986), *Representation Theory of Semisimple Groups: An Overview Based on Examples*, Princeton University Press, Princeton, NJ.

R. LAUTERBACH (1986), *An example of symmetry-breaking with submaximal isotropy*, in Multiparameter Bifurcation Theory, M. Golubitsky and J. Guckenheimer, eds., Contemp. Math., 56, American Mathematical Society, RI, pp. 217–222.

T. LEVI-CIVITA (1925), *Détermination rigoreuse des ondes permanantes d'ampleur finie*, Math. Anal., 93, pp. 264–314.

G. MOORE AND A. SPENCE (1980), *The calculation of turning points of nonlinear equations*, SIAM J. Numer. Anal., 17, pp. 567–576.

A. I. NEKRASOV (1920), *On Stokes waves*, Izv. Ivan.-Voznesesck Politekh. Inst., pp. 81–91.

H. OKAMOTO (1990), *On the problem of water waves of permanent configuration*, Nonlinear Anal., Theory, Methods, Appl., 14, pp. 469–481.

W. C. RHEINBOLDT (1986), *Numerical Analysis of Parameterised Nonlinear Equations*, Wiley-Interscience, New York.

P. G. SAFFMAN (1980), *Long wavelength bifurcation of gravity waves on deep water*, J. Fluid Mech., 101, pp. 567–581.

D. H. SATTINGER (1973), *Transformation groups and bifurcation at multiple eigenvalues*, Bull. Amer. Math. Soc., 79, pp. 709–711.

———, (1977), *Group representation theory and branch points of nonlinear functional equations*, SIAM J. Math. Anal., 8, pp. 179–201.

———, (1979), *Group Theoretic Methods in Bifurcation Theory*, Springer-Verlag, Berlin, New York.

J. C. SCOVEL, I. G. KEVREKIDIS, AND B. NICOLAENKO (1988), *Scaling laws and the prediction of bifurcations in systems modelling pattern formation*, Phys. Lett. A, 130, pp. 73–80.

M. TANAKA (1983), *The stability of steep gravity waves*, J. Phys. Soc. Japan, 52, pp. 3047–3055.

J. F. TOLAND AND M. C. W. JONES (1985), *The bifurcation and secondary bifurcation of capillary-gravity waves*, Proc. Roy. Soc. London Ser. A, 399, pp. 391–417.

A. VANDERBAUWHEDE (1982), *Local Bifurcation and Symmetry*, Res. Notes in Math. 75, Pitman, Boston, London.

B. WERNER (1988), *Computational methods for bifurcation problems with symmetries and applications to steady-states of n-box reaction-diffusion models*, in 1987 Dundee Conference on Numerical Analysis, D. F. Griffiths and G. A. Watson, eds., Pitman, Boston, London.

B. WERNER AND A. SPENCE (1984), *The computation of symmetry breaking bifurcation points*, SIAM J. Numer. Anal., 21, pp. 388–399.

J. A. ZUFIRIA (1987a), *Weakly nonlinear non-symmetric gravity waves on water of finite depth*, J. Fluid Mech., 180, pp. 371–385.

———, (1987b), *Non-symmetric gravity waves on water of infinite depth*, J. Fluid Mech., 181, pp. 17–39.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 213-237, January 1991

013

UNIFORM SIMPLIFICATION AT A TRANSITION POINT AND THE PROBLEM OF RESONANCE*

SHIGEMI OHKOHCHI†

This article is dedicated to Professor Tosihusa Kimura on the occasion of his 60th birthday.

Abstract. The sufficiency of the Matkowsky condition concerning the differential equation $\varepsilon y''+f(x,\varepsilon)y'+g(x,\varepsilon)y=0$ is considered under the assumption that $f(0,\varepsilon)=f'(0,\varepsilon)=\cdots=f^{(m-1)}(0,\varepsilon)=0$ and $f^{(m)}(0,\varepsilon)\neq 0$. It is proved that the Matkowsky condition implies resonance in the sense of N. Kopell. Y. Sibuya has proved such a problem if $f(0,\varepsilon)=0$ and $f'(0,\varepsilon)\neq 0$ $[m=1]$.

Key words. ordinary differential equation, turning point, singular perturbation

AMS(MOS) subject classification. 34E20

1. Introduction.

In this paper we consider a differential equation

$$\varepsilon\frac{d^2v}{dx^2}+F(x,\varepsilon)\frac{dv}{dx}+G(x,\varepsilon)v=0, \tag{1.1}$$

where F and G are holomorphic in two complex variables x and ε in a domain

$$x\in D_0,\ |\varepsilon|<\rho_0, \tag{1.2}$$

where D_0 is a domain in the x-plane and ρ_0 is a positive number. We assume that D_0 contains a real interval

$$I_0=\{x;\ -a\leqq \operatorname{Re} x\leqq b,\ \operatorname{Im} x=0\}, \tag{1.3}$$

where a and b are positive numbers. We also assume that

$$F(x,0)=-2x^m, \tag{1.4}$$

where m is a positive integer.

We say that the differential equation (1.1) satisfies the *Matkowsky condition on* I_0, if there exists a nontrivial formal power series solution of (1.1),

$$v(x,\varepsilon)=\sum_{n=0}^{\infty}v_n(x)\varepsilon^n, \tag{1.5}$$

such that all the $v_n(x)$ are bounded on the real interval I_0. We also say that the differential equation (1.1) exhibits *resonance in the sense of N. Kopell on* I_0, if there exists a solution $v(x,\varepsilon)$ that converges uniformly on I_0 as $\varepsilon\to+0$ to a nontrivial solution of the reduced equation

$$F(x,0)\frac{dv}{dx}+G(x,0)v=0. \tag{1.6}$$

We will prove the following main theorem.

THEOREM 1.1. *The Matkowsky condition implies resonance in the sense of N. Kopell.*

Y. Sibuya [9] has proved that the Matkowsky condition implies resonance for the case in which

$$F(x,0)=-2x, \tag{1.4}_{m=1}$$

* Received by the editors August 31, 1987; accepted for publication (in revised form) January 12, 1990.
† Faculty of Engineering, Oita University, Oita 870-11, Japan.

and D_0 is a disk with the center at $x=0$, i.e.,

$$D_0=\{x;\,|x|<r_0\} \quad \text{for some } r_0>0, \tag{1.7}$$

by using the properties of Whittaker's parabolic cylinder functions and Hermite polynomials. Furthermore, in that proof, Theorem 1.2 below, and a theorem concerning uniform simplification [8], play an important part.

THEOREM 1.2 (Y. Sibuya [9]). *Let*

$$S_j=\{\varepsilon;\,a_j<\arg\varepsilon<b_j,\,0<|\varepsilon|<\rho\},\qquad j=1,2,\cdots,\nu \tag{1.8}$$

be sectors in the complex ε-plane, where ρ is a positive number and the a's and the b's are real numbers. Let $\delta_1(\varepsilon),\cdots,\delta_\nu(\varepsilon)$ be functions of ε. Assume that

(i) $S_1\cup S_2\cup\cdots\cup S_\nu=\{\varepsilon;\,0<|\varepsilon|<\rho\}$.

(ii) *$\delta_j(\varepsilon)$ is holomorphic in S_j.*

(iii) *$\delta_j(\varepsilon)$ is asymptotically zero as $\varepsilon\to 0$ in S_j, i.e.,*

$$|\delta_j(\varepsilon)|\leqq K_N|\varepsilon|^N,\qquad N=0,1,\cdots \quad \text{in } S_j$$

for some positive numbers K_N.

(iv) *If $S_j\cap S_k\neq\varnothing$, we have*

$$|\delta_j(\varepsilon)-\delta_k(\varepsilon)|\leqq c_0\exp\left[-c_1/|\varepsilon|^\lambda\right] \quad \text{in } S_j\cap S_k \tag{1.9}$$

for some positive numbers c_0, c_1, and λ.

Then there exists a positive number H such that

$$|\delta_j(\varepsilon)|\leqq H\exp\left[-c_1/|\varepsilon|^\lambda\right] \quad \text{in } S_j \quad j=1,2,\cdots,\nu. \tag{1.10}$$

Lin [4] has shown that the condition that D_0 is a disk is not necessary, so that the sufficiency of the Matkowsky condition is proved for the general case for $m=1$, by the use of Lin's cohomological theorem, as follows.

THEOREM 1.3 (C. H. Lin [5]). *Let $S_j=\{\varepsilon;\,\alpha_j<\arg\varepsilon<\beta_j,\,0<|\varepsilon|<\rho\}$, $j=1,\cdots,\nu$, be sectors in the right half complex ε-plane, where $\rho>0$:*

$$-\tfrac{1}{2}\pi<-\omega_0<\alpha_1<\alpha_2<\beta_1<\alpha_3<\beta_2<\alpha_4<\beta_3<\cdots<\alpha_\nu<\beta_{\nu-1}<\beta_\nu=\omega_0<\tfrac{1}{2}\pi.$$

Let $\phi^1(\varepsilon),\phi^2(\varepsilon),\cdots,\phi^\nu(\varepsilon)$ be functions of ε. Assume that

(i) *$\phi^j(\varepsilon)$ is holomorphic in S_j and continuous on*

$$S_j^*=\{\varepsilon;\,\alpha_j\leqq\arg\varepsilon\leqq\beta_j,\,0<|\varepsilon|\leqq\rho\}.$$

(ii) *$\phi^j(\varepsilon)$ is asymptotically zero as ε tends to zero in S_j.*

(iii)
$$|\phi^1(\varepsilon)|\leqq H_1\exp\{-\mu\cdot\mathrm{Re}\,[1/|\varepsilon|]\}$$

on the line segment $\arg\varepsilon=-\omega_0$, $0<|\varepsilon|<\rho$;

$$|\phi^\nu(\varepsilon)|\leqq H_1\exp\{-\mu\cdot\mathrm{Re}\,[1/|\varepsilon|]\}$$

on the line segment $\arg\varepsilon=-\omega_0$, $0<|\varepsilon|<\rho$, *for some positive numbers μ and H_1.*

(iv)
$$|\phi^j(\varepsilon)-\phi^{j+1}(\varepsilon)|\leqq H_1\exp\{-\mu\cdot\mathrm{Re}\,[1/|\varepsilon|]\} \quad \text{in } S_j\cap S_{j+1} \quad (j=1,2,\cdots,\nu).$$

Then there exists a positive number H such that

$$|\phi^j(\varepsilon)|\leqq H\cdot\exp\{-\mu\cdot\mathrm{Re}\,[1/|\varepsilon|]\} \quad \text{in } S_j \quad (j=1,2,\cdots,\nu).$$

Matkowsky [6] has demonstrated that the resonance phenomenon occurs for functions $F(x, \varepsilon)$ more general than those having a single simple zero in the interval by the following example:

$$\varepsilon t'' - x^3(x^2-1)(x-2)^2 y' = 0, \qquad -a \leqq x \leqq b, \quad b > 1, \tag{1.11}$$

$$y(-2) = \alpha, \quad y(b) = \beta, \quad n \neq 2. \tag{1.12}$$

In this paper we will prove the theorem concerning uniform simplification in a full neighborhood of a higher-order transition point and treat the problem of resonance for the case in which $F(x, \varepsilon)(F(x, 0) = -2x''')$ has a higher-order zero in the interval I_0 and D_0 is a domain (*not disk*) in the x-plane containing the real interval $[-a, b]$. In this analysis we will use Theorems 1.2 and 1.3.

2. Reduction to a standard form. Let ρ_0 be a positive number, and let D be a domain in the complex ξ-plane which contains a real interval

$$I = \{\xi: -\alpha \leqq \operatorname{Re} \xi \leqq \beta, \operatorname{Im} \xi = 0\}, \tag{2.1}$$

where α and β are positive numbers.

We will consider a linear differential equation

$$\varepsilon \frac{d^2 v}{d\xi^2} + f(x, \varepsilon) \frac{dv}{d\xi} + g(x, \varepsilon) v = 0, \tag{2.2}$$

where f and g are holomorphic in two variables ξ and ξ in the domain

$$\xi \in D, \qquad |\varepsilon| < \rho_0. \tag{2.3}$$

Set

$$f_0(\xi) = f(\xi, 0). \tag{2.4}$$

We assume that

$$f_0(0) = f_0'(0) = \cdots = f_0^{(m-1)}(0) = 0, \qquad f_0^{(m)}(0) \neq 0, \tag{2.5}$$

$$\xi^m f_0(\xi) < 0 \quad \text{for } \xi \in I \quad \text{if } \xi \neq 0. \tag{2.6}$$

Under this situation, we can write f_0 as

$$f_0(\xi) = \xi^m h(\xi), \tag{2.7}$$

where $h(\xi)$ is holomorphic in the domain D and

$$h(\xi) < 0 \quad \text{for } \xi \in I. \tag{2.8}$$

Let us change the independent variable by

$$x = \phi(\xi) = \left[-\int_0^{\xi} \frac{m+1}{2} f_0(t)\, dt \right]^{1/(m+1)}. \tag{2.9}$$

Then (2.2) becomes

$$\varepsilon \frac{d^2 v}{dx^2} + F(x, \varepsilon) \frac{dv}{dx} + G(x, \varepsilon) v = 0, \tag{2.10}$$

where

$$\begin{aligned} F(\phi, \varepsilon) &= \{\phi'(\xi)\}^{-2}[\phi'(\xi) f(\xi, \varepsilon) + \varepsilon \phi''(\xi)], \\ G(\phi, \varepsilon) &= \{\phi'(\xi)\}^{-2} g(\xi, \varepsilon). \end{aligned} \tag{2.11}$$

Since

$$(m+1)\{\phi(\xi)\}^m\phi'(\xi) = -\frac{m+1}{2}\cdot f_0(\xi),$$

we have

$$F(x,\varepsilon) = -2x^m + \varepsilon k(x,\varepsilon), \tag{2.12}$$

and $k(x,\varepsilon)$ and $G(x,\varepsilon)$ are holomorphic in the domain

$$x \in D_0, \qquad |\varepsilon| < \rho_0, \tag{2.13}$$

where D_0 is a domain in the x-plane which contains the real interval

$$I_0 = \{x: -a \leqq \operatorname{Re} x \leqq b, \operatorname{Im} x = 0\}, \tag{2.14}$$

where

$$a = \left[-\int_0^{-\alpha} f_0(t)\,dt\right]^{1/(m+1)}, \qquad b = \left[-\int_0^{\beta} f_0(t)\,dt\right]^{1/(m+1)}.$$

Let r_2 be a sufficiently small positive number and r_1 be a positive number with $r_1 > b, a$, respectively. Set

$$D_1 = \{x = x_1 + ix_2: -r_2 < x_2 < r_2, -(r_1^2 - x_2^2)^{1/2} < x_1 < (r_1^2 - x_2^2)^{1/2}\}. \tag{2.15}$$

Then D_1 is a simply connected domain in the complex x-plane which contains the real interval I_0. We can choose r_1 and r_2 so that $D_1 \subset D_0$.

Another transformation,

$$v = w\cdot\exp\left[-\frac{1}{2\varepsilon}\int_0^x F(t,\varepsilon)\,dt\right], \tag{2.16}$$

takes (2.10) to

$$\varepsilon^2\frac{d^2w}{dx^2} - \left[\frac{1}{4}F(x,\varepsilon)^2 + \varepsilon\left[\frac{1}{2}\frac{\partial F}{\partial x}(x,\varepsilon) - G(x,\varepsilon)\right]\right]w = 0. \tag{2.17}$$

Note that

$$\frac{1}{4}F^2 + \varepsilon\left[\frac{1}{2}\frac{\partial F}{\partial x} - G\right] = x^{2m} + \varepsilon R(x,\varepsilon), \tag{2.18}$$

where $R(x,\varepsilon)$ is holomorphic in a domain (2.13).

3. Formal simplification and outer expansions. We consider the system

$$\varepsilon\frac{d}{dx}\begin{pmatrix} w \\ \varepsilon\dfrac{dw}{dx}\end{pmatrix} = \begin{pmatrix} 0 & 1 \\ x^{2m}+\varepsilon R(x,\varepsilon) & 0\end{pmatrix}\begin{pmatrix} w \\ \varepsilon\dfrac{dw}{dx}\end{pmatrix}, \tag{3.1}$$

which is equivalent to the single linear differential equation (2.17). It is well known among experts that there exists a two-by-two matrix

$$T(x,\varepsilon) = \sum_{h=0}^{\infty} T_h(x)\varepsilon^h \tag{3.2}$$

whose components are formal power series in ε such that

(i) The components of the two-by-two matrices $T_h(x)$ are holomorphic in the domain D_0.

(ii) $\det T_0(x) \equiv 1$ in the domain D_0.

(iii) The formal transformation

$$\begin{pmatrix} w \\ \varepsilon \dfrac{dw}{dx} \end{pmatrix} = T(x, \varepsilon) \begin{pmatrix} u \\ \varepsilon \dfrac{du}{dx} \end{pmatrix}, \tag{3.3}$$

reduces (3.1) to

$$\varepsilon \frac{d}{dx} \begin{pmatrix} u \\ \varepsilon \dfrac{du}{dx} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ x^{2m} + \varepsilon \sum_{r=0}^{2m-1} a_r(\varepsilon) x^r & 0 \end{pmatrix} \begin{pmatrix} u \\ \varepsilon \dfrac{du}{dz} \end{pmatrix} \tag{3.4}$$

and each $a_r(\varepsilon)$ is a formal power series in ε with constant coefficients. Furthermore, if we put

$$a_r(x) = \sum_{h=0}^{\infty} a_{r,h} \varepsilon^h, \tag{3.5}$$

we have

$$R(x, 0) - \sum_{r=0}^{2m-2} a_{r,0} x^r = O(x^{2m-1}) \tag{3.6}$$

in the neighborhood of $x = 0$ [1], [2].

Note that (3.4) is equivalent to the single linear differential equation

$$\varepsilon^2 \frac{d^2u}{dx^2} - \left[x^{2m} + \varepsilon \sum_{r=0}^{2m-2} a_r(\varepsilon) x^r \right] u = 0. \tag{3.7}$$

A formal power series in ε:

$$v(x, \varepsilon) = \sum_{n=0}^{\infty} v_n(x) \varepsilon^n \tag{3.8}$$

is called an *outer expansion*, associated with the differential equation (2.10), if (3.8) formally satisfies (2.10). The power series (3.8) is an outer expansion if and only if

$$-2x^m \frac{dv_0(x)}{dx} + G_0(x) v_0(x) = 0, \tag{3.9}$$

$$-2x^m \frac{dv_k(x)}{dx} + G_0(x) v_k(x) = L_k(x) - \frac{d^2 v_{k-1}(x)}{dx^2}, \qquad k = 1, 2, \cdots, \tag{3.10}$$

where $G_0(x) = G(x, 0)$ and $L_k(x)$ is linearly homogeneous in $v_0, v_1, \cdots, v_{k-1}$ and $(dv_0/dx), (dv_1/dx), \cdots, (dv_{k-1}/dx)$ with coefficients holomorphic in D_0.

DEFINITION 3.1. The differential equation (2.10) is said to satisfy the *Matkowsky condition* if there exists a nontrivial outer expansion (3.8) such that all the $v_k(x)$ are bounded on the interval I_0.

LEMMA 3.2. *The differential equation* (2.10) *satisfies the Matkowsky condition. Then the differential equation* (3.7) *satisfies*

(i) $a_{m-1,0} + m$ *is a nonpositive even integer*;

(ii) $a_{r,0} = 0$ $(r = 0, 1, \cdots, m-2)$.

Proof. The transformation

$$u = y \cdot \exp\left[-x^{m+1}/(m+1)\varepsilon\right] \tag{3.11}$$

changes (3.7) into

$$\varepsilon y''-2x^m y'-\left[\sum_{r=0}^{2m-2}\left[\sum_{j=0}^{\infty} a_{r,j}\varepsilon^j\right]x^r+mx^{m-1}\right]y=0. \tag{3.12}$$

If the differential equation (2.10) satisfies the Matkowsky condition, then, by manipulating the transformations (2.16) and (3.3) together with (3.11), we can show that the differential equation (3.12) satisfies the same condition. It follows from (3.12) that

$$2x^m v_0'(x)+\left[\sum_{r=0}^{2m-2} a_{r,0}x^r+mx^{m-1}\right]v_0(x)=0; \tag{3.13}$$

that is,

$$v_0(x)=c\cdot x^{-1/2a_{m-1,0}-m/2}\exp\left[-\sum_{r=0,r\neq m-1}^{2m-2}\frac{1}{2(r+1-m)}a_{r,0}x^{r+1-m}\right], \tag{3.14}$$

where c is a constant.

Hence, if $v_0(x)$ is to be holomorphic at $x=0$, it must hold that

(i) $-\frac{1}{2}(a_{m-1,0}+m)$ is a nonnegative integer;

(ii) $a_{r,0}=0$ $(r=0,1,\cdots,m-2)$.

This completes the proof of Lemma 3.2.

4. Uniform simplification. We will assume that the transformed differential equation (3.7) satisfies conditions (i) and (ii) of Lemma 3.2. Then (3.7) is called fully reducible by stretching and shearing (Wasow [10]). Let us change the independent variable by the stretching transformation

$$x=\mu^{-2}t, \qquad \varepsilon^{-1}=\mu^{2m+2}. \tag{4.1}$$

Then (3.7) becomes

$$\frac{d^2u}{dt^2}-\left[t^{2m}+\sum_{r=0}^{2m-2}\mu^{2m-2r-2}\cdot a_r(\varepsilon)t^r\right]u=0, \tag{4.2}$$

where

$$\mu^{2m-2r-2}\cdot a_r(\varepsilon)=\begin{cases}o(1), & r\neq m-1,\\ a_{m-1,0}+o(1), & r=m-1.\end{cases} \tag{4.3}$$

Let $b_1, b_2, \cdots, b_{2m}$ be complex parameters and b be a *real* constant, and consider the differential equation of the form

$$y''-[z^{2m}+b_1z^{2m-1}+\cdots+[b+b_{m+1}]z^{m-1}+\cdots+b_{2m}]y=0. \tag{4.4}$$

The differential equation (4.4) has a solution

$$y(z, b, b_j)=y(z, b; b_1, \cdots, b_{m+1}, \cdots, b_{2m}),$$

which is an entire function of $z, b, b_1, \cdots, b_{2m}$. Respectively, $y(z, b, b_j)$ and $y'(z, b, b_j)$ admit the asymptotic representations

$$\begin{aligned} y(z, b, b_j)&=z^{-\alpha_{m+1}(b,b_j)-m/2}\exp\left[\sum_{0\leqq k<m+1}\frac{-\alpha_k(b, b_j)}{m+1-k}z^{m+1-k}\right]\\ &\quad\cdot\{1+O(z^{-1/2})\},\\ y'(z, b, b_j)&=z^{-\alpha_{m+1}(b,b_j)+m/2}\exp\left[\sum_{0\leqq k<m+1}\frac{-\alpha_k(b, b_j)}{m+1-k}z^{m+1-k}\right]\\ &\quad\cdot\{-1+O(z^{-1/2})\}\end{aligned} \tag{4.5}$$

uniformly on each compact set in the $(b_1, \cdots, b_{2m})$-space, as z tend to infinity in any closed subsector of the open sector $|\arg z| < 3\pi/(2m+2)$. Here

$$\left[1+\sum_{j=1, j\neq m+1}^{2m} b_j \cdot z^{-j} + (b+b_{m+1})z^{-m-1}\right]^{1/2} = 1+\sum_{h=1}^{\infty} \alpha_h(b, b_j) z^{-h}. \tag{4.6}$$

The solution $y(z, b, b_j)$ is subdominant on $|\arg z| < \pi/(2m+2)$ and therefore it is uniquely determined. Furthermore, if we put

$$y(z, b, b_j) = \eta(z, b) + \sum_{j=1}^{2m} \eta_j(z, b) b_j + O(|b_1|^2 + \cdots + |b_{2m}|^2) \tag{4.7}$$

in the neighborhood of $(b_1, \cdots, b_{m+1}, \cdots, b_{2m}) = (0, \cdots, 0)$, we can obtain the following properties of $\eta(z, b)$.

LEMMA 4.1 (Ohkohchi [7]). *If z is positive real and the constant b is real, then $\eta(z, b)$ is a real-valued function. Furthermore, $\eta(z, b)$ satisfies the following conditions*:

(i) *$\eta(z, b)$ is an entire function of (z, b).*

(ii) *$\eta(z, b)$ and $\eta'(z, b)$ admit, respectively, the asymptotic representations*

$$\eta(z, b) = z^{-b/2-m/2} \exp\left[-\frac{1}{m+1} z^{m+1}\right][1+O(z^{-1/2})], \tag{4.8}$$

$$\eta'(z, b) = z^{-b/2+m/2} \exp\left[-\frac{1}{m+1} z^{m+1}\right][-1+O(z^{-1/2})] \tag{4.9}$$

uniformly on each compact set in the b-space, as z tends to infinity in any closed subsector of the open sector $|\arg z| < 3\pi/(2m+2)$.

(iii)

$$\eta(0, b) = 2^{(m+b)/(2m+2)}(m+1)^{-(m+b)/(2m+2)} \frac{\Gamma(1/(m+1))}{\Gamma((b+m+2)/(2m+2))}, \tag{4.10}$$

$$\eta'(0, b) = 2^{(b+m+2)/(2m+2)}(m+1)^{-1-(b+m)/(2m+2)} \frac{\Gamma(-1/(m+1))}{\Gamma((b+m)/(2m+2))}. \tag{4.11}$$

If we put

$$y_k(z, b, b_j) = y(\omega^{-k} z, (-1)^k b, G^k(b_j)), \tag{4.12}$$

where

$$\omega = \exp\left[\frac{\pi i}{m+1}\right] \quad \text{and} \quad G^k(b_j) = (\omega^{-k} b_1, \cdots, \omega^{-2mk} b_{2m}),$$

then, for each integer k, $y_k(z, b, b_j)$ is a solution of the differential equation (4.4) (see Sibuya [8]). Note that

$$y_k(z, b, b_j) = y_h(z, b, b_j) \quad \text{if } k \equiv h \pmod{2m+2}.$$

It holds from (4.5) that

$$y_k(z, b, b_j) = (\omega^{-k} z)^{(-1)^{k+1}\alpha_{m+1}(b, b_j)-(m/2)} \cdot [1+O(z^{-1/2})] \cdot \exp\left[(-1)^{k+1} \sum_{0 \leqq p < m+1} \frac{\alpha_p(b, b_j)}{m+1-p} z^{m+1-p}\right],$$

$$y_k'(z, b, b_j) = \omega^{-k\{(-1)^{k+1}\alpha_{m+1}(b, b_j)-(m/2)\}} z^{(m/2)+(-1)^{k+1}\alpha_{m+1}(b, b_j)} \cdot \exp\left[(-1)^{k+1} \sum_{0 \leqq p < m+1} \frac{\alpha_p(b, b_j)}{m+1-p} z^{m+1-p}\right]\{(-1)^{k+1} + O(z^{-1/2})\}.$$

Here we used the identities

$$\alpha_h((-1)^k b, G^k(b_j)) = \omega^{-hk} \cdot \alpha_h(b, b_j) \qquad (h = 1, 2, \cdots).$$

The solution $y_k(z, b, b_j) \to 0$ as $x \to \infty$ in the sector

$$\left| \arg x - \frac{k}{m+1} \pi \right| < \frac{\pi}{2m+2}. \tag{4.13}$$

Therefore, the solution $y_k(z, b, b_j)$ is called a subdominant solution in sector (4.13).

The two solutions $y_{k+1}(z, b, b_j)$ and $y_{k+2}(z, b, b_j)$ of the differential equation (4.4) are linearly independent (see Lemma 4.2 below). Therefore, $y_k(z, b, b_j)$ is a linear combination of $y_{k+1}(z, b, b_j)$ and $y_{k+2}(z, b, b_j)$.

Set

$$y_k(z, b, b_j) = C_k(b, b_j) y_{k+1}(z, b, b_j) + \tilde{C}_k(b, b_j) y_{k+2}(z, b, b_j). \tag{4.14}$$

Relation (4.14) is a connection formula for $y_k(z, b, b_j)$. The coefficients $C_k(b, b_j)$ and $\tilde{C}_k(b, b_j)$ are the Stokes multipliers for $y_k(z, b, b_j)$ with respect to $y_{k+1}(z, b, b_j)$ and $y_{k+2}(z, b, b_j)$. We will study the Stokes multipliers $C_k(b, b_j)$ and $\tilde{C}_k(b, b_j)$ as functions of $b_1, \cdots, b_j, \cdots, b_{2m}$.

Let us put

$$C_0(b, b_j) = C(b, b_j), \qquad \tilde{C}_0(b, b_j) = \tilde{C}(b, b_j), \tag{4.15}$$

$$C_1(b, b_j) = C^*(b, b_j), \qquad \tilde{C}_1(b, b_j) = \tilde{C}^*(b, b_j). \tag{4.16}$$

Then we have

$$y_0(z, b, b_j) = C(b, b_j) y_1(z, b, b_j) + \tilde{C}(b, b_j) y_2(z, b, b_j), \tag{4.17}$$

$$y_1(z, b, b_j) = C^*(b, b_j) y_2(z, b, b_j) + \tilde{C}^*(b, b_j) y_3(z, b, b_j). \tag{4.18}$$

From (4.12), (4.17), and (4.18) we easily derive

$$\begin{aligned} y_{2k}(z, b, b_j) &= C((-1)^{2k} b, G^{2k}(b_j)) y_{2k+1}(z, b, b_j) \\ &\quad + \tilde{C}((-1)^{2k} b, G^{2k}(b_j)) y_{2k+2}(z, b, b_j), \end{aligned} \tag{4.19}$$

$$\begin{aligned} y_{2k+1}(z, b, b_j) &= C^*((-1)^{2k} b, G^{2k}(b_j)) y_{2k+2}(z, b, b_j) \\ &\quad + \tilde{C}^*((-1)^{2k} b, G^{2k}(b_j)) y_{2k+3}(z, b, b_j), \end{aligned} \tag{4.20}$$

$$(k = 0, 1, \cdots, m).$$

From (4.12), (4.19), and (4.20) we obtain

$$\begin{aligned} C_{2k}(b, b_j) &= C((-1)^{2k} b, G^{2k}(b_j)), \\ \tilde{C}_{2k}(b, b_j) &= \tilde{C}((-1)^{2k} b, G^{2k}(b_j)), \\ C_{2k+1}(b, b_j) &= C^*((-1)^{2k} b, G^{2k}(b_j)), \\ \tilde{C}_{2k+1}(b, b_j) &= \tilde{C}^*((-1)^{2k} b, G^{2k}(b_j)), \end{aligned} \tag{4.21}$$

$$(k = 0, 1, 2, \cdots, m).$$

Next, let us put

$$W_{k,h}(b, b_j) = \begin{vmatrix} y_k(z, b, b_j) & y_h(z, b, b_j) \\ y_k'(z, b, b_j) & y_h'(z, b, b_j) \end{vmatrix}. \tag{4.22}$$

The determinant is known as the *Wronskian* of the functions y_k and y_h. It follows from (4.4) that the right-hand member of (4.22) is independent of x. Then we have the following lemmas.

LEMMA 4.2 (Hsieh and Sibuya [3]).

$$W_{k,k+1}(b, b_j) = 2 \cdot \omega^{-\alpha_{m+1}((-1)^k b, G^k(b))+m/2-k}. \tag{4.23}$$

LEMMA 4.3.

$$\left.\frac{\partial \alpha_{m+1}(b, b_j)}{\partial b_j}\right|_{b_2=\cdots=b_{2m}=0} = \begin{cases} 0 & (j \neq m+1), \\ 1/2 & (j = m+1). \end{cases} \tag{4.24}$$

Proof. Differentiating (4.6) with respect to b_j and putting $b_1 = \cdots = b_{2m} = 0$, we derive

$$\frac{1}{2}[1 + b \cdot z^{-m-1}]^{-1/2} \cdot z^{-j} = \sum_{h=1}^{\infty} A_j^h \cdot z^{-h}, \tag{4.25}$$

where

$$A_j^h = \left.\frac{\partial \alpha_h(b)}{\partial b_j}\right|_{b_1=\cdots=b_{2m}=0}.$$

It follows from (4.25) that

$$A_j^1 = \cdots = A_j^{j-1} = 0, \quad A_j^j = \tfrac{1}{2}, \quad A_j^{j+1} = \cdots = A_j^{j+m} = 0.$$

Thus we get (4.24).

Using these results, we can obtain the properties of the Stokes multipliers $C(b, b_j)$, $C^*(b, b_j)$, $\tilde{C}(b, b_j)$, and $\tilde{C}^*(b, b_j)$ as follows.

LEMMA 4.4.

$$C(b, 0) = 0 \qquad \text{for } b + m = -(2m+2)k, -(2m+2)k-2, \quad k = 0, 1, 2, \cdots,$$

$$C^*(b, 0) \neq 0 \qquad \text{for } b + m = -2k, \quad k = 0, 1, 2, \cdots. \tag{4.26}$$

LEMMA 4.5.

$$\tilde{C}(b, b_j) = -\omega^{1-2\cdot\alpha_{m+1}(b,b_j)},$$
$$\tilde{C}^*(b, b_j) = -\omega^{1+2\cdot\alpha_{m+1}(b,b_j)}.$$

LEMMA 4.6.

$$\det \begin{vmatrix} \dfrac{\partial C(b, b_j)}{\partial b_p} & \dfrac{\partial C(b, b_j)}{\partial b_{p+m+1}} \\ \dfrac{\partial C^*(b, b_j)}{\partial b_p} & \dfrac{\partial C^*(b, b_j)}{\partial b_{p+m+1}} \end{vmatrix} \neq 0 \tag{4.27}$$

for

$$b = -m - 2k(m+1), -m-2-2k(m+1) \qquad (k = 0, 1, 2, \cdots),$$
$$(b_2, b_3, \cdots, b_{2m}) = (0, 0, \cdots, 0), \tag{4.28}$$
$$p = 1, 2, \cdots, m-1.$$

LEMMA 4.7.

$$\left.\frac{\partial C(b, b_j)}{\partial b_p}\right|_{\substack{b_1=\cdots=b_{2m}=0, \\ b=-m-2k(m+1),-m-2-2k(m+1)}} \neq 0 \qquad (p \neq m).$$

Using Lemmas 4.5–4.7, we can obtain the following lemma. The proof is similar to that of Sibuya [8, Case iv].

LEMMA 4.8. *There exist two positive numbers* ρ_1 *and* ρ_2, *functions* $\delta_1(\varepsilon)$, $\delta_2(\varepsilon), \cdots, \delta_{2m}(\varepsilon)$, *and a two-by-two matrix* $P(x, \varepsilon)$ *such that*

(i) $\delta_j(\varepsilon)$ $(j=1, 2, \cdots, 2m)$ *are holomorphic in the sector*

$$S=\{\varepsilon; |\arg \varepsilon|<\rho_1, 0<|\varepsilon|<\rho_2\}; \tag{4.29}$$

(ii) $\delta_j(\varepsilon)$ *are asymptotically zero as* $\varepsilon \to 0$ *in the sector* S,

$$|\delta_j(\varepsilon)| \leqq K_N|\varepsilon|^N, \qquad N=0, 1, 2, \cdots \quad \text{in } S \tag{4.30}$$

for some positive numbers K_N.

(iii) *Entries of* $P(x, \varepsilon)$ *and* $P^{-1}(x, \varepsilon)$ *are holomorphic in the domain*

$$x \in D_2=\{x; |x|<r\}, \qquad \varepsilon \in S. \tag{4.31}$$

(iv) $P(x, \varepsilon)$ *admits the formal transformation matrix* $T(x, \varepsilon)$ (3.2) *as an asymptotic expansion as* $\varepsilon \to 0$ *in* S *which is valid uniformly in* D_2.

(v) *The transformation*

$$W=P(x, \varepsilon)V \tag{4.32}$$

takes (3.1) *into*

$$\varepsilon \frac{dV}{dx}=\begin{pmatrix} 0 & 1 \\ x^{2m}+\varepsilon \cdot \sum_{p=1}^{2m}\{a_p(\varepsilon)+\delta_p(\varepsilon)\}x^{2m-p} & 0 \end{pmatrix} \cdot V \tag{4.33}$$

in the domain (4.31).

Using this result and manipulating with rotations of the disk D_2, we can prove the following lemma.

LEMMA 4.9. *There exist sectors*

$$S_j=\{\varepsilon: \alpha_j<\arg \varepsilon<\beta_j, 0<|\varepsilon|<\rho_3\}, \qquad j=1, 2, \cdots, k, \tag{4.34-j}$$

where ρ_3 *is a positive number and the* α*'s and* β*'s are real numbers, functions* $\delta_p^j(\varepsilon)$ $(j=1, 2, \cdots, k; p=1, 2, \cdots, 2m)$, *and two-by-two matrices* $P_1(x, \varepsilon)$, $P_2(x, \varepsilon)$, $\cdots, P_k(x, \varepsilon)$ *such that*

(i) $S_1 \cup S_2 \cup \cdots \cup S_k=\{\varepsilon; 0<|\varepsilon|<\rho_3\}$.

(ii) $\delta_p^j(\varepsilon)$ $(p=1, 2, \cdots, 2m; j=1, 2, \cdots, k)$ *is holomorphic in the sector* S_j.

(iii) $\delta_p^j(\varepsilon)$ $(p=1, 2, \cdots, 2m; j=1, 2, \cdots, k)$ *is asymptotically zero as* $\varepsilon \to 0$ *in the sector* S_j.

(iv) *Entries of* $P_j(x, \varepsilon)$ *and* $P_j^{-1}(x, \varepsilon)$ *are holomorphic in the domain*

$$x \in D_2, \qquad \varepsilon \in S_j. \tag{4.35-j}$$

(v) $P_j(x, \varepsilon)$ *admits the formal transformation matrix* $T(x, \varepsilon)$ *as an asymptotic expansion as* $\varepsilon \to 0$ *in the sector* S_j *which is valid uniformly in the domain* D_2.

(vi) *The transformation*

$$W=P_j(x, \varepsilon)V_j \tag{4.36}$$

takes (3.1) *into*

$$\varepsilon \frac{dV_j}{dx}=\begin{pmatrix} 0 & 1 \\ x^{2m}+\varepsilon \cdot \sum_{p=1}^{2m}[a_p(\varepsilon)+\delta_p^j(\varepsilon)]x^{2m-p} & 0 \end{pmatrix} V_j \tag{4.37-j}$$

in the domain (4.35-j).

Remark 4.10. In Lemmas 4.8 and 4.9,

$$\delta_m(\varepsilon), \qquad \delta_m^j(\varepsilon)=0.$$

5. An estimate for $\delta(\varepsilon)$. In the previous section we have obtained that $\delta_p^j(\varepsilon)$ is asymptotically zero as $\varepsilon \to 0$ in the sector S_j. In this section we will derive an estimate

$$|\delta_p^j(\varepsilon)| \leqq H_{j,p} \exp\left[-\frac{r^{m+1}}{|\varepsilon|}\right] \quad \text{for } \varepsilon \in S_j \quad \text{and} \quad p = 1, 2, \cdots, 2m, \tag{5.1}$$

in the domain D_2(disk), where $H_{j,p}$ is a positive number. To do this we need the following theorem, which is easily obtained from Theorem 1.2.

THEOREM 5.1. *Let*

$$S_j = \{\varepsilon : \alpha_j < \arg \varepsilon < \beta_j, 0 < |\varepsilon| < \rho_3\}, \qquad j = 1, 2, \cdots, k, \tag{5.2}$$

be sectors in the complex ε-plane, where ρ_3 is a positive number and the α's and β's are real numbers. Let $\delta_p^j(\varepsilon)$ $(j = 1, 2, \cdots, k;\ p = 1, 2, \cdots, 2m)$ be functions of ε. Assume that

(i) $S_1 \cup S_2 \cup \cdots \cup S_k = \{\varepsilon : 0 < |\varepsilon| < \rho_3\}$;

(ii) *$\delta_p^j(\varepsilon)$ are holomorphic in the sector S_j.*

(iii) *$\delta_p^j(\varepsilon)$ are asymptotically zero as $\varepsilon \to 0$ in the sector S_j;*

$$|\delta_p^j(\varepsilon)| \leqq K_{N,p} |\varepsilon|^N, \quad N = 0, 1, \cdots \quad \text{in } S_j \tag{5.3}$$

for some positive numbers $K_{N,p}$;

(iv) *If $S_j \cap S_h \neq \varnothing$, we have*

$$|\delta_p^j(\varepsilon) - \delta_p^h(\varepsilon)| \leqq c_0 \exp\left[-c_1/|\varepsilon|^\lambda\right] \quad \text{in } S_j \cap S_h \tag{5.4}$$

for some positive numbers c_0, c_1, and λ. Then, there exist positive numbers H_p $(p = 1, 2, \cdots, 2m)$ such that

$$|\delta_p^j(\varepsilon)| \leqq H_p \cdot \exp\left[-c_1/|\varepsilon|^\lambda\right] \quad \text{in } S_j, \qquad j = 1, 2, \cdots, k. \tag{5.5}$$

To derive (5.1), we have only to prove that, if $S_j \cap S_l \neq \varnothing$, we have for each p $(p = 1, 2, \cdots, 2m)$

$$|\delta_p^j(\varepsilon) - \delta_p^l(\varepsilon)| \leqq M_{j,l} \cdot \exp\left[-r^{m+1}/|\varepsilon|\right] \quad \text{for } \varepsilon \in S_j \cap S_l, \tag{5.6}$$

where $M_{j,l}$ is a positive number. To derive an estimate (5.6), we need some preparations.

Roughly speaking, we shall derive in this section that the difference between two Stokes multipliers of the differential equations (4.37-j) and (4.37-l) is exponentially small. Then, using the implicit function theorem, we show that the difference $\delta^j(\varepsilon) - \delta^l(\varepsilon)$ is exponentially small.

Let us consider the differential equation (4.4). This equation admits solutions $y_k(z, b, b_j)$ $(k = 0, 1, \cdots, 2m+1)$. Set

$$\Psi_k(z, b, b_j) = \begin{pmatrix} y_k(z, b, b_j) & y_{k+1}(z, b, b_j) \\ y_k'(z, b, b_j) & y_{k+1}'(z, b, b_j) \end{pmatrix}, \tag{5.7-k}$$

where $'$ denotes $\partial/\partial z$. These matrices (5.7-k) are fundamental matrix solutions of (4.4) and

$$\Psi_k(z, b, b_j) = \Psi_h(z, b, b_j) \quad \text{if } k \equiv h \bmod 2m+2.$$

Set

$$\Gamma_k(b, b_j) = \begin{pmatrix} C_k(b, b_j) & 1 \\ \tilde{C}_k(b, b_j) & 0 \end{pmatrix} \qquad (k = 0, 1, \cdots, 2m+1). \tag{5.8}$$

Then, we have from (4.14) that

$$\Psi_k(z, b, b_j) = \Psi_{k+1}(z, b, b_j)\Gamma_k(b, b_j). \tag{5.9}$$

Fix l and j so that $S_l \cap S_j \neq \emptyset$. Choose a branch of $\varepsilon^{1/(m+1)}$ in the sector $S_l \cap S_j$. Set

$$\Lambda(\varepsilon) = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon^{1/(m+1)} \end{pmatrix}. \tag{5.10}$$

Now, we consider (4.37-j) and (4.37-l). If we put

$$\begin{aligned} \Phi_{l,h}(x, \varepsilon) &= \Lambda(\varepsilon)\Psi_h(x\varepsilon^{-1/(m+1)}, a, a_1 + \delta_1^l, \cdots, a_{2m} + \delta_{2m}^l) \\ &= \Lambda(\varepsilon)\Psi_h(x\varepsilon^{-1/(m+1)}, a, a_p + \delta_p^l), \qquad \text{(5.11-}h\text{)} \\ \Phi_{j,h}(x, \varepsilon) &= \Lambda(\varepsilon)\Psi_h(x\varepsilon^{-1/(m+1)}, a, a_p + \delta_p^j) \qquad (h = 0, 1, 2, \cdots, 2m+1), \end{aligned}$$

then $\Phi_{l,h}(x, \varepsilon)$ and $\Phi_{j,h}(x, \varepsilon)$ are fundamental matrix solutions of (4.37-l) and (4.37-j), respectively. Furthermore we have

$$\Phi_{l,h}(x, \varepsilon) = \Phi_{l,h+1}(x, \varepsilon)\Gamma_h(a, a_p + \delta_p^l), \tag{5.12}$$

$$\Phi_{j,h}(x, \varepsilon) = \Phi_{j,h+1}(x, \varepsilon)\Gamma_h(a, a_p + \delta_p^j) \qquad (h = 0, 1, 2, \cdots, 2m+1). \tag{5.13}$$

Set

$$J = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{5.14}$$

$$\begin{aligned} Q_{l,h}(x, \varepsilon) &= \Phi_{l,h}(x, \varepsilon) \exp\{(-1)^h E(z, a_p + \delta_p^l)J\} \\ Q_{j,h}(x, \varepsilon) &= \Phi_{j,h}(x, \varepsilon) \exp\{(-1)^h E(z, a_p + \delta_p^j)J\} \qquad \text{(5.15)} \\ &\qquad\qquad (h = 0, 1, \cdots, 2m+1), \end{aligned}$$

where

$$\begin{aligned} z &= x\varepsilon^{-1/(m+1)}, \\ E(z, a_p + \delta_p^j) &= \sum_{0 \leqq k < m+1} \frac{\alpha_k(a, a_p + \delta_p^j)}{m+1-k} z^{m+1-k}. \end{aligned} \tag{5.16}$$

It is known that, if (x, ε) is in a domain

$$\left| \arg x\varepsilon^{-1/(m+1)} - \frac{h}{m+1}\pi \right| \leqq \frac{3\pi}{2m+2} - v, \qquad x \in D_2, \quad \varepsilon \in S_l \cap S_j, \tag{5.17-}h$$

where v is a small positive number, we have

$$\|Q_{j,h}(x, \varepsilon)\| \leqq H \cdot |\varepsilon|^q, \qquad \|Q_{j,h}(x, \varepsilon)^{-1}\| \leqq H \cdot |\varepsilon|^q, \tag{5.18}$$

where H is a positive number depending on v, q is a real number, and $\|\cdot\|$ denotes a usual norm of matrices. Furthermore, the matrix

$$Q_{j,h}(x, \varepsilon) - Q_{l,h}(x, \varepsilon) \tag{5.19}$$

is asymptotically zero as $\varepsilon \to 0$ in the sector $S_l \cap S_j$ uniformly in the domain (5.17-h) (see Sibuya [8]).

Let $P_l(x, \varepsilon)$ and $P_j(x, \varepsilon)$ be the matrices given in Lemma 4.9. Then $P_l(x, \varepsilon)\Phi_{l,h}(x, \varepsilon)$ and $P_j(x, \varepsilon)\Phi_{j,h}(x, \varepsilon)$ are two fundamental matrix solutions of (3.1) in the domain

$$x \in D_2, \qquad \varepsilon \in S_l \cap S_j. \tag{5.20}$$

Therefore, there exists a two-by-two matrix $L_h(\varepsilon)$ such that

$$P_l(x, \varepsilon)\Phi_{l,h}(x, \varepsilon) = P_j(x, \varepsilon)\Phi_{j,h}(x, \varepsilon)L_h(\varepsilon). \tag{5.21}$$

Note that $L_h(\varepsilon)$ does not depend on x. It follows from (5.21) that

$$\begin{aligned}(5.22)\qquad &\exp\{-E(z, a_p+\delta_p^j)\cdot J\}\cdot L_h(\varepsilon)\cdot\exp\{E(z, a_p+\delta_p^l)\cdot J\}\\ &\quad = Q_{j,h}(x,\varepsilon)^{-1}P_j(x,\varepsilon)^{-1}P_l(x,\varepsilon)Q_{l,h}(x,\varepsilon).\end{aligned}$$

Hence, the matrix

$$(5.23)\qquad \exp\{-E(z, a_p+\delta_p^j)J\}\cdot L_h(\varepsilon)\cdot\exp\{E(z, a_p+\delta_p^l)J\}-I_2$$

is asymptotically zero as $\varepsilon\to 0$ in the sector $S_l\cap S_j$ uniformly in the domain (5.17-h), where I_2 is the two-by-two identity matrix.

LEMMA 5.2. *Let*

$$(5.24)\qquad L_h(\varepsilon)=\begin{pmatrix} d_{11}^h(\varepsilon) & d_{12}^h(\varepsilon)\\ d_{21}^h(\varepsilon) & d_{22}^h(\varepsilon)\end{pmatrix}.$$

Then

$$(5.25)\qquad d_{11}^h(\varepsilon)-1,\qquad d_{22}^h(\varepsilon)-1\simeq 0\quad \text{as } \varepsilon\to 0\quad \text{in } S_l\cap S_j$$

$$(5.26)\qquad |d_{12}^h(\varepsilon)|\leqq c\cdot\exp\left\{-\frac{2r^{m+1}}{(m+1)|\varepsilon|}\right\},\qquad |d_{21}^h(\varepsilon)|\leqq c\cdot\exp\left\{-\frac{2r^{m+1}}{(m+1)|\varepsilon|}\right\}$$

for $\varepsilon\in S_l\cap S_j$, where c is a positive constant.

Proof. Using (5.24), we get from (5.23) that

$$\begin{aligned}&\begin{pmatrix}\exp[-E(z, a_p+\delta_p^j)] & 0\\ 0 & \exp[E(z, a_p+\delta_p^j)]\end{pmatrix}\begin{pmatrix} d_{11}^h(\varepsilon) & d_{12}^h(\varepsilon)\\ d_{21}^h(\varepsilon) & d_{22}^h(\varepsilon)\end{pmatrix}\\ &\quad\cdot\begin{pmatrix}\exp[E(z, a_p+\delta_p^l)] & 0\\ 0 & \exp[-E(z, a_p+\delta_p^l)]\end{pmatrix}-\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\\ &=\begin{pmatrix} d_{11}^h(\varepsilon)\exp\{-E(z, a_p+\delta_p^j)+E(z, a_p+\delta_p^l)\}-1 & d_{12}^h(\varepsilon)\exp\{-E(z, a_p+\delta_p^j)-E(z, a_p+\delta_p^l)\}\\ d_{21}^h(\varepsilon)\exp\{E(z, a_p+\delta_p^j)+E(z, a_p+\delta_p^l)\} & d_{22}^h(\varepsilon)\exp\{E(z, a_p+\delta_p^j)-E(z, a_p+\delta_p^l)\}-1\end{pmatrix}\\ &\simeq 0\end{aligned}$$

as $\varepsilon\to 0$ in $S_l\cap S_j$.

On the other hand,

$$\begin{aligned}E(z, a_p+\delta_p^j)&=E(x\varepsilon^{-1/(m+1)}.\ a_p+\delta_p^j)\\ &=\frac{1}{(m+1)\varepsilon}x^{m+1}+\sum_{k=1}^{m}\frac{1}{m+1-k}\alpha_k(a, a_p+\delta_p^j)\frac{x^{m+1-k}}{\varepsilon^{(m+1-k)/(m+1)}}\end{aligned}$$

and, since $\delta_p^l(\varepsilon)$ and $\delta_p^j(\varepsilon)$ $(p=1, 2, \cdots, 2m)$ are asymptotically zero as $\varepsilon\to 0$, it holds that

$$\alpha_k(a, a_p+\delta_p^l)-\alpha_k(a, a_p+\delta_p^j)\simeq 0\qquad(\varepsilon\to 0).$$

Note that there are values of $x\in D_2$ and of $\varepsilon\in S_l\cap S_j$ for which (5.17-h) is true and $z^{m+1}=\{x\varepsilon^{-1/(m+1)}\}^{m+1}$ is positive, as well as other such values for which z^{m+1} is negative. Choosing such values with $|x|=r$, we can obtain Lemma 5.1.

DEFINITION 5.3. To shorten the expressions for the calculation that follows, the notation

$$a(\varepsilon) \approx b(\varepsilon) \quad \text{in the sector } S$$

will be introduced to indicate that

$$[a(\varepsilon) - b(\varepsilon)] \exp [2r^{m+1}/(m+1)|\varepsilon|]$$

remains bounded, as $\varepsilon \to 0$ in the sector S (see Wasow [10]). Furthermore, the notation

$$a(\varepsilon) :\approx b(\varepsilon) \quad \text{in the sector } S$$

will be interpreted as the sense that, for $0 < r' < r$,

$$\{a(\varepsilon) - b(\varepsilon)\} \exp \{2r'^{m+1}/(m+1)|\varepsilon|\} \to 0,$$

as $\varepsilon \to 0$ in the sector S.

We easily obtain from (5.12), (5.13), and (5.21) that

$$L_{h+1}(\varepsilon)\Gamma_h(a, a_p + \delta_p^l) = \Gamma_h(a, a_p + \delta_p^j) L_h(\varepsilon). \tag{5.27}$$

That is

$$\begin{aligned} &\begin{pmatrix} d_{11}^{h+1}(\varepsilon) & d_{12}^{h+1}(\varepsilon) \\ d_{21}^{h+1}(\varepsilon) & d_{22}^{h+1}(\varepsilon) \end{pmatrix} \begin{pmatrix} C_h(a, a_p + \delta_p^l) & 1 \\ \tilde{C}_h(a, a_p + \delta_p^l) & 0 \end{pmatrix} \\ &\quad = \begin{pmatrix} C_h(a, a_p + \delta_p^j) & 1 \\ \tilde{C}_h(a, a_p + \delta_p^j) & 0 \end{pmatrix} \begin{pmatrix} d_{11}^h(\varepsilon) & d_{12}^h(\varepsilon) \\ d_{21}^h(\varepsilon) & d_{22}^h(\varepsilon) \end{pmatrix}. \end{aligned} \tag{5.28}$$

This relation is important in the following analysis.

LEMMA 5.4.

$$\tilde{C}_h(a, a_p + \delta_p^l) - \tilde{C}_h(a, a_p + \delta_p^j) \approx 0 \quad \textit{in the sector } S_j \cap S_l. \tag{5.29}$$

Proof. The determinants of both sides of (5.28) imply that

$$\begin{aligned} &\{d_{11}^{h+1}(\varepsilon) d_{22}^{h+1}(\varepsilon) - d_{12}^{h+1}(\varepsilon) d_{21}^{h+1}(\varepsilon)\} \tilde{C}_h(a, a_p + \delta_p^l) \\ &\quad = \tilde{C}_h(a, a_p + \delta_p^j) \{d_{11}^h(\varepsilon) d_{22}^h(\varepsilon) - d_{12}^h(\varepsilon) d_{21}^h(\varepsilon)\}. \end{aligned}$$

That is

$$d_{11}^{h+1}(\varepsilon) d_{22}^{h+1}(\varepsilon) \tilde{C}_h(a, a_p + \delta_p^l) \approx \tilde{C}_h(a, a_p + \delta_p^j) d_{11}^h(\varepsilon) d_{22}^h(\varepsilon). \tag{5.30}$$

Here we used (5.26).

On the other hand, we see from the (1, 2)-element of the matrix relation (5.28) that

$$d_{11}^{h+1}(\varepsilon) = C_h(a, a_p + \delta_p^j) d_{12}^h(\varepsilon) + d_{22}^h(\varepsilon).$$

That is

$$d_{11}^{h+1}(\varepsilon) \approx d_{22}^h(\varepsilon). \tag{5.31}$$

It follows from (5.30) and (5.31) that

$$d_{22}^{h+1}(\varepsilon) \tilde{C}_h(a, a_p + \delta_p^l) \approx \tilde{C}_h(a, a_p + \delta_p^j) d_{11}^h(\varepsilon).$$

Furthermore, using (5.31) ($d_{22}^{h+1}(\varepsilon) \approx d_{11}^{h+2}(\varepsilon)$), we get

$$d_{11}^{h+2}(\varepsilon) \tilde{C}_h(a, a_p + \delta_p^l) \approx d_{11}^h(\varepsilon) \tilde{C}_h(a, a_p + \delta_p^j).$$

That is

$$\begin{aligned} &d_{11}^{h+2}(\varepsilon) \{\tilde{C}_h(a, a_p + \delta_p^l) - \tilde{C}_h(a, a_p + \delta_p^j)\} \\ &\quad - \tilde{C}_h(a, a_p + \delta_p^j) \{d_{11}^h(\varepsilon) - d_{11}^{h+2}(\varepsilon)\} \approx 0. \end{aligned}$$

Hence, we can obtain

$$d_{11}^{h}(\varepsilon)-d_{11}^{h+2}(\varepsilon)\approx\frac{d_{11}^{h+2}(\varepsilon)}{\tilde{C}_h(a,a_p+\delta_p^j)}\{\tilde{C}_h(a,a_p+\delta_p^l)-\tilde{C}_h(a,a_p+\delta+\delta_p^j)\}.$$

Similarly, we get

$$\begin{aligned}&d_{11}^{h+2k}(\varepsilon)-d_{11}^{h+2k+2}(\varepsilon)\\&\quad\approx\frac{d_{11}^{h+2k+2}(\varepsilon)}{\tilde{C}_{h+2k}(a,a_p+\delta_p^j)}\cdot\{\tilde{C}_{h+2k}(a,a_p+\delta_p^l)-\tilde{C}_{h+2k}(a,a_p+\delta_p^j)\},\end{aligned}\tag{5.32}$$

$$(k=0,1,\cdots,m).$$

Note that

$$d_{11}^{h}(\varepsilon)=d_{11}^{h+2m+2}(\varepsilon),$$
$$\tilde{C}_{h+2m}(a,a_p+\delta_p^j)=\tilde{C}_{h+2m-2}(a,a_p+\delta_p^j)=\cdots=\tilde{C}_h(a,a_p+\delta_p^j)$$

(cf. Sibuya [9, form. (11.47)]). We can obtain from (5.32) that

$$\{\tilde{C}_h(a,a_p+\delta_p^l)-\tilde{C}_h(a,a_p+\delta_p^j)\}\frac{1}{\tilde{C}_h(a,a_p+\delta_p^j)}\cdot\sum_{s=1}^{m+1}d_{11}^{h+2s}(\varepsilon)\approx 0.$$

Since

$$\tilde{C}_h(a,a_p+\delta_p^j)\neq 0\quad\text{and}\quad d_{11}^{h+2s}(\varepsilon)\simeq 1,$$

we can get (5.29).

LEMMA 5.5.

$$C_1(a,a_p+\delta_p^l)\cdot C_{2h+2}(a,a_p+\delta_p^l):\approx C_1(a,a_p+\delta_p^j)\cdot C_{2h+2}(a,a_p+\delta_p^j)\tag{5.33}$$

$$(h=0,1,\cdots,m-2).$$

Proof. It follows from (5.27) that

$$\begin{aligned}&L_{h+2}(\varepsilon)\Gamma_{h+1}(a,a_p+\delta_p^l)\Gamma_h(a,a_p+\delta_p^l)\\&\quad=\Gamma_{h+1}(a,a_p+\delta_p^j)L_{h+1}(\varepsilon)\Gamma_h(a,a_p+\delta_p^j)\\&\quad=\Gamma_{h+1}(a,a_p+\delta_p^j)\Gamma_h(a,a_p+\delta_p^j)L_h(\varepsilon).\end{aligned}\tag{5.34}$$

That is

$$\begin{aligned}&\begin{pmatrix}d_{11}^{h+2}(\varepsilon)&d_{12}^{h+2}(\varepsilon)\\d_{21}^{h+2}(\varepsilon)&d_{22}^{h+2}(\varepsilon)\end{pmatrix}\begin{pmatrix}C_{h+1}(a,a_p+\delta_p^l)&1\\\tilde{C}_{h+1}(a,a_p+\delta_p^l)&0\end{pmatrix}\begin{pmatrix}C_h(a,a_p+\delta_p^1)&1\\\tilde{C}_h(a,a_p+\delta_p^l)&0\end{pmatrix}\\&\quad=\begin{pmatrix}C_{h+1}(a,a_p+\delta_p^j)&1\\\tilde{C}_{h+1}(a,a_p+\delta_p^j)&0\end{pmatrix}\begin{pmatrix}C_h(a,a_p+\delta_p^j)&1\\\tilde{C}_h(a,a_p+\delta_p^j)&0\end{pmatrix}\begin{pmatrix}d^{h_{11}}(\varepsilon)&d_{12}^{h}(\varepsilon)\\d_{21}^{h}(\varepsilon)&d_{22}^{h}(\varepsilon)\end{pmatrix}.\end{aligned}\tag{5.35}$$

We see from the (1, 1)-element of the matrix relation (5.35) that

$$\begin{aligned}&d_{11}^{h+2}(\varepsilon)\{C_{h+1}(a,a_p+\delta_p^l)C_h(a,a_p+\delta_p^l)+\tilde{C}_h(a,a_p+\delta_p^l)\}\\&\quad+d_{12}^{h+2}(\varepsilon)\tilde{C}_{h+1}(a,a_p+\delta_p^l)C_h(a,a_p+\delta_p^l)\\&=d_{11}^{h}(\varepsilon)\{C_{h+1}(a,a_p+\delta_p^j)C_h(a,a_p+\delta_p^j)+\tilde{C}_h(a,a_p+\delta_p^j)\}\\&\quad+d_{21}^{h}(\varepsilon)C_{h+1}(a,a_p+\delta_p^j).\end{aligned}\tag{5.36}$$

Note that

$$d_{11}^{h+2}(\varepsilon)\approx d_{11}^{h}(\varepsilon)\quad\text{in } S_l\cap S_j,$$

which follows from (5.32). Using (5.29) and (5.26), we can obtain from (5.36) that

$$C_{h+1}(a,a_p+\delta_p^l)C_h(a,a_p+\delta_p^l)\approx C_{h+1}(a,a_p+\delta_p^j)C_h(a,a_p+\delta_p^j)\quad\text{in } S_l\cap S_j.\tag{5.37}$$

So we have proved $(5.33)_{h=0}$.

Putting $h=1, 3$ in (5.37), we can get

$$C_1(a, a_p+\delta_p^l)C_2(a, a_p+\delta_p^l)C_3(a, a_p+\delta_p^l)C_4(a, a_p+\delta_p^l)$$
$$\approx C_1(a, a_p+\delta_p^j)C_2(a, a_p+\delta_p^j)C_3(a, a_p+\delta_p^j)C_4(a, a_p+\delta_p^j) \quad \text{in } S_j \cap S_l.$$

Hence we get

$$\{C_1(a, a_p+\delta_p^l)C_4(a, a_p+\delta_p^l)-C_1(a, a_p+\delta_p^j)C_4(a, a_p+\delta_p^j)\}$$
$$\cdot C_2(a, a_p+\delta_p^l)C_3(a, a_p+\delta_p^l)$$
$$\approx\{C_2(a, a_p+\delta_p^j)C_3(a, a_p+\delta_p^j)-C_2(a, a_p+\delta_p^l)C_3(a, a_p+\delta_p^l)\} \tag{5.38}$$
$$\cdot C_1(a, a_p+\delta_p^j)C_4(a, a_p+\delta_p^j)$$
$$\text{in the sector } S_l \cap S_j.$$

It holds from Lemmas 4.4 and 4.7 that

$$C_{2k}(a, a_p+\delta_p^l)|_{\varepsilon=0}=0 \quad \text{and} \quad \left.\frac{\partial C_{2k}(a, a_p+\delta_p^l)}{\partial(a_p+\delta_p^l)}\right|_{\varepsilon=0} \neq 0.$$

Hence we get

$$C_{2k}(a, a_p+\delta_p^l)=O(a_p+\delta_p^l)=O(a_p)=O(\varepsilon^q) \quad \text{as } \varepsilon \to 0, \tag{5.39}$$

where q is a some real number. Since

$$C_{2k+1}(a, a_p+\delta_p^l)|_{\varepsilon=0} \neq 0, \tag{5.40}$$

we obtain

$$\frac{C_1(a, a_p+\delta_p^j)C_4(a, a_p+\delta_p^j)}{C_2(a, a_p+\delta_p^l)C_3(a, a_p+\delta_p^l)}=O(\varepsilon^{q'}) \quad \text{as } \varepsilon \to 0,$$

where q' is a real number.

Therefore, from (5.38) we get

$$\{C_1(a, a_p+\delta_p^l)C_4(a, a_p+\delta_p^l)-C_1(a, a_p+\delta_p^j)C_4(a, a_p+\delta_p^j)\} \cdot \exp\left[\frac{2r'^{m+1}}{(m+1)|\varepsilon|}\right]$$
$$=O(\varepsilon^{q'}) \cdot \exp[\{2r'^{m+1}-2r^{m+1}\}/(m+1)|\varepsilon|]$$
$$\to 0 \quad \text{as } \varepsilon \to 0,$$

where r' is positive and $r-r'$ is a sufficiently small positive number. Here we used $(5.37)_{h=2}$. Thus we have proved $(5.33)_{h=2}$.

Similarly, we can prove Lemma 5.5.

LEMMA 5.6.

$$\frac{C_{2h+1}(a, a_p+\delta_p^l)}{C_1(a, a_p+\delta_p^l)} :\approx \frac{C_{2h+1}(a, a_p+\delta_p^j)}{C_1(a, a_p+\delta_p^j)} \tag{5.41}$$
$$\text{in } S_l \cap S_j \quad (h=1, 2, \cdots, m-1).$$

Proof. Putting $h=1, 2$ in (5.37), we can get

$$C_1(a, a_p+\delta_p^l)C_2(a, a_p+\delta_p^l)C_2(a, a_p+\delta_p^l)C_1(a, a_p+\delta_p^l)$$
$$\cdot \frac{C_3(a, a_p+\delta_p^l)}{C_1(a, a_p+\delta_p^l)}$$
$$\approx C_1(a, a_p+\delta_p^j)C_2(a, a_p+\delta_p^j)C_2(a, a_p+\delta_p^j)C_1(a, a_p+\delta_p^j) \tag{5.42}$$
$$\cdot \frac{C_3(a, a_p+\delta_p^j)}{C_1(a, a_p+\delta_p^j)} \quad \text{in } S_j \cap S_l.$$

Hence

$$
\begin{aligned}
&\{C_1(a, a_p+\delta_p^l)C_2(a, a_p+\delta_p^l)\}^2 \cdot \left[\frac{C_3(a, a_p+\delta_p^l)}{C_1(a, a_p+\delta_p^l)}-\frac{C_3(a, a_p+\delta_p^j)}{C_1(a, a_p+\delta_p^j)}\right] \\
&\approx \frac{C_3(a, a_p+\delta_p^j)}{C_1(a, a_p+\delta_p^j)} \\
&\quad \cdot \{C_1(a, a_p+\delta_p^j)C_2(a, a_p+\delta_p^j)+C_1(a, a_p+\delta_p^l)C_2(a, a_p+\delta_p^l)\} \\
&\quad \cdot \{C_1(a, a_p+\delta_p^j)C_2(a, a_p+\delta_p^j)-C_1(a, a_p+\delta_p^l)C_2(a, a_p+\delta_p^l)\} \\
&\qquad\qquad \text{in } S_j \cap S_l.
\end{aligned} \tag{5.43}
$$

So, by an argument similar to the proof of Lemma 5.5, we can obtain

$$
\frac{C_3(a, a_p+\delta_p^l)}{C_1(a, a_p+\delta_p^l)}-\frac{C_3(a, a_p+\delta_p^j)}{C_1(a, a_p+\delta_p^j)} :\approx 0 \quad \text{in } S_l \cap S_j.
$$

Here we used (5.39), (5.40), and $(5.37)_{h=1}$. Thus we have proved $(5.41)_{h=1}$. Similarly, we can prove Lemma 5.6.

We put

$$
\begin{aligned}
s &= (s_1, \cdots, s_{m-1}, s_{m+1}, s_{m+2}, \cdots, s_{2m}) \\
&= (\delta_1^l-\delta_1^j, \cdots, \delta_{m-1}^l-\delta_{m-1}^j, \delta_{m+1}^l-\delta_{m+1}^j, \cdots, \delta_{2m}^l-\delta_{2m}^j).
\end{aligned} \tag{5.44}
$$

Note that, if $a(\varepsilon) \approx b(\varepsilon)$ in S, then $a(\varepsilon) :\approx b(\varepsilon)$ in S. We have obtained in Lemmas 5.4–5.6 that

$$
\begin{aligned}
&\tilde{C}_h(a, a_p+\delta_p^j+s_p)-\tilde{C}_h(a, a_p+\delta_p^j) \\
&\quad = -\omega^{1-(-1)^h 2\alpha_{m+1}(a, a_p+\delta_p^j+s_p)}+\omega^{1-(-1)^h 2\alpha_{m+1}(a, a_p+\delta_p^j)} \\
&\quad :\approx 0 \quad \text{in } S_l \cap S_j,
\end{aligned} \tag{5.45}
$$

$$
\begin{aligned}
&C^*(a, a_p+\delta_p^j+s_p)C(a, G^{2h+2}(a_p+\delta_p^j+s_p)) \\
&\qquad -C^*(a, a_p+\delta_p^j)C(a, G^{2h+2}(a_p+\delta_p^j)) \\
&\quad = C^*(a, a_p+\delta_p^j+s_p)C^*(a, G^{2h+1}(a_p+\delta_p^j+s_p)) \\
&\qquad -C^*(a, a_p+\delta_p^j)C^*(a, G^{2h+1}(a_p+\delta_p^j)) \\
&\quad :\approx 0 \quad \text{in } S_j \cap S_l \quad (h=0, 1, \cdots, m-2),
\end{aligned} \tag{5.46}
$$

$$
\begin{aligned}
&\frac{C^*(a, G^{2h}(a_p+\delta_p^j+s_p))}{C^*(a, a_p+\delta_p^j+s_p)}-\frac{C^*(a, G^{2h}(a_p+\delta_p^j))}{C^*(a, a_p+\delta_p^j)} :\approx 0 \\
&\qquad\qquad \text{in } S_l \cap S_j \quad (h=1, 2, \cdots, m-1).
\end{aligned} \tag{5.47}
$$

LEMMA 5.7. *Let*

$$
\begin{aligned}
f_{2h}(a_p, s_p) &= \frac{C^*(a, G^{2h}(a_p+\delta_p^j+s_p))}{C^*(a, a_p+\delta_p^j+s_p)}-\frac{C^*(a, G^{2h}(a_p+\delta_p^j)}{C^*(a, a_p+\delta_p^j)} \qquad (h=1, 2, \cdots, m-1), \\
f_{2h+1}(a_p, s_p) &= C^*(a, a_p+\delta_p^j+s_p)C(a, G^{2h+2}(a_p+\delta_p^j+s_p)) \\
&\quad -C^*(a, a_p+\delta_p^j)C(a, G^{2h+2}(a_p+\delta_p^j)) \qquad (h=0, 1, \cdots, m-2), \\
f_{2m-1}(a_p, s_p) &= -\omega^{1+\alpha_{m+1}(a, a_p+\delta_p^j+s_p)}+\omega^{1-\alpha_{m+1}(a, a_p+\delta_p^j)}.
\end{aligned}
$$

Assume that ε_1 and ε_2 are sufficiently small positive numbers. Then, if

$$
\begin{aligned}
&|a_1|+|a_2|+\cdots+|a_{2m}| \leqq \varepsilon_1, \\
&|a_1'|+|a_2'|+\cdots+|a_{2m-1}'| \leqq \varepsilon_2,
\end{aligned} \tag{5.48}
$$

there exists a unique solution

$$s = g(a_p, a'_p)$$
$$= (g_1(a_p, a'_p), \cdots, g_{m-1}(a_p, a'_p), g_{m+1}(a_p, a'_p), \cdots, g_{2m}(a_p, a'_p))$$

of the system of equations

$$f_k(a_p, s_p) = a'_k \qquad (k = 1, 2, \cdots, 2m-1;\ p = 1, \cdots, m-1, m+1, \cdots, 2m), \tag{5.49}$$

such that $g_1, g_2, \cdots, g_{m-1}, g_{m+1}, \cdots, g_{2m}$ *are holomorphic in the domain* (5.48) *and*

$$g_k(a_p, 0) = 0 \qquad (k = 1, \cdots, m-1, m+1, \cdots, 2m).$$

Proof. It holds that

$$\left.\frac{\partial f_{2h}(a_p, s_p)}{\partial s_k}\right|_{a_p=0, s_p=0} = (\omega^{-2hk} - 1)\frac{1}{C^*(a, 0)}\left.\frac{\partial C^*(a, a_p)}{\partial a_k}\right|_{a_p=0} \tag{5.50}$$
$$(h = 1, 2, \cdots, m-1)$$

$$\left.\frac{\partial f_{2h+1}(a_p, s_p)}{\partial s_k}\right|_{a_p=0, s_p=0} = \omega^{-(2h+2)k} \cdot C^*(a, 0)\left.\frac{\partial C(a, a_p)}{\partial a_k}\right|_{a_p=0} \tag{5.51}$$
$$(h = 0, 1, \cdots, m-2),$$

$$\left.\frac{\partial f_{2m-1}(a_p, s_p)}{\partial s_k}\right|_{a_p=0, s_p=0} = \begin{cases} 0 & (k \neq m+1), \\ -\omega^{1+a} & (k = m+1). \end{cases} \tag{5.52}$$

Here we used (4.25), (4.26), and (4.24).

Now we consider the Jacobian determinant of system (5.49) with respect to $s_1, \cdots, s_{m-1}, s_{m+1}, \cdots, s_{2m}$ at $a_p = 0$, $s_p = 0$.

$$V = \det\left(\frac{\partial f_k}{\partial s_p}\right)_{a_p=0, s_p=0}$$

$$= \begin{vmatrix} & \dfrac{\partial f_{2h+2}}{\partial s_k} & & 0 \\ & & & \vdots \\ & & & 0 \\ 0 & \cdots & 0 \quad \dfrac{\partial f_{2m-1}}{\partial s_{m+1}} \quad 0 & \cdots \quad 0 \\ & \dfrac{\partial f_{2h+1}}{\partial s_k} & & 0 \\ & & & \vdots \\ & & & 0 \end{vmatrix} \qquad \begin{matrix} & \dfrac{\partial f_{2h+2}}{\partial s_{k+m+1}} & \\ & & \\ & & \\ & & \\ & \dfrac{\partial f_{2h+1}}{\partial s_{k+m+1}} & \\ & & \end{matrix}$$

$$(h = 0, 1, \cdots, m-2;\ k = 1, 2, \cdots, m-1),$$

$$= -\omega^{1+a} \cdot \begin{vmatrix} & \vdots & & \vdots & \\ \cdots & (\omega^{-2jk} - 1)\dfrac{1}{C^*}\dfrac{\partial C^*}{\partial a_k} & \cdots & (\omega^{-2jk} - 1)\dfrac{1}{C^*}\dfrac{\partial C^*}{\partial a_{k+m+1}} & \cdots \\ & \vdots & & \vdots & \\ \cdots & \omega^{-2jk} \cdot C^*\dfrac{\partial C}{\partial a_k} & \cdots & \omega^{-2jk} \cdot C^*\dfrac{\partial C}{\partial a_{k+m+1}} & \cdots \\ & \vdots & & \vdots & \end{vmatrix}$$

$$(j, k = 1, 2, \cdots, m-1),$$

$$= -\omega^{1+a} \cdot \begin{vmatrix} V_1 & 0 \\ 0 & V_2 \end{vmatrix}.$$

(5.53)
$$\begin{vmatrix} \ddots & & & \\ \dfrac{\partial C^*}{\partial a_k} & \ddots & \dfrac{\partial C^*}{\partial a_{k+m+1}} & \ddots \\ \ddots & & & \\ \dfrac{\partial C}{\partial a_k} & \ddots & \dfrac{\partial C}{\partial a_{k+m+1}} & \ddots \end{vmatrix} \begin{matrix} \\ m-1 \\ \\ m-1 \end{matrix},$$

where V_1 and V_2 are the $(m-1)$-by-$(m-1)$ matrices whose components are $\omega^{-2jk}-1$ and ω^{-2jk} $(j, k=1, 2, \cdots, m-1)$.

Hence we get

(5.54)
$$V=\frac{1}{2}\cdot|V_1|\cdot|V_2|\cdot\prod_{k=1}^{m-1}\begin{vmatrix} \dfrac{\partial C^*}{\partial a_k} & \dfrac{\partial C^*}{\partial a_{k+m+1}} \\ \dfrac{\partial C}{\partial a_k} & \dfrac{\partial C}{\partial a_{k+m+1}} \end{vmatrix}.$$

Note that

$$|V_1|=\begin{vmatrix} 1 & 1 & \cdots & 1 \\ \omega^{-2} & \omega^{-4} & \cdots & \omega^{-2(m-1)}) \\ \omega^{-4} & \omega^{-8} & \cdots & \omega^{-4(m-1)} \\ \vdots & \vdots & & \vdots \\ \omega^{-2(m-2)} & \omega^{-4(m-2)} & \cdots & \omega^{-2(m-2)(m-1)} \end{vmatrix}$$
$$\cdot\begin{vmatrix} \omega^{-2}-1 & & & \\ & \omega^{-4}-1 & & \\ & & \ddots & \\ & & & \omega^{-2(m-1)}-1 \end{vmatrix}\neq 0.$$

Therefore, by virtue of (4.27), the Jacobian determinant V is different from zero. Thus we have proved this lemma.

Now we apply this lemma to (5.45)-(5.47). If we put

$$a_k'(\varepsilon):\approx 0 \quad \text{in the sector } S_l\cap S_j,$$

we can obtain from Lemma 5.7 that

$$s=g(a_p, a_p')=O(|a_p'|).$$

That is

$$|\delta_k^l(\varepsilon)-\delta_k^j(\varepsilon)|\leqq H\cdot\exp[-2r'^{m+1}/(m+1)|\varepsilon|] \quad \text{in } S_l\cap S_j$$
$$(k=1, \cdots, m-1, m+1, \cdots, 2m)$$

for some positive number H, where $r-r'$ is a sufficiently small positive number.

Therefore, thanks to Theorem 5.1, we can obtain

$$|\delta_k^j(\varepsilon)|\leqq H_{j,k}\cdot\exp[-2r'^{m+1}/(m+1)|\varepsilon|] \quad \text{for } \varepsilon\in S_j,$$

where $H_{j,k}$ are positive numbers.

6. Proof of the main theorem. In this section, as an application of Theorem 1.3, we will prove the main theorem. Let S_j be sectors in Theorem 1.3. Furthermore, for simplicity, we assume that $m+1$ is even.

To prove the main theorem we need the following lemmas which were proved by Lin [3]. We state them here.

LEMMA 6.1. *Let $f_j(\varepsilon)$ be holomorphic in $S_j \cap S_{j+1}$ and asymptotically* 1 *as ε tends to zero in $S_j \cap S_{j+1}$.*

Then there exist functions $g_j(\varepsilon)$ in S_j that are, respectively, holomorphic in S_j and asymptotically 1 *as ε tends to zero in S_j such that*

$$f_j(\varepsilon) = g_{j+1}(\varepsilon)/g_j(\varepsilon) \qquad (j = 1, 2, \cdots, \nu - 1) \quad \text{in } S_j \cap S_{j+1}.$$

We are now in a position to prove the main theorem, Theorem 1.1.

Case I. $x > 0$. The transformation

$$u = y \cdot \exp\left[-x^{m+1}/(m+1)|\varepsilon|\right] \tag{6.1}$$

changes the formally transformed equation (3.7) into

$$\varepsilon y'' - 2x^m y' - \left\{ \sum_{r=0}^{2m-1} a_r(\varepsilon) x^r + m x^{m-1} \right\} y = 0. \tag{6.2}$$

If the differential equation (2.10) satisfies the Matkowsky condition, then, by manipulating the transformations, we can show that the differential equation (6.2) satisfies the Matkowsky condition. So (6.2) has a formal power series solution

$$y(x, \varepsilon) = \sum_{n=0}^{\infty} y_n(x) \varepsilon^n. \tag{6.3}$$

Therefore, (3.7) has a formal solution of the form

$$\exp\left[-x^{m+1}/(m+1)\varepsilon\right] \cdot \sum_{n=0}^{\infty} y_n(x) \varepsilon^n, \tag{6.4}$$

which is subdominant on the positive real axis.

On the other hand, it follows from (4.2) that (3.7) has two independent solutions $y_0(x\varepsilon^{-1/(m+1)}, b, b_p)$ and $y_1(x\varepsilon^{-1/(m+1)}, b, b_p)$. Since y_0 is subdominant and y_1 is dominant on the positive real axis, there exists $c(\varepsilon)$, which does not depend on x, such that

$$\sum_{n=0}^{\infty} y_n(x) \varepsilon^n = c(\varepsilon) \cdot y_0(x\varepsilon^{-1/(m+1)}, b, b_p) \exp\left[\frac{x^{m+1}}{(m+1)\varepsilon}\right]. \tag{6.5}$$

We can choose S_j such that the sector S_j contains $\arg \varepsilon = 0$. Set

$$y^{\#}(x, \varepsilon) = y_0(x\varepsilon^{-1/(m+1)}, b, b_p + \delta_p^j) \exp\left[x^{m+1}/(m+1)\varepsilon\right]. \tag{6.6}$$

Note that (5.16) and

$$\begin{aligned} &\left|\{y_k(z, b, b_p + \delta_p) - y_k(z, b, b_p)\} \exp\{(-1)^k E(z, b_p)\}\right| \\ &\qquad \leqq K|\varepsilon|^q \cdot \sum_{p=1}^{2m} |\delta_p(\varepsilon)| \\ &\qquad\qquad\qquad\qquad (k = 0, 1, \cdots, 2m+1), \end{aligned} \tag{6.7}$$

where K is a positive number and q is some real number (see Sibuya [9, Lemma 3.1]). It follows that

$$\begin{aligned} &\frac{x^{m+1}}{(m+1)\varepsilon} - E(x\varepsilon^{-1/(m+1)}, b_p(\varepsilon)) \\ &\qquad = \sum_{k=1}^{m} \frac{\alpha_k(b, b_p)}{m+1-k} x^{m+1-k} \varepsilon^{-(m+1-k)/(m+1)} \end{aligned} \tag{6.8}$$

remains bounded as $\varepsilon \to 0$. Here we used

$$\alpha_h(b, b_p) = O(b_h(\varepsilon)) = O(\mu^{2h-(2m+2)}) = O(\varepsilon^{1-h/(m+1)}) \qquad (h = 2, \cdots, m), \qquad \alpha_1(b, b_p) = 0.$$

So we get from (6.5) and (6.6) that

$$\begin{aligned} c(\varepsilon)y^{\#}(x, \varepsilon) - \sum_{n=0}^{\infty} y_n(x)\varepsilon^n \\ = c(\varepsilon)\{y_0(z, b, b_p + \delta_p^j) - y_0(z, b, b_p)\} \exp\{E(z, b_p\} \\ \cdot \exp[x^{m+1}/(m+1)\varepsilon - E(z, b_p)] \\ = O(\delta_p^j(\varepsilon)) \simeq 0 \end{aligned}$$

as $\varepsilon \to +0$ uniformly on $x > 0$. We have shown that the actual solution $c(\varepsilon)y^{\#}(x, \varepsilon)$ tends to the formal solution $\sum y_n(x)\varepsilon^n$ as $\varepsilon \to 0$ on the positive real axis $x > 0$.

Case II. $x < 0$. We consider the asymptotic behavior of $c(\varepsilon)y^{\#}(x, \varepsilon)$ on the negative real axis $x < 0$, by using the connection formulas for $y_k(z, b, b_p)$ and $y_k(z, b, b_p + \delta_p)$.

Let us put

$$y^{\#\#}(x, \varepsilon) = y_1(z, b, b_p) \exp[x^{m+1}/(m+1)\varepsilon],$$

$$Y^{\#}(x, \varepsilon) = \begin{pmatrix} y^{\#}(x, \varepsilon) & y^{\#\#}(x, \varepsilon) \\ \dfrac{d}{dx} y^{\#}(x, \varepsilon) + x^m \varepsilon^{-1} y^{\#}(x, \varepsilon) & \dfrac{d}{dx} y^{\#\#}(x, \varepsilon) + x^m \varepsilon^{-1} y^{\#\#}(x, \varepsilon) \end{pmatrix}$$

$$C(\varepsilon) = \begin{pmatrix} c(\varepsilon) \\ 0 \end{pmatrix}.$$

So we have

$$\begin{aligned} Y^{\#}(x, \varepsilon)C(\varepsilon) &= \Phi_{j,0}(x, \varepsilon) \exp[x^{m+1}/(m+1)\varepsilon] \cdot I_2 \cdot C(\varepsilon) \\ &= \Phi_{j,m+1}(x, \varepsilon) \cdot \prod_{h=0}^{m} \Gamma(b, b_p(\varepsilon) + \delta_p^j(\varepsilon)) \\ &\quad \cdot \exp\left[\frac{x^{m+1}}{(m+1)\varepsilon}\right] \cdot I_2 \cdot C(\varepsilon). \end{aligned} \tag{6.9}$$

Similarly, putting

$$\begin{aligned} y^*(x, \varepsilon) &= y_0(z, b, b_p(\varepsilon)) \exp[x^{m+1}/(m+1)\varepsilon] \\ &= c(\varepsilon)^{-1} \cdot \sum_{n=0}^{\infty} y_n(x)\varepsilon^n, \end{aligned}$$

$$y^{**}(x, \varepsilon) = y_1(z, b, b_p(\varepsilon)) \exp[x^{m+1}/(m+1)\varepsilon],$$

$$Y^*(x, \varepsilon) = \begin{pmatrix} y^*(x, \varepsilon) & y^{**}(x, \varepsilon) \\ \dfrac{d}{dx} y^*(x, \varepsilon) + x^m \varepsilon^{-1} y^*(x, \varepsilon) & \dfrac{d}{dx} y^{**}(x, \varepsilon) + x^m \varepsilon^{-1} y^{**}(x, \varepsilon) \end{pmatrix},$$

we get

$$
\begin{aligned}
Y^*(x,\varepsilon)C(\varepsilon) &= \tilde{\Phi}_0(x,\varepsilon)\cdot\exp[x^{m+1}/(m+1)\varepsilon]\cdot I_2\cdot C(\varepsilon)\\
&= \tilde{\Phi}_{m+1}(x,\varepsilon)\cdot\prod_{h=0}^{m}\Gamma(b,b_p(\varepsilon))\cdot\exp\left[\frac{x^{m+1}}{(m+1)\varepsilon}\right]\cdot I_2\cdot C(\varepsilon),
\end{aligned}
\tag{6.10}
$$

where

$$
\tilde{\Phi}_h(x,\varepsilon)=\Lambda(\varepsilon)\Psi(x\varepsilon^{-1/(m+1)\varepsilon},\, b,\, b_p)\qquad (h=0,1,\cdots,2m+1),
$$

which is a fundamental matrix solution of (3.4). Note that $\Phi_{m+1}(x,\varepsilon)$ and $\tilde{\Phi}_{m+1}(x,\varepsilon)$ are asymptotically known on the negative real axis $\arg x=\pi$.

If we put

$$
\prod_{h=0}^{m}\Gamma(b,b_p(\varepsilon))=\begin{pmatrix}A_{11}(b_p) & A_{12}(b_p)\\ A_{21}(b_p) & A_{22}(b_p)\end{pmatrix},
\tag{6.11}
$$

we can obtain from (6.10) that

$$
\begin{aligned}
\sum_{n=0}^{\infty}y_n(x)\varepsilon^n &= c(\varepsilon)y^*(x,\varepsilon)\\
&= c(\varepsilon)\{A_{11}(b_p)y_{m+1}(z,b,b_p)+A_{21}(b_p)y_{m+2}(z,b,b_p)\}\\
&\quad\cdot\exp[x^{m+1}/(m+1)\varepsilon].
\end{aligned}
\tag{6.12}
$$

On the other hand, we have from (6.9) that

$$
\begin{aligned}
c(\varepsilon)y^{\#}(x,\varepsilon) = c(\varepsilon)\{&A_{11}(b_p+\delta_p^j)y_{m+1}(z,b,b_p+\delta_p^j)\\
&+A_{21}(b_p+\delta_p^j)y_{m+2}(z,b,b_p+\delta_p^j)\}\\
&\cdot\exp[x^{m+1}/(m+1)\varepsilon].
\end{aligned}
\tag{6.13}
$$

Since $m+1$ is even, it follows from (6.12) that

$$
A_{21}(b_p)\cdot\exp[2x^{m+1}/(m+1)\varepsilon]
$$

remains bounded as $\varepsilon\to+0$ on $x<0$. So, as an application of Theorem 1.3, we prove that

$$
A_{21}(b_p+\delta_p^j)\cdot\exp[2x^{m+1}/(m+1)\varepsilon]
$$

remains bounded as $\varepsilon\to+0$ on $x<0$. To shorten the expressions for the calculation that follows, the notation $a(\varepsilon)\approx; b(\varepsilon)$ in S will be introduced to indicate that

$$
\{a(\varepsilon)-b(\varepsilon)\}\cdot\exp[2r_0^{m+1}/(m+1)\varepsilon]
$$

remains bounded, as $\varepsilon\to 0$ in the sector S for $\underline{a,\, b<r_0<r_1}$ (cf. Definition 5.3). That is

$$
\{a(\varepsilon)-b(\varepsilon)\}\cdot\exp[2r_0^{m+1}/(m+1)\cdot\mathrm{Re}(1/\varepsilon)]
$$

remains bounded as $\varepsilon\to 0$ in S.

Noting that

$$
\alpha_k(b,b_p+\delta_p^{l+1})-\alpha_k(b,b_p+\delta_p^{l})\simeq 0\qquad(\varepsilon\to 0 \text{ in } S_l\cap S_{l+1}),
$$

by the same argument as Lemma 5.2, we can easily obtain

$$
d_{ii}^h(\varepsilon)\simeq 1\qquad(\varepsilon\to 0 \text{ in } S_l\cap S_{l+1}),\quad(i=1,2;\, h=0,1,\cdots,2m+1),
\tag{6.14}
$$

$$d_{12}^{h}(\varepsilon)\cdot\exp\left[\frac{2x^{m+1}}{(m+1)\varepsilon}\right]\simeq 0,\qquad d_{21}^{h}(\varepsilon)\cdot\exp\left[\frac{2x^{m+1}}{(m+1)\varepsilon}\right]\simeq 0 \tag{6.15}$$

as $\varepsilon\to 0$ in $S_l\cap S_{l+1}$ uniformly in (5.17-h). In (6.15), we consider the cases in which $h=0$ and $h=m+1$. Since $-\pi/2<\arg\varepsilon<\pi/2$, it follows from (5.17-0) that $-\pi/(m+1)<\arg x<\pi/(m+1)$. So we can choose values of $x\in D_1$ such that $\arg x=0$. Hence we have from (6.15) that

$$d_{21}^{0}(\varepsilon)\approx;0\quad\text{in } S_l\cap S_{l+1}. \tag{6.16}$$

Similarly, in (5.17-($m+1$)) we can choose a value $x\in D_1$ such that $\arg x=\pi$. Then we get from (6.15) that

$$d_{21}^{m+1}(\varepsilon)\approx;0\quad\text{in } S_l\cap S_{l+1}. \tag{6.17}$$

Since

$$L_{h+1}(\varepsilon)\Gamma_h(b,b_p+\delta_p^l)=\Gamma_h(b,b_p+\delta_p^{l+1})L_h(\varepsilon), \tag{5.27}$$

we get

$$L_{m+1}(\varepsilon)\cdot\prod_{h=0}^{m}\Gamma_h(b\cdot b_p+\delta_p^l)=\prod_{h=0}^{m}\Gamma_h(b,b_p+\delta_p^{l+1})\cdot L_0(\varepsilon).$$

That is

$$\begin{aligned}&\begin{pmatrix}d_{11}^{m+1}(\varepsilon) & d_{12}^{m+1}(\varepsilon)\\ d_{21}^{m+1}(\varepsilon) & d_{22}^{m+1}(\varepsilon)\end{pmatrix}\begin{pmatrix}A_{11}(b_p+\delta_p^l) & A_{12}(b_p+\delta_p^l)\\ A_{21}(b_p+\delta_p^l) & A_{22}(b_p+\delta_p^l)\end{pmatrix}\\ &\quad=\begin{pmatrix}A_{11}(b_p+\delta_p^{l+1}) & A_{12}(b_p+\delta_p^{l+1})\\ A_{21}(b_p+\delta_p^{l+1}) & A_{22}(b_p+\delta_p^{l+1})\end{pmatrix}\begin{pmatrix}d_{11}^{0}(\varepsilon) & d_{12}^{0}(\varepsilon)\\ d_{21}^{0}(\varepsilon) & d_{22}^{0}(\varepsilon)\end{pmatrix}.\end{aligned} \tag{6.18}$$

We see from the (2, 1)-element of the matrix relation (6.18) that

$$\begin{aligned}&d_{21}^{m+1}(\varepsilon)\cdot A_{11}(b_p+\delta_p^l)+d_{22}^{m+1}(\varepsilon)\cdot A_{21}(b_p+\delta_p^l)\\ &\quad=d_{11}^{0}(\varepsilon)\cdot A_{21}(b_p+\delta_p^{l+1})+d_{21}^{0}(\varepsilon)\cdot A_{22}(b_p+\delta_p^{l+1}).\end{aligned}$$

That is

$$\begin{aligned}&A_{21}(b_p+\delta_p^l)-A_{21}(b_p+\delta_p^{l+1})\cdot\frac{d_{11}^{0}(\varepsilon)}{d_{22}^{m+1}(\varepsilon)}\\ &=\frac{1}{d_{22}^{m+1}(\varepsilon)}[d_{21}^{0}(\varepsilon)\cdot A_{22}(b_p+\delta_p^{l+1})-d_{21}^{m+1}(\varepsilon)\cdot A_{11}(b_p+\delta_p^l)]\\ &\approx;0\quad\text{in } S_l\cap S_{l+1}.\end{aligned} \tag{6.19}$$

Here we used (6.16) and (6.17). Since

$$d_{11}^{0}(\varepsilon)/d_{22}^{m+1}(\varepsilon)\simeq 1\qquad(\varepsilon\to 0\text{ in }S_l),$$

which follows from (6.14), by virtue of Lemma 6.1 there exist functions $g_l(\varepsilon)$ ($l=1,2,\cdots,\nu$) in S_l such that

$$g_l(\varepsilon)\simeq 1\qquad(\varepsilon\to 0\text{ in }S_l), \tag{6.20}$$

$$d_{11}^{0}(\varepsilon)/d_{22}^{m+1}(\varepsilon)=g_{l+1}(\varepsilon)/g_l(\varepsilon)\quad\text{in } S_l\cap S_{l+1}. \tag{6.21}$$

Then (6.19) reduces to

$$g_l(\varepsilon)A_{21}(b_p+\delta_p^l)-g_{l+1}(\varepsilon)A_{21}(b_p+\delta_p^{l+1})\approx;0 \quad \text{in } S_l\cap S_{l+1}. \tag{6.22}$$

In § 5 we considered the case in which the domain D is a disk with center at zero. In this section, the domain of x contains a disk with center at zero, although the radius of the disk is small. So we get

$$|\delta_p^l(\varepsilon)|\leqq H_1\cdot\exp[-2r_2^{m+1}/(m+1)|\varepsilon|] \quad \text{in } S_l$$
$$(l=1,2,\cdots,\nu;\ p=1,2,\cdots,2m),$$

for some positive number H_1.

Let r_0 be a positive number with $a<r_0<r_1$ and choose ω_0 with $0<\omega_0<\pi/2$ such that $r_3^{m+1}\geqq r_0^{m+1}\cdot\cos\omega_0>0$. Then if $\arg\varepsilon=\omega_0$ (or $-\omega_0$), we get

$$r_2^{m+1}/|\varepsilon|\geqq r_0^{m+1}\cdot\mathrm{Re}[1/\varepsilon].$$

Therefore, for some sector that contains the line segment $\arg\varepsilon=\omega_0$ (or $-\omega_0$), say S_ν (or S_1), we have

$$|\delta_p^1(\varepsilon)|,|\delta_p^\nu(\varepsilon)|\leqq H_1\cdot\exp\left[-2r_0^{m+1}/(m+1)\cdot\mathrm{Re}\left(\frac{1}{\varepsilon}\right)\right],\qquad (p=1,2,\cdots,2m).$$

That is

$$\begin{aligned}&\delta_p^1(\varepsilon)\approx;0 \quad \text{in } \arg\varepsilon=-\omega_0,\\ &\delta_p^\nu(\varepsilon)\approx;0 \quad \text{in } \arg\varepsilon=\omega_0 \quad (p=1,2,\cdots,2m).\end{aligned} \tag{6.23}$$

On the other hand, it holds that

$$A_{21}(b_p+\delta_p^l)-A_{21}(b_p)=O(\delta_p^l(\varepsilon)),$$
$$A_{21}(b_p)\approx;0.$$

Putting $l=1$, ν and using (6.23) and (6.20), we obtain

$$\begin{aligned}&g_1(\varepsilon)\cdot A_{21}(b_p+\delta_p^1)\approx;0 \quad \text{in } \arg\varepsilon=-\omega_0,\\ &g_\nu(\varepsilon)\cdot A_{21}(b_p+\delta_p^\nu)\approx;0 \quad \text{in } \arg\varepsilon=\omega_0.\end{aligned} \tag{6.24}$$

If we put $\phi_j(\varepsilon)=g_j(\varepsilon)\cdot A_{21}(b_p(\varepsilon)+\delta_p^j(\varepsilon))$ and apply Theorem 1.3 to (6.23) and (6.24), we can obtain

$$g_j(\varepsilon)\cdot A_{21}(b_p(\varepsilon)+\delta_p^j(\varepsilon))\approx;0 \quad \text{in } S_j.$$

That is

$$A_{21}(b_p(\varepsilon)+\delta_p^j(\varepsilon))\approx;0 \quad \text{in } S_j \quad (j=1,2,\cdots,\nu). \tag{6.25}$$

Here we used (6.20).

Now, by virtue of Cauchy's integral representation theorem, we can prove that

$$\begin{aligned}&|\{A_{21}(b_p(\varepsilon)+\delta_p^j(\varepsilon))-A_{21}(b_p(\varepsilon))\}\exp[2r_{r0}^{m+1}/(m+1)\varepsilon]|\\ &\leqq H\cdot\sum_{p=1}^{2m}|\delta_p^j(\varepsilon)|,\end{aligned} \tag{6.26}$$

where H is a positive number (cf. the proof of Lemma 3.1 in [8]).

We now return to (6.12) and (6.13), and use (6.25) and (6.26). Then we get

$$
\begin{aligned}
& c(\varepsilon)y^{\#}(x,\varepsilon)-\sum_{n=0}^{\infty}y_n(x)\varepsilon^n \\
&= c(\varepsilon)[\{A_{11}(b_p+\delta_p^j)-A_{11}(b_p)\}y_{m+1}(z,b,b_p+\delta_p^j) \\
&\qquad +A_{11}(b_p)\{y_{m+1}(z,b,b_p+\delta_p^j)-y_{m+1}(z,b,b_p)\}] \\
&\quad \cdot \exp\{+E(z,b_p)\}\cdot\exp[-E(z,b_p)+x^{m+1}/(m+1)\varepsilon] \\
&\quad +c(\varepsilon)[\{A_{21}(b_p+\delta_p^j)-A_{21}(b_p)\}y_{m+2}(z,b,b_p+\delta_p^j) \\
&\qquad +A_{21}(b_p)\{y_{m+2}(z,b,b_p+\delta_p^j)-y_{m+2}(z,b,b_p)\}] \\
&\quad \cdot \exp\{-E(z,b_p)\}\cdot\exp[+E(z,b_p)+x^{m+1}/(m+1)\varepsilon] \\
&= c(\varepsilon)\cdot O(\delta_p^j(\varepsilon))\cdot y_{m+1}(\cdot,\cdot,\cdot)\exp\{E(z,b_p)\}+c(\varepsilon)\cdot A_{11}(b_p)\cdot O(\delta_p^j(\varepsilon)) \\
&\quad +c(\varepsilon)\cdot O(\delta_p^j(\varepsilon))\cdot y_{m+2}(\cdot,\cdot,\cdot)\exp\{-E(z,b,b_p)\} \\
&\qquad +c(\varepsilon)\cdot A_{21}(b_p)\exp[2x^{m+1}/(m+1)\varepsilon]\cdot O(\delta_p^j(\varepsilon)) \\
&\simeq 0
\end{aligned}
$$

as $\varepsilon \to +0$ uniformly on $x<0$.

Thus we have proved that $c(\varepsilon)y^{\#}(x,\varepsilon)$ converges uniformly on a real interval to a nontrivial solution as ε tends to zero on the positive real axis.

Acknowledgment. I thank Professor Y. Sibuya for his suggestions contributing to this research.

REFERENCES

[1] R. J. HANSON, *Simplification of second-order systems of ordinary differential equation with a turning point*, SIAM J. Appl. Math., 16 (1968), pp. 1059-1080.

[2] R. J. HANSON AND D. L. RUSSEL, *Classification and a reduction of second order systems at a turning point*, J. Math. Phys., 46 (1967), pp. 74-92.

[3] P. F. HSIEH AND Y. SIBUYA, *On the asymptotic integration of second order linear ordinary differential equation with polynomial coefficients*, J. Math. Anal. Appl., 16 (1966), pp. 84-103.

[4] C. H. LIN, *The sufficiency of the Matkowsky condition in the problem of resonance*, Trans. Amer. Math. Soc., 278 (1983), pp. 647-670.

[5] ———, *Phragmen-Lindelof theorem in a cohomological form*, Proc. Amer. Math. Soc., 89 (1983), pp. 589-597.

[6] B. J. MATKOWSKY, *On boundary layer problems exhibiting resonance*, SIAM Rev., 17 (1975), pp. 82-100.

[7] S. OHKOHCHI, *An extension of Weber's equation*, Kumamoto J. Math., 3 (1990), pp. 69-80.

[8] Y. SIBUYA, *Uniform simplification in a full neighborhood of a transition point*, Mem. Amer. Math. Soc., 149 (1974).

[9] ———, *A theorem concerning uniform simplification at a transition point and the problem of resonance*, SIAM J. Math. Anal., 12 (1981), pp. 653-668.

[10] W. WASOW, *Linear Turning Point Theory*, Springer-Verlag, Berlin, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 238-259, January 1991

014

REDUCTION TO CANONICAL FORMS AND THE STOKES PHENOMENON IN THE THEORY OF LINEAR DIFFERENCE EQUATIONS*

G. K. IMMINK†

Abstract. Previous results concerning the existence of right inverses of linear difference operators on Banach Spaces of holomorphic functions are extended. The Stokes phenomenon is analyzed for a class of very singular linear difference equations.

Key words. linear difference operator, right inverse, asymptotic expansion, canonical form, Stokes phenomenon, connection matrix

AMS(MOS) subject classification. 39A10

Introduction. This paper is concerned with homogeneous linear difference equations of the type

$$y(s+1)-A(s)y(s)=0, \tag{0.1}$$

where s is a complex variable and A is an $n\times n$ matrix function, meromorphic at ∞. More precisely, we shall assume that $A\in \mathrm{Gl}(n;\mathbb{C}\{s^{-1}\}[s])$, $n\in\mathbb{N}$. We are interested in the global asymptotic properties of solutions of (0.1), i.e., their behaviour as $s\to\infty$ in an arbitrary direction.

It should be noted here that, in general, solutions of (0.1) are not analytic in a (reduced) neighbourhood of ∞. However, it is easily seen from (0.1) that any solution, analytic in a left half plane, can be continued analytically to a region of the form $-U(R)$, where

$$U(R)=\{s\in\mathbb{C}: |s+x|\geqq R\forall x\geqq 0\},\qquad R>0. \tag{0.2}$$

Similarly, the relation

$$y(s)=A(s)^{-1}y(s+1)$$

implied by (0.1), shows that any solution, analytic in a right halfplane, can be continued analytically to a region of the form $U(R)$, $R>0$. Therefore, we shall consider the asymptotic behaviour of solutions of (0.1) in regions of either type.

The usual approach to this kind of problem is the following. First, the existence of fundamental solutions of the equation with a prescribed asymptotic behaviour, in different sectors covering a neighbourhood of ∞ is established. Next, the relations between these solutions are studied.

For example, let Y_1 and Y_2 be holomorphic fundamental solutions of (0.1), admitting the same asymptotic representation in a left and an upper halfplane, respectively. The connection matrix P is defined by

$$Y_1=Y_2P. \tag{0.3}$$

As $Y_1(s+1)Y_1(s)^{-1}=Y_2(s+1)Y_2(s)^{-1}=A(s)$, P is a periodic function of period 1. Obviously, (0.3) defines the analytic continuation of Y_1 to an upper halfplane and knowledge of P implies knowledge of the asymptotic behaviour of Y_1 in this upper half plane.

* Received by the editors March 23, 1988; accepted for publication (in revised form) March 12, 1990.
† University of Groningen, Institute of Econometrics, P.O. Box 800, 9700 AV, Groningen, the Netherlands.

The asymptotic representations that play a role in the study of (0.1) are the so-called formal solutions of this equation. It is known that (0.1) possesses a formal fundamental solution of the form

$$\hat{Y}(s) = \hat{F}(s) \operatorname{diag}\{s^{G_1} \exp q_1(s), \cdots, s^{G_m} \exp q_m(s)\},$$

where $\hat{F} \in \mathrm{Gl}(n; \mathbb{C}[[s^{-1/p}]][s^{1/p}])$ for some $p \in \mathbb{N}$, $m \in \mathbb{N}(m \leqq n)$, G_j is a constant matrix and

$$q_j(s) = d_j s \log s + \sum_{h=1}^{p} \mu_{j,h} s^{h/p} \tag{0.4}$$

with $d_j \in \mathbb{Q}$ and $\mu_{j,h} \in \mathbb{C}$ for each $j \in \{1, \cdots, m\}$, $h \in \{1, \cdots, p\}$ (cf. [2], [13], [16]). All constants figuring in this representation are uniquely determined by the matrix function A, except for $\mu_{j,p}$, which is determined up to a multiple of $2\pi i$ (this is related to the fact that any matrix solution of (0.1), when multiplied from the right with a periodic matrix function, remains a solution of (0.1)).

If $d_i = d_j$ for all $i, j \in \{1, \cdots, m\}$, the numbers

$$\frac{1}{p} \operatorname{degr}(q_i - q_j)$$

will be called the "levels" of the difference equation. The most difficult case to deal with is when $d_i \neq d_j$ for at least one pair (i, j) with $i \neq j$. In this case we will say that the "level 1^+" is present in (0.1). If $d_i \neq d_j$ for all $i \neq j$, we shall, with a slight abuse of terminology, speak of a difference equation of level 1^+ (cf. [4]).

Existence theorems for solutions of homogeneous linear difference equations with a prescribed asymptotic behaviour have been the subject of various studies since the beginning of this century (cf. [1], [3], [5], [7], [12]). The most general result so far is a theorem by Birkhoff and Trjitzinsky (cf. [1]). It states that, in every quadrant Γ of the form

$$\Gamma = \left\{ s \in \mathbb{C}: k\frac{\pi}{2} \leqq \arg(s - s_0) \leqq (k+1)\frac{\pi}{2}; |s| \geqq R \right\}, \qquad k \in \mathbb{Z}, \quad s_0 \in \mathbb{C}, \quad R > 0,$$

there exists a holomorphic fundamental solution of (0.1), represented asymptotically by a given formal fundamental solution as $s \to \infty$ in Γ, provided R is sufficiently large. However, the proof of this result contains some inaccuracies and its correctness has been questioned.

In [7] we have derived existence theorems for both linear and nonlinear difference equations using a method developed by Hukuhara, Sibuya, Malgrange and others, for analogous problems in the theory of differential equations (cf. [6], [11], [14], [17]). It is based on the existence of right inverses of linear difference operators on Banach spaces of functions that are holomorphic in suitable ("proper") regions of the complex plane. Later we realized that the class of "proper" regions considered in [7], at least in the presence of level 1^+, was too restricted, and that our results could be easily extended. This is explained in § 2 of the present paper. These generalizations permit us to give a straightforward proof of the theorem of Birkhoff and Trjitzinsky, thereby settling the first half of our problem.

Next, we turn to the second part: the connection problem or Stokes phenomenon, i.e., the change in asymptotic behaviour of solutions of the equation, as they are continued analytically beyond certain "maximal regions." This phenomenon has been studied in [8], under "generic" conditions, and in particular the link with the theory of resurgent functions was explained. It is closely related to the problem of analytic

classification of difference equations, which has been solved by Ecalle in [4]. Nevertheless, the precise nature of the Stokes phenomenon in the most general case of (0.1) has remained somewhat mysterious. The results presented in § 2 also contribute to a better understanding of this phenomenon.

The maximal regions in which a solution of (0.1) may be represented asymptotically by a given formal solution are bounded by curves of the form

$$\operatorname{Re}\{q_i(s)-q_j(s)\}=c, \qquad i,j\in\{1,\cdots,m\}, \quad i\neq j.$$

We shall call these curves Stokes curves of level k if $d_i=d_j$ and $1/p \operatorname{degr}(q_i-q_j)=k$, and Stokes curves of level 1^+ if $d_i\neq d_j$. Due to the infinite number of possible determinations of the $\mu_{j,p}$ in (0.4), there is a countably infinite number of Stokes curves of the levels 1 and 1^+ (we do not distinguish between curves that differ only in the value of c). Whereas two Stokes curves of a level less than or equal to 1 generally have distinct limiting directions, those of level 1^+ all have the same limiting directions, viz. those of the positive and negative imaginary axis. This makes the analysis of the Stokes phenomenon more delicate in the presence of the level 1^+ than otherwise. In order to distinguish between different solutions, we must take into consideration their behaviour along curves of the type

$$\operatorname{Re} s(\log s+i\theta)=c, \qquad \theta\in\mathbb{R}.$$

Section 4 deals with the Stokes phenomenon of homogeneous linear difference equations of level 1^+. We study the properties of the periodic matrix functions connecting two fundamental solutions of (0.1) represented asymptotically by the formal fundamental solution in different maximal regions (a method to compute the leading parts of these connection matrices from the asymptotic behaviour of the coefficients of the formal fundamental solution is discussed in [9]).

Throughout this paper, we have restricted ourselves to "classical" asymptotic expansions, just as in [7], however, all statements remain valid (with slight modifications) when these expansions are replaced by asymptotic expansions with suitable Gevrey-type error bounds (cf. also [10]).

1. Definitions and notation.

1.1. Classes of holomorphic functions admitting an asymptotic power series representation. In order to describe the asymptotic properties of solutions of (0.1), we introduce families of closed unbounded regions of $\mathbb{C}$, indexed by a parameter $R\in\mathbb{R}^+$, which measures the distance to the origin. We define classes of holomorphic functions on these regions, admitting an asymptotic expansion with uniform error bounds.

We shall restrict our attention to subregions of $\mathbb{C}\backslash\mathbb{R}^-$. All results obtained for this type of regions can be easily "translated" into analogous statements for corresponding regions in $\mathbb{C}\backslash\mathbb{R}^+$, by means of the following equality:

$$y(s+1)-A(s)y(s)=-A(s)\{\tilde{y}(-s)-\tilde{A}(-s-1)\tilde{y}(-s-1)\},$$

where $\tilde{A}$ and $\tilde{y}$ are defined by

$$\tilde{A}(s)=A(-s-1)^{-1}, \qquad \tilde{y}(s)=y(-s).$$

DEFINITION 1.1.1. An asymptotic set S of closed regions is a decreasing set of closed unbounded regions $S(R)$ of the complex plane, defined for all $R>1$, with the property that $d(S(R),0)\to\infty$ as $R\to\infty$.

DEFINITION 1.1.2. Let I be an index set and let $S_i=\{S_i(R), R>1\}$ be an asymptotic set of closed regions for all $i\in I$. By $\bigcup_{i\in I} S_i$ and $\bigcap_{i\in I} S_i$ we shall denote

the sets

$$\left\{\bigcup_{i\in I} S_i(R),\ R>1\right\},\qquad \left\{\bigcap_{i\in I} S_i(R),\ R>1\right\},$$

respectively. A set of the first type will be called an asymptotic set.

If I is a finite set then $\bigcup_{i\in I} S_i$ is again an asymptotic set of closed regions. If, in addition, $\bigcap_{i\in I} S_i(R)\neq\emptyset$ for all $R>0$, then also $\bigcap_{i\in I} S_i$ is an asymptotic set of closed regions.

DEFINITION 1.1.3. Let S be an asymptotic set of closed regions. By $\mathcal{M}(S)$ we shall denote the set of all functions f on $\mathbb{C}$ with the following properties:

(i) There exists a positive number R such that f is continuous on $S(R)$ and holomorphic in int $S(R)$.

(ii) There exist an integer h_0, a positive integer p and complex numbers a_h, $h\in\mathbb{Z}$, such that, for all $N\in\mathbb{N}$,

$$\sup_{s\in S(R)}\left|s^{N/P}\left\{f(s)-\sum_{h=h_0}^{N-1} a_h s^{-h/p}\right\}\right|<\infty.$$

If these conditions are fulfilled we write

$$\sum_{h=h_0}^{\infty} a_h s^{-h/p}=\hat{f}\quad\text{and}\quad f\sim\tilde{f},\quad s\to\infty,\quad s\in S(R).$$

Remark. Different functions that coincide on $S(R)$ for some $R>0$, will be identified. More precisely, a "function" f is to be thought of as a representative of the equivalence class of all functions g with the property that there exists a positive number R such that $g(s)=f(s)$ for all $s\in S(R)$.

Let I be a finite index set and suppose that, for every $i\in I$, S_i is an asymptotic set of closed regions. Then, obviously, $\mathcal{M}(\bigcup_{i\in I} S_i)$ is contained in $\bigcap_{i\in I}\mathcal{M}(S_i)$. If, in addition, int $\bigcup_{i\in I} S_i(R)=\bigcup_{i\in I}$ int $S_i(R)$ for all $R>1$, then it immediately follows that $\mathcal{M}(\bigcup_{i\in I} S_i)=\bigcap_{i\in I}\mathcal{M}(S_i)$.

DEFINITION 1.1.4. Let I be an index set and, for every $i\in I$, let S_i be an asymptotic set of closed regions such that

$$\text{int}\bigcup_{i\in I} S_i(R)=\bigcup_{i\in I}\text{int}\,S_i(R)\quad\text{for all } R>1.$$

Let $S=\bigcup_{i\in I} S_i$. By $\mathcal{M}(S)$ we shall denote the set

$$\mathcal{M}(S)=\bigcap_{i\in I}\mathcal{M}(S_i),$$

by $\mathcal{A}(S)$ the set of all $f\in\mathcal{M}(S)$ such that

$$\hat{f}\in\bigcup_{p\in\mathbb{N}}\mathbb{C}[\![s^{-1/p}]\!],$$

and by $\mathcal{A}_0(S)$ the set of $f\in\mathcal{M}(S)$ with the property that $\hat{f}=0$.

In the following definition some particular examples of asymptotic sets are given.

DEFINITION 1.1.5. For all α_0, $\beta_0\in[-\pi,\pi]$ such that $\alpha_0<\beta_0$, we define the following asymptotic sets:

(i) The "closed sector" $S[\alpha_0,\beta_0]$ defined by

$$S[\alpha_0,\beta_0](R)=\{s\in\mathbb{C}\backslash\mathbb{R}^-\colon \alpha_0\leqq\arg(s+\mathrm{Re}^{i\alpha})\leqq\beta_0\text{ for all }\alpha\in[-\pi,\pi]\};$$

(ii) The "half-open sectors" $S[\alpha_0,\beta_0)=\bigcup_{\alpha_0<\beta<\beta_0} S[\alpha_0,\beta]$ and $S(\alpha_0,\beta_0]=\bigcup_{\alpha_0<\alpha<\beta_0} S[\alpha,\beta_0]$;

(iii) The "open sector" $S(\alpha_0,\beta_0)=\bigcup_{\alpha_0<\alpha<\beta<\beta_0} S[\alpha,\beta]$.

If S is any of the sectors defined in (ii) and (iii), then $\bar{S}$ will denote the closed sector $S[\alpha_0, \beta_0]$.

1.2. Canonical forms and formal invariants. We use the following notation:

$$K = \bigcup_{p \in N} \mathbb{C}\{s^{-1/p}\}[s^{1/p}], \qquad \hat{K} = \bigcup_{p \in \mathbb{N}} \mathbb{C}[\![s^{-1/p}]\!][s^{1/p}].$$

Let S be an asymptotic set. If $A \in \mathrm{End}(n; \mathcal{M}(S))$ and $F \in \mathrm{Gl}(n; \mathcal{M}(S))$, or $A \in \mathrm{End}(n; \hat{K})$ and $F \in \mathrm{Gl}(n; \hat{K})$ we shall denote by A^F the matrix function

$$A^F(s) = F(s+1)^{-1} A(s) F(s).$$

By Δ_A we shall denote the linear difference operator defined by

$$\Delta_A y(s) = y(s+1) - A(s) y(s),$$

where y belongs to a suitable space of n-dimensional vector functions.

Let $A, B \in \mathrm{Gl}(n; \mathcal{M}(S))$. The difference operators Δ_A and Δ_B are said to be formally equivalent if there exists a matrix function $F \in \mathrm{Gl}(n; \hat{K})$ with the property that

$$\hat{A}^F = \hat{B}.$$

If $A \in \mathrm{Gl}(n; \mathcal{M}(S))$ the difference operator Δ_A is known to be formally equivalent to a difference operator $\Delta_{\overset{c}{A}}$ of the following particular type. The matrix function $\overset{c}{A}$ is block diagonal,

$$\overset{c}{A} = \mathrm{diag}\{\overset{c}{A}_1, \cdots, \overset{c}{A}_m\}, \qquad m \in \mathbb{N}$$

with diagonal blocks of the form

$$\overset{c}{A}_j(s) = \exp\{q_j(s+1) - q_j(s)\}(1 + 1/s)^{G_j},$$

where, for $j \in \{1, \cdots, m\}$,

(1.2.1)
$$q_j(s) = d_j s \log s + \sum_{h=1}^{p} \mu_{j,h} s^{h/p}, \qquad p \in \mathbb{N}, \quad d_j \in \mathbb{Z}/p, \quad \mu_{j,1}, \cdots, \mu_{j,p} \in \mathbb{C},$$

$0 \leqq \mathrm{Im}\, \mu_{j,p} < 2\pi$, and

$G_j = \gamma_j I_{n_j} + N_j$, $\gamma_j \in \mathbb{C}$, $0 \leqq \mathrm{Re}\, \gamma_j < 1/p$, $n_j \in \mathbb{N}$, N_j is a nilpotent $n_j \times n_j$ matrix.

The matrix function $\overset{c}{A}$ is uniquely determined by A up to permutations of the diagonal blocks. We shall assume that the blocks are arranged in such a way that

$$d_i \leqq d_j \quad \text{if } i < j, \quad i, j \in \{1, \cdots, m\}.$$

DEFINITION 1.2.2. A matrix function $\overset{c}{A}$ with the above-mentioned properties will be called a canonical matrix or a canonical form of A.

The numbers d_j, γ_j, and $\mu_{j,h}$ ($j \in \{1, \cdots, m\}$, $h \in \{1, \cdots, p\}$) are formal invariants of the difference equation $\Delta_A y = 0$. They are uniquely determined by the matrix function A.

We shall further use the following notation:

$Q = \mathrm{diag}\{q_1 I_{n_1}, \cdots, q_m I_{n_m}\}$,

$\overset{c}{Y}(s) = e^{Q(s)} s^G$ (note that $\overset{c}{A}(s) = \overset{c}{Y}(s+1)\overset{c}{Y}(s)^{-1}$),

$d(A) = \{d_1, \cdots, d_m\}$,

$k(A) = \{\mathrm{degr}\, q_j : j \in \{1, \cdots, m\}$ such that $d_j = 0\}$ (degr q_j here is understood to be a rational number not exceeding 1).

DEFINITION 1.2.3. Let $k \in k(A)$, $k \neq 0$. By $\Sigma_k(A)$ we shall denote the set of all real numbers α with the property that there is a $j \in \{1, \cdots, m\}$ such that $d_j = 0$, degr $q_j = k$ and

$$\alpha = \frac{1}{k}\left(\frac{\pi}{2} - \arg \mu_{j,pk}\right) \quad \text{if } k \neq 1,$$

$$\alpha = \frac{\pi}{2} - \arg (\mu_{j,p} \bmod 2\pi i) \quad \text{if } k = 1.$$

Furthermore, we define: $\Sigma_0(A) = \varnothing$ and $\Sigma(A) = \bigcup_{k \in k(A)} \Sigma_k(A)$.

The elements of $\Sigma(A)$ are comparable to the Stokes directions in the theory of homogeneous linear differential equations.

DEFINITION 1.2.4. By $\theta(A)$ we shall denote the set of all real numbers θ with the property that there is a $j \in \{1, \cdots, m\}$ such that $d_j \neq 0$ and

$$d_j \theta = \operatorname{Im} \mu_{j,p} \bmod 2\pi.$$

The problem of transforming a given matrix function A into a canonical form $\overset{c}{A}$ is equivalent to that of finding a solution of the equation

$$Y(s+1) = A(s)\, Y(s) \overset{c}{A}(s)^{-1}$$

in some appropriate set of matrix functions Y. By $\sigma(A)$ we shall denote the matrix function corresponding to the linear mapping

$$Y \to A Y \overset{c}{A}^{-1}.$$

It is easily verified that $\sigma(\overset{c}{A})$ is a canonical form of $\sigma(A)$. Hence it follows, for example, that

$$d(\sigma(A)) = \{d - d' \colon d, d' \in d(A)\}.$$

It will often prove convenient to partition other matrices in the same way as a given canonical matrix $\overset{c}{A}$ (cf. (1.2.1)) associated with the particular problem under consideration. If M is an $n \times n$ matrix, the notation M_{ij} $(i, j \in \{1, \cdots, m\})$ will always refer to a block of M in that partition and not to a single matrix element.

2. Existence theorems.

2.1. Preliminaries. The following sections are mainly concerned with the existence of right inverses of linear difference operators Δ_A defined on Banach spaces of holomorphic functions of the type described in the definition below.

DEFINITION 2.1.1. Let $r \in \mathbb{R}$ and let G be a closed region of the complex plane. By $B_r(G)$ we denote the Banach space of all functions f with the following properties:

(i) f is continuous on G and holomorphic in int G;

(ii) $\|f\|_r \equiv \sup_{s \in G} |s^r f(s)| < \infty$.

If S is an asymptotic set of closed regions the following two statements are equivalent:

(i) $f \in \mathscr{A}_0(S)$;

(ii) There exists a positive number R such that $f \in B_r(S(R))$ for all $r \in \mathbb{R}$.

Let S be an asymptotic set of closed regions $S(R)$ with the additional property that $s \in S(R)$ implies $s+1 \in S(R)$ for all $R > 1$ and let $A \in \mathrm{Gl}(n; \mathscr{M}(S))$. It is easily seen that there exists a real number u such that the difference operator Δ_A maps $B_r(S(R))^n$ into $B_{r-u}(S(R))^n$ for all $r \in \mathbb{R}$ and all sufficiently large R.

DEFINITION 2.1.2. An asymptotic set of closed regions S is proper for the difference operator Δ if there exist real numbers r_0 and v, positive numbers R_0 and K, and linear mappings

$$\Lambda_{r,R} : B_r(S(R))^n \to B_{r-v}(S(R))^n$$

defined for all $r \geqq r_0$ and all $R \geqq R_0$ such that

(i) $\Delta\Lambda_{r,R} f = \Lambda_{r-u,R}\Delta f = f$ for all $f \in B_r(S(R))^n$,
(ii) $\|\Lambda_{r,R} f\|_{r-v} \leqq K\|f\|_r$,
(iii) If $r' > r \geqq r_0$, then $\Lambda_{r,R}|_{B_{r'}(S(R))^n} = \Lambda_{r',R}$.

2.2. The asymptotic sets S_θ. Let $C > 0$, $\theta \in \mathbb{R}$. We shall consider regions of the complex plane bounded by curves of the following type:

$$\sigma_C(\theta) = \{s \in \mathbb{C} : \operatorname{Re}(s \log s e^{i\theta}) = C\},$$

where $\log s$ has the principal value.

We begin by deriving some properties of these curves. First of all, note that a change from θ to $-\theta$ is equivalent to a reflection of $\sigma_C(\theta)$ with respect to the real axis. Therefore, we shall restrict the discussion to nonnegative values of θ.

If $|s| < 1$, then $\operatorname{Re} s \log|s| < 1/e$ and, consequently, $\operatorname{Re}(s \log s e^{i\theta}) < 1/e + \pi + \theta$. From now on we shall always assume that C is so large that $d(\sigma_C(\theta), 0) > 1$. We put

$$\operatorname{Im} s = x, \qquad \operatorname{Re} s = \rho.$$

On the set $V = \{(x, \rho) \in \mathbb{R}^2 : x \neq 0 \text{ if } \rho \leqq 0\}$ we define a function F by

$$F(x, \rho) = \rho \log\sqrt{(\rho^2 + x^2)} - x\{\arg(\rho + ix) + \theta\}.$$

By assumption, $F(x, \rho) < C$ for all (x, ρ) with the property that $\rho^2 + x^2 \leqq 1$. Furthermore, we have

$$D_\rho F(x, \rho) = 1 + \log\sqrt{(\rho^2 + x^2)}.$$

Consequently, $D_\rho F(x, \rho) > 1$ whenever $\rho^2 + x^2 \geqq 1$. Hence it follows that, for every $x \in \mathbb{R}$, there is a unique $\rho \in \mathbb{R}$ such that

$$F(x, \rho) = C. \tag{2.2.1}$$

Indicating this number by $\rho(x)$ and differentiating (2.2.1) with respect to x we obtain

$$\rho'(x) = \{\theta + \arg(\rho(x) + ix)\}\{1 + \log\sqrt{(\rho(x)^2 + x^2)}\}^{-1}, \qquad x \in (-\infty, \infty). \tag{2.2.2}$$

Obviously, the function $\rho(x)/x \log\sqrt{(\rho(x)^2 + x^2)}$ is bounded on $|x| > 1$ and hence

$$\rho(x) = O\left(\frac{x}{\log x}\right), \qquad |x| \to \infty. \tag{2.2.3}$$

For a more detailed description of the curve $\sigma_C(\theta)$ it is convenient to distinguish the following three cases.

Case 1. $0 \leqq \theta < \pi/2$. The corresponding class of curves (reflected with respect to the imaginary axis) has been studied in [7]. These curves are completely contained in the right halfplane $\rho > 0$. $\rho(x)$ has an absolute minimum which is attained when $\arg(\rho(x) + ix) = -\theta$. The symmetrical case ($\theta = 0$) is represented in Fig. 1.

Case 2. $\theta = \pi/2$. In this case $\rho'(x)$ is positive for all $x \in \mathbb{R}$, but tends to 0 as $x \to -\infty$. The curve $\sigma_C(\pi/2)$ is contained in the right halfplane $\rho > 0$ and is asymptotic to the negative imaginary axis. It is easily seen that $\rho(x) = O(1/\log x)$ as $x \to -\infty$ (see Fig. 2).

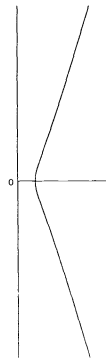


FIG. 1

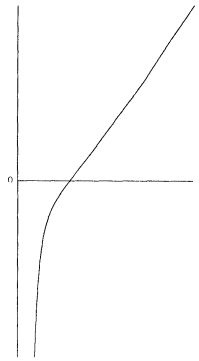


FIG. 2

Case 3. $\theta > \pi/2$. The curves in this class intersect the negative imaginary axis at $x = -C(\theta - \pi/2)^{-1}$. Furthermore, we have

$$\arg(\rho(x) + ix) + \theta > 0 \quad \text{for all } x \in \mathbb{R}$$

in this case, as can be seen from the following argument. Suppose that $\arg(\rho(x) + ix) + \theta < 0$ for some $x < 0$. This would imply that

$$\rho(x) \log \sqrt{(\rho(x)^2 + x^2)} = C + x\{\arg(\rho(x) + ix) + \theta\} > 0,$$

but this is in contradiction with the fact that $\arg(\rho(x) + ix) < -\theta < -\pi/2$ and, consequently, $\rho(x) < 0$. Hence, by (2.2.2), it follows that $\rho'(x) > 0$ for all $x \in \mathbb{R}$ (see Fig. 3).

Now let $\theta \in \mathbb{R}$ and $R > 1$. By $C(\theta, R)$ we shall denote the value of C such that $d(\sigma_C(\theta), 0) = R$.

DEFINITION 2.2.4. Let $\theta \in \mathbb{R}$. S will denote the asymptotic set of closed regions $\{S_\theta(R), R > 1\}$, where

$$S_\theta(R) = \{s \in \mathbb{C}: \operatorname{Re}(s \log s\, e^{i\theta}) \geqq C(\theta, R)\}.$$

Furthermore, we shall put

$$S(-\pi, 0] = S_\infty, \quad S[0, \pi) = S_{-\infty}, \quad S(0, \pi] = \tilde{S}_\infty, \quad S[-\pi, 0) = \tilde{S}_{-\infty}.$$

2.3. Proper asymptotic sets. Let $\theta \in \mathbb{R}$ and $\Delta = \Delta_A$, where $A \in \mathrm{Gl}(n, \mathcal{M}(S_\theta))$. In order to prove the existence of linear mappings $\Lambda_{r,R}$ defined on the Banach spaces $B_r(S_\theta(R))^n$ and possessing the properties mentioned in Definition 2.1.2, we proceed exactly as in [7, § 12]. Thus we obtain the following addition to Proposition 4.12 in [7].

PROPOSITION 2.3.1. *Let $S = \bigcup_{\theta \in [\alpha, \beta]} S_\theta$, where α and β are real numbers such that $\alpha \leqq \beta$. Let $A \in \mathrm{Gl}(n; \mathcal{M}(S))$ and assume that*

(i) $[\pi/2 - \pi/k, -\pi/2] \cap \Sigma_k(A) = \varnothing$ *for all* $k \in k(A)$ *such that* $k \neq 0$;

(ii) $[\alpha, \beta] \cap \theta(A) = \varnothing$.

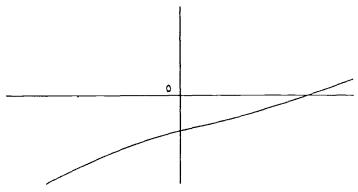


FIG. 3

Then S is proper for the difference operator Δ_A.

The next proposition is concerned with the cases $S=\bar{S}_\infty$ and $S=\bar{S}_{-\infty}$ (note that $\bar{S}_\infty$ is the set of lower halfplanes $\{s\in\mathbb{C}: \operatorname{Im} s\leqq -R\}$, $R>1$, whereas $\bar{S}_{-\infty}$ is a set of upper halfplanes). Although it shows some resemblance to Proposition 4.16 in [7], the assumptions made here are much more restrictive and therefore a stronger statement can be made.

PROPOSITION 2.3.2. *Let $S=\bar{S}_\infty$ or $S=\bar{S}_{-\infty}$ and $A\in \mathrm{Gl}(n;\mathcal{M}(S))$. Assume that there exists a positive number α less than 1 and a positive number R_0 such that, for all $s\in S(R_0)$ either of the following inequalities holds:*

$$|A(s)|\leqq\alpha \quad \text{or} \quad |A(s)^{-1}|\leqq\alpha.$$

Then S is proper for the difference operator Δ_A.

Proof. Without loss of generality we may assume that A is continuous on $S(R_0)$ and holomorphic in int $S(R_0)$. For every $R\geqq R_0$ and $r\geqq 0$ we define a linear mapping $\Lambda_{r,R}$ by either of the following expressions: for all $f\in B_r(S(R))^n$,

(i) $\Lambda_{r,R}f(s)=f(s-1)+\sum_{h=1}^{\infty}A(s-1)\cdots A(s-h)f(s-h-1)$ if the first inequality holds, and

(ii) $\Lambda_{r,R}f(s)=-\sum_{h=0}^{\infty}A(s)^{-1}\cdots A(s+h)^{-1}f(s+h)$ if the second inequality holds.

Consider the first case. For all $s\in S(R)$ we have

$$|s^r\Lambda_{r,R}f(s)|\leqq\sum_{h=0}^{\infty}\alpha^h\left|\frac{s}{s-h-1}\right|^r\sup_{s\in S(R)}|s^rf(s)|$$

hence

$$\|\Lambda_{r,R}f\|_r\leqq\sum_{h=0}^{\infty}\alpha^h\left(1+\frac{h+1}{R_0}\right)^r\|f\|_r.$$

With the aid of these estimates we readily verify that the mappings $\Lambda_{r,R}$ possess the required properties. The second case can be dealt with analogously.

2.4. Analytic simplification of linear difference operators. Propositions 2.3.1 and 2.3.2 can be used to derive existence theorems for solutions of nonlinear difference equations, admitting asymptotic power series expansions in appropriate regions of the complex plane. These, in their turn, can be applied to achieve analytic simplification of homogeneous linear difference systems. Thus, for example, the statements made in Theorems 15.16 (concerning a class of nonlinear equations) and 17.13 (on block diagonalization) of [7] can now be extended immediately to all proper asymptotic sets mentioned in Proposition 2.3.1. We shall not explicitly state the generalized versions of these theorems here, but refer the reader to [7].

One immediate consequence of Proposition 2.3.1 is the following theorem.

THEOREM 2.4.1. *Let $S=\bigcup_{\theta\in[\alpha,\beta]}S_\theta$, where α and β are real numbers such that $\alpha\leqq\beta$. Let $A\in\mathrm{Gl}(n;\mathcal{M}(S))$, let $\overset{c}{A}$ be a canonical form of A, and let $\Phi\in\mathrm{Gl}(n;\hat{K})$ such that $A^{\Phi}=\overset{c}{A}$. Furthermore, assume that*

(i) $[\pi/2-\pi/k,\pi/2]\cap\Sigma_k(\sigma(A))=\varnothing$ *for all $k\in k(\sigma(A))$ such that $k\neq 0$;*

(ii) $[\alpha,\beta]\cap\theta(\sigma(A))=\varnothing$.

Then there exists a unique matrix function $F\in\mathrm{Gl}(n;\mathcal{M}(S))$ such that $\hat{F}=\Phi$ and $A^F=\overset{c}{A}$.

With the aid of the above-mentioned extension of Theorem 17.13 in [7] the following result is obtained.

THEOREM 2.4.2. *Let $A\in\mathrm{Gl}(n;\mathcal{M}(S(-\pi,\pi)))$ and let $\overset{c}{A}$ be a canonical form of A.*

(i) *If* $-\pi/2 \notin \Sigma(\sigma(A))$ *there exists a matrix function* $F \in \mathrm{Gl}\,(n; \mathcal{M}(S[-\pi/2, \pi/2)))$ *such that* $A^F = \overset{c}{A}$.

(ii) *If* $\pi/2 \notin \Sigma(\sigma(A))$ *there exists a matrix function* $F \in \mathrm{Gl}\,(n; \mathcal{M}(S(-\pi/2, \pi/2])))$ *such that* $A^F = \overset{c}{A}$.

Remark. The condition $A \in \mathrm{Gl}\,(n; \mathcal{M}(S(-\pi, \pi)))$ may be replaced by $A \in \mathrm{Gl}\,(n; \mathcal{M}(S_\theta))$ *for some* θ *greater than* $\pi/2$ in case (i) or less than $-\pi/2$ in case (ii).

The proof of Theorem 2.4.2 is roughly analogous to that of Theorem 18.13 in [7]. The difference with the latter theorem consists in the fact that here the asymptotic expansion of the matrix function F is also valid as $s \to \infty$ in the direction of the negative (case (i)) or positive (case (ii)) imaginary axis. The asymptotic behaviour of F in the opposite direction can be deduced from the following lemma (by identifying F with a vector solution of the equation $\Delta_{\sigma(A)} y = 0$).

LEMMA 2.4.3. *Let* $\theta \in \mathbb{R}$ *and* $A \in \mathrm{Gl}\,(n; \mathcal{M}(S_\theta \cup S[-\pi/2, \pi/2]))$. *Assume that* f *is a solution of the equation* $\Delta_A y = 0$ *with the following properties*:

(i) f *is continuous on the set* $S_\theta(R) \cup S[-\pi/2, \pi/2](R)$ *and holomorphic in its interior, for some* $R > 0$.

(ii) f *grows at most exponentially of order* 1 *as* $s \to \infty$ *in* $S_\theta(R)$. *Then* f *grows at most exponentially of order* 1 *as* $s \to \infty$ *in* $S[-\pi/2, \pi/2](R)$ *provided* R *is sufficiently large.*

Proof. Let $s \in S[-\pi/2, \pi/2](R)$ and suppose that $|\mathrm{Im}\, s| > e$ and $s \notin S_\theta(R)$. Let $n(s)$ be the smallest integer such that $s + n(s) \in S_\theta(R)$. Suppose that R is so large that A^{-1} is continuous on the set $S_\theta(R) \cup S[-\pi/2, \pi/2](R)$ and holomorphic in its interior. Then we have

$$y(s) = A(s)^{-1} A(s+1)^{-1} \cdots A(s+n(s)-1)^{-1} y(s+n(s)).$$

By assumption, there exist positive numbers c and C such that, for all $\zeta \in S_\theta(R)$,

$$|y(\zeta)| \leqq C e^{c|\zeta|}.$$

Moreover, there exist positive constants d and D such that, for all $\zeta \in S_\theta(R) \cup S[-\pi/2, \pi/2](R)$,

$$|A(\zeta)^{-1}| \leqq D|\zeta|^d.$$

Hence it follows that

$$|y(s)| \leqq C D^{n(s)} |s + n(s)|^{dn(s)}\, e^{c|s+n(s)|}. \tag{2.4.4}$$

According to (2.2.3) there exists a positive constant K, independent of s, such that

$$\mathrm{Re}\,(s + n(s)) \leqq K \frac{|\mathrm{Im}\, s|}{\log |\mathrm{Im}\, s|}.$$

Since $|\mathrm{Im}\, s| > e$, this implies that

$$|s + n(s)| \leqq (K+1)|\mathrm{Im}\, s|.$$

Hence

$$n(s) \log |s + n(s)| \leqq K |\mathrm{Im}\, s| \{1 + \log (K+1)\}.$$

The proof is completed by inserting the last two estimates into (2.4.4).

The next two theorems are based on Proposition 2.3.2. They can be proved by the familiar method of successive block diagonalizing transformations. The existence (and uniqueness) of these transformations can be deduced from Proposition 2.3.2 in the usual manner.

THEOREM 2.4.5. *Let* $S = \bar{S}_\infty$ *or* $\bar{S}_{-\infty}$, *let* $A \in \mathrm{Gl}\,(n; \mathcal{M}(S))$, *and let* $\overset{c}{A}$ *be a canonical form of* A *of the form* (1.2.1). *Assume that, for all* $i, j \in \{1, \cdots, m\}$ *such that* $i \neq j$ *and* $d_i = d_j$,

$$\mathrm{degr}\, \mathrm{Re}\,(q_i - q_j) = 1.$$

Then there exists a matrix function $F \in \mathrm{Gl}(n; \mathcal{M}(S))$ *such that*

$$A^F(s) = \overset{c}{A}(s)(I_n + s^{-r}B(s))$$

where $r \in \mathbb{Q}$, $r > 1$, $B \in \mathrm{End}(n; \mathcal{A}(S))$, *and* $B = \mathrm{diag}\{B_{11}, \cdots, B_{mm}\}$. *Moreover, this matrix function* F *is determined uniquely by its asymptotic expansion.*

A complete reduction of A to the canonical form in general is possible only if we drop the requirement that the asymptotic expansion of F be valid as $s \to \infty$ in both horizontal directions. The following theorem will be needed in § 4.

THEOREM 2.4.6. *Let* $S \in \{S_\infty, S_{-\infty}, \tilde{S}_\infty, \tilde{S}_{-\infty}\}$. *Let* $A \in \mathrm{Gl}(n; \mathcal{M}(\bar{S}))$ *and let* $\overset{c}{A}$ *be a canonical form of* A. *Assume that the conditions of Theorem* 2.4.5 *are satisfied. For every* $\Phi \in \mathrm{Gl}(n; \hat{K})$ *such that* $A^\Phi = \overset{c}{A}$, *there exists a unique matrix function* $F \in \mathrm{Gl}(n; \mathcal{M}(S))$ *with the following properties*:

(i) $A^F = \overset{c}{A}$ *and* $\hat{F} = \Phi$;

(ii) *There exist positive numbers* k *and* R *such that both* $F(s) = O(s^k)$ *and* $F(s)^{-1} = O(s^k)$, $s \to \infty$, $s \in \bar{S}(R)$.

Remark. If A is upper or lower block triangular in the usual partition, then the same is true of F.

3. A result of Birkhoff and Trjitzinsky. Let $A \in \mathrm{Gl}(n; \mathbb{C}\{s^{-1}\}[s])$, or, more generally, $A \in \mathrm{Gl}(n; K)$, and let $\overset{c}{A}$ be a canonical form of A. As we mentioned in § 1.2, there exists a matrix function $\Phi \in \mathrm{Gl}(n; \hat{K})$ with the property that

$$A^\Phi = \overset{c}{A}.$$

This section deals with the following problem.

If α is any direction in the complex plane, does there exist a matrix function F, analytic in a sector containing a half-line with direction α, such that

$$A^F = \overset{c}{A} \quad \text{and} \quad \hat{F} = \Phi?$$

So far we have been able to answer this question in the affirmative for all directions except those of the positive and negative imaginary axis. In order to include the latter directions we had to impose a rather mild condition, viz. that either $d(\sigma(A)) = \{0\}$, or else that 0 or π or both $-\pi/2$ and $\pi/2$ do not belong to $\Sigma(\sigma(A))$ (cf. Theorem 2.4.2 above and Theorem 18.18 in [7]). It is the purpose of this section to remove this last restrictive condition.

DEFINITION 3.1. A quadrant Γ is an asymptotic set of closed regions $\Gamma(R)$ $(R > 1)$ of the following type:

$$\Gamma(R) = \left\{ s \in \mathbb{C} : \alpha \leqq \arg(s - s_0) \leqq \alpha + \frac{\pi}{2}, |s| \geqq R \right\},$$

where $s_0 \in \mathbb{C}$, $\alpha = l(\pi/2)$, $l \in \mathbb{Z}$.

THEOREM 3.2. *Let* $A \in \mathrm{Gl}(n; K)$ *and let* $\overset{c}{A}$ *be a canonical form of* A. *Let* Γ *be any quadrant. There exists a matrix function* $F \in \mathrm{Gl}(n; \mathcal{M}(\Gamma))$ *such that*

$$A^F = \overset{c}{A}.$$

As a matter of fact this theorem was proved by Birkhoff and Trjitzinsky in [1]. Unfortunately, their methods are not very transparent and the argument is very hard

to follow. Here we have tried to remedy certain inaccuracies contained in this paper and have sketched a considerably simplified version of their proof.

We shall consider the case that $\alpha = -\pi/2$. All other cases can be proved analogously.

The proof consists of two steps. The first step is to carry the matrix function A into a block-triangular form by a suitable transformation. In the paper by Birkhoff and Trjitzinsky this is achieved by a rather particular method which has been exposed in earlier papers by Birkhoff alone. It is quite different from the one we used in [7], but the results are essentially the same. These can be stated as follows (cf. Lemma 9 of [1]).

PROPOSITION 3.3. *Let* $\hat{S}_0 = S(-\pi, \pi)$ *and* $A \in \mathrm{Gl}(n; \mathscr{M}(\hat{S}_0))$. *Let* $\overset{c}{A}$ *be a canonical form of* A. *There exists a matrix function* $F_1 \in \mathrm{Gl}(n; \mathscr{M}(\hat{S}_0))$ *such that*

$$A^{F_1} = \overset{c}{A}(I_n + s^{-r}B(s)),$$

where $r > 1$, $B \in \mathrm{End}(n; \mathscr{A}(\hat{S}_0))$ *and* $B_{ij} \equiv 0$ *if* $i > j$, $i, j \in \{1, \cdots, m\}$.

For the proof of this proposition we refer the reader to Proposition 18.15 of [7].

The second and more delicate step is the final transformation of A^{F_1} into $\overset{c}{A}$. Put $A^{F_1} = A_1$. We now search a matrix function $F \in \mathrm{Gl}(n; \mathscr{A}(\Gamma))$ with the following properties:

(i) $A_1^F = \overset{c}{A}$;

(ii) $F_{ij} \equiv 0$ if $i > j$, $i, j \in \{1, \cdots, m\}$.

If $j \in \{1, \cdots, m\}$ and $i \leqq j$, the block F_{ij} must satisfy the following inhomogeneous difference equation:

$$Z(s+1) = (A_1)_{ii}(s)Z(s)\overset{c}{A}_j(s)^{-1} + \sum_{i<h\leqq j} (A_1)_{ih}(s)F_{hj}(s)\overset{c}{A}_j(s)^{-1}. \tag{3.4}$$

For every $j > 1$ there are j equations that we can solve successively, beginning with F_{jj}, then $F_{j-1\,j}$, etc. First, let us suppose that $d_i = d_j$ for some $i < j$. As $d_l \leqq d_j$ for all $l < j$, it follows that $d_l = d_j$ for all l such that $i \leqq l \leqq j$.

LEMMA 3.5. *Let* $i \leqq j$ *and assume that* $d_i = d_j$. *Then there exists a number* $\varepsilon \in (0, \pi/2)$ *and, for every* l *such that* $i \leqq l \leqq j$, *a matrix function* $F_{lj} \in \mathrm{Hom}(\mathscr{A}(\hat{S}_\varepsilon)^{n_j}, \mathscr{A}(\hat{S}_\varepsilon)^{n_l})$, *where* $\hat{S}_\varepsilon = S(-\pi + \varepsilon, \varepsilon)$, *satisfying* (3.4) *with* i *replaced by* l.

The statement above can easily be deduced from Theorem 15.1 of [7] by means of induction on $j - l$. If $d_i = d_j$ for all $i, j \in \{1, \cdots, m\}$ the assertion of Theorem 3.2 follows immediately. Now suppose that $d_i < d_j$ for some $j \in \{1, \cdots, m\}$ and some $i < j$. Consequently, $d_l < d_j$ for all $l \leqq i$. In that case the proof is completed by repeated application of Lemma 3.7 below.

DEFINITION 3.6. Let $\theta, x_0 \in \mathbb{R}$. By Γ_{θ,x_0} we shall denote the asymptotic set of closed regions $\Gamma_{\theta,x_0}(R)$ $(R > 1)$ defined by

$$\Gamma_{\theta,x_0}(R) = \{s \in S_\theta(R) : \mathrm{Im}\, s \leqq x_0\}.$$

LEMMA 3.7. *Let* $\theta, x_0 \in \mathbb{R}$, $B \in \mathrm{End}(n; \mathscr{A}(\Gamma_{\theta,x_0}))$, *and* $h \in (\mathscr{A}(\Gamma_{\theta,x_0}))^n$. *Let*

$$A(s) = \exp\{q(s+1) - q(s)\}(I_n + s^{-1}B(s)),$$

where $q(s) = ds \log s + \sum_{h=1}^{p} \mu_h s^{h/p}$, $p \in \mathbb{N}$, $d \in \mathbb{Z}/p$, $\mu_h \in \mathbb{C}$ *for all* $h \in \{1, \cdots, p\}$. *Suppose that* $d < 0$ *and* $d\theta \neq \mathrm{Im}\, \mu_p \bmod 2\pi$. *Then the equation*

$$\Delta_A y = h \tag{3.8}$$

possesses a solution $y \in \bigcap_{x<x_0} \mathscr{A}(\Gamma_{\theta,x})^n$.

Proof. The homogeneous equation $\Delta_A y = 0$ possesses a fundamental matrix Y of the form

$$Y(s) = Z(s)s^G e^{q(s)},$$

where G is a constant matrix and $Z \in \mathrm{Gl}(n; \mathscr{A}(\Gamma_{\theta,x_0}))$. Since, by assumption, $\theta \neq 1/d$ ($\mathrm{Im}\ \mu_p \bmod 2\pi$) and as μ_p obviously is determined modulo $2\pi i$ by A, we may assume that

$$d\theta < \mathrm{Im}\ \mu_p < d\theta + 2\pi.$$

Let R be a positive number such that Z is holomorphic in $\mathrm{int}\,\Gamma_{\theta,x_0}(R)$ and represented asymptotically by $\hat{Z}$ as $s \to \infty$ in $\Gamma_{\theta,x_0}(R)$. Let s_0 denote the point on the boundary of $S_\theta(R)$ with the property that $\mathrm{Im}\ s_0 = x_0$. Consider the linear mapping Λ of $B_0(\Gamma_{\theta,x_0}(R))$ (cf. Definition 2.1.1) defined by the following formula:

$$\Lambda f(s) = Y(s) \int_{C(s)} d\zeta \frac{Y(\zeta+1)^{-1} f(\zeta)}{1-\exp\{2\pi i(s-\zeta)\}} + A(s)^{-1} f(s),$$

where $f \in B_0(\Gamma_{\theta,x_0}(R))$, $s \in \Gamma_{\theta,x_0}(R)$ such that $\mathrm{Im}\ s < x_0$ and $C(s)$ is a path going from s_0 to infinity in such a way that it intersects the line $\mathrm{Im}\ \zeta = \mathrm{Im}\ s$ exactly once, in a point between s and $s+1$.

It can easily be verified that the vector function Λh is a solution of the equation (3.8). In order to prove that it has the desired asymptotic properties, we shall show that, for all $r > 0$ and all $x < x_0$, Λ maps the Banach space $B_r(\Gamma_{\theta,x_0}(R))$ into $B_{r+v}(\Gamma_{\theta,x}(R))$, where v is some fixed real number.

For all $s \in \Gamma_{\theta,x_0}(R)$ let s' denote the point on the boundary of Γ_{θ,x_0} with the property that

$$\mathrm{Re}\,(s' \log s' e^{i\theta}) = \mathrm{Re}\,\{(s+\tfrac{1}{2}) \log (s+\tfrac{1}{2})\, e^{i\theta}\}.$$

Let $C_1(s)$ be the path from ∞ to s' such that

$$\mathrm{Re}\,(\zeta \log \zeta e^{i\theta}) = \mathrm{Re}\,(s' \log s'\, e^{i\theta}), \qquad \zeta \in C_1(s)$$

and let $C_2(s)$ denote the directed line segment from s' to s_0 (see Fig. 4). Let $r > 0$, $x < x_0$, and $f \in B_r(\Gamma_{\theta,x}(R))$. Putting

$$Y(s) \int_{C_i(s)} d\zeta \frac{Y(\zeta+1)^{-1} f(\zeta)}{1-\exp\{2\pi i(s-\zeta)\}} = I_i(s), \qquad i = 1, 2$$

we have

$$\Lambda f(s) = I_1(s) + I_2(s) + A(s)^{-1} f(s). \tag{3.9}$$

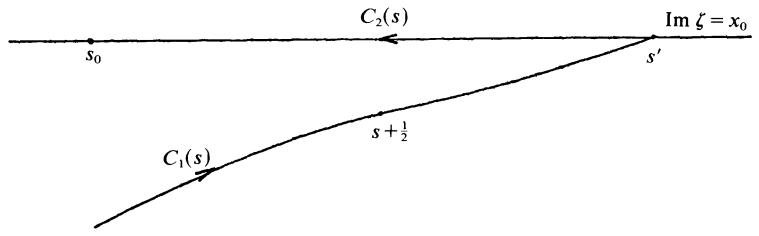


FIG. 4

The second term may be written as follows:

$$I_2(s) = Y(s)\,Y(s')^{-1} \exp\{2\pi i(s'-s)\} \exp\{2\pi i(s-s')\} I_2(s'). \tag{3.10}$$

Noting that, for all $s \in \Gamma_{\theta,x}(R)$ and all $\zeta \in C_2(s)$,

$$\left| \frac{\exp\{2\pi i(s-s')\}}{1-\exp\{2\pi i(s-\zeta)\}} \right| = |\exp\{-2\pi i(s-\zeta)\}-1|^{-1} \leqq [1-\exp\{2\pi(x-x_0)\}]^{-1}$$

and taking into account the rapid decrease of $Y(\zeta+1)^{-1}$ along $C_2(s)$ we readily verify that

$$\sup_{s \in \Gamma_{\theta,x}(R)} |(s')^{r+d} \exp\{2\pi i(s-s')\} I_2(s')| < \infty.$$

The term $I_1(s)$ in (3.9) and the product

$$Y(s)\,Y(s')^{-1} \exp\{2\pi i(s'-s)\}$$

in (3.10) can be dealt with by the methods used in [7, § 12]. Indeed, the integral $I_1(s)$ is similar to the one figuring in (12.2) of [7], where the above product can be estimated in much the same way as the integrand of $I_+(s)$ defined on p. 71 of [7]. Thus we find that $\Lambda f \in B_{r+d}(\Gamma_{\theta,x}(R))$, provided R is sufficiently large. Hence it follows that $\Lambda h \in \mathscr{A}(\Gamma_{\theta,x})^n$ for all $x < x_0$. This concludes the proof of the lemma.

Remark. With the aid of Lemma 3.7 we can prove a slightly stronger statement than the one made in Theorem 3.2, namely, the existence, for all $x \in \mathbb{R}$ and all $\theta \notin \theta(\sigma(A))$, of a matrix function $F \in \mathrm{Gl}\,(n; \mathscr{M}(\Gamma_{\theta,x}))$ with the property that $A^F = \overset{c}{A}$.

4. The Stokes phenomenon.

4.1. A preliminary transformation. In the remaining sections we shall determine the "maximal asymptotic sets" for and study the connection between different fundamental matrix solutions of the linear homogeneous difference equation

$$\Delta_A y = 0 \tag{4.1.1}$$

where $A \in \mathrm{Gl}\,(n; \mathbb{C}\{s^{-1}\}[s])$.

In order to avoid the complications caused by the intermingling of different types of Stokes phenomena (associated with different levels) we shall make the simplifying assumption that the set $d(A)$ defined in § 1.2 has m distinct elements.

Let $\overset{c}{A}$ be a canonical form of A and let U denote the asymptotic set of closed regions $U(R)(R>1)$, defined by (0.2). According to Theorem 18.16 in [7] there exists a matrix function $T \in \mathrm{Gl}\,(n; \mathscr{M}(U))$ such that

$$A^T(s) = \overset{c}{A}(s)(I_n + s^{-r}B(s)),$$

where $r>1$, $B \in \mathrm{End}\,(n; \mathscr{A}(U))$ and $B_{ij} \equiv 0$ if $i>j$, $i, j \in \{1, \cdots, m\}$. (Actually, Theorem 18.16 only states the existence of matrix functions $T_1 \in \mathrm{Gl}\,(n; \mathscr{M}(S[-\pi, \pi))$ and $T_2 \in \mathrm{Gl}\,(n; \mathscr{M}(S(-\pi, \pi]))$ with analogous properties, but these can be seen to coincide in some sector, provided $\hat{T}_1 = \hat{T}_2$.) Let S be an asymptotic set of closed regions with the property that $S(R) \subset U(R)$ for all $R>1$. In the following sections we shall consider fundamental matrix solutions of (4.1.1) of the form

$$Y = TF\overset{c}{Y},$$

where F is a solution of the equation

$$Z(s+1) = A^T(s)Z(s)\overset{c}{A}(s)^{-1} \tag{4.1.2}$$

with the properties that $F \in \mathrm{Gl}(n; \mathscr{A}(S))$, $F(\infty) = I_n$, and $F_{ij} \equiv 0$ if $i > j$, $i, j \in \{1, \cdots, m\}$.

4.2. Connection matrices. Let G_1 and G_2 be open regions in $U(1)$ such that $G_1 \cap G_2 \neq \emptyset$. Let F_1 and F_2 be solutions of (4.1.2), holomorphic in G_1 and G_2, respectively.

DEFINITION 4.2.1. The connection matrix of the pair (F_1, F_2) is the matrix function P defined by the expression

$$P = \overset{c}{Y}^{-1} F_1^{-1} F_2 \overset{c}{Y}.$$

Obviously, P is a periodic matrix function of period 1. If both F_1 and F_2 are upper block-triangular in the usual partition, then so is P.

In what follows we shall always be concerned with the case that one of the two regions G_1 and G_2 is a lower or an upper halfplane. It can easily be verified that (4.1.2) possesses a unique formal solution $\sum_{h=0}^{\infty} F_h s^{-h/p} (p \in \mathbb{N})$, with the property that $F_0 = I_n$. Hence, according to Theorem 2.4.6, there exist four unique matrix functions

$$F^{\infty} \in \mathrm{Gl}(n; \mathscr{A}(S_{\infty})), \qquad F^{-\infty} \in \mathrm{Gl}(n; \mathscr{A}(S_{-\infty})),$$
$$\tilde{F}^{\infty} \in \mathrm{Gl}(n; \mathscr{A}(\tilde{S}_{\infty})), \qquad \tilde{F}^{-\infty} \in \mathrm{Gl}(n; \mathscr{A}(\tilde{S}_{-\infty})),$$

with the properties mentioned in the theorem. Moreover, all four matrix functions are upper block triangular (cf. the remark below Theorem 2.4.6). Let P^{∞} and $P^{-\infty}$ denote the connection matrices of the pairs $(F^{\infty}, \tilde{F}^{-\infty})$ and $(F^{-\infty}, \tilde{F}^{\infty})$, respectively. Obviously, these too are upper block triangular. Let $\overset{c}{A}$ have the form (1.2.1). For all $i, j \in \{1, \cdots, m\}$ we have

$$P_{ij}^{\infty}(s) = \exp(q_j(s) - q_i(s)) s^{-G_i} (F^{\infty}(s)^{-1} \tilde{F}^{-\infty}(s))_{ij} s^{G_j}.$$

If $i < j$, due to the fact that $d_i < d_j$, the first factor on the right-hand side of this identity decreases very rapidly as $\mathrm{Re}\, s \to -\infty$. In view of the growth properties of F^{∞} and $\tilde{F}^{-\infty}$ this implies that P_{ij}^{∞} tends to zero as $\mathrm{Re}\, s \to -\infty$ and hence must vanish identically.

Similarly, it can be seen that $P_{ij}^{-\infty} \equiv 0$ if $i \neq j$. Now consider the diagonal blocks F_{ii}^{∞} and $\tilde{F}_{ii}^{-\infty}$, $i \in \{1, \cdots, m\}$. Both satisfy the equation

$$Z(s+1) = A_{ii}^{T}(s) Z(s) \overset{c}{A}_i(s)^{-1}$$
$$= \left(1 + \frac{1}{s}\right)^{N_i} (I_{n_i} + s^{-r} B_{ii}(s)) Z(s) \left(1 + \frac{1}{s}\right)^{-N_i}.$$

This equation has a solution $F_i \in \mathrm{Gl}(n_i; \mathscr{A}(S(-\pi, \pi)))$ represented by the infinite product

$$F_i(s) = s^{N_i} \prod_{n=0}^{\infty} (I_{n_i} + B_i(s+n))^{-1} s^{-N_i}, \qquad s \in U(R_0) \tag{4.2.2}$$

and a solution $\tilde{F}_i \in \mathrm{Gl}(n_i; \mathscr{A}(\tilde{S}_{\infty}) \cap \mathscr{A}(\tilde{S}_{-\infty}))$ represented by

$$\tilde{F}_i(s) = s^{N_i} \prod_{n=1}^{\infty} (I_{n_i} + B_i(s-n)) s^{-N_i}, \qquad s \in \bar{S}_{\infty}(R_0) \cup \bar{S}_{-\infty}(R_0),$$

where $B_i(s) = s^{-r} s^{-N_i} B_{ii}(s) s^{N_i}$ and R_0 is some sufficiently large positive number. It is easily seen that

$$\lim_{\mathrm{Res} \to \infty} F_i(s) = \lim_{\mathrm{Res} \to -\infty} \tilde{F}_i(s) = I_{n_i}.$$

Furthermore, if R_0 is sufficiently large, F_i and $\tilde{F}_i$ as well as their inverses are bounded on $\bar{S}_\infty(R_0) \cup \bar{S}_{-\infty}(R_0)$. Hence, by Theorem 2.4.6, F_i must coincide with F_{ii}^∞ in a lower halfplane and with $F_{ii}^{-\infty}$ in an upper halfplane, whereas $\tilde{F}_i$ must coincide with $\tilde{F}_{ii}^{-\infty}$ in a lower halfplane and with $\tilde{F}_{ii}^\infty$ in an upper halfplane. Consequently, the connection matrices P_{ii}^∞ and $P_{ii}^{-\infty}$ can both be represented by the infinite product

$$\prod_{n=-\infty}^{\infty} (I_{n_i} + B_i(s-n)), \qquad i \in \{1, \cdots, m\}$$

in a lower and an upper halfplane, respectively. We shall assume that R_0 is so large that F^∞ and $(F^\infty)^{-1}$ are bounded on the (closed) halfplane $\bar{S}_\infty(R_0)$, and, moreover, $F^{-\infty}$ and its inverse have the same properties with respect to $\bar{S}_{-\infty}(R_0)$.

Now let G be an open region in $U(R_0)$ such that

$$G = \bigcup_{x \in \mathbb{R}^+} G + x,$$

and $\sup_{s \in G} \operatorname{Im} s = -\inf_{s \in G} \operatorname{Im} s = \infty$.

Let F be a solution of (4.1.2), holomorphic in G and with diagonal blocks $F_{ii} = F_i$, where F_i is defined by (4.2.2), for all $i \in \{1, \cdots, m\}$. F may be continued analytically to a holomorphic function in

$$\bigcup_{x \in \mathbb{R}} G + x \cap U(R_0).$$

This function will again be denoted by F.

In the following lemma we consider the connection matrices of (F^∞, F) and $(F^{-\infty}, F)$.

LEMMA 4.2.3. *Let* $Y_i = F_i \overset{c}{Y}_i$, $i \in \{1, \cdots, m\}$ *and let* P *denote the matrix function defined by*

$$P(s) = \begin{cases} \overset{c}{Y}(s)^{-1} F^\infty(s)^{-1} F(s) \overset{c}{Y}(s) & \text{if } s \in \bar{S}_\infty(R_0), \\ \overset{c}{Y}(s)^{-1} F^{-\infty}(s)^{-1} F(s) \overset{c}{Y}(s) & \text{if } s \in \bar{S}_{-\infty}(R_0). \end{cases}$$

For all $i, j \in \{1, \cdots, m\}$ *and all* $s \in \bar{S}_\infty(R_0) \cup \bar{S}_{-\infty}(R_0)$ *the following identity holds*:

$$P_{ij}(s) = \lim_{n \to \infty} Y_i(s-n)^{-1} F_{ij}(s-n) \overset{c}{Y}_j(s-n).$$

Proof. We shall prove the statement for all $s \in \bar{S}_\infty(R_0)$, by means of induction on $j - i$. If $j - i = 1$ we have, for all $s \in \bar{S}_\infty(R_0)$ and all $n \in \mathbb{Z}$,

$$P_{ij}(s) = Y_i(s-n)^{-1} F_{ij}(s-n) \overset{c}{Y}_j(s-n) + \overset{c}{Y}_i(s-n)^{-1} (F^\infty)_{ij}^{-1}(s-n) Y_j(s-n).$$

Due to the fact that $d_i < d_j$ while $(F^\infty)^{-1}$ is bounded on $\bar{S}_\infty(R_0)$, the second term on the right-hand side of this identity tends to zero as $n \to \infty$.

Now suppose that $j - i > 1$ and that the statement is true for all pairs of indices (k, l) such that $l - k < j - i$. Then we have

$$\begin{aligned} &P_{ij}(s) - Y_i(s-n)^{-1} F_{ij}(s-n) \overset{c}{Y}_j(s-n) \\ &\quad = \sum_{i < h \leqq j} \overset{c}{Y}_i(s-n)^{-1} (F^\infty)_{ih}^{-1}(s-n) Y_h(s-n) Y_h(s-n)^{-1} F_{hj}(s-n) \overset{c}{Y}_j(s-n). \end{aligned}$$

Again the product $\overset{c}{Y}_i(s-n)^{-1} (F^\infty)_{ih}^{-1}(s-n)^{-1} Y_h(s-n)$ tends to zero as $n \to \infty$ for all $h > i$. By assumption, for all $h < j$ the product $Y_h(s-n)^{-1} F_{hj}(s-n) \overset{c}{Y}_j(s-n)$ tends to a finite limit, namely, $P_{hj}(s)$, as $n \to \infty$. Consequently, the right-hand side of the above

identity tends to zero as $n \to \infty$ and the result follows. For $s \in \bar{S}_{-\infty}(R_0)$ the proof is analogous.

4.3. Maximal asymptotic sets. Let $\theta_0 = \max\{\theta \in \theta(\sigma(A)) : \theta \leqq 0\}$ and let all elements of $\theta(\sigma(A))$ be numbered in such a way that $\theta_N < \theta_{N+1}$, $N \in \mathbb{Z}$. For all $N \in \mathbb{Z}$ and all $i, j \in \{1, \cdots, m\}$, $n_{ij}(N)$ will denote the smallest integer not less than

$$\frac{1}{2\pi}\{(d_i - d_j)\theta_N - \operatorname{Im}(\mu_{i,p} - \mu_{j,p})\}.$$

By the definition of $\theta(\sigma(A))$, there exists an integer M such that

$$n_{ij}(N) = \frac{1}{2\pi}\{(d_i - d_j)\theta_M - \operatorname{Im}(\mu_{i,p} - \mu_{j,p})\}$$

and an integer M' such that

$$n_{ij}(N) + 1 = \frac{1}{2\pi}\{(d_i - d_j)\theta_{M'} - \operatorname{Im}(\mu_{i,p} - \mu_{j,p})\}.$$

Suppose that $i < j$. This implies that $d_i < d_j$. Then, obviously $M' < M \leqq N$. Hence it follows that

$$0 \leqq n_{ij}(N-1) - n_{ij}(N) \leqq 1, \qquad i, j \in \{1, \cdots, m\}, \quad i < j, \quad N \in \mathbb{Z}. \tag{4.3.1}$$

For all $N \in \mathbb{Z}$ we define an upper block-triangular matrix function F_N by means of the following recursive relation for the blocks $(F_N)_{ij}$:

$$(F_N)_{ii} = F_i \quad \text{(defined in (4.2.2))},$$

$$(F_N)_{ij}(s) = Y_i(s) \int_{C(s)} d\zeta \frac{\exp\{2n_{ij}(N)\pi i(s-\zeta)\}}{1 - \exp\{2\pi i(s-\zeta)\}} I_{ij}^N(\zeta) \overset{c}{Y}_j(s)^{-1}, \qquad i < j, \tag{4.3.2}$$

$s \in \operatorname{int} U(R_0)$, where

$$I_{ij}^N(\zeta) = Y_i(\zeta+1)^{-1} \sum_{i < h \leqq j} A_{ih}^T(\zeta)(F_N)_{hj}(\zeta) \overset{c}{Y}_j(\zeta) \tag{4.3.3}$$

and $C(s)$ is a contour in $U(R_0)$, enclosing the negative imaginary axis as well as the points $s - n$, $n \in \mathbb{N}$, but not s (see Fig. 5).

LEMMA 4.3.4. *Let R and R' be positive numbers such that $R' > R > R_0$. Let $U_0 = U(R) \setminus U(R')$. There exist positive constants c and C such that, for all $i, j \in \{1, \cdots, m\}$ and all $N \in \mathbb{Z}$,*

$$\sup_{s \in U_0} |Y_i(s)^{-1}(F_N)_{ij}(s) \overset{c}{Y}_j(s)| \leqq C e^{c|N|}.$$

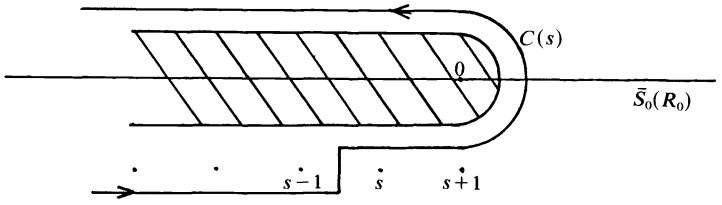


FIG. 5

Proof. For all $k \in \{1, \cdots, m\}$ we define $U_k = U(R - k\varepsilon) \setminus U(R' + k\varepsilon)$, where ε is a sufficiently small positive number such that $U_m \subset U(R_0)$. For all $(s, \zeta) \in U_m \times U_m$ we have

$$|\exp\{2n_{ij}(N)\pi i(s-\zeta)\}| \leqq \exp\{4(R' + m\varepsilon)\pi |n_{ij}(N)|\}.$$

Hence, in view of (4.3.1), we obtain the inequality

$$|\exp\{2n_{ij}(N)\pi i(s-\zeta)\}| \leqq C_0\, e^{c_0|N|}, \tag{4.3.5}$$

where c_0 and C_0 are positive numbers independent of N.

If $s \in U_k$ for some $k \in \{1, \cdots, m-1\}$ the contour $C(s)$ in (4.3.2) may be chosen in such a way that $C(s) \subset U_{k+1}$ and $d(\zeta, s + \mathbb{Z}) \geqq \varepsilon$ for all $\zeta \in C(s)$. Then there exists a positive number K such that

$$|1 - \exp\{2\pi i(s-\zeta)\}|^{-1} \leqq K \quad \text{for all } s \in U_{m-1} \text{ and all } \zeta \in C(s). \tag{4.3.6}$$

We shall prove the lemma by means of induction on $j - i$. If $j - i = 1$ the expression in (4.3.3) is reduced to

$$I_{ij}^N(\zeta) = Y_i(\zeta + 1)^{-1} A_{ij}^T(\zeta)\, Y_j(\zeta).$$

Due to the fact that $d_i < d_j$, this function tends to zero very rapidly as $\operatorname{Re}\zeta \to -\infty$, uniformly on U_m. Consequently, the integrals

$$\int_{C(s)} |I_{ij}^N(\zeta)||d\zeta|, \qquad s \in U_{m-1}$$

exist and are bounded by a constant, independent of N. With (4.3.5) and (4.3.6) it follows that, for all $N \in \mathbb{Z}$,

$$\sup_{s \in U_{m-1}} |Y_i(s)^{-1}(F_N)_{ij}(s)\overset{c}{Y}_j(s)| \leqq C_1\, e^{c_1|N|},$$

where c_1 and C_1 are positive constants independent of N.

Now let $j - i = k > 1$ and suppose that for all $l < k$ there exist positive numbers c_l and C_l such that

$$\sup_{s \in U_{m-l}} |Y_g(s)^{-1}(F_N)_{gh}(s)\overset{c}{Y}_h(s)| \leqq C_l\, e^{c_l|N|} \tag{4.3.7}$$

for all $N \in \mathbb{Z}$ and all $h, g \in \{1, \cdots, m\}$ such that $h - g = l$. Then we have, for all $N \in \mathbb{Z}$, all $h \in \{1, \cdots, m\}$ such that $i < h < j$ and all $s \in U_{m-j+h}$,

$$|Y_i(s+1)^{-1}A_{ih}^T(s)(F_N)_{hj}(s)\overset{c}{Y}_j(s)| \leqq |Y_i(s+1)^{-1}A_{ih}^T(s)\, Y_h(s)|C_{j-h}\, e^{c_{j-h}|N|}.$$

Inserting this into (4.3.3) and using the estimates (4.3.5) and (4.3.6) we conclude that there exist positive numbers c_k and C_k such that (4.3.7) holds for $l = k$ as well. Hence it follows that for all $l \in \{1, \cdots, m-1\}$ there exist positive numbers c_l and C_l such that (4.3.7) holds. Since $U_0 \subset U_l$ for all $l \in \{1, \cdots, m-1\}$ this proves the lemma.

With the aid of residue calculus it is readily verified that, for all $i, j \in \{1, \cdots, m\}$ such that $i < j$ and all $N \in \mathbb{Z}$, the matrix function $(F_N)_{ij}$ satisfies the equation

$$Z(s+1) = A_{ii}^T(s)Z(s)\overset{c}{A}_j(s)^{-1} + \sum_{i < h \leqq j} A_{ih}^T(s)(F_N)_{hj}(s)\overset{c}{A}_j(s)^{-1}.$$

Hence it follows that, for all $N \in \mathbb{Z}$, F_N is a solution of (4.1.2).

THEOREM 4.3.8. *For all $N \in \mathbb{Z}$ let P_N^+ and P_N^- denote the connection matrices of (F^∞, F_N) and $(F^{-\infty}, F_N)$, respectively. Let $R_1 > R_0$. There exist positive numbers d and D such that, for all $j \in \{1, \cdots, m\}$, all $i < j$ and all $N \in \mathbb{Z}$, the following inequalities hold:*

$$\sup_{s \in \bar{S}_\infty(R_1)} |(P_N^+)_{ij}(s) \exp\{-2(n_{ij}(N)-1)\pi i s\}| \leqq D\, e^{d|N|},$$

$$\sup_{s \in \bar{S}_{-\infty}(R_1)} |(P_N^-)_{ij}(s) \exp\{-2n_{ij}(N)\pi i s\}| \leqq D\, e^{d|N|}.$$

COROLLARY 4.3.9. *There exists a positive number R such that*

$$\lim_{N \to \infty} F_N(s) = F^\infty(s) \quad \text{if } \operatorname{Im} s \leqq -R,$$

$$\lim_{N \to -\infty} F_N(s) = F^{-\infty}(s) \quad \text{if } \operatorname{Im} s \geqq R.$$

Proof. For all $N \in \mathbb{Z}$, all $i, j \in \{1, \cdots, m\}$ such that $i < j$ and all $s \in \bar{S}_\infty(R_0)$ we have

$$(F_N)_{ij}(s) = F_{ij}^\infty(s) + \sum_{i \leqq h < j} F_{ih}^\infty(s) \overset{c}{Y}_h(s) (P_N^+)_{hj}(s) \overset{c}{Y}_j(s)^{-1}. \tag{4.3.10}$$

Applying Theorem 4.3.8 and noting that, according to (4.3.1),

$$n_{hj}(0) - n_{hj}(N) \leqq N \quad \text{for all } h < j, \quad N \in \mathbb{Z},$$

we obtain the following inequality for all $s \in \bar{S}_\infty(R_1)$:

$$|(F_N)_{ij}(s) - F_{ij}^\infty(s)| \leqq \sum_{i \leqq h < j} |F_{ih}^\infty(s) \overset{c}{Y}_h(s)|\, |\overset{c}{Y}_j(s)^{-1}| D_{hj}(s) \exp\{N(d - 2\pi \operatorname{Im} s)\},$$

where $D_{hj}(s) = D \exp\{2\pi(n_{hj}(0)+1) \operatorname{Im} s\}$. If we take $R > \max(R_1, d/2\pi)$ the first statement follows immediately. The second statement is proved analogously.

Proof of Theorem 4.3.8. Let C_0 be a U-shaped contour in the interior of $U(R_1) \setminus U(R_0)$ enclosing the negative imaginary axis. By means of residue calculus it is easily shown that

$$Y_i(s)^{-1} (F_N)_{ij} \overset{c}{Y}_j(s)^{-1} = (P_N)_{ij}(s) + \sum_{l=1}^{\infty} I_{ij}^N(s-l),$$

where I_{ij}^N is defined by (4.3.3) and

$$(P_N)_{ij}(s) = \int_{C_0} d\zeta \frac{\exp\{2n_{ij}(N)\pi i(s-\zeta)\}}{1 - \exp\{2\pi i(s-\zeta)\}} I_{ij}^N(\zeta),$$

for all $N \in \mathbb{Z}$, all $i, j \in \{1, \cdots, m\}$ such that $i < j$ and all s with the property that $|\operatorname{Im} s| \geqq R_1$.

Consequently, the following identity holds for all $n \in \mathbb{N}$:

$$Y_i(s-n)^{-1} (F_N)_{ij}(s-n) \overset{c}{Y}_j(s-n) = (P_N)_{ij}(s) + \sum_{l=n}^{\infty} I_{ij}^N(s-l).$$

Making $n \to \infty$ and applying Lemma 4.2.3 we find

$$(P_N)_{ij}(s) = \begin{cases} (P_N^+)_{ij}(s) & \text{if } \operatorname{Im} s \leqq -R_1, \\ (P_N^-)_{ij}(s) & \text{if } \operatorname{Im} s \geqq R_1. \end{cases}$$

Hence it follows that

$$\exp\{-2(n_{ij}(N)-1)\pi i s\}(P_N^+)_{ij}(s) = \int_{C_0} d\zeta \frac{\exp\{-2(n_{ij}(N)-1)\pi i \zeta\}}{\exp\{-2\pi i(s-\zeta)\} - 1} I_{ij}^N(\zeta) \tag{4.3.11}$$

for all $s \in \bar{S}_{\infty}(R_1)$, whereas

$$\exp\{-2n_{ij}(N)\pi i s\}(P_N^-)_{ij}(s) = \int_{C_0} d\zeta \frac{\exp\{-2n_{ij}(N)\pi i\zeta\}}{1-\exp\{2\pi i(s-\zeta)\}} I_{ij}^N(\zeta) \tag{4.3.12}$$

for all $s \in \bar{S}_{-\infty}(R_1)$. Writing

$$I_{ij}^N(\zeta) = \sum_{i<h\leqq j} Y_i(\zeta+1)^{-1} A_{ih}^T(\zeta)\, Y_h(\zeta)\, Y_h(\zeta)^{-1} (F_N)_{hj}(\zeta) \overset{c}{Y}_j(\zeta)$$

and using Lemma 4.3.4 we obtain, for all $\zeta \in C_0$, the estimate

$$I_{ij}^N(\zeta) \leqq C\, e^{c|N|} \sum_{i<h\leqq j} |Y_i(\zeta+1)^{-1} A_{ih}^T(\zeta)\, Y_h(\zeta)|,$$

where c and C are positive constants. As $d_i < d_h$ if $i<h$, the integral

$$\int_{C_0} |d\zeta|\, |Y_i(\zeta+1)^{-1} A_{ih}^T(\zeta)\, Y_h(\zeta)|$$

is convergent for all $i<h$. Furthermore, the functions

$$\varphi(s,\zeta) = [\exp\{-2\pi i(s-\zeta)\}-1]^{-1}$$

and

$$\psi(s,\zeta) = [1-\exp\{2\pi i(s-\zeta)\}]^{-1}$$

are obviously bounded on $\bar{S}_{\infty}(R_1)\times C_0$ and $\bar{S}_{-\infty}(R_1)\times C_0$, respectively. With (4.3.1) it is now easily verified that the expressions in (4.3.11) and (4.3.12) lead to estimates of the required form.

Remark. The matrix elements of P_N^{-1} (i.e., the connection matrix of (F_N, F^{∞}) and $(F_N, F^{-\infty})$) can be computed from F_N by means of the following recursive relation:

$$(P_N^{-1})_{ij} = \delta_{ij} \quad \text{if } i \geqq j,$$

$$= \delta_{ij} - \sum_{i<l\leqq j} \int_{C_0} d\zeta \frac{\exp\{2n_{il}(N)\pi i(s-\zeta)\}}{1-\exp\{2\pi i(s-\zeta)\}} I_{il}^N(\zeta)(P_N^{-1})_{lj} \quad \text{if } i<j$$

where I_{il}^N is defined by (4.3.3).

The next, and final, proposition is concerned with the asymptotic behaviour of the matrix functions F_N.

PROPOSITION 4.3.13. *Let $N \in \mathbb{Z}$ and $S_N^* = \cup_{\theta_N<\theta<\theta_{N+1}} S_\theta$. We have*

$$F_N \in \mathrm{Gl}(n; \mathscr{A}(S_N^*)).$$

Moreover, F_N is the unique solution of (4.1.2) *which is analytic in a right half plane and possesses the properties mentioned in Theorem* 4.3.8.

Proof. Let $\theta \in (\theta_N, \theta_{N+1})$ and let

$$S_\theta^- = S_\theta \cap \bar{S}_{\infty}.$$

Putting

$$d_h - d_j = d_{hj}, \qquad \mu_h - \mu_j + 2(n_{hj}(N)-1)\pi i = \mu_{hj}^N$$

for all $h, j \in \{1, \cdots, m\}$ and all $N \in \mathbb{Z}$, and using Theorem 4.3.8 we find

$$|\overset{c}{Y}_h(s)(P_N^+)_{hj}(s)\overset{c}{Y}_j(s)^{-1}| = \exp\{(d_{hj}s\log s + \mu_{hj}^N s)(1+o(1))\}, \quad s\to\infty, \quad s \in S_\theta^-(R_1),$$

for all $j \in \{1, \cdots, m\}$ and all $h<j$. By Definition 2.2.4,

$$\mathrm{Re}\, s\log s\, e^{i\theta} \geqq C(\theta, R_1)$$

for all $s \in S_\theta(R_1)$, hence

$$\operatorname{Re}(d_{hj} s \log s + \mu_{hj}^N s) \leqq d_{hj} C(\theta, R_1) + \operatorname{Re}\{(\mu_{hj}^N - i d_{hj}\theta) s\}$$

if $h < j$ and $s \in S_\theta(R_1)$. Now, $\theta \in (\theta_N, \theta_{N+1})$ implies that

$$0 < d_{hj}\theta - \operatorname{Im} \mu_{hj}^N < 2\pi.$$

With the aid of the above inequalities we readily verify that

$$\overset{c}{Y}_h(s)(P_N^+)_{hj}(s)\overset{c}{Y}_j(s)^{-1} \sim 0, \qquad s \to \infty, \quad s \in S_\theta^-(R),$$

if $h < j$ and R is a sufficiently large number. With (4.3.10) we conclude that

$$F_N - F^\infty \in \mathrm{Gl}(n; \mathscr{A}_0(S_\theta^-)).$$

In a similar manner we prove that

$$F_N - F^{-\infty} \in \mathrm{Gl}(n; \mathscr{A}_0(S_\theta \cap \bar{S}_{-\infty})).$$

As F^∞ and $F^{-\infty}$ are represented asymptotically by the series $\sum_{h=0}^\infty F_h s^{-h/p}$ as $s \to \infty$ in $\bar{S}_\infty(R_1)$ and $\bar{S}_{-\infty}(R_1)$, respectively, it remains to be shown that F_N admits the same asymptotic expansion as $s \to \infty$ in a "strip" of the form $S_\theta(R) \cap \{s \in \mathbb{C} : |\operatorname{Im} s| \leqq R\}$, where R is a suitable positive number. This follows immediately from a Phragmén-Lindelöf-type of argument (cf. [15, p. 180]). Thus we conclude that $F_N \in \mathrm{Gl}(n; \mathscr{A}(S_\theta))$ for all $\theta \in (\theta_N, \theta_{N+1})$ and, consequently, $F_N \in \mathrm{Gl}(n; \mathscr{A}(S_N^*))$. Since here we have not used any properties of F_N, except for those mentioned in Theorem 4.3.8, the second statement now follows from Theorem 2.4.1.

In conclusion we can say that, unless $P_{N-1} = P_N$ or $P_N = P_{N+1}$, the asymptotic sets S_N^* mentioned in Proposition 4.3.13 are maximal in the following sense: for every $N \in \mathbb{Z}$ there exists a (unique) matrix function $F_N \in \mathrm{Gl}(n; \mathscr{A}(S_N^*))$ with the property that $A^{F_N} = \overset{c}{A}$, whereas $F_N \notin \mathrm{Gl}(n; \mathscr{A}(S_{\theta_N}))$ and $F_N \notin \mathrm{Gl}(n; \mathscr{A}(S_{\theta_{N+1}}))$. Analogously, there exist maximal asymptotic sets $\tilde{S}_N$, defined by

$$\tilde{S}_N = \bigcup_{\theta_N < \theta < \theta_{N+1}} e^{i\pi} S_{\theta+\pi}, \qquad N \in \mathbb{Z}$$

and unique matrix functions $\tilde{F}_N \in \mathrm{Gl}(n; \mathscr{A}(\tilde{S}_N))$ such that $A^{\tilde{F}_N} = \overset{c}{A}$. The Stokes phenomenon in the class of difference equations considered in this section can be completely described by determining the connection matrices of (F_N, F_{N+1}) and $(\tilde{F}_N, \tilde{F}_{N+1})$, $N \in \mathbb{Z}$, and the connection matrices P^∞ and $P^{-\infty}$ defined below Definition 4.2.1.

REFERENCES

[1] G. D. Birkhoff and W. J. Trjitzinsky, *Analytic theory of singular difference equations*, Acta Math., 60 (1933), pp. 1-89.

[2] A. Duval, *Lemmes de Hensel et factorisation formelle pour les opérateurs aux différences*, Funkcial. Ekvac., 26 (1983), pp. 349-368.

[3] ———, *Equations aux différences dans le champ complexe*, thesis, Publ. Institut de Recherche Mathématique Avancée, Strasbourg, France, 1984.

[4] J. Ecalle, *Les fonctions résurgentes*, Tôme III, Publ. Math. d'Orsay, Université de Paris-Sud, France, 1985.

[5] J. Horn, *Zur Theorie der linearen Differenzengleichungen*, Jahresber. Deutsch. Math.-Verein., 24 (1915), pp. 210-225.

[6] M. Hukuhara, *Sur les points singuliers des équations différentielles linéaires*, II, J. Fac. Sci. Hokkaido Imp. Univ., 5 (1937), pp. 123-166; III, Mém. Fac. Sci. Kyushu Univ., Ser. A, 2 (1941), pp. 125-137.

[7] G. K. Immink, *Asymptotics of Analytic Difference Equations*, Lecture Notes in Math. 1085, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1984.

[8] G. K. IMMINK, *Resurgent functions and connection matrices for a linear homogeneous system of difference equations*, Funkcial. Ekvac., 31 (1988), pp. 197–219.

[9] ———, *On the asymptotic behavior of the coefficients of asymptotic power series and its relevance to Stokes phenomena*, Memorandum 738, University of Twente, Twente, the Netherlands, 1988; SIAM J. Math. Anal., 22 (1991).

[10] ———, *Asymptotic expansions with error bounds for solutions of difference equations of 'level* 1^+,' to appear in *Equations différentielles dans le champ complex*, Colloque Franco-Japonnais 1985, Vol. 1, Publ. Institut de Recherche Mathématique Advancée, Strasbourg, France, 1988, pp. 35–60.

[11] B. MALGRANGE, *Sur les points singuliers des équations différentielles*, l'Enseign. Math., 20 (1974), pp. 147–176.

[12] N. E. NÖRLUND, *Vorlesungen über Differenzenrechnung*, Chelsea, New York, 1954.

[13] C. PRAAGMAN, *The formal classification of linear difference operators*, Proc. Kon. Ned. Ac. van Wetensch. Ser. A., 86 (1983), pp. 249–261.

[14] Y. SIBUYA, *Simplification of a system of linear ordinary differential equations about a singular point*, Funkcial. Ekvac., 4 (1962), pp. 29–56.

[15] E. C. TITCHMARSH, *The Theory of Functions*, Second edition, Oxford University Press, Oxford, 1939.

[16] H. L. TURRITTIN, *The formal theory of systems of irregular homogeneous linear difference and differential equations*, Bol. Soc. Mat. Mexicana (2), 5 (1960), pp. 255–264.

[17] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, Interscience, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 260-267, January 1991

015

INEQUALITIES FOR THE ZEROS OF THE AIRY FUNCTIONS*

GIOVANNA PITTALUGA† AND LAURA SACRIPANTE†

Abstract. By using a theorem of Sturm type the authors show that the approximations obtained by truncating the asymptotic series for the real zeros of the Airy functions Ai (x) or Bi (x) are in fact lower and upper bounds. A lower bound for the zeros of the derivatives of these functions is also derived.

Key words. Airy functions, zeros, inequalities, Sturm comparison theorem

AMS(MOS) subject classifications. primary 33A40; secondary 34C10

1. Introduction. It is well known that the Airy functions Ai (x) and Bi (x) play an important rôle in the asymptotic theory of differential equations (see Olver [7, Chap. 11]). They satisfy the so-called Airy differential equation

$$y''-xy=0, \tag{1.1}$$

which arises naturally as an approximation to the general second-order differential equation reduced to the normal form [5, p. B4].

The functions Ai (x) and Bi (x) have infinitely many real zeros on the negative axis:

$$0>a_1>a_2>\cdots \quad \text{and} \quad 0>b_1>b_2>\cdots,$$

respectively.

For these zeros the following asymptotic expansions, due to Miller [5], hold:

$$a_n \cong -f(\lambda_n), \tag{1.2}$$

$$b_n \cong -f(\mu_n), \tag{1.3}$$

where

$$\lambda_n=\frac{3(4n-1)\pi}{8}, \quad \mu_n=\frac{3(4n-3)\pi}{8}, \quad n=1,2,\cdots,$$
$$f(z)=z^{2/3}\left(1+\frac{5}{48z^2}-\frac{5}{36z^4}+\frac{77125}{82944z^6}-\frac{108056875}{6967296z^8}+\frac{162375596875}{334430208z^{10}}+\cdots\right). \tag{1.4}$$

These approximations, even if truncated to the first terms, give good numerical results also for moderate values of n, but do not furnish any information about the error. Moreover,

$$a_1=-2.33810741\cdots, \qquad b_1=-1.17371322\cdots, \tag{1.5}$$

and, as $n\to\infty$,

$$a_n \cong -\left[\frac{3}{8}(4n-1)\pi\right]^{2/3}, \qquad b_n \cong -\left[\frac{3}{8}(4n-3)\pi\right]^{2/3}. \tag{1.6}$$

For the properties of the Airy functions, we refer to Miller [5], Olver [7], and Abramowitz and Stegun [1].

* Received by the editors June 13, 1988; accepted for publication (in revised form) December 12, 1989. This work was supported by the Ministero della Pubblica Istruzione of Italy.

† Università di Torino, Dipartimento di Matematica, 10123 Torino, Italy.

One- and two-term asymptotic representations for the zeros a_n, with bounds for the error term, can be found in a paper of Hethcote [2]. In a more recent paper, Laforgia [4] has found an analogous one-term asymptotic representation for the zeros b_n.

Inequalities for the zeros a_n and b_n can be obtained by means of the following theorem of Sturm type, due to Hethcote [3].

THEOREM 1.1. *Let c_n and d_n, $n=1, 2, \cdots$, be the zeros (in increasing order) of $u(x)$ and $v(x)$, respectively, which are nontrivial solutions of $u''+p(x)u=0$ and $v''+q(x)v=0$ with continuous $p(x)$ and $q(x)$. If $c_n-d_n\to 0$ as $n\to\infty$, $p(x)\geqq q(x)$ and either $p(x)$ or $q(x)$ is nonincreasing, then $c_n\geqq d_n$ for $n=1, 2, \cdots$.*

By applying this theorem, Hethcote [3] has found the following inequalities for the nth zero a_n of the function Ai (x):

$$-\left(\lambda_n+\frac{3}{2}\arctan\frac{5}{48\lambda_n}\right)^{2/3}\leqq a_n\leqq -\lambda_n^{2/3}, \tag{1.7}$$

where $\lambda_n=3(4n-1)\pi/8$ and $n=1, 2, \cdots$.

The Hethcote result shows that a_n is greater than the two-term approximation obtained by truncating asymptotic expansion (1.2). Indeed, it is easily verified that

$$a_n\geqq -\left(\lambda_n+\frac{3}{2}\arctan\frac{5}{48\lambda_n}\right)^{2/3}>-\lambda_n^{2/3}\left(1+\frac{5}{32\lambda_n^2}\right)^{2/3}>-\lambda_n^{2/3}\left(1+\frac{5}{48\lambda_n^2}\right). \tag{1.8}$$

The aim of this paper is to improve (1.7) by means of Theorem 1.1. We also derive, in a similar way, lower and upper bounds for the zeros b_n, $n=1, 2, \cdots$, of the function Bi (x), by truncating the asymptotic series (1.3).

The main results are stated in the following theorem.

THEOREM 1.2. *The error in stopping at any of the first five terms of Miller's expansions* (1.2) *and* (1.3) *is bounded by the next term and has the same sign. Furthermore, the following inequalities hold*:

$$a_n>-F(\lambda_n),\qquad b_n>-F(\mu_n),$$

where

$$\lambda_n=\frac{3(4n-1)\pi}{8},\quad \mu_n=\frac{3(4n-3)\pi}{8},\quad n=1, 2, \cdots,$$

$$F(z)=z^{2/3}\left(1+\frac{5}{48z^2}-\frac{5}{36z^4}+\frac{77125}{82944z^6}-\frac{108056875}{6967296z^8}+\frac{162375596875}{334430208z^{10}}\right).$$

To establish these results, it is necessary to prove, step by step, that the approximations obtained by considering an even or odd number of terms in the asymptotic series (1.2) and (1.3) are in fact lower and upper bounds, respectively. The essence of each successive step is quite similar, but details are progressively more complicated. So, in the following two sections, we will give only the outline of the proof for the first of the required results related to the zeros a_n of Ai (x) that improve the Hethcote inequality (1.8).

We have established the complete proof of Theorem 1.2 by using the symbolic system MAPLE, as we will explain at the end of § 3.

2. The comparison equation. In order to apply Theorem 1.1 we will refer to, instead of (1.1), the differential equation

$$u''+\left(1+\frac{5}{36\xi^2}\right)u=0, \tag{2.1}$$

which is satisfied by the functions

$$A(\xi) = (\tfrac{3}{2}\xi)^{1/6}\,\mathrm{Ai}\,[-(\tfrac{3}{2}\xi)^{2/3}],$$
$$B(\xi) = (\tfrac{3}{2}\xi)^{1/6}\,\mathrm{Bi}\,[-(\tfrac{3}{2}\xi)^{2/3}]. \tag{2.2}$$

If α_n and β_n, $n = 1, 2, \cdots$, denote the positive zeros (in increasing order) of $A(\xi)$ and $B(\xi)$, respectively, then we have

$$\alpha_n = \tfrac{2}{3}(-a_n)^{3/2}, \quad \beta_n = \tfrac{2}{3}(-b_n)^{3/2}, \quad n = 1, 2, \cdots. \tag{2.3}$$

Hence, from the asymptotic expansions (1.2) and (1.3), we get

$$\alpha_n \cong T(\lambda_n), \qquad \beta_n \cong T(\mu_n), \tag{2.4}$$

where λ_n and μ_n have the previous values (1.4), and

$$T(z) = \frac{2}{3}z\left(1 + \frac{5}{32z^2} - \frac{1255}{6144z^4} + \frac{272075}{196608z^6} - \frac{4084035425}{176160768z^8} + \frac{4098107432375}{5637144576z^{10}} + \cdots\right).$$

Furthermore, (1.5) and (1.6) give

$$\alpha_1 = 2.383446612\cdots, \qquad \beta_1 = 0.8477186453\cdots,$$

and, as $n \to \infty$,

$$\alpha_n \cong \frac{\pi}{4}(4n-1), \qquad \beta_n \cong \frac{\pi}{4}(4n-3).$$

The differential equation (2.1) will be compared with the differential equation

$$v'' + G(\xi)v = 0, \tag{2.5}$$

where

$$G(\xi) = \frac{1}{2}\frac{g'''}{g'} - \frac{3}{4}\left(\frac{g''}{g'}\right)^2 + g'^2, \tag{2.6}$$

which is satisfied by

$$v(\xi) = (g')^{-1/2}\cos g(\xi). \tag{2.7}$$

As we will see, (2.1) and (2.5) are "very close" for a proper choice of the function $g(\xi)$. For this choice it will be necessary to take into account that the asymptotic expansions (2.4) yield

$$\lambda_n \cong S(\alpha_n), \qquad \mu_n \cong S(\beta_n), \tag{2.8}$$

where

$$S(z) = \frac{3}{2}z\left(1 - \frac{5}{72z^2} + \frac{1105}{31104z^4} - \frac{82825}{746496z^6} + \frac{1282031525}{1504935936z^8} - \frac{1683480621875}{139314069504z^{10}}\cdots\right).$$

3. The first approximation for the zeros of Ai (x). In this section we will prove the following inequality for the zeros a_n of Ai (x):

$$a_n < -\lambda_n^{2/3}\left(1 + \frac{5}{48\lambda_n^2} - \frac{5}{36\lambda_n^4}\right). \tag{3.1}$$

For this purpose we use in (2.6)

$$g(\xi) = \xi - \frac{5}{72\xi} + \frac{1105}{31104\xi^3} - \frac{\pi}{4}, \tag{3.2}$$

as suggested by (2.8). From (2.6) we obtain

$$G(\xi) - \left(1 + \frac{5}{36\xi^2}\right) = \frac{y^3}{[g'(y)]^2}p(y),$$

where $y = 5/72\xi^2$ and

$$p(y) = \frac{221^4}{10^4} y^5 - \frac{221^3}{250} y^4 - \frac{103 \cdot 221^2}{125} y^3 - \frac{10487 \cdot 221}{250} y^2 + \frac{194801}{100} y - 3313.$$

So, we easily get

$$G(\xi) - \left(1 + \frac{5}{36\xi^2}\right) < 0,$$

for $\xi > \eta$, with $\eta = 0.335 \cdots$.

Now we note that the function (2.7), with $g(\xi)$ given by (3.2), has infinitely many positive zeros ξ_n, $n = 1, 2, \cdots$, belonging to the interval $\eta < \xi < \infty$. Hence, observing that

$$\lim_{n\to\infty} \xi_n = +\infty, \qquad \lim_{n\to\infty} (\xi_n - \alpha_n) = 0,$$

and that the coefficient of u in (2.1) is a continuous decreasing function in $0 < \xi < \infty$, we see that the conditions required in Theorem 1.1 are satisfied. Thus we may conclude that

$$\xi_n \leqq \alpha_n \quad \text{for } n \geqq 1.$$

Now we consider the equation

$$\xi - \frac{5}{72\xi} + \frac{1105}{31104\xi^3} - \frac{\pi}{4} = (2n-1)\frac{\pi}{2}. \tag{3.3}$$

This can be written in the form

$$\xi = \psi_n(\xi),$$

with

$$\psi_n(\xi) = \frac{5}{72\xi} - \frac{1105}{31104\xi^3} + \frac{2}{3}\lambda_n,$$

where λ_n is defined in (1.4).

We observe that the root of (3.3), which we have previously denoted by ξ_n, lies in the interval $\sqrt{221}/12 < \xi < \infty$, where the function $\psi_n(\xi)$ decreases. Since from (1.8) and (2.3)

$$\alpha_n < \delta_n \equiv \frac{2}{3}\lambda_n\left(1 + \frac{5}{32\lambda_n^2}\right),$$

we have $\xi_n \leqq \alpha_n < \delta_n$, and consequently,

$$\alpha_n \geqq \psi_n(\xi_n) > \psi_n(\delta_n).$$

Then, taking into account that

$$\alpha_n > \frac{2}{3}\lambda_n\left(1 + \frac{5}{32\lambda_n^2} - \frac{1255}{6144\lambda_n^4}\right),$$

we obtain

$$a_n < -\lambda_n^{2/3}\left(1 + \frac{5}{32\lambda_n^2} - \frac{1255}{6144\lambda_n^4}\right)^{2/3}.$$

Finally, by observing that

$$0 < \frac{5}{32\lambda_n^2} - \frac{1255}{6144\lambda_n^4} < 1$$

and using a well-known inequality [6, p. 35], i.e.,

$$(1+h)^{\alpha} > 1+\alpha h+\binom{\alpha}{2}h^2, \quad 0<h<1, \quad 0<\alpha<1,$$

we establish (3.1).

Now, to complete the proof of Theorem 1.2 it is necessary to repeat the previous procedure three times. Since the algebraic calculations become progressively harder, we have overcome this difficulty by using a symbolic mathematical system. Therefore, rather than go into full details, we will limit ourselves to some remarks.

First, we add a term to $g(\xi)$. Consequently, using the same notation,

(a) We find the function $G(\xi)$ and, to apply the Hethcote theorem, we prove that the difference $G(\xi)-(1+(5/36\xi^2))$ does not change sign in the interval of the zeros.

(b) We prove that, in the same interval, the function $\psi_n(\xi)$ is always decreasing.

(c) We reach a new inequality by evaluating $\psi_n(\xi)$ in the bound previously established.

The most important functions and the inequalities arising at each step of our proof are:

1. *First step.*

$$g(\xi)=\xi-\frac{5}{72\xi}+\frac{1105}{31104\xi^3}-\frac{82825}{746496\xi^5}-\frac{\pi}{4},$$

$$G(\xi)-\left(1+\frac{5}{36\xi^2}\right)=\frac{y^4}{[g'(y)]^2}p_1(y),$$

$$a_n > -\lambda_n^{2/3}\left(1+\frac{5}{48\lambda_n^2}-\frac{5}{36\lambda_n^4}+\frac{77125}{82944\lambda_n^6}\right).$$

2. *Second step.*

$$g(\xi)=\xi-\frac{5}{72\xi}+\frac{1105}{31104\xi^3}-\frac{82825}{746496\xi^5}+\frac{1282031525}{1504935936\xi^7}-\frac{\pi}{4},$$

$$G(\xi)-\left(1+\frac{5}{36\xi^2}\right)=\frac{y^5}{[g'(y)]^2}p_2(y),$$

$$a_n < -\lambda_n^{2/3}\left(1+\frac{5}{48\lambda_n^2}-\frac{5}{36\lambda_n^4}+\frac{77125}{82944\lambda_n^6}-\frac{108056875}{6967296\lambda_n^8}\right).$$

3. *Third step.*

$$g(\xi)=\xi-\frac{5}{72\xi}+\frac{1105}{31104\xi^3}-\frac{82825}{746496\xi^5}+\frac{1282031525}{1504935936\xi^7}-\frac{1683480621875}{139314069504\xi^9}-\frac{\pi}{4},$$

$$G(\xi)-\left(1+\frac{5}{36\xi^2}\right)=\frac{y^6}{[g'(y)]^2}p_3(y),$$

$$a_n > -\lambda_n^{2/3}\left(1+\frac{5}{48\lambda_n^2}-\frac{5}{36\lambda_n^4}+\frac{77125}{82944\lambda_n^6}-\frac{108056875}{6967296\lambda_n^8}+\frac{162375596875}{334430208\lambda_n^{10}}\right),$$

where $y=5/72\xi^2$ and

$$p_1(y)=\frac{51281261}{100}-\frac{27372926}{125}y+\frac{7794996033}{500}y^2-\frac{50103718291}{250}y^3$$
$$-\frac{7012956328919}{10000}y^4+\frac{8691232890407}{500}y^5+\frac{5244570435487}{200}y^6$$

$$-\frac{8036308150637}{20}y^7+\frac{120471895488961}{16}y^8,$$

$$p_2(y)=-538713799/4+\frac{14820773151}{1000}y-\frac{22493384194527}{5000}y^2$$
$$+\frac{3278666826931429}{8000}y^3-\frac{1046709128741315107}{100000}y^4$$
$$-\frac{1261062200876664583}{50000}y^5+\frac{6390468125205053 4991}{50000}y^6$$
$$-\frac{65735260766514133374699}{1000000}y^7-\frac{201944954419818577655571}{2000000}y^8$$
$$+\frac{514517617846345251536371 51}{20000000}y^9-\frac{44678390902083129973969 0853}{4000000}y^{10}$$
$$+\frac{69156783124351054388530995 14641}{1600000000}y^{11},$$

$$p_3(y)=\frac{53751522223401}{1000}+\frac{20773109330799}{2500}y$$
$$+\frac{11109302866435629}{8000}y^2-\frac{10150465536010571741}{50000}y^3$$
$$+\frac{5891049024457820012593}{200000}y^4-\frac{1248226635247820542 90243}{100000}y^5$$
$$-\frac{19155021338580508161711 99}{1000000}y^6$$
$$+\frac{17640498460982180295968469}{125000}y^7$$
$$-\frac{27553584174318977467203014 8599}{20000000}y^8$$
$$+\frac{24072387103655940191454086 96631}{2000000}y^9$$
$$+\frac{27147944733527654973290247 82474641}{1600000000}y^{10}$$
$$-\frac{87433525211356346353713691 60970219}{16000000}y^{11}$$
$$+\frac{48793719536427679447591219 47186563}{1280000}y^{12}$$
$$-\frac{80173897396755453632596736 92504139}{25600}y^{13}$$
$$+\frac{84223328375724498468133493 019784801}{4096}y^{14}.$$

4. Bounds for the zeros of Bi (x). In this section we will consider again the differential equation (2.1), but we will refer to the second solution, i.e., to the function $B(\xi)$ defined in (2.2). Let us remember that we have denoted by β_n, $n=1, 2, \cdots$, the positive increasing zeros of $B(\xi)$ that are related to the zeros b_n, $n=1, 2, \cdots$, by the second relation in (2.3).

The procedure for deriving inequalities satisfied by the zeros b_n is perfectly analogous to that used for the zeros a_n. We give only a sketch of this procedure.

In the first instance, let us consider the simple case

$$g(\xi)=\xi+\frac{\pi}{4}.$$

Here it is very easy to verify that Theorem 1.1 can be applied. So we obtain

$$\beta_n \geqq \tfrac{2}{3}\mu_n, \qquad n=1, 2, \cdots, \tag{4.1}$$

i.e.,

$$b_n \leqq -\mu_n^{2/3}, \qquad n=1, 2, \cdots.$$

The next step is to choose

$$g(\xi)=\xi-\frac{5}{72\xi}+\frac{\pi}{4}.$$

We observe that

$$G(\xi)-\left(1+\frac{5}{36\xi^2}\right)>0 \quad \text{if } \xi>0,$$

and that, from (4.1), the zeros β_n lie in the interval $((\pi/4), +\infty)$.

Now, as in § 3, it is possible to obtain a lower bound for b_n by applying Theorem 1.1, that is,

$$b_n > -\mu_n^{2/3}\left(1+\frac{5}{48\mu_n^2}\right), \qquad n=1, 2, \cdots.$$

The next steps parallel the ones for the function Ai (x).

5. A bound for the zeros of Ai′ (x) and Bi′ (x). To complete the study of the zeros of the Airy functions, it is natural to ask whether the same analysis can be adapted to establish similar bounds for the zeros a_n' and b_n' of the derivatives Ai′ (x) and Bi′ (x).

In fact, it is well known that the following asymptotic expansions due to Miller hold:

$$a_n' \cong -\bar{f}(\mu_n), \qquad b_n' \cong -\bar{f}(\lambda_n),$$

where

$$\bar{f}(z)=z^{2/3}\left(1-\frac{7}{48z^2}+\frac{35}{288z^4}-\cdots\right),$$

while μ_n and λ_n are given by (1.4).

Unfortunately we can prove only the following theorem.

THEOREM 5.1. *For the zeros a_n' and b_n' of the derivatives of the Airy functions* Ai′ (x) *and* Bi′ (x), *respectively, the following inequalities hold*:

$$a_n' \geqq -[\tfrac{3}{8}(4n-3)\pi]^{2/3}, \quad b_n' \geqq -[\tfrac{3}{8}(4n-1)\pi]^{2/3}, \quad n=1, 2, \cdots. \tag{5.1}$$

Indeed, if we denote by α'_n, $n=1, 2, \cdots$, the positive zeros (in increasing order) of the function

$$\bar{A}(\xi)=(\tfrac{3}{2}\xi)^{-1/6}\,\mathrm{Ai}'\,[-(\tfrac{3}{2}\xi)^{2/3}],$$

which satisfies the differential equation

$$u''+\left(1-\frac{7}{36\xi^2}\right)u=0, \tag{5.2}$$

we have

$$\alpha'_n\cong\frac{2}{3}\,\mu_n\left(1-\frac{7}{32\mu_n^2}+\frac{1169}{6144\mu_n^4}+\cdots\right),$$

and consequently

$$\mu_n\cong\frac{3}{2}\,\alpha'_n\left(1+\frac{7}{72\alpha_n'^2}-\frac{1463}{31104\alpha_n'^4}+\cdots\right).$$

As in the previous section, we first compare (5.2) with the differential equation (2.5), by assuming that in (2.6)

$$g(\xi)=\xi+\frac{\pi}{4}.$$

We find that $G(\xi)=1$, and so we can apply the Hethcote theorem and easily establish the first of the two inequalities (5.1).

Then, if we add a term to $g(\xi)$, i.e.,

$$g(\xi)=\xi+\frac{7}{72\xi}+\frac{\pi}{4},$$

we find

$$G(\xi)=\left(1-\frac{7}{72\xi^2}\right)^2-\frac{1512}{(72\xi^2-7)^2}.$$

Since in this case $G(\xi)$ is an increasing function, the conditions required in Theorem 1.1 are not satisfied.

The same remarks hold for the zeros of $\mathrm{Bi}'(x)$.

Acknowledgment. The authors are grateful to Professor F. W. J. Olver for his careful reading of the manuscript and useful suggestions for improved formulation of the paper.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Applied Mathematics Series 55, National Bureau of Standards, Washington, D.C., 1964.

[2] H. W. Hethcote, *Error bounds for asymptotic approximations of zeros of transcendental functions*, SIAM J. Math. Anal., 1 (1970), pp. 147-152.

[3] ———, *Bounds for zeros of some special functions*, Proc. Amer. Math. Soc., 25 (1970), pp. 72-74.

[4] A. Laforgia, *Disuguaglianze asintotiche per gli zeri delle funzioni* $\mathrm{Bi}(x)$, $\mathrm{Ai}'(x)$, $\mathrm{Bi}'(x)$, Pubbl. Istituto per l'applicazione del Calcolo, Rome, Ser. III, 178 (1979).

[5] J. C. P. Miller, *The Airy Integral*, Mathematical Tables, Vol. B, British Association for the Advancement of Science, Cambridge University Press, Cambridge, U.K., 1946.

[6] D. S. Mitrinović, *Analytic Inequalities*, Springer-Verlag, Berlin, New York, 1970.

[7] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 268-271, January 1991

016

INDUCTIVELY GENERATING THE SPHERICAL HARMONICS*

ALLAN FRYANT†

Abstract. The spherical harmonics of degree less than or equal to k in R^{n-1} are used to generate the spherical harmonics of degree k in R^n.

Key words. spherical harmonics, generating function

AMS(MOS) subject classifications. 33A45, 31B99

A homogeneous polynomial of degree k in the real variables, $x_1, x_2, \cdots, x_n$, that satisfies Laplace's equation,

$$\frac{\partial^2 H}{\partial x_1^2}+\frac{\partial^2 H}{\partial x_2^2}+\cdots+\frac{\partial^2 H}{\partial x_n^2}=0,$$

is called a spherical harmonic of degree k in R^n.

The set of all spherical harmonics of degree k in R^n forms a vector space over the field of complex numbers. We denote this vector space by H_n^k. Let $d_n^k = \dim H_n^k$. Then

$$d_n^k = (n+2k-2)\frac{(n+k-3)!}{k!(n-2)!}$$

[2, p. 140]. An elementary argument by induction on k shows that

$$d_n^k = \sum_{j=0}^{k} d_{n-1}^j, \qquad n=3,4,5,\cdots.$$

This dimensional equality suggests a deep relationship between the spherical harmonics of degree less than or equal to k in R^{n-1} and the spherical harmonics of degree k in R^n. We investigate this relationship here.

Let $\sum_{n-1}$ denote the unit sphere in R^n, and $\langle f, g\rangle$ be the usual inner product on $\sum_{n-1}$. That is,

$$\langle f, g\rangle = \int_{\sum_{n-1}} f(x)g(x)\,dx, \tag{1}$$

where

$$\sum_{n-1} = \{(x_1, x_2, \cdots, x_n): x_1^2+x_2^2+\cdots+x_n^2=1\} \subset R^n.$$

Spherical harmonics of different degrees are orthogonal with respect to the inner product (1) [2, p. 144]. Thus, letting $\{P_j^1, P_j^2, \cdots, P_j^{d_{n-1}^j}\}$ be orthonormal bases for each of the vector spaces H_{n-1}^j, on taking the union of these bases over $j \leqq k$ we obtain a set of d_n^k orthonormal spherical harmonics that forms a basis for the vector space of harmonic polynomials of degree not exceeding k in R^{n-1}. Re-indexing, we let $\{P_1, P_2, \cdots, P_{d_n^k}\}$ denote such a basis.

The following theorem shows how the spherical harmonics in R^{n-1} can be used to generate the spherical harmonics in R^n, $n=2, 3, 4, \cdots$.

* Received by the editors April 4, 1988; accepted for publication (in revised form) October 30, 1989.
† Jamestown College, Box 6037, Jamestown, North Dakota 58401.

THEOREM. *Let* $P_1, P_2, \cdots, P_{d_n^k}$ *be orthonormal spherical harmonics in* R^{n-1} *of degree less than or equal to* k. *If* $x=(x_1, x_2, \cdots, x_n)\in R^n$, *and* $t=(t_1, t_2, \cdots, t_{n-1})\in \sum_{n-2}$, *then*

$$(x_1+ix_2t_1+ix_3t_2+\cdots+ix_nt_{n-1})^k=\sum_{j=1}^{d_n^k} Y_k^j(x)P_j(t),$$

where Y_k^j, $j=1, 2, \cdots, d_n^k$, *are homogeneous harmonic polynomials of degree* k *in* R^n, *and* $\langle Y_k^j, Y_k^l\rangle=0$ *if* $j\neq l$.

Proof. Consider the multinomial expansion

$$(x_1+ix_2t_1+\cdots+ix_nt_{n-1})^k=\sum_{\nu=1}^{l} c_\nu x^{\alpha_\nu}t^{\beta_\nu}.$$

Here $\alpha_\nu=(\alpha_1^\nu, \alpha_2^\nu, \cdots, \alpha_n^\nu)$, and $x^{\alpha_\nu}=x_1^{\alpha_1^\nu}x_2^{\alpha_2^\nu}\cdots x_n^{\alpha_n^\nu}$, and similarly for t^{β_ν}. When restricted to the unit sphere $\sum_{n-2}$, any polynomial of degree j in $t_1, t_2, \cdots, t_{n-1}$ is a linear combination of spherical harmonics of degree not exceeding j [2, p. 140]. Thus, on restricting t to the unit sphere, each monomial t^{β_ν} can be written as a linear combination of the spherical harmonics $P_1, P_2, \cdots, P_{d_n^k}$. A rearrangement then yields

$$(x_1+ix_2t_1+\cdots+ix_nt_{n-1})^k=\sum_{j=1}^{d_n^k} Y_k^j(x)P_j(t).$$

The Y_k^j are clearly homogeneous polynomials of degree k in $x_1, x_2, \cdots, x_n$. Further, since $t_1^2+t_2^2+\cdots+t_{n-1}^2=1$, they are harmonic. To see that these Y_k^j are orthogonal, consider the function

$$G(u, b)=\int_{\sum_{n-1}}(\xi_1u_1+\xi_2u_2+\cdots+\xi_nu_n)^k(\xi_1b_1+\xi_2b_2+\cdots+\xi_nb_n)^k\,d\xi$$

$$=\int_{\Sigma_{n-1}}(u, \xi)^k(b, \xi)^k\,d\xi.$$

The function G is an orthogonal invariant. That is, $G(Ou, Ob)=G(u, b)$ for any orthogonal transformation O of R^n. Thus, there exists a polynomial $\Phi(\lambda, \mu, \nu)$ such that

$$G(u, b)=\Phi[(u, u), (u, b), (b, b)]$$

[1, pp. 244–245]. Further, since $G(u, b)$ is homogeneous and of degree k in the components of u and b, it follows that

$$\Phi[(u, u), (u, b), (b, b)]=\sum c_{\alpha\beta\gamma}(u, u)^\alpha(u, b)^\beta(b, b)^\gamma,$$

where the sum is over all nonnegative integers α, β, γ such that $2\alpha+\beta=k$ and $2\gamma+\beta=k$. Thus we have the polynomial equation

$$\sum c_{\alpha\beta\gamma}(u, u)^\alpha(u, b)^\beta(b, b)^\gamma=\int_{\sum_{n-1}}(u, \xi)^k(b, \xi)^k\,d\xi. \tag{2}$$

Now let

$$u=(1, it_1, it_2, \cdots, it_{n-1}), \qquad t\in\sum_{n-2},$$

$$b=(1, -is_1, -is_2, \cdots, -is_{n-1}), \qquad s\in\sum_{n-2}.$$

Then $(u, u) = (b, b) = 0$, and (2) becomes

$$c_k(u, b)^k = \int_{\Sigma_{n-1}} (\xi_1 + i\xi_2 t_1 + \cdots + i\xi_n t_{n-1})^k (\xi_1 - i\xi_2 s_1 - \cdots - i\xi_n s_{n-1})^k \, dx$$

$$= \int_{\Sigma_{n-1}} \left[\sum_{j=1}^{d_n^k} Y_k^j(x) P_j(t) \right] \left[\sum_{l=1}^{d_n^k} \overline{Y_k^l(x) P_l(s)} \right] dx$$

$$= \sum_{j=1}^{d_n^k} \sum_{l=1}^{d_n^k} \left[\int_{\Sigma_{n-1}} Y_k^j(x) \overline{Y_k^l(x)} \, dx \right] P_j(t) \overline{P_l(s)}.$$

But

$$(u, b)^k = (1 + s_1 t_1 + s_2 t_2 + \cdots + s_{n-1} t_{n-1})^k.$$

Further, appealing to the fact that any polynomial of degree j when restricted to the unit sphere is a linear combination of spherical harmonics of degree not exceeding j, and using the Funk–Hecke theorem [1, p. 247], it follows that

$$(1 + s_1 t_1 + s_2 t_2 + \cdots + s_{n-1} t_{n-1})^k = \sum_{j=1}^{d_n^k} \lambda_j P_j(t) \overline{P_j(s)},$$

where the λ_j are constants. Thus we have

$$\sum_{j=1}^{d_n^k} \sum_{l=1}^{d_n^k} \left[\int_{\Sigma_{n-1}} Y_k^j(\xi) \overline{Y_k^l(\xi)} \, d\xi \right] P_j(t) \overline{P_l(s)} = c_k \sum_{j=1}^{d_n^k} \lambda_j P_j(t) \overline{P_j(s)},$$

from which it follows that

$$\int_{\Sigma_{n-1}} Y_k^j(\xi) \overline{Y_k^l(\xi)} \, d\xi = 0 \quad \text{for } j \neq l.$$

This completes the proof.

It is perhaps interesting to note that the generating function given in the preceding theorem produces the same spherical harmonics in R^2 as does taking the real and imaginary parts of $(x + iy)^k$. Laplace's equation in R^1 is $\partial^2 H / \partial x^2 = 0$, and thus the only harmonic functions in R^1 are 1 and t. These are the spherical harmonics in R^1. The unit sphere in R^1 has equation $t^2 = 1$, and the result of the theorem in this case is

$$(x + iyt)^k = Y_k^1(x, y) \cdot 1 + Y_k^2(x, y) \cdot t.$$

Thus, $Y_k^1(x, y) = \operatorname{Re}\{(x + iy)^k\}$ and $Y_k^2(x, y) = i \operatorname{Im}\{(x + iy)^k\}$. Our dimension-invariant result therefore suggests that use of the complex $i = \sqrt{-1}$ provides a simple means of generating the spherical harmonics in R^2 only because i behaves as does t when t is restricted to the unit sphere $t^2 = 1$.

The result of our theorem can be used to develop integral operators for harmonic functions in R^n [3]. That is, if $x = (x_1, x_2, \cdots, x_n) \in R^n$ and $H(x)$ satisfies

$$\frac{\partial^2 H}{\partial x_1^2} + \frac{\partial^2 H}{\partial x_2^2} + \cdots + \frac{\partial^2 H}{\partial x_n^2} = 0, \qquad \|x\| < R,$$

then there exists a function $h(z, t)$, analytic in the complex variable z in the disk $|z| < R$, and continuous in $t \in R^{n-1}$ on the unit sphere $\|t\| = 1$, such that

$$H(x) = F_n(h)$$

$$= \int_{t=1} h(x_1 + ix_2 t_1 + ix_3 t_2 + \cdots + ix_n t_{n-1}, t) \, dt.$$

In the case $n = 3$, this integral operator reduces to the Bergman B_3 integral operator [4].

In view of the inductive relation that obtains for the spherical harmonics in increasing dimensions, it appears that the generating function presented here is, in a sense, natural. It is certainly far simpler than classical generating functions for the spherical harmonics [5]. In particular, the simplicity of this generating function makes it easy to construct the spherical harmonics in rectangular coordinates, as polynomials in their variables. This can be done by hand, or using a symbol manipulation computer language such as MACSYMA.

REFERENCES

[1] A. ERDÉLYI, *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York, 1953.

[2] E. M. STEIN AND G. WEISS, *Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.

[3] A. FRYANT, *Integral operators for harmonic functions*, in The Mathematical Heritage of C. F. Gauss, G. Rassias, ed., World Scientific, Singapore, to appear.

[4] S. BERGMAN, *Integral Operators in the Theory of Linear Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1969.

[5] P. APPELL AND J. KAMPÉ DE FÉRIET, *Fonctions hypergéométriques et hypersphériques*, Gauthier-Villars, Paris, 1926.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 272–294, January 1991

017

FAMILIES OF ORTHOGONAL AND BIORTHOGONAL POLYNOMIALS ON THE N-SPHERE*

E. G. KALNINS†, WILLARD MILLER, JR.‡, AND M. V. TRATNIK§

Abstract. The Laplace–Beltrami eigenvalue equation $H\Phi = \lambda\Phi$ on the n-sphere is studied, with an added vector potential term motivated by the differential equations for the polynomial Lauricella functions F_A. The operator H is self-adjoint with respect to the natural inner product induced on the sphere and, in certain special coordinates, it admits a spectral decomposition with eigenspaces composed entirely of polynomials. The eigenvalues are degenerate but the degeneracy can be broken through use of the possible separable coordinate systems on the n-sphere. Then a basis for each eigenspace can be selected in terms of the simultaneous eigenfunctions of a family of commuting second-order differential operators that also commute with H. The results provide a multiplicity of n-variable orthogonal and biorthogonal families of polynomials that generalize classical results for one and two variable families of Jacobi polynomials on intervals, disks, and paraboloids.

Key words. multivariable orthogonal and biorthogonal polynomials, the n-sphere

AMS(MOS) subject classifications. 22E70, 33A65, 33A75

1. Introduction.

Orthogonal polynomials in one variable which also satisfy second-order ordinary differential or difference equations have proven extraordinarily useful in the development of special function theory and in the practical approximation of functions (e.g., Askey [2]). Orthogonal and biorthogonal families of polynomials in several variables which satisfy second-order partial differential or difference equations are similarly very useful but there is as yet no general theory and more examples are needed. In this paper we will study such families which are related to the Laplace–Beltrami eigenvalue equation on the n-sphere. Our procedure provides a uniform setting within which to classify several known examples related to the n-sphere and to generate many new examples. Our approach falls within the theory of Dunkl's differential-difference operators (Dunkl, [5], [6]); the main contribution of our paper is to point out the power of separation of variable methods in this theory. (Note: There is also a considerable literature on discrete analogues of the Laplace–Beltrami eigenvalue equation on the sphere in which the symmetry groups are finite, e.g., Stanton [25].)

It was shown by Lam and Tratnik [21] that the Lauricella functions

$$\Phi = F_A\left[\begin{matrix} M+G-1; & -m_1,\cdots,-m_n \\ & \gamma_1,\cdots,\gamma_n \end{matrix}; x_1,\cdots,x_n\right] \tag{1.1}$$

and

*Received by the editors July 10, 1989; accepted for publication (in revised form) December 20, 1989.

†Mathematics Department, University of Waikato, Hamilton, New Zealand.

‡School of Mathematics and Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported in part by National Science Foundation grant DMS 86–00372.

§Center for Nonlinear Studies and T7, Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

$$(1.2)\qquad (1-x)^M F_A\left[\begin{matrix}-M-\gamma_{n+1}+1; & -m_1,\cdots,-m_n \\ & \gamma_1,\cdots,\gamma_n\end{matrix}; -\frac{x_1}{1-x},\cdots,-\frac{x_n}{1-x}\right]$$

form a biorthogonal polynomial family where $m_i = 0,1,2,\cdots$, $M = \sum_{k=1}^n m_i$, $G = \sum_{\ell=1}^{n+1}\gamma_\ell$, $x = \sum_{k=1}^n x_i$, and the γ_ℓ are positive real numbers. (We will derive the inner product later.) Here, the Lauricella function F_A is defined by the series

$$(1.3)\qquad \begin{aligned} &F_A\left[\begin{matrix}a; & b_1,\cdots,b_n \\ & c_1,\cdots,c_n\end{matrix}; z_1,\cdots,z_n\right] \\ &\quad = \sum_{m_1,\cdots,m_n=0}^{\infty} \frac{(a)_{m_1+\cdots+m_n}(b_1)_{m_1}\cdots(b_n)_{m_n} z_1^{m_1}\cdots z_n^{m_n}}{(c_1)_{m_1}\cdots(c_n)_{m_n} m_1!\cdots m_n!}, \end{aligned}$$

where

$$(a)_m = \begin{cases} 1 & \text{if } m = 0 \\ a(a+1)\cdots(a+m-1) & \text{if } m \geq 1. \end{cases}$$

As is easily verified by adding the standard partial differential equations for the F_A (Appell and Kampe de Feriet [1]), these polynomial functions Φ satisfy the eigenvalue equation

$$(1.4)\qquad H\Phi = -M(M+G-1)\Phi,$$

where

$$(1.5)\qquad H = \sum_{i,j=1}^n (x_i\delta_{ij} - x_i x_j)\partial_{x_i x_j} + \sum_{i=1}^n (\gamma_i - G x_i)\partial_{x_i}.$$

Here δ_{ij} is the Kronecker delta. Note that H maps polynomials of maximum order m_i in x_i to polynomials of the same type. It is easy to see that as the m_i range over all nonnegative integers the functions (1.1) form a basis for the space of all polynomials in variables $x_1,\cdots,x_n$, and that the spectrum of H acting on this space is exactly

$$\{-M(M+G-1) : M = 0,1,2,\cdots\}.$$

(For $n = 2$ equation (1.4) appears in the classification by Krall and Sheffer [20] of all second-order partial differential operators such that the Mth order orthogonal polynomials in two variables, with respect to some weight function, are eigenfunctions of the operator.) We will look for other bases of solutions to equation (1.4), both orthogonal and biorthogonal with respect to a natural inner product.

Equation (1.4) is closely related to the Laplace–Beltrami eigenvalue equation on the n-sphere (Eisenhart [7]). To see this, consider the contravariant metric determined by the second derivative terms in H:

$$(1.6)\qquad g^{ij} = \delta_{ij}x_i - x_i x_j, \qquad 1 \leq i,j \leq n.$$

Then $\det(g^{ij}) = g^{-1} = x_1 x_2 \cdots x_n(1-x)$ and

$$g_{ij} = \frac{1}{1-x} + \frac{\delta_{ij}}{x_i}. \tag{1.7}$$

Note that

$$\sum_{j=1}^{n} g^{ij} g_{jk} = \delta_k^i = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{otherwise.} \end{cases}$$

Thus

$$ds^2 = \sum_{i,j=1}^{n} g_{ij} dx_i dx_j$$

determines a metric on a Riemannian space with associated Laplace–Beltrami operator

$$\Delta_n = \frac{1}{\sqrt{g}} \sum_{i,j=1}^{n} \partial_{x_i}(g^{ij}\sqrt{g}\, \partial_{x_j}). \tag{1.8}$$

A straightforward computation yields

$$H = \Delta_n + \Lambda_n \tag{1.9}$$

where

$$\Lambda_n = \sum_{j=1}^{n} \left[\gamma_j - \frac{1}{2} + \left(\frac{n+1}{2} - G\right) x_j\right] \partial_{x_j}. \tag{1.10}$$

Thus if $\gamma_1 = \cdots = \gamma_{n+1} = \frac{1}{2}$ then $H \equiv \Delta_n$, but in general H differs from Δ_n by the first-order differential operator Λ_n.

To identify the Riemannian space we introduce Cartesian coordinates $z_0, z_1, \cdots, z_n$ in $n+1$ dimensional Euclidean space and restrict these coordinates by the conditions

$$\begin{aligned} z_0^2 &= 1 - \sum_{i=1}^{n} x_i = 1 - x \\ z_1^2 &= x_1 \\ z_2^2 &= x_2 \\ &\vdots \\ z_n^2 &= x_n. \end{aligned} \tag{1.11}$$

Note that $z_0^2 + z_1^2 + \cdots + z_n^2 = 1$. Defining a metric ds^2 by

$$ds^2 = \sum_{m=0}^{n} (dz_m)^2$$

we find

$$ds^2 = \frac{1}{4}\sum_{i,j=1}^{n}\left(\frac{1}{1-x}+\frac{\delta_{ij}}{x_i}\right)dx_i dx_j. \tag{1.12}$$

Thus the space corresponds to a portion of the n-sphere S^n. We can consider the coordinates $\{x_i\}$ for $0 \le x_i$ and $x \le 1$ as covering the portion of the n-sphere given by $0 \le z_i$, $\sum_{k=1}^{n} z_k^2 = 1$.

We can transfer the Schrödinger equation (1.4) with vector potential Λ_n to one with a scalar potential V_n through the use of a multiplier transformation ρ. Setting $\Phi(\mathbf{x}) = \rho(\mathbf{x})\Psi(\mathbf{x})$ for a nonzero scalar function ρ we find

$$\begin{aligned}(\Delta_n + \Lambda_n)\Phi &= -M(M+G-1)\Phi \\ &\iff (\Delta_n + V_n(\mathbf{x}))\Psi = -M(M+G-1)\Psi,\end{aligned}$$

provided

$$\rho^{-1} = x_1^{\gamma_1/2-1/4}\cdots x_n^{\gamma_n/2-1/4}(1-x)^{\gamma_{n+1}/2-1/4}. \tag{1.13}$$

A straightforward but tedious computation gives for the scalar potential:

$$\begin{aligned}V_n = &-\frac{1}{4}\sum_{i=1}^{n}\frac{(\gamma_i-\frac{1}{2})(\gamma_i-\frac{3}{2})}{x_i} - \frac{1}{4}\frac{(\gamma_{n+1}-\frac{1}{2})(\gamma_{n+1}-\frac{3}{2})}{1-x} \\ &+\frac{1}{4}\left[(1-G)^2 - 1 - \frac{(n-3)(n+1)}{4}\right]\end{aligned} \tag{1.14}$$

or, in terms of Cartesian coordinates,

$$\begin{aligned}V_n = &-\frac{1}{4}\sum_{i=1}^{n}\frac{(\gamma_i-\frac{1}{2})(\gamma_i-\frac{3}{2})}{z_i^2} - \frac{1}{4}\frac{(\gamma_{n+1}-\frac{1}{2})(\gamma_{n+1}-\frac{3}{2})}{z_0^2} \\ &+\frac{1}{4}\left[(1-G)^2 - 1 - \frac{(n-3)(n+1)}{4}\right].\end{aligned} \tag{1.15}$$

The equation $H'\Psi \equiv (\Delta_n + V_n)\Psi = \lambda\Psi$ has a natural Riemannian metric

$$d\omega = g^{\frac{1}{2}}dx_1\cdots dx_n = x_1^{-\frac{1}{2}}\cdots x_n^{-\frac{1}{2}}(1-x)^{-\frac{1}{2}}dx_1\cdots dx_n \tag{1.16}$$

(Eisenhart [7]). Furthermore, the operator $H' = \rho^{-1}H\rho = \Delta_n + V_n$ is formally self-adjoint with respect to the inner product

$$\langle\Psi_1,\Psi_2\rangle = \int\cdots\int_{x_i>0,x<1}\Psi_1(\mathbf{x})\overline{\Psi_2}(\mathbf{x})d\omega \tag{1.17}$$

where Ψ_1, Ψ_2 are twice continuously differentiable functions of the x_j which take complex values:

$$\langle H'\Psi_1, \Psi_2\rangle = \langle \Psi_1, H'\Psi_2\rangle.$$

This induces an inner product on the space of polynomial functions $\Phi(\mathbf{x}) = \rho\Psi$, with respect to which H is self-adjoint:

$$\begin{aligned}(\Phi_1, \Phi_2) &\equiv \langle \Psi_1, \Psi_2\rangle \\ &= \int\cdots\int_{x_i>0, x<1} \Phi_1(\mathbf{x})\overline{\Phi_2}(\mathbf{x})\rho^{-2}(\mathbf{x})\, d\omega \\ &= \int\cdots\int_{x_i>0, x<1} \Phi_1\overline{\Phi_2}\, d\tilde{\omega},\end{aligned} \tag{1.18}$$

$$d\tilde{\omega} = x_1^{\gamma_1-1}\dots x_n^{\gamma_n-1}(1-x)^{\gamma_{n+1}-1}dx_1\dots dx_n,$$
$$(H\Phi_1, \Phi_2) = (\Phi_1, H\Phi_2).$$

(Indeed, H is clearly formally self-adjoint and the boundary terms obviously vanish for the γ_i sufficiently large. The result can then be extended to all $\gamma_i > 0$ by analytic continuation.) Thus $(\cdot,\cdot)$ is the natural inner product associated with equation (1.4).

A *first-order symmetry operator* for the equation $H\Phi = \lambda\Phi$ is a differential operator

$$K = \sum_{i=1}^{n} f_i(\mathbf{x})\partial_{x_i} + g(\mathbf{x})$$

such that

$$[H, K] \equiv HK - KH = 0$$

(Miller [22]). The first-order symmetry operators form a real Lie algebra under addition of operators, multiplication of an operator by a real scalar, and the commutator bracket $[A, B] = AB - BA$. If $\gamma_1 = \gamma_2 = \cdots = \gamma_{n+1} = \frac{1}{2}$ then $H = \Delta_n$ and it is well known (Eisenhart [7], [8]) that the Lie algebra of real symmetry operators of Δ_n is $so(n+1)$, with dimension $n(n+1)/2$ and a basis of the form $\{L_{\ell k}\}$ where $0 \le \ell < k \le n$ and $L_{\ell k} = -L_{k\ell}$. Explicitly,

$$L_{\ell k} = z_\ell\partial_{z_k} - z_k\partial_{z_\ell} \tag{1.19}$$

and

$$\begin{aligned}L_{ij} &= 2\sqrt{x_i x_j}(\partial_{x_j} - \partial_{x_i}), \qquad 1 \le i, j \le n, \\ L_{0i} &= 2\sqrt{x_i(1-x)}\partial_{x_i}, \qquad 1 \le i \le n.\end{aligned} \tag{1.20}$$

Furthermore, all real second-order differential operators S that commute with Δ_n can be expressed as linear combinations over $\mathbb{R}$ of real constants, elements $L_{\ell k}$ and elements $L_{\ell k}L_{\ell' k'}$. For $\gamma_1, \cdots, \gamma_{n+1}$ arbitrary, however, we have the following lemma.

LEMMA 1. *If K is a first-order operator such that $[K, H] = 0$ then $K = c$, multiplication by the real constant c. The second-order operators*

$$\begin{aligned} S_{ij} &\equiv 4x_ix_j(\partial_{x_i} - \partial_{x_j})^2 + 4(\gamma_i x_j - \gamma_j x_i)(\partial_{x_i} - \partial_{x_j}) \\ &= L_{ij}^2 + 4[(\gamma_i - \tfrac{1}{2})x_j - (\gamma_j - \tfrac{1}{2})x_i](\partial_{x_i} - \partial_{x_-j}) \\ &= S_{ji}, \qquad 1 \le i < j \le n, \end{aligned} \tag{1.21}$$

$$\begin{aligned} S_{0i} &\equiv 4x_i(1-x)\partial_{x_i}^2 + 4[\gamma_i(1-x) - \gamma_{n+1}x_i]\partial_{x_i} \\ &= L_{0i}^2 + 4[(\gamma_i - \tfrac{1}{2})(1-x) - (\gamma_{n+1} - \tfrac{1}{2})x_i]\partial_{x_i} \\ &= S_{i0}, \qquad 1 \le i \le n, \end{aligned} \tag{1.22}$$

do commute with H: $[S_{ij}, H] = [S_{0i}, H] = 0$. Also

$$8H \equiv \sum_{i,j=1}^{n} S_{ij} + 2\sum_{i=1}^{n} S_{0i}. \tag{1.23}$$

We conjecture, but have not proven, that linear combinations of the S_{ij} and S_{0i} are the only second-order operators commuting with H.

If S is a second-order symmetry operator for H then $S' = \rho^{-1}S\rho$ is a second-order symmetry for $H' = \Delta_n + V_n$ and, necessarily, $S' = \Upsilon + f$ where Υ is a second-order symmetry for Δ_n and f is a real-valued function. Thus S' is a formally self-adjoint operator with respect to the inner product $\langle \cdot, \cdot \rangle$ and S is formally self-adjoint with respect to $(\cdot, \cdot)$.

2. Orthogonal bases of separable solutions. In the paper (Kalnins and Miller [13]) and in the book (Kalnins [11]) all separable coordinates for the equation $\Delta_n\Psi = \lambda\Psi$ are constructed, where Δ_n is the Laplace–Beltrami operator on S^n. It is shown that all separable coordinates are orthogonal and that for each separable coordinate system the corresponding separated solutions are characterized as simultaneous eigenfunctions of a set of n second-order commuting symmetry operators for Δ_n. These operators are real linear combinations of the symmetries L_{ij}^2, $1 \le i < j \le n+1$, where L_{ij} is a rotational generator in so$(n+1)$. For $n = 2$ there are two separable systems (ellipsoidal and spherical coordinates), while for $n = 3$ there are six systems. The number of separable systems grows rapidly with n, but all systems can be constructed through a simple graphical procedure. (In general, the possible separable systems are the various polyspherical coordinates (Vilenkin [26]), the basic ellipsoidal coordinates, and combinations of polyspherical and ellipsoidal coordinates.) Moreover, the equation $(\Delta_n + V_n)\Psi = \lambda\Psi$, where the scalar potential takes the form

$$V_n = \sum_{i=1}^{n} \frac{\alpha_i}{z_i^2} + \frac{\alpha_0}{z_0^2}, \qquad \alpha_0, \alpha_1, \cdots, \alpha_n \text{ const.}, \tag{2.1}$$

is separable in *all* the coordinate systems in which the Laplace–Beltrami eigenvalue equation is separable. (That is, V_n of this form is a *Stäckel multiplier* for all separable coordinate systems on S^n; see Boyer, Kalnins, and Miller [3].) Indeed, the equation

with potential (2.1) is separable in general ellipsoidal coordinates. Since all other coordinates are limiting cases of ellipsoidal coordinates, the conclusion follows. (Note: If each $\alpha_j = -\frac{1}{4}k_j(k_j + m_j - 1)$ where k_j and m_j are nonnegative integers with $m_j \geq 1$, then the equation $(\Delta_n + V_n)\Psi = \lambda\Psi$ can be viewed as a resriction of the Laplace–Beltrami eigenvalue equation $(\Delta_M + V_M)\Psi' = \lambda\Psi'$ on the N-sphere where $N = \sum_{j=0}^{n} m_j + n$, in which the variable dependence on the subspheres S^{m_j} has already been factored out. Moreover, using the canonical equation technique found in Kalnins, Manocha, and Miller [12] we can show that all solutions of the above equation for general γ_i are solutions of the flat-space wave equation in $2n + 2$ dimensions with signature $(n + 1, n + 1)$. Thus the conformal symmetry algebra of the wave equation can be expected to transform solutions of the eigenvalue equations among themselves. Lemma 2 and Corollary 1 below are examples of this action.)

The results of Kalnins and Miller, characterizing separable systems by symmetry operators, can easily be translated to the present case. In those references (for $V_n = 0$) the symmetry operators are given explicitly as linear combinations of the symmetries L_{ij}^2. The results for the potential (1.14) are similar: L_{ij}^2 is replaced by $S'_{ij} = \rho^{-1}S_{ij}\rho$ and takes the same linear combinations. Moreover, since the defining symmetry operators for a separable system are real linear combinations of the L_{ij}^2 plus scalar functions, they are formally self-adjoint with respect to the inner product $\langle \cdot, \cdot \rangle$.

These results can now easily be extended to results for solutions of

$$(\Delta_n + \Lambda_n)\Phi = \lambda\Phi \tag{2.2}$$

through the mappings

$$\begin{aligned} &\Delta_n + \Lambda_n = \rho(\Delta_n + V_n)\rho^{-1} \\ &S_{ij} = \rho S'_{ij}\rho^{-1} \\ &\Phi = \rho\Psi. \end{aligned} \tag{2.3}$$

Thus all separable solutions Ψ map to R-separable solutions Φ of (2.2) (Miller [22]). The R-separable coordinates and solutions are determined by commuting symmetry operators S of $\Delta_n + \Lambda_n$ which are obtained from expressions in Kalnins and Miller [13] or Kalnins [11], where each occurrence of L_{ij}^2 is replaced by S_{ij}. The defining symmetry operators are all formally self-adjoint with respect to the inner product $(\cdot, \cdot)$. Finally, since each S_{ij} maps polynomials of maximum order m_k in x_k to polynomials of the same type, it follows that a basis of separated solutions can be expressed as *polynomials* in the x_i. Since the symmetry operators are self-adjoint, the basis of simultaneous eigenfunctions can be chosen to be *orthogonal*.

We conclude from this argument that every separable coordinate system for the Laplace–Beltrami eigenvalue equation on the n-sphere yields an orthogonal basis of polynomial solutions of equation (1.4), hence an orthogonal basis for all n-variable polynomials with inner product (1.18).

As an example we work out the separation equations for spherical coordinates $\{u_i\}$ on S^n:

$$
\begin{aligned}
z_0^2 &= 1 - x = 1 - u_n \\
z_1^2 &= x_1 = u_1 u_2 \dots u_n \\
z_2^2 &= x_2 = (1 - u_1) u_2 \dots u_n \\
&\vdots \\
z_{n-1}^2 &= x_{n-1} = (1 - u_{n-2}) u_{n-1} u_n \\
z_n^2 &= x_n = (1 - u_{n-1}) u_n .
\end{aligned} \tag{2.4}
$$

(Note that in terms of angles $\{\theta_i\}$ we usually set $u_i = \sin^2 \theta_i$.) It follows that

$$
u_j = \begin{cases} w_j / w_{j+1}, & j = 1, \cdots, n-1 \\ w_n, & j = n \end{cases}, \tag{2.5}
$$

where

$$
w_\ell = \sum_{i=1}^{\ell} x_i .
$$

In terms of the $\{u_i\}$, the operator (1.5) becomes

$$
H = \sum_{i=1}^{n} \frac{1}{u_{i+1} \cdots u_n} \left[u_i (1 - u_i) \partial_{u_i}^2 + \left(\sum_{j=1}^{i} \gamma_j - \left(\sum_{p=1}^{i+1} \gamma_p \right) u_i \right) \partial_{u_i} \right] . \tag{2.6}
$$

Equation (1.4) is separable in these coordinates with separation equations

$$
\begin{aligned}
& u_1 (1 - u_1) \partial_{u_1}^2 \Theta_1 + [\gamma_1 - (\gamma_1 + \gamma_2) u_1] \partial_{u_1} \Theta_1 = c_1 \Theta_1 , \\
& \left[\frac{c_{k-1}}{u_k} + u_k (1 - u_k) \partial_{u_k}^2 \right] \Theta_k + \left[\sum_{j=1}^{k} \gamma_j - \left(\sum_{p=1}^{k+1} \gamma_p \right) u_k \right] \partial_{u_k} \Theta_k = c_k \Theta_k , \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 2, 3, \cdots, n .
\end{aligned} \tag{2.7}
$$

Here $\Theta = \prod_{k=1}^{n} \Theta_k(u_k)$ and the c_i are the separation constants, with $c_n = -M(M + G - 1)$.

Noting that the hypergeometric equation

$$
u(1 - u) \frac{d^2 g}{du^2} + [c - (a + b + 1) u] \frac{dg}{du} - abg = 0
$$

admits the solution

$$
g = {}_2F_1 \left(\begin{matrix} a, & b \\ & c \end{matrix} ; u \right) = \sum_{m=1}^{\infty} \frac{(a)_m (b)_m}{(c)_m m!} u^m ,
$$

a polynomial for $a = 0, -1, -2, \cdots$, and requiring that Θ be a polynomial in the $\{x_i\}$ we obtain the solutions

(2.8)

$$\Theta_1(u_1) = {}_2F_1\left(\begin{matrix}-\ell_1, & \ell_1+\gamma_1+\gamma_2-1\\ & \gamma_1\end{matrix}; u_1\right)$$

$$c_1 = -\ell_1(\ell_1+\gamma_1+\gamma_2-1),$$

$$\Theta_k(u_k) = u_k^{\ell_1+\ell_2+\cdots+\ell_{k-1}}\,{}_2F_1\left(\begin{matrix}-\ell_k, & 2(\ell_1+\cdots+\ell_{k-1})+\ell_k+\gamma_1+\cdots+\gamma_{k+1}-1\\ & 2(\ell_1+\cdots+\ell_{k-1})+\gamma_1+\cdots+\gamma_k\end{matrix}; u_k\right),$$

$$c_k = -(\ell_1+\cdots+\ell_k)(\ell_1+\cdots+\ell_k+\gamma_1+\cdots+\gamma_{k+1}-1),$$

$$k = 2,3,\cdots,n,$$

where $\sum_{i=1}^n \ell_i = M$ and $\ell_i = 0,1,2\cdots$. This determines Θ to within a normalization factor.

In the special case $n = 2$ we have the result of (Proriol [23]) and of (Karlin and McGregor [14]):

$$\begin{aligned}\Theta_{\ell_1,\ell_2}(x_1,x_2) =& {}_2F_1\left(\begin{matrix}-\ell_1, & \ell_1+\gamma_1+\gamma_2-1\\ & \gamma_1\end{matrix}; \frac{x_1}{x_1+x_2}\right)(x_1+x_2)^{\ell_1}\\ &\times {}_2F_1\left(\begin{matrix}-\ell_2, & 2\ell_1+\ell_2+\gamma_1+\gamma_2+\gamma_3-1\\ & 2\ell_1+\gamma_1+\gamma_2\end{matrix}; x_1+x_2\right)\\ &\sim P_{\ell_2}^{(\gamma_3-1,\gamma_1+\gamma_2+2\ell_1-1)}(2x_1+2x_2-1)\,(x_1+x_2)^{\ell_1}\\ &\times P_{\ell_1}^{(\gamma_2-1,\gamma_1-1)}\left(\frac{2x_1}{x_1+x_2}-1\right),\end{aligned} \tag{2.9}$$

where $P_k^{(\alpha,\beta)}(x)$ is a Jacobi polynomial.

Returning to the general case, we have the eigenvalue equations

$$S_\ell\Theta_\ell = c_\ell\Theta_\ell, \qquad \ell = 1,\cdots,n, \tag{2.10}$$

where

$$\begin{aligned}S_1 &= u_1(1-u_1)\partial_{u_1}^2 + [\gamma_1-(\gamma_1+\gamma_2)u_1]\,\partial_{u_1},\\ S_k &= \frac{1}{u_k}S_{k-1} + u_k(1-u_k)\partial_{u_k}^2 + [\gamma_1+\cdots+\gamma_k-(\gamma_1+\cdots+\gamma_{k+1})u_k]\,\partial_{u_k},\\ &\qquad k = 2,3,\cdots,n,\end{aligned} \tag{2.11}$$

and $S_n = H$. Furthermore, $[S_i, S_j] = 0$ and the S_i are self-adjoint with respect to the inner product $(\cdot,\cdot)$. It follows immediately that

$$(\Theta_\ell, \Theta_{\mathbf{m}}) = 0$$

unless $\ell_1 = m_1$, $\ell_2 = m_2,\cdots$, $\ell_n = m_n$. The measure $d\tilde{\omega}$ becomes in these coordinates

$$d\tilde{\omega} = u_1^{\gamma_1-1}u_2^{\gamma_1+\gamma_2-1}\cdots u_n^{\gamma_1+\cdots+\gamma_n-1}(1-u_1)^{\gamma_2-1}(1-u_2)^{\gamma_3-1}\cdots(1-u_n)^{\gamma_{n+1}-1}du_1\cdots du_n,$$

where $0 < u_i < 1$. In terms of the symmetries S_{ij}, S_{0i}, (1.21)–(1.22), we have

$$S_k = \frac{1}{8}\sum_{i,j=1}^{k+1} S_{ij}, \qquad k = 1, \cdots, n-1 \tag{2.12}$$

$$S_n = H = \frac{1}{8}\left(\sum_{h,p=0}^{n} S_{hp}\right),$$

where we set $S_{hh} = 0$.

3. Orthogonal bases for another space of polynomials. Now we make the change of coordinates $x_i = y_i^2$, $1 \le i \le n$, and look for solutions of (1.4) that are polynomials in the y_i. In general, H does not map polynomials in the y_i to polynomials, but in the special case $\gamma_1 = \gamma_2 = \cdots = \gamma_n = \frac{1}{2}$, $G = \gamma_{n+1} + n/2 = s/2 + (n+1)/2$, we have

$$H = \frac{1}{4}\sum_{i,j=1}^{n}(\delta_{ij} - y_i y_j)\partial_{y_i y_j} + \frac{1}{2}\left(\frac{1}{2} - G\right)\sum_{j=1}^{n} y_j \partial_{y_j}, \tag{3.1}$$

and H does map polynomials to polynomials of at most the same degree. Moreover, the differential operators

$$L_{ij} = -L_{ji} = y_i\partial_{y_j} - y_j\partial_{y_i}, \qquad 1 \le i < j \le n, \tag{3.2}$$

commute with H and form a basis for the symmetry algebra so(n). The special second-order symmetries take the form $S_{ij} = L_{ij}^2$, $1 \le i < j \le n$, and

$$S_{0i} = L_{0i}^2 - 2\left(G - \frac{1}{2}\right) y_i\partial_{y_i} = \left(1 - \sum_{j=1}^{n} y_j^2\right)\partial_{y_i}^2 - 2G y_i \partial_{y_i},$$

and clearly map polynomials to polynomials of at most the same degree. The measure takes the form

$$d\tilde{\omega} = (1 - y_1^2 - \cdots - y_n^2)^{s/2-(1/2)} dy_1 \cdots dy_n, \tag{3.3}$$

where $-1 \le y_i \le 1$ and

$$\rho^{-1} = (1 - y_1^2 - \cdots - y_n^2)^{s/4}. \tag{3.4}$$

Again, H and the $S_{\ell k}$ are formally self-adjoint with respect to the inner product

$$(\Phi_1, \Phi_2) = \int \cdots \int_{\sum y_i^2 < 1} \Phi_1(\mathbf{y})\overline{\Phi_2}(\mathbf{y})\, d\tilde{\omega}, \tag{3.5}$$

where Φ_1, Φ_2 are polynomials in the y_i.

Every separable coordinate system for the equation

$$H\Phi = -M(M+G-1)\Phi, \qquad 2M \text{ a nonnegative integer,} \tag{3.6}$$

where H is given by (3.1) yields an orthogonal basis of multivariable polynomials with respect to the inner product $(\cdot,\cdot)$. (For $n = 2$ this equation is also on the list of Krall and Sheffer [20].) Indeed, for spherical coordinates $u_i = \sin^2\theta_i$ we obtain the orthogonal basis of polynomials in $\mathbf{y}$:

(3.7)

$$e^{\pm 2i\ell_1\theta_1}\prod_{k=2}^{n-1}[\sin\theta_k]^{2(\ell_1+\cdots+\ell_{k-1})}C_{2\ell_k}^{2(\ell_1+\cdots+\ell_{k-1})+(k-1)/2}(\cos\theta_k)$$
$$\times u_n^{\ell_1+\cdots+\ell_{n-1}}{}_2F_1\left(\begin{matrix}-\ell_n, & 2(\ell_1+\cdots+\ell_{n-1})+\ell_n+(n-1)/2+s/2\\ & 2(\ell_1+\cdots+\ell_{n-1})+n/2\end{matrix};u_n\right),$$

where $2\ell_i = 0,1,2,\cdots$ for $1 \le i \le n-1$, $\ell_n = 0,1,2,\cdots$, and the $C_k^\lambda(x)$ are Gegenbauer polynomials

$$C_k^\lambda(x) = \frac{(2\lambda)_k}{k!}{}_2F_1\left(\begin{matrix}-k, & k+2\lambda\\ & \lambda+\frac{1}{2}\end{matrix};\tfrac{1}{2}-x/2\right)$$

(Erdélyi et al. [9]). (The eigenvalues are defined as before.)

Using the results of [11] or [13], many other orthogonal bases can be worked out. Moreover, the symmetry group $\mathrm{SO}(n)$ permits the derivation of addition theorems for the basis elements, related to the addition theorem for Gegenbauer polynomials and Koornwinder's addition theorem [15]–[17].

Next we relate the Cartesian coordinates z_ℓ and the y_a via

$$\begin{aligned} z_0^2 &= y_1^2\\ z_1^2 &= y_2^2\\ &\vdots\\ z_{n-1}^2 &= y_n^2\\ z_n^2 &= 1-y_1^2-\cdots-y_n^2, \end{aligned} \tag{3.8}$$

a simple permutation of the relations (2.4), so that the (separable) spherical coordinates v_i are associated with the y_a through

$$v_1 = \frac{y_2^2}{y_2^2+y_3^2}, \quad v_2 = \frac{y_2^2+y_3^2}{y_2^2+y_3^2+y_4^2}, \quad \cdots, \quad v_{n-2} = \frac{y_2^2+\cdots+y_{n-1}^2}{y_2^2+\cdots+y_n^2},$$

$$v_{n-1} = \frac{y_2^2+\cdots+y_n^2}{1-y_1^2}, \qquad v_n = 1-y_1^2. \tag{3.9}$$

From the point of view of separability for the Laplace–Beltrami eigenvalue equation, these v_i coordinates are equivalent to the u_i coordinates introduced earlier, since one system can be obtained from the other through the action of an element of the $\mathrm{SO}(n+1)$ symmetry group for this equation. However, the term Λ_n breaks this

symmetry so that from the viewpoint of the eigenvalue equation for H, with $\gamma_1 = \gamma_2 = \cdots \gamma_n = \frac{1}{2}$, these are distinct coordinates. The separation equations for the v_i are identical to those for the u_i if we interchange $\gamma_n = \frac{1}{2}$ and $\gamma_{n+1} = s/2 + \frac{1}{2}$. For $n = 2$ the orthogonal basis of polynomials is

$$C_{2\ell_1}^{s/2}(\sin\theta_1) C_{2\ell_2}^{2\ell_1+s/2+\frac{1}{2}}(\cos\theta_2)\sin^{2\ell_1}\theta_2, \tag{3.10}$$

where $\ell_1, \ell_2 = 0, \frac{1}{2}, 1, \frac{3}{2}, \cdots$, $\ell_1 + \ell_2 = N$ and $v_j = \sin^2\theta_j$. This is in agreement with the basis of Koschmieder [18], [19].

For $n > 2$ we have an orthogonal basis of polynomials of the form $\Theta = \prod_{k=1}^{n} \Theta_k$, where

(3.11)

$$\Theta_1 = e^{\pm 2i\ell_1\theta_1}, \quad \ell_1 = 0, 1/2, 1, 3/2, \cdots,$$

$$\Theta_k = [\sin\theta_k]^{2(\ell_1+\cdots+\ell_{k-1})} C_{2\ell_k}^{2(\ell_1+\cdots+\ell_{k-1})+(k-1)/2}(\cos\theta_k)$$

$$\ell_k = 0, \tfrac{1}{2}, 1, \cdots, \qquad 1 < k < n-1,$$

$$\Theta_{n-1} = v_{n-1}^{\ell_1+\cdots+\ell_{n-2}} {}_2F_1\left(\begin{matrix} -\ell_{n-1}, & 2(\ell_1+\cdots+\ell_{n-2}) + \ell_{n-1} + (n-2)/2 + s/2 \\ & 2(\ell_1+\cdots+\ell_{n-2}) + (n-1)/2 \end{matrix}; v_{n-1}\right)$$

$$\ell_{n-1} = 0, 1, 2, \cdots,$$

$$\Theta_n = [\sin\theta_n]^{2(\ell_1+\cdots+\ell_{n-1})} C_{2\ell_n}^{2(\ell_1+\cdots+\ell_{n-1})+(n-1)/2+s/2}(\cos\theta_n)$$

$$\ell_n = 0, \tfrac{1}{2}, 1, \cdots,$$

where $\ell_1 + \cdots + \ell_n = M$ and $v_j = \sin^2\theta_j$.

4. The "mixed" case. Next we consider the more general mixed case with variables $x_1, \cdots, x_{n_1}, y_1, \cdots, y_{n_2}$, $n_1 + n_2 = n$, where

$$\begin{aligned} z_0^2 &= 1 - \sum_{i=1}^{n_1} x_i - \sum_{a=1}^{n_2} y_a^2, \\ z_1^2 &= x_1, \\ &\vdots \\ z_{n_1}^2 &= x_{n_1}, \\ z_{n_1+1}^2 &= y_1^2, \\ &\vdots \\ z_n^2 &= y_{n_2}^2, \end{aligned} \tag{4.1}$$

and look for polynomial solutions in x_i, y_a of the equation

$$H\Phi(\mathbf{x}, \mathbf{y}) = -M(M + G - 1)\Phi(\mathbf{x}, \mathbf{y}), \tag{4.2}$$

where $\gamma_{n_1+1} = \gamma_{n_1+2} = \cdots = \gamma_n = \frac{1}{2}$ and

$$
\begin{aligned}
H &= \frac{1}{4}\sum_{a,b}(\delta_{ab}-y_ay_b)\partial_{y_ay_b}+\sum_{i,j}(\delta_{ij}x_i-x_ix_j)\partial_{x_ix_j}\\
&\quad -\sum_{a,i}y_ax_i\partial_{y_ax_i}+\sum_i(\gamma_i-Gx_i)\partial_{x_i}+\frac{1}{2}\sum_a\left(\frac{1}{2}-G\right)y_a\partial_{y_a},\\
G &= \frac{n_2+1}{2}+\sum_i\gamma_i+\frac{s}{2}, \qquad 2M \text{ a nonnegative integer.}
\end{aligned}
\tag{4.3}
$$

For reference,

$$
\begin{aligned}
\Delta_n &= \frac{1}{4}\sum_{a,b}(\delta_{ab}-y_ay_b)\partial_{y_ay_b}+\sum_{i,j}(\delta_{ij}x_i-x_ix_j)\partial_{x_ix_j}\\
&\quad -\sum_{a,i}y_ax_i\partial_{y_ax_i}+\frac{1}{2}\sum_i(1-(n+1)x_i)\,\partial_{x_i}-\frac{n}{4}\sum_a y_a\partial_{y_a},\\
\Lambda_n &= \sum_i\left[\gamma_i-\frac{1}{2}+\left(\frac{n+1}{2}-G\right)x_i\right]\partial_{x_i}+\frac{1}{2}\sum_a\left(\frac{n+1}{2}-G\right)y_a\partial_{y_a}.
\end{aligned}
\tag{4.4}
$$

Note that H maps polynomials in x_i, y_a to polynomials of at most the same order. The induced measure is

$$
\begin{aligned}
&d\tilde{\omega} = x_2^{\gamma_1-1}\cdots x_{n_2}^{\gamma_{n_1}-1}(1-\sum x_i-\sum y_a^2)^{s/2-\frac{1}{2}}dx_1\cdots dx_{n_1}dy_1\cdots dy_{n_2},\\
&0<x_i, \quad -1<y_a<1, \quad \sum_i x_i+\sum_a y_a^2<1,
\end{aligned}
\tag{4.5}
$$

and

$$
\rho^{-1} = x_1^{\gamma_1/2-1/4}\cdots x_{n_1}^{\gamma_{n_1}/2-1/4}\left(1-\sum_i x_i-\sum_a y_a^2\right)^{s/4}.
$$

Equation (4.2) admits the symmetry algebra $\mathrm{so}(n_2)$ with basis

$$
L_{ab} = -L_{ba} = y_a\partial_{y_b}-y_b\partial_{y_a}, \qquad 1\le a<b\le n_2.
$$

The operators H and S_{mk} are formally self-adjoint on the space of polynomials in x_i, y_a with respect to the inner product

$$
(\Phi_1,\Phi_2) = \int\cdots\int_{0<x_i,\sum_i x_i+\sum_a y_a^2<1}\Phi_1(\mathbf{x},\mathbf{y})\overline{\Phi_2}(\mathbf{x},\mathbf{y})\,d\tilde{\omega}.
$$

However, in general the S_{mk} do not map a polynomial to one of the same or lower order in each variable, e.g.,

$$
S_{ia} = 4x_iy_a^2\partial_{x_i}^2+x_i\partial_{y_a}^2-4x_iy_a\partial_{x_iy_a}+4\gamma_iy_a^2\partial_{x_i}-2x_i\partial_{x_i}-2\gamma_iy_a\partial_{y_a},
$$

although they do map polynomials to polynomials. It is still true that each symmetry operator S maps a polynomial eigenspace of H into itself.

It follows that all separable coordinate systems for the n-sphere yield bases of orthogonal polynomials in the mixed case (indeed multiple sets of such bases, depending on the ordering of the variables x_i, y_a). For example, if we choose spherical coordinates $u_\ell = \sin^2\theta_\ell$ in the form

$$u_\ell = \frac{w_\ell}{w_{\ell+1}},$$

where

$$w_\ell = \begin{cases} \sum_{i=1}^{\ell} x_i, & \ell = 1,\cdots,n_1 \\ \sum_{i=1}^{n_1} x_i + \sum_{a=1}^{\ell-n_1} y_a^2, & \ell = n_1+1,\cdots,n_1+n_2 \\ 1, & \ell = n_1+n_2+1 \end{cases} \tag{4.6}$$

we find the orthogonal basis of polynomials:

$$\Theta = \prod_{k=1}^{n} \Theta_k(u_k),$$

where

(4.7)

$$\Theta_k(u_k) = u_k^{\ell_1+\cdots+\ell_{k-1}}\,{}_2F_1\left(\begin{matrix}-\ell_k, & 2(\ell_1+\cdots+\ell_{k-1})+\ell_k+\gamma_1+\cdots+\gamma_{k+1}-1 \\ & 2(\ell_1+\cdots+\ell_{k-1})+\gamma_1+\cdots+\gamma_k\end{matrix}; u_k\right),$$

$$c_k = -(\ell_1+\cdots+\ell_k)(\ell_1+\cdots+\ell_k+\gamma_1+\cdots+\gamma_{k+1}-1),$$

$$k = 1,\cdots,n_1,$$

$$\Theta_k(u_k) = [\sin\theta_k]^{2(\ell_1+\cdots+\ell_{k-1})} C_{2\ell_k}^{2(\ell_1+\cdots+\ell_{k-1})+\sum_{i=1}^{k}\gamma_i-\frac{1}{2}}(\cos\theta_k),$$

$$c_k = -(\ell_1+\cdots+\ell_k)(\ell_1+\cdots+\ell_k+\gamma_1+\cdots+\gamma_{k+1}-1),$$

$$k = n_1+1,\cdots,n_2+n_1-1,$$

$$\Theta_n(u_n) = u_n^{\ell_1+\cdots+\ell_{n-1}}\,{}_2F_2\left(\begin{matrix}-\ell_n, & 2(\ell_1+\cdots+\ell_{n-1})+\ell_n+G-2 \\ & 2(\ell_1+\cdots+\ell_{n-1})+G-s/1-\frac{1}{2}\end{matrix}; u_n\right).$$

Here $\ell_1,\cdots,\ell_{n_1}$, ℓ_n and $2\ell_{n_1+1},\cdots,2\ell_{n_1+n_2-1}$ are nonnegative integers. (Recall that $\gamma_{n_1+1} = \gamma_{n_1+2} = \cdots = \gamma_{n_1+n_2} = \frac{1}{2}$.)

5. Biorthogonal families of polynomials on S^n. We begin this section with a simplified proof of the biorthogonality of the polynomials (1.1) and (1.2) with respect to the inner product $(\cdot,\cdot)_\gamma$; see (1.18), (1.19). Let $\mathscr{S}_\gamma$ be the space of all polynomials in $x_1,\cdots,x_n$ with respect to this inner product and let $\mathscr{H}_{\gamma,M}$ be the subspace of $\mathscr{S}_\gamma$ consisting of solutions Φ to the eigenvalue equation

$$H\Phi = -M(M+G-1)\Phi, \tag{5.1}$$

where H is given by (1.5). Since the functions

$$D_{\mathbf{m}}^{\gamma}(\mathbf{x}) = F_A\left[\begin{matrix}M+G-1; & -m_1,\cdots,-m_n \\ & \gamma_1,\cdots,\gamma_n\end{matrix}; x_1,\cdots,x_n\right] \tag{5.2}$$

clearly satisfy (5.1) for $M = \sum_{i=1}^{n} m_i$ and since the highest order monomial in these solutions is $x_1^{m_1}\cdots x_n^{m_n}$ it follows that the $D_{\mathbf{m}}^{\gamma}$ for $m_1+\cdots+m_n \equiv m = M$ form

a basis for $\mathscr{H}_{\gamma,M}$ and, as the m_i range over all nonnegative integers, a basis for $\mathscr{S}_\gamma$. (Note: $\dim \mathscr{H}_{\gamma,M} = \binom{M+n-1}{n-1}$.) Since H is self-adjoint we have $\mathscr{H}_{\gamma,M} \perp \mathscr{H}_{\gamma,M'}$ for $M' \neq M$. Thus

$$(D^\gamma_{\mathbf{m}'}, D^\gamma_{\mathbf{m}})_\gamma = 0 \quad \text{for } m' \neq m. \tag{5.3}$$

It is simple to verify the recurrence relation

$$\partial_{x_i} D^\gamma_{\mathbf{m}}(\mathbf{x}) = \frac{(M+G-1)(-m_i)}{\gamma_i} D^{\hat{\gamma}}_{\hat{\mathbf{m}}}(\mathbf{x}), \tag{5.4}$$

where

$$\hat{\gamma}_j = \begin{cases} \gamma_j & \text{for } j \neq i,\ 1 \leq j \leq n \\ \gamma_i + 1 & \text{for } j = i \\ \gamma_{n+1} + 1 & \text{for } j = n+1 \end{cases}$$

$$\hat{m}_j = \begin{cases} m_j & \text{for } j \neq i,\ 1 \leq j \leq n \\ m_i - 1 & \text{for } j = i \end{cases} \tag{5.5}$$

$$\hat{M} = M - 1, \qquad \hat{G} = G + 2.$$

We can consider $P_i = \partial_{x_i}$ as an operator

$$P_i : \mathscr{S}_\gamma \to \mathscr{S}_{\hat{\gamma}}.$$

Indeed we have the following lemma.

LEMMA 2. *P_i, $(1 \leq i \leq n)$, maps $\mathscr{H}_{\gamma,M}$ onto $\mathscr{H}_{\hat{\gamma},\hat{M}}$.*

Proof. The proof is immediate from (5.4). For a basis free proof we can easily verify the operator identity

$$\hat{H}P_i = GP_i + P_iH, \tag{5.6}$$

where $\hat{H}$ is the operator H with the γ_j replaced by $\hat{\gamma}_j$. Then if $H\Phi = -M(M+G-1)\Phi$ we have $\hat{H}(P_i\Phi) = -\hat{M}(\hat{M}+\hat{G}-1)(P_i\Phi)$. The null space of P_i acting on $\mathscr{H}_{\gamma,M}$ is of dimension $\binom{M+n-2}{n-2}$ for $n \geq 2$, hence the dimension of the range of P_i is

$$\binom{M+n-1}{n-1} - \binom{M+n-2}{n-2} = \binom{M+n-2}{n-1} = \dim \mathscr{H}_{\hat{\gamma},\hat{M}}. \qquad \square$$

COROLLARY 1. *The operator $P_i - P_j$ maps $\mathscr{H}_{\gamma,M}$ into $\mathscr{H}_{\tilde{\gamma},\tilde{M}}$, where $1 \leq i < j \leq n$ and*

$$\tilde{\gamma}_k = \begin{cases} \gamma_k & \text{for } 1 \leq k \leq n+2,\ k \neq i,j \\ \gamma_i + 1 & \text{for } k = i \\ \gamma_j + 1 & \text{for } k = j \end{cases}$$

$$\tilde{M} = M - 1, \qquad \tilde{G} = G + 2.$$

Proof.

$$\tilde{H}(P_i - P_j) = G(P_i - P_j) + (P_i - P_j)H.$$

Thus if $H\Phi = -M(M+G-1)\Phi$ we have

$$\tilde{H}([P_i - P_j]\Phi) = -\tilde{M}(\tilde{M}+\tilde{G}-1)[P_i - P_j]\Phi. \qquad \square$$

The operator P_i induces an adjoint operator $P_i^* : \mathscr{S}_{\hat{\gamma}} \to \mathscr{S}_\gamma$, defined by

$$(P_i^*\Phi, \Phi')_\gamma = (\Phi, P_i\Phi')_{\hat{\gamma}}$$

for all $\Phi \in \mathscr{S}_{\hat{\gamma}}$, $\Phi' \in \mathscr{S}_\gamma$. A straightforward computation yields

$$P_i^* = -x_i(1-x)\partial_{x_i} - \gamma_i(1-x) + \gamma_{n+1}x_i. \tag{5.7}$$

THEOREM 1. *P_i^* is a 1-1 map of $\mathscr{H}_{\hat{\gamma},\hat{M}}$ into $\mathscr{H}_{\gamma,M}$.*

Proof. Taking the adjoint of the relation (5.4) we obtain

$$P_i^*\hat{H} = GP_i^* + HP_i^*.$$

Furthermore, P_i^* is 1-1 since P_i is onto. □

Let

$$C_{\mathbf{m}}^\gamma(\mathbf{x}) = (P_1^*)^{m_1} \cdots (P_n^*)^{m_n} 1 \in \mathscr{S}_\gamma$$

be the result of applying m_n operators $P_n^*, \cdots, m_1$ operators P_1^*, one at a time, to the function $1 \in \mathscr{S}_{\gamma'}$, where

$$\gamma_i' = \gamma_i + m_i, \qquad 1 \le i \le n,$$
$$\gamma_{n+1}' = \gamma_{n+1} + m.$$

(Each time an operator P_j^* is applied it lowers γ_j and γ_{n+1} by 1 and leaves the other γ_k's unchanged. The order in which these operators are applied makes no difference in the result.) It follows from the recurrence relation

$$\left(x_i \sum_{j=1}^n x_j\partial_{x_j} - x_i\partial_{x_i} + x_i(-M-\gamma_{n+1}+1) - \gamma_i + 1\right) \tag{5.8}$$
$$\times F_A\left[\begin{matrix} -M-\gamma_{n+1}+1; & -m_1, \cdots, -m_n \\ & \gamma_1, \cdots, \gamma_n \end{matrix}; x_1, \cdots, x_n\right]$$
$$= (1-\gamma_i)F_A\left[\begin{matrix} -(M+1)-(\gamma_{n+1}-1)+1; & -m_1, \cdots, -(m_i+1), \cdots, -m_n \\ & \gamma_1, \cdots, \gamma_i - 1, \cdots, \gamma_n \end{matrix}; x_1, \cdots, x_n\right]$$

and a simple induction argument that

$$C_{\mathbf{m}}^\gamma(\mathbf{x}) = c_{\gamma,m}(1-x)^M F_A\left[\begin{matrix} -M-\gamma_{n+1}+1; & -m_1, \cdots, -m_n \\ & \gamma_1, \cdots, \gamma_n \end{matrix}; -\frac{x_1}{1-x}, \cdots, -\frac{x_n}{1-x}\right], \tag{5.9}$$

where $c_{\gamma,m}$ is a nonzero constant. It follows from Theorem 1 that the $C_{\mathbf{m}}^\gamma$ belong to $\mathscr{H}_{\gamma,M}$ for $M = m_1 + \cdots + m_n$. Since there are $\binom{M+n-1}{n-1}$ of these functions for fixed M and since they are clearly linearly independent, they form a basis for $\mathscr{H}_{\gamma,M}$.

Now consider the inner product

$$(C_{\mathbf{m}}^\gamma, D_{\mathbf{m}'}^\gamma)_\gamma.$$

If $m = M \neq m' = M'$ the inner product vanishes, since $\mathscr{H}_{\gamma,M} \perp \mathscr{H}_{\gamma,M'}$. If $m = m'$ but $\mathbf{m} \not\equiv \mathbf{m}'$ then $m_i > m_i'$ for some i. Thus

$$(C_{\mathbf{m}}^\gamma, D_{\mathbf{m}'}^\gamma)_\gamma = \kappa\,(1, P_1^{m_1} \cdots P_n^{m_n} D_{\mathbf{m}'}^\gamma)_{\gamma'} = 0$$

since $P_i^{m_i} D_{\mathbf{m}'}^\gamma = 0$. (Here, κ is a nonzero constant.) We conclude that the set $\{C_{\mathbf{m}}^\gamma, D_{\mathbf{m}'}^\gamma\}$ is biorthogonal. (This family is a generalization of biorthogonal polynomials in two variables studied by Appell and Kampé de Fériet [1] and extended by Fackerell and Littler [10].)

Note that the norm of the weight function is

(5.10)
$$\int_0^1 dx_1 \int_0^{1-x_1} dx_2 \cdots \int_0^{1-x_1-\cdots-x_{n-1}} dx_n \left[\prod_{k=1}^n x_k^{\gamma_k-1}\right](1-x)^{\gamma_{n+1}-1}$$
$$= (1,1)_\gamma = \frac{\left[\prod_{k=1}^{n+1}\Gamma(\gamma_k)\right]}{\Gamma(G)}.$$

The relation
$$(P_i^* C_{\hat{\mathbf{m}}}^{\hat{\gamma}}, D_{\mathbf{m}'}^{\gamma})_\gamma = (C_{\hat{\mathbf{m}}}^{\hat{\gamma}}, P_i D_{\mathbf{m}'}^{\gamma})_{\hat{\gamma}}$$
yields (for $\mathbf{m} = \mathbf{m}'$) the recurrence relation
$$(C_{\mathbf{m}}^{\gamma}, D_{\mathbf{m}}^{\gamma})_\gamma = -\frac{m_i(M+G-1)}{\gamma_i}(C_{\hat{\mathbf{m}}}^{\hat{\gamma}}, D_{\hat{\mathbf{m}}}^{\hat{\gamma}})_{\hat{\gamma}}.$$
The normalization of the biorthogonal basis can be obtained from this result and (5.10).

Now we extend the biorthogonality relations to the full n-sphere. We make the change of variables
$$x_k = y_k^2, \qquad k = 1, 2, \cdots, n$$
in (5.10) and extend the domain of integration to negative values of y_k, since the integrand is even in all variables, to get
$$\int_{-1}^1 dy_1 \int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} dy_2 \cdots \int_{-\sqrt{1-y_1^2-\cdots-y_{n-1}^2}}^{\sqrt{1-y_1^2-\cdots-y_{n-1}^2}} dy_n$$
$$\cdot \left[\prod_{k=1}^n (y_k^2)^{\gamma_k-1/2}\right](1-y_1^2-\cdots-y_n^2)^{s/2-1/2}$$
$$= (1,1)'_\gamma = \frac{[\prod_{k=1}^n \Gamma(\gamma_k)]\Gamma(\frac{s}{2}+\frac{1}{2})}{\Gamma(\gamma_1+\cdots+\gamma_n+\frac{s}{2}+\frac{1}{2})}. \tag{5.11}$$

Here we have set $\gamma_{n+1} = \frac{s}{2} + \frac{1}{2}$. (This is a generalization of the weight function for the biorthogonal family $\{V_m^{(s)}(\mathbf{x}), U_m^{(s)}(\mathbf{x})\}$ on the n-sphere of [1], which is obtained by setting $\gamma_1 = \cdots = \gamma_n = \frac{1}{2}$.) Under this change of variables the polynomials $\{C_{\mathbf{m}}^{\gamma}, D_{\mathbf{m}'}^{\gamma}\}$ become

(5.12)
$$U_{2m}^{(\gamma,s)}(\mathbf{y}) = (1-y_1^2-\cdots-y_n^2)^M$$
$$\times F_A\left(\begin{matrix}-M-s/2+\frac{1}{2}; & -m_1,\cdots,-m_n \\ \gamma_1,\cdots,\gamma_n & \end{matrix}; \frac{-y_1^2}{1-y_1^2-\cdots-y_n^2},\cdots,\frac{-y_n^2}{1-y_1^2-\cdots-y_n^2}\right),$$
$$V_{2m}^{(\gamma,s)}(\mathbf{y}) = F_A\left(\begin{matrix}M+\gamma_1+\cdots+\gamma_n+s/2-\frac{1}{2}; & -m_1,\cdots,-m_n \\ \gamma_1,\cdots,\gamma_n & \end{matrix}; y_1^2,\cdots,y_n^2\right).$$

In the special case $\gamma_2 = \cdots = \gamma_n = \frac{1}{2}$ these are exactly the $U_m^{(s)}(\mathbf{y})$ and $V_m^{(s)}(\mathbf{y})$ of [1, p. 269]. (To see this transform $m_k \to m'_k/2$, reverse the order of the sums in F_A by transforming the summation indices as $j_k \to m'_k/2 - j_k$, and then use the reflection formula $\Gamma(z)\Gamma(1-z) = \pi/\sin(\pi z)$ to represent these polynomials in terms of F_B, as given on p. 269.) The biorthogonality demonstration given above immediately implies
$$(V_{2m}^{(\gamma,s)}, U_{2m'}^{(\gamma,s)})'_\gamma \sim \prod_{k=1}^n \delta_{m_k m'_k}, \tag{5.13}$$

where

$$(V,U)'_\gamma = \int\cdots\int_{y_1^2+\cdots+y_n^2<1}\left[\prod_{k=1}^{n}(y_k^2)^{\gamma_k-1/2}\right] \cdot(1-y_1^2-\cdots-y_n^2)^{s/2-1/2}V(\mathbf{y})\overline{U}(\mathbf{y})dy_1\cdots dy_n. \tag{5.14}$$

Also, since the operator H is self-adjoint with respect to this inner product and since $U_m^{(s)}$ and $V_m^{(s)}$ are eigenfunctions of H we have

$$(V_{2m}^{(\gamma,s)}, V_{2m'}^{(\gamma,s)})'_\gamma = (U_{2m}^{(\gamma,s)}, U_{2m'}^{(\gamma,s)})'_\gamma = 0 \quad \text{if } M \neq M'. \tag{5.15}$$

Here, $U_{2m}^{(s)}(\mathbf{y})$ and $V_{2m}^{(s)}(\mathbf{y})$ are strictly even degree in all the variables y_k with total degree $2M$. We define odd degree polynomials as follows:

$$V_{2m+1}^{(\gamma,s)}(\mathbf{y}) \equiv \left[\prod_{k\in Q} y_k\right] V_{2m}^{(\gamma',s)}(\mathbf{y}),$$

$$U_{2m+1}^{(\gamma,s)}(\mathbf{y}) \equiv \left[\prod_{k\in Q} y_k\right] U_{2m}^{(\gamma',s)}(\mathbf{y}), \tag{5.16}$$

where Q is any subset of $(1,2,\cdots,n)$, and

$$\begin{aligned} \gamma'_k &= \gamma_k + 1 \quad \text{if } k \in Q, \\ \gamma'_k &= \gamma_k \quad \text{if } k \notin Q. \end{aligned} \tag{5.17}$$

Since the weight function is even in all variables and the odd degree polynomials are odd in the variables y_k, $k \in Q$, we immediately deduce by parity

$$\begin{aligned} \left(V_{2m+1}^{(\gamma,s)}, V_{2m'}^{(\gamma,s)}\right)'_\gamma &= \left(U_{2m+1}^{(\gamma,s)}, U_{2m'}^{(\gamma,s)}\right)'_\gamma = 0, \\ \left(V_{2m+1}^{(\gamma,s)}, U_{2m'}^{(\gamma,s)}\right)'_\gamma &= \left(V_{2m}^{(\gamma,s)}, U_{2m'+1}^{(\gamma,s)}\right)'_\gamma = 0. \end{aligned} \tag{5.18}$$

Also, $(V_{2m+1}^{(\gamma,s)}, U_{2m'+1}^{(\gamma,s)})'_\gamma$ vanishes by parity unless both polynomials are odd in exactly the same variables, in which case it is easy to verify that

$$\left(V_{2m+1}^{(\gamma,s)}, U_{2m'+1}^{(\gamma,s)}\right)'_\gamma = \left(V_{2m+1}^{(\gamma',s)}, U_{2m'+1}^{(\gamma',s)}\right)'_{\gamma'} \sim \prod_{k=1}^{n}\delta_{m_k m'_k}. \tag{5.19}$$

Similarly,

$$\left(V_{2m+1}^{(\gamma,s)}, V_{2m'+1}^{(\gamma,s)}\right)'_\gamma = \left(U_{2m+1}^{(\gamma,s)}, U_{2m'}^{(\gamma,s)}\right)'_\gamma = 0 \quad \text{if } M \neq M'. \tag{5.20}$$

THEOREM 2. *Let*

$$V_m^{(\gamma,s)}(\mathbf{y}) \equiv \begin{cases} V_{2q}^{(\gamma,s)}(\mathbf{y}) \\ V_{2q+1}^{(\gamma,s)}(\mathbf{y}), \end{cases}$$

$$U_m^{(\gamma,s)}(\mathbf{y}) \equiv \begin{cases} U_{2q}^{(\gamma,s)}(\mathbf{y}) \\ U_{2q+1}^{(\gamma,s)}(\mathbf{y}). \end{cases}$$

Then

$$\left(V_m^{(\gamma,s)}, U_{m'}^{(\gamma,s)}\right)'_\gamma \sim \prod_{k=1}^{n} \delta_{m_k m'_k},$$
$$\left(V_m^{(\gamma,s)}, V_{m'}^{(\gamma,s)}\right)'_\gamma = \left(U_m^{(\gamma,s)}, U_{m'}^{(\gamma,s)}\right)'_\gamma = 0 \quad \text{if } M \neq M'.$$

In the case $n = 1$ the biorthogonal polynomials are orthogonal:

$$V_{2m}^{(\gamma,s)}(y) = U_{2m}^{(\gamma,s)}(y) = {}_2F_1\left(\begin{matrix} m+\gamma+s/2-\frac{1}{2}, -m \\ \gamma \end{matrix}; y^2\right), \tag{5.21}$$
$$V_{2m+1}^{(\gamma,s)}(y) = U_{2m+1}^{(\gamma,s)}(y) = y\ {}_2F_1\left(\begin{matrix} m+\gamma+s/2+\frac{1}{2}, -m \\ \gamma+1 \end{matrix}; y^2\right).$$

The measure on the interval $-1 \leq y \leq 1$ is

$$d\omega(y) = (y^2)^{\gamma-1/2}(1-y^2)^{s/2-1/2}\, dy.$$

For $\gamma = \frac{1}{2}$ these are exactly the Gegenbauer polynomials. For general γ they are a generalization of these polynomials [4, p. 256].

The same construction with $U_m = V_m$ can be carried out for *all* the orthogonal systems of polynomials in the variables x_k as found in §2 to obtain orthogonal polynomials in the variables y_k on the full n-sphere. In general, something is lost in this construction, however. The polynomials $U_m = V_m$ are (except for the even case) no longer eigenfunctions of H. Indeed, we have the following lemma.

LEMMA 3. *Let $\Phi(\mathbf{y})$ be a polynomial eigenfunction of H:*

$$H\Phi = -M(M+G-1)\Phi$$

in the coordinates y_k, where $x_k = y_k^2$, $1 \leq k \leq n$, and let Q be a subset of $\{1, 2, \cdots, n\}$ with $|Q| > 0$ elements. Then $\Psi_Q \equiv [\prod_{i\in Q} y_i]\Phi(\mathbf{y})$ is an eigenfunction of the operator H' corresponding to parameters $\gamma'_k, \gamma'_{n+1}, G'$ if and only if

$$\gamma_k = \tfrac{3}{2} \quad \text{for } k \in Q$$

and

$$\gamma'_k = \tfrac{1}{2} \quad \text{for } k \in Q,$$
$$\gamma'_k = \gamma_k \quad \text{for } k \notin Q,$$
$$G' = G - |Q|, \qquad M' = M + \frac{|Q|}{2}.$$

Then

$$H'\Psi_Q = -M'(M'+G'-2)\Psi_Q.$$

It follows from this result that in the case where $\gamma_1 = \cdots = \gamma_n = \frac{1}{2}$, the construction leading to Theorem 2 yields the biorthogonal polynomials $U_m^{(s)}(\mathbf{y})$ and $V_m^{(s)}(\mathbf{y})$ of [1]. These polynomials are all eigenfunctions of H. Similarly, for $\gamma_1 = \cdots = \gamma_n = \frac{1}{2}$ the same construction applied to the families of orthogonal polynomials in x_k, found in §2, leads to the families of orthogonal polynomials in y_k, found in §3, all eigenfunctions of H.

As a referee has kindly pointed out, Lemma 3 can be generalized if we use Dunkl's differential-difference operator [5]. In the coordinates y_i and for general $\gamma_1, \cdots, \gamma_{n+1}$,

Dunkl's operator $\tilde{H}$ is defined as

$$\tilde{H}p(\mathbf{y}) = \frac{1}{4}\Big[\sum_{i,j=1}^{n}(\delta_{ij} - y_i y_j)\partial_{y_i y_j} p + (1-2G)\sum_{j=1}^{n} y_j \partial_{y_j} p + \sum_{j=1}^{n}\left(\gamma_j - \frac{1}{2}\right)\left(\frac{2}{y_j}\partial_{y_j} p - \frac{p(\mathbf{y}) - p(\cdots, -y_j, \cdots)}{y_j^2}\right)\Big].$$

(This differs from the operator (1.5) with $x_j = y_j^2$ only in the last term.) The eigenvalue equation is

$$\tilde{H}p(\mathbf{y}) = -M(M+G-1)p(\mathbf{y}).$$

Note that $\tilde{H}$ always maps polynomials in the y_i to polynomials and that $\tilde{H}p \equiv Hp$ for polynomials p which are even in each of the variables y_j and $\tilde{H} \equiv H$ if $\gamma_j = \frac{1}{2}$ for all j. Furthermore, since the operators I_j, which map $p(\mathbf{y})$ to $p(y_1, \cdots, -y_j, \cdots, y_n)$ for $j = 1, \cdots, n$, commute with $\tilde{H}$, we can assume, without loss of generality, that each eigenfunction is either even or odd in every one of its variables y_j. We have the following generalization of Lemma 3.

LEMMA 3′. *Let $\Phi(\mathbf{y})$ be a polynomial eigenfunction of $\tilde{H}$:*

$$\tilde{H}\Phi = -M(M+G-1)\Phi$$

in the coordinates y_k, where $x_k = y_k^2$, $1 \le k \le n$, and let Q be a subset of $\{1, 2, \cdots, n\}$ with $|Q| > 0$ elements. Then $\Psi_Q \equiv [\prod_{i\in Q} y_i]\Phi(\mathbf{y})$ is an eigenfunction of the operator $\tilde{H}'$ corresponding to parameters $\gamma_k', \gamma_{n+1}', G'$ if and only if

$$\gamma_k' = \gamma_k - 1 \quad \text{for } k \in Q,$$

and

$$\gamma_k' = \gamma_k \quad \text{for } k \notin Q,$$
$$G' = G - |Q|, \qquad M' = M + \frac{|Q|}{2}.$$

Then

$$\tilde{H}'\Psi_Q = -M'(M'+G'-1)\Psi_Q.$$

Similar comments apply to the "mixed" case in §6.

6. The "mixed" biorthogonal case. Using the techniques introduced in §5 we can now easily determine a biorthogonal basis of polynomials in the mixed case with coordinates (4.1). We set

$$n_1 + n_2 = n, \quad x \equiv \sum_{k=1}^{n_1} x_k, \quad y^2 \equiv \sum_{k=1}^{n_2} y_k^2,$$
$$M \equiv \sum_{k=1}^{n_1} m_k, \quad \tilde{M} \equiv \sum_{k=1}^{n_2} \tilde{m}_k.$$

The basic building blocks are the polynomials

(6.1)
$$C_{m,2\tilde{m}}^{(\gamma,s)}(\mathbf{x},\mathbf{y}) = (1-x-y^2)^{M+\tilde{M}} F_A \cdot \left(\begin{matrix} -M-\tilde{M}-s/2+\frac{1}{2}; -m_k, -\tilde{m}_k \\ \gamma_k, \quad s_k \end{matrix}; \frac{-x_k}{1-x-y^2}, \frac{-y_k^2}{1-x-y^2}\right)$$

and

(6.2)

$$D^{(\gamma,s)}_{m,2\tilde{m}}(\mathbf{x},\mathbf{y}) = F_A\left(\begin{matrix} M+\tilde{M}+\gamma_1+\cdots+\gamma_{n_1}+s_1+\cdots+s_{n_2}+s/2-\frac{1}{2}; -m_k, -\tilde{m}_k \\ \gamma_k, \quad s_k \end{matrix}; x_k, y_k^2\right).$$

The weight function is

$$w(\mathbf{x},\mathbf{y}) = \left[\prod_{k=1}^{n_1} x_k^{\gamma_k-1}\right]\left[\prod_{k=1}^{n_2}(y_k^2)^{s_k-1/2}\right](1-x-y^2)^{s/2-1/2} \tag{6.3}$$

with $\gamma_k, s_k > 0$ and $s > -1$. The inner product is

(6.4)

$$\langle \Phi_2, \Phi_2\rangle_{\gamma,s} = \int\cdots\int_{0<x_i,x+y^2<1} \Phi_1(\mathbf{x},\mathbf{y})\overline{\Phi_2}(\mathbf{x},\mathbf{y})\; w(\mathbf{x},\mathbf{y})dx_1\cdots dx_{n_1}dy_1\cdots dy_{n_2}.$$

Furthermore,

$$\langle 1,1\rangle_{\gamma,s} = \frac{[\prod_{k=1}^{n_1}\Gamma(\gamma_k)][\prod_{k=1}^{n_2}\Gamma(s_k)]\Gamma(s/2+1/2)}{\Gamma(\gamma_1+\cdots+\gamma_{n_1}+s_1+\cdots+s_{n_2}+s/2+1/2)}. \tag{6.5}$$

It follows from the results immediately preceding (5.10) that the polynomial sets (6.1) and (6.2) are biorthogonal. However, since they are even functions of the y_k they do not form a basis for all polynomial functions in the variables x_k, y_k. To construct such a basis we define functions

$$\begin{aligned} C^{(\gamma,s)}_{m,2\tilde{m}+1}(\mathbf{x},\mathbf{y}) &= \left[\prod_{k\in Q} y_k\right] C^{(\gamma,s_k+1,s)}_{m,2\tilde{m}}(\mathbf{x},\mathbf{y}), \\ D^{(\gamma,s)}_{m,2\tilde{m}+1}(\mathbf{x},\mathbf{y}) &= \left[\prod_{k\in Q} y_k\right] D^{(\gamma,s_k+1,s)}_{m,2\tilde{m}}(\mathbf{x},\mathbf{y}), \end{aligned} \tag{6.6}$$

where Q is any nonempty subset of $(1,2,\cdots,n_2)$.

By parity we have

$$\begin{aligned} &\langle C^{(\gamma,s)}_{m,2\tilde{m}+1}, D^{(\gamma,s)}_{m',2\tilde{m}'}\rangle_{\gamma,s} = 0, \qquad \langle C^{(\gamma,s)}_{m,2\tilde{m}}, D^{(\gamma,s)}_{m',2\tilde{m}'+1}\rangle_{\gamma,s} = 0, \\ &\langle C^{(\gamma,s)}_{m,2\tilde{m}+1}, D^{(\gamma,s)}_{m',2\tilde{m}'+1}\rangle_{\gamma,s} = 0 \quad \text{if } Q \neq Q'. \end{aligned}$$

If $Q = Q'$ a simple computation yields

$$\langle C^{(\gamma,s)}_{m,2\tilde{m}+1}, D^{(\gamma,s)}_{m',2\tilde{m}'+1}\rangle_{\gamma,s} = \langle C^{(\gamma,s_k+1,s)}_{m,2\tilde{m}}, D^{(\gamma,s_k+1,s)}_{m',2\tilde{m}'}\rangle_{\gamma,s_k+1,s} \sim \prod_{k=1}^{n_1}\delta_{m_k m'_k}\prod_{k=1}^{n_2}\delta_{\tilde{m}_k,\tilde{m}'_k}.$$

Since $C^{(\gamma,s)}_{m,2\tilde{m}}$ and $D^{(\gamma,s)}_{m,2\tilde{m}}$ are eigenfunctions of H there are additional orthogonality relations obeyed by the C's alone and by the D's alone. Collecting all these results, we have the following theorem.

THEOREM 3. *Let*

$$\begin{aligned} C^{(\gamma,s)}_{m,\tilde{m}}(\mathbf{x},\mathbf{y}) &\equiv \begin{cases} C^{(\gamma,s)}_{m,2\tilde{q}}(\mathbf{x},\mathbf{y}) \\ C^{(\gamma,s)}_{m,2\tilde{q}+1}(\mathbf{x},\mathbf{y}), \end{cases} \\ D^{(\gamma,s)}_{m,\tilde{m}}(\mathbf{x},\mathbf{y}) &\equiv \begin{cases} D^{(\gamma,s)}_{m,2\tilde{q}}(\mathbf{x},\mathbf{y}) \\ D^{(\gamma,s)}_{m,2\tilde{q}+1}(\mathbf{x},\mathbf{y}). \end{cases} \end{aligned}$$

Then

$$\langle C^{(\gamma,s)}_{m,\tilde{m}}, D^{(\gamma,s)}_{m',\tilde{m}'}\rangle_{\gamma,s} \sim \prod_{k=1}^{n_1} \delta_{m_k m'_k} \prod_{k=1}^{n_2} \delta_{\tilde{m}_k \tilde{m}'_k},$$

$$\langle C^{(\gamma,s)}_{m,\tilde{m}}, C^{(\gamma,s)}_{m',\tilde{m}'}\rangle_{\gamma,s} = 0 \quad \text{if } M + \tilde{M} \neq M' + \tilde{M}',$$

$$\langle D^{(\gamma,s)}_{m,\tilde{m}}, D^{(\gamma,s)}_{m',\tilde{m}'}\rangle_{\gamma,s} = 0 \quad \text{if } M + \tilde{M} \neq M' + \tilde{M}'.$$

In general, the biorthogonal polynomials listed in Theorem 3 are not eigenfunctions of H. However, in the case $s_1 = \cdots = s_{n_2} = \frac{1}{2}$ it follows from Lemma 3 that each of the polynomials satisfies the eigenvalue equation

$$H\Phi = -(M + \tilde{M})(M + \tilde{M} + G - 1)\Phi,$$

where $G = \sum_{k=1}^{n_1} \gamma_k + (n_2 + 1)/2 + s$.

Similarly, the above procedure when applied to any one of the orthogonal bases discussed in §2 leads to an orthogonal polynomial basis with respect to the inner product $\langle \cdot, \cdot \rangle_{\gamma,s}$. Restriction to the case $s_1 = \cdots = s_{n_2} = \frac{1}{2}$ yields eigenfunctions of H and coincides with the results of §4.

REFERENCES

[1] P. APPELL AND J. KAMPÉ DE FÉRIET, *Fonctions Hypergéométriques et Hypersphériques - Polynomes D'Hermite*, Gauthier-Villars et Cie, Paris, 1926.

[2] R. ASKEY, ED., *Theory and Application of Special Functions*, Academic Press, New York, 1975.

[3] C. P. BOYER, E. G. KALNINS, AND W. MILLER, JR., *Stäckel-equivalent integrable Hamiltonian systems*, SIAM J. Math. Anal., 17 (1986), pp. 778–797.

[4] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[5] C. F. DUNKL, *Reflection groups and orthogonal polynomials on the sphere*, Math. Z., 197 (1988), pp. 33–60.

[6] ———, *Harmonic polynomials and peak sets of refection groups*, 1987.

[7] L. P. EISENHART, *Riemannian Geometry*, Princeton University Press, Second edition, 1949.

[8] ———, *Continuous Groups of Transformations*, Dover Reprint, Dover, Delaware, 1961.

[9] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions, Vol.* II, McGraw-Hill, New York, 1953.

[10] E. D. FACKERELL AND R. A. LITTLER, *Polynomials biorthogonal to Appell's polynomials*, Bull. Austral. Math. Soc., 11 (1974), pp. 181–195.

[11] E. G. KALNINS, *Separation of Variables for Riemannian Spaces of Constant Curvature*, Pitman, Monographs and Surveys in Pure and Applied Mathematics 28, Longman, Essex, England, 1986.

[12] E. G. KALNIS, MANOCHA, AND W. MILLER, JR., Stud. Appl. Math., 62 (1980), pp. 143–173.

[13] E. G. KALNINS AND W. MILLER, JR., *Separation of variables on n-dimensional Riemannian manifolds* 1. *The n-sphere S_n and Euclidean n-space R_n*, J. Math. Phys, 27 (1986), pp. 1721–1736.

[14] S. KARLIN AND J. MCGREGOR, *On some stochastic models in genetics*, Stochastic Models in Medicine and Biology, J. Gurland, ed., University of Wisconsin Press, Madison, WI, 1964, pp. 245–271.

[15] T. H. KOORNWINDER, *The addition formula for Jacobi polynomials*, I, *Summary of results*, Indag. Math., 34 (1972), pp. 188–191.

[16] ———, *Jacobi polynomials*, III. *An analytic proof of the addition formula*, SIAM J. Math. Anal., 6 (1975), pp. 533–543.

[17] ———, *Two-variable analogues of the classical orthogonal polynomials*, Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 435–496.

[18] L. KOSCHMIEDER, *Orthogonal polynomials on certain simple domains in the plane and in space*, Tecnica Rev. Fac. Ci. Ex. Tec. Univ. Nac. Tucuman, 1 (1951), pp. 173–181. (In Spanish.)

[19] ———, *A generator of orthogonal polynomials in the circle and in the triangle*, Rev. Mat. Hisp.-Amer., (4) 17 (1957), pp. 291–298. (In Spanish.)

[20] H. L. KRALL AND I. M. SHEFFER, *Orthogonal polynomials in two variables*, Ann. Mat. Pura Appl., (4) 76 (1967), pp. 325–376.

[21] C. S. LAM AND M. V. TRATNIK, *Conformally invariant operator-product expansions of any number of operators of arbitrary spin*, Canad. J. Phys., 63 (1985), pp. 1427–1437.

[22] W. MILLER, JR., *Symmetry and Separation of Variables*, Addison-Wesley, Reading, MA, 1977.

[23] J. PRORIOL, *Sur une famille de polynomes à deux variables orthogonaux dans un triangle*, C. R. Acad. Sci. Paris, 245 (1957), pp. 2459–2461.

[24] L.J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.

[25] D. STANTON, *Orthogonal polynomials and Chevalley groups*, Special Functions: Group Theoretical Aspects and Applications, R.A. Askey, T.H. Koornwinder, and W. Schempp, eds., D. Reidel, Boston, 1984, pp. 87–128.

[26] N. VILENKIN, *Special Functions and the Theory of Group Representations*, Amer. Math. Soc. Transl., American Mathematical Society, Providence, RI, 1968.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 295-301, January 1991

018

THE ADDITION FORMULA FOR LITTLE q-LEGENDRE POLYNOMIALS AND THE $SU(2)$ QUANTUM GROUP*

TOM H. KOORNWINDER†

Abstract. From the interpretation of little q-Jacobi polynomials as matrix elements of the irreducible unitary representations of the SU(2) quantum group an addition formula is derived for the little q-Legendre polynomials. It involves an expansion in terms of Wall polynomials. A product formula for little q-Legendre polynomials follows by q-integration.

Key words. quantum groups, SU(2), little q-Jacobi polynomials, little q-Legendre polynomials, Wall polynomials, addition formula, product formula

AMS(MOS) subject classifications. 33A65, 33A75, 22E70

1. Introduction. Quantum groups were recently introduced by Drinfeld [5] and Woronowicz [17]. Interesting examples are provided by deformation into a noncommutative Hopf algebra of some suitable commutative Hopf algebra of functions on a specific group. The most elementary nontrivial example comes from deformation of the algebra of polynomials on $SU(2)$ (cf. Woronowicz [18]). We denote the resulting quantum group by $SU_\mu(2)$.

It has been proved by Vaksman and Soibelman [15], Masuda et al. [11], [12], and Koornwinder [10] that the matrix elements of the irreducible unitary representations of the quantum group $SU_\mu(2)$ can be expressed in terms of the little q-Jacobi polynomials. In the present paper we use this interpretation to derive an addition formula for the little q-Legendre polynomials, i.e., for the q-analogues of the Legendre polynomials within the class of little q-Jacobi polynomials. The derivation of this formula is straightforward and analogous to a proof of the addition formula for Legendre polynomials using irreducible representations of $SU(2)$ (cf. Vilenkin [16, Chap. 3]). However, the resulting formula involves noncommuting variables. It is less easy to rewrite it equivalently as a formula involving only commuting variables. We can do this by using an infinite-dimensional irreducible $*$-representation of the Hopf algebra considered as a $*$-algebra. The result (Theorem 4.1) gives an expansion in terms of Wall polynomials (little q-Jacobi polynomials with the second parameter equal to zero).

Our addition formula somewhat resembles an addition formula for (continuous) q-ultraspherical polynomials derived by Rahman and Verma [14], but for that formula a (quantum) group-theoretic interpretation is not yet known. It would have been hard to find our formula without guidance from the quantum group. Indeed, that it is possible to obtain such a formula demonstrates the power and depth of the quantum group-theoretic interpretation of special functions. It turns out to be highly nontrivial to prove this formula analytically or to show that its limit case for $q \uparrow 1$ is the addition formula for Legendre polynomials. The analytic proof has been done by Rahman [13] and the limit result is proved by Van Assche and Koornwinder [2].

The contents of this paper are as follows. In §§ 2 and 3 the preliminaries about q-hypergeometric orthogonal polynomials and quantum groups, respectively, are presented. The derivation of the addition formula is given in § 4. Finally, a product formula for little q-Legendre polynomials is derived from the addition formula in § 5.

* Received by the editors March 13, 1989; accepted for publication (in revised form) October 10, 1989.
† CWI: Centre for Mathematics and Computer Science, Postbus 4079, 1009 AB Amsterdam, the Netherlands.

2. Some q-hypergeometric orthogonal polynomials. Let $1 \neq q \in \mathbb{C}$. We use the familiar definitions and notation for q-shifted factorials and q-hypergeometric functions (cf. [6, Chap. 1], [10, § 2]). The *little q-Jacobi polynomials*

$$p_n(x; a, b \mid q) := {}_2\phi_1\left(\begin{matrix} q^{-n}, abq^{n+1} \\ aq \end{matrix}; q, qx\right) \tag{2.1}$$

occur as part of a classification by Hahn [7] of orthogonal polynomials satisfying q-difference equations. Their detailed properties as orthogonal polynomials are given by Andrews and Askey [1]. For $a = b = 1$ we will call these polynomials *little q-Legendre polynomials.*

The special little q-Jacobi polynomials obtained by putting $b := 0$ in (2.1) are known as *Wall polynomials*; cf. Chihara [3, § 5, Case I], [4, Chap. 6, § 11]:

$$p_n(x; a, 0 \mid q) = {}_2\phi_1(q^{-n}, 0; aq; q, qx) \tag{2.2}$$

(Chihara uses another notation). We will put $a := q^\alpha$. These polynomials can be viewed as one of the many q-analogues of the Laguerre polynomials in view of the limit formula

$$\lim_{q \uparrow 1} p_n((1-q)x; q^\alpha, 0 \mid q) = L_n^\alpha(x)/L_n^\alpha(0).$$

By specialization of the orthogonality relations for the little q-Jacobi polynomials we obtain the orthogonality relations for the Wall polynomials:

$$\frac{(q^{\alpha+1}; q)_\infty}{(1-q)(q; q)_\infty} \int_0^1 p_n(t; q^\alpha, 0 \mid q) p_m(t; q^\alpha, 0 \mid q) t^\alpha (qt; q)_\infty \, d_q t = \delta_{n,m} \frac{q^{n(\alpha+1)}(q; q)_n}{(q^{\alpha+1}; q)_n}, \tag{2.3}$$

where the q-integral is defined by

$$\int_0^1 f(t) \, d_q t := \sum_{k=0}^{\infty} f(q^k)(q^k - q^{k+1}),$$

and where we suppose that $0 < q < 1$ and $\alpha > -1$. From [3, Forms. (5.1), (5.2)] together with (2.2) we obtain the three-term recurrence relation

$$\begin{aligned} x p_n(x; a, 0 \mid q) = {} & -q^n(1 - aq^{n+1}) p_{n+1}(x; a, 0 \mid q) \\ & + q^n(1 + a - aq^n - aq^{n+1}) p_n(x; a, 0 \mid q) \\ & - q^n(a - aq^n) p_{n-1}(x; a, 0 \mid q), \\ p_{-1}(x; a, 0 \mid q) = 0, \qquad & p_0(x; a, 0 \mid q) = 1. \end{aligned} \tag{2.4}$$

Put

$$\begin{aligned} P_n(q^k; q^\alpha \mid q) := {} & \left(\frac{(q^{\alpha+1}; q)_\infty (q^{k+1}; q)_\infty (q^{\alpha+1}; q)_n}{(q; q)_\infty (q; q)_n} \right)^{1/2} \\ & \cdot (-1)^n q^{(k-n)(\alpha+1)/2} p_n(q^k; q^\alpha, 0 \mid q). \end{aligned} \tag{2.5}$$

Then (2.3) can be rewritten as

$$\sum_{k=0}^{\infty} P_n(q^k; q^\alpha \mid q) P_m(q^k; q^\alpha \mid q) = \delta_{n,m}, \qquad n, m = 0, 1, 2, \cdots. \tag{2.6}$$

Since the orthogonality measure in (2.3) has compact support, the orthonormal system

$$\{P_n(q^k; q^\alpha \mid q)\}_{k=0,1,2,\cdots}, \qquad n = 0, 1, 2, \cdots,$$

is complete in the Hilbert space l^2. Hence, we have also the dual orthogonality relations

$$\sum_{n=0}^{\infty} P_n(q^k; q^\alpha \mid q) P_n(q^l; q^\alpha \mid q) = \delta_{k,l}, \qquad k, l = 0, 1, 2, \cdots. \tag{2.7}$$

We conclude this section with an expression of Wall polynomials in terms of a ${}_3\phi_2$:

$$p_n(x; q^\alpha, 0 \mid q) = \frac{(-1)^n q^{n(n+2\alpha+1)/2} x^n}{(q^{\alpha+1}; q)_n} {}_3\phi_2\left(\begin{matrix} q^{-n}, q^{-n-\alpha}, x^{-1} \\ 0, 0 \end{matrix}; q, q\right). \tag{2.8}$$

This follows by putting $b := 0$ in

$$\begin{aligned} {}_2\phi_1\left(\begin{matrix} q^{-n}, abq^{n+1} \\ aq \end{matrix}; q, qx\right) &= (q^{-n+1}x; q)_n \, {}_2\phi_2\left(\begin{matrix} q^{-n}, q^{-n}b^{-1} \\ qa, q^{-n+1}x \end{matrix}; q, q^{n+2}abx\right) \\ &= \frac{(qb; q)_n q^{n(n-1)/2}(-aqx)^n}{(qa; q)_n} {}_3\phi_2\left(\begin{matrix} q^{-n}, q^{-n}\alpha^{-1}, x^{-1} \\ qb, 0 \end{matrix}; q, q\right). \end{aligned}$$

Here we have used a transformation formula for ${}_2\phi_1$ (cf. [6, Chap. 1]) in the first equality and reversion of summation order in the second equality (see also [8]).

3. The quantum group $SU_\mu(2)$. In the rest of this paper we fix $0 \neq \mu \in (-1, 1)$. Let $\mathscr{A}$ be the unital $*$-algebra generated by the two elements α and γ satisfying the relations

$$\alpha\gamma = \mu\gamma\alpha, \quad \alpha\gamma^* = \mu\gamma^*\alpha, \quad \gamma\gamma^* = \gamma^*\gamma,$$
$$\alpha^*\alpha + \gamma\gamma^* = I, \qquad \alpha\alpha^* + \mu^2\gamma\gamma^* = I.$$

Let $\Phi: \mathscr{A} \to \mathscr{A} \otimes \mathscr{A}$ be the unital $*$-homomorphism such that

$$\Phi(\alpha) = \alpha \otimes \alpha - \mu\gamma^* \otimes \gamma, \qquad \Phi(\gamma) = \gamma \otimes \alpha + \alpha^* \otimes \gamma.$$

Then Φ acts as a comultiplication and $\mathscr{A}$ thus becomes a Hopf algebra with involution which we say to be *associated with the compact matrix quantum group* $SU_\mu(2)$. In the limit for $\mu \uparrow 1$, $\mathscr{A}$ becomes the algebra of polynomials in the matrix elements of the natural representation of $SU(2)$, and the comultiplication is then induced by the group structure of $SU(2)$.

It is possible to embed the $*$-algebra $\mathscr{A}$ as a dense $*$-subalgebra of a C^*-algebra by a universal construction. The C^*-algebra approach is emphasized in particular by Woronowicz [17], [18]. In this paper we could work with the C^*-algebra, but only elements of the dense $*$-subalgebra $\mathscr{A}$ will be needed, so we will use this latter algebra. Other references for $SU_\mu(2)$ besides [17] and [18], are [5], [15], [11], and [12].

The irreducible unitary co-representations of $\mathscr{A}$ (which are called irreducible unitary representations of $SU_\mu(2)$ in [18] and [10]) have been completely classified [18], [9], [15], [11], [12], [10]. Up to equivalence, there is one such co-representation for each finite dimension. We will denote the co-representation of dimension $2l+1$ by $t^{l,\mu}$ ($l = 0, \frac{1}{2}, 1, \cdots$), and its matrix elements with respect to a suitable orthonormal basis corresponding to the quantum subgroup $U(1)$ by $t^{l,\mu}_{n,m}$ ($n, m = -l, -l+1, \cdots, l$). Then the co-representation property of $t^{l,\mu}$ is expressed by

$$\Phi(t^{l,\mu}_{n,m}) = \sum_{k=-l}^{l} t^{l,\mu}_{n,k} \otimes t^{l,\mu}_{k,m}. \tag{3.1}$$

The $t^{l,\mu}_{n,m}$ have been computed explicitly in terms of little q-Jacobi polynomials (cf. [15], [11], [12], [10]). Here we will only need the cases that $l = 0, 1, 2, \cdots$ and m or $n = 0$. Put

$$p^\mu_{l,k}(x) := \begin{bmatrix} l \\ k \end{bmatrix}^{1/2}_{\mu^2} \begin{bmatrix} l+k \\ k \end{bmatrix}^{1/2}_{\mu^2} \mu^{-k(l-k)} p_{l-k}(x; \mu^{2k}, \mu^{2k} \mid \mu^2), \tag{3.2}$$

$$p^\mu_l(x) := p^\mu_{l,0}(x) = p_l(x; 1, 1 \mid \mu^2). \tag{3.3}$$

Here the notation (2.1) for the little q-Jacobi polynomials is used and (3.3) gives a little q-Legendre polynomial. Then (cf. [10, Thm. 5.3]), for $k=0,1,\cdots,l$,

$$
\begin{aligned}
t^{l,\mu}_{k,0} &= (\alpha^*)^k p^{\mu}_{l,k}(\gamma\gamma^*)\gamma^k,\\
t^{l,\mu}_{0,k} &= (\alpha^*)^k p^{\mu}_{l,k}(\gamma\gamma^*)(-\mu\gamma^*)^k,\\
t^{l,\mu}_{-k,0} &= (-\mu\gamma^*)^k p^{\mu}_{l,k}(\gamma\gamma^*)\alpha^k,\\
t^{l,\mu}_{0,-k} &= \gamma^k p^{\mu}_{l,k}(\gamma\gamma^*)\alpha^k.
\end{aligned}
\tag{3.4}
$$

Hence

$$
t^{l,\mu}_{0,0} = p^{\mu}_l(\gamma\gamma^*). \tag{3.5}
$$

The irreducible $*$-representations of the $*$-algebra $\mathscr{A}$ on a Hilbert space are classified in [15, Thm. 3.2]. There is a family of one-dimensional and a family of infinite-dimensional representations, both parametrized by the unit circle. We pick one of these infinite-dimensional representations: Let $\mathscr{H}$ be a Hilbert space with orthonormal basis $e_0, e_1, \cdots$. Put $e_{-1}, e_{-2}, \cdots := 0$. We define a $*$-representation of τ of $\mathscr{A}$ on $\mathscr{H}$ by specifying the action of the generators of $\mathscr{A}$:

$$
\begin{aligned}
\tau(\alpha)e_n &:= (1-\mu^{2n})^{1/2}e_{n-1},\\
\tau(\alpha^*)e_n &:= (1-\mu^{2n+2})^{1/2}e_{n+1},\\
\tau(\gamma)e_n &:= \mu^n e_n,\\
\tau(\gamma^*)e_n &:= \mu^n e_n.
\end{aligned}
\tag{3.6}
$$

4. Proof of the addition formula. Let $l=0,1,2,\cdots$. A special case of (3.1) is

$$
\Phi(t^{l,\mu}_{0,0}) = \sum_{k=-l}^{l} t^{l,\mu}_{0,k}\otimes t^{l,\mu}_{k,0}.
$$

Hence, by (3.4) and (3.5),

$$
\begin{aligned}
p^{\mu}_l(\Phi(\gamma\gamma^*)) &= p^{\mu}_l(\gamma\gamma^*)\otimes p^{\mu}_l(\gamma\gamma^*)\\
&\quad + \sum_{k=1}^{l} (\alpha^*)^k p^{\mu}_{l,k}(\gamma\gamma^*)(-\mu\gamma^*)^k \otimes (\alpha^*)^k p^{\mu}_{l,k}(\gamma\gamma^*)\gamma^k\\
&\quad + \sum_{k=1}^{l} \gamma^k p^{\mu}_{l,k}(\gamma\gamma^*)\alpha^k \otimes (-\mu\gamma^*)^k p^{\mu}_{l,k}(\gamma\gamma^*)\alpha^k.
\end{aligned}
\tag{4.1}
$$

This formula might already be called an addition formula for little q-Legendre polynomials p^{μ}_l. It involves noncommuting variables. In the limit, for $\mu\uparrow 1$, the variables commute and (4.1) becomes the classical addition formula for Legendre polynomials. In three steps we will rewrite (4.1) into a formula involving commuting variables: First, we represent (4.1) as an operator identity on the Hilbert space $\mathscr{H}\otimes\mathscr{H}$ by using the representation $\tau\otimes\tau$ (cf. (3.6)). Second, we let these operators act on the standard basis of $\mathscr{H}\otimes\mathscr{H}$. Thus we obtain a family of vector identities in $\mathscr{H}\otimes\mathscr{H}$. Third, we take inner products with respect to another suitable orthonormal basis of $\mathscr{H}\otimes\mathscr{H}$. This will yield a family of scalar identities.

Apply $\tau\otimes\tau$ to both sides of (4.1) and let both sides of the resulting operator equality act on $e_{x+y}\otimes e_y$. Then

$$
\begin{aligned}
p_l^{\mu}((\tau\otimes\tau)\Phi(\gamma\gamma^*))\, e_{x+y}\otimes e_y &= p_l^{\mu}(\mu^{2x+2y})p_l^{\mu}(\mu^{2y})\, e_{x+y}\otimes e_y \\
&\quad + \sum_{k=1}^{l} (-1)^k \mu^{k(x+2y+1)}(\mu^{2(x+y+1)};\mu^2)_k^{1/2}(\mu^{2(y+1)};\mu^2)_k^{1/2} \\
&\quad \cdot p_{l,k}^{\mu}(\mu^{2x+2y})p_{l,k}^{\mu}(\mu^{2y})\, e_{x+y+k}\otimes e_{y+k} \\
&\quad + \sum_{k=1}^{l} (-1)^k \mu^{k(x+2y-2k+1)}(\mu^{2(x+y)};\mu^{-2})_k^{1/2}(\mu^{2y};\mu^{-2})_k^{1/2} \\
&\quad \cdot p_{l,k}^{\mu}(\mu^{2x+2y-2k})p_{l,k}^{\mu}(\mu^{2y-2k})\, e_{x+y-k}\otimes e_{y-k}.
\end{aligned} \tag{4.2}
$$

(Remember the convention that $e_n = 0$ for $n < 0$.)

In order to say more about the left-hand side of (4.2) we consider the action of

$$\Phi(\gamma\gamma^*) = (\gamma\otimes\alpha + \alpha^*\otimes\gamma)(\gamma^*\otimes\alpha^* + \alpha\otimes\gamma^*)$$

on $e_{x+y}\otimes e_y$. We obtain

$$
\begin{aligned}
&(\tau\otimes\tau)(\Phi(\gamma\gamma^*))\, e_{x+y}\otimes e_y \\
&\quad = \mu^{x+2y+1}(1-\mu^{2x+2y+2})^{1/2}(1-\mu^{2y+2})^{1/2}\, e_{x+y+1}\otimes e_{y+1} \\
&\qquad + (\mu^{2x+2y} + \mu^{2y} - \mu^{2x+4y} - \mu^{2x+4y+2})\, e_{x+y}\otimes e_y \\
&\qquad + \mu^{x+2y-1}(1-\mu^{2x+2y})^{1/2}(1-\mu^{2y})^{1/2}\, e_{x+y-1}\otimes e_{y-1}.
\end{aligned}
$$

Hence, if

$$f := \sum_{y=0}^{\infty} c_y\, e_{x+y}\otimes e_y$$

belongs to $\mathcal{H}\otimes\mathcal{H}$, then

$$
\begin{aligned}
(\tau\otimes\tau)(\Phi(\gamma\gamma^*))f = \sum_{y=0}^{\infty} [\, & c_{y-1}\mu^{x+2y-1}(1-\mu^{2x+2y})^{1/2}(1-\mu^{2y})^{1/2} \\
& + c_y(\mu^{2x+2y} + \mu^{2y} - \mu^{2x+4y} - \mu^{2x+4y+2}) \\
& + c_{y+1}\mu^{x+2y+1}(1-\mu^{2x+2y+2})^{1/2}(1-\mu^{2y+2})^{1/2}]\, e_{x+y}\otimes e_y.
\end{aligned} \tag{4.3}
$$

Now choose

$$c_y := P_y(\mu^{2z};\mu^{2x}|\mu^2),$$

where P_y is defined in terms of Wall polynomials by (2.5). Then $f\in\mathcal{H}\otimes\mathcal{H}$ and, by (2.4), the expression in square brackets on the right-hand side of (4.3) is equal to $\mu^{2z}c_y$. Define

$$f_z^x := \sum_{y=0}^{\infty} P_y(\mu^{2z};\mu^{2x}|\mu^2)\, e_{x+y}\otimes e_y. \tag{4.4}$$

Then, by the orthogonality relations (2.7), the vectors $\{f_z^x\}_{z=0,1,2,\ldots}$ form an orthonormal basis of

$$\bigoplus_{y=0}^{\infty} \mathbb{C}\, e_{x+y}\otimes e_y. \tag{4.5}$$

We also have

$$(\tau\otimes\tau)(\Phi(\gamma\gamma^*))f_z^x = \mu^{2z}f_z^x. \tag{4.6}$$

Now take the inner product of both sides of (4.2) with respect to f_z^x and apply (4.4), (4.6), and the self-adjointness of $\Phi(\gamma\gamma^*)$ acting on $\mathscr{H}\otimes\mathscr{H}$. Assume also the convention that $P_n := 0$ and $p_n := 0$ for $n<0$. Then we obtain

$$
\begin{aligned}
p_l^{\mu}(\mu^{2z})P_y(\mu^{2z};\mu^{2x}|\mu^2) &= p_l^{\mu}(\mu^{2x+2y})p_l^{\mu}(\mu^{2y})P_y(\mu^{2z};\mu^{2x}|\mu^2)\\
&\quad+\sum_{k=1}^{l}(-1)^k\mu^{k(x+2y+1)}(\mu^{2(x+y+1)};\mu^2)_k^{1/2}(\mu^{2(y+1)};\mu^2)_k^{1/2}\\
&\quad\cdot p_{l,k}^{\mu}(\mu^{2x+2y})p_{l,k}^{\mu}(\mu^{2y})P_{y+k}(\mu^{2z};\mu^{2x}|\mu^2)\\
&\quad+\sum_{k=1}^{l}(-1)^k\mu^{k(x+2y-2k+1)}(\mu^{2(x+y)};\mu^{-2})_k^{1/2}(\mu^{2y};\mu^{-2})_k^{1/2}\\
&\quad\cdot p_{l,k}^{\mu}(\mu^{2x+2y-2k})p_{l,k}^{\mu}(\mu^{2y-2k})P_{y-k}(\mu^{2z};\mu^{2x}|\mu^2).
\end{aligned}
\tag{4.7}
$$

Finally substitute (3.2), (3.3), and (2.5) in (4.7) and replace μ^2 by q. Then we obtain the following theorem.

THEOREM 4.1 (addition formula for little q-Legendre polynomials). *For $x, y, z = 0, 1, 2, \cdots$ we have*

$$
\begin{aligned}
&p_l(q^z;1,1|q)p_y(q^z;q^x,0|q)\\
&\quad=p_l(q^{x+y};1,1|q)p_l(q^y;1,1|q)p_y(q^z;q^x,0|q)\\
&\qquad+\sum_{k=1}^{l}\frac{(q;q)_{x+y+k}(q;q)_{l+k}q^{k(y-l+k)}}{(q;q)_{x+y}(q;q)_{l-k}(q;q)_k^2}\\
&\qquad\cdot p_{l-k}(q^{x+y};q^k,q^k|q)p_{l-k}(q^y;q^k,q^k|q)p_{y+k}(q^z;q^x,0|q)\\
&\qquad+\sum_{k=1}^{l}\frac{(q;q)_y(q;q)_{l+k}q^{k(x+y-l+1)}}{(q;q)_{y-k}(q;q)_{l-k}(q;q)_k^2}\\
&\qquad\cdot p_{l-k}(q^{x+y-k};q^k,q^k|q)p_{l-k}(q^{y-k};q^k,q^k|q)p_{y-k}(q^z;q^x,0|q).
\end{aligned}
\tag{4.8}
$$

Remark 4.1. We could have derived the final result (4.8) from (4.1) as well by using one of the other members of the series of infinite-dimensional irreducible $*$-representations of $\mathscr{A}$ as given by [15, Thm. 3.2]. The same result would also have been obtained by use of the faithful $*$-representation of $\mathscr{A}$ given in [18, Thm. 1.2]. Furthermore, it can be shown that formula (4.8) taken for all $x, y, z = 0, 1, 2, \cdots$ and with $q=\mu^2$ is equivalent to (4.1).

Remark 4.2. It is possible to give a more conceptual interpretation of the occurrence of Wall polynomials in the addition formula. Namely, it can be shown that Wall polynomials have an interpretation as Clebsch–Gordan coefficients for the decomposition of the $*$-representation $(\tau\otimes\tau)\circ\Phi$ of $\mathscr{A}$ as a direct integral of irreducible $*$-representations of $\mathscr{A}$.

5. The product formula for little q-Legendre polynomials. If we multiply both sides of (4.7) with $P_y(\mu^{2z};\mu^{2x}|\mu^2)$ and sum over z, then by (2.6) we obtain the product formula

$$
p_l^{\mu}(\mu^{2x+2y})p_l^{\mu}(\mu^{2y})=\sum_{z=0}^{\infty}p_l^{\mu}(\mu^{2x})(P_y(\mu^{2z};\mu^{2x}|\mu^2))^2.
$$

After substituting (3.3), (2.5), and (2.8) and after replacing x by $x-y$ and μ^2 by q, we get the desired product formula.

THEOREM 5.1. *For $x, y = 0, 1, 2, \cdots$ we have*

$$p_l(q^x; 1, 1 \mid q) p_l(q^y; 1, 1 \mid q) = (1-q) \sum_{z=0}^{\infty} p_l(q^z; 1, 1 \mid q) K(q^x, q^y, q^z \mid q) q^z$$

with

$$K(q^x, q^y, q^z \mid q) := \frac{(q^{x+1}; q)_\infty (q^{y+1}; q)_\infty (q^{z+1}; q)_\infty q^{xy+xz+yz}}{(q; q)_\infty^2 (1-q)} \cdot \left\{ {}_3\phi_2 \left(\begin{matrix} q^{-x}, q^{-y}, q^{-z} \\ 0, 0 \end{matrix} ; q, q \right) \right\}^2.$$

REFERENCES

[1] G. E. ANDREWS AND R. ASKEY, *Enumeration of partitions: The role of Eulerian series and q-orthogonal polynomials*, in Higher Combinatorics, M. Aigner, ed., D. Reidel, Dordrecht, the Netherlands, 1977, pp. 3-26.

[2] W. VAN ASSCHE AND T. H. KOORNWINDER, *Asymptotic behaviour for Wall polynomials and the addition formula for little q-Legendre polynomials*, SIAM J. Math. Anal., this issue (1991), pp. 301-302.

[3] T. S. Chihara, *Orthogonal polynomials with Brenke type generating functions*, Duke Math. J., 35 (1968), pp. 505-518.

[4] ———, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[5] V. G. DRINFELD, *Quantum groups*, in Proc. International Congress of Mathematicians, Berkeley, 1986, American Mathematical Society, Providence, RI, 1987, pp. 798-820.

[6] G. GASPER AND M. RAHMAN, *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, U.K., 1990.

[7] W. HAHN, *Über Orthogonalpolynome, die q-Differenzengleichungen genügen*, Math. Nachr., 2 (1949), pp. 4-34, 379.

[8] M. E. H. ISMAIL AND J. A. WILSON, *Asymptotic and generating relations for the q-Jacobi and ${}_4\phi_3$ polynomials*, J. Approx. Theory, 36 (1982), pp. 43-54.

[9] M. JIMBO, *A q-analogue of $U(g)$ and the Yang-Baxter equation*, Lett. Math. Phys., 10 (1985), pp. 63-69.

[10] T. H. KOORNWINDER, *Representations of the twisted $SU(2)$ quantum group and some q-hypergeometric orthogonal polynomials*, Nederl. Akad. Wetensch. Indag. Math., 51 (1989), pp. 97-117.

[11] T. MASUDA, K. MIMACHI, Y. NAKAGAMI, M. NOUMI, AND K. UENO, *Representations of quantum groups and a q-analogue of orthogonal polynomials*, C. R. Acad. Sci. Paris Sér. I Math., 307 (1988), pp. 559-564.

[12] ———, *Representations of the quantum group $SU_q(2)$ and the little q-Jacobi polynomials*, J. Funct. Anal., to appear.

[13] M. RAHMAN, *A simple proof of Koornwinder's addition formula for the little q-Legendre polynomials*, Proc. Amer. Math. Soc., 107 (1989), pp. 373-381.

[14] M. RAHMAN AND A. VERMA, *Product and addition formula for the continuous q-ultraspherical polynomials*, SIAM J. Math. Anal., 17 (1986), pp. 1461-1474.

[15] L. L. VAKSMAN AND YA. S. SOIBELMAN, *Algebra of functions on the quantum group $SU(2)$*, Funct. Anal. Appl., 22 (1988), pp. 170-181.

[16] N. YA. VILENKIN, *Special functions and the theory of group representations*, Amer. Math. Soc. Transl. 22, American Mathematical Society, Providence, RI, 1968.

[17] S. L. WORONOWICZ, *Compact matrix pseudogroups*, Comm. Math. Phys., 111 (1987), pp. 613-665.

[18] ———, *Twisted $SU(2)$ group. An example of a non-commutative differential calculus*, Publ. Res. Inst. Math. Sci., 23 (1987), pp. 117-181.

SIAM J. MATH. ANAL.
Vol. 22, No. 1, pp. 302-311, January 1991

019

ASYMPTOTIC BEHAVIOUR FOR WALL POLYNOMIALS AND THE ADDITION FORMULA FOR LITTLE q-LEGENDRE POLYNOMIALS*

WALTER VAN ASSCHE† AND TOM H. KOORNWINDER‡

Abstract. Wall polynomials $W_n(x; b, q)$ are considered and their asymptotic behaviour is described when $q = c^{1/n}$ and n tends to infinity. The results are then used to derive the addition and product formulas for the Legendre polynomials from the recently obtained addition and product formulas for little q-Legendre polynomials.

Key words. Wall polynomials, addition formula, product formula, basic hypergeometric polynomials, Legendre polynomials

AMS(MOS) subject classifications. 33A65, 42C05

1. Introduction.

The Wall polynomials $W_n(x; b, q)$ are defined by the recurrence formula

$$W_{n+1}(x; b, q) = \{x - [b + q - (1+q)bq^n]q^n\} W_n(x; b, q) - b(1-q^n)(1-bq^{n-1})q^{2n} W_{n-1}(x; b, q), \qquad n = 0, 1, 2, \cdots \tag{1.1}$$

with initial values $W_{-1} = 0$ and $W_0 = 1$. Clearly $W_n(x; b, q)$ is a monic polynomial of degree n in the variable x. Some properties of Wall polynomials are given in Chihara's book [4, p. 198]. These polynomials are closely related to the continued fraction

$$1 + \frac{x}{1+} \frac{(1-b)qx}{1} + \frac{(1-q)bqx}{1} + \frac{(1-bq)q^2x}{1} + \cdots,$$

which was studied by H. S. Wall [16]. The Wall polynomials were also studied by Chihara [5] because they have a Brenke-type generating function, i.e.,

$$\sum_{n=0}^{\infty} W_n(x; b, q) \frac{z^n}{(b; q)_n (q; q)_n} = A(z)B(zx),$$

where

$$A(z) = \sum_{n=0}^{\infty} (-1)^n q^{n(n+1)/2} \frac{z^n}{(q; q)_n} = (zq; q)_\infty,$$

$$B(z) = \sum_{n=0}^{\infty} \frac{z^n}{(b; q)_n (q; q)_n}.$$

We have used the notation

$$(b; q)_n = (1-b)(1-bq) \cdots (1-bq^{n-1}),$$

$$(b; q)_\infty = \lim_{n \to \infty} (b; q)_n;$$

* Received by the editors March 13, 1989; accepted for publication (in revised form) October 10, 1989.
† Catholic University of Leuven, Department of Mathematics, Celestijnenlaan 200B, B-3030 Leuven, Belgium. This author is a Research Associate of the Belgium National Fund for Scientific Research.

‡ Centre for Mathematics and Computer Science, P.O. Box 4079, NL-1009 AB Amsterdam, the Netherlands.

the latter limit exists whenever $|q|<1$. From this generating function we easily find

$$W_n(x;b,q)=(-1)^n(b;q)_n q^{n(n+1)/2}\sum_{k=0}^{n}\frac{(q;q)_n}{(q;q)_{n-k}(q;q)_k}q^{k(k-1)/2}\frac{(-q^{-n}x)^k}{(b;q)_k}$$
$$=(-1)^n(b;q)_n q^{n(n+1)/2}{}_2\phi_1(q^{-n},0;b;q,x), \tag{1.2}$$

where the q-hypergeometric (or basic hypergeometric [6]) function is defined by

$${}_{r+1}\phi_r(a_1,\cdots,a_{r+1};b_1,\cdots,b_r;q,z)=\sum_{k=0}^{\infty}\frac{(a_1;q)_k\cdots(a_{r+1};q)_k}{(b_1;q)_k\cdots(b_r;q)_k}\frac{z^k}{(q;q)_k}.$$

If $0<q<1$ and $0<b<1$ then the Wall polynomials are orthogonal with respect to a positive measure supported on the geometric sequence $\{q^n: n=1,2,3,\cdots\}$ and we have

$$\sum_{k=0}^{\infty}W_n(q^{k+1};b,q)W_m(q^{k+1};b,q)\frac{b^k}{(q;q)_k}=0,\qquad n\neq m.$$

The orthonormal polynomials are given by

$$w_n(x;b,q)=\frac{q^{-n(n+1)/2}}{\sqrt{b^n(q;q)_n(b;q)_n}}W_n(x;b,q), \tag{1.3}$$

and they satisfy

$$(b;q)_\infty\sum_{k=0}^{\infty}w_n(q^{k+1};b,q)w_m(q^{k+1};b,q)\frac{b^k}{(q;q)_k}=\delta_{n,m},\qquad n,m\geqq 0 \tag{1.4}$$

and the three-term recurrence relation (1.1) becomes

$$xw_n(x;b,q)=a_{n+1}w_{n+1}(x;b,q)+b_nw_n(x;b,q)+a_nw_{n-1}(x;b,q) \tag{1.5}$$

with $w_{-1}=0$, $w_0=1$, and

$$a_n=a_n(b,q)=q^n\sqrt{b(1-q^n)(1-bq^{n-1})},\qquad n=1,2,3,\cdots,$$
$$b_n=b_n(b,q)=q^n[b+q-(1+q)bq^n],\qquad n=0,1,2,\cdots. \tag{1.6}$$

Sometimes it is convenient to use the notation

$$(b;q)_\infty\sum_{k=0}^{\infty}f(q^{k+1})\frac{b^k}{(q;q)_k}=\int_0^1 f(z)\,d\mu(z;b,q),\qquad f\in C[0,1] \tag{1.7}$$

so that $\mu(\cdot;b,q)$ is the orthogonality measure for the Wall polynomials $W_n(x;b,q)$.

Recently Koornwinder [8] obtained the addition formula for little q-Legendre polynomials by using the fact that the matrix elements of the irreducible unitary representations of the quantum group $S_\mu U(2)$ (see, e.g., Woronowicz [17], [18]) can be expressed in terms of little q-Jacobi polynomials (Masuda et al. [9], Vaksman and Soibelman [13], Koornwinder [7]). The little q-Jacobi polynomials are defined in terms of q-hypergeometric functions by

$$p_n(x;a,b\,|\,q)={}_2\phi_1(q^{-n},abq^{n+1};aq;q,qx).$$

If $a=q^\alpha$ and $b=q^\beta$ then these little q-Jacobi polynomials approach the Jacobi polynomials $P_n^{(\alpha,\beta)}(1-2x)/P_n^{(\alpha,\beta)}(1)$ as q tends to 1 [1], [3]. If $a=b=1$ then we have

the little q-Legendre polynomials. Notice that for $b=0$ we essentially have the Wall polynomials:

$$p_n\left(\frac{x}{q};\frac{b}{q},0\,|\,q\right)=(-1)^n\frac{q^{-n(n+1)/2}}{(b;q)_n}W_n(x;b,q)$$
$$=(-1)^n\left\{\frac{b^n(q;q)_n}{(b;q)_n}\right\}^{1/2}w_n(x;b,q). \tag{1.8}$$

The addition formula for little q-Legendre polynomials is

$$\begin{aligned}
&p_m(q^z;1,1\,|\,q)p_y(q^z;q^x,0\,|\,q)\\
&\quad=p_m(q^{x+y};1,1\,|\,q)p_m(q^y;1,1\,|\,q)p_y(q^z;q^x,0\,|\,q)\\
&\qquad+\sum_{k=1}^{m}\frac{(q;q)_{x+y+k}(q;q)_{m+k}q^{k(y-m+k)}}{(q;q)_{x+y}(q;q)_{m-k}(q;q)_k^2}p_{m-k}(q^{x+y};q^k,q^k\,|\,q)\\
&\qquad\cdot p_{m-k}(q^y;q^k,q^k\,|\,q)p_{y+k}(q^z;q^x,0\,|\,q)\\
&\qquad+\sum_{k=1}^{m}\frac{(q;q)_y(q;q)_{m+k}q^{k(x+y-m+1)}}{(q;q)_{y-k}(q;q)_{m-k}(q;q)_k^2}p_{m-k}(q^{x+y-k};q^k,q^k\,|\,q)\\
&\qquad\cdot p_{m-k}(q^{y-k};q^k,q^k\,|\,q)p_{y-k}(q^z;q^x,0\,|\,q)
\end{aligned} \tag{1.9}$$

with $x,y,z=0,1,2,\cdots$. Rahman [11] has given an analytic proof of this addition formula while Rahman and Verma [12] have given similar formulas for the continuous q-ultraspherical polynomials. The right-hand side of the above formula can be considered as an expansion of the left-hand side in terms of Wall polynomials. For $q\uparrow 1$ we should get the familiar addition formula for Legendre polynomials (see, e.g., [2, pp. 29-38]), but this limit involves some interesting asymptotic formulas for the Wall polynomials $W_n(x;b,c^{1/n})$ with $0<c<1$ and n tending to infinity. This was the main reason for investigating such asymptotic formulas for Wall polynomials.

In §2 we establish some weak asymptotics for Wall polynomials. In § 3 we show how the addition formula for Legendre polynomials can be obtained from the addition formula for little q-Legendre polynomials by letting $q\to 1$, and in § 4 we obtain the familiar product formulas for Legendre polynomials from the product formulas for little q-Legendre polynomials.

2. Weak asymptotics for Wall polynomials. For little q-Jacobi polynomials $p_n(x;a,b\,|\,q)$ we can put $a=q^\alpha$ and $b=q^\beta$ and let $q\uparrow 1$ to find Jacobi polynomials on $[0,1]$. However, if either a or b is zero, which is exactly what happens for Wall polynomials, then the limit as $q\uparrow 1$ is $(1+(x/(a-1))^n$. Therefore another approach is needed to handle the behaviour of Wall polynomials as $q\uparrow 1$. It turns out that we can find some relevant results if we consider the polynomials $W_n(x;b,c^{1/n})$ for $n\to\infty$. We will prove a more general result for orthonormal polynomials $\{p_k(x;n)\colon k=0,1,2,\cdots;\ n\in\mathbb{N}\}$, where k is the degree of the polynomial and n an extra (discrete) parameter. The recurrence formula for these polynomials is given by

$$xp_k(x;n)=a_{k+1,n}p_{k+1}(x;n)+b_{k,n}p_k(x;n)+a_{k,n}p_{k-1}(x;n), \tag{2.1}$$

where $a_{k,n}>0$, $b_{k,n}\in\mathbb{R}$, $p_0(x;n)=1$, and $p_{-1}(x;n)=0$. Orthogonal polynomials with regularly varying recurrence coefficients [15] are of this type.

THEOREM 1. *Assume that* $[r, s]$ *is a finite interval that, for all* n, *contains the support of the orthogonality measure for* $\{p_k(x; n)\}$. *Assume moreover that*

$$\lim_{n\to\infty} a_{n,n} = A > 0, \qquad \lim_{n\to\infty} b_{n,n} = B \in \mathbb{R} \tag{2.2}$$

and that

$$\lim_{n\to\infty} (a_{k,n}^2 - a_{k-1,n}^2) = 0, \qquad \lim_{n\to\infty} (b_{k,n} - b_{k-1,n}) = 0, \tag{2.3}$$

uniformly in k, *then*

$$\lim_{n\to\infty} \frac{p_{n+1}(x; n)}{p_n(x; n)} = \rho\left(\frac{x-B}{2A}\right), \tag{2.4}$$

uniformly on compact sets of $\mathbb{C}\backslash[r, s]$, *where* $\rho(x) = x + \sqrt{x^2-1}$ (*the square root here is defined to be the one for which* $|\rho(x)| > 1$ *for* $x \in \mathbb{C}\backslash[-1, 1]$).

Proof. Let K be a compact set in $\mathbb{C}\backslash[r, s]$; then the distance between K and $[r, s]$ is strictly positive. Denote this distance by $\delta > 0$. A decomposition into partial fractions gives

$$\frac{p_{k-1}(x; n)}{p_k(x; n)} = a_{k,n} \sum_{j=1}^{k} \frac{d_{j,k}}{x - x_{j,k}},$$

where $\{x_{j,k}: 1 \leqq j \leqq k\}$ are the zeros of $p_k(x; n)$ and $\{d_{j,k}: 1 \leqq j \leqq k\}$ are positive numbers adding up to 1. Since all the zeros of $p_k(x; n)$ are in $[r, s]$ we have $|x - x_{j,k}| > \delta$ for $x \in K$ and therefore

$$\left|\frac{p_{k-1}(x; n)}{p_k(x, n)}\right| < \frac{a_{k,n}}{\delta} \tag{2.5}$$

holds uniformly for $x \in K$. Consider the Turán determinant

$$D_k(x; n) = p_k^2(x; n) - \frac{a_{k+1,n}}{a_{k,n}} p_{k+1}(x; n) p_{k-1}(x; n).$$

By using the recurrence relation (2.1) we find

$$\begin{aligned} D_k(x; n) = D_{k-1}(x; n) &+ \frac{b_{k,n} - b_{k-1,n}}{a_{k,n}} p_k(x; n) p_{k-1}(x; n) \\ &+ \frac{a_{k,n}^2 - a_{k-1,n}^2}{a_{k,n} a_{k-1,n}} p_{k-2}(x; n) p_k(x; n) \end{aligned} \tag{2.6}$$

(see [14, Thm. 4.10, p. 117]). If we define

$$R_{k,n}(x) = \frac{D_k(x; n)}{p_{k+1}(x; n) p_k(x; n)},$$

then by (2.6)

$$\begin{aligned} |R_{k,n}(x)| \leqq |R_{k-1,n}(x)| &\left|\frac{p_{k-1}(x; n)}{p_{k+1}(x; n)}\right| + \frac{|b_{k,n} - b_{k-1,n}|}{a_{k,n}} \left|\frac{p_{k-1}(x; n)}{p_{k+1}(x; n)}\right| \\ &+ \frac{|a_{k,n}^2 - a_{k-1,n}^2|}{a_{k,n} a_{k-1,n}} \left|\frac{p_{k-2}(x; n)}{p_{k+1}(x; n)}\right|, \end{aligned}$$

so that by (2.5) we have for $x \in K$

$$|R_{k,n}(x)| \leqq \frac{a_{k,n} a_{k+1,n}}{\delta^2} |R_{k-1,n}(x)| + |b_{k,n} - b_{k-1,n}| \frac{a_{k+1,n}}{\delta^2} + |a_{k,n}^2 - a_{k-1,n}^2| \frac{a_{k+1,n}}{\delta^3}.$$

By the conditions imposed there exists a constant C such that $a_{k,n} < C$ for every n and k (cf. [4, Chap. IV, Example 2.12]). Therefore, by (2.3),

$$|R_{k,n}(x)| \leqq \left(\frac{C}{\delta}\right)^2 |R_{k-1,n}(x)| + A_n, \qquad x \in K,$$

where $A_n \to 0$ as $n \to \infty$. Iteration gives

$$|R_{n,n}(x)| \leqq A_n \frac{(C/\delta)^{2n}-1}{(C/\delta)^2-1} + |R_{0,n}(x)|(C/\delta)^{2n}, \qquad x \in K.$$

If $\delta > C$ then obviously $R_{n,n}(x) \to 0$ as $n \to \infty$ (use $|R_{0,n}| = |p_0(x; n)/p_1(x; n)| < a_{1,n}/\delta$), which by (2.2), (2.3), and (2.5) leads to

$$\lim_{n\to\infty} \left| \frac{p_n(x; n)}{p_{n+1}(x; n)} - \frac{p_{n-1}(x; n)}{p_n(x; n)} \right| = 0, \tag{2.7}$$

uniformly for $x \in K$ (provided $\delta > C$). By (2.5) the sequence of analytic functions $p_n(x; n)/p_{n+1}(x; n)$ is uniformly bounded on compact sets of $\mathbb{C}\backslash[r, s]$ and thus there exists a subsequence converging to some function $L(x)$, uniformly on K. Use the recurrence formula (2.1) and the properties (2.2), (2.3), and (2.7) to find that this limit satisfies

$$x = \frac{A}{L(x)} + B + AL(x),$$

and since $|p_n(x; n)/p_{n+1}(x; n)| < C/\delta < 1$ for $x \in K$ by (2.5) we have

$$\frac{1}{L(x)} = \rho\left(\frac{x-B}{2A}\right).$$

This gives the result for $\delta > C$. This can be extended to hold for $\delta > 0$ by using the Stieltjes-Vitali theorem (cf. [4, p. 121]) and the uniform bound (2.5). □

Remark. The asymptotic behaviour actually holds uniformly on compact sets of $\mathbb{C}\backslash\Omega$, where Ω is the closure of the set of zeros of $p_n(x; n)$ as n runs through the integers. Clearly, Ω is a subset of $[r, s]$ since the zeros of $p_n(x; n)$ are all inside the interval $[r, s]$. The condition that the joint supports of the orthogonality measures should be contained in the finite interval $[r, s]$ can also be relaxed. Only the zeros of $p_k(x; n)(k \leqq n+1, n = 0, 1, 2, \cdots)$ must lie in $[r, s]$.

COROLLARY 1. *Suppose* $0 < b < 1$ *and* $0 < c < 1$. *Then*

$$\lim_{n\to\infty} \frac{W_{n+k}(x; b, c^{1/n})}{W_n(x; b, c^{1/n})} = \{b(1-c)(1-bc)c^2\}^{k/2} \rho^k\left(\frac{x-[b+1-2bc]c}{2c\sqrt{b(1-c)(1-bc)}}\right) \tag{2.8}$$

uniformly on compact sets of $\mathbb{C}\backslash[0, 1]$.

Proof. The proof follows immediately from

$$\lim_{n\to\infty} \frac{W_{n+k}(x; b, c^{1/n})}{W_{n+k-1}(x; b, c^{1/n})} - \{b(1-c)(1-bc)c^2\}^{1/2} \rho\left(\frac{x-[b+1-2bc]c}{2c\sqrt{b(1-c)(1-bc)}}\right),$$

which in turn can be proved by using Theorem 1 with recurrence coefficients $a_{k,n} = a_k(b, c^{1/n})$ and $b_{k,n} = b_k(b, c^{1/n})$ given by (1.6). □

COROLLARY 2. *Suppose* $0 < b < 1$ *and* $0 < c < 1$. *Then*

$$\lim_{n\to\infty} \frac{p_{n+k}(z; b, 0|c^{1/n})}{p_n(z; b, 0|c^{1/n})} = (-1)^k \left\{\frac{b(1-c)}{1-bc}\right\}^{k/2} \rho^k\left(\frac{z-[b+1-2bc]c}{2c\sqrt{b(1-c)(1-bc)}}\right) \tag{2.9}$$

uniformly for z on compact subsets of $\mathbb{C}\backslash[0,1]$, where $p_n(x; a, b \mid q)$ are the little q-Jacobi polynomials.

Proof. This follows immediately from (1.8) and Corollary 1. □

It is important in the asymptotic formula (2.4) that the variable x stays away from the zeros of $p_n(x; n)$. On the set Ω, the closure of the zeros of $p_n(x; n)$, the orthogonal polynomials will oscillate. The following theorem gives a result about the weak convergence of measures involving the polynomials $p_k(x; n)$ on $[r, s]$ in terms of their orthogonality measures.

THEOREM 2. *Assume that $[r, s]$ is a finite interval that, for all n, contains the support of the orthogonality measure μ_n for the orthonormal polynomials $\{p_k(x; n): k = 0, 1, 2, \cdots\}$. Assume, moreover, that for all $k \in \mathbb{Z}$*

$$\lim_{n\to\infty} a_{n+k,n} = A, \qquad \lim_{n\to\infty} b_{n+k,n} = B; \tag{2.10}$$

then for every continuous function f on $[r, s]$

$$\lim_{n\to\infty} \int_r^s f(z)p_n(z; n)p_{n+k}(z; n)d\mu_n(z) = \frac{1}{\pi}\int_{B-2A}^{B+2A} \frac{f(z)T_k((z-B)/(2A))}{\sqrt{4A^2-(z-B)^2}}\,dz,$$

where $T_n(x)$ are the Chebyshev polynomials of the first kind.

Proof. We follow the ideas of Nevai and Dehesa [10, Lemma 3]. Let m be a positive integer and apply the recurrence formula (2.1) repeatedly to get

$$z^m p_n(z; n) = \sum_{\substack{-1\le k_i\le 1\\ i=1,2,\cdots,m}} \alpha_{n,n+k_1}\alpha_{n+k_1,n+k_1+k_2}\cdots\alpha_{n+k_1+\cdots+k_{m-1},n+k_1+\cdots+k_m}p_{n+k_1+\cdots+k_m}(z; n),$$

where

$$\alpha_{j,k} = \begin{cases} a_{j,n} & \text{if } k = j-1,\\ b_{j,n} & \text{if } k = j,\\ a_{j+1,n} & \text{if } k = j+1. \end{cases}$$

Hence

$$\int_r^s z^m p_n(z; n)p_{n+k}(z; n)d\mu_n(z) = \sum_{\substack{-1\le k_i\le 1\\ i=1,2,\cdots,m\\ k_1+\cdots+k_m=k}} \alpha_{n,n+k_1}\alpha_{n+k_1,n+k_1+k_2}\cdots\alpha_{n+k_1+\cdots+k_{m-1},n+k}.$$

Because of this equation and by (2.10) it follows that the limit as $n\to\infty$ of $\int_r^s z^m p_n(z; n)p_{n+k}(z; n)d\mu_n(z)$ is the same as the limit of

$$\frac{1}{2A^2\pi}\int_{B-2A}^{B+2A} z^m U_n\left(\frac{z-B}{2A}\right)U_{n+k}\left(\frac{z-B}{2A}\right)\sqrt{4A^2-(z-B)^2}\,dz$$

since the Chebyshev polynomials of the second kind $U_n((z-B)/2A)$ are the orthogonal polynomials with constant recurrence coefficients $a_n = A$ and $b_n = B$. Use the identity

$$U_n(x)U_{n+k}(x) = \frac{1}{2}\frac{T_k(x) - T_{2n+k+2}(x)}{1-x^2}$$

to find

$$\frac{1}{2A^2\pi}\int_{B-2A}^{B+2A} z^m U_n\left(\frac{z-B}{2A}\right)U_{n+k}\left(\frac{z-B}{2A}\right)\sqrt{4A^2-(z-B)^2}\,dz$$
$$= \frac{1}{\pi}\int_{B-2A}^{B+2A} \frac{z^m T_k((z-B)/2A)}{\sqrt{4A^2-(z-B)^2}}\,dz - \frac{1}{\pi}\int_{B-2A}^{B+2A} \frac{z^m T_{2n+k+2}((z-B)/2A)}{\sqrt{4A^2-(z-B)^2}}\,dz.$$

If $2n+k+2>m$ then the second term on the right-hand side vanishes because of orthogonality, and thus we have the result when $f(x)=x^m$. The general result follows from the Hahn-Banach theorem: let the operators $L_{k,n}(k, n=0, 1, 2, \cdots)$, defined on the Banach space $C[r, s]$ of continuous functions equipped with the supremum norm, be given by

$$L_{k,n}f=\int_r^s f(z)p_n(z; n)p_{n+k}(z; n)\, d\mu_n(z).$$

These are uniformly bounded operators because, by Schwarz's inequality and the orthonormality,

$$\begin{aligned}\left|\int_r^s f(z)p_n(z; n)p_{n+k}(z; n)\, d\mu_n(z)\right|^2 &\\ \leqq \int_r^s |f(z)|p_n^2(z; n)\, d\mu_n(z) \int_r^s |f(z)|p_{n+k}^2(z; n)\, d\mu_n(z) &\\ \leqq \|f\|_\infty^2.&\end{aligned}$$

Now use Weierstrass's result that the polynomials form a dense subspace of $C[r, s]$. □

COROLLARY 3. *Suppose* $0<b<1$ *and* $0<c<1$. *Then for every continuous function* f *on* $[0, 1]$

$$\begin{aligned}\lim_{n\to\infty}\int_0^1 f(z)w_n(z; b, c^{1/n})w_{n+k}(z; b, c^{1/n})d\mu(z; b, c^{1/n})&\\ =\frac{1}{\pi}\int_{B-2A}^{B+2A}\frac{f(z)T_k((z-B)/(2A)}{\sqrt{4A^2-(z-B)^2}}\, dz,&\end{aligned}$$

where $A=c\sqrt{b(1-c)(1-bc)}$, $B=(b+1-2bc)c$, *and* $T_n(x)$ *are the Chebyshev polynomials of the first kind.*

Proof. The proof follows because the Wall polynomials $w_n(x; b, c^{1/n})$ satisfy the conditions of Theorem 2, with recurrence coefficients $a_{k,n}=a_k(b, c^{1/n})$ and $b_k(b, c^{1/n})$ given by (1.6). □

3. The addition formula. The little q-Legendre polynomials $p_n(z; 1, 1|q)$ and the Wall polynomials $p_n(z; a, 0|q)$ are analytic functions of z and the addition formula (1.9) holds for every $z\in\{q^n: n=0, 1, 2, \cdots\}$ (which is a set with an accumulation point). Therefore it follows that

$$\begin{aligned}&p_m(z; 1, 1|q)p_y(z; q^x, 0|q)\\ &\quad=p_m(q^{x+y}; 1, 1|q)p_m(q^y; 1, 1|q)p_y(z; q^x, 0|q)\\ &\qquad+\sum_{k=1}^m\frac{(q; q)_{x+y+k}(q; q)_{m+k}q^{k(y-m+k)}}{(q; q)_{x+y}(q; q)_{m-k}(q; q)_k^2}p_{m-k}(q^{x+y}; q^k, q^k|q)\\ &\qquad\cdot p_{m-k}(q^y; q^k, q^k|q)p_{y+k}(z; q^x, 0|q)\\ &\qquad+\sum_{k=1}^m\frac{(q; q)_y(q; q)_{m+k}q^{k(x+y-m+1)}}{(q; q)_{y-k}(q; q)_{m-k}(q; q)_k^2}p_{m-k}(q^{x+y-k}; q^k, q^k|q)\\ &\qquad\cdot p_{m-k}(q^{y-k}; q^k, q^k|q)p_{y-k}(z; q^x, 0|q)\end{aligned}\tag{3.1}$$

holds for every $z\in\mathbb{C}$ and $x, y=0, 1, 2, \cdots$. It is well known that

$$\lim_{q\uparrow 1}p_n(z; q^\alpha, q^\beta|q)=R_n^{(\alpha,\beta)}(1-2z),\tag{3.2}$$

where $R_n^{(\alpha,\beta)}(x)$ are Jacobi polynomials with the normalization $R_n^{(\alpha,\beta)}(1)=1$, i.e., $R_n^{(\alpha,\beta)}(x)=P_n^{(\alpha,\beta)}(x)/P_n^{(\alpha,\beta)}(1)$. Fix b, c in $(0,1)$ such that $\log b/\log c=\beta/\gamma$ with β, γ positive integers, substitute in (3.1) $q=b^{1/(n\beta)}=c^{1/(n\gamma)}$, $x=n\beta$, $y=n\gamma$, and let $n\to\infty$ through the integers. Then by (2.9), (3.1), and (3.2)

$$\begin{aligned}R_m^{(0,0)}(1-2z)&=R_m^{(0,0)}(1-2bc)R_m^{(0,0)}(1-2b)\\&\quad+\sum_{k=1}^{m}\frac{(m+k)!}{(m-k)!(k!)^2}(1-bc)^k c^k R_{m-k}^{(k,k)}(1-2bc)R_{m-k}^{(k,k)}(1-2c)\\&\quad\cdot(-1)^k\left\{\frac{b(1-c)}{1-bc}\right\}^{k/2}\rho^k\left(\frac{z-[b+1-2bc]c}{2c\sqrt{b(1-c)(1-bc)}}\right)\\&\quad+\sum_{k=1}^{m}\frac{(m+k)!}{(m-k)!(k!)^2}(1-c)^k(bc)^k R_{m-k}^{(k,k)}(1-2bc)R_{m-k}^{(k,k)}(1-2c)\\&\quad\cdot(-1)^k\left\{\frac{1-bc}{b(1-c)}\right\}^{k/2}\rho^{-k}\left(\frac{z-[b+1-2bc]c}{2c\sqrt{b(1-c)(1-bc)}}\right).\end{aligned}$$

Now use the formula $T_k(x)=[\rho^k(x)+\rho^{-k}(x)]/2$; then

$$\begin{aligned}R_m^{(0,0)}(1-2z)&=R_m^{(0,0)}(1-2bc)R_m^{(0,0)}(1-2b)\\&\quad+2\sum_{k=1}^{m}(-1)^k\frac{(m+k)!}{(m-k)!(k!)^2}c^k[b(1-c)(1-bc)]^{k/2}\\&\quad\cdot R_{m-k}^{(k,k)}(1-2bc)R_{m-k}^{(k,k)}(1-2c)\,T_k\left(\frac{z-[b+1-2bc]c}{2c\sqrt{b(1-c)(1-bc)}}\right).\end{aligned}$$

Finally, choose

$$\begin{aligned}1-2z&=xy-\sqrt{1-x^2}\sqrt{1-y^2}\,t,\\1-2bc&=x,\\1-2c&=y;\end{aligned}$$

then

$$\begin{aligned}R_m^{(0,0)}(xy-\sqrt{1-x^2}\sqrt{1-y^2}\,t)&=R_m^{(0,0)}(x)R_m^{(0,0)}(y)\\&\quad+2\sum_{k=1}^{m}(-1)^k\frac{(m+k)!}{(m-k)!(k!)^2}2^{-2k}\{\sqrt{1-x^2}\sqrt{1-y^2}\}^k\\&\quad\cdot R_{m-k}^{(k,k)}(x)R_{m-k}^{(k,k)}(y)\,T_k(t),\end{aligned}$$

which is the familiar addition formula for Legendre polynomials. By our method of proof this formula only holds for $t\in\mathbb{C}\backslash\mathbb{R}$ (because we use Corollary 2), but since all the functions considered are analytic in t, the result definitely holds for every $t\in\mathbb{C}$.

4. Product formulas. If we multiply both sides of the addition formula (1.9) by $p_{y+k}(q^z;q^x,0|q)q^{(x+1)z}/(q;q)_z$ and sum from $z=0$ to ∞, then by the orthogonality (1.4) and by (1.8)

$$\begin{aligned}&\sum_{z=0}^{\infty}p_m(q^z;1,1|q)p_y(q^z;q^x,0|q)p_{y+k}(q^z;q^x,0|q)\frac{q^{(x+1)z}}{(q;q)_z}\\&\quad=\frac{(q;q)_{x+y+k}(q;q)_{m+k}q^{k(y-m+k)}}{(q;q)_{x+y}(q;q)_{m-k}(q;q)_k^2}p_{m-k}(q^{x+y};q^k,q^k|q)p_{m-k}(q^y;q^k,q^k|q)\\&\quad\cdot\sum_{z=0}^{\infty}p_{y+k}^2(q^z;q^x,0|q)\frac{q^{(x+1)z}}{(q;q)_z},\end{aligned}$$

which holds whenever $k \in \{0, 1, \cdots, m\}$. In terms of orthonormal Wall polynomials we have by (1.8)

$$
\begin{aligned}
& p_{m-k}(q^{x+y}; q^k, q^k \mid q) p_{m-k}(q^y; q^k, q^k \mid q) \\
& \quad = (-1)^k \frac{(q;q)_{m-k}(q;q)_k^2}{(q;q)_{m+k}} q^{-k(y+k-m)} \left\{ q^{-k(x+1)} \frac{(q;q)_y (q;q)_{x+y}}{(q;q)_{y+k}(q;q)_{x+y+k}} \right\}^{1/2} \\
& \quad \cdot (q^{x+1}; q)_\infty \sum_{z=0}^{\infty} p_m(q^z; 1, 1 \mid q) w_y(q^{z+1}; q^{x+1}, q) \\
& \quad \cdot w_{y+k}(q^{z+1}; q^{x+1}, q) \frac{q^{(x+1)z}}{(q;q)_z},
\end{aligned} \tag{4.1}
$$

which can be considered as a product formula for the little q-Legendre polynomials and which (for $k=0$) is equivalent with the product formula given by Koornwinder [8]. If we use the notation (1.7) then

$$
\begin{aligned}
& (q^{x+1}; q)_\infty \sum_{z=0}^{\infty} p_m(q^z; 1, 1 \mid q) w_y(q^{z+1}; q^{x+1}, q) w_{y+k}(q^{z+1}; q^{x+1}, q) \frac{q^{(x+1)z}}{(q;q)_z} \\
& \quad = \int_0^1 p_m\left(\frac{z}{q}; 1, 1 \mid q\right) w_y(z; q^{x+1}, q) w_{y+k}(z; q^{x+1}, q)\, d\mu(z; q^{x+1}, q).
\end{aligned}
$$

Fix b, c in $(0,1)$ such that $\log b / \log c = \beta/\gamma$ with β and γ positive integers and let $q = b^{1/(n\beta)} = c^{1/(n\gamma)}$, $1+x = n\beta$, $y = n\gamma$. Then as $n \to \infty$ we have by Corollary 3 and by the uniform convergence in (3.2) (keep in mind that $p_m((z/q); 1, 1 \mid q)$ is a polynomial of degree m)

$$
\begin{aligned}
R_{m-k}^{(k,k)}(1-2bc) R_{m-k}^{(k,k)}(1-2c) &= (-1)^k \frac{(m-k)!(k!)^2}{(m+k)!} c^{-k} \{b(1-c)(1-bc)\}^{-k/2} \\
&\quad \cdot \frac{1}{\pi} \int_{B-2A}^{B+2A} R_m^{(0,0)}(1-2z) \frac{T_k((z-B)/2A)}{\sqrt{4A^2-(z-B)^2}}\, dz,
\end{aligned}
$$

where $A = c\sqrt{b(1-c)(1-bc)}$ and $B = (b+1-2bc)c$. Setting $bc = x$, $c = y$ gives the familiar product formulas for Legendre polynomials:

$$
\begin{aligned}
R_{m-k}^{(k,k)}(1-2x) R_{m-k}^{(k,k)}(1-2y) &= (-1)^k \frac{(m-k)!(k!)^2}{(m+k)!} \{xy(1-y)(1-x)\}^{-k/2} \\
&\quad \cdot \frac{1}{\pi} \int_{B-2A}^{B+2A} R_m^{(0,0)}(1-2z) \frac{T_k((z-B)/2A)}{\sqrt{4A^2-(z-B)^2}}\, dz,
\end{aligned}
$$

with $A = \sqrt{xy(1-x)(1-y)}$ and $B = x+y-2xy$.

REFERENCES

[1] G. ANDREWS AND R. ASKEY, *Enumeration of partitions: The role of Eulerian series and q-orthogonal polynomials*, in Higher Combinatorics, M. Aigner, ed., D. Reidel, Dordrecht, the Netherlands, 1977, pp. 3-26.

[2] R. ASKEY, *Orthogonal Polynomials and Special Functions*, CBMS-NSF Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[3] R. ASKEY AND J. WILSON, *Some Basic Hypergeometric Orthogonal Polynomials that Generalize Jacobi Polynomials*, Mem. Amer. Math. Soc. 319, Providence, RI , 1985.

[4] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[5] ———, *Orthogonal polynomials with Brenke type generating functions*, Duke Math. J., 35 (1968), pp. 505-518.

[6] G. GASPER AND M. RAHMAN, *Basic Hypergeometric Series*, in Encyclopedia of Mathematics and Its Applications, Vol. 35, Cambridge University Press, Cambridge, 1990.

[7] T. H. KOORNWINDER, *Representations of the twisted* SU(2) *quantum group and some q-hypergeometric orthogonal polynomials*, Indag. Math., 51 (1989), pp. 97–117.

[8] ———, *The addition formula for little q-Legendre polynomials and the* SU(2) *quantum group*, SIAM J. Math. Anal., this issue (1991), pp. 292–301.

[9] T. MASUDA, K. MIMACHI, Y. NAKAGAMI, M. NOUMI AND K. UENO, *Representations of quantum groups and a q-analogue of orthogonal polynomials*, C.R. Acad. Sci. Paris Sér. 1. Math., 307 (1988), pp. 559–564.

[10] P. G. NEVAI AND J. S. DEHESA, *On asymptotic average properties of zeros of orthogonal polynomials*, SIAM J. Math. Anal., 10 (1979), pp. 1184–1192.

[11] M. RAHMAN, *A simple proof of Koornwinder's addition formula for the little q-Legendre polynomials*, Proc. Amer. Math. Soc., 107 (1989), pp. 373–381.

[12] M. RAHMAN AND A. VERMA, *Product and addition formula for the continuous q-ultraspherical polynomials*, SIAM J. Math. Anal., 17 (1986), pp. 1461–1474.

[13] L. L. VAKSMAN AND YA. S. SOIBELMAN, *Function algebra on the quantum group* SU(2), Funktsional. Anal. i Prilozhen, 22 (1988), pp. 1–14. (In Russian.) Funct. Anal. Appl., 22 (1988), pp. 170–181.

[14] W. VAN ASSCHE, *Asymptotics for Orthogonal Polynomials*, Lecture Notes in Math., 1265, Springer-Verlag, Berlin, New York, 1987.

[15] W. VAN ASSCHE AND J. S. GERONIMO, *Asymptotics for orthogonal polynomials with regularly varying recurrence coefficients*, Rocky Mountain J. Math., 19 (1989), pp. 39–49.

[16] H. S. WALL, *A continued fraction related to some partition formulas of Euler*, Amer. Math. Monthly, 48 (1941), pp. 102–108.

[17] S. L. WORONOWICZ, *Compact matrix pseudogroups*, Comm. Math. Phys., 111 (1987), pp. 613–665.

[18] ———, *Twisted* SU(2) *group. An example of a non-commutative differential calculus*, Publ. Res. Inst. Math. Sci., 23 (1987), pp. 117–181.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 313–327, March 1991

AN EXISTENCE THEOREM FOR MODEL EQUATIONS RESULTING FROM KINETIC THEORIES OF POLYMER SOLUTIONS*

MICHAEL RENARDY†

Abstract. A local existence and uniqueness theorem is proved for a set of partial differential equations modelling the flow of polymer solutions. The constitutive relations considered here are motivated by kinetic theory. The stress tensor is given by an integral which involves the solution of a linear diffusion equation. The coefficients of this diffusion equation depend on the velocity gradient.

Key words. polymer rheology, kinetic theories, local existence

AMS(MOS) subject classifications. 35K15, 35L20, 35Q99, 76A10

1. Introduction.

General existence theorems for the partial differential equations of continuum mechanics are relatively recent, even for the case of elasticity. The initial value problem for compressible elastic materials occupying all of space was solved in [6], and the Dirichlet initial-boundary value problem is treated in [7], [2], and [3]. Existence results for incompressible materials were obtained in [4] and [12] for the initial value problem on all of space and in [5] for the Dirichlet problem.

Viscoelastic fluids such as polymers are characterized by a constitutive relation which gives the stress as a functional of the deformation history. Most popular models fall into two categories: integral and differential models. Integral models give the stress by an integral expression involving the history of the deformation gradient; differential models give the stress as the solution of a differential equation which involves the stress and velocity gradient. There is, however, a class of models motivated by considerations of kinetic theory, which can in general not be represented in either integral or differential form. Instead, such models require the solution of a diffusion equation in order to determine the stress [1]. Because of the practical impossibility—except in special cases—of solving this diffusion equation, such models have only had limited applications in solving flow problems. For the case of dumbbell models, however, computers may soon have the capacity to obtain solutions, at least in two-dimensional flow problems.

For materials with instantaneous elasticity, the terms of leading differential order in the equations of motion are like those for elasticity. A natural approach to a mathematical existence theory is therefore to regard the problem as essentially "elastic" and to treat the memory as a perturbation. This has been successfully carried out for integral models; see, e.g., Chapter III of [9], and also for differential models [11]. In this paper, we shall discuss the models from kinetic theory in a similar fashion. We shall in part be able to rely on the results of [11], but the solution of the diffusion equation raises new issues.

*Received by the editors October 18, 1989; accepted for publication (in revised form) February 19, 1990. The author began this research while visiting the Institute for Mathematics and Its Applications, University of Minnesota. This research was supported in part by the Institute for Mathematics and Its Applications and by National Science Foundation grant DMS-8796241.

†Department of Mathematics and Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0123.

The equations of motion for an incompressible fluid are

$$\rho\Big(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}\Big) = \operatorname{div}\mathbf{T} - \nabla p + \mathbf{f}, \qquad \operatorname{div}\mathbf{v} = 0. \tag{1}$$

Here $\mathbf{v}$ denotes the velocity, p the pressure, $\mathbf{T}$ the extra stress tensor, ρ the density (assumed constant), and $\mathbf{f}$ is a given body force. We investigate molecular models which are based on treating polymer molecules as dumbbells consisting of two beads connected by a spring and floating in a Newtonian solvent. We shall assume that the solvent contribution to the stress is small relative to the polymer contribution and can be neglected; if this is not the case, the equations of motion can be treated as a perturbation of the Stokes equations, and existence results can be obtained along the lines of [10]. Let $\mathbf{R}$ denote the vector between the two ends of the dumbbell, and let $\psi(\mathbf{R},\mathbf{x},t)$ denote the probability density for dumbbells in $\mathbf{R}$-space (this probability density depends on the point $\mathbf{x}$ in space and the time t; since ψ has the interpretation of a probability density, it will always be nonnegative and $\int \psi(\mathbf{R},\mathbf{x},t)\,d\mathbf{R} = 1$). The spring force is given by $\mathbf{F}(\mathbf{R}) = -\gamma(|\mathbf{R}|^2)\mathbf{R}$, and the stress is given by

$$T_{ij}(\mathbf{x},t) = -n\int R_i F_j(\mathbf{R})\psi(\mathbf{R},\mathbf{x},t)\,d\mathbf{R}. \tag{2}$$

Here n is the number density of dumbbells and the integral extends over all possible values of $\mathbf{R}$. We shall consider infinitely extensible as well as finitely extensible dumbbells. For infinitely extensible dumbbells, $\mathbf{R}$ can be any vector in $\mathbb{R}^3$. For finitely extensible dumbbells, the potential associated with $\mathbf{F}(\mathbf{R})$ becomes infinite at a finite value R_0 of $|\mathbf{R}|$, and the integral in (2) extends only over the ball $|\mathbf{R}| \le R_0$. The probability density $\psi(\mathbf{R},\mathbf{x},t)$ obeys the diffusion equation

$$\Big(\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla_{\mathbf{x}})\Big)\psi = \alpha\Delta_{\mathbf{R}}\psi + \operatorname{div}_{\mathbf{R}}\Big(-(\nabla_{\mathbf{x}}\mathbf{v})\cdot\mathbf{R}\psi - \beta\mathbf{F}(\mathbf{R})\psi\Big). \tag{3}$$

Here we have used subscripts $\mathbf{R}$ and $\mathbf{x}$ to indicate the variables with respect to which derivatives are being taken. Our convention for the gradient of a vector is that the first index refers to the component of the vector and the second index to the direction of differentiation. The quantities α and β are positive constants. The reader is referred to [1] for the derivation of (2) and (3) and for a discussion of possible generalizations.

We seek solutions of (1)–(3) for $\mathbf{x}\in\Omega$, $t>0$, where Ω is a bounded domain in $\mathbb{R}^3$ with smooth boundary. We impose Dirichlet boundary conditions,

$$\mathbf{v}(\mathbf{x},t) = \mathbf{0}, \quad \mathbf{x}\in\partial\Omega, \quad t\ge 0, \tag{4}$$

and initial conditions,

$$\mathbf{v}(\mathbf{x},0) = \mathbf{v}_0(\mathbf{x}), \qquad \psi(\mathbf{R},\mathbf{x},0) = \psi_0(\mathbf{R},\mathbf{x}). \tag{5}$$

Concerning the spring force, we shall make one of the following two assumptions:

(F) The function γ is C^∞-smooth from $[0,\infty)$ to $(0,\infty)$, $\gamma' \ge 0$, and there exist numbers $\sigma \ge 0$ and $k > 0$ such that $\lim_{|\mathbf{R}|\to\infty}\gamma(|\mathbf{R}|^2)/|\mathbf{R}|^\sigma = k$. Moreover, $\limsup_{|\mathbf{R}|\to\infty}\gamma'(|\mathbf{R}|^2)/|\mathbf{R}|^{\sigma-2}$ is finite and higher derivatives of γ have at most polynomial growth as $|\mathbf{R}|\to\infty$.

(F′) For some $R_0 > 0$, the function γ is C^∞-smooth from $[0, R_0^2)$ to $(0, \infty)$, $\gamma' \geq 0$, and there exist numbers $\sigma > 1$ and $k > 0$ such that $\lim_{|\mathbf{R}| \to R_0} \gamma(|\mathbf{R}|^2)(R_0 - |\mathbf{R}|)^\sigma = k$. Moreover, $\limsup_{|\mathbf{R}| \to R_0} \gamma'(|\mathbf{R}|^2)(R_0 - |\mathbf{R}|)^{\sigma+1}$ is finite and higher derivatives of γ grow at most like powers of $(R_0 - |\mathbf{R}|)^{-1}$ as $|\mathbf{R}| \to R_0$.

In the first case, the dumbbell is infinitely extensible, in the second case it is finitely extensible. The assumption $\gamma' \geq 0$ means that the spring gets stiffer as the dumbbell is extended; this assumption is usually made in molecular models. The analysis that follows can be modified to allow faster growth of γ as $|\mathbf{R}| \to \infty$ or, respectively, $|\mathbf{R}| \to R_0$, but slower growth would pose difficulties. In particular, we have to exclude $\sigma = 1$ in (F′). This case corresponds to the most popular model of finitely extensible dumbbells (see [1]). There is, however, no specific reason to assume $\sigma = 1$ rather than a larger value of σ.

The goal of this paper is an existence and uniqueness theorem for solutions of the problem defined by (1)–(5). The paper is organized as follows. In §2, we define an iteration scheme which is used to construct the solution. The scheme alternates between solving an equation of the same type as encountered in incompressible elasticity and solving a linear diffusion equation such as (3). In §3 we define the function spaces used to carry out the analysis and state a precise theorem. The proof is given in §4.

2. Iterative construction of solution. We apply the operation $\partial/\partial t + (\mathbf{v}\cdot\nabla) + (\nabla\mathbf{v})^T$ to the equation of motion (1). The resulting equation, written in components, reads (here and in the rest of the paper, the summation convention is used)

$$
\begin{aligned}
\rho\left[\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla)\right]^2 v_i = &\frac{\partial}{\partial x_j}\left[\left(\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla)\right)T_{ij}\right] - \frac{\partial q}{\partial x_i} \\
&- \frac{\partial v_k}{\partial x_j}\frac{\partial T_{ij}}{\partial x_k} + \frac{\partial v_k}{\partial x_i}\frac{\partial T_{kj}}{\partial x_j} - \rho\frac{\partial v_j}{\partial x_i}\left[\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla)\right]v_j \\
&+ \left[\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla)\right]f_i + \frac{\partial v_j}{\partial x_i}f_j,
\end{aligned} \tag{6}
$$

where $q = \partial p/\partial t + (\mathbf{v}\cdot\nabla)p$. Using (2), we find that

$$
\left[\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla)\right]T_{ij} = -n\int R_iF_j(\mathbf{R})\left[\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla_{\mathbf{x}})\right]\psi(\mathbf{R},\mathbf{x},t)\,d\mathbf{R}, \tag{7}
$$

and using (3), we find further

$$
\begin{aligned}
&\left[\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla)\right]T_{ij} \\
&\quad = -n\int R_iF_j(\mathbf{R})\left[\alpha\Delta_{\mathbf{R}}\psi + \operatorname{div}_{\mathbf{R}}\left(-(\nabla_{\mathbf{x}}\mathbf{v})\cdot\mathbf{R}\psi - \beta\mathbf{F}(\mathbf{R})\psi\right)\right]d\mathbf{R} \\
&\quad = -n\alpha\int \Delta_{\mathbf{R}}(R_iF_j(\mathbf{R}))\psi\,d\mathbf{R} - n\beta\int\frac{\partial}{\partial R_k}(R_iF_j(\mathbf{R}))F_k(\mathbf{R})\psi\,d\mathbf{R} \\
&\qquad - n\int\frac{\partial}{\partial R_k}(R_iF_j(\mathbf{R}))\frac{\partial v_k}{\partial x_l}R_l\psi\,d\mathbf{R}.
\end{aligned} \tag{8}
$$

Next we note that

$$
\begin{aligned}
\frac{\partial}{\partial R_k}(R_iF_j(\mathbf{R})) &= -\frac{\partial}{\partial R_k}(R_iR_j\gamma(|\mathbf{R}|^2)) \\
&= -\gamma(|\mathbf{R}|^2)(R_j\delta_{ik} + R_i\delta_{jk}) - 2R_iR_jR_k\gamma'(|\mathbf{R}|^2).
\end{aligned} \tag{9}
$$

By combining (8) and (9), we find

$$
\begin{aligned}
&\frac{\partial}{\partial x_j}\left[\Big(\frac{\partial}{\partial t}+(\mathbf{v}\cdot\nabla)\Big)T_{ij}\right] \\
&\quad = n\frac{\partial^2 v_k}{\partial x_j \partial x_l}\int(\gamma(|\mathbf{R}|^2)R_jR_l\delta_{ik}+2\gamma'(|\mathbf{R}|^2)R_iR_jR_kR_l)\psi\ d\mathbf{R} \\
&\quad\quad - n\alpha\int\Delta_{\mathbf{R}}(R_iF_j(\mathbf{R}))\frac{\partial\psi}{\partial x_j}\ d\mathbf{R} - n\beta\int\frac{\partial}{\partial R_k}(R_iF_j(\mathbf{R}))F_k(\mathbf{R})\frac{\partial\psi}{\partial x_j}\ d\mathbf{R} \\
&\quad\quad - n\frac{\partial v_k}{\partial x_l}\int\frac{\partial}{\partial R_k}(R_iF_j(\mathbf{R}))R_l\frac{\partial\psi}{\partial x_j}\ d\mathbf{R}.
\end{aligned}
\tag{10}
$$

By inserting (10) into (6), we finally obtain

$$
\begin{aligned}
&\rho\left[\frac{\partial}{\partial t}+(\mathbf{v}\cdot\nabla)\right]^2 v_i \\
&\quad = n\frac{\partial^2 v_k}{\partial x_j \partial x_l}\int(\gamma(|\mathbf{R}|^2)R_jR_l\delta_{ik}+2\gamma'(|\mathbf{R}|^2)R_iR_jR_kR_l)\psi\ d\mathbf{R} - \frac{\partial q}{\partial x_i} \\
&\quad\quad - n\alpha\int\Delta_{\mathbf{R}}(R_iF_j(\mathbf{R}))\frac{\partial\psi}{\partial x_j}\ d\mathbf{R} - n\beta\int\frac{\partial}{\partial R_k}(R_iF_j(\mathbf{R}))F_k(\mathbf{R})\frac{\partial\psi}{\partial x_j}\ d\mathbf{R} \\
&\quad\quad - n\frac{\partial v_k}{\partial x_l}\int\frac{\partial}{\partial R_k}(R_iF_j(\mathbf{R}))R_l\frac{\partial\psi}{\partial x_j}\ d\mathbf{R} \\
&\quad\quad - \frac{\partial v_k}{\partial x_j}\frac{\partial T_{ij}}{\partial x_k}+\frac{\partial v_k}{\partial x_i}\frac{\partial T_{kj}}{\partial x_j} - \rho\frac{\partial v_j}{\partial x_i}\left[\frac{\partial}{\partial t}+(\mathbf{v}\cdot\nabla)\right]v_j \\
&\quad\quad + \left[\frac{\partial}{\partial t}+(\mathbf{v}\cdot\nabla)\right]f_i+\frac{\partial v_j}{\partial x_i}f_j.
\end{aligned}
\tag{11}
$$

The iteration is based on equations (3) and (11). Given an iterate $\mathbf{v}^m$, we determine ψ^m by solving equation (3),

$$
\left(\frac{\partial}{\partial t}+(\mathbf{v}^m\cdot\nabla_{\mathbf{x}})\right)\psi^m = \alpha\Delta_{\mathbf{R}}\psi^m+\operatorname{div}_{\mathbf{R}}\ (-(\nabla_{\mathbf{x}}\mathbf{v}^m)\cdot\mathbf{R}\psi^m-\beta\mathbf{F}(\mathbf{R})\psi^m),
\tag{12a}
$$

$$
\psi^m(\mathbf{R},\mathbf{x},0)=\psi_0(\mathbf{R},\mathbf{x}).
\tag{12b}
$$

We define $\mathbf{T}^m$ in terms of ψ^m by equation (2) and then determine a new velocity field $\mathbf{v}^{m+1}$ by solving the problem

$$
\begin{aligned}
&\rho\left[\frac{\partial}{\partial t}+(\mathbf{v}^m\cdot\nabla)\right]^2 v_i^{m+1} \\
&= n\frac{\partial^2 v_k^{m+1}}{\partial x_j \partial x_l}\int(\gamma(|\mathbf{R}|^2)R_jR_l\delta_{ik}+2\gamma'(|\mathbf{R}|^2)R_iR_jR_kR_l)\psi^m\ d\mathbf{R} - \frac{\partial q^{m+1}}{\partial x_i} \\
&\quad - n\alpha\int\Delta_{\mathbf{R}}(R_iF_j(\mathbf{R}))\frac{\partial\psi^m}{\partial x_j}\ d\mathbf{R} - n\beta\int\frac{\partial}{\partial R_k}(R_iF_j(\mathbf{R}))F_k(\mathbf{R})\frac{\partial\psi^m}{\partial x_j}\ d\mathbf{R} \\
&\quad - n\frac{\partial v_k^m}{\partial x_l}\int\frac{\partial}{\partial R_k}(R_iF_j(\mathbf{R}))R_l\frac{\partial\psi^m}{\partial x_j}\ d\mathbf{R} \\
&\quad - \frac{\partial v_k^m}{\partial x_j}\frac{\partial T_{ij}^m}{\partial x_k}+\frac{\partial v_k^m}{\partial x_i}\frac{\partial T_{kj}^m}{\partial x_j} - \rho\frac{\partial v_j^m}{\partial x_i}\left[\frac{\partial}{\partial t}+(\mathbf{v}^m\cdot\nabla)\right]v_j^m \\
&\quad + \left[\frac{\partial}{\partial t}+(\mathbf{v}^m\cdot\nabla)\right]f_i+\frac{\partial v_j^m}{\partial x_i}f_j,
\end{aligned}
\tag{13a}
$$

$$\text{div } \mathbf{v}^{m+1} = 0, \tag{13b}$$

$$\mathbf{v}^{m+1} = \mathbf{0}, \quad x \in \partial\Omega, \quad t \geq 0. \tag{13c}$$

$$\mathbf{v}^{m+1}(\mathbf{x}, 0) = \mathbf{v}_0(\mathbf{x}), \quad \frac{\partial \mathbf{v}^{m+1}}{\partial t}(\mathbf{x}, 0) = \mathbf{v}_1(\mathbf{x}). \tag{13d}$$

Here $\mathbf{v}_1$ is the initial value of $\partial \mathbf{v}/\partial t$, which can be computed by applying the Hodge projection to (1). Equations (13) also yield a new iterate for q, from which an approximation to the pressure p may be recovered. However, this is not necessary in order to proceed to the next step of the iteration.

3. Definition of function spaces and statement of results. Since ψ has the meaning of a probability density, L^1-spaces are natural for the $\mathbf{R}$-dependence. However, the definition of $\mathbf{T}$ in (2) involves integrals of ψ against increasing functions of $\mathbf{R}$, and we shall therefore use weighted L^1-spaces. We define

$$X_{n,0} = \left\{ \psi : \ \mathbb{R}^3 \to \mathbb{R} \ \middle| \ \int (1 + |\mathbf{R}|^n)|\psi(\mathbf{R})| \ d\mathbf{R} < \infty \right\}, \tag{14a}$$

in the case of infinitely extensible dumbbells, and

$$X_{n,0} = \left\{ \psi : \ B \to \mathbb{R} \ \middle| \ \int (R_0 - |\mathbf{R}|)^{-n}|\psi(\mathbf{R})| \ d\mathbf{R} < \infty \right\}, \tag{14b}$$

in the case of finitely extensible dummbells. Here B denotes the ball of radius R_0. Moreover, we let $X_{n,k}$ be the space of all ψ whose derivatives up to order k lie in $X_{n,0}$. Finally, let

$$X_k = \bigcap_{n=0}^{\infty} X_{n,k}, \tag{15}$$

with the natural topology of a Fréchet space. The space X_k consists of functions which, together with their derivatives, vanish at infinite order as $|\mathbf{R}| \to \infty$ or $|\mathbf{R}| \to R_0$, respectively. Instead of working with this Fréchet space, it would be possible to work with Banach spaces of functions vanishing at sufficiently high order; however, different orders of decay would have to be chosen for higher derivatives of ψ than for ψ itself. This would make the precise statement of a theorem rather laborious. The analysis that follows is not substantially complicated by the fact that X_k is a Fréchet space rather than a Banach space.

We make the following smoothness assumptions:

(S1) The domain $\Omega \subset \mathbb{R}^3$ is bounded and $\partial\Omega$ is of class C^5.

(S2) $\mathbf{v}_0 \in H^4(\Omega)$.

(S3) $\psi_0 \in \bigcap_{k=0}^{4} H^k(\Omega; X_{8-2k})$, where $H^k(\Omega; X_l)$ stands for $\bigcap_{n=0}^{\infty} H^k(\Omega; X_{n,l})$. Moreover, $\psi_0 \geq 0$ and $\int \psi_0(\mathbf{R}, \mathbf{x}) \ d\mathbf{R} = 1$ for every $\mathbf{x} \in \Omega$.

(S4) For some $T > 0$, we have $\mathbf{f} \in \bigcap_{k=0}^{4} W^{k,1}([0, T]; H^{4-k}(\Omega))$.

In addition, we need compatibility conditions between the initial data and the incompressibility and boundary conditions. Note that by applying the Hodge projection

operator to equation (1), we can obtain an initial value for $\partial \mathbf{v}/\partial t$, and, after differentiating (1) with respect to t, we can obtain initial values of higher time derivatives in an analogous fashion. We shall denote the initial value of $\partial^i \mathbf{v}/\partial t^i$ by $\mathbf{v}_i$. These initial values for time derivatives of $\mathbf{v}$ satisfy the incompressibility condition because of their construction, but we still need to require that they satisfy the boundary conditions. We assume the following compatibility conditions:

(C1) div $\mathbf{v}_0 = 0$ and $\mathbf{v}_0 = \mathbf{0}$ on $\partial\Omega$.

(C2) $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ vanish on $\partial\Omega$.

The goal of the paper is the following result.

THEOREM. *Assume that* (S1)–(S4), (C1), (C2), *and* (F) *or, respectively,* (F′) *hold. Then there is a* $T' \in (0,T]$ *such that the problem* (1)–(5) *has a unique solution with the regularity*

$$\mathbf{v} \in \bigcap_{k=0}^{4} C^k([0,T']; H^{4-k}(\Omega)); \quad \mathbf{T} \in \bigcap_{k=0}^{3} C^k([0,T']; H^{3-k}(\Omega));$$
$$\psi \in \bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} C^k([0,T']; H^l(\Omega; X_{8-2k-2l})). \tag{16}$$

The proof will be based on showing that the mapping $\Sigma : \mathbf{v}^m \to \mathbf{v}^{m+1}$ defined by (12) and (13) is a contraction in an appropriate complete space of functions. This space of functions, denoted by $Z(M,T')$, is defined as the set of all functions $\mathbf{v} : \Omega \times [0,T'] \to \mathbb{R}^3$ with the following properties:

$$\mathbf{v} \in \bigcap_{k=0}^{4} W^{k,\infty}([0,T']; H^{4-k}(\Omega)), \tag{17a}$$

$$\|\mathbf{v}\|_{0,4} + \|\mathbf{v}\|_{1,3} + \|\mathbf{v}\|_{2,2} + \|\mathbf{v}\|_{3,1} + \|\mathbf{v}\|_{4,0} \le M, \tag{17b}$$

$$\operatorname{div} \mathbf{v} = 0, \tag{17c}$$

$$\mathbf{v}|_{\partial\Omega} = \mathbf{0}, \tag{17d}$$

(17e)
$$\mathbf{v}(\mathbf{x},0) = \mathbf{v}_0(\mathbf{x}), \quad \frac{\partial \mathbf{v}}{\partial t}(\mathbf{x},0) = \mathbf{v}_1(\mathbf{x}), \quad \frac{\partial^2 \mathbf{v}}{\partial t^2}(\mathbf{x},0) = \mathbf{v}_2(\mathbf{x}), \quad \frac{\partial^3 \mathbf{v}}{\partial t^3}(\mathbf{x},0) = \mathbf{v}_3(\mathbf{x}).$$

Here $\|\cdot\|_{k,l}$ denotes the norm in $W^{k,\infty}([0,T']; H^l(\Omega))$; below we shall also use the notation $\|\cdot\|_{k,l,p}$ for the norm in $W^{k,p}([0,T']; H^l(\Omega))$. In $Z(M,T')$, we introduce the metric

$$d(\mathbf{v},\mathbf{w}) = \|\mathbf{v}-\mathbf{w}\|_{0,3} + \|\mathbf{v}-\mathbf{w}\|_{1,2} + \|\mathbf{v}-\mathbf{w}\|_{2,1} + \|\mathbf{v}-\mathbf{w}\|_{3,0}. \tag{18}$$

It is easy to see that $Z(M,T')$ with this metric is complete. We refer to [11] for a proof that it is not empty (provided M is large enough).

We note that (13) is of the same type as problem (16) considered in [11]. The quantity

$$\int (\gamma(|\mathbf{R}|^2) R_j R_l \delta_{ik} + 2\gamma'(|\mathbf{R}|^2) R_i R_j R_k R_l) \psi^m \, d\mathbf{R} \tag{19}$$

assumes the role of C_{ijkl} in [11]. We note that the symmetry condition

$$C_{ijkl} = C_{klij} \tag{20}$$

and the strong ellipticity condition

$$C_{ijkl}\zeta_i\zeta_k\eta_j\eta_l \geq \kappa|\zeta|^2|\eta|^2 \quad \forall\zeta,\eta \in \mathbb{R}^3, \tag{21}$$

with $\kappa > 0$ depending continuously on $\psi \in \{\psi \in L^1 \mid \psi \geq 0, \int \psi\, d\mathbf{R} = 1\}$, are satisfied, because we have assumed γ to be an increasing function of $|\mathbf{R}|^2$. Hence the results of [11] are applicable to (13), and most of the proof in the following section will be concerned with solving the diffusion equation (12).

4. Proof of the theorem. In order to show that Σ is a contraction in $Z(M,T')$, we shall have to choose M sufficiently large and T' sufficiently small. The estimates used in the proof will involve bounds of the form $K(M,T',a_1,a_2,\cdots)$, where the size of a_1, a_2 etc., can be kept within given bounds. It is important that the size of K for large M can be controlled by choosing T' small enough. This leads us to the following definition.

DEFINITION. A continuous function $K(M,T',a_1,a_2,\cdots) : \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \cdots \to \mathbb{R}^+$ is called controllable, if there are continuous positively valued functions $\tau(M,a_1,a_2,\cdots)$ and $\omega(a_1,a_2,\cdots)$ such that $K(M,T',a_1,a_2,\cdots) \leq \omega(a_1,a_2,\cdots)$ as long as $T' \leq \tau(M,a_1,a_2,\cdots)$.

We begin by stating the results of [11], which can be applied to problem (13). This problem is of the form

$$\begin{gathered}\rho\left[\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla)\right]^2 w_i = -\frac{\partial q}{\partial x_i} + C_{ijkl}(\mathbf{x},t)\frac{\partial^2 w_k}{\partial x_j \partial x_l} + h_i, \\ \operatorname{div}\mathbf{w} = 0, \quad \mathbf{w}|_{\partial\Omega} = \mathbf{0}, \quad \mathbf{w}(\mathbf{x},0) = \mathbf{w}_0(\mathbf{x}), \quad \dot{\mathbf{w}}(\mathbf{x},0) = \mathbf{w}_1(\mathbf{x}).\end{gathered} \tag{22}$$

LEMMA 1. *Let* $\mathbf{v}$, $\mathbf{h}$ *and* $\mathbf{C}$ *be given such that* $\operatorname{div}\mathbf{v} = 0$, $\mathbf{v}|_{\partial\Omega} = 0$ *and* $\mathbf{C}$ *satisfies the symmetry condition* (20) *and the strong ellipticity condition* (21) *with a positive lower bound* γ *for* κ *in* (21). *Moreover, assume that bounds of the following kind hold:*

$$\|\mathbf{v}\|_{0,4} + \|\mathbf{v}\|_{1,3} + \|\mathbf{v}\|_{2,2} + \|\mathbf{v}\|_{3,1} + \|\mathbf{v}\|_{4,0} \leq M, \tag{23a}$$

$$\|\mathbf{v}\|_{0,3} + \|\mathbf{v}\|_{1,2} + \|\mathbf{v}\|_{2,1} + \|\mathbf{v}\|_{3,0} \leq K, \tag{23b}$$

$$\|\mathbf{C}\|_{0,3} + \|\mathbf{C}\|_{1,2} + \|\mathbf{C}\|_{2,1} + \|\mathbf{C}\|_{3,0} \leq M, \tag{23c}$$

$$\|\mathbf{C}\|_{0,2} + \|\mathbf{C}\|_{1,1} + \|\mathbf{C}\|_{2,0} \leq K, \tag{23d}$$

$$\|\mathbf{h}\|_{0,2} + \|\mathbf{h}\|_{1,1} + \|\mathbf{h}\|_{2,0} \leq K, \tag{23e}$$

$$\left\|\frac{\partial\mathbf{h}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{h}\right\|_{0,2,1} + \left\|\frac{\partial\mathbf{h}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{h}\right\|_{1,1,1} + \left\|\frac{\partial\mathbf{h}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{h}\right\|_{2,0,1} \leq K. \tag{23f}$$

Finally, assume that $\mathbf{w}_0 \in H^4(\Omega)$, $\mathbf{w}_1 \in H^3(\Omega)$ *with norms bounded by* K, *that* $\operatorname{div} \mathbf{w}_0 = \operatorname{div} \mathbf{w}_1 = 0$, *and that* $\mathbf{w}_0$, $\mathbf{w}_1$, *and the initial values of* $\partial^2\mathbf{w}/\partial t^2$ *and* $\partial^3\mathbf{w}/\partial t^3$, *as determined by* (22), *are compatible with the boundary condition* $\mathbf{w}|_{\partial\Omega} = \mathbf{0}$. *Then* (22) *has a solution*

$$\mathbf{w} \in \bigcap_{k=0}^{4} C^k([0,T']; H^{4-k}(\Omega)),$$

and we have

$$\|\mathbf{w}\|_{0,4} + \|\mathbf{w}\|_{1,3} + \|\mathbf{w}\|_{2,2} + \|\mathbf{w}\|_{3,1} + \|\mathbf{w}\|_{4,0} \le \phi(M,T',K,\gamma), \tag{24}$$

where ϕ *is controllable.*

LEMMA 2. *Consider* (22) *and a second equation*

$$\begin{aligned} &\rho\left[\frac{\partial}{\partial t} + (\tilde{\mathbf{v}}\cdot\nabla)\right]^2 \tilde{w}_i = -\frac{\partial \tilde{q}}{\partial x_i} + \tilde{C}_{ijkl}(\mathbf{x},t)\frac{\partial^2 \tilde{w}_k}{\partial x_j \partial x_l} + \tilde{h}_i, \\ &\operatorname{div} \tilde{\mathbf{w}} = 0, \quad \tilde{\mathbf{w}}|_{\partial\Omega} = \mathbf{0}, \quad \tilde{\mathbf{w}}(\mathbf{x},0) = \mathbf{w}_0(\mathbf{x}), \quad \dot{\tilde{\mathbf{w}}}(\mathbf{x},0) = \mathbf{w}_1(\mathbf{x}). \end{aligned} \tag{25}$$

Assume that the assumptions of Lemma 1 *also hold for* (25) (*with the same constants* M, K *and* γ). *Moreover, assume that for* $t = 0$ *we have*

$$\mathbf{v} = \tilde{\mathbf{v}}, \quad \dot{\mathbf{v}} = \dot{\tilde{\mathbf{v}}}, \quad \ddot{\mathbf{v}} = \ddot{\tilde{\mathbf{v}}}, \quad \mathbf{C} = \tilde{\mathbf{C}}, \quad \dot{\mathbf{C}} = \dot{\tilde{\mathbf{C}}}, \quad \mathbf{h} = \tilde{\mathbf{h}}, \quad \dot{\mathbf{h}} = \dot{\tilde{\mathbf{h}}}. \tag{26}$$

Then we have an estimate of the form

$$\begin{aligned} &\|\mathbf{w}-\tilde{\mathbf{w}}\|_{0,3} + \|\mathbf{w}-\tilde{\mathbf{w}}\|_{1,2} + \|\mathbf{w}-\tilde{\mathbf{w}}\|_{2,1} + \|\mathbf{w}-\tilde{\mathbf{w}}\|_{3,0} \le \psi(M,T',K,\gamma) \\ &\quad\cdot\Big[\|\mathbf{v}-\tilde{\mathbf{v}}\|_{0,3} + \|\mathbf{v}-\tilde{\mathbf{v}}\|_{1,2} + \|\mathbf{v}-\tilde{\mathbf{v}}\|_{2,1} + \|\mathbf{v}-\tilde{\mathbf{v}}\|_{3,0} \\ &\qquad + \|\mathbf{C}-\tilde{\mathbf{C}}\|_{0,2} + \|\mathbf{C}-\tilde{\mathbf{C}}\|_{1,1} + \|\mathbf{C}-\tilde{\mathbf{C}}\|_{2,0} + \|\mathbf{h}-\tilde{\mathbf{h}}\|_{0,1} + \|\mathbf{h}-\tilde{\mathbf{h}}\|_{1,0} \\ &\qquad + \left\|\frac{\partial \mathbf{h}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{h} - \frac{\partial \tilde{\mathbf{h}}}{\partial t} - (\tilde{\mathbf{v}}\cdot\nabla)\tilde{\mathbf{h}}\right\|_{0,1,1} \\ &\qquad + \left\|\frac{\partial \mathbf{h}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{h} - \frac{\partial \tilde{\mathbf{h}}}{\partial t} - (\tilde{\mathbf{v}}\cdot\nabla)\tilde{\mathbf{h}}\right\|_{1,0,1}\Big]. \end{aligned} \tag{27}$$

The function $\psi(M,T',K,\gamma)$ *tends to zero as* $T' \to 0$.

We need to establish analogous results for (12). For given $\mathbf{v} \in Z(M,T')$, we have to find ψ satisfying

$$\begin{aligned} \left(\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla_{\mathbf{x}})\right)\psi &= \alpha\Delta_{\mathbf{R}}\psi + \operatorname{div}_{\mathbf{R}}(-(\nabla_{\mathbf{x}}\mathbf{v})\cdot\mathbf{R}\psi - \beta\mathbf{F}(\mathbf{R})\psi), \\ \psi(\mathbf{R},\mathbf{x},0) &= \psi_0(\mathbf{R},\mathbf{x}). \end{aligned} \tag{28}$$

The following lemma holds.

LEMMA 3. *Given* $\mathbf{v} \in Z(M,T')$, *there exists a unique solution of* (28) *which has the regularity*

$$\psi \in \bigcap_{k=0}^{3}\bigcap_{l=0}^{3-k} C^k([0,T']; H^l(\Omega; X_{8-2k-2l})). \tag{29}$$

Let $\|\cdot\|_{(n)}$ *be the norm in*

$$\bigcap_{k=0}^{3}\bigcap_{l=0}^{3-k} C^k([0,T'];H^l(\Omega;X_{n,8-2k-2l})). \tag{30}$$

For every n, *we have an estimate of the form*

$$\|\psi\|_{(n)} \leq C_n \cdot K(M,T',n), \tag{31}$$

where C_n *depends only on the initial data and* K *is controllable.*

We note that it is in general neither possible nor necessary to obtain a bound for $K(M,T',n)$ which is independent of n. By using (29) and (28), we find, moreover, that

$$\frac{\partial\psi}{\partial t} + (\mathbf{v}\cdot\nabla_{\mathbf{x}})\psi \in \bigcap_{k=0}^{3}\bigcap_{l=0}^{3-k} W^{k,\infty}([0,T'];H^l(\Omega;X_{6-2k-2l})). \tag{32}$$

In addition to (28), let us consider a second equation of the same form:

$$\begin{aligned}\Big(\frac{\partial}{\partial t} + (\tilde{\mathbf{v}}\cdot\nabla_{\mathbf{x}})\Big)\tilde{\psi} &= \alpha\Delta_{\mathbf{R}}\tilde{\psi} + \operatorname{div}_{\mathbf{R}}((-\nabla_{\mathbf{x}}\tilde{\mathbf{v}})\cdot\mathbf{R}\tilde{\psi} - \beta\mathbf{F}(\mathbf{R})\tilde{\psi}),\\ \tilde{\psi}(\mathbf{R},\mathbf{x},0) &= \psi_0(\mathbf{R},\mathbf{x}).\end{aligned} \tag{33}$$

The following result holds.

LEMMA 4. *Let* $\mathbf{v}$, $\tilde{\mathbf{v}} \in Z(M,T')$ *be given. Let* $\|\cdot\|_{[n]}$ *denote the norm in*

$$\bigcap_{k=0}^{2}\bigcap_{l=0}^{2-k} C^k([0,T'];H^l(\Omega;X_{6-2k-2l})). \tag{34}$$

Then for every n, *we have an estimate of the form*

$$\|\psi - \tilde{\psi}\|_{[n]} \leq K(M,T',n)d(\mathbf{v},\tilde{\mathbf{v}}). \tag{35}$$

Here $d(\mathbf{v},\tilde{\mathbf{v}})$ *is as defined in* (18). *The function* $K(M,T',n)$ *tends to zero as* $T' \to 0$.

Using the equations, we find, as a consequence, a similar estimate for

$$\Big(\frac{\partial}{\partial t} + (\mathbf{v}\cdot\nabla_{\mathbf{x}})\Big)\psi - \Big(\frac{\partial}{\partial t} + (\tilde{\mathbf{v}}\cdot\nabla_{\mathbf{x}})\Big)\tilde{\psi} \tag{36}$$

in the space

$$\bigcap_{k=0}^{2}\bigcap_{l=0}^{2-k} C^k([0,T'];H^l(\Omega;X_{4-2k-2l})). \tag{37}$$

By combining Lemmas 1 through 4, it follows easily that Σ is a contraction in $Z(M,T')$ if M is chosen sufficiently large and T' is chosen sufficiently small. The theorem follows immediately. The rest of the paper will therefore be concerned with the proofs of Lemmas 3 and 4.

We introduce Lagrangian coordinates, which we shall denote by $\mathbf{X}$. Let $\mathbf{x}(\mathbf{X},t)$ denote the Eulerian coordinate corresponding to $\mathbf{X}$, i.e., the solution of the problem

$$\frac{\partial}{\partial t}\mathbf{x}(\mathbf{X},t) = \mathbf{v}(\mathbf{x}(\mathbf{X},t),t), \qquad \mathbf{x}(\mathbf{X},0) = \mathbf{X}. \tag{38}$$

Let $\phi(\mathbf{R},\mathbf{X},t) = \psi(\mathbf{R},\mathbf{x}(\mathbf{X},t),t)$. Equation (28) can then be written in the alternative form

$$\begin{aligned} &\frac{\partial}{\partial t}\phi(\mathbf{R},\mathbf{X},t) = \alpha\Delta_{\mathbf{R}}\phi + \operatorname{div}_{\mathbf{R}}(-(\nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x}(\mathbf{X},t),t))\cdot\mathbf{R}\phi - \beta\mathbf{F}(\mathbf{R})\phi), \\ &\phi(\mathbf{R},\mathbf{X},0) = \psi_0(\mathbf{R},\mathbf{X}). \end{aligned} \tag{39}$$

This is a parabolic equation for ϕ, in which the only derivatives are with respect to $\mathbf{R}$ and t, while $\mathbf{X}$ only appears as a parameter in the coefficients and initial conditions. We note that the transformation $(\mathbf{X},t) \to (\mathbf{x}(\mathbf{X},t),t)$ is smooth enough so that if ψ has the regularity claimed by Lemma 3, then so does ϕ and vice versa.

The fact that solutions of (39) with the regularity claimed by Lemma 3 are unique follows from a straightforward application of the maximum principle. It also follows from the maximum principle that positivity is preserved, and by integrating both sides of (39) we find that $\int \phi(\mathbf{R},\mathbf{X},t)\,d\mathbf{R} = \int \psi(\mathbf{R},\mathbf{x},t)\,d\mathbf{R} = 1$ for all t. To show the existence of solutions, we cannot use standard results available in the literature because the coefficients of the equation are unbounded. To get around this difficulty, we use a sequence of approximating problems with bounded coefficients, for which we derive uniform estimates. The cases of infinitely and, respectively, finitely extensible dumbbells have to be treated somewhat differently, and we shall first deal with infinitely extensible dumbbells.

Let $\chi(\mathbf{R})$ be a C^∞-function such that $\chi(\mathbf{0}) = 1$, χ is a monotone decreasing function of $|\mathbf{R}|$, and $\chi(\mathbf{R}) = |\mathbf{R}|^{-\nu}$ for large $|\mathbf{R}|$, where ν is a sufficiently large number. For $N \in \mathbb{N}$, let $\chi_N(\mathbf{R}) = \chi(\mathbf{R}/N)$. We now replace (39) by the approximate problem

$$\begin{aligned} &\frac{\partial\phi_N}{\partial t} = \alpha\Delta_{\mathbf{R}}\phi_N + \operatorname{div}_{\mathbf{R}}\Big[\chi_N(\mathbf{R})\big(-(\nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x}(\mathbf{X},t),t))\cdot\mathbf{R}\phi_N - \beta\mathbf{F}(\mathbf{R})\phi_N\big)\Big], \\ &\phi_N(\mathbf{R},\mathbf{X},0) = \psi_0(\mathbf{R},\mathbf{X}). \end{aligned} \tag{40}$$

We note that equation (40) has bounded coefficients. Hence the existence of solutions and their decay as $|\mathbf{R}| \to \infty$ are easily established. We now multiply (40) by $|\mathbf{R}|^{2n}$ and integrate. This yields after an integration by parts

$$\begin{aligned} &\frac{\partial}{\partial t}\int |\mathbf{R}|^{2n}\phi_N(\mathbf{R},\mathbf{X},t)\,d\mathbf{R} \\ &\quad = \alpha(4n^2+2n)\int |\mathbf{R}|^{2n-2}\phi_N(\mathbf{R},\mathbf{X},t)\,d\mathbf{R} \\ &\qquad + 2n\int |\mathbf{R}|^{2n-2}\chi_N(\mathbf{R})\Big[\mathbf{R}\cdot(\nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x}(\mathbf{X},t),t))\cdot\mathbf{R} \\ &\qquad\qquad + \beta\mathbf{R}\cdot\mathbf{F}(\mathbf{R})\Big]\phi_N(\mathbf{R},\mathbf{X},t)\,d\mathbf{R}. \end{aligned} \tag{41}$$

The term $\mathbf{R}\cdot\mathbf{F}(\mathbf{R})$ is negative, and all other terms on the right-hand side of (41) can be estimated by a constant (depending on n but not on N) times $\int(1+|\mathbf{R}|^{2n})$ $\phi_N(\mathbf{R},\mathbf{X},t)\,d\mathbf{R}$. Hence (41) yields a bound of the form

$$\int |\mathbf{R}|^{2n}\phi_N(\mathbf{R},\mathbf{X},t) \le K(n,Mt)\int(1+|\mathbf{R}|^{2n})\psi_0(\mathbf{R},\mathbf{X})\,d\mathbf{R}. \tag{42}$$

We next derive estimates for derivatives with respect to $\mathbf{R}$. We differentiate (40) with respect to R_i and obtain

$$\begin{aligned}\frac{\partial^2\phi_N}{\partial R_i\partial t} &= \alpha\Delta_{\mathbf{R}}\frac{\partial\phi_N}{\partial R_i} + \operatorname{div}_{\mathbf{R}}\Big[\chi_N(\mathbf{R})\big(-(\nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x}(\mathbf{X},t),t))\cdot\mathbf{R}-\beta\mathbf{F}(\mathbf{R})\big)\frac{\partial\phi_N}{\partial R_i}\Big]\\ &\quad + \operatorname{div}_{\mathbf{R}}\Big[\frac{\partial}{\partial R_i}\Big(\chi_N(\mathbf{R})(-(\nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x}(\mathbf{X},t),t))\cdot\mathbf{R}-\beta\mathbf{F}(\mathbf{R}))\Big)\phi_N\Big],\\ \frac{\partial\phi_N}{\partial R_i}(\mathbf{R},\mathbf{X},0) &= \frac{\partial\psi_0}{\partial R_i}(\mathbf{R},\mathbf{X}).\end{aligned}\tag{43}$$

The last term on the right-hand side of this equation can be further transformed as follows:

$$\begin{aligned}&\operatorname{div}_{\mathbf{R}}\Big[\frac{\partial}{\partial R_i}\Big(\chi_N(\mathbf{R})(-(\nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x}(\mathbf{X},t),t))\cdot\mathbf{R}-\beta\mathbf{F}(\mathbf{R}))\Big)\phi_N\Big]\\ &\quad=\phi_N\operatorname{div}_{\mathbf{R}}\Big[\frac{\partial}{\partial R_i}\Big(\chi_N(\mathbf{R})(-(\nabla_{\mathbf{x}}\mathbf{v})\cdot\mathbf{R}-\beta\mathbf{F}(\mathbf{R}))\Big)\Big]\\ &\qquad+\frac{\partial\phi_N}{\partial R_j}\Big[\frac{\partial\chi_N}{\partial R_i}\Big(-\frac{\partial v_j}{\partial x_k}R_k-\beta F_j(\mathbf{R})\Big)+\chi_N(\mathbf{R})\Big(-\frac{\partial v_j}{\partial x_i}-\beta\frac{\partial F_j}{\partial R_i}\Big)\Big].\end{aligned}\tag{44}$$

Since $\partial\phi_N/\partial R_i$ is not positive, we cannot proceed in exact analogy to (41) above. Instead, we first decompose $\partial\phi_N/\partial R_i=\pi_i^+-\pi_i^-$, where π_i^+ and π_i^- are solutions to the problems,

$$\begin{aligned}\frac{\partial\pi_i^+}{\partial t} &= \alpha\Delta_{\mathbf{R}}\pi_i^+ + \operatorname{div}_{\mathbf{R}}\Big[\chi_N(\mathbf{R})\big(-(\nabla_{\mathbf{x}}\mathbf{v})\cdot\mathbf{R}\pi_i^+-\beta\mathbf{F}(\mathbf{R})\pi_i^+\big)\Big]+s_i^+,\\ \pi_i^+(\mathbf{R},\mathbf{X},0) &= \max\Big(\frac{\partial\psi_0}{\partial R_i}(\mathbf{R},\mathbf{X}),0\Big),\end{aligned}\tag{45}$$

and, respectively,

$$\begin{aligned}\frac{\partial\pi_i^-}{\partial t} &= \alpha\Delta_{\mathbf{R}}\pi_i^- + \operatorname{div}_{\mathbf{R}}\Big[\chi_N(\mathbf{R})\big(-(\nabla_{\mathbf{x}}\mathbf{v})\cdot\mathbf{R}\pi_i^--\beta\mathbf{F}(\mathbf{R})\pi_i^-\big)\Big]-s_i^-,\\ \pi_i^-(\mathbf{R},\mathbf{X},0) &= \max\Big(-\frac{\partial\psi_0}{\partial R_i}(\mathbf{R},\mathbf{X}),0\Big).\end{aligned}\tag{46}$$

Here s_i^+ and s_i^- denote the positive and negative parts of the expression in (44). Since π_i^+ and π_i^- are positive, we can now proceed as above. In place of (41), we obtain

$$\begin{aligned}&\frac{\partial}{\partial t}\int|\mathbf{R}|^{2n}\pi_i^+(\mathbf{R},\mathbf{X},t)\,d\mathbf{R}\\ &=\alpha(4n^2+2n)\int|\mathbf{R}|^{2n-2}\pi_i^+(\mathbf{R},\mathbf{X},t)\,d\mathbf{R}+2n\int|\mathbf{R}|^{2n-2}\chi_N(\mathbf{R})\\ &\quad\cdot\Big[\mathbf{R}\cdot(\nabla_{\mathbf{x}}\mathbf{v}(\mathbf{x}(\mathbf{X},t),t))\cdot\mathbf{R}+\beta\mathbf{R}\cdot\mathbf{F}(\mathbf{R})\Big]\pi_i^+(\mathbf{R},\mathbf{X},t)\,d\mathbf{R}+\int|\mathbf{R}|^{2n}s_i^+\,d\mathbf{R},\end{aligned}\tag{47}$$

and an analogous equation for π_i^-. (A minor problem in deriving (47) arises from the fact that $\Delta\pi_i^+$ does not necessarily exist at $t=0$, but we may justify the argument by

approximating the initial value of π_i^+ by a sequence of smoother functions and then passing to the limit.) Using (44), we find that

$$\begin{aligned}
&\int |\mathbf{R}|^{2n} s_i^+ \, d\mathbf{R} \\
&\qquad \le \int |\mathbf{R}|^{2n} \phi_N \left| \operatorname{div}_{\mathbf{R}} \left[\frac{\partial}{\partial R_i} \Big(\chi_N(\mathbf{R})(-(\nabla_{\mathbf{x}} \mathbf{v}) \cdot \mathbf{R} - \beta \mathbf{F}(\mathbf{R})) \Big) \right] \right| d\mathbf{R} \\
&\qquad + \int |\mathbf{R}|^{2n} (\pi_j^+ + \pi_j^-) \left[\left| \frac{\partial \chi_N}{\partial R_i} \right| \left(\left| \frac{\partial v_j}{\partial x_k} R_k \right| + \beta |F_j(\mathbf{R})| \right) \right. \\
&\qquad \left. + \chi_N(\mathbf{R}) \left(\left| \frac{\partial v_j}{\partial x_i} \right| + \beta \left| \frac{\partial F_j}{\partial R_i} \right| \right) \right] d\mathbf{R}.
\end{aligned} \tag{48}$$

The first integral on the right-hand side of (48) involves only ϕ_N and no derivatives of ϕ_N; it can hence be estimated using (42) above. The integrand in the second integral can be estimated for large $|\mathbf{R}|$ by a constant times $\chi_N(\mathbf{R})(M|\mathbf{R}|^{2n} + |\mathbf{R}|^{2n-1}|\mathbf{F}(\mathbf{R})|) \sum_j (\pi_j^+ + \pi_j^-)$. From (47), we therefore obtain an estimate of the form

$$\begin{aligned}
&\frac{\partial}{\partial t} \int |\mathbf{R}|^{2n} \pi_i^+(\mathbf{R}, \mathbf{X}, t) \, d\mathbf{R} \\
&\qquad \le K_1(n, M) \int (1 + |\mathbf{R}|^{2n}) \pi_i^+(\mathbf{R}, \mathbf{X}, t) \, d\mathbf{R} + K_2(n, M) \\
&\qquad + 2n \int |\mathbf{R}|^{2n-2} \chi_N(\mathbf{R}) \beta \mathbf{R} \cdot \mathbf{F}(\mathbf{R}) \pi_i^+(\mathbf{R}, \mathbf{X}, t) \, d\mathbf{R} \\
&\qquad + K_3 \int (1 + |\mathbf{R}|^{2n-1} \chi_N(\mathbf{R}) |\mathbf{F}(\mathbf{R})| + M \chi_N(\mathbf{R}) |\mathbf{R}|^{2n}) \\
&\qquad \cdot \sum_j (\pi_j^+ + \pi_j^-)(\mathbf{R}, \mathbf{X}, t) \, d\mathbf{R}.
\end{aligned} \tag{49}$$

Note that K_3 does not depend on n, but that the previous term in (49), which is negative, has a factor n in front. By summing the inequalities corresponding to (49) over all values of i and all choices of $+$ or $-$, we can therefore get bounds for $\int |\mathbf{R}|^{2n} \pi_i^+ \, d\mathbf{R}$. As a consequence, we have found estimates for the first derivatives of ϕ_N. Bounds for higher derivatives are obtained in an analogous fashion after differentiating (40) with respect to the components of $\mathbf{R}$. The calculation involves higher derivatives of ϕ_N than are known to exist at $t = 0$, but the argument can be justified by approximating the initial datum by smoother functions and passing to the limit. Proceeding in this fashion, we find that

$$\phi_N \in L^\infty([0, T']; L^2(\Omega; X_8)). \tag{50}$$

Derivatives of ϕ_N with respect to $\mathbf{X}$ can be estimated by taking difference quotients in (40), estimating the difference quotients of ϕ_N, and then taking limits. Problems near the boundary of Ω are avoided by first extending $\mathbf{v}$ and ψ_0 so that they are defined for $\mathbf{X}$ outside Ω. This can be done in such a fashion that the extended functions have the same level of regularity as the original ones; for example, the method of reflection across the boundary can be used to obtain such extensions (cf. [8], p. 38).

Mixed derivatives with respect to $\mathbf{R}$ and $\mathbf{X}$ can be estimated by taking difference quotients with respect to $\mathbf{X}$ after first differentiating with respect to $\mathbf{R}$. The calculations are tedious but straightforward and we omit them. We find in this way that

$$\phi_N \in \bigcap_{l=0}^{3} L^\infty([0,T']; H^l(\Omega; X_{8-2l})). \tag{51}$$

Finally, temporal derivatives of ϕ_N can be estimated by using the equation itself. This yields

$$\phi_N \in \bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} W^{k,\infty}([0,T']; H^l(\Omega; X_{8-2k-2l})). \tag{52}$$

All these estimates are uniform in N.

Using the uniform bounds on the ϕ_N, we can extract a subsequence which converges, in the sense of distributions, to a solution ϕ of (39). Unfortunately, it is not a priori clear that

$$\phi \in \bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} W^{k,\infty}([0,T']; H^l(\Omega; X_{8-2k-2l})). \tag{53}$$

The difficulty is that the spaces X_k are based on L^1-type norms, and the limit of a distributionally convergent bounded sequence in L^1 may not be an L^1-function, but a singular measure. To show (53), we need a separate argument. We approximate the initial datum ψ_0 by a sequence ψ_0^M such that ψ_0^M has the additional regularity

$$\psi_0^M \in \bigcap_{k=0}^{4} H^k(\Omega; X_{10-2k}), \tag{54}$$

and $\psi_0^M \to \psi_0$ in the topology of

$$\bigcap_{k=0}^{4} H^k(\Omega; X_{8-2k}). \tag{55}$$

Let ϕ_N^M be the solution of (40) which is obtained when the initial datum is replaced by ψ_0^M. Then the same argument as above yields

$$\phi_N^M \in \bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} W^{k,\infty}([0,T']; H^l(\Omega; X_{10-2k-2l})); \tag{56}$$

moreover, in the topology of

$$\bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} W^{k,\infty}([0,T']; H^l(\Omega; X_{8-2k-2l})), \tag{57}$$

we have $\phi_N^M \to \phi_N$ uniformly in N. From (56) and (40), we conclude that

$$\frac{\partial \phi_N^M}{\partial t} \in \bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} W^{k,\infty}([0,T']; H^l(\Omega; X_{8-2k-2l})). \tag{58}$$

From (56) and (58) it follows that

$$\phi_N^M \in \bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} C^k([0,T']; H^l(\Omega; X_{8-2k-2l})). \tag{59}$$

For fixed M, the bounds for ϕ_N^M are again uniform in N. Let now Π_n be a reflexive Banach space such that

$$\begin{aligned} &\left[\bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} W^{k,\infty}([0,T'];H^l(\Omega; X_{n,10-2k-2l}))\right] \\ &\qquad \cap \left[\bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} W^{k+1,\infty}([0,T']; H^l(\Omega; X_{n,8-2k-2l}))\right] \\ &\qquad \subset \Pi_n \subset \bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} C^k([0,T']; H^l(\Omega; X_{n-1,8-2k-2l})). \end{aligned} \tag{60}$$

Such a space Π_n can be constructed using appropriate Sobolev norms. Let $\Pi = \bigcap_{n\in\mathbb{N}} \Pi_n$. Then, for every M, we have a uniform bound for ϕ_N^M in the norm of each Π_n, hence we can extract a subsequence which converges weakly in each Π_n to an element ϕ^M of Π. Since, for $M \to \infty$, ϕ_N^M converges to ϕ_N uniformly in N, we find that the sequence ϕ^M also converges in

$$\bigcap_{k=0}^{3} \bigcap_{l=0}^{3-k} C^k([0,T']; H^l(\Omega; X_{8-2k-2l})). \tag{61}$$

By taking $\phi = \lim_{M\to\infty} \phi^M$, we obtain a solution with the desired regularity. This concludes the proof of Lemma 3 for the case of the infinitely extensible dumbbell.

The proof of Lemma 4 is based on the same type of estimates, applied to the function $\psi - \tilde{\psi}$. We omit the details.

For the case of finitely extensible dumbbells, we use approximation by infinitely extensible dumbbells. We extend the initial datum ψ_0 by zero for $|\mathbf{R}| > R_0$, and we define an approximation γ_ϵ to the spring constant γ as follows. Let χ be a C^∞-function $[0,1] \to [0,1]$ such that $\chi(s) = 1$ in a neighborhood of 0 and $\chi(s) = 0$ in a neighborhood of 1. Let $\chi_\epsilon(\mathbf{R}) = \chi(\frac{2}{\epsilon}(|\mathbf{R}| - R_0 + \epsilon))$. Then we define

$$\gamma_\epsilon(|\mathbf{R}|^2) = \begin{cases} \gamma(|\mathbf{R}|^2) & \text{if } |\mathbf{R}| \le R_0 - \epsilon, \\ \gamma((R_0 - \frac{\epsilon}{2})^2) & \text{if } |\mathbf{R}| \ge R_0 - \frac{\epsilon}{2}, \\ \chi_\epsilon(\mathbf{R})\gamma(|\mathbf{R}|^2) + (1 - \chi_\epsilon(\mathbf{R}))\gamma((R_0 - \frac{\epsilon}{2})^2) & \text{if } R_0 - \epsilon < |\mathbf{R}| \\ & \qquad < R_0 - \frac{\epsilon}{2}. \end{cases} \tag{62}$$

We now consider the problem with spring constant γ_ϵ, which is infinitely extensible and satisfies the assumptions which we required for the infinitely extensible case. We now define

$$w_\epsilon(\mathbf{R}) = \begin{cases} \frac{1}{R_0^2 - |\mathbf{R}|^2} & \text{if } |\mathbf{R}| \le R_0 - \epsilon, \\ \frac{1}{\epsilon(2R_0 - \epsilon)} + \frac{2(R_0-\epsilon)}{\epsilon^2(2R_0-\epsilon)^2}(|\mathbf{R}| - R_0 + \epsilon) & \text{if } |\mathbf{R}| > R_0 - \epsilon. \end{cases} \tag{63}$$

The function w_ϵ and its gradient are continuous at $|\mathbf{R}| = R_0$, and w_ϵ has linear growth at infinity. We now repeat the estimates above, but instead of multiplying the equation

by powers of $|\mathbf{R}|$, we multiply by powers of $w_\epsilon(|\mathbf{R}|)$. The resulting estimates turn out to be uniform in ϵ as $\epsilon \to 0$, and Lemma 3 follows by passing to the limit.

REFERENCES

[1] R. B. BIRD, O. HASSAGER, R. C. ARMSTRONG, AND C. F. CURTISS, *Dynamics of Polymeric Liquids, Vol. 2: Kinetic Theory*, John Wiley, New York, 1977 and 1987.

[2] C. CHEN AND W. VON WAHL, *Das Rand-Anfangswertproblem für quasilineare Wellengleichungen in Sobolevräumen niedriger Ordnung*, J. Reine Angew. Math., 337 (1982), pp. 77–112.

[3] C. M. DAFERMOS AND W. J. HRUSA, *Energy methods for quasilinear hyperbolic initial-boundary value problems. Applications to elastodynamics*, Arch. Rational Mech. Anal., 87 (1985), pp. 267–292.

[4] E. G. EBIN AND R. A. SAXTON, *The initial value problem for elastodynamics of incompressible bodies*, Arch. Rational Mech. Anal., 94 (1986), pp. 15–38.

[5] W. J. HRUSA AND M. RENARDY, *An existence theorem for the Dirichlet problem in the elastodynamics of incompressible materials*, Arch. Rational Mech. Anal., 102 (1988), pp. 95–117; Corrigendum: Arch. Rational Mech. Anal., 110 (1990), pp. 373–375.

[6] T. J. R. HUGHES, T. KATO, AND J. E. MARSDEN, *Well-posed quasi-linear second-order hyperbolic systems with applications to nonlinear elastodynamics and general relativity*, Arch. Rational Mech. Anal., 63 (1976), pp. 273–284.

[7] T. KATO, *Linear and quasi-linear equations of hyperbolic type*, in Hyperbolicity, G. da Prato and G. Geymonat, eds., Centro Internazionale Matematico Estivo, II ciclo, Cortona, 1976, pp. 125–191.

[8] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications* I, Springer-Verlag, Berlin, New York, 1972.

[9] M. RENARDY, W. J. HRUSA, AND J. A. NOHEL, *Mathematical Problems in Viscoelasticity*, John Wiley, New York, and Longman Press, London, 1987.

[10] M. RENARDY, *Local existence theorems for the first and second initial-boundary value problems for a weakly non-Newtonian fluid*, Arch. Rational Mech. Anal., 83 (1983), pp. 229–244.

[11] ———, *Local existence of solutions of the Dirichlet initial-boundary value problem for incompressible hypoelastic materials*, SIAM J. Math. Anal., 21 (1990), pp. 1369–1385.

[12] S. SCHOCHET, *The incompressible limit in nonlinear elasticity*, Comm. Math. Phys., 102 (1985), pp. 207–215.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 328-343, March 1991

002

THE EFFECT OF TEMPERATURE DEPENDENT VISCOSITY ON SHEAR FLOW OF INCOMPRESSIBLE FLUIDS*

M. BERTSCH†, L. A. PELETIER‡, AND S. M. VERDUYN LUNEL§

Abstract. This paper studies the effect of temperature dependence of the viscosity on the stability of classical Couette flow between two parallel plates. When the viscosity increases with temperature, it is well known that the flow becomes increasingly stable. With decreasing viscosity, however, if the dependence is sufficiently strong, the flow is found to become unstable.

Key words. fluid dynamics, plane Couette flow, temperature dependent viscosity, reaction-diffusion equations, invariant sets

AMS(MOS) subject classifications. 35B30, 35K57, 76E30

1. Introduction. In this paper we consider a rectilinear flow of an incompressible Newtonian viscous fluid, and assume that the viscosity is temperature dependent. In particular, we are interested in situations where the viscosity decreases with increasing temperature.

In general, the dissipation caused by viscosity has a stabilizing effect. However, when the heat generated by this mechanism raises the temperature, thus causing the viscosity to decrease, the flow may nevertheless become unstable. It is the object of this note to investigate if, and when, this possibility might arise. We do this for a simple problem of an adiabatic rectilinear shearing flow between parallel plates, one moving with respect to another, at constant distance from one another.

Choosing Cartesian coordinates, with the x-axis perpendicular to the plates, we choose units such that the plates are situated in the planes $x=0$ and $x=1$. We assume that the plate at $x=0$ is at rest and that the plate at $x=1$ moves with the constant velocity V in the direction of the positive y-axis. Between the plates the flow is parallel to the plates and uniform in the y- and the z-direction. Thus it is completely described by the velocity $v(x, t)$ and the temperature $\theta(x, t)$.

If we normalize units so that the density and the specific heat both become unity, conservation of momentum and energy yield the equations

$$v_t = \sigma_x, \qquad 0<x<1, \quad t>0, \tag{1.1}$$

$$\theta_t = \sigma v_x, \qquad 0<x<1, \quad t>0. \tag{1.2}$$

Here σ denotes the shear stress. We shall assume that it is related to the velocity gradient through the linear expression

$$\sigma = \mu(\theta) v_x, \tag{1.3}$$

where μ denotes the (temperature-dependent) viscosity.

At the plates we have

$$v(0, t) = 0, \qquad v(1, t) = V \tag{1.4}$$

and at $t=0$ we prescribe the velocity and the temperature

$$v(x, 0) = v_0(x), \quad \theta(x, 0) = \theta_0(x), \quad 0<x<1. \tag{1.5}$$

* Received by the editors December 1, 1989; accepted for publication March 9, 1990.

† Dipartimento di Matematica, Università di Torino, Via Principe Amedeo 8, 10123 Torino, Italy.

‡ Mathematical Institute, Leiden University, P.O. Box 9512, 2300 RA Leiden, the Netherlands.

§ Center for Dynamical Systems and Non-Linear Studies, Georgia Institute of Technology, Atlanta, Georgia 30332.

Here v_0 and θ_0 are given functions such that $v_0 > 0$ when $0 < x \leqq 1$ and $\theta_0 > 0$ when $0 \leqq x \leqq 1$, and v_0 is taken compatible with the boundary conditions (1.4), that is, $v_0(0) = 0$ and $v_0(1) = V$.

Note that the boundary conditions (1.4) imply the following *compatibility condition* for σ and θ:

$$\int_0^1 \frac{\sigma(x, t)}{\mu(\theta(x, t))}\, dx = V. \tag{1.6}$$

If $v_0(x) = Vx$ and $\theta_0(x) = a$, where a is a positive number, then the solution of (1.1)-(1.5) can be written down explicitly:

$$v(x, t) = Vx, \qquad \theta(x, t) = h(t), \tag{1.7}$$

where $h(t)$ is determined by

$$\int_a^{h(t)} \frac{ds}{\mu(s)} = V^2 t. \tag{1.8}$$

Since $v_x(x, t) \equiv V$ for this solution, it describes a *uniform* shearing flow.

Problem (1.1)-(1.5) was first studied in some detail by Dafermos and Hsiao [DH] and subsequently by Tzavaras [T1]-[T3]. These authors have shown that if $\mu(\theta)$ tends monotonically to some *finite positive constant* as $\theta \to \infty$, and either μ^2 is concave or μ is convex, then for sufficiently smooth initial data, problem (1.1)-(1.5) has a unique solution, which converges to the uniform shearing flow as $t \to \infty$.

If, however, $\mu(\theta)$ tends to zero as $\theta \to \infty$, the situation becomes much more delicate [DH], [T1]. Thus, if μ is taken to be

$$\mu(\theta) = \theta^{-\alpha}, \qquad \alpha > 0, \tag{1.9}$$

then global existence is still ensured provided $\alpha < 1$, i.e., provided μ does not decrease too rapidly, and the flow converges to the uniform shearing flow as $t \to \infty$ [DH], [T1]. However, when $\alpha \geqq 1$, no such results are known. It is the object of this paper to explore in particular this parameter range.

It is natural to distinguish three cases:

(I) $\alpha \in (0, 1)$,

(II) $\alpha = 1$,

(III) $\alpha > 1$.

Assuming throughout that $v_0 \in W^{2,2}(0, 1)$ and $\theta_0 \in W^{1,2}(0, 1)$ we shall prove the following theorems.

THEOREM A. *Suppose that* $0 < \alpha < 1$. *Then* (1.1)-(1.5) *has a unique global solution* $(v(x, t), \theta(x, t))$ *that converges asymptotically to the uniform shearing flow*:

$$v_x(x, t) = V + O(t^{-(1-\alpha)/(1+\alpha)}) \quad \text{as } t \to \infty,$$

$$\theta(x, t) = \{(1+\alpha) V^2 t\}^{1/(1+\alpha)} [1 + O(t^{-(1-\alpha)/(1+\alpha)})] \quad \text{as } t \to \infty.$$

This theorem is not new. It has been proved in [DH] and [T1] by means of energy methods. Here we use maximum principle arguments to identify appropriate invariant regions, in order to establish existence and asymptotic behaviour. This method has recently also been used for this purpose by Tzavaras [T3].

THEOREM B. *Suppose that* $\alpha = 1$. *Then* (1.1)-(1.5) *has a unique global solution* $(v(x, t), \theta(x, t))$ *which behaves asymptotically as*

$$\lim_{t\to\infty} v_x(x, t) = \frac{V}{\sqrt{2}}\theta^*(x),$$

$$\lim_{t\to\infty} t^{-1/2}\theta(x, t) = \theta^*(x),$$

where θ^* *is a positive function that depends on the initial data. In addition*

$$\sigma(x, t) = \frac{V}{\sqrt{2t}}[1 + O(t^{-1})] \quad \text{as } t\to\infty$$

for every $x \in [0, 1]$.

Remark. In view of the compatibility condition (1.6), we can say about θ^* that

$$\int_0^1 \theta^*(x)\, dx = V\sqrt{2}.$$

For $\alpha > 1$ we will show that the situation becomes completely different. Specifically, we will prove the following result.

THEOREM C. *Suppose* $\alpha > 1$. *Then the uniform shearing flow is unstable.*

The proof is based on the study of the solutions emanating from a particular family of initial functions (v_0, θ_0). For such solutions we prove the following dichotomy: *Either* there exists a point $x_0 \in [0, 1]$ and a finite time T such that

$$\theta(x_0, t)\to\infty \quad \text{as } t\uparrow T,$$

or

$$v(x, t)\to V \quad \text{as } t\to\infty$$

on a set $[0, 1]\setminus\mathscr{E}$, where the exceptional set $\mathscr{E}$ can be made arbitrary small by a suitable choice of initial data.

To prove these results it will be more convenient to use σ instead of v as a dependent variable. By (1.3) we have, when we use (1.9),

$$v_x = \theta^\alpha\sigma, \tag{1.10}$$

and so, when we differentiate (1.1) with respect to x and use (1.2) we obtain after some rearrangement

$$\sigma_t = -\alpha\theta^{\alpha-1}\sigma^3 + \theta^{-\alpha}\sigma_{xx} \quad \text{in } Q, \tag{1.11}$$

$$\theta_t = \theta^\alpha\sigma^2 \quad \text{in } Q, \tag{1.12}$$

where $Q = (0, 1)\times\mathbf{R}^+$. The boundary conditions become in terms of σ, in view of (1.1):

$$\sigma_x(0, t) = 0, \quad \sigma_x(1, t) = 0, \quad t > 0, \tag{1.13}$$

and the initial conditions,

$$\sigma(x, 0) = \sigma_0(x), \quad \theta(x, 0) = \theta_0(x), \quad x \in [0, 1], \tag{1.14}$$

where $\sigma_0 = \theta_0^{-\alpha}v_{0x}$. The remainder of the paper will be devoted to the study of system (1.11)-(1.14) for the three ranges of the parameter α.

We note that by a simple scaling of the variables it is possible to set the velocity of the top plate V equal to unity. Specifically, if $(v(x, t), \theta(x, t))$ is a solution of (1.1)-(1.5) with μ given by (1.9), then $(\hat{v}(x, \hat{t}), \hat{\theta}(x, \hat{t}))$ is a solution of (1.1)-(1.5) with $V=1$ if we set

$$\hat{t} = V^{-2\alpha} t, \quad \hat{v}(x, \hat{t}) = \frac{1}{V} v(x, t), \quad \hat{\theta}(x, \hat{t}) = V^{-2}\theta(x, t) \tag{1.15}$$

and scale the initial data appropriately. Thus setting $V=1$ in §§ 2 and 3 involves no loss of generality. In § 4, where we construct specific initial data, it is more convenient to choose V appropriately.

2. The case $0 < \alpha < 1$. For $\alpha \in (0, 1)$ it is convenient to replace θ by a new variable

$$q = \frac{1}{1-\alpha} \theta^{1-\alpha}. \tag{2.1}$$

Then system (1.11)-(1.14) becomes

$$\sigma_t = -\frac{\alpha}{1-\alpha} \frac{\sigma^3}{q} + \{(1-\alpha)q\}^{-\alpha/(1-\alpha)} \sigma_{xx} \quad \text{in } Q, \tag{2.2}$$

$$q_t = \sigma^2 \quad \text{in } Q, \tag{2.3}$$

$$\sigma_x(0, t) = 0, \quad \sigma_x(1, t) = 0, \quad t > 0, \tag{2.4}$$

$$\sigma(x, 0) = \sigma_0(x), \quad q(x, 0) = q_0(x), \quad x \in [0, 1], \tag{2.5}$$

where $q_0 = \theta_0^{1-\alpha}/(1-\alpha)$.

Seeking solutions of (2.2)-(2.3) which are independent of x, we arrive at the following system of ordinary differential equations:

$$y' = -\frac{1+\alpha}{1-\alpha} y^2, \qquad q' = qy, \tag{2.6}$$

where primes denote differentiation with respect to t and we have set

$$y = \frac{\sigma^2}{q}. \tag{2.7}$$

This system can be solved explicitly. It yields

$$y(t) = \frac{1-\alpha}{1+\alpha} \frac{1}{t+A}, \qquad q(t) = B(t+A)^{(1-\alpha)/(1+\alpha)}, \tag{2.8}$$

where A and B are positive constants. Hence

$$\sigma(t) = \left(\frac{1-\alpha}{1+\alpha} B\right)^{1/2} (t+A)^{-\alpha/(1+\alpha)}. \tag{2.9}$$

Plainly (2.8)-(2.9) is a solution of (2.2)-(2.5) for suitably chosen initial values of σ and q.

The solution (2.8)-(2.9) suggests new dependent variables

$$\tilde{r}(x, t) = (t+1)^{-(1-\alpha)/(1+\alpha)} q(x, t), \tag{2.10}$$

$$\tilde{s}(x, t) = (t+1)^{\alpha/(1+\alpha)} \sigma(x, t). \tag{2.11}$$

If we rewrite (2.2)-(2.5) in these variables, define the new time

$$\tau = \log(t+1) \tag{2.12}$$

and set

$$r(x, \tau) = \tilde{r}(x, t), \qquad s(x, \tau) = \tilde{s}(x, t),$$

we obtain the following system of parabolic equations:

$$\begin{aligned} r_\tau &= -r\left(\frac{1-\alpha}{1+\alpha} - \frac{s^2}{r}\right) \quad \text{in } Q, \\ s_\tau &= +\frac{\alpha}{1-\alpha}\, s\left(\frac{1-\alpha}{1+\alpha} - \frac{s^2}{r}\right) + \{(1-\alpha)r\}^{-\alpha/(1-\alpha)}\, e^{\mu\tau} s_{xx} \quad \text{in } Q, \end{aligned} \tag{2.13a}$$

together with the initial and boundary conditions

$$\begin{aligned} &s_x(0, t) = 0, \qquad s_x(1, t) = 0, \qquad t > 0 \\ &r(x, 0) = r_0(x), \qquad s(x, 0) = s_0(x), \qquad x \in [0, 1], \end{aligned} \tag{2.13b}$$

where $\mu = 1/(1+\alpha)$ and r_0 and s_0 are positive functions on $[0, 1]$. If $r_0, s_0 \in W^{1,2}(0, 1)$, then the local existence and the uniqueness of a classical solution on an interval $0 \leqq t \leqq T_0$, where T_0 depends only on the supremum norm of r_0 and s_0, and of $1/r_0$ and $1/s_0$, follow from a straightforward contraction argument [S]. Thus, if we obtain a priori bounds from above and below for r and s, which are uniform in time, it follows that the problem (2.13) has a solution that exists for all $t \geqq 0$.

To prove the a priori bounds, we use the notion of invariant regions. An application of Theorem 4.1 of [CCS] yields that the rectangles

$$\Sigma_{ab} = \left\{(r, s):\ a < r < b,\ \left(\frac{1-\alpha}{1+\alpha}\, a\right)^{1/2} < s < \left(\frac{1-\alpha}{1+\alpha}\, b\right)^{1/2}\right\}, \qquad 0 < a < b,$$

(see Fig. 1) are invariant regions for (2.13).

Consequently problem (2.13) admits arbitrarily large invariant rectangles. Thus, if $r_0, s_0 \in W^{1,2}(0, 1)$ and there exist positive constants r^0, r^1, s^0, and s^1 such that

$$r^0 \leqq r_0(x) \leqq r^1, \qquad s^0 \leqq s_0(x) \leqq s^1,$$

then (2.13) has a unique global solution $(r(x, t), s(x, \tau))$. If in addition we assume that the constants r^k, s^k $(k = 0, 1)$ are related through the expression

$$(s^k)^2 = (1-\alpha) r^k / (1+\alpha),$$

which involves no loss of generality, then

$$\begin{aligned} &r^0 \leqq r(x, \tau) \leqq r^1, \\ &s^0 \leqq s(x, \tau) \leqq s^1 \end{aligned} \qquad \text{in } Q. \tag{2.14}$$

Thus we have proved the following theorem.

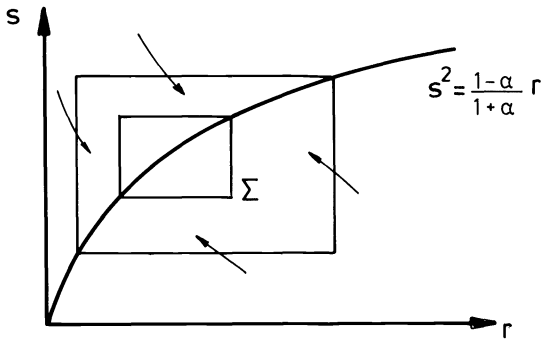


FIG. 1. *Invariant regions.*

THEOREM 2.1. *Problem* (1.11)-(1.14) *with initial data* $\theta_0, \sigma_0 \in W^{1,2}(0,1)$ *such that*

$$0 < \sigma^0 \leqq \sigma_0(x) \leqq \sigma^1, \qquad 0 \leq \theta^0 \leqq \theta_0(x) \leqq \theta^1$$

has a unique classical solution on $\bar{Q}$. *In addition*

$$\sigma^0(t+1)^{-\alpha/(1+\alpha)} \leqq \sigma(x,t) \leqq \sigma^1(t+1)^{-\alpha/(1+\alpha)} \quad \text{in } \bar{Q},$$

$$\theta^0(t+1)^{1/(1+\alpha)} \leqq \theta(x,t) \leqq \theta^1(t+1)^{1/(1+\alpha)} \quad \text{in } \bar{Q},$$

when the constants $\sigma^k, \theta^k (k=0,1)$ *satisfy the relation* $(\sigma^k)^2 = (\theta^k)^{1-\alpha}/(1+\alpha)$.

Having established global bounds for σ and θ in Q, we can now turn to the question of large-time behaviour. We begin with the observation that *both* σ and θ converge to their mean values $\bar{\sigma}$ and $\bar{\theta}$ defined by

$$\bar{\sigma}(t) = \int_0^1 \sigma(x,t)\,dx, \qquad \bar{\theta}(t) = \int_0^1 \theta(x,t)\,dx.$$

Defining $\bar{s}(t)$ and $\bar{r}(t)$ in a similar fashion we obtain the following estimate.

LEMMA 2.2. *We have*

(a) $$\|s(\cdot,\tau) - \bar{s}(\tau)\|_\infty = O(e^{-\mu r}) \quad \text{as } \tau \to \infty,$$

(b) $$\|r(\cdot,\tau) - \bar{r}(\tau)\|_\infty = O(e^{-\gamma\tau}) \quad \text{as } \tau \to \infty,$$

where

$$\mu = \frac{1}{1+\alpha}, \qquad \gamma = \frac{1-\alpha}{1+\alpha}$$

and $\|\cdot\|_\infty$ *denotes the norm in* $L^\infty(0,1)$.

Proof. We multiply the equation for s by $-s_{xx}$ and integrate over $(0,1)$. This yields, when we integrate the left-hand side by parts and use the boundary conditions,

$$\frac{1}{2}\frac{d}{dt}\int_0^1 s_x^2\,dx = -e^{\mu\tau}\int_0^1 \{(1-\alpha)r\}^{-\alpha/(1-\alpha)} s_{xx}^2\,dx - \int_0^1 g(s,r)s_{xx}\,dx,$$

where

$$g(s,r) = \frac{\alpha}{1-\alpha}\, s\left(\frac{1-\alpha}{1+\alpha} - \frac{s^2}{r}\right).$$

Because $r(x,\tau)$ is bounded above by r^1 in Q, we have

$$\{(1-\alpha)r(x,\tau)\}^{-\alpha/(1-\alpha)} \geqq \{(1-\alpha)r^1\}^{-\alpha/(1-\alpha)} \stackrel{\text{def}}{=} a_0$$

and so

$$\frac{1}{2}\frac{d}{dt}\int_0^1 s_x^2\,dx \leqq -a_0\, e^{\mu\tau}\int_0^1 s_{xx}^2\,dx + \frac{\lambda}{2}\int_0^1 s_{xx}^2\,dx + \frac{1}{2\lambda}\int_0^1 g^2\,dx, \tag{2.15}$$

where λ is a positive number, which may depend on τ. If we choose $\lambda = a_0\, e^{\mu\tau}$, (2.15) becomes

$$\frac{1}{2}\frac{d}{dt}\int_0^1 s_x^2\,dx \leqq -\frac{1}{2}\,a_0\, e^{\mu\tau}\int_0^1 s_{xx}^2\,dx + \frac{1}{2a_0}\,e^{-\mu\tau}\int_0^1 g^2\,dx. \tag{2.16}$$

Finally, we observe that in view of the boundary conditions $s_x \in H_0^1(0,1)$, and we can use Poincaré's inequality to estimate the first term on the right of (2.16). This yields in the end

$$\frac{d}{dt}\int_0^1 s_x^2\,dx \leqq -A\, e^{\mu\tau}\int_0^1 s_x^2\,dx + B\, e^{-\mu\tau}, \tag{2.17}$$

where A and B are positive constants, because g is uniformly bounded in Q by (2.14). An elementary computation now reveals that (2.17) implies that

$$\int_0^1 s_x^2(x, \tau)\, dx \leqq C e^{-2\mu\tau}, \qquad \tau \geqq 0, \tag{2.18}$$

where C is some positive constant.

To complete the proof of part (a), we note that for some $\hat{x}(\tau) \in [0, 1]$ we have $s(\hat{x}(\tau), \tau) = \bar{s}(\tau)$. Hence

$$s(x, \tau) - \bar{s}(\tau) = \int_{\hat{x}(\tau)}^{x} s_x(\xi, \tau)\, d\xi$$

and so

$$|s(x, \tau) - \bar{s}(\tau)| \leqq \int_0^1 |s_x(\xi, \tau)|\, d\xi \leqq \left(\int_0^1 s_x^2(\xi, \tau)\, d\xi \right)^{1/2}. \tag{2.19}$$

Substituting (2.18) into (2.19), we arrive at the desired bound.

To prove part (b), we solve the equation for r in terms of s. This is possible since the equation is linear in r:

$$r_\tau = -\gamma r + s^2.$$

We obtain

$$r(x, \tau) = e^{-\gamma\tau} r(x, 0) + e^{-\gamma\tau} \int_0^\tau e^{\gamma p} s^2(x, p)\, dp.$$

However, by part (a) and the uniform lower bound for s,

$$s^2(x, \tau) = \bar{s}^2(\tau) + \varepsilon(x, \tau),$$

where

$$\|\varepsilon(\cdot, \tau)\|_\infty \leqq e^{-\mu\tau}.$$

Hence

$$r(x, \tau) = e^{-\gamma\tau} r(x, 0) + e^{-\gamma\tau} \int_0^\tau e^{\gamma p} \bar{s}^2(p)\, dp + \rho(x, \tau), \tag{2.20}$$

where

$$\|\rho(\cdot, \tau)\|_\infty \leqq C e^{-\mu\tau}.$$

If we integrate (2.20) over $(0, 1)$, we obtain

$$\bar{r}(\tau) = e^{-\gamma\tau} \bar{r}(0) + e^{-\gamma\tau} \int_0^\tau e^{\gamma p} \bar{s}^2(p)\, dp + \int_0^1 \rho(x, \tau)\, dx, \tag{2.21}$$

whence

$$r(x, \tau) - \bar{r}(\tau) = e^{-\gamma\tau} \{ r(x, 0) - \bar{r}(0) \} - \int_0^1 \rho(x, \tau)\, dx,$$

i.e.,

$$\|r(\cdot, \tau) - \bar{r}(\tau)\|_\infty \leqq C e^{-\gamma\tau},$$

because $\gamma < \mu$.

Remark. Equation (2.21) yields a relation between $\bar{r}$ and $\bar{s}$. In view of the estimate for $\rho(\cdot, \tau)$ we can write it as

$$\bar{r}(\tau) = e^{-\gamma\tau} \int_0^\tau e^{\gamma p} \bar{s}^2(p)\, dp + O(e^{-\gamma\tau}). \tag{2.22}$$

Another relation between $\bar{r}$ and $\bar{s}$ is supplied by the compatibility relation (1.6) in which we have set $V = 1$. This yields after transformation

$$\int_0^1 s(x, \tau)\{r(x, \tau)\}^{\alpha/(1-\alpha)}\, dx = (1-\alpha)^{-\alpha/(1-\alpha)}. \tag{2.23}$$

Applying Lemma 2.2, we deduce from (2.23) that

$$\bar{s}(\tau)\{\bar{r}(\tau)\}^{\alpha/(1-\alpha)} = (1-\alpha)^{-\alpha/(1-\alpha)} + O(e^{-\gamma\tau}). \tag{2.24}$$

It now remains to determine the asymptotic behaviour of $\bar{r}(\tau)$ and $\bar{s}(\tau)$ as $\tau \to \infty$. This is done by means of the equation for $r(x, \tau)$. It yields, upon integration over $(0, 1)$,

$$\bar{r}' = -\gamma\bar{r} + \int_0^1 s^2\, dx = -\gamma\bar{r} + \bar{s}^2(\tau) + \omega(\tau),$$

where the prime denotes differentiation with respect to τ and $\omega(\tau) = O(e^{-\mu\tau})$ as $\tau \to \infty$. Eliminating $\bar{s}$ by means of (2.24) we finally obtain the equation

$$\bar{r}' = g(\bar{r}) + \tilde{\omega}(\tau), \tag{2.25}$$

where

$$g(\xi) = -\gamma\xi + \{(1-\alpha)\xi\}^{-2\alpha/(1-\alpha)}, \qquad \xi > 0$$

and $\tilde{\omega}(\tau) = O(e^{-\gamma\tau})$ as $\tau \to \infty$. Plainly, g has a unique zero ξ_0 given by

$$\xi_0 = \frac{1}{1-\alpha}(1+\alpha)^{(1-\alpha)/(1+\alpha)},$$

$$g(\xi) \begin{cases} > 0 & \text{if } 0 < \xi < \xi_0, \\ < 0 & \text{if } \xi_0 < \xi < \infty. \end{cases}$$

Thus

$$\bar{r}(\tau) \to \xi_0 \quad \text{as } \tau \to \infty$$

and, because $g'(\xi_0) < -\gamma$,

$$\bar{r}(\tau) = \xi_0 + O(e^{-\gamma\tau}) \quad \text{as } \tau \to \infty.$$

The limiting behaviour of $\bar{s}(\tau)$ now follows from (2.24):

$$\bar{s}(\tau) = \eta_0 + O(e^{-\gamma\tau}) \quad \text{as } \tau \to \infty,$$

where

$$\eta_0 = \{(1-\alpha)\xi_0\}^{-\alpha/(1-\alpha)} = (1+\alpha)^{-\alpha/(1-\alpha)}.$$

In summary, we have proved the following theorem.

THEOREM 2.3. *Let the initial data σ_0 and θ_0 satisfy the conditions of Theorem* 2.1 *and let $V = 1$. Then the asymptotic behaviour of the solution (σ, θ) of problem* (1.15)–(1.18) *is given by*

$$\sigma(x, t) = \{(1+\alpha)t\}^{-\alpha/(1+\alpha)}[1 + O(t^{-1/(1+\alpha)})] \qquad \text{as } t \to \infty,$$

$$\theta(x, t) = \{(1+\alpha)t\}^{1/(1+\alpha)}[1 + O(t^{-(1-\alpha)/(1+\alpha)})] \qquad \text{as } t \to \infty,$$

and v_x satisfies

$$v_x(x, t) = 1 + O(t^{-(1-\alpha)/(1+\alpha)}) \quad \text{as } t \to \infty.$$

The error terms are all uniform with respect to $x \in [0, 1]$.

Remark. It is readily seen from the solution (y, q) of the system of ordinary differential equations (2.6) that the powers in the error terms in Theorem 2.3 are optimal.

3. The case $\alpha = 1$. When $\alpha = 1$, Problem (1.11)-(1.14) becomes

$$\sigma_t = -\sigma^3 + \theta^{-1}\sigma_{xx} \quad \text{in } Q, \tag{3.1}$$

$$\theta_t = \theta\sigma^2 \quad \text{in } Q \tag{3.2}$$

with

$$\sigma_x(0, t) = 0, \qquad \sigma_x(1, t) = 0, \qquad t > 0, \tag{3.3}$$

$$\sigma(x, 0) = \sigma_0(x), \qquad \theta(x, 0) = \theta_0(x), \qquad x \in [0, 1]. \tag{3.4}$$

As in § 2, we rescale the variables, writing

$$\tilde{r}(x, t) = (t+1)^{-1/2}\theta(x, t), \qquad \tilde{s}(x, t) = (t+1)^{1/2}\sigma(x, t).$$

If in addition we introduce again the new time variable $\tau = \log(t+1)$, we obtain for

$$r(x, \tau) = \tilde{r}(x, t) \quad \text{and} \quad s(x, \tau) = \tilde{s}(x, t)$$

the system of equations

$$s_\tau = -s\left(s^2 - \frac{1}{2}\right) + \frac{1}{r}e^{\tau/2}s_{xx}, \tag{3.5}$$

$$r_\tau = r\left(s^2 - \frac{1}{2}\right) \tag{3.6}$$

with boundary conditions

$$s_x(0, t) = 0, \quad s_x(1, t) = 0, \quad t > 0 \tag{3.7}$$

and initial conditions

$$s(x, 0) = \sigma_0(x), \quad r(x, 0) = \theta_0(x), \quad x \in [0, 1]. \tag{3.8}$$

Observe that because

$$v_x = rs,$$

the boundary conditions (1.4), with $V = 1$, yield the identity

$$\int_0^1 r(x, t)s(x, t)\, dx = 1. \tag{3.9}$$

To prove the a priori bounds on r and s, we no longer can use the method of invariant regions. However, since the lower-order terms in (3.5) for s now do not depend on r, we can still use a maximum principle argument to obtain the desired bounds. As before, we consider the pair of ordinary differential equations obtained from (3.5)-(3.6) by omitting the diffusion term:

$$y' = -y(y^2 - \tfrac{1}{2}), \tag{3.10}$$

$$z' = z(y^2 - \tfrac{1}{2}). \tag{3.11}$$

Here primes denote differentiation with respect to τ. Let $y(\tau; a)$ be the solution of (3.10) with initial value $y(0; a) = a$ and let $z(\tau; a, b)$ be the solution of (3.11) with $y = y(\tau; a)$ and initial value $z(0; a, b) = b$. For any $a \in \mathbf{R}$ the solution $y(\tau; a)$ exists, is unique and bounded for all $\tau \geqq 0$. It can actually be found explicitly and for any $a > 0$ its large-time behaviour is given by

$$y(\tau; a) = \frac{1}{\sqrt{2}} + O(e^{-\tau}) \quad \text{as } \tau \to \infty. \tag{3.12}$$

Because z satisfies a linear equation with bounded coefficients, $z(\tau; a, b)$ also exists for all time.

By uniqueness we have

$$a_1 < a_2 \Rightarrow y(\tau; a_1) < y(\tau; a_2) \quad \text{for } \tau \geqq 0 \tag{3.13}$$

and because the right-hand side of (3.11) is increasing in y when $y > 0$, we also have

$$\begin{aligned} &a_1 < a_2, \qquad b_1 < b_2 \\ &\Rightarrow z(\tau; a_1, b_1) < z(\tau; a_2, b_2) \quad \text{for } \tau \geqq 0. \end{aligned} \tag{3.14}$$

Set $a_1 = \min \sigma_0$ and $a_2 = \max \sigma_0$. Then by the maximum principle

$$y(\tau; a_1) \leqq s(x, \tau) \leqq y(\tau; a_2) \quad \text{in } \bar{Q},$$

and so, by (3.12),

$$s(x, \tau) = \frac{1}{\sqrt{2}} + O(e^{-\tau}) \quad \text{as } \tau \to \infty, \tag{3.15}$$

uniformly in $[0, 1]$.

To estimate r we set $b_1 = \min \theta_0$ and $b_2 = \max \theta_0$, leaving a_1 and a_2 unchanged. Then for any $x \in [0, 1]$ we have $\theta_0 \in [b_1, b_2]$ and hence

$$z(\tau; a_1, b_1) \leqq r(x, \tau) \leqq z(\tau; a_2, b_2) \quad \text{in } \bar{Q}.$$

Note that (3.10) and (3.11) imply that

$$\frac{z'}{z} = -\frac{y'}{y}.$$

Therefore

$$z(\tau; a, b) = \frac{ab}{y(\tau; a)}$$

and hence

$$z(\tau; a, b) = \sqrt{2}\, ab + O(e^{-\tau}) \quad \text{as } \tau \to \infty.$$

Thus we have established the existence of a global solution (r, s) of (3.5)-(3.8) as well as uniform bounds in $\bar{Q}$.

Reformulating this result in terms of the solution (σ, θ) of (3.1)-(3.4) we obtain the following theorem.

THEOREM 3.1. *Problem* (3.1)-(3.4), *with initial data* $\sigma_0, \theta_0 \in W^{1,2}(0, 1)$ *that are positive in* $[0, 1]$, *has a unique solution* $(\sigma(x, t), \theta(x, t))$ *on* $\bar{Q}$. *In addition there exist positive constants* $A^{\pm}$, $B^{\pm}$ *such that*

$$A^-(t+1)^{-1/2} \leqq \sigma(x, t) \leqq A^+(t+1)^{-1/2} \quad \text{in } \bar{Q},$$

$$B^-(t+1)^{1/2} \leqq \theta(x, t) \leqq B^+(t+1)^{1/2} \quad \text{in } \bar{Q}.$$

For the asymptotic behaviour of solutions of (3.1)-(3.4) we have only a partial answer.

THEOREM 3.2. *Let* $(\sigma(x, t), \theta(x, t))$ *be the solution of* (3.1)-(3.4). *Then*

(a) $$\sigma(x, t) = \frac{1}{\sqrt{2t}}[1 + O(t^{-1})] \quad \text{as } t \to \infty,$$

(b) $$\theta(x, t) = \theta^*(x)\sqrt{t}[1 + o(1)] \quad \text{as } t \to \infty,$$

uniformly with respect to $x \in [0, 1]$, *where* θ^* *is a positive function that depends on the initial data.*

Proof. Part (a) is an immediate consequence of (3.12).

To prove part (b), we observe that

$$\frac{d}{d\tau}\{\log r(x, \tau)\} = s^2(x, \tau) - \frac{1}{2} = \nu(x, \tau) \tag{3.16}$$

and recall that $\nu(x, \tau) = O(e^{-\tau})$ as $\tau \to \infty$ uniformly in x. Integration of (3.16) over $(0, \tau)$ yields

$$r(x, \tau) = r(x, 0) \exp\left(\int_0^\tau \nu(x, s)\, ds\right), \tag{3.17}$$

and hence, because $\nu(x, \cdot) \in L^1(0, \infty)$ for every $x \in [0, 1]$,

$$\lim_{\tau \to \infty} r(x, \tau) \text{ exists.}$$

Returning to θ gives the second limit.

Theorem 3.2, (1.3), and (1.8) yield the following corollary.

COROLLARY 3.3. *Let* $(\sigma(x, t), \theta(x, t))$ *be the solution of* (3.1)-(3.4). *Then*

$$\lim_{t \to \infty} v_x(x, t) = \frac{1}{\sqrt{2}}\theta^*(x), \qquad 0 \leqq x \leqq 1.$$

This completes the proof of Theorem B.

Remarks. 1. If s_0 does not depend on x, then neither do $s(x, t)$ nor, by (3.16), do $\nu(x, t)$ depend on x. Hence we deduce from (3.17) that in that case θ^* is a constant if and only if θ_0 is a constant. This example shows that in general we cannot expect convergence to the uniform shearing flow if $\alpha = 1$.

2. Because $\theta^*(x) > 0$ for every $x \in [0, 1]$ it follows from Theorem 3.2 that, as when $0 < \alpha < 1$, $\theta(x, t) \to \infty$ as $t \to \infty$ for every $x \in [0, 1]$. As we will see in the next section, this will cease to be true when α becomes greater than 1.

4. The case $\alpha > 1$. In this section we shall study a particular family of solutions of the problem defined by (1.11)-(1.14).

As in the case $0 < \alpha < 1$, we introduce the function

$$q = \frac{1}{\alpha - 1}\theta^{1-\alpha}$$

and use q to rewrite system (1.11)-(1.14) as

$$\sigma_t = -\frac{\alpha}{\alpha - 1}\frac{\sigma^3}{q} + \{(\alpha - 1)q\}^{\alpha/(\alpha-1)}\sigma_{xx} \quad \text{in } Q, \tag{4.1}$$

$$q_t = -\sigma^2 \quad \text{in } Q \tag{4.2}$$

with boundary and initial conditions

$$\sigma_x(0, t) = 0, \quad \sigma_x(1, t) = 0, \quad t > 0, \tag{4.3}$$

$$\sigma(x, 0) = \sigma_0(x), \quad q(x, 0) = q_0(x), \quad x \in [0, 1], \tag{4.4}$$

where $\sigma_0 = \theta_0^{-\alpha} v_{0x}$. For positive initial values σ_0, $\theta_0 \in W^{1,2}(0, 1)$ the local existence and uniqueness of a solution (σ, θ) of (4.1)-(4.4) are readily established.

Let $0 < c < 1$ be a constant to be determined appropriately and $0 < \gamma < 1$ a parameter. Define

$$\beta = \frac{\alpha}{\alpha - 1} \quad \text{and} \quad \delta = c^{\beta}\gamma.$$

We define the following family of initial functions:

$$\sigma_0(x) = (\alpha - 1)^{\beta/2} \quad \text{for } 0 \leqq x \leqq 1, \tag{4.5}$$

$$q_{0,\gamma}(x) = \begin{cases} c & \text{for } x \in [0, \delta], \\ 1 & \text{for } x \in (\delta, 1]. \end{cases} \tag{4.6}$$

For what follows it will be convenient to introduce the integral

$$I_\gamma \overset{\text{def}}{=} \int_0^1 \sigma_0^2(x)\{(\alpha - 1)q_{0,\gamma}(x)\}^{-\beta}\, dx. \tag{4.7}$$

Note that

$$I_\gamma = (1 - c^{\beta})\gamma + 1 \quad \text{and} \quad V_\gamma = (\alpha - 1)^{-\beta/2} I_\gamma$$

and that

$$I_\gamma \to 1 \quad \text{and} \quad V_\gamma \to V_0 \overset{\text{def}}{=} (\alpha - 1)^{-\beta/2} \quad \text{as } \gamma \to 0.$$

In terms of the original variables v and θ the initial data are now given by

$$v_{0\gamma}(x) = \int_0^x \{(\alpha - 1)q_{0,\gamma}(s)\}^{-\beta}\sigma_0(s)\, ds, \qquad x \in [0, 1],$$

$$\theta_{0,\gamma}(x) = \{(\alpha - 1)q_{0,\gamma}(x)\}^{-1/(\alpha-1)}, \qquad x \in [0, 1].$$

It is readily verified that

$$v_{0,\gamma}(x) \to V_0 x \quad \text{in } L^2(0, 1) \quad \text{as } \gamma \to 0,$$

$$\theta_{0,\gamma}(x) \to (\alpha - 1)^{-1/(\alpha-1)} \quad \text{in } L^2(0, 1) \quad \text{as } \gamma \to 0.$$

The idea underlying this choice of initial functions is based on the observation that if we set q small compared to σ in the neighbourhood of $x = 0$, it will soon reach zero, according to (4.2), unless σ drops sufficiently fast. However, diffusion of σ towards the origin may prevent this from happening and thus cause q to remain positive for all time. Which of these two possibilities will occur is still an open question.

In this section we will use the family of solutions introduced above to show that the uniform shear flow $(\hat{v}, \hat{\theta}(t))$ is unstable. By this we mean that there exists a $\rho > 0$ such that for any $\varepsilon > 0$ there exist initial functions $(v_{0,\gamma}, \theta_{0,\gamma})$ from the family introduced above such that if

$$\|v_{0,\gamma} - \hat{v}\|_{L^2} + \|\theta_{0,\gamma} - \hat{\theta}(0)\|_{L^2} + |V_\gamma - V_0| < \varepsilon,$$

then there exists a finite time T_ε such that *either* the solution ceases to exist at $t = T_\varepsilon$, *or*

$$\|v_\gamma(t) - \hat{v}\|_{L^2} + (1+t)^{-1/(\alpha+1)} \|\theta_\gamma(t) - \hat{\theta}(t)\|_{L^2} + |V_\gamma - V_0| > \rho \quad \text{for } t = T_\varepsilon.$$

More precisely, we will prove the following theorem.

THEOREM 4.1. *Let* $\alpha > 1$ *and let for some* $b \in (0, 1)$,

$$c = \frac{b(1-b)}{8(1+b)}.$$

Let $v_{0,\gamma}$, $\theta_{0,\gamma}$ *and* V_γ *be defined as before and let* $(v_\gamma(x, t), \theta_\gamma(x, t))$ *be the corresponding solution of* (1.1)-(1.4), (1.9) *on a maximal interval* $[0, T_\gamma)$, *where* $0 < T_\gamma \leqq \infty$. *If*

$$0 < \gamma < \frac{4c}{b} \quad \text{and} \quad T_\gamma = \infty,$$

then

(a) $$|v_\gamma(x, t) - V_\gamma| = O(t^{-1/2}) \quad \text{as } t \to \infty$$

uniformly on $(\delta, 1]$;

(b) $\theta(x, t)$ *is uniformly bounded on* $(\delta, 1] \times [0, \infty)$.

Using the definition of $v_{0,\gamma}$, $\theta_{0,\gamma}$ and V_γ, we can readily check that the uniform shear flow $(\hat{v}, \hat{\theta}(t))$ corresponding to the initial and boundary conditions $v_0(x) = V_0 x$ and $\theta_0 = (\alpha - 1)^{-1/(\alpha-1)}$ is unstable in the sense formulated above. In particular, Theorem C follows from Theorem 4.1.

The proof of Theorem 4.1 is based on estimates for the solution pair (σ, q) (from now on we shall omit the subscript γ). We begin with an upper bound for q.

LEMMA 4.2. *Let* (σ, θ) *be the solution of* (4.1)-(4.4). *Then for* $0 \leqq \xi < \eta \leqq 1$ *and* $b > 0$ *we have*

$$q(\xi, t) \leqq q(\xi, 0) - \frac{b}{1+b} \int_0^t \sigma^2(\eta, \tau)\, d\tau + b \int_0^t \int_0^1 \sigma_x^2(x, \tau)\, dx\, d\tau. \tag{4.8}$$

Proof. Integration of σ_x over (ξ, η) yields

$$\sigma(\eta, t) - \sigma(\xi, t) = \int_\xi^\eta \sigma_x(x, t)\, dx \leqq \left(\int_0^1 \sigma_x^2(x, t)\, dx \right)^{1/2}$$

or

$$\sigma(\xi, t) \geqq \sigma(\eta, t) - \left(\int_0^1 \sigma_x^2(x, t)\, dx \right)^{1/2}. \tag{4.9}$$

Hence, for small t, when in view of the initial data the right-hand side of (4.9) is nonnegative,

$$\begin{aligned} \sigma^2(\xi, t) &\geqq \sigma^2(\eta, t) \left\{ 1 - \frac{1}{\sigma(\eta, t)} \left(\int_0^1 \sigma_x^2(x, t)\, dx \right)^{1/2} \right\}^2 \\ &\geqq \frac{b}{1+b} \sigma^2(\eta, t) - b \int_0^1 \sigma_x^2(x, t)\, dx, \end{aligned} \tag{4.10}$$

where we have used the inequality

$$(1-x)^2 \geqq \frac{b}{1+b} - bx^2, \qquad 0 \leqq x \leqq 1, \quad b > 0. \tag{4.11}$$

Note that if the right-hand side of (4.9) is negative, the right-hand side of (4.10) is also negative, and hence (4.10) is still satisfied.

Integration of (4.2) finally yields, together with (4.10),

$$q(\xi, t) = q(\xi, 0) - \int_0^t \sigma^2(\xi, \tau)\, d\tau$$

$$\leqq q(\xi, 0) - \frac{b}{1+b} \int_0^t \sigma^2(\eta, \tau)\, d\tau + b \int_0^t \int_0^1 \sigma_x^2(x, \tau)\, dx\, d\tau,$$

which we set out to prove.

In the next lemma we show that the diffusion term in (4.8) is uniformly bounded with respect to γ for bounded γ.

LEMMA 4.3. *Let* (σ, q) *be the solution of* (4.1), (4.2) *with initial data* (σ_0, q_0) *given by* (4.5) *and* (4.6). *Then*

$$\int_0^t \int_0^1 \sigma_x^2(x, \tau)\, dx\, d\tau < \frac{1}{2} I_\gamma. \tag{4.12}$$

Proof. Because of boundary condition (4.3), we have

$$\begin{aligned} \int_0^1 \sigma_x^2(x, t)\, dx &= -\int_0^1 \sigma(x, t)\sigma_{xx}(x, t)\, dx \\ &= -\int_0^t \sigma\{(\alpha-1)q\}^{-\beta}\left(\sigma_t + \beta\frac{\sigma^3}{q}\right) dx \qquad (4.13) \\ &= -\int_0^1 \sigma\sigma_t\{(\alpha-1)q\}^{-\beta}\, dx - \alpha\int_0^1 \sigma^4\{(\alpha-1)q\}^{-\beta-1}\, dx. \end{aligned}$$

We now integrate (4.13) over $(0, t)$. Because

$$\begin{aligned} \int_0^t \int_0^1 \sigma\sigma_t\{(\alpha-1)q\}^{-\beta}\, dx\, d\tau &= \frac{1}{2}\int_0^1 \sigma^2\{(\alpha-1)q\}^{-\beta}\, dx \\ &\quad -\frac{1}{2}\int_0^1 \sigma_0^2\{(\alpha-1)q_0\}^{-\beta}\, dx \qquad (4.14) \\ &\quad -\frac{1}{2}\alpha\int_0^t \int_0^1 \sigma^4\{(\alpha-1)q\}^{-\beta-1}\, dx\, d\tau, \end{aligned}$$

this yields in the end

$$\int_0^t \int_0^1 \sigma_x^2(x, \tau)\, dx\, d\tau \leqq \frac{1}{2}\int_0^1 \sigma_0^2\{(\alpha-1)q_0\}^{-\beta}\, dx,$$

and the lemma follows from (4.7).

Observe that by (4.2)

$$\int_0^t \sigma^2(x, \tau)\, d\tau = 1 - q(x, t), \qquad \delta < x \leqq 1.$$

Hence, putting the results of Lemmas 4.2 and 4.3 together, and remembering that

$q(\xi, 0) = c$ if $\xi \leqq \delta$, we obtain the upper bound

$$
\begin{aligned}
q(\xi, t) &\leqq c - \frac{b}{1+b}\{1 - q(\eta, t)\} + \frac{b}{2} I_\gamma \\
&\leqq c - \frac{b}{1+b}\{1 - q(\eta, t)\} + \frac{b}{2}(1+\gamma) \qquad (4.15) \\
&= c - \frac{b}{2(1+b)}\{1 - b - \gamma(1+b) - 2q(\eta, t)\}
\end{aligned}
$$

for $0 \leqq \xi < \delta$ and $\xi < \eta \leqq 1$. We now choose

$$
0 < b < 1, \quad c = \frac{b(1-b)}{8(1+b)}, \quad \gamma = \frac{4c}{b}
$$

for $0 \leqq \xi < \delta$ and $\xi < \eta \leqq 1$. Then (4.15) becomes

$$
q(\xi, t) \leqq \frac{b}{1+b}\left\{ q(\eta, t) - \frac{1}{8}(1-b) \right\}.
$$

We distinguish two cases.

Case I. There exists a point $\eta_1 \in (\delta, 1]$ and a time $T_1 > 0$ such that

$$
q(\eta_1, T_1) < \frac{1-b}{8}.
$$

This means that $q(\xi, t)$ must have vanished at some time $t = T_0 < T_1$.

Case II. We have

$$
q(x, t) \geqq \frac{1-b}{8} \quad \text{for all } \delta < x \leqq 1, \quad t \geqq 0.
$$

In the first case, the solution ceases to exist at $t = T_0$, or possibly even before.

To prove Theorem 4.1, we may assume that the solution exists for all time and we are necessarily in the second case. In view of the definition of q this means that

$$
\theta(x, t) \leqq \{\tfrac{1}{8}(1-b)(\alpha-1)\}^{-1/(\alpha-1)} \quad \text{for } \delta < x \leqq 1, \quad t \geqq 0 \qquad (4.16)
$$

and so, according to (1.3) and (1.9),

$$
0 \leqq v_x(x, t) \leqq K\sigma(x, t), \qquad \delta < x \leqq 1, \quad t \geqq 0 \qquad (4.17)
$$

for some positive constant K.

It now remains to show that $\sigma(x, t) \to 0$ as $t \to 0$. From the differential equation (4.2) for q, we see that q is decreasing in t. Therefore $q(x, t) < q(x, 0) \leqq 1$. This means that the function $\bar{\sigma}(x, t) = y(t)$ defined as the solution of the initial value problem

$$
\begin{aligned}
y' &= -\beta y^3, \qquad t > 0, \\
y(0) &= (\alpha - 1)^{\beta/2}
\end{aligned}
$$

is a supersolution of (4.1), (4.3), and (4.4). Hence

$$
0 < \sigma(x, t) \leqq (c_0 + 2\beta t)^{-1/2}, \qquad 0 \leqq x \leqq 1, \quad t \geqq 0,
$$

where $c_0 = (\alpha - 1)^{-\beta}$, and thus by (4.17),

$$
|v_x(x, t)| \leqq K(c_0 + 2\beta t)^{-1/2}, \qquad \delta < x \leqq 1, \quad t \geqq 0.
$$

Therefore

$$
|v(x, t) - V| = O(t^{-1/2}) \quad \text{as } t \to \infty
$$

uniformly on $(\delta, 1]$. Since θ is uniformly bounded on $(\delta, 1]$ by (4.16), the proof of Theorem 4.1 is complete.

Remark. The initial function $q_{0,\gamma}$ defined by (4.6) does not belong to $W^{1,2}(0, 1)$ as has been assumed throughout this paper. This assumption was made for computational reasons. However, if for some $\gamma > 0$, $q_{0,\gamma}$ is an initial value for which the solution blows up in finite time, so is any function $\tilde{q}$ with the properties (i) $\tilde{q} = c$ for $0 \leqq x \leqq \delta$; (ii) $c \leqq \tilde{q}(x) \leqq 1$; and (iii) $\|\tilde{q} - q_{0,\gamma}\|_{L^2}$ sufficiently small.

REFERENCES

[CCS] K. CHUEH, C. CONLEY, AND J. SMOLLER, *Positive invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373-392.

[DH] C. M. DAFERMOS AND L. HSIAO, *Adiabatic shearing of incompressible fluids with temperature dependent viscosity*, Quart. Appl. Math., 41 (1983), pp. 45-58.

[S] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, New York, 1982.

[T1] A. E. TZAVARAS, *Shearing of materials exhibiting thermal softening or temperature dependent viscosity*, Quart. Appl. Math., 44 (1986), pp. 1-12.

[T2] ———, *Effect of thermal softening in shearing of strain-rate dependent materials*, Arch. Rational Mech. Anal., XXX (1987), pp. 349-374.

[T3] ———, *Strain softening in viscoelasticity of the rate type*, J. Integral Equations Appl., to appear.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 344-351, March 1991

003

THE NONCONVEX MULTI-DIMENSIONAL RIEMANN PROBLEM FOR HAMILTON–JACOBI EQUATIONS*

MARTINO BARDI† AND STANLEY OSHER‡

Abstract. Simple inequalities are presented for the viscosity solution of a Hamilton–Jacobi equation in N space dimension when neither the initial data nor the Hamiltonian need be convex (or concave). The initial data are uniformly Lipschitz and can be written as the sum of a convex function in a group of variables and a concave function in the remaining variables, therefore including the nonconvex Riemann problem. The inequalities become equalities wherever a "maxmin" equals a "minmax" and thus a representation formula for this problem is then obtained, generalizing the classical Hopf's formulas.

Key words. Hamilton–Jacobi equations, viscosity solutions, Riemann problem, Godunov's scheme, Hopf's representation formulas

AMS(MOS) subject classifications. 35L99, 35L65, 65M15, 65M10

1. Introduction.

We are concerned with viscosity solutions (see Crandall and Lions [3], Crandall, Evans, and Lions [2], Lions [12]) to the following partial differential equation:

$$\varphi_t + H(D_x\varphi) = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty), \tag{H-J}$$

satisfying the initial data

$$\varphi(x, 0) = \varphi_0(x) \quad \text{in } \mathbb{R}^N, \tag{IC}$$

where $H \in C(\mathbb{R}^N)$, $D_x\varphi = (\varphi_{x_1}, \cdots, \varphi_{x_N})$ is the spatial gradient of φ, and φ_0 is at least uniformly continuous. This Cauchy problem has, for any $T > 0$, a unique viscosity solution $\varphi(x, t)$ in the space $UC_x(\mathbb{R}^N \times [0, T])$ of the continuous functions which are uniformly continuous in $x \in \mathbb{R}^N$ uniformly in $t \in [0, T]$, see Ishii [10] or Crandall and Lions [5].

We are interested in giving explicit pointwise upper and lower bounds for the solution, providing in some cases a representation formula for φ, for some special initial data but without extra assumptions on the Hamiltonian H.

Some general representation formulas for viscosity solutions of Cauchy problems for Hamilton–Jacobi equations are due to Evans [6] and Evans and Souganidis [7]. However, they either involve an infinite number of max-min operations over $\mathbb{R}^N$ [6], or a single max-min operation over infinite-dimensional sets of "controls" and "strategies" [7]. Two simpler formulas solving almost everywhere (H-J)(IC), one dual of the other, were derived by Hopf [9] for two special cases. The first one holds for convex Hamiltonians and general (Lipschitz) initial data, and it is well known in the theory of conservation laws in the case $N = 1$ (it is often called the Lax formula). It was shown to give the viscosity solution to the problem by Lions [12], Evans [6], Bardi and Evans [1], with different proofs and slightly different assumptions. The second

* Received by the editors January 30, 1989; accepted for publication February 6, 1990.

† Dipartimento di Matematica Pura ed Applicata, Università di Padova, via Belzoni 7, 35131 Padova, Italy. The work of this author was partially supported by the Italian National Project "Equazioni di evoluzione e applicazioni fisicomatematiche."

‡ Mathematics Department, University of California, Los Angeles, California 90024. The work of this author was partially supported by National Science Foundation grant DMS88-11863, Defense Advanced Research Projects Agency grant in the ACMP Program, Office of Naval Research grant N00014-86-K-0691, and NASA Langley grant NAG1-270.

Hopf's formula is valid for general Hamiltonians and convex or concave (Lipschitz) initial data φ_0, and it is

$$\varphi(x,t)=\sup_{v\in\mathbb{R}^N}\{x\cdot v-\varphi_0^*(v)-tH(v)\} \tag{1.1}$$

for φ_0 convex, and

$$\varphi(x,t)=\inf_{v\in\mathbb{R}^N}\{x\cdot v-\varphi_0^*(-v)-tH(v)\} \tag{1.2}$$

for φ_0 concave, where φ_0^* is the Legendre transform (or Fenchel conjugate) of φ_0, that is

$$\varphi_0^*(v):=\sup_{x\in\mathbb{R}^N}\{x\cdot v-\varphi_0(x)\}\leqq+\infty$$

for φ_0 convex, while for φ_0 concave it is

$$\varphi_0^*(v):=-(-\varphi_0)^*(v)=\inf_{x\in\mathbb{R}^N}\{-x\cdot v-\varphi_0(x)\}\geqq-\infty.$$

Osher [14] rederived for the viscosity solution of (H-J) the special case of formula (1.1) occurring when the initial data are of Riemann type (and convex), i.e., they are piecewise affine with one jump in the derivative across a plane. Bardi and Evans [1] showed the connection between Osher's formulas for convex Riemann data and Hopf's formulas, and proved that (1.1) and (1.2) give the viscosity solution of (H-J)(IC) in the general case. Lions and Rochet [13] gave a different proof under slightly more general assumptions.

We are now going to describe our main result. Let j be an integer, $0\leqq j\leqq N$, and for any $v\in\mathbb{R}^N$ set

$$v=(v_A,v_B),\quad v_A:=(v_1,\cdots,v_j)\in\mathbb{R}^j,\quad v_B:=(v_{j+1},\cdots,v_N)\in\mathbb{R}^{N-j}.$$

THEOREM 1. *Assume* $H\in C(\mathbb{R}^N)$, $\varphi_1:\mathbb{R}^j\to\mathbb{R}$ *uniformly Lipschitz and convex,* $\varphi_2:\mathbb{R}^{N-j}\to\mathbb{R}$ *uniformly Lipschitz and concave. Then the unique viscosity solution* $\varphi\in UC_x(\mathbb{R}^N\times[0,T])$ *of* (H-J) *taking on the initial data*

$$\varphi(x,0)=\varphi_1(x_A)+\varphi_2(x_B)$$

satisfies for all $x\in\mathbb{R}^N$ *and* $t\geqq0$

$$\sup_{v_A\in\mathbb{R}^j}\inf_{v_B\in\mathbb{R}^{N-j}}G(v,x,t)\leqq\varphi(x,t)\leqq\inf_{v_B\in\mathbb{R}^{N-j}}\sup_{v_A\in\mathbb{R}^j}G(v,x,t), \tag{1.3}$$

where

$$G(v,x,t):=x\cdot v-\varphi_1^*(v_A)-\varphi_2^*(-v_B)-tH(v).$$

Note that the pointwise estimate (1.3) gives a representation formula for the solution whenever the first and last terms are equal (as they are for $t=0$). A trivial case where this occurs is for $j=N$ or $j=0$, because (1.3) reduces to Hopf's formulas (1.1) or (1.2). A more interesting case occurs when the Hamiltonian separates the variables v_A and v_B, that is,

$$H(v)=H_1(v_A)+H_2(v_B).$$

In this case we get

$$\varphi(x,t)=\sup_{v_A\in\mathbb{R}^j}\{x_A\cdot v_A-\varphi_1^*(v_A)-tH_1(v_A)\}+\inf_{v_B\in\mathbb{R}^{N-j}}\{x_B\cdot v_B-\varphi_2^*(-v_B)-tH_2(v_B)\},$$

which is the superposition of the solutions to the problems

$$\varphi_t + H_i(D_x\varphi) = 0, \qquad \varphi(\cdot, 0) = \varphi_i$$

for $i = 1, 2$.

Next we specialize formula (1.3) to a particular class of (Riemann) initial data. Let A, u_i^+, u_i^- be constants and define

$$u_i(x) := \begin{cases} u_i^+ & \text{if } x_i > 0, \\ u_i^- & \text{if } x_i < 0 \end{cases}$$

for $i = 1, \cdots, N$. Then take

$$\varphi_0(x) = A + \sum_{i=1}^{N} x_i u_i(x) = A + x \cdot u(x). \tag{1.4}$$

These data correspond to a Riemann problem for the system of conservation laws satisfied (formally) by the spatial gradient of φ; see Remark 2.2. Let, for $i = 1, \cdots, N$,

$$\Omega_i := \{s \mid \min(u_i^+, u_i^-) \leqq s \leqq \max(u_i^+, u_i^-)\},$$
$$\chi_i := \operatorname{sign}(u_i^+ - u_i^-),$$

and reorder the indices, without loss of generality, so that

$$\chi_i = 1 \quad \text{for } i = 1, \cdots, j; \qquad \chi_i = -1 \quad \text{for } i = j+1, \cdots, N \tag{1.5}$$

$(0 \leqq j \leqq N)$. Finally set

$$\Omega_A := \Omega_1 \times \cdots \times \Omega_j; \quad \Omega_B := \Omega_{j+1} \times \cdots \times \Omega_N; \quad \Omega := \Omega_A \times \Omega_B.$$

COROLLARY 2. *The viscosity solution to* (H-J)(IC) *with the initial data given by* (1.4) *under the convention* (1.5) *satisfies*

$$A + \max_{v_A \in \Omega_A} \min_{v_B \in \Omega_B} \{x \cdot v - tH(v)\} \leqq \varphi(x, t) \leqq A + \min_{v_B \in \Omega_B} \max_{v_A \in \Omega_A} \{x \cdot v - tH(v)\}. \tag{1.6}$$

The rest of the paper is organized as follows. In § 2, as motivation, we show how formula (1.6) was first (formally) derived in connection with numerical approximation schemes for Hamilton–Jacobi equations and for conservation laws. In § 3 we give the proofs of Theorem 1 and Corollary 2, which are quite different from the previous derivation, and rather simple, in that they make use only of Hopf's formulas (1.1), (1.2) and a comparison argument.

2. A derivation of (1.6) by means of Godunov's Hamiltonians. The purpose of this section is to motivate Corollary 2 and to explain its connection with approximation schemes for (H-J). The rigorous proofs will be given in § 3. We assume that the solutions of (H-J) have the following properties:

(P1) The solution $\varphi(x, t)$ is a nondecreasing function of the initial data.

(P2) The partial derivatives φ_{x_i} satisfy a maximum principle at points of continuity, i.e., for $i = 1, \cdots, N$:

$$\min(u_i^-, u_i^+) \leqq \varphi_{x_i} \leqq \max(u_i^-, u_i^+).$$

(P3) The speed of propagation is finite.

(P4) If $\psi(x_2, \cdots, x_N, t)$ is a viscosity solution of

$$\psi_t + H(v_1, \psi_{x_2}, \cdots, \psi_{x_N}) = 0$$

for a constant v_1 then

$$\varphi(x, t) = v_1 x_1 + \psi(x_2, \cdots, x_N, t)$$

is a viscosity solution to (H-J).

It is easy to see formally that the solution to the Cauchy problem (H-J)(IC), with initial data given by (1.4), satisfies

$$\varphi(x,t) = tg\left(\frac{x}{t}\right) + A = tg(\zeta) + A, \tag{2.1}$$

where g satisfies:

$$g = \zeta \cdot D_\zeta g - H(D_\zeta g) \tag{2.2}$$

whenever $D_\zeta g$ is continuous.

In (H-J), we let $\tau = t$, $y_i = x_i - \zeta_i t$ for ζ fixed. (H-J) becomes

$$\varphi_\tau + H(D_y\varphi) - \zeta D_y\varphi = \varphi_\tau + H^1(D_y\varphi) = 0 \qquad \text{(defining } H^1(D_y\varphi)\text{)} \tag{H-J^1}$$

with the same initial data (1.4).

Thus, by (2.2), to evaluate $g(\zeta)$ we need only evaluate $-H^1(D_y(g(y)))$ at $y=0$ for any $t>0$. From (P2) above we know that $(D_y g)_{y=0}$ lies in Ω for $t>0$. Moreover, if we integrate (H-J)^1 from $\tau=0$ to $\tau=\Delta t$ we have

$$\begin{aligned}\varphi(0,\Delta t) &= A - \Delta t H^1((D_y g)_{y=0}) \\ &= \varphi_0(0) - \Delta t \tilde{H}^1(D_+^{x_1}\varphi_0(0), D_-^{x_1}\varphi_0(0); D_+^{x_2}\varphi_0(0), D_-^{x_2}\varphi_0(0); \cdots \\ &\quad \cdots; D_+^{x_N}\varphi_0(0), D_-^{x_N}\varphi_0(0)).\end{aligned} \tag{2.3}$$

Here

$$D_\pm^{x_i}\varphi_0(0) = \pm\frac{(\varphi_0(\pm h e_i) - \varphi_0(0))}{h} = u_i^\pm \tag{2.4}$$

where $e_i = \{0, 0, \cdots, 1, 0, \cdots\}$, the ith unit vector, and $\tilde{H}^1(u_1^+, u_1^-; u_2^+, u_2^-; \cdots; u_N^+, u_N^-)$ is determined by (2.3).

This formula can be interpreted as a numerical algorithm. Suppose we are given a grid

$$x_{j_i}^i = j_i h, \qquad i = 1, \cdots, N; \qquad j_i = 0, \pm 1, \cdots$$

and values of a discrete function $\psi_j = \psi_{j_1 j_2 \cdots j_N}$. Then for each j, we construct the piecewise affine function which, in each of the 2^N orthants centered at j, interpolates ψ_j and its N nearest neighbors, $\psi_{j\pm e_i}$ for $i = 1, \cdots, N$. From (P3), if

$$\text{(CFL)} \qquad \frac{\Delta t}{h} \max_{\substack{v\in\Omega^{(j)} \\ i=1,\cdots,N}} |H_{u_i}^1| \leqq \frac{1}{N^{1/2}}, \tag{2.5}$$

where $\Omega^{(j)}$ is the same as Ω with each u_i^-, u_i^+ replaced by $D_-^{x_i}\psi_j$, $D_+^{x_i}\psi_j$, then the solution to the initial value problem (H-J^1) with the above affine initial data in the diamond centered at j when evaluated at $x = x_j$ and $t = \Delta t$ is independent of the values of the initial data outside of this diamond.

Thus (2.3) (with $\varphi_0(0)$ replaced by ψ_j^n and $\varphi(0,\Delta t)$ by ψ_j^{n+1}), gives us a monotone finite difference scheme approximating (H-J^1) which is in differenced form with numerical Hamiltonian $\tilde{H}^1$. These concepts were introduced in [4]. The scheme is monotone, which means that the right side of (2.3) is an increasing function of all the $\varphi_{j\pm e_i}$, because of property (P1). The function $\tilde{H}^1$ is called Godunov's Hamiltonian by analogy with the definition of Godunov's scheme for conservation laws in one space dimension [8]. The scheme is consistent, which means

$$\tilde{H}^1(u_1, u_1; u_2, u_2; \cdots; u_N, u_N) = H^1(u_1, u_2, \cdots, u_N).$$

Monotonicity implies that

$$\tilde{H}^1(u_1^+, u_1^-; u_2^+, u_2^-; \cdots; u_N^+, u_N^-)$$

is a nonincreasing function of all the u_i^+ and a nondecreasing function of all the u_i. In particular, for $N=1$, this means for any $v_1 \in \Omega = \Omega_1$:

$$\begin{aligned} \operatorname{sgn}(u_1^+ - u_1^-)[\tilde{H}^1(u_1^+, u_1^-) - H^1(v_1)] &= \operatorname{sgn}(u_1^+ - v_1)[\tilde{H}^1(u_1^+, u_1^-) - \tilde{H}^1(v_1, u_1^-)] \\ &\quad + \operatorname{sgn}(v_1 - u_1^-)[\tilde{H}^1(v_1, u_1^-) - \tilde{H}^1(v_1, v_1)] \\ &\leqq 0. \end{aligned} \tag{2.6}$$

But, by (P2), $\tilde{H}^1(u_1^+, u_1^-) = H^1(\tilde{u}_1)$ for some $\tilde{u}_1$ in Ω. Thus we have

$$\tilde{H}^1(u_1^+, u_1^-) = \chi_1 \min_{v_1 \in \Omega_1} \chi_1 H^1(v_1). \tag{2.7}$$

(This formula was obtained earlier in [14].) Now we proceed inductively. Suppose, for $N \leqq M-1$, we have

$$\begin{aligned} &\max_{v_{j+1} \in \Omega_{j+1}} \cdots \max_{v_N \in \Omega_N} \min_{v_1 \in \Omega_1} \cdots \min_{v_j \in \Omega_j} H^1(v_1, v_2, \cdots, v_N) \\ &\leqq \tilde{H}^1(u_1^+, u_1^-; u_2^+, u_2^-; \cdots; u_N^+, u_N^-) \\ &\leqq \min_{v_1 \in \Omega_1} \cdots \min_{v_j \in \Omega_j} \max_{v_{j+1} \in \Omega_{j+1}} \cdots \max_{v_N \in \Omega_N} H^1(v_1, v_2, \cdots, v_N), \end{aligned} \tag{2.8}$$

where

$$\begin{aligned} \chi_i &= 1, \qquad i = 1, \cdots, j, \\ \chi_i &= -1, \qquad i = j+1, \cdots, N. \end{aligned}$$

Next we have, $N = M$ and for any $v_1 \in \Omega_1$:

$$\chi_1[\tilde{H}^1(u_1^+, u_1^-; u_2^+, u_2^-; \cdots; u_M^+, u_M^-) - \tilde{H}^1(v_1, v_1; u_2^+, u_2^-; \cdots; u_M^+, u_M^-)] \leqq 0, \tag{2.9}$$

using the same argument as in (2.6).

Now, for any fixed v_1, $\tilde{H}^1(v_1, v_1; u_2^+ u_2^-, \cdots; u_M^+, u_M^-)$ is Godunov's Hamiltonian when the initial data for (H-J^1) has a constant x_1 derivative,

$$\frac{\partial \varphi_0}{\partial x_1}(x) \equiv v_1.$$

Then it follows from (P4) that

$$g\left(\frac{x}{t}\right) = \frac{x_1}{t} v_1 + \tilde{g}\left(\frac{x_2}{t}, \frac{x_3}{t}, \cdots, \frac{x_M}{t}\right)$$

(where $\tilde{g}$ also depends on v_1).

By the induction hypothesis, this means we have

$$\begin{aligned} &\chi_1 \tilde{H}^1(u_1^+, u_1^-; u_2^+, u_2^-; \cdots; u_M^+, u_M^-) \\ &\quad \leqq \chi_1 \tilde{H}^1(v_1, v_1; u_2^+, u_2^-; \cdots; u_M^+, u_M^-) \\ &\quad = -\chi_1 \tilde{g}(0, 0, \cdots, 0) \\ &\quad \leqq \chi_1 \chi_2 \min_{v_2 \in \Omega_2} \chi_2 \cdots \chi_M \min_{v_M \in \Omega_M} \chi_M H^1(v_1, v_2 \cdots v_M) \\ &\quad = \chi_1 H^1(v_1, \tilde{v}_2, \cdots, \tilde{v}_M) \end{aligned} \tag{2.10}$$

where the extrema is taken on at $\tilde{v}_2, \cdots, \tilde{v}_M$, which depends on v_1. The vector $(v_1, \tilde{v}_2, \cdots, \tilde{v}_N) \in \Omega_1$, where $v_1 \in \Omega_1$ is arbitrary. We next take $\min_{v_1}$ of the expression in (2.10). If all the $\chi_i \equiv 1$ or all the $\chi_i \equiv -1$ we have equality by (P2). Otherwise, $\chi_i \equiv 1$, $1 \leqq i \leqq j$, $\chi_i \equiv -1$, $j+1 \leqq i \leqq M$, and we have the right-hand inequality in (2.8). Next we have, for any $v_{j+1} \in \Omega_{j+1}$, following the argument above:

$$
\begin{aligned}
&\tilde{H}^1(u_1^+, u_1^-; u_2^+, u_2^-; \cdots; u_M^+, u_M^-) \\
&\quad \geqq \tilde{H}^1(u_1^+, u_1^-; \cdots; v_{j+1}; \cdots; u_M^+, u_M^-) \\
&\quad \geqq \chi_{j+2} \min_{v_{j+2} \in \Omega_{j+2}} \chi_{j+2} \cdots \chi_M \min_{v_M \in \Omega_M} \chi_M \chi_1 \min_{v_1 \in \Omega_1} \chi_1 \cdots \chi_j \\
&\qquad \cdot \min_{v_j \in \Omega_j} \chi_j H^1(v_1, v_2, \cdots, v_M).
\end{aligned} \tag{2.11}
$$

We next take the $\max_{v_{j+1}}$ of the expression in (P4) which gives us the left-hand inequality in (2.8).

We have now obtained formula (2.8) for any N; using (2.1) and (2.2) gives us our intuitive derivation of (1.6).

Remark 2.1. We note that (2.8) validates the conjecture about Godunov's Hamiltonian in [15] when the inequalities in (2.10) and (2.11) become equalities. That paper also discusses the high-order accurate nonoscillatory numerical solution of (H-J) in some detail. See also [16] for a further discussion of these issues.

Remark 2.2. If we take the space gradient of (H-J) and call $u_1 = \varphi_{x_1}$, $u_2 = \varphi_{x_2}$, etc., we arrive at the system of conservation laws

$$
(u_i)_t + \frac{\partial}{\partial_{x_i}} H(u_1, \cdots, u_N) = 0, \qquad i = 1, \cdots, N \tag{2.12}
$$

with initial data:

$$
u_i(x, 0) = \begin{cases} u_i^+ & \text{if } x_i > 0 \\ u_i^- & \text{if } x_i < 0, \end{cases} \qquad i = 1, \cdots, N.
$$

Thus (1.6) gives us information about the solution to this special Riemann problem for a special system of conservation laws.

3. Proofs.

Proof of Theorem 1. Since φ_2 can be written as the Legendre transform of its Legendre transform

$$
\varphi_2(x_B) = \inf_{v_B \in \mathbb{R}^N} \{-x_B \cdot v_B - \varphi_2^*(v_B)\},
$$

we will first solve (H-J) with the initial data

$$
\psi_0(v_B, x) = \varphi_1(x_A) - x_B \cdot v_B - \varphi_2^*(v_B), \tag{3.1}
$$

and then take the infimum as v_B varies in $\mathbb{R}^{N-j}$. Since the initial data ψ_0 are convex in x for each choice of v_B, we can write Hopf's formula for the solution $\psi(v_B, x, t)$ of (H-J) plus

$$
\psi(v_B, x, 0) = \psi_0(v_B, x) \quad \text{for all } x \in \mathbb{R}^N.
$$

To do this we compute the Legendre transform with respect to x of ψ_0:

$$
\begin{aligned}
\psi_0^*(v_B, y) &= \sup_{x \in \mathbb{R}^N} \{x_A \cdot y_A + x_B \cdot (y_B + v_B) - \varphi_1(x_A) + \varphi_2^*(v_B)\} \\
&= \begin{cases} +\infty & \text{if } y_B \neq -v_B, \\ \varphi_1^*(y_A) + \varphi_2^*(v_B) & \text{if } y_B = -v_B, \end{cases}
\end{aligned}
$$

and then apply (1.1) to get

$$\psi(v_B, x, t) = \sup_{y_A \in \mathbb{R}^j} \{x_A \cdot y_A - x_B \cdot v_B - \varphi_1^*(y_A) - \varphi_2^*(v_B) - tH(y_A, -v_B)\}$$
$$= \sup_{v_A \in \mathbb{R}^j} G(-v, x, t).$$

Since $\psi(v_B, \cdot, \cdot) \in UC_x(\mathbb{R}^N \times [0, T])$ for all v_B and $\psi(v_B, x, 0) \geqq \varphi(x, 0)$, a standard comparison theorem for unbounded viscosity solutions [10], [5] gives

$$\psi(v_B, x, t) \geqq \varphi(x, t) \quad \text{for all } (x, t) \in \mathbb{R}^N \times [0, T], \quad v_B \in \mathbb{R}^{N-j}.$$

Then

$$\inf_{v_B \in \mathbb{R}^{N-j}} \sup_{v_A \in \mathbb{R}^j} G(v, x, t) = \inf_{v_B \in \mathbb{R}^{N-j}} \psi(v_B, x, t) \geqq \varphi(x, t),$$

which is the second inequality in (1.3).

The first inequality is proved in a similar way. We apply Hopf's formula (1.2) to compute the solution $\psi(v_A, x, t)$ of (H-J) with the concave initial condition

$$\psi(v_A, x, 0) = \varphi_2(x_B) + x_A \cdot v_A - \varphi_1^*(v_A) \leqq \varphi(x, 0).$$

Since

$$\psi^*(v_A, y, 0) = \begin{cases} -\infty & \text{if } y_A \neq -v_A, \\ \varphi_1^*(v_A) + \varphi_2^*(y_B) & \text{if } y_A = -v_B, \end{cases}$$

we get

$$\psi(v_A, x, t) = \inf_{v_B \in \mathbb{R}^{N-j}} G(v, x, t),$$

and, as before, we conclude by means of a comparison theorem. □

Remark 3.1. The first and the third member of (1.3) coincide with φ at $t = 0$, but in general it is not clear whether they are continuous. However, they are anyway respectively a subsolution and a supersolution of (H-J) in the generalized viscosity sense of Ishii [11]. This follows from Proposition 2.4 in [11], because they are, respectively, a supremum and an infimum of solutions of (H-J).

Proof of Corollary 2. We set

$$\varphi_1(x_A) = A + x_A \cdot u(x_A), \qquad \varphi_2(x_B) = x_B \cdot u(x_B),$$

and compute the Legendre transforms

$$\varphi_1^*(v_A) = -A + \sup_{x_A \in \mathbb{R}^j} x_A \cdot (v_A - u(x_A))$$
$$= \begin{cases} +\infty & \text{if } v_i > u_i^+ \text{ or } v_i < u_i^- \quad \text{for some } 1 \leqq i \leqq j, \\ -A & \text{if } u_i^- \leqq v_i \leqq u_i^+ \quad \text{for all } i = 1, \cdots, j \end{cases}$$
$$= \begin{cases} +\infty & \text{if } v_A \notin \Omega_A, \\ -A & \text{if } v_A \in \Omega_A; \end{cases}$$

$$\varphi_2^*(-v_B) = \begin{cases} -\infty & \text{if } v_B \notin \Omega_B, \\ 0 & \text{if } v_B \in \Omega_B; \end{cases}$$

which substituted in (1.3) give immediately (1.6). □

REFERENCES

[1] M. BARDI AND L. C. EVANS, *On Hopf's formulas for solutions of Hamilton-Jacobi equations*, Nonlinear Anal. TMA, 8 (1984), pp. 1373-1381.

[2] M. G. CRANDALL, L. C. EVANS, AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487-502.
[3] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.
[4] ———, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 45 (1984), pp. 1-19.
[5] ———, *On existence and uniqueness of solutions of Hamilton-Jacobi equations*, Nonlinear Anal. TMA, 10 (1986), pp. 353-370.
[6] L. C. EVANS, *Some min-max methods for the Hamilton-Jacobi equation*, Indiana Univ. Math. J., 33 (1984), pp. 31-50.
[7] L. C. EVANS AND P. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773-797.
[8] S. K. GODUNOV, *A finite difference method for the numerical solution of discontinuous solutions of the equations of fluid dynamics*, Math. Sb., 47 (1959), pp. 271-291.
[9] E. HOPF, *Generalized solutions of non-linear equations of first order*, J. Math. Mech., 14 (1965), pp. 951-973.
[10] H. ISHII, *Remarks on existence of viscosity solutions of Hamilton-Jacobi equations*, Bull. Fac. Sci. Engrg. Chuo Univ., 26 (1983), pp. 5-24.
[11] ———, *Perron's method for Hamilton-Jacobi equations*, Duke Math. J., 55 (1987), pp. 369-384.
[12] P. L. LIONS, *Generalized solutions of Hamilton-Jacobi equations*, Pitman, London, 1982.
[13] P. L. LIONS AND J.-C. ROCHET, *Hopf formula and multi-time Hamilton-Jacobi equations*, Proc. Amer. Math. Soc., 96 (1986), pp. 79-84.
[14] S. OSHER, *The Riemann problem for nonconvex scalar conservation laws and Hamilton-Jacobi equations*, Proc. Amer. Math. Soc., 89 (1983), pp. 641-646.
[15] S. OSHER AND J. A. SETHIAN, *Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12-49.
[16] S. OSHER AND C.-W. SHU, *High order essentially nonoscillatory schemes for Hamilton-Jacobi equations*, SIAM J. Numer. Anal. (1989), submitted.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 352-378, March 1991

004

THE ANALYSIS OF A MODEL FOR WAVE MOTION IN A LIQUID SEMICONDUCTOR: BOUNDARY INTERACTION AND VARIABLE CONDUCTIVITY*

WILLIAM V. SMITH

Abstract. The theory of conducting fluids in relative motion with small conductivity is studied with a model including the Maxwell displacement current. The model is linearized, and the interaction of waves with a plane boundary in three space is studied for two orientations of the external magnetic field. It is found that two *families* of boundary conditions preserve energy in one orientation (external field orthogonal to the boundary), while in the other (external field parallel to the boundary) only one condition exists which preserves energy. It is shown that generalized Fourier transforms exist, generated from the generalized eigenfunction expansions. Further, it is shown that surface waves are not supported by this model, indicating that their presence is unstable when relative motion of the fluid is allowed (surface waves exist in the still fluid case). Finally, the problem of variable conductivity (decaying to zero at infinity) is studied and steady-state and time dependent solutions are shown to exist for certain force terms.

Key words. eigenfunction expansions, energy preserving boundaries, variable conductivity, Maxwell's equations, magnetofluiddynamics, liquid semiconductor

AMS(MOS) subject classifications. 35L50, 76W05

1. Introduction. The theoretical modeling of problems in "magnetofluiddynamics" is a rich source of interesting and unusual systems of partial differential equations and corresponding wave motions [LL]. The problem we consider here involves waves of *finite energy* in a fluid-like conducting medium which we assume to be a relatively dense poor conductor, and we treat the *Maxwell displacement current as significant.* The model (like most in this area) can be said to be "physical" only in a certain range of the parameters. For example, in the model studied here, frequency must be relatively high but not high enough to require a particle treatment. (It may also be assumed that permitivity is high relative to free space.) The problem we study here is also of physical interest in a true gas where the constitutive equations (see (1.2)) are much simpler, but we want to examine the fluid case first as, perhaps, a kind of transition state (the theory of liquid semiconductors is still in a rather primitive state with few settled issues [C]). The conductivity appears in only one of the model equations explicitly (see (1.17)). Our model, at apparent zero conductivity, does not reduce (formally) to the uncoupled Maxwell equatoins and fluid motion equations. This is because the (finite) conductivity is implicitly present in the equations containing E'. Mixed frame equations of this type are useful in studying dissipative nonlinear processes since they remove terms which are second order in time. It is this fact that makes the model useful to consider for the undamped behavior of small amplitude waves in a rather dense poor conductor, as this effect makes it possible to study the essentially dissipative problem (1.17) as the bounded perturbation of a symmetric problem. Hence the solutions of (1.17) will be like those of its associated symmetric problem modulo an exponentially decreasing (with time) factor. Elsewhere [S1], [S2], we have already studied the MHD fluid case (perfect conductor), and we refer the reader there for a more detailed

* Received by the editors January 23, 1989; accepted for publication (in revised form) April 2, 1990. This work was partially supported by a grant from Centre National de Recherche Scientifique, France; a scientific investigation grant from CTRC, Orem, Utah; and by a faculty research grant from Brigham Young University.

† Mathematics Department, Brigham Young University, Provo, Utah 84602.

exposition at various points of our treatment here. We note that while this introductory section is mathematically informal (but rather typical of the theoretical treatments of the subject) the following sections are completely rigorous in nature and are founded securely on functional analysis and in particular the Hilbert space theory of differential equations. However, at the end of this section we shall give a brief summary of the results contained in this paper, a comparison to related problems and some comments about the computational difficulties in discussing the differential equations.

The derivation of the problem considered here is founded on Maxwell's equations. The properties of the medium are assumed to be enough like those of a fluid that the continuum approach is reasonably close to reality. We will assume that all fluid *velocities are nonrelativistic*, and that acceleration is small in magnitude (compared to the velocity of light). In order to illustrate the differences between our model and the classical case of a perfect conductor, we will indicate the contrasting assumptions that lead to these two models in their respective derivations.

The Maxwell equations in RMKS units are

$$\nabla \circ D = \rho_e, \qquad \nabla \circ B = 0,$$

$$\nabla \times E = -\frac{\partial B}{\partial t}, \qquad \nabla \times H = J + \frac{\partial D}{\partial t}, \tag{1.1}$$

which hold in any frame of reference, either the rest frame (with respect to the fluid) or the laboratory frame. The constitutive equations must be used and these are (in the laboratory frame):

$$D = \varepsilon_0 E + \varepsilon_0 V \times B - \frac{1}{c^2} V \times H,$$

$$B = \mu_0 H - \mu_0 V \times D + \frac{1}{c^2} V \times E. \tag{1.2}$$

Here as usual, ε_0 is the electric permitivity (for free space), μ_0 the magnetic permeability, E the electric field, H the magnetic field, B the magnetic flux, and D the electric flux. We assume a linear isotropic medium so that μ_0 and ε_0 are scalar constants (this can be modified somewhat—see §§ 4 and 5). c is the speed of light. V is the fluid velocity. We now assume the electric fields to be of order $V \times B$, that is, of the order of magnitude of the induced effects. In other words, the induced magnetic field is much smaller than the externally applied magnetic field. From this it is easily shown that the magnetic induction is the same in all reference frames. Of course, because B is the same in all frames of reference does not mean the same is true for H, but we will see that this is the case under our assumptions on V and the electric field. Let us write H' for the rest frame field and H the laboratory frame. By our Newtonian assumption and the Lorentz transformation,

$$H' = H - V \times (\varepsilon(E + V \times B) - 1/c^2(V \times H)), \tag{1.3}$$

so that $H = H'$ is valid if the magnitude of $V\varepsilon\mu$ is $\ll 1$ ($\varepsilon\mu = 1/c^2$), or in other words, E is approximated by $V \times B$ (here we have used $B = \mu H$). For E' we have

$$E' = E + V \times B. \tag{1.4}$$

E' must always be considered, since to get $H' = H$ as noted, E and $V \times B$ must have the same order of magnitude. We now assume that the period of variation of the fields is large compared to the mean free time of the conduction electrons and that the

Larmor frequency is small compared with the mean free time of the conduction electrons. (In rarefied gases this may break down). This allows the assumption of a constant conductivity σ [LL] (see § 5 below). (We will also assume it is a scalar quantity to begin with—see the remarks on ρ_e below.)

In the MHD approximation, the displacement current $\partial D/\partial t$ would be neglected compared to J, at least when σ is significant. (In a dielectric J is virtually zero.) Here we assume that the displacement current is *not* trivial (in a true metal, for example, the displacement current is essentially meaningless, except at frequencies where the other hypotheses we use begin to break down). In Ohm's law, ρ_e (the space charge) may usually be neglected in a liquid (it must be retained in some gases—we ignore this); hence we have

$$J = \sigma E' + \rho_e V = \sigma(E + V \times B) + (V \circ J/c^2)V. \tag{1.5}$$

The second term is small compared to the first (the coefficient of V in the second term being ρ_e). Thus we take

$$J = \sigma(E + V \times B).$$

Now the Maxwell equations become (note that ρ_e is not present now)

$$\begin{aligned} &-\nabla \times E' + \nabla \times (V \times B) = \frac{\partial B}{\partial t}, \\ &\nabla \times H = \sigma E' + \frac{\partial D}{\partial t}, \\ &\nabla \circ J = 0, \\ &\nabla \circ B = 0, \end{aligned} \tag{1.6}$$

Ohm's law:

$$J = \sigma(E + V \times B).$$

The fluid equations are

$$\frac{\partial p}{\partial t} + \nabla \circ (\rho V) = 0 \qquad \text{(continuity)}, \tag{1.7}$$

(1.8)

$$\rho\left(\frac{\partial V}{\partial t} + \nabla\left(\frac{V \circ V}{2}\right) - V \times (\nabla \times V)\right) = -\nabla P - \rho \nabla \psi + \nabla \circ \tau' + J \times B + \Sigma \qquad \text{(motion)},$$

with the other terms on the right of the motion equation (Σ) depending on the displacement current if it is considered in a fluid—in a gas Σ is zero (see below). Here ψ is the gravitational potential and τ' is the shear part of the mechanical stress tensor τ. From the Maxwell equations,

$$J \times B = (\sigma E' \times B) = -\sigma B \times E' \tag{1.9}$$

(*in the perfect conductor case, we would use the Maxwell equation to get J here—see below* (1.12)) and so the motion equation becomes

$$\rho\left(\frac{\partial V}{\partial t} + (V \circ \nabla)V\right) = -\nabla P - \rho \nabla \psi + \nabla \circ \tau' - \sigma B \times E', \tag{1.10}$$

$$\rho\left(\frac{\partial V}{\partial t} + (V \circ \nabla)V\right) = -\nabla P - \rho \nabla \psi + \nabla \circ \tau' - \mu H \times (\nabla \times H) \tag{1.11}$$

and replacing τ' by its value in terms of velocity and viscosity,

(1.12)
$$\rho\left(\frac{\partial V}{\partial t}+(V\circ\nabla)V\right)=-\nabla P-\rho\nabla\psi+\rho\nu\nabla^2 H+\left(\zeta+\frac{1}{3}\rho\nu\right)\nabla(\nabla\circ V)+\mu(\nabla\times H)\times H,$$

where the second and third equations are the perfect conductor case ($\partial D/\partial t=0$ and so $J=\nabla\times H$). We write this down because it turns out that (1.10) is correct for the case of a gas only, while (1.12) is correct for a fluid because it includes the complete body force (see (1.16) below). Here we have used P for pressure; ν and ζ are, respectively, the first and second coefficients of viscosity.

In many problems it is useful to make the assumption of infinite conductivity in order to obtain qualitative information about physical situations, since this assumption generally allows a much simpler mathematical formulation. An important application of the concept of infinite conductivity is in high temperature plasma studies, such as those associated with fusion devices. In interstellar matter, the decay of the magnetic fields is so slow that infinite conductivity gives a good approximation. In such cases (where infinite conductivity is assumed) the results differ from physical reality by a damping term. When the displacement current is neglected, we can combine the two curl Maxwell equations using Ohm's law and the divergence equation for B to obtain

(1.13)
$$\frac{\partial B}{\partial t}=\eta\nabla^2 B+\nabla\times(V\times B),$$

the so-called magnetic transport equation. Here $\eta=(1/\sigma\mu)$ and is called the magnetic diffusivity by obvious analogy. *For $\sigma=\infty$ we have, formally, that B becomes "frozen into" the fluid (the transport term vanishes).* The neglect of gravitational force and viscosity together with the appropriate equation of state yields the well-known equations of magnetohydrodynamics:

(1.14)
$$\begin{aligned}
&\frac{\partial B}{\partial t}=\nabla\times(V\times B),\\
&\rho\frac{DV}{Dt}=-a^2\nabla\rho+\mu(\nabla\times H)\times H,\\
&\frac{D\rho}{Dt}=0,\\
&\nabla\circ B=0,
\end{aligned}$$

where we have used the convective derivative notation in the second and third equations. a is the sound speed from the equation of state.

Equations (1.14) are mentioned only for comparison's sake since the case of interest here is when σ is relatively small. So, supposing that the displacement current is significant compared to the conduction current, and if we note the equation for D given above, that is,

$$D=\varepsilon_0 E'-\frac{1}{c^2}V\times H,$$

then we have

(1.15)
$$\varepsilon_0\frac{\partial E'}{\partial t}=\nabla\times H-\sigma E'+\frac{1}{c^2}\frac{\partial V\times H}{\partial t}.$$

The essential assumptions made so far are that V is relatively small, external forces (gravity) may be neglected, and dissipation from viscous effects is small. Now, using (1.15), and the constitutive expression for D (there is some question as to the proper form for the body force here but we take the one implied by the Abraham tensor [LL])

and considering the motion equation with the body force term, assuming no steady D field, and neglecting space charge effects, we have for the body force density (last terms in the motion equation)

$$J\times B+\frac{\partial}{\partial t}(D\times B)-\frac{1}{c^2}\frac{\partial}{\partial t}(E\times H)=J\times B+\Sigma. \tag{1.16}$$

Now if the permitivity ε of the medium is comparable to frequency over conductivity (in sea water for example, at frequencies above about 200 megacycles), (or else $\varepsilon \gg \varepsilon_0$), we may neglect the last term on the left side of (1.16) compared to the others. If the magnitude of acceleration is small compared to c we can neglect the last term in (1.15) and so finally obtain

$$\begin{aligned}
&\varepsilon_0\frac{\partial E'}{\partial t}=\nabla\times H-\sigma E',\\
&\frac{\partial B}{\partial t}=\nabla\times(V\times B)-\nabla\times E',\\
&\rho\frac{DV}{Dt}=-a^2\nabla\rho+(\nabla\times B)\times\mu^{-1}B,\\
&\frac{D\rho}{Dt}=0,\\
&\nabla\circ B=0,\qquad \nabla\circ J=0.
\end{aligned} \tag{1.17}$$

The reader will notice that formally, (1.17) reduces to (1.1) when $V=0$ or to (1.14) when $\sigma=\infty$.

Introducing small disturbances about a steady-state condition and neglecting the second-order terms, we finally arrive at a form to observe in terms of wave motion (see (2.1)). It is well known that in the MHD approximation (1.14) ($\eta=0$, or $\sigma=\infty$) there exist essentially three modes of propagation, namely, Alfven waves, and the slow and fast magnetosonic waves [LL], [A], [K], etc. The Alfven waves do not involve acoustic effects but are simply disturbances in the velocity and magnetic fields. As we will see below, the "Alfven waves" in (1.17) degenerate in the sense that they appear as disturbances which are like sound waves (the external field is not "frozen into" the medium) but move more rapidly in the direction of the external field.

It has often been stated that surface wave phenomena are important in the physics of conducting fluids. But as noted in [A], for example, and shown rigorously in [S2], (1.14) does not support surface waves. The displacement current term is needed to generate surface wave phenomena but as we will see, the *presence of surface waves is unstable*: *Whether conductivity is high* (MHD *case*) *or low* (*the case studied here*) *surface waves do not exist when the fluid is in motion. We may say that such surface disturbances are convected away by the fluid. At zero velocity however, a type* I *boundary for orthogonal external fields or the boundary for horizontal external field both support surface waves.* These matters are fully explained below.

As to boundary conditions that are appropriate for the system (1.17), these may be derived from the boundary conditions for Maxwell fields plus the appropriate conditions on the fluid equations. We remark here that the boundary conditions discovered in [Sc1] are related to those derived here in the case where the external magnetic field is parallel to the boundary. This might be expected since the derivation of (1.17) is based on the Maxwell equations. In fact, the boundary conditions in this case are (taking account of the larger number of variables) the so-called "strange"

boundary conditions of [Sc1]. It is perhaps then of some surprise that no surface waves exist in this configuration. This is the instability just mentioned.

In terms of the confinement of fusion plasmas, a number of simple conditions have been studied. No matter the shape of the confinement device for a conducting fluid the boundary problem may frequently be reduced to the consideration of a half-space [J]. That is, the problem may be studied as though the medium occupies a volume with a plane boundary, at least locally.

The boundary conditions for (1.17) with $\sigma = 0$ in a half-space, which are energy conserving, are particularly useful in the study of the dissipative problem $\sigma > 0$. These are derived in the next section.

We have said that one of the main (negative?) results we prove is the absence of surface waves for this model. Another is the absence of steady-state motion for low frequencies. This is proved even for the anisotropic case (see § 5). The reader may consult [S3] and [S4] for a treatment of general systems of the type considered here.

The results given below require very complex computations involving large symbolic matrices and polynomials in several variables. Nearly all of these were carried out using a combination of certain observations about the structure of the matrices involved and certain computer-based symbolic algebra routines constructed by the author as well as those standard packages available from commercial vendors, most computations being done in MAPLE and MACSYMA and a few in MATHEMATICA. A frontal attack on the problems leads nowhere, however, and considerable pattern recognition/reduction is required on the human side. Such techniques are nearly always very specific to the problem and are of an ad hoc nature. Once required objects were derived, checking was done by essentially the same methods, i.e., a combination of human observation and machine interaction. There are several methods of computing the large eigenprojector matrices used here. But they are based on the following facts.

Suppose $A(p)$ is a real symmetric matrix depending on the parameter(s) $p \neq 0$. The spectrum of $A(p)$ is real for all p. $A(p)$ is assumed to have the property that all its entries are linear combinations of the parameter(s) p. The positive and negative eigenvalues of $A(p)$ are equal in number and as continuous functions of p may be enumerated as an ordered list (counting possible multiplicities) as

$$\lambda_k(p) \geqq \lambda_{k-1}(p) \geqq \cdots \geqq 0 (= \lambda_0(p)) \geqq \lambda_{-1}(p) \geqq \cdots \geqq \lambda_{-k}(p).$$

The $\lambda_j(p)$ have the two properties

(1) $\lambda_j(\alpha p) = \alpha \lambda_j(p)$ for all $\alpha > 0$,

(2) $\lambda_j(-p) = -\lambda_{-j}(p)$.

The $\lambda_j(p)$ are roots of the minimal polynomial for $A(p)$ which has the form

$$S = \lambda^{\pi(r(p))}\{\lambda^{2\alpha(p)} + S_1(p)\lambda^{2\alpha(p)-1} + \cdots + S_{2\alpha(p)}(p)\}.$$

In case $A(p)$ has constant rank, $r(p)$ (=dimension of A minus the rank of A) and $\alpha(p)$ are constant. ($\pi(p) = 0$ or 1 depending on whether A is of full rank or not.) We need only deal with the constant rank case in our problem. $D(p)$, the discriminate of S in λ is a homogeneous polynomial and hence the set $\beta = \{p \mid D = 0\}$ is an algebraic cone (in n space for some n). β is the locus of points p where one or more of the functions $\lambda_j(p)$ coincide (and is a set of Lebesgue measure zero). $\lambda_j(p)$ is an analytic function of p on $R^n - \beta$. The orthogonal projection of C^m (A is $m \times m$ and m is related to k in the obvious way) onto the eigenspace for $\lambda_j(p)$ is given by

$$P_j(p) = -\frac{1}{2}\pi i \int_{\gamma_j(p)} (A(p) - z)^{-1}\, dz, \qquad p \in R^n - \beta,$$

where $\gamma_j(p) = \{z \,||z - \lambda_j(p)| = \rho_j(p)\}$ and the $\rho_j(p)$ are chosen so small that the $\gamma_j(p)$

do not intersect. The $P_j(p)$ have the properties ($p \neq 0$)

(1) $P_j(p)$ is analytic on $R^n - \beta$,

(2) $P_j(p) = P_j(\alpha p)$ for all $\alpha > 0$ and $p \in R^n - \beta$,

(3) $P_j(-p) = P_{-j}(p)$ for $p \in R^n - \beta$,

(4) $\sum_j P_j(p) = I$ (the identity matrix),

(5) $A(p)P_j(p) = \lambda_j(p)P_j(p)$.

Each of these facts plays a role in the actual computation of the matrices $P_j(p)$ the results of which are given in § 3 below for a certain $A(p)$ defined by the system of partial differential equations studied here. The path integral for $P_j(p)$ may be computed in a number of ways in a given example; the Cauchy integral theorem is an obvious method of attack. Many of the wave propagation problems of classical physics present with symbols ($A(p)$) of a particularly simple and useful form (the nonzero entries are contained in two nonintersecting submatrices each being the transpose of the other) [Sc1] but the problem we consider here is one of the interesting exceptions to that rule. Hence the computations are more difficult and resolution of the problem requires more basic methods, particularly since we need to extend one of the real parameters p into the complex plane.

2. Boundary conditions. The energy-preserving boundary conditions for the case of a perfect conductor (1.14), $\sigma = \infty$, were characterized in [S1] (see also [S2] and [S5]). The computations are somewhat more complex for the case of (1.17), and since they are carried out in essentially the same manner as in [S1], we will not give the complete details. We will nevertheless construct a complete set of boundary conditions. The divergence equations in (1.17) are contained in the other equations and so will not be needed here. It may be expected that E is divergence free as well. In fact, by the Lorentz transformation of E, the divergence of E' will be related to the divergence of V (the reader will recall that the space charge was neglected in the derivation of (1.17)). This requires that $\nabla \circ (E' - V \times B) = 0$. We will discuss this further below (see (4.18), (4.19)).

Since there is a boundary to consider, the direction of the external magnetic field may not be trivialized by the choice of a convenient coordinate system. The complications arising by treatment of general magnitude and direction of the external field require a great deal of space; we will treat two special cases, namely, external fields which are either parallel or orthogonal to the boundary plane. Oblique fields may be considered at a later time, provided a way can be discovered to sufficiently compress the expressions in a meaningful way.

The linearized version of (1.17) is

$$
\begin{aligned}
&\frac{\partial B}{\partial t} = \nabla \times (V \times B_0) - \nabla \times E', \\
&\varepsilon_0 \frac{\partial E'}{\partial t} = \nabla \times H - \sigma E', \\
&\rho_0 \frac{\partial V}{\partial t} = -a^2 \nabla \rho + (\nabla \times B) \times \mu^{-1} B_0, \\
&\frac{\partial \rho}{\partial t} = -\rho_0 \nabla \circ V.
\end{aligned}
\tag{2.1}
$$

Here, ρ_0 is the equilibrium density, $B_0 = (h_1, h_2, h_3)$ is the external magnetic field, a is the equilibrium speed of sound, μ is the magnetic permeability, B is the internal magnetic field, V is the velocity field, and ρ is the density. If we choose units in which $\varsigma\mu^2 \approx \sqrt{\varepsilon_0}/a^3$ and $h_i \approx \sqrt{\varepsilon_0}/\sqrt{\mu\rho_0}$ numerically, (these may be nonstandard units for

these quantities) then a change of variables in (2.1) allows us to take ρ_0, ε_0, a, and μ as unity, and h_3 or $h_2 = 1$ depending on which external field we consider. *We will assume this to be the case from now on* except where it is necessary to record the location of the external field components. σ has a somewhat different expression, but this is unimportant for our purposes and we still refer to it with the same notation.

We may then write (2.1) in matrix form (we have rearranged the order of the equations in (2.1) as indicated by the definition of u below) as

$$
\begin{aligned}
&-i\frac{\partial u}{\partial t} = -i\sum_{j=1}^{3} A_j \frac{\partial u}{\partial x_j} + iKu, \\
&K = \operatorname{diag}(0, 0, 0, 0, 0, 0, 0, \sigma, \sigma, \sigma), \\
&u = (V, B, \rho, E)^t.
\end{aligned} \tag{2.2}
$$

The superscipt t in (2.2) means transpose and $i = \sqrt{-1}$ (this is added for later convenience), "diag" means the square diagonal matrix with entries as shown and the A_j are given by

$$
A_1 = \begin{bmatrix}
0 & 0 & 0 & 0 & -h_2 & -h_3 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & h_1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & h_1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-h_2 & h_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
-h_3 & 0 & h_1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}, \tag{2.3}
$$

$$
A_2 = \begin{bmatrix}
0 & 0 & 0 & h_2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -h_1 & 0 & -h_3 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & h_2 & 0 & 0 & 0 & 0 \\
h_2 & -h_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -h_3 & h_2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}, \tag{2.4}
$$

$$
A_3 = \begin{bmatrix}
0 & 0 & 0 & h_3 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & h_3 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -h_1 & -h_2 & 0 & -1 & 0 & 0 & 0 \\
h_3 & 0 & -h_1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & h_3 & -h_2 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}. \tag{2.5}
$$

We write D_j for $-i\partial/\partial x_j$ and $D=(D_1, D_2, D_3)$. For the right-hand side of (2.2) (taking $K=0$), we write $A(D)$. As we noted above, we will consider the case of a half-space where the vector B_0 is given by either $(0, 0, h_3)$ or $(0, h_2, 0)$. When it is necessary to distinguish between these two cases, we will do so by using a superscript as $A^3(D)$ and $A^2(D)$. By n, we mean the inward unit normal vector to ∂G (=boundary of G), where G is some domain in $\mathbb{R}^3$. From here on, $G=\mathbb{R}^3_+=\{x \mid x=(x_1, x_2, x_3),\ x_3>0\}$.

DEFINITION 2.1 [LP]. A subspace $\mathcal{S}(n)$ of $\mathbb{R}^{10}$ is a maximal conservative boundary space for $A(D)$ in G if and only if $\zeta \circ A(n)\zeta=0$ for all ζ in $\mathcal{S}(n)$ and $\mathcal{S}(n)$ is maximal with respect to this property.

To proceed further, it is necessary to consider the eigenvalues of the symbol of $A(D)$. These are the solutions to the equation $\det(A(p)-\lambda I)=0$, where $p=(p_1, p_2, p_3)\in\mathbb{R}^3\backslash\{0\}$ (the plane wave speeds). They are given by (for A^3):

$$
\begin{aligned}
&{}_3\lambda_0(p)=0 \qquad \text{(multiplicity 4)},\\
&{}_3\lambda_{\pm1}(p)=\pm(2p_3^2+|n|^2)^{1/2} \qquad \text{(each with multiplicity 1)},\\
&{}_3\lambda_{\pm2}(p)=\frac{\pm(-(p_3^2(p_3^2+6|n|^2)+5|n|^4)^{1/2}+3(p_3^2+|n|^2))^{1/2}}{\sqrt{2}},\\
&{}_3\lambda_{\pm3}(p)=\frac{\pm((p_3^2(p_3^2+6|n|^2)+5|n|^4)^{1/2}+3(p_3^2+|n|^2))^{1/2}}{\sqrt{2}}
\end{aligned}
\tag{2.6}
$$

(each with multiplicity 1 for almost all p). Here we have used the notation $|n|=(p_1^2+p_2^2)^{1/2}$ and in (2.7), $|n_1|=(p_1^2+p_3^2)^{1/2}$. For A^2,

$$
\begin{aligned}
&{}_2\lambda_0(p)=0 \qquad \text{(multiplicity 4)},\\
&{}_2\lambda_{\pm1}(p)=\pm(|n_1|^2+2p_2^2)^{1/2},
\end{aligned}
\tag{2.7}
$$

and similarly for the rest, exchanging p_2 and p_3, n and n_1 in (2.6). For future reference, we record the following: (cf. (3.24) and also (3.33))

$$
\begin{aligned}
&\frac{\partial {}_3\lambda_{\pm1}}{\partial\tau}=\frac{2\tau}{{}_3\lambda_{\pm1}},\\
&\frac{\partial {}_3\lambda_{\pm2}}{\partial\tau}=\frac{-1}{2\,{}_3\lambda_{\pm2}}\left(\frac{2\tau(\tau^2+6|n|^2)+2\tau^3)}{\tau^2(\tau^2+6|n|^2)+5|n|^4)^{1/2}}+3\tau\right),\\
&\frac{\partial {}_3\lambda_{\pm3}}{\partial\tau}=\frac{1}{2\,{}_3\lambda_{\pm3}}\left(\frac{2\tau(\tau^2+6|n|^2)+2\tau^3)}{\tau^2(\tau^2+6|n|^2)+5|n|^4)^{1/2}}+3\tau\right)
\end{aligned}
\tag{2.6a}
$$

with the expressions for ${}_2\lambda_{\pm j}$ obtained in a similar fashion. The multiplicity of the second and third eigenvalues in (2.6) may change for p of certain direction and magnitude ($p=(0, 0, \pm1)$). This is important for the application of Lemma 2.2 below. We will refer to ${}_i\lambda_{\pm1}$ (2.6), (2.7) as the quasi-Alfven wave speeds since the constant speed surfaces of these waves have the same relation to the electromagnetosonic constant speed surfaces as do Alfven waves for the MHD slow and fast magnetosonic waves (i.e., roughly speaking, first the fast wave arrives, then the Alfven wave, and finally the slow wave; in the direction of the external field, the Alfven wave may arrive at the same time as either the slow or fast wave depending upon certain relationships of the parameters (see [CH]).

It is evident from (2.6), (2.7) that $A(D)$ is strongly propagative [Wi]. It is instructive to compare this with the MHD case [S2]. For MHD there are (almost everywhere)

three nonzero plane wave speeds, and they are given (using the notation above) by

$$
\begin{aligned}
&{}_3\lambda_0 \equiv 0,\\
&{}_3\lambda_{\pm 1} \equiv \pm p_3,\\
&{}_3\lambda_{\pm 2} \equiv \pm(|p|^2 - |n||p|)^{1/2},\\
&{}_3\lambda_{\pm 3} \equiv \pm(|p|^2 + |n||p|)^{1/2},
\end{aligned}
\tag{2.8}
$$

and

$$
\begin{aligned}
&{}_2\lambda_0 \equiv 0,\\
&{}_2\lambda_{\pm 1} \equiv \pm p_2,\\
&{}_2\lambda_{\pm 2} \equiv \pm(|p|^2 - |n_1||p|)^{1/2},\\
&{}_2\lambda_{\pm 3} \equiv \pm(|p|^2 + |n_1||p|)^{1/2}.
\end{aligned}
\tag{2.9}
$$

Here, if p is orthogonal to B_0 ($=(0, 0, 1)$ or $(0, 1, 0)$ in (2.8) or (2.9), respectively), then $\lambda_{\pm 1,2}$ vanish. $\lambda_{\pm 1,2}$ are the Alfven and slow magnetosonic wave speeds, respectively [CH]. It is instructive to consider the slow magnetosonic speed profile (see Fig. 1) (the normal surface or "slowness surface" [CH], [Wi]) compared to the electromagnetosonic profiles of (2.1) (see Figs. 2, 3, and 4—the grids in these figures are the same relative size). These are just the unit level surfaces of the functions ${}_i\lambda_j(p)$ in p space—in Figs. 2–4, the plane is the p_1p_2 plane, $i = 3$. From Fig. 1 we see the constant speed (normal) surface for the slow magnetosonic wave. It is unbounded (it tends to the direction of the Alfven wave surface (a plane!) to which it is parallel at ∞), while that for the slow electromagnetosonic wave (Fig. 2) is roughly inverse to that of Fig. 1; it is bounded. The quasi-Alfven surface is caught between the slow and fast surfaces just as for MHD (see the illustrations on page 615 of [CH] for a two-dimensional cross

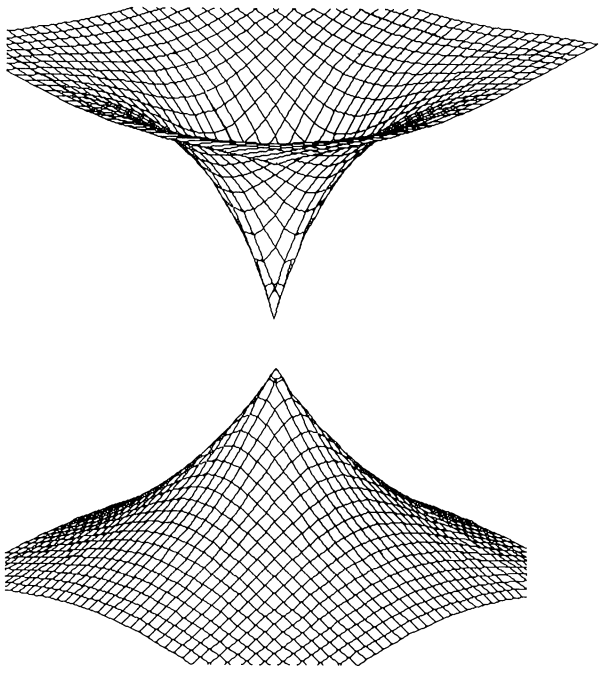


FIG. 1

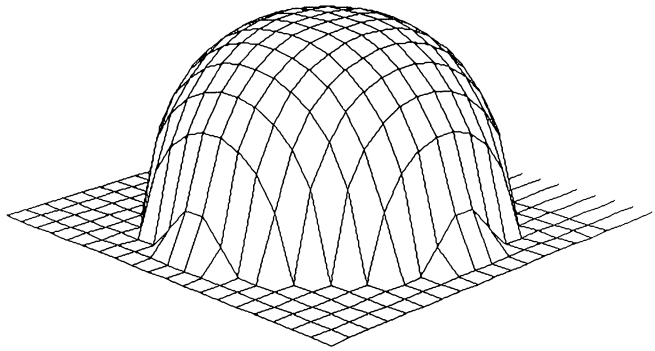


FIG. 2

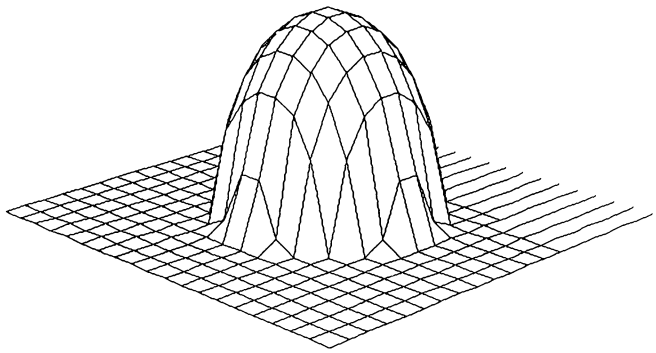


FIG. 3

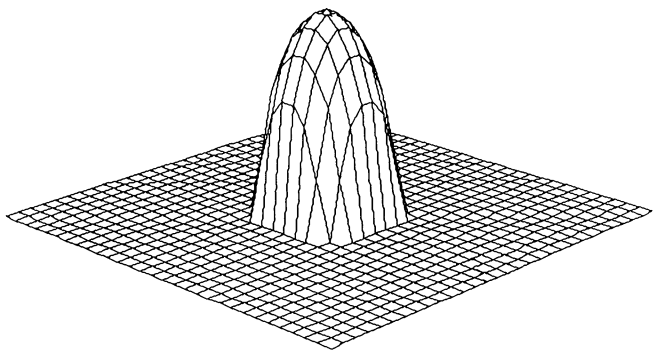


FIG. 4

section for MHD). In our choice of units and external field intensity, the quasi-Alfven surface meets the fast wave surface at the external magnetic axis (vertical in all figures) and is disjoint from the slow wave surface.

By the positive and negative eigenvectors we mean those corresponding to the positive and negative eigenvalues, respectively. Their number depends on B_0.

LEMMA 2.2 [Sc1]. *Let $\mathcal{N}(A(n))$, $\mathcal{X}(n)$, $\mathcal{Y}(n)$ denote, respectively, the null space of $A(n)$, the subspace spanned by the positive eigenvectors of $A(n)$, and the subspace spanned by the negative eigenvectors of $A(n)$. Let ζ_j be any orthonormal base of $\mathcal{N}(A(n))$. Let ξ_j be any base of $\mathcal{X}(n)$ that is orthonormal with respect to $A(n)$, i.e., $\xi_i \circ A(n)\xi_j = \delta_{ij}$, and*

η_j be any base of $\mathcal{Y}(n)$ orthonormal with respect to $-A(n)$ (j as for $\mathcal{X}(n)$). Suppose $\mathcal{S}^{3,2}(n)$ is the subspace of $\mathbb{R}^{10}$ spanned by $\{\zeta_j, \xi_j+\eta_j \text{ (all } j)\}$. Then $\mathcal{S}^{3,2}(n)$ is a maximal conservative boundary space for $A(D)$ and any such boundary space may be constructed in this way.

The lemma is obvious when the eigenvalues of $A(n)$ ($=A_3$, see (2.5)) are computed. (Recall that $n=(0,0,1)$.)

To classify such spaces, we proceed as in [S1] and [S5]. Consider any basis of $\mathcal{X}\oplus\mathcal{Y}$, say for A^3. We have ξ_1, ξ_2, ξ_3 and η_1, η_2, η_3 with $\lambda_1\leftrightarrow\xi_1$, $\lambda_2\leftrightarrow\xi_2$, $\lambda_3\leftrightarrow\xi_3$, etc. Let e_2^1, e_2^2, e_1^3, e_{-2}^1, e_{-2}^2, e_{-1}^3 be any such fixed basis. Then we have

$$
\begin{aligned}
&\eta_i = d_{i1}e_{-2}^1 + d_{i2}e_{-2}^2 \qquad (i=1,2),\\
&\eta_3 = d_3 e_{-1}^3,\\
&\xi_i = c_{i1}e_2^1 + c_{i2}e_2^2 \qquad (i=1,2),\\
&\xi_3 = c_3 e_1^3,
\end{aligned}
\tag{2.10}
$$

In order that the orthonormality conditions be satisfied, it must be that $c_{i1}c_{j1}+c_{i2}c_{j2}=\delta_{ij}$, and thus the matrix $[c_{ij}]$ must be orthogonal and the same is true of $[d_{ij}]$. The constants d_3 and c_3 must have the value 1. Thus by letting $C=[c_{ij}]$ and $D=[d_{ij}]$ run through all possible such matrices, we obtain all possible orientations of the boundary spaces associated with A^3. This allows us to compute operators whose kernels identify the boundary spaces for A^3 (and by a similar process, for A^2). For the details, we refer the reader to [S1] and [Sc1]. In any case, the boundary operators obtained by this process for A^3 consist of two one-parameter families which can be written (here we include the effect of the external field intensity) as

$$
\begin{aligned}
G_{1,\lambda} &= \begin{bmatrix} h_3\lambda & -h_3 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda & 0 \\ 0 & 0 & 0 & 1 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix},\\
G_{2,\lambda} &= \begin{bmatrix} h_3\lambda & 0 & 0 & 0 & -1 & 0 & 0 & 0 & \lambda & 0 \\ 0 & h_3\lambda & 0 & 1 & 0 & 0 & 0 & -\lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.
\end{aligned}
\tag{2.11}
$$

One thing which is immediately apparent from (2.11) is that for orthogonal external magnetic fields, the fluid density must vanish at the boundary, if the energy is confined to a half space. This has been proposed in the physical literature, see [A] for example. This fact is in contrast to the orthogonal field case in MHD, where the component of velocity orthogonal to the boundary must vanish ([S1] or [S2]) (nothing is required of the density) and the boundary conditions do not depend on the field intensity. $G_{1,\lambda}$ couples the velocity and electric fields at the boundary while $G_{2,\lambda}$ couples all three fields. Neither condition requires anything from the induced field components orthogonal to the boundary.

For the case of the parallel external field, there is but a single boundary condition for which energy is preserved, and it does not depend on the external field intensity. The boundary condition is

$$
G_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.
\tag{2.12}
$$

Once again we see that the density must vanish at the boundary. The reader should compare (1.7) with the "strange" boundary condition of [Sc2] to see that they are the same, modulo the density term. Again there is no condition on the velocity field, while the magnetic and electric field components in the direction of the external field must vanish at the boundary. The parallel field case in MHD also gives a single energy-preserving boundary condition but there the boundary condition depends on the external field intensity. However (see [S2]), in MHD the density and induced field component in the direction of the external field are coupled at the boundary ($h_2H_2 + r = 0$ at $x_3 = 0$). Thus in the MHD case, if the density does vanish at the boundary, so also must H_2, which is reminiscent of G_2. One other comparison between MHD and the present system should be noted: the modes are uncoupled in the parallel case for MHD (an incident slow wave generates only a slow wave, etc. [S2, Thm. 3.6], and they are here, too.

DEFINITION 2.3. The operators $A^{3,2}$ in $L_2(\mathbb{R}^2 \times \mathbb{R}_+, \mathbb{C}^{10}) = \mathscr{H}$ with their associated boundary spaces $\mathscr{S}^{3,2}$ are defined with domains:

$$D(A^{3,2}) = \{u \mid u, A^{3,2}u \text{ are in } \mathscr{H} \text{ and } G_{i,\lambda}u \text{ or } G_2 u = 0 \text{ if } x_3 = 0 \ (i = 1 \text{ or } 2)\}.$$

The proof of self-adjointness is essentially the same as in Theorem 3.1 of [Sc2] and will not be repeated here. We note the following, which may be proved in a manner similar to that of [Sc1].

THEOREM 2.4. *If u is in $\mathscr{D}(A) \cap \mathscr{N}(A)^{\perp}$, then the D_3 derivative of u lies in $L_2(\mathbb{R}_+, \mathscr{H}^{-1})$ where $\mathscr{H}^{-1}$ is the usual Sobolev space, $u(\cdot, 0)$ is in $\mathscr{S}$ in the $\mathscr{H}^{-1/2}$ sense and there exists a sequence*

$$\{u_k\} \subset \mathscr{H}^1(\mathbb{R}^2 \times \mathbb{R}_+, \mathbb{C}^{10}) \cap C(\mathbb{R}^2 \times (\mathbb{R}_+ \backslash \{0\}), \mathbb{C}^{10}) \cap \mathscr{D}(A) \cap \mathscr{N}(A)^{\perp}$$

such that $u_k(x_1, x_2, 0)$ is in $\mathscr{S}$ (with either orientation), $u_k(\cdot, 0) \to u(\cdot, 0)$ in $\mathscr{H}^{-1/2}$, and $u_k \to u$ in graph norm.

3. Resolvent kernels. The analysis here is based on Stone's theorem for the construction of the spectral family of a self-adjoint operator A in a Hilbert space $\mathscr{H}$ with inner product $(,)$. Let $R(\lambda) = (A - \lambda I)^{-1}$ (the resolvent of A) and let $E(\lambda)$ be the (right continuous) spectral family of A. Then for (a, b) a finite interval and for f, g in $\mathscr{H}$,

$$\begin{aligned} &\left(\frac{(E(b) + E(b-))f}{2} - \frac{(E(a) + E(a-))f}{2}, g \right) \\ &\quad = \lim_{\varepsilon \to 0} \int_a^b ((R(k + i\varepsilon) - R(k - i\varepsilon))f, g) \frac{dk}{2\pi i}. \end{aligned} \tag{3.1}$$

Using the well-known relations (* signifies adjoint operator),

$$\begin{aligned} &R^*(\lambda) = R(\bar{\lambda}), \\ &R(\lambda_1) - R(\lambda_2) = (\lambda_1 - \lambda_2) R(\lambda_1) R(\lambda_2). \end{aligned} \tag{3.2}$$

Using the second equation of (3.2), the integral of the right-hand side of (3.1) may be rewritten as

$$\lim_{\varepsilon \to 0^+} \int_a^b (R(k - i\varepsilon)f, R(k - i\varepsilon)g) \, dk(\varepsilon/\pi). \tag{3.3}$$

Taking $f = g$ and using the first equation of (3.2) we have

$$\left(\frac{(E(b) + E(b-))}{2} f - \frac{(E(a) + E(a-))}{2} f, f \right) = \lim_{\varepsilon \to 0^+} \frac{\varepsilon}{\pi} \int_a^b |R(k - i\varepsilon)f|^2 \, dk \tag{3.4}$$

with $|\cdot|$ representing the norm in $\mathcal{H}$. Equation (3.4) gives (2.1) upon polarization. Therefore, we seek to compute (3.4) for A^3 and A^2. When it is not necessary to distinguish these operators we simply write A.

We will need the Fourier transform. On $\mathcal{S}(\mathbb{R}^n, \mathbb{C}^m)$, the space of smooth, rapidly decreasing $\mathbb{C}^m$-valued functions on $\mathbb{R}^n$, the Fourier transform is defined ($x \circ y = \sum x_i y_i$) as:

$$\Phi_n f(p) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-ix\circ p} f(x)\, dx \tag{3.5}$$

with $\Phi_n^{-1} = \Phi_n^*$ defined by

$$(\Phi_n^{-1} f)(p) = (\Phi_n f)(-p). \tag{3.6}$$

Φ_n is an isomorphism on $\mathcal{S}$ which extends by duality to $\mathcal{S}'$ the continuous dual of $\mathcal{S}$ and by continuity to $L^2(\mathbb{R}^n, \mathbb{C}^m)$ (see [R], for example). We will employ the notation $\mathcal{H}$ for $L^2(\mathbb{R}^3_+, \mathbb{C}^7)$. Now, using Parseval's formula in the case of Φ_3, (3.4) may be written (here and below, χ_c is the characteristic function of the set c) as

$$\lim_{\varepsilon \to 0^+} \frac{\varepsilon}{\pi} \int_a^b |\Phi_3(\chi_{R^3_+} R(k - i\varepsilon) f)|^2\, dk \tag{3.7}$$

$$= \lim_{\varepsilon \to 0^+} \int_{R^3} \frac{\varepsilon}{\pi} \int_a^b |(\Phi_3 \chi_{R^3_+} R(k - i\varepsilon) f)(p)|^2\, dk\, dp. \tag{3.8}$$

We first wish to obtain

$$\frac{\varepsilon}{\pi} (\Phi_3 \chi_{\mathbb{R}^3} R(k - i\varepsilon) f)(p) \tag{3.9}$$

in a form which can be studied as $\varepsilon \to 0+$. To this end, we need to compute the "resolvent kernel" of $R(\lambda)$. This is a function $R(x, y; z)$ such that for f in $\mathcal{H}$,

$$R(z) f(x) = \int_{\mathbb{R}^3_+} R(x, y; z) f(y)\, dy. \tag{3.10}$$

The idea is to seek $R(x, y; z)$ in the form

$$\mathcal{E}(x - y; z) - F(x, y; z), \tag{3.11}$$

where $\mathcal{E}(x - y; z)$ is a solution in $\mathcal{S}'$ of

$$(A(D) - zI)\mathcal{E}(x; z) = \delta(x) I_{10\times 10}, \tag{3.12}$$

and F satisfies the three conditions:

$$(A(D) - zI) F(x, y; z) = 0, \qquad x, y \in \mathbb{R}^3_+ \text{ (differentiation on } x), \tag{3.13}$$

$$G^i_{\lambda,j} F(x_1, x_2, 0, y; z) = G^i_{\lambda,j} \mathcal{E}(x - y; z)|_{x_3 = 0}, \qquad y \in \mathbb{R}^3_+, \tag{3.14}$$

$$\int_{\mathbb{R}^3_+} F(x, y; z) f(y)\, dy \text{ is in } \mathcal{H} \text{ for } f \text{ in } \mathcal{H}. \tag{3.15}$$

Let us define $A(p)$ to be $\sum A_j p_j$ for all nonzero p in $\mathbb{R}^3$. Then it is clear from our definition of Φ_3 that in $\mathcal{S}'$,

$$\mathcal{E}(\cdot; z) = (2\pi)^{-3/2} \Phi_3^* (A(p) - zI)^{-1} \Phi_3. \tag{3.16}$$

Taking the Fourier transform Φ_2 (on (x_1, x_2)) of (3.13) and (3.14) results in a first-order initial value problem in x_3. To solve this, it is necessary to compute $\Phi_2 G^i_{\lambda,j}\mathscr{E}(x-y; z)|_{x_3=0}$. It is evident that we will need $\Phi_2\mathscr{E}(x-y; z)|_{x_3=0}$ explicitly. In $\mathscr{S}'$ this means the evaluation of the integral (im $(z) \neq 0$) (which may be regarded as a member of $\mathscr{S}'$ or it may be computed in the usual way, by insertion of an appropriate exponential factor $e^{-p_3\varepsilon}\chi_{(0,\infty)}(p_3)$, for example, then letting $\varepsilon \to 0$)

$$(3.17) \qquad (2\pi)^{-2} e^{-iy'\circ n} \int_{-\infty}^{\infty} e^{-iy_3 p_3}[A(n, p_3) - z]^{-1}\, dp_3,$$

where we have used the notation $n = (p_1, p_2)$, $y' = (y_1, y_2)$. This will be done by means of the residue theorem through deforming the integration into the lower half plane. It is therefore necessary to consider the integrand as being extended as a function of p_3 into $\mathbb{C}$; n, z are not zero.

We write $\tau = p_3 + i\alpha$. We must consider the zeros of

$$(3.18) \qquad \det([A(n, \tau) - z])$$

in τ. These occur in the upper and lower half plane at values $\tau_\pm$, respectively. We consider the cases $A = A^3$ and $A = A^2$ separately now. The roots of $\det(A(p) - \lambda I)$ are given by (2.6), (2.7) above.

For $i = 2, 3$ and $j = 0, 1, 2, 3$ let ${}_iP_{\pm j}(p)$ be the associated eigenprojectors on $\mathbb{C}^{10}$ of $A^i(p)$. By the spectral theorem,

$$(3.19) \qquad [A^i(p) - z]^{-1} = \sum_{j=-3}^{3} ({}_i\lambda_j(p) - z)^{-1}\, {}_iP_j(p).$$

We wish to extend (in single-valued fashion) ${}_i\lambda_j(n, p_3)$ to ${}_i\lambda_j(n, \tau)$ and likewise ${}_iP_j(n, p_3)$ to ${}_iP_j(n, \tau)$ so that (3.19) remains valid, with all poles determined by the coefficients $({}_i\lambda_j(n, \tau) - z)^{-1}$. For ${}_i\lambda_{\pm 1}$, ${}_i\lambda_{\pm 2}$ and ${}_i\lambda_{\pm 3}$ we will make branchcuts in the τ plane (see Fig. 5) along the intervals $[(-i\infty, -i\sqrt{(2p_2^2 + p_1^2)}), (i\sqrt{(2p_2^2 + p_1^2)}, i\infty)]$, $[(-i\infty, -i\sqrt{(2p_2^2 + p_1^2)}), (i\sqrt{(2p_2^2 + p_1^2)}, i\infty)]$, $[(-i\infty, -in), (in, i\infty)]$, respectively, for A^2 and $[(-i\infty, -in/\sqrt{2}), (in/\sqrt{2}, i\infty)]$, $[(-i\infty, -in/\sqrt{2}), (in/\sqrt{2}, i\infty)]$, $[(-i\infty, -in), (in, i\infty)]$, respectively, for A^3.

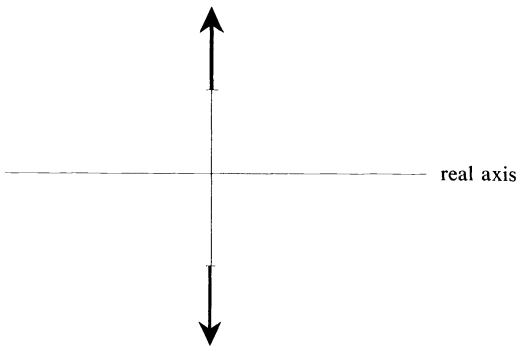


FIG. 5. *τ-plane branchcuts.*

The (τ) zeros of ${}_i\lambda_j(n,\tau)-z$ in the plane are given by

$$
\begin{aligned}
{}_2\tau_{\pm 1} &= \pm(z^2-(2p_2^2+p_1^2))^{1/2},\\
{}_2\tau_{\pm 2} &= \pm\frac{((5z^4-6p_2^2z^2+p_2^4)^{1/2}+3z^2-3p_2^2-2p_1^2)^{1/2}}{\sqrt{2}},\\
{}_2\tau_{\pm 3} &= \pm\frac{(-(5z^4-6p_2^2z^2+p_2^4)^{1/2}+3z^2-3p_2^2-2p_1^2)^{1/2}}{\sqrt{2}},
\end{aligned}
\tag{3.20}
$$

$$
\begin{aligned}
{}_3\tau_{\pm 1} &= \frac{\pm(z^2-n^2)^{1/2}}{\sqrt{2}},\\
{}_3\tau_{\pm 2} &= \frac{((z^4+6n^2z^2+n^4)^{1/2}+3z^2-3n^2)^{1/2}}{2},\\
{}_3\tau_{\pm 3} &= \frac{(-(z^4+6n^2z^2+n^4)^{1/2}+3z^2-3n^2)^{1/2}}{2}.
\end{aligned}
\tag{3.21}
$$

Here branchcuts are made for (3.20) $(A2)$ on the intervals

$$
\begin{aligned}
&[(-\infty,-\sqrt{(2p_2^2+p_1^2)}),(\sqrt{(2p_2^2+p_1^2)},\infty)],\\
&\left[\left(-\infty,\frac{-3n^2-(9n^4-4p_2^2n^2)^{1/2})^{1/2}}{\sqrt{2}}\right),\left(\frac{(3n^2-(9n^4-4p_2^2n^2)^{1/2})^{1/2}}{\sqrt{2}},\infty\right)\right],\\
&\left[\left(-\infty,\frac{(3n^2+(9n^4-4p_2^2n^2)^{1/2})^{1/2}}{\sqrt{2}}\right),\right.\\
&\left(\frac{(3n^2-(9n^4-4p_2^2n^2)^{1/2})^{1/2}}{\sqrt{2}},\frac{(3n^2-(9n^4-4p_2^2n^2)^{1/2})^{1/2}}{\sqrt{2}}\right)\\
&\left.\left(\frac{(3n^2+(9n^4-4p_2^2n^2)^{1/2})^{1/2}}{\sqrt{2}},\infty\right)\right],
\end{aligned}
\tag{3.22}
$$

respectively, and for (3.21) (A^3) we make the branchcuts (see Fig. 6)

$$
\begin{aligned}
&[(-\infty,-n),(n,\infty)]\\
&\left[\left(-\infty,-\left(\frac{3-\sqrt{5}}{2}\right)^{1/2}n\right),\left(\left(\frac{3-\sqrt{5}}{2}\right)^{1/2}n,\infty\right)\right]\\
&\left[\left(-\infty,-\left(\frac{3+\sqrt{5}}{2}\right)^{1/2}n\right),\left(-\left(\frac{3-\sqrt{5}}{2}\right)^{1/2}n,\left(\frac{3-\sqrt{5}}{2}\right)^{1/2}n\right),\left(\left(\frac{3+\sqrt{5}}{2}\right)^{1/2}n,\infty\right)\right].
\end{aligned}
\tag{3.23}
$$

It is easily verified that $\pm\operatorname{im}({}_i\tau_{\pm j})>0$. Using the residue theorem, we obtain for (3.17) the expression (see Fig. 7):

$$
-(2\pi i)^{-1}\,e^{-y'\circ n}\sum_{j=1}^{3}e^{iy_i\tau_j}\,{}_ic_j\,{}_iP_j(n,-{}_i\tau_j,z),
\tag{3.24}
$$

where the expression ${}_ic_j$ is determined by l'Hopital's rule as the reciprocal of

$$
\left.\frac{\partial_i\lambda_j}{\partial\tau}\right|_{\tau={}_i\tau_j}.
$$

The matrices ${}_iP_j(n,-{}_i\tau_j,z)$ are obtained from (3.19) by substitution of ${}_i\tau_j$ for p_3. They are given here (note that in ${}_2P_2$ and ${}_2P_3$, $a_1=s_{12}^2-z^2$, $a_2=s_{12}^2-2z^2$, $a_3=s_{12}^2-3z^2$,

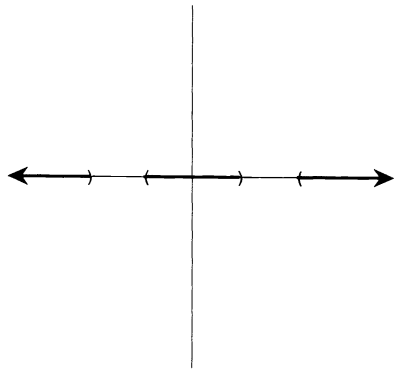


FIG. 6. *z-plane branchcuts (center cut used only for* λ_3*).*

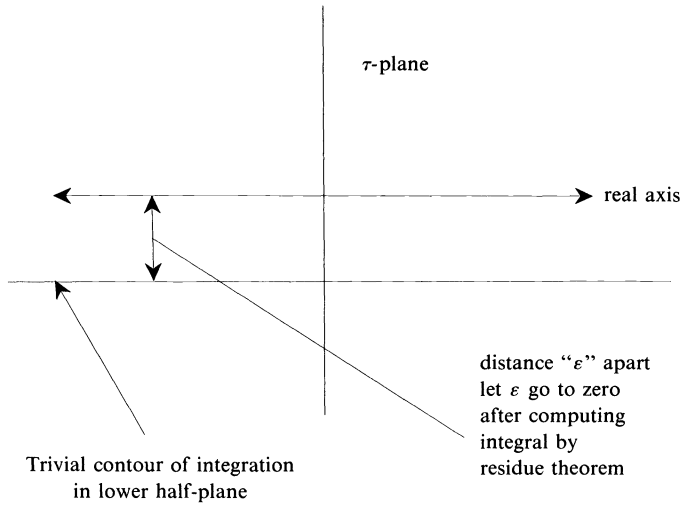


FIG. 7

$b_1 = s_{13}^2 - z^2$, $b_2 = s_{13}^2 - 2z^2$, $b_3 = s_{13}^2 - 3z^2$, $s_{ij} = {}_2\lambda_1({}_i\tau_j)$ and in ${}_3P_2$, ${}_3P_3$, $a_1 = n^2 + {}_3\tau_2^2 - z^2$, $a_2 = 2_3\tau_2^2 + n^2 - 2z^2$, $b_1 = z^2 - {}_3\tau_3^2 - n^2$, $b_2 = 2z^2 - 2_2\tau_3^2 - n^2$, the functions f_{ij} are normalization factors)(see Figs. 8-10).

We are able to write down the resolvent kernel now. First, we note that in the solution of (3.13), (3.14) we have

$$\Phi_{x'}F(n, x_3, y, z) = \frac{-1}{2\pi i} e^{-iy'\circ n} \sum_{j=1}^{3} e^{ix_3 {}_i\tau_j} {}_ic_j \, {}_iP_j \, {}_iM_j,$$

where the matrices ${}_iM_j$ are selected so that (3.14) is satisfied. Generally there are many possible choices for the ${}_iM_j$. The idea is to select the simplest among these for each of the boundary conditions. The ${}_iM_j$ are functions of z, t, n and are bounded except near points z where the so-called Lopatinski determinant vanishes. These (real) points yield the speeds of any surface waves. We discuss this further in the next section. We note that for $h_3 \neq 1$ or $h_2 \neq 1$, the development above is completely parallel except for the explicit formulas of the ${}_ic_j$ and ${}_iP_j$.

DEFINITION. For $k \neq j \neq 0$, let β be the set of points in p space where any ${}_i\lambda_j$ coincides with another ${}_i\lambda_k$. It is easy to see that this is a set of measure zero in p space.

$$
(3.25)\qquad {}_2P_1=\frac{1}{2(\tau_1^2+p_1^2)z^2}=\begin{bmatrix}
p_2^2\tau_1 & 0 & -p_1p_2^2\tau_1 & p_2\tau_1^2z & 0 & -p_1p_2^2\tau_1 & 0 & -p_1p_2^2\tau_1 & n_1^2p_2\tau_1 & -p_2^2\tau_1^2\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
-p_1p_2^2\tau_1 & 0 & p_1^2p_2^2 & -p_1p_2\tau_1z & 0 & p_1^2p_2z & 0 & p_1^2p_2^2 & -n_1^2p_1p_2 & p_1p_2^2\tau_1\\
p_2\tau_1^2z & 0 & -p_1p_2\tau_1z & \tau_1^2z^2 & 0 & -p_1\tau_1z^2 & 0 & -p_1p_2\tau_1z & n_1^2\tau_1z & -p_2\tau_1^2z\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
-p_1p_2\tau_1z & 0 & p_1^2p_2z & -p_1\tau_1z^2 & 0 & p_1^2z^2 & 0 & p_1^2p_2z & -n_1^2p_2z & (p_1p_2\tau_1z)\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
-p_1p_2^2\tau_1 & 0 & p_1^2p_2^2 & -p_1p_2\tau_1z & 0 & p_1^2p_2z & 0 & p_1^2p_2^2 & -n_1^2p_1p_2 & p_1p_2^2\tau_1\\
n_1^2p_2\tau_1 & 0 & -n_1^2p_1p_2 & n_1^2\tau_1z & 0 & -n_1^2p_1z & 0 & -n_1^2p_1p_2 & n_1^4 & -n_1^2p_2\tau_1\\
-p_2^2\tau_1^2 & 0 & p_1p_2^2\tau_1 & -p_2\tau_1^2z & 0 & p_1p_2\tau_1z & 0 & p_1p_2^2\tau_1 & -n_1^2p_2\tau_1 & p_2^2\tau_1^2
\end{bmatrix}.
$$

$$
(3.26)\qquad {}_2p_2=\frac{1}{{}_2f_2(z,n)}=\begin{bmatrix}
a_2^2p_1^2z^2 & a_1a_2p_1p_2z^2 & a_2^2p_1\tau_2z^2 & a_2a_3p_1^2p_2z & a_1a_2^2p_1z & a_2a_3p_1p_2\tau_2z & a_1a_2p_1z^3 & -a_2p_1\tau_2z^4 & 0 & a_2p_1^2z^4\\
a_1a_2p_1p_2z^2 & a_1^2p_2^2z^2 & a_1a_2p_2\tau_2z^2 & a_1a_3p_1p_2^2z & a_1^2a_2p_2z & a_1a_3p_2^2\tau_2z & a_1^2p_2z^3 & -a_1p_2\tau_2z^4 & 0 & a_1p_1p_2z^4\\
a_2^2p_1z_2a^2 & a_1a_2p_2\tau_2z^2 & a_2^2\tau_2^2z^2 & a_2a_3p_1p_2\tau_2z & a_1a_2^2\tau_2z & a_2a_3p_2\tau_2^2z & a_1a_2\tau_2z^3 & -a_2\tau_2^2z^4 & 0 & a_2p_1\tau_2z^4\\
a_2a_3p_1^2p_2z & a_1a_3p_1p_2^2z & a_2a_3p_1p_2\tau_2z & a_3^2p_1^2p_2^2 & a_1a_2a_3p_1p_2 & a_3^2p_1p_2^2\tau_2 & a_1a_3p_1p_2z^2 & -a_3p_1p_2\tau_2z^3 & 0 & a_3p_1^2p_2z^3\\
a_1a_2^2p_1z & a_1^2a_2p_2z & a_1a_2^2\tau_2z & a_1a_2a_3p_1p_2 & a_1^2a_3^2 & a_1a_2a_3p_3\tau & a_1^2a_2z^2 & -a_1a_2\tau_2z^3 & 0 & a_1a_2p_1z^3\\
a_2a_3p_1p_2\tau_2z & a_1a_3p_2^2\tau_2z & a_2a_3p_2\tau_2^2z & a_3^2p_1p_2^2\tau_2 & a_1a_2a_3p_2\tau_2 & a_3^2p_3^2\tau_2^2 & a_1a_3p_2\tau_2z^2 & -a_3p_2\tau_2^2z^3 & 0 & a_3p_1p_2\tau_2z^3\\
a_1a_2p_1z^3 & a_1^2p_2z^3 & a_1a_2\tau_2z^3 & a_1a_3p_1p_2z^2 & a_1^2a_2z^2 & a_1a_3p_2\tau_2z & a_1\tau_2z^5 & -a_1\tau_2z^5 & 0 & a_1p_1z^5\\
-a_2p_1\tau_2z^4 & -a_1p_2\tau_2z^4 & -a_2\tau_2^2z^4 & -a_3p_1p_2\tau_2z^3 & -a_1a_2\tau_2z^3 & -a_3p_2\tau_2^2z^3 & -a_1\tau_2z^5 & \tau_2^2z^6 & 0 & -p_1\tau_2z^6\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
a_2p_1^2z^4 & a_1p_1p_2z^4 & a_1p_1\tau_2z^4 & a_3p_1^2p_2z^3 & a_3p_1\tau_2z^3 & a_3p_1p_2\tau_2z^3 & a_1p_1z^5 & -p_1\tau_2z^6 & 0 & p_1^2z^6
\end{bmatrix}
$$

FIG. 8

(3.27) $$ {}_2p_3 = \frac{1}{({}_2f_3(z))} = \begin{bmatrix} b_2^2p_1^2z^2 & b_1b_2p_1p_2z^2 & b_2^2p_1\tau_3z^2 & b_2b_3p_1^2p_2z & b_1b_2^2p_1z & {}_2b_3p_1p_2\tau_3z & -b_1b_2p_1z^3 & -b_2p_1\tau_3z^4 & 0 & b_2p_1^2z^4 \\ b_1b_2p_1p_2z^2 & b_1^2p_2^2z^2 & b_1b_2p_2\tau_3z^2 & b_1b_3p_1p_2^2z & b_1b_2p_2z & b_1b_3p_2^2\tau_3z & -b_1^2p_2z^3 & -b_1p_2\tau_3z^4 & 0 & b_1p_1p_2z^4 \\ b_2^2p_1\tau_3z^2 & b_1b_2p_2\tau_3z^2 & b_2^2\tau_3^2z^2 & b_2b_3p_1p_2\tau_3z & b_1b_2^2\tau_3z & b_2b_3p_2\tau_3^2z & -b_1b_2\tau_3z^3 & -b_2\tau_3^2z^4 & 0 & b_2p_1\tau_3 \\ b_2b_3p_1^2p_2z & b_1b_3p_1p_2^2z & b_2b_3p_1p_2\tau_3z & b_3^2p_1^2p_2^2 & b_1b_2b_3p_1p_2 & b_3^2p_1p_2^2\tau_3 & -b_1b_3p_1p_2z^2 & -b_3p_1p_2\tau_3z^3 & 0 & b_3p_1^2z^3 \\ b_1b_2^2p_1z & b_1^2b_2p_2z & b_1b_2^2\tau_3z & b_1b_2b_3p_1p_2 & b_1^2b_2^2 & b_1b_2b_3p_2\tau_3 & -b_1^2b_2z^2 & -b_1b_2\tau_3z^3 & 0 & b_1b_2p_1z^3 \\ b_2b_3p_1p_2\tau_3z & b_1b_3p_2^2\tau_3z & b_2b_3p_2\tau_3^2z & b_3^2p_1p_2^2\tau_3 & b_1b_2b_3p_2\tau_3 & b_3^2p_2^2\tau_3^2 & -b_1b_3p_2\tau_3z^2 & -b_1p_2\tau_3^2z^3 & 0 & b_3p_1p_2\tau_3z^3 \\ -b_1b_2p_1z^3 & -b_1^2p_2z^3 & -b_1b_2\tau_3z^3 & -b_1b_3p_1p_2z^2 & -b_1^2b_2z^2 & -b_1b_3p_2\tau_3z^2 & b_1^2z^4 & b_1\tau_3z^5 & 0 & -b_1p_1z^5 \\ -b_2p_1\tau_3z^4 & -b_1p_2\tau_3z^4 & -b_2\tau_3^2z^4 & -b_3p_1p_2\tau_3z^3 & -b_1b_2\tau_3z^3 & -b_3p_2\tau_3^2z^3 & b_1\tau_3z^5 & \tau_3^2z^6 & 0 & -p_1\tau_3z^6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b_2p_1^2z^4 & b_1p_2p_2z^4 & b_2p_1\tau_3z^4 & b_1b_2p_1z^3 & b_3p_1p_2z^3 & b_3p_1p_2\tau_3z^3 & -b_1p_1z^5 & -p_1\tau_3z^6 & 0 & p_1^2z^6 \end{bmatrix} $$

(3.28) $$ {}_3f_1(z)\,{}_3P_1 = \begin{bmatrix} a_2^2\tau_1^2 & -p_1p_2\tau_1^2 & 0 & p_2^2\tau_1z & -p_1p_2\tau_1z & 0 & 0 & p_1p_2\tau_1^2 & p_2^2\tau_1^2 & -n^2p_2\tau_1 \\ -p_1p_2\tau_1^2 & p_1^2\tau_1^2 & 0 & -p_1p_2\tau_1z & p_1^2\tau_1z & 0 & 0 & -p_1^2\tau_1^2 & -p_1p_2\tau_1^2 & n^2p_2\tau_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_2^2\tau_1z & -p_1p_2\tau_1z & 0 & p_2^2z^2 & -p_1p_2z^2 & 0 & 0 & p_1p_2\tau_1z & p_2^2\tau_1z & -n^2p_2z \\ -p_1p_2\tau_1z & p_1^2\tau_1z & 0 & -p_1p_2z^2 & p_1^2z^2 & 0 & 0 & -p_1^2\tau_1z & -p_1p_2\tau_1z & n^2p_1z \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_1p_2\tau_1^2 & -p_1^2\tau_1^2 & 0 & p_1p_2\tau_1z & -p_1^2\tau_1z & 0 & 0 & p_1^2\tau_1^2 & p_1p_2\tau_1^2 & -n^2p_1\tau_1 \\ p_2^2\tau_1^2 & -p_1p_2\tau_1^2 & 0 & p_2^2\tau_1z & -p_1p_2\tau_1z & 0 & 0 & p_1p_2\tau_1^2 & p_2^2\tau_1^2 & -n^2p_2\tau_1 \\ -n^2p_2\tau_1 & n^2p_1\tau_1 & 0 & -n^2p_2 & n^2p_1z & 0 & 0 & -n^2p_2\tau_1 & -n^2p_2\tau_1 & n^4 \end{bmatrix} $$

FIG. 9

$$(3.29)\quad \frac{1}{{}_3f_2(z)}\,{}_3P_2 = \begin{bmatrix}
a_1^2a_2^2p_1^2 & a_1^2a_2^2p_1p_2 & a_1a_2n^2p_1\tau_2z^2 & -a_1a_2^2p_1^2\tau_2z & -a_1a_2^2p_1p_2\tau_2z & a_1a_2^2n^2p_1z & -a_1a_2n^2p_1z^3 & a_1^2a_2p_1p_2z^2 & -a_1a_2^2p_1^2(\tau_2^2+n^2) & 0\\
a_1^2a_2^2p_1p_2 & a_1^2a_2^2p_2^2 & a_1a_2n^2p_1\tau_2z^2 & -a_1a_2^2p_1p_2\tau_2z & -a_1a_2^2p_2^2\tau_2z & a_1a_2^2n^2p_2z & -a_1a_2n^2p_2z^3 & a_1^2a_2p_2^2z^2 & -a_1a_2^2p_1p_2(\tau_2^2+n^2) & 0\\
a_1a_2n^2p_1\tau_2z^2 & a_1a_2n^2p_2\tau_2z^2 & n^4\tau_2^2z^4 & -a_2n^2p_1p_2\tau_2^2z^3 & -a_2n^2p_2\tau_2^2z^3 & a_2n^4\tau_2z^3 & -n^4\tau_2z^5 & a_1n^2p_2\tau_2z^4 & -a_2n^2p_1\tau_2(\tau_2^2+n^2)z^2 & 0\\
-a_1a_2^2p_1^2\tau_2z & -a_1a_2^2p_1p_2\tau_2z & -a_2n^2p_1\tau_2^2z^3 & a_2^2p_1^2\tau_2^2z^2 & a_2^2p_1p_2\tau_2^2z^2 & -a_2^2n^2p_1\tau_2z^2 & a_1n^2p_1\tau_2z^4 & -a_1a_2p_1p_2\tau_2z^3 & a_2^2p_1^2\tau_2(\tau_2^2+n^2)z & 0\\
-a_1a_2^2p_1p_2\tau_2z & -a_1a_2^2p_2^2\tau_2z & -a_2n^2p_2\tau_2^2z^3 & a_2^2p_1p_2\tau_2^2z^2 & a_2^2p_2^2\tau_2^2z^2 & -a_2^2n^2p_2\tau_2z^2 & a_2n^2p_2\tau_2z^4 & -a_1a_2p_2^2\tau_2z^3 & a_2^2p_1p_2\tau_2(\tau_2^2+n^2)z & 0\\
a_1a_2^2n^2p_1z & a_1a_2^2n^2p_2z & a_2n^4\tau_2z^3 & -a_2^2n^2p_1\tau_2z^2 & -a_2^2n^2p_2\tau_2z^2 & a_2^2n^4z^2 & -a_2n^4z^4 & a_1a_2n^2p_2z^3 & -a_2^2n^2p_1(\tau_2^2+n^2)z & 0\\
-a_1a_2n^2p_1z^3 & -1_1a_2n^2p_2z^3 & -n^4\tau_2z^5 & a_2n^2p_1\tau_2z^4 & a_2n^2p_2\tau_2z^4 & -a_2n^4z^4 & n^4z^6 & -a_1n^2p_2z^5 & a_2n^2p_1(\tau_2^2+n^2)z^3 & 0\\
a_1^2a_2p_1p_2z^2 & a_1a_2p_2^2z^2 & a_1n^2p_2\tau_2z^4 & -a_1a_2p_1p_2\tau_2z^3 & -a_1a_2p_2^2\tau_2z^3 & a_1a_2n^2p_2z^3 & -a_1n^2p_2z^5 & a_1^2p_2^2z^4 & -a_1a_2p_1p_2(\tau_2^2+n^2)z^2 & 0\\
-a_1a_2^2p_1^2(\tau_2^2+n^2) & -a_1a_2^2p_1p_2(\tau_2^2+n^2) & -a_2n^2p_1\tau_2(\tau_2+n^2)z^2 & a_2^2p_1^2\tau_2(\tau_2^2+n^2)z & a_2^2p_1p_2\tau_2(\tau_2^2+n^2)z & -a_2^2n^2p_1(\tau_2^2+n^2)z & a_2n^2p_1(\tau_2^2+n^2)z^3 & -a_1a_2p_1p_2(\tau_2^2+n^2)z^2 & a_2^2p_1^2(\tau_2^2+n^2)^2 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

$$(3.30)\quad {}_3f_3(z)\,{}_3P_3 = \begin{bmatrix}
b_1^2b_2^2p_1^2 & b_1^2b_2^2p_1p_2 & b_1b_2n^2p_1\tau_3z^2 & b_1b_2^2p_1^2\tau_3z & b_1b_2^2p_1p_2z & -b_1b_2^2n^2p_1z & -b_1b_2n^2p_1z^3 & -b_1^2b_2p_1p_2z^2 & b_1b_2^2p_1^2(\tau_3^2+n^2) & 0\\
b_1^2b_2^2p_1p_2 & b_1^2b_2^2p_2^2 & b_1b_2n^2p_2\tau z^2 & b_1b_2^2p_1p_2\tau_3z & b_1b_2^2p_2^2\tau_3z & -b_1b_2^2n^2p_2z & -b_1b_2n^2p_2z^3 & -b_1^2b_2p_2^2z^2 & b_1b_2^2p_1p_2(\tau_3^2+n^2) & 0\\
b_1b_2n^2p_1\tau_3z^2 & b_1b_2n^2p_2\tau_3z^2 & n^4t_3^2z^4 & b_2n^2p_1\tau_3^2z^3 & b_2n^2p_2\tau_3^2z^3 & -b_2n^4\tau_3z^3 & -n^4\tau_3z^5 & -b_1n^2p_2\tau_3z^4 & b_2n^2p_1\tau_3(\tau_3^2+n^2)z^2 & 0\\
b_1b_2^2p_1^2\tau z & b_1b_2^2p_1p_2\tau_3z & b_2n^2p_1\tau_3^2z^3 & b_2^2p_1^2\tau_3^2z^2 & b_2^2p_1p_2\tau_3^2z^2 & -b_2^2n^2p_1\tau_3z^2 & -b_2n^2p_1\tau_3z^4 & -b_1b_2p_1p_2\tau_3z^3 & b_2^2p_1^2\tau_3(\tau_3^2+n^2)z & 0\\
b_1b_2^2p_1p_2\tau_3z & b_1b_2^2p_2^2\tau_3z & b_2n^2p_2\tau_3^2z^3 & b_2^2p_1p_2\tau_3^2z^2 & b_2^2p_2^2\tau_3^2z^2 & -b_2^2n^2p_2\tau_3z^2 & -b_2n^2p_2\tau_3z^4 & -b_1b_2p_2^2\tau_3z^3 & b_2^2p_1p_2\tau_3(\tau_3^2+n^2)z & 0\\
-b_1b_2^2n^2p_1z & -b_1b_2^2n^2p_2z & -b_2n^4\tau_3z^3 & -b_2^2n^2p_1\tau_3z^2 & -b_2^2n^2p_2\tau_3z^2 & -b_2^2n^4z^2 & b_2n^4z^4 & b_1b_2n^2p_2z^3 & -b_2^2n^2p_1(\tau_3^2+n^2)z & 0\\
-b_1b_2n^2p_1z^3 & -b_1b_2n^2z_2z^3 & -n^4\tau_3z^5 & -b_2n^2p_1\tau_3z^4 & -b_2n^2p_2\tau_3z^4 & b_2n^4z^4 & n^4z^6 & b_1n^2p_2z^5 & -b_2n^2p_1(\tau_3^2+n^2)z^3 & 0\\
-b_1^2b_2p_1p_2z^2 & -b_1^2b_2p_2^2z^2 & -b_1n^2p_2\tau_3z^4 & -b_1b_2p_1p_2\tau_3z^3 & -b_1b_2p_2^2\tau_3z^3 & b_1b_2n^2p_2z^3 & b_1n^2p_2z^5 & b_1^2p_2^2z^4 & -b_1b_2p_1p_2(\tau_3^2+n^2)z^2 & 0\\
b_1b_2^2p_1^2(\tau_3+n^2) & b_1b_2^2p_1p_2(\tau_3^2+n^2) & b_2n^2p_1\tau_3(\tau_3^2+n^2)z^2 & b_2^2p_1^2\tau_3(\tau_3^2+n^2)z & b_2^2p_1p_2\tau_3(\tau_3^2+n^2)z & -b_2^2p_1(\tau_3^2+n^2)z & -b_2n^2p_1(\tau_3^2+n^2)z^3 & -b_1b_2p_1p_2(\tau_3^2+n^2)z^2 & b_2^2p_1(\tau_3^2+n^2)^2 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

FIG. 10

We may now write down $\Phi_3 \chi_{\mathbb{R}^3_+} R(p, y, z)$ by applying Φ_1 in x_3 to (3.31), using (3.11) to get

$$
\begin{aligned}
&(2\pi)^{-3/2} e^{-iy'\circ n}(e^{-iy_3 p_3}[A^3(p) - zI]^{-1} + \sum_{j=1}^{3} \frac{e^{iy_3 {}_3\tau_j}}{({}_3\tau_j + p_3)} {}_3c_j \, {}_3P_j(n, -{}_3\tau_j, z)) \\
&\quad + 2\pi e^{-iy'\circ n} \sum_{j=1}^{3} \left(\frac{{}_3c_j}{{}_3\tau_j - p_3}\right) {}_3P_j(n, {}_3\tau_j, z) {}_3M_j
\end{aligned} \tag{3.32}
$$

with a completely analogous expression for A^2. Here it is helpful to note that the functions ${}_ic_j$ are not singular. For later reference we also note the facts:

$$
\begin{aligned}
&\text{(i)} \quad \lim_{z \to {}_i\lambda_{kj}(p) \pm i0} {}_i\tau_j(p', z) = \pm k |p_3|, \qquad k = \pm 1, \\
&\text{(ii)} \quad \lim_{z \to {}_i\lambda_{kl}(p) \pm i0} {}_i\tau_j(p', z) \neq \pm k |p_3|, \qquad l \neq j.
\end{aligned} \tag{3.33}
$$

4. Eigenfunction expansions. In the computation of the spectral families of the various operators A^i_λ arising from the different combinations of external fields and boundary conditions, the (first-order) singularities of the resolvent kernel give rise to the terms of the spectral family. These singularities include the eigenvalues ${}_i\lambda_j$ but may also include singularities of the matrices ${}_iM_j$. The singularities of the ${}_iM_j$ are the surface wave speeds and may be computed directly from (3.14). This reduces to the search for real zeros (in z) of the Lopatinski determinant [Wa]. This is defined as follows.

DEFINITION 4.1 (A^3). The Lopatinski determinant is the family of determinants $\det[G_3P_1^j, G_3P_2^k, G_3P_3^l]$. A number $s(n, z)$ is a zero of the Lopatinski determinant if it is a zero for each member of the above family. Here, G is fixed to be one of the A^3 boundary conditions (2.11) and $G_3P_k^j$ is the jth column of G_3P_k. The definition for A^2 is entirely similar.

THEOREM 4.2. *The Lopatinski determinant has no real zeros for either A^3 or A^2, and hence neither of these supports surface waves.*

The proof is rather tedious but is just a matter of finding one of each family of determinants that has no real zeros. We give the computation for A^2 as an example. The matrices G_2P_j, $j = 1, 2, 3$, respectively, are given as in Fig. 11. The Lopatinski determinant is seen to be essentially $a_2 + b_2$. This has no real zeros. In fact, we may give (up to a nice scalar factor determined by (3.14) and (3.31)) the matrices ${}_2M_j$, $j = 1, 2, 3$ (respectively) as

$$
\begin{aligned}
&\text{diag}\,(-1, a, 1, -1, a, 1, a, 1, 1, -1), \\
&\text{diag}\,(1, 1, -1, 1, 1, -1, 1, -1, a, 1), \\
&\text{diag}\,(1, 1, -1, 1, 1, -1, 1, -1, a, 1),
\end{aligned} \tag{4.1}
$$

where "a" means the entry is arbitrary. Combining this with the fact that (1.17) reduces to the Maxwell equations when $V = 0$, and results of [Sc2] together with the remarks above (cf. (2.11)) we have the following corollary.

COROLLARY 4.3. *The presence of surface waves is unstable in a liquid semiconductor modeled by* (1.17) *for either parallel or orthogonal external fields.*

DEFINITION 4.4.

$$
{}_i\psi_j^*(p, y, z^*) = ({}_i\lambda_j(p) - z^*)\chi_{\mathbb{R}^3 \setminus \beta}(p) \Phi_3 \chi_{\mathbb{R}^3_+} R(p, y, z^*), \tag{4.2}
$$

$$
{}_i\hat{f}_j(p, z) = \int_{\mathbb{R}^3_+} {}_i\psi_j^*(p, y, z) f(y)\, dy, \tag{4.3}
$$

$$G_2P_1=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ m_1^2p_2\tau_1 & 0 & -n_1^2p_1p_2 & n_1^2\tau_1z & 0 & -n_1^2p_1z & 0 & 0 & -n_1^2\tau_1p_2 & n_1^4 & -n_1^2p_2\tau_1 \end{bmatrix},$$

$$G_2P_2=\begin{bmatrix} a_1a_2^2p_1z & a_1^2a_2p_2z & a_1a_2^2\tau_2z & a_1a_2a_3p_1p_2 & a_1^2a_2^2 & a_1a_2a_3p_2\tau_2 & a_1^2a_2z^2 & -a_1a_2\tau_2z^3 & 0 & a_1a_2p_1z^3 \\ a_1a_2p_1z & a_1^2p_2z^3 & a_1a_2\tau_2z^3 & a_1a_3p_1p_2z^2 & a_1^2a_2z^2 & a_1a_3p_2\tau_2z^2 & a_1^2z^4 & -a_1\tau_2z^5 & 0 & a_1p_1z^5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$G_2P_3=\begin{bmatrix} b_1b_2^2p_1z & b_1^2b_2p_2z & b_1b_2^2\tau_3z & b_1b_2b_3p_2 & b_1^2b_2^2 & b_1b_2b_3p_2\tau_3 & -b_1^2b_2z^2 & -b_1b_2\tau_3z^3 & 0 & b_1b_2p_1z^3 \\ -b_1b_2p_1z^3 & -b_1^2p_2z^3 & -b_1b_2\tau_3z^3 & -b_1b_3p_1p_2z^2 & -b_1^2b_2z^2 & -b_1b_3p_2\tau_3z^2 & b_1^2z^4 & b_1\tau_3z^5 & 0 & -b_1p_1z^5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

FIG. 11

where f is smooth and has bounded support.

Set $k = \pm 1$ and

$$(4.4)\qquad {}_i\psi^*_{kj}(p, y) = \lim_{z \to {}_i\lambda_{kj}(p)+i0} {}_i\psi^*_j(p, y, z)$$

$$(4.5)\qquad \begin{aligned} &= (2\pi)^{-3/2}\chi_{\mathbb{R}^3\setminus\beta}(p)\chi_{\mathbb{R}_{-k}}(p_3)\,{}_iP_j(p) \\ &\quad \times \{e^{iy\circ p}I - {}_iM_j(n, {}_i\lambda_{kj}(p) - i0, y_3)\, e^{-iy'\circ n}\}. \end{aligned}$$

LEMMA 4.5. *If f is smooth and has compact support, then*

$$(4.6)\qquad {}_i\hat{f}_{\pm j}(p) = \lim_{z \to {}_i\lambda_{\pm j}(p)+i0} {}_j\hat{f}_j(p, z) = \int_{\mathbb{R}^3_+} {}_i\psi^*_{\pm j}(p, y) f(y)\, dy$$

defines a function which is smooth and rapidly decreasing almost everywhere.

Proof. Equations (4.2) and (4.5) show that the function on the left-hand side of (4.3) in this case converges by the definition of ${}_iM_j$ and the dominated convergence theorem as indicated. The fact that the Fourier transform of f is smooth and rapidly decreasing together with (4.5) gives the result.

Most of the results in this section are essentially independent of external field direction, at least in their statements. So that the notation does not become unwieldy, we will omit the front subscript from most expressions. The generalized Fourier transforms are defined by (4.6). These will also be denoted by expression Φ. Whether the ordinary or generalized transform is meant should be clear from the context.

LEMMA 4.6.

$$(4.7)\qquad \begin{aligned} &\lim_{\varepsilon\to 0} \int_{\mathbb{R}^3} \frac{\varepsilon}{\pi} \int_a^b |\Phi_3 \chi_{\mathbb{R}^3_+} R(k - i\varepsilon) f(p)|^2\, dk\, dp \\ &\quad = \int_{\mathbb{R}^3_+} \lim_{\varepsilon\to 0^+} \frac{\varepsilon}{\pi} \int_a^b |\Phi_3 \chi_{\mathbb{R}^3_+} R(k - i\varepsilon) f(p)|^2\, dk\, dp. \end{aligned}$$

There is no problem in switching the order of integration for positive e, since the integrand is continuous in k and measurable in p and nonnegative. The proof of this lemma is tedious but straightforward.

THEOREM 4.7.

$$(4.8)\qquad \lim_{\varepsilon\to 0^+} \frac{\varepsilon}{\pi} \int_a^b |\Phi_3 \chi_{\mathbb{R}^3_+} R(k - i\varepsilon) f(p)|^2\, dk = \sum_{j\neq 0} \chi_{\lambda_j(p)\in(a,b)}(p) |\hat{f}_j(p)|^2$$

for all f in the orthogonal complement of the null space of A^i.

THEOREM 4.8. *The modes of propagation are uncoupled for A^2 and for A^3 with type* II *boundary at $\lambda = \infty$. For A^3 with type* II *boundary at $\lambda = 0$, the quasi-Alfven mode is uncoupled.*

Proof of Theorem 4.7. We apply the classical elementary fact:

$$(4.9)\qquad \lim_{\varepsilon\to 0^+} \frac{\varepsilon}{\pi} \int_a^b \frac{f(x)}{(k-x)^2 + \varepsilon^2}\, dk = \chi_{(a,b)}(x) f(x)$$

for any continuous f.

For $p \notin \beta$, and δ small, the sets

$$(4.10)\qquad \Delta_j = (a, b) \cap (\lambda_j(p) - \delta, \lambda_j(p) + \delta)$$

are pairwise disjoint. Making the appropriate substitution for f above, we have $(z^* = k - i\varepsilon)$

$$(4.11)\qquad \sum_j \lim_{\varepsilon\to 0^+} \frac{1}{\pi}\int_{\Delta_j} \frac{\varepsilon}{(\lambda_j(p)-k)^2+\varepsilon^2}|(\lambda_j(p)-z^*)\Phi_3\chi_{\mathbb{R}^3_+}R(k-i\varepsilon)f(p)|^2\,dk$$

from which the result follows.

Proof of Theorem 4.8. Here we must use (3.14) to obtain the equations for M_j as follows:

$$(4.12)\qquad \sum_{j=1}^{3} G(e^{iy_3\tau_j}c_j(-\tau_j)P_j(n,-\tau_j,z)-c_jP_j(n,\tau_j,z)M_j)=0.$$

Now from (4.5) it follows that one mode is uncoupled from the others when M_j can be found for that j so that

$$(4.13)\qquad G(e^{iy_3\tau_j}c_j(-\tau_j)P_j(n,-\tau_j,z)-c_jP_j(n,\tau_j,z)M_j)=0.$$

The result for A^2 now follows from (4.1) and for A^3, $\lambda=\infty$,

$$\begin{aligned}&\text{diag}\,(1,1,a,-1,-1,a,a,1,1,-1),\\ &\text{diag}\,(1,1,-1,-1,-1,1,1,1,1,a),\\ &\text{diag}\,(1,1,-1,-1,-1,1,1,1,1,a),\end{aligned}\qquad(4.14)$$

and for $\lambda=0$ for

$$(4.15)\qquad \text{diag}\,(-1,-1,1,1,a,a,-1,-1,1).$$

To check that coupling occurs for the other boundary conditions is a straightforward computation and is omitted.

In order to continue, we must define the null spaces associated with the operators A^2 and A^3. These null spaces are determined, respectively, by the sums of two collections of orthoprojectors given by the pseudodifferential operator kernels

$$(4.16)\qquad {}_2P_{01}=\frac{1}{p^2}(0,0,0,p_1,p_2,p_3,0,0,0,0)\otimes(0,0,0,p_1,p_2,p_3,0,0,0,0),$$

$$(4.17)\qquad {}_2P_{02}=\frac{1}{p^2}(0,0,0,0,0,0,0,p_1,p_2,p_3)\otimes(0,0,0,0,0,0,0,p_1p_2,p_3),$$

$$(4.18)\qquad {}_2P_{03}={}_2m_3\otimes{}_2m_3,$$

$$(4.19)\qquad {}_2P_{04}={}_2m_4\otimes{}_2m_4,$$

where

$$(4.20)\qquad {}_2m_3=(-p^2p_2p_3,p^2p_1p_3,0,0,0,0,0,p_1p_2p_3^2,p_2^2p_3^2,-n^2p_2p_3)/({}_2\lambda_1 p|p_3|n),$$

$$(4.21)\qquad {}_2m_4=(-p_1^2p_3,-p_1p_2p_3,n^2p_1,0,0,0,0,-p_1p_2^2,p_1^2p_2,0)/({}_2\lambda_1 n|p_1|),$$

and for A^3

$$(4.22)\qquad {}_3P_{01}={}_2P_{01},$$

$$(4.23)\qquad {}_3P_{02}={}_2P_{02},$$

$$(4.24)\qquad {}_3m_3=(-p^2p_2,p^2p_1,0,0,0,0,0,p_1p_3^2,p_2p_3^2,-n^2p_3)/({}_3\lambda_1 np),$$

$$(4.25)\qquad {}_4m_4=(-p_1^2p_3,-p_1p_2p_3,n^2p_1,0,0,0,0,-p_1p_2p_3,p_1^2p_3,0)N({}_3\lambda_1 n|p_1|),$$

From this we see that the usual Maxwell divergence equations continue to hold for B and E' for both A^2 and A^3. The other two auxiliary conditions are more complex, relating E' and V.

Defining ${}_2P_0 = {}_2P_{01} + {}_2P_{02} + {}_2P_{03} + {}_2P_{04}$ and similarly for ${}_3P_0$, we obtain the following result.

THEOREM 4.9. *If $g, h \in (I - P_0)\mathscr{H}$ then*

$$(E(I)g, h) = \sum_{j \neq 0} \int_{\mathbb{R}^3} \chi_{(\lambda_j(p) \in I)}(p) \hat{g}_j(p)^* \hat{h}_j(p)\, dp, \tag{4.26}$$

where E is the spectral family for A and I is any subinterval of $\mathbb{R}$.

Proof. This follows from (4.9) and the polarization identity. (See Lemma 4.5 for the notation $\hat{}$.) We can define the generalized transforms now.

DEFINITION 4.10. For $g \in \mathscr{D}(\mathbb{R}^3_+)$ define

$$\Phi_j g(p) = \hat{g}_j(p) \tag{4.27}$$

and by Theorem 4.9 extend to all of $\mathscr{H}$. The adjoints of the maps Φ_j are given by

$$\Phi_j^* f(x) = \int_{\mathbb{R}^3} \psi_j^{**}(p, x) f(p)\, dp. \tag{4.28}$$

This follows easily for functions in $\mathscr{D}$ by definition and the general case follows by extension. The maps Φ_j yield the reduction of the unitary groups e^{-itA^i}. To check that they are orthogonal in the sense that the range of Φ_j^* is in the null space of Φ_j, $k \neq j$, suppose f is smooth and rapidly decreasing. Then the expression

$$\begin{aligned} \Phi_j f(r) &= \int_{\mathbb{R}^3_+} \psi_j^*(x, r) f(x)\, dx \\ &= (2\pi)^{-3/2} \chi_{\mathbb{R}^3 \setminus \beta} \chi_{\mathbb{R}}(r_3) P_j(r) \int_{\mathbb{R}^3_+} \{e^{-ix \circ r} I - M_j(r)\, e^{-ix' \circ r'}\} f(x)\, dx \end{aligned} \tag{4.29}$$

makes sense pointwise and further (here we have assumed $j, k > 0$),

$$\begin{aligned} \Phi_k^* g(x) &= \int_{\mathbb{R}^3} \psi_k^{**}(x, s) g(s)\, ds \\ &= (2\pi)^{-3/2} \int_{\mathbb{R}^3} \chi_{\mathbb{R}_-}(s_3) \chi_{\mathbb{R}^3 \setminus \beta}(s) \{e^{ixs} I - e^{ix' \circ s'} M_k^*(s) P_k(s) g(s)\, ds \end{aligned} \tag{4.30}$$

is a smooth rapidly decreasing function if g is, and if g vanishes in a neighborhood of β for a fixed p on a neighborhood of the set of s such that $\lambda_k(s) = \lambda_j(p)$. If F satisfies this condition, then $g(s) = F(s)/(\lambda_j(p) - \lambda_k(s))$ also satisfies the same condition.

Let $\mathscr{D}_1 = \{f \in \mathscr{D}(\mathbb{R}^3, \mathbb{C}^7) \mid \beta \cap \operatorname{supp}(f) = \varnothing\}$. Fix $F \in \mathscr{D}_1$ and $p \in \mathbb{R}^3$, and set $g(s)$ as above; $g \in \mathscr{D}_1$. Then $\Phi_k^* g$ is smooth, rapidly decreasing, and satisfies the boundary conditions and so is in the domain of A, and $A\Phi_k^* g = \Phi_k^* \lambda_k(\cdot) g$. Hence $\Phi_j A \Phi_k^* g = \Phi_j \Phi_k^* \lambda_k(\cdot) g$. But also $\Phi_j A \Phi_k^* g = \lambda_j(p) \Phi_j \Phi_k^* g(p)$. Subtracting, we obtain $\Phi_j \Phi_k^* F(p) = 0$. Since p is arbitrary and $\mathscr{D}_1$ is dense, this proves the required relation.

It follows from the preceding that the maps $[{}_i\Phi_j^*\, {}_i\Phi_j]$ are projections on $\mathscr{H}$.

In a similar way, we may show the spectral representation

$$e^{-itA^i} f = \sum_{j \neq 0} {}_i\Phi_j^*\, e^{-i|p|t}\, {}_i\Phi_j f. \tag{4.31}$$

The theory of potential scattering in a half-space may be studied for the operators A^i. We will not do this here. The interested reader may consult Theorems 3.8–3.10 of [S2], where this was done in the perfect conductor case. The method is entirely similar for the present problem. Instead, we take up the problem of when $\sigma = \sigma(x)$, or $\sigma = \sigma(t, x)$ is nonzero but decays at infinity in an appropriate sense. The case of σ that do not decay will be studied elsewhere.

5. Variable conductivity. We now wish to consider the problem of nonvanishing conductivity which may vary in space and/or time. First we consider the spatial variation only. We allow for possible anisotropy of the medium.

Assumption. Let $\sigma(x)$ be any two-tensor of dimension 3 whose components σ_{ij} $(i, j = 1, 2, 3)$ are almost everywhere uniformly bounded and satisfy the condition

$$|\sigma_{ij}(x)| < 0(|x|^{-1-\varepsilon}) \quad \text{as } |x| \to \infty \tag{5.1}$$

for some $\varepsilon > 0$.

We wish to study the operator determined by the right-hand side of (2.2) but where $B = i[\sigma_{ij}]$. We write $\Lambda(D, x)u = A(D)u + B(x)u$, where $A(D)$ is given by the first terms on the right side of (2.2). It is easily established that $\Lambda(D, x)$ is maximal dissipative in $L^2(\mathbb{R}^3, \mathbb{C}^{10})$. We will show that steady-state solutions of

$$-i\frac{\partial u}{\partial t} = \Lambda(D, x)u \tag{5.2}$$

exist in certain weighted spaces when the initial disturbance lies in the dual of the given weighted space. The interesting concept of "spectral barrier" arises here (see the appendix of [S3]).

We define the weighted spaces $L_{2,\alpha}(\mathbb{R}^3, \mathbb{C}^{10})$ as

$$L_{2,\alpha}(\mathbb{R}^3, \mathbb{C}^{10}) = \left\{ f \,\middle|\, \int_{\mathbb{R}^3} (1+|x|^2)^{\alpha} |f(x)|^2 \, dx < \infty, \, f : \mathbb{R}^3 \to \mathbb{C}^{10} \right\}.$$

It is noted here that the λ_j satisfy the "strongly propagative" hypothesis (λ_j is either bounded away from zero or is identically zero—see [Wi]).

The steady-state form of (5.2) at frequency λ is given by

$$A(D)u + B(x)u - \lambda u = f, \tag{5.3}$$

where f is assumed to belong to $L_{2,\alpha}$ with $\alpha > \frac{1}{2}$ and u is sought in a space $L_{2,-\beta}$, $\beta > \frac{1}{2}$. Note that $L_{2,-\beta} \supseteq L^2 \supseteq L_{2,\alpha}$. Note that B, considered as a multiplication operator, maps $L_{2,-\beta}$ to $L_{2,\alpha}$ if α and β are sufficiently close to $\frac{1}{2}$ (we will assume they are from now on). We will use the notation $\mathbb{C}^{\pm} = \{z \in \mathbb{C} \mid \pm\text{imaginary part of } z > 0\}$. Without loss of generality, we may assume $A(D) = A^3(D)$ by choice of coordinates since we are working in all of $\mathbb{R}^3$. Let $P_1 = I - {}_3P_0$, $P_0 = {}_3P_0$ in L^2. Assume $\lambda \in \mathbb{C}^{\pm}$ and operate on both sides of (5.2) with $(A(D) - \lambda I)^{-1}$ in the sense of L^2. This makes sense because $A(D)$ is self-adjoint. From (2.9) of [We] we may conclude that $P_1(A(D) - \lambda I)^{-1}$, thought of as mapping $L_{2,\alpha}$ to $L_{2,-\beta}$ is continuous in $\mathbb{C}^{\pm}$ and has continuous extensions $P_1(A(D) - \lambda I)_{\pm}^{-1}$ to the closure of $\mathbb{C}^+$ or $\mathbb{C}^-$ (i.e., down to or up to the real axis), that assume compact values as operators from $L_{2,\alpha}$ to $L_{2,-\beta}$. P_0 has a bounded extension to $L_{2,-\beta}$. We may "solve" for u now when λ is real as

$$u_{\pm}(x, \lambda) = (I - P_0(B/\lambda) + \lambda P_1(A(D) - \lambda I)_{\pm}^{-1}(B/\lambda))^{-1}(A(D) - \lambda I)_{\pm}^{-1} f.$$

The Fredholm theory (see [S4], for example) now allows us to say $u_{\pm}$ exists (in $L_{2,-\beta}$) when $|\lambda|$ is sufficiently large. (There may be some other exceptional values of λ besides

the "small" values for which a solution fails to exist—these form a countable nowhere dense set of linear measure zero [S4].) The difficulty for small λ is that the operator $(I - P_0(B/\lambda))^{-1}$ may not exist. In fact, using the explicit formula for P_0 given above, it is possible to construct examples exhibiting this difficulty. $u_\pm$ exists provided λ does not belong to the set of exceptional values or to the spectrum of P_0B (the spectrum of P_0B is the "spectral barrier."

For $\sigma = \sigma(t, x)$, a similar technique can be employed. We quote the following result from [S3, Thm. 4.2], adapted to the present situation.

THEOREM 5.1. *Suppose $B(t, x)$ is measurable in (t, x) and $t \to B(t, x)$ is a continuous map from $\mathbb{R}$ to the set of bounded operators on L^2. If $|B(t, x)| \leqq C(1+|t|)^{-1-\varepsilon}$ $(\varepsilon > 0)$ then for any $f(t, x)$ in the space $L_{2,\alpha}(\mathbb{R}, L^2(\mathbb{R}^3, \mathbb{C}^{10}))(\alpha > \frac{1}{2})$ there is a solution $u(t, x)$ of*

$$-i\frac{\partial u}{\partial t} = \Lambda(D, x)u + f(t, x)$$

in the space $L_{2,-\beta}(\mathbb{R}, L^2(\mathbb{R}^3, \mathbb{C}^{10}))$.

In fact, since the medium is a semiconductor, we may assume that C in the statement of Theorem 5.1 is small. In that case, the continuity hypothesis on B may be discarded (see Theorem 4.1 of [S3]).

Acknowledgment. The author would like to thank the referees for many helpful comments and corrections.

REFERENCES

[A] W. P. ALLIS, S. J. BUCHSBAUM, AND A. BERS, *Waves in Anisotropic Plasmas*, M.I.T. Press, Cambridge, MA, 1963.

[C] M. CUTLER, *Liquid Semiconductors*, Academic Press, New York, 1977.

[CH] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics, Vol.* II, Wiley, New York, 1963.

[J] F. JOHN, *Partial Differential Equations*, Fourth edition, Springer-Verlag, Berlin, 1982.

[K] W. B. KUNKEL, ED., *Plasma Physics in Theory and Application*, McGraw-Hill, New York, 1966.

[LL] L. D. LANDAU AND E. M. LIFSHITZ, *Electrodynamics of Continuous Media*, Addison-Wesley, New York, 1960.

[LP] P. LAX AND R. S. PHILLIPS, *Scattering Theory*, Academic Press, New York, 1967.

[R] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.

[S1] W. V. SMITH, *Boundary conditions for the MHD equations*, in Differential Equations and Applications, II, Ohio University Press, Columbus, OH, 1989, pp. 412-415.

[S2] ———, *Waves in a perfectly conducting fluid filling a half-space*, IMA J. Appl. Math., 43 (1989), pp. 47-69.

[S3] ———, *Average stability and decay properties of forced solutions of the wave propagation problems of classical physics in energy and mean norms*, J. Math. Anal. Appl., 143 (1989), pp. 148-186.

[S4] ———, *A local limiting absorption principle in a singular dispersive medium*, Quart. J. Appl. Math. Mech., 39 (1986), pp. 453-466.

[S5] ———, *Energy preserving boundary conditions for plasma in a half-space*, Proc. International Conference on Differential Equations, Columbus, OH, 1988, to appear.

[Sc1] J. R. SCHULENBERGER, *On conservative boundary conditions for operators of constant deficit: the Maxwell operator*, J. Math. Anal. Appl., 48 (1974), pp. 223-248.

[Sc2] ———, *Boundary waves on perfect conductors*, J. Math. Anal. Appl., 66 (1978), pp. 514-549.

[Wa] S. WAKABAYASHI, *Eigenfunction expansions for symmetric systems of first order in the half-space* $\mathbb{R}^n_+$, Publ. RIMS, Kyoto Univ. 11 (1975), pp. 67-147.

[We] R. WEDER, *Analyticity of the scattering matrix for wave propagation in crystals*, J. Math. Pures Appl., 64 (1985), pp. 121-148.

[Wi] C. H. WILCOX, *Asymptotic wave functions and energy distributions in strongly propagative media*, J. Math. Pures Appl., 57 (1978), pp. 275-321.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 379-391, March 1991

005

QUALITATIVE THEORY OF THE CAUCHY PROBLEM FOR A ONE-STEP REACTION MODEL ON BOUNDED DOMAINS*

JOEL D. AVRIN†

Abstract. A system of reaction-diffusion equations on a bounded domain arising as a model of laminar flames in a premixed reactive gas is considered. The equations couple the temperature T and the concentration Y subject to Arrhenius kinetics. After establishing the existence of unique global strong solutions for arbitrary nonnegative initial data in L^p, the bulk of the paper is devoted to an examination of the qualitative behavior of solutions subject to various boundary conditions, showing in particular that both T and Y remain bounded in most of the cases discussed.

In the no-flux case in which both T and Y satisfy zero Neumann boundary conditions, it is shown that if the average of the initial temperature over the domain is larger than ignition temperature, then eventually T is uniformly above ignition temperature and Y eventually decays exponentially to zero. This situation is called complete asymptotic burning, and an example is given to show that it does not always occur if the averaging condition is not met; if Q is the chemical heat release it is in fact shown in the case of equal diffusion coefficients that there exist parameters Q^* and Q_* with $Q_* \leqq Q^*$ such that complete asymptotic burning or eventual flame quenching occurs if, respectively, $Q > Q^*$ or $Q < Q_*$. In this case convergence to constant steady states is also established if $Q > Q^*$ or $Q < Q_*$.

For T satisfying fixed Dirichlet boundary conditions an example of complete asymptotic burning, and an example of flame quenching in at least portions of the domain are constructed. When Y satisfies fixed Dirichlet boundary condtions, cases are constructed where Y is bounded away from zero in portions of the domain if a certain parameter appearing in the Arrenius rate law is small enough.

Key words. laminar flames, boundary conditions, complete asymptotic burning, flame quenching

AMS(MOS) subject classifications. 35B40, 35K57, 80A25

1. Introduction. For the one-step reaction $A \to B$, let $Y = Y(x, t)$ denote the concentration of A and let $T = T(x, t)$ denote the temperature, where for a spatial domain Ω and time t we have $(x, t) \in \Omega \times \mathbb{R}^+$. The following reaction-diffusion equations model the dynamic interaction of T and Y subject to first-order Arrhenius kinetics:

$$T_t = d_0 \Delta T + QYf(T), \tag{1.1a}$$

$$Y_t = d_1 \Delta Y - Yf(T). \tag{1.1b}$$

Here d_0, d_1, and Q are positive constants, with d_0 and d_1 representing the thermal and mass diffusivities, respectively, while Q is proportional to the chemical heat release. For positive constants B and E, $f(T)$ is given by the usual Arrhenius rate law

$$f(T) = \begin{cases} 0, & T \leqq T_I, \\ B \exp\left(-E/(T - T_I)\right), & T > T_I \end{cases} \tag{1.2}$$

where the nonnegative constant T_I represents the ignition temperature. Here B is the preexponential factor and E is proportional to the activation energy. For a physical background of these equations, see Williams [22].

With f as in (1.2), the system (1.1) has appeared in a variety of contexts. For the case $\Omega = \mathbb{R}$, traveling wave solutions have been studied as models of propagating flame fronts. Existence of traveling wave solutions was established in [5] for $T_I > 0$ (for the case $T_I = 0$, see [13], [14], [20]). It is interesting to note that for Lewis number $L = d_1/d_0$ far from one, these solutions are known to be unstable. This fact was shown by formal

* Received by the editors July 5, 1989; accepted for publication (in revised form) April 11, 1990.
† Department of Mathematics, University of North Carolina, Charlotte, North Carolina 28223.

asymptotics in [18] and [6], and a rigorous proof has recently been found [21]. Meanwhile the solutions have been shown to be stable for L close to one [18], [6]. As traveling wave solutions are not stable in all cases, it is worthwhile to consider the Cauchy problem for (1.1) on $\Omega = \mathbb{R}$ with arbitrary initial data. Basic qualitative theory is obtained for the Cauchy problem for wide classes of initial data in [4], [12], and [17] with some overlap of results (see, e.g., §§ 2–5 of [4] and § 2 of [17]).

Additional qualitative theory is developed in §§ 6 and 7 of [4], in which the important case of ignition at one end only is treated. It is shown in § 6 of [4] that if the average of the initial temperature values at $+\infty$ and $-\infty$ is above ignition temperature, then on any ray coming from the ignition end the temperature is eventually uniformly above ignition temperature and the concentration decays uniformly to zero. As this ray can be continually enlarged by making t large enough, a rough sense of flame propagation is thus obtained. If the "averaging condition" on the initial temperature is not met, examples of both flame propagation or flame quenching can be constructed (see § 7 of [4]).

In this paper we consider the system (1.1) when Ω is a bounded domain in $\mathbb{R}^n$, $n \geqq 1$, with the usual smoothness assumptions on the boundary Γ. Various numerical studies exist [19] and for the steady-state problem for a porous pellet with exothermic reaction asymptotic solutions were derived and studied in [10]. We consider here the Cauchy problem for (1.1) with various boundary conditions and arbitrary nonnegative initial data in $L^p(\Omega)$ for any real $p > 1$. After establishing the existence and uniqueness of global strong solutions we will examine their qualitative behavior.

Consider first the case of zero Neumann boundary conditions for both T and Y, i.e., $\partial T/\partial \nu = \partial Y/\partial \nu = 0$ on Γ, where ν is the outward normal. Let T_{AV} denote the average of the initial temperature $T(x, 0) \equiv T_0(x)$ over Ω, i.e.,

$$T_{AV} = \frac{1}{|\Omega|} \int_\Omega T_0(x)\, dx \tag{1.3}$$

where $|\Omega|$ is the volume of Ω. We will show that if $T_{AV} > T_I$ then eventually T is uniformly above ignition temperature and Y eventually decays exponentially to zero. We will refer to this phenomenon as complete asymptotic burning. If $T_{AV} < T_I$, we will show that complete asymptotic burning still occurs if the initial data $Y_0(x)$ for Y is bounded from below by a positive constant and Q is large enough. In the case where $d_0 = d_1$ and with $T_{AV} < T_I$ we will show that eventual uniform flame quenching occurs (i.e., T is eventually uniformly below T_I) if Q is small enough. When $T_{AV} < T_I$, $d_0 = d_1$, and $Y_0(x)$ is bounded below by a positive constant we will show that there exist positive numbers Q_* and Q^* with $Q_* \leqq Q^*$ such that complete asymptotic burning occurs whenever $Q > Q^*$ while eventual flame quenching occurs if $Q < Q_*$.

The assumption $d_0 = d_1$ makes it easy for us to establish steady-state convergence results in these cases of complete asymptotic burning or flame quenching. If $T_{AV} > T_I$ not only does $Y(t)$ converge uniformly to zero, but $T(t)$ converges uniformly to the average of $T_0 + QY_0$ on Ω. If $T_{AV} < T_I$ and we add the assumption that $Y_0(x) \geqq \gamma$ for some $\gamma > 0$, then, as noted above, there exists a constant Q^* such that $Q > Q^*$ implies complete asymptotic burning and a constant Q_* such that $Q < Q_*$ implies eventual flame quenching; thus if $Q > Q_*$ the steady-state convergence is the same as in the case $T_{AV} > T_I$. If $Q < Q_*$ we will show that there exist constants T_1 and Y_1 such that $T(t)$ and $Y(t)$ converge to T_1 and Y_1, respectively; moreover, the averages of $T_1 + QY_1$ and $T_0 + QY_0$ over Ω are equal.

Our results for these boundary conditions closely parallel the results in [4] for the case $\Omega = \mathbb{R}$. In fact, our results here are somewhat better as we are able to show

that T is bounded as well as Y when $T_{AV} > T_I$, and we obtain a uniform exponential decay estimate for Y in that case.

Specializing to the case where $T_I = 0$ we note that unless T_0 is zero throughout Ω we have that $T_{AV} > 0$. Hence, whenever we have nontrivial solutions in the case $T_I = 0$, we have that $T_{AV} > T_I$, so that regardless of all other parameters (e.g., Q) $T(t)$ remains bounded for all $t \geqq 0$ and Y decays exponentially to zero. Our results in this case relate somewhat to those in [15], in which the examples treated include the system (1.1) with $T_I = 0$ in (1.2) and $Yf(T)$ is replaced by the more general case $Y^p f(T)$ where p is a positive integer. Marion shows that there exists a maximal attractor which attracts the bounded sets of $L^2(\Omega)$, and she estimates the fractal Hausdorff dimensions of this attractor (the attractor is compact in L^2 and bounded in the closed subset of H^1 appropriate for the boundary conditions). These results also hold in the case of periodic or zero Dirichlet boundary conditions, so there is also some relationship with our results described below.

Consider next the assumptions that T satisfies a fixed Dirichlet boundary condition and Y satisfies either a zero Neumann boundary condition or a fixed Dirichlet condition on Γ. The boundary condition on T corresponds physically to the case of an exothermic reaction in which the exterior heat source is fixed. Meanwhile the boundary conditions for Y correspond, respectively, to the no-flux case mentioned earlier and physical situations of which the permeable pellet as studied in [10] is a special case.

For these situations we will show that both T and Y remain bounded for all time; moreover, if the values of T on Γ are uniformly above ignition temperature and Y satisfies zero Neumann boundary conditions then complete asymptotic burning occurs. This latter condition imposed on the boundary values of T can occur, for example, in engines equipped with a precombustion chamber [19]. If on some nonempty open portions of Γ we have that T is below T_I we will show that if Q is small enough eventual flame quenching occurs in certain portions of Ω. This last result will be obtained regardless of the initial values of T and Y or the boundary conditions on Y.

In our final example we will allow either arbitrary fixed Dirichlet or Neumann conditions for T, and fixed nonzero Dirichlet conditions for Y. For each choice of initial conditions on T and Y we will show that if B is small enough then in portions of Ω Y is bounded uniformly above zero. Thus the rate at which Y burns is small enough to allow the boundary conditions to maintain a positive level for Y in subsets of Ω.

These latter results will be established in § 5 below. The exothermic case mentioned earlier will be handled in § 4, while the case of zero Neumann conditions for T and Y on Γ will be developed in § 3. We begin in § 2 with some preliminary observations and the construction of unique global strong solutions for (1.1).

2. Preliminaries and global existence. Let B_0, B_1 be a pair of boundary value operators; in particular, for $u \in C^2(\Omega)$, $B_i u = 0$ means that u satisfies either zero Dirichlet or zero Neumann boundary conditions, $i = 0, 1$. The indexing will correspond in what follows to the boundary conditions desired for T and Y, respectively. For $i = 0, 1$ set

$$C^2_{B_i} = \{u \in C^2(\Omega) \mid B_i u = 0\}.$$

We will correspondingly write $L^p(\Omega)$ and the usual Sobolev spaces $W^{m,p}(\Omega)$ in the abbreviated form L^p and $W^{m,p}$.

Each of the operators $d_0\Delta$ and $d_1\Delta$ is a continuous linear map from $C^2_{B_i}$ to $C(\Omega)$. It is well known that $C^2_{B_i}$ is dense in L^p for $1 < p < +\infty$ and that the closure of both

of the above operators is well defined and generates an analytic semigroup on L^p, $1<p<+\infty$. These facts are also true when $p=1$ (see, e.g., [3] where this is shown for a general class of second-order elliptic operators). Let A_0 and A_1 denote the closures of $d_0\Delta$ and $d_1\Delta$ in L^1 with domains $D(A_0)$ and $D(A_1)$. It is also shown in [3] that (for $i=0, 1$) $D(A_i)\cap L^p$ is the domain of the closure of the operator $d_i\Delta$ in L^p, $1<p<+\infty$, and A_i restricted to $D(A_i)\cap L^p$ (which we call $A_{i,p}$) equals this closure. We denote the domains $D(A_i)\cap L^p$ by $D_p(A_i)$; meanwhile we will denote the semigroups generated by $A_{0,p}$ and $A_{1,p}$ by $W_0(t)$ and $W_1(t)$. With $B_0T=B_1Y=0$ on Γ the system (1.1) then has the corresponding integral equations

$$T(t)=W_0(t)T_0+Q\int_0^t W_0(t-s)Y(s)f(T(s))\,ds, \tag{2.1a}$$

$$Y(t)=W_1(t)Y_0-\int_0^t W_1(t-s)Y(s)f(T(s))\,ds \tag{2.1b}$$

where $T_0=T(x,0)$ and $Y_0=Y(x,0)$ are in L^p. We have followed the usual convention of suppressing the dependence on x in writing (2.1).

The following existence result for $B_0T=B_1Y=0$ on Γ follows easily by considering the usual types of contraction maps on appropriate metric spaces defined by the right-hand sides of (2.2), once we note that $Yf(T)$ is a globally Lipschitz continuous map in Y and T by l'Hôpital's rule as follows.

THEOREM 2.1. *Let T_0 and Y_0 be in L^p for any $p\in(1,+\infty)$; then there exists an $S>0$ such that* (2.1a), (2.1b) *have unique solutions T, $Y\in C([0,S];L^p)$.*

COROLLARY 2.1. *The mild solutions found in Theorem* 2.1 *are, in fact, global solutions, i.e., T, $Y\in C([0,+\infty);L^p)$.*

Corollary 2.1, in particular, follows easily by exploiting the globally Lipschitz continuous properties of the nonlinearities and employing the usual contradiction argument on the hypothesis that finite-time blowup occurs. Meanwhile, the case of nonzero boundary conditions will be handled later in this section.

The next result establishes regularity for $t>0$; again the proof is straightforward, using standard semigroup techniques and the fact that every derivative of f is continuous and bounded.

THEOREM 2.2. *For each j, $k\geqq 1$ we have that T, $Y\in C^j([0,+\infty);C^k(\bar{\Omega}))\cap C([0,+\infty);L^p)$.*

The proof of the results above, in fact, follow directly from, or are simple adaptations of, the corresponding existence and regularity results found in [4] and [12].

The results summarized in the next theorem hold also when $\Omega=\mathbb{R}$ (see [4, § 2]) and easily follow from the maximum principle (see, e.g., [16]), (2.1a) and (1.2).

THEOREM 2.3. *Let T and Y be as above. If in addition T_0 and Y_0 are nonnegative, then $T(t)$ and $Y(t)$ remain nonnegative for all $t>0$. Also, if T_0, $Y_0\in C(\bar{\Omega})$, then for all $t>0$*

$$\|Y(t)\|_\infty\leqq\|Y_0\|_\infty, \tag{2.2a}$$

$$\|T(t)\|_\infty\leqq\|T_0\|_\infty+Q\|Y_0\|_\infty Bt. \tag{2.2b}$$

We now indicate how to modify Theorems 2.1–2.3 and Corollary 2.1 to handle cases with nonzero boundary conditions (if $n=1$ we do this only for Dirichlet conditions or with the appropriate matching conditions imposed in the Neumann case). Let g and h be measurable (e.g., continuous) functions on Γ such that the elliptic problem $\Delta w=0$ in Ω with respective boundary conditions $B_0w=g$ and $B_1w=h$ on Γ has a

unique strong solution. Denote these solutions by w_0 and w_1, respectively. The corresponding integral equations such that T and Y satisfy these boundary conditions are

$$T(t) = W_0(t)(T_0 - w_0) + w_0 + Q\int_0^t W_0(t-s)\,Y(s)f(T(s))\,ds, \tag{2.3a}$$

$$Y(t) = W_1(t)(Y_0 - w_1) + w_1 - \int_0^t W_1(t-s)\,Y(s)f(T(s))\,ds. \tag{2.3b}$$

Standard regularity theory (see, e.g., [7]) allows us to select g and h smooth enough so that w_0 and w_1 are in $C^k(\bar{\Omega})$ for any given $k \geqq 0$. Then if $W_0(t)T_0$ and $W_1(t)Y_0$ are replaced by $W_0(t)(T_0 - w_0) + w_0$ and $W_1(t)(Y_0 - w_1) + w_1$, respectively, the above existence and regularity techniques apply to (2.3) instead of (2.1) with only minor modification. In this way we obtain the following result.

THEOREM 2.4. *Given an integer $k \geqq 2$ select g and h as above such that $w_i \in C^k(\bar{\Omega})$, $i = 0, 1$. Then for each choice of T_0 and Y_0 in L^p for any $p \in (1, +\infty)$ there exist unique global strong solutions T and Y of (1.1) with $B_0 T = g$ and $B_1 Y = h$ on Γ such that T, $Y \in C^j((0, +\infty); C^k(\bar{\Omega})) \cap C([0, +\infty); L^p)$ for each $j \geqq 1$.*

From the maximum principle, (1.2), and (2.3) we can modify Theorem 2.3 to handle $B_0 T = g$ and $B_1 Y = h$ in the case where B_0 and B_1 are Dirichlet boundary operators on Γ as follows.

THEOREM 2.5. *Let T and Y be as in Theorem 2.4. Then if T_0, $Y_0 \geqq 0$ then $T(t)$, $Y(t) \geqq 0$ for all $t > 0$. If in addition T_0, $Y_0 \in C(\bar{\Omega})$, then for all $t > 0$ and g and h as in Theorem 2.4 with $g \geqq 0$ and $h \geqq 0$ we have that*

$$\| Y(t) \|_\infty \leqq \| Y_0 \|_\infty + \sup_\Gamma h, \tag{2.4}$$

$$\| T(t) \|_\infty \leqq \| T_0 \|_\infty + \sup_\Gamma g + QBt\left(\| Y_0 \|_\infty + \sup_\Gamma h \right). \tag{2.5}$$

3. Zero Neumann boundary conditions for T and Y. Let ν denote the outward normal on Ω; then we assume throughout this section that $B_0 = B_1 = \partial/(\partial\nu)$, i.e.,

$$\frac{\partial T}{\partial \nu} = \frac{\partial Y}{\partial \nu} = 0 \quad \text{on } \Gamma. \tag{3.1}$$

Let T_{AV} be as in (1.3). We first note a fairly standard fact about the Laplace operator with zero Neumann boundary conditions.

PROPOSITION 3.1. *If $B_0 T = \partial T/(\partial\nu) = 0$ on Γ we have that*

$$\lim_{t \to \infty} W_0(t) T_0 = T_{AV} \tag{3.2}$$

uniformly in x.

Proof. For $T_0 \in C^1(\bar{\Omega})$ the result follows from expanding $W_0(t)T_0$ in terms of the eigenfunctions of $-A_{0,2}$. The first term in the expansion is T_{AV}, corresponding to the first eigenfunction $1/|\Omega|$ with eigenvalue zero. The sum of the next terms decays uniformly to zero as $O(\exp(-\lambda_2 t))$, where λ_2 is the first nonzero eigenvalue of $-A_{0,2}$. For $T_0 \in L^p$ with $1 < p < +\infty$, we note that $W_0(t)$ conserves L^1-norms; this can be seen by integrating the heat equation satisfied by $W_0(t)T_0$ on both sides with respect to x, and then applying the divergence theorem (recall that T_0 is nonnegative and $W_0(t)$ is positivity preserving). Thus we can replace T_0 in (3.2) by any $T(t_1)$ with $t_1 > 0$. But $T(t_1)$ is $C^1(\bar{\Omega})$ by parabolic regularity, and thus the result holds for general $T_0 \in L^p$. □

The next proposition asserts that Y decays uniformly to zero exponentially if the temperature starts out uniformly above ignition temperature throughout Ω.

PROPOSITION 3.2. *Let T and Y be as in Theorem* 2.2 *such that* (3.1) *holds. Suppose that there exists an α such that $T_0(x) \geqq \alpha > T_I$ for all $x \in \Omega$. Then for $\beta = f(\alpha)$*

$$\| Y(t) \|_p \leqq \| Y_0 \|_p e^{-\beta t}, \qquad t \geqq 0. \tag{3.3}$$

If in addition $Y_0 \in C(\bar{\Omega})$, then

$$\| Y(t) \|_\infty \leqq \| Y_0 \|_\infty e^{-\beta t}, \qquad t \geqq 0. \tag{3.3a}$$

Proof. Since the integral part of (2.1a) is nonnegative for all $t \geqq 0$, we have that $T(x, t) \geqq \alpha$ for all $t \geqq 0$ and all $x \in \Omega$. Since $f(T)$ is strictly increasing in T for $T > 0$, we thus have that $f(T(x, t)) \geqq f(\alpha) = \beta$ for all $t \geqq 0$ and all $x \in \Omega$. Let $V(t)$ denote multiplication by $f(T(t))$ and let $U(t, s)$ be the fundamental solution generated by the (time-dependent) operator $A_{1,p} - V(t)$, i.e., $U(t, s)$ is the unique operator-valued function satisfying $\partial U(t, s)/(\partial t) = (A_{1,p} - V(t)) U(t, s)$, $U(s, s) = I$ for every $0 \leqq s \leqq t$; we have that $Y(t) = U(t, 0) Y_0$. If $U_\beta(t, s)$ is the fundamental solution generated by $A_{1,p} - V(t) + \beta I$, then

$$\| U_\beta(t, 0) Y_0 \|_p \leqq \| Y_0 \|_p. \tag{3.4}$$

Inequality (3.3) now follows since

$$U(t, 0) = U_\beta(t, 0) e^{-\beta t}. \tag{3.5}$$

If $Y_0 \in C(\bar{\Omega})$ we obtain (3.3a) by replacing p and ∞ in (3.4) and then using (3.5). □

Our main result of this section follows by combining the above propositions.

THEOREM 3.1. *Let T and Y be as in Theorem* 2.2 *such that* (3.1) *holds. Then if $T_{AV} > T_I$ we have that Y eventually decays to zero exponentially; that is, there exists a $t_1 > 0$ and a $\beta > 0$ such that for all $t \geqq t_1$*

$$\| Y(t) \|_p \leqq \| Y_0 \|_p e^{-\beta(t - t_1)}. \tag{3.6}$$

If in addition $Y_0 \in C(\bar{\Omega})$, then for all $t \geqq t_1$

$$\| Y(t) |_\infty \leqq \| Y_0 \|_\infty e^{-\beta(t - t_1)}. \tag{3.6a}$$

Proof. Let α be a constant such that $T_{AV} > \alpha > T_I$. By Proposition 3.1 there exists a $t_1 > 0$ such that $t \geqq t_1$ implies that $(W_0(t) T_0)(x) \geqq \alpha$ for all $x \in \Omega$. Again, from (1.2) and (2.1a) this implies that $T(x, t) \geqq \alpha$ for all $t \geqq t_1$ and all $x \in \Omega$. If we now consider (1.1) with initial data $T(t_1)$, $Y(t_1)$, the inequality (3.6) now follows from Proposition 3.2, noting that $\| Y(t_1) \|_p \leqq \| Y_0 \|_p$ since $U(t, s)$ (as in the proof of Proposition 3.2) is nonexpansive on L^p. If $Y_0 \in C(\bar{\Omega})$ (3.6a) follows similarly, making use of (2.2a). □

The next result asserts that if T_0, Y_0 are in $C(\bar{\Omega})$ then under the above conditions $T(t)$ (as well as $Y(t)$) is bounded for all t.

THEOREM 3.2. *Under the conditions of Theorem* 3.1 *assume in addition that T_0 and Y_0 are in $C(\bar{\Omega})$. Then there exists a constant K such that $\| T(t) \|_\infty \leqq K$ for all $t \geqq 0$.*

Proof. Let t_1 be as in Theorem 3.1. From (2.1a) and (3.6a) we have for $t \geqq t_1$ that

$$\begin{aligned} 0 \leqq T(t) &= W_0(t) T_0 + Q \int_0^{t_1} W_0(t - s) Y(s) f(T(s))\, ds \\ &\quad + Q \int_{t_1}^{t} W_0(t - s) Y(s) f(T(s))\, ds \\ &\leqq \| T_0 \|_\infty + Q \| Y_0 \|_\infty B t_1 + Q \int_{t_1}^{t} \| Y(s) \|_\infty B\, ds \\ &\leqq \| T_0 \|_\infty + Q \| Y_0 \|_\infty B \left[t_1 + \int_{t_1}^{\infty} e^{-\beta(s - t_1)}\, ds \right] \end{aligned} \tag{3.7}$$

$$= \| T_0 \|_\infty + Q \| Y_0 \|_\infty B[t_1 + (1/\beta)].$$

Thus the result is established with K taken as the right-hand side of (3.7). □

Note that Theorem 3.1 asserts that in the case where $T_I = 0$ we have complete asymptotic burning whenever T_0 is nonnegative in Ω and positive on a subset of positive measure in Ω. However, if $T_I > 0$, different results can occur depending on the value of other parameters. We illustrate below with two examples in which eventual flame extinction occurs in the first case for a small enough value of Q and in the second case complete asymptotic burning occurs if Q is large enough.

PROPOSITION 3.3. *Under the conditions of Theorem* 2.3 *with* T *and* Y *satisfying* (3.1), *suppose in addition that* $d_0 = d_1$ *and* $T_{AV} < T_I$; *then for small enough* Q *we have that* $T(t)$ *eventually stays below ignition temperature.*

Proof. From (2.1b) and the nonnegativity of Y and $f(T)$ we have for all $t \geqq 0$ that

$$0 \leqq \int_0^t W_1(t-s)\, Y(s) f(T(s))\, ds \leqq W_1(t)\, Y_0 \leqq \| Y_0 \|_\infty. \tag{3.8}$$

Since $d_0 = d_1$ we have that $W_0(t) = W_1(t)$, hence from (2.1a) and (3.8) it follows that

$$0 \leqq T(t) \leqq W_0(t)\, T_0 + Q \| Y_0 \|_\infty. \tag{3.9}$$

Given γ such that $T_{AV} < \gamma < T_I$, by Proposition 3.1 there exists a t_2 such that $t \geqq t_2$ implies that $W_0(t) T_0 \leqq \gamma$. Thus if Q is such that $Q \| Y_0 \|_\infty < T_I - \gamma$ we have from (3.9) that $T(t) < T_I$ for all $t \geqq t_2$. Thus the proposition is established. □

We next construct a case in which we can have $T_{AV} < T_I$ but complete asymptotic burning prevails if Q is large enough. To simplify the proof, we place some restrictions on the initial data. In this case no restrictions are placed on the diffusion coefficients.

PROPOSITION 3.4. *Under conditions* (3.1) *assume in addition that there exists a constant* $\gamma > 0$ *such that* $Y_0(x) \geqq \gamma$ *for all* $x \in \Omega$. *Assume also that there exists a constant* $\alpha > T_I$ *and a subset* E_0 *of* Ω *such that* E_0 *has positive measure and* $T_0(x) \geqq \alpha$ *for all* $x \in E_0$. *Then if* Q *is large enough we have complete asymptotic burning.*

Proof. Select t_0 small enough so that $t_0 B \| Y_0 \|_\infty \leqq \gamma/2$; then from (1.2), (2.1b), and (2.2a) we have that

$$Y(t) \geqq W_1(t)\, Y_0 - tB \| Y_0 \|_\infty \geqq W_1(t)\, Y_0 - \gamma/2 \tag{3.10}$$

whenever $0 \leqq t \leqq t_0$. Since $W_1(t)$ is positivity preserving we have that $(W_1(t) Y_0)(x) \geqq \gamma$ for all $x \in \Omega$, hence it follows from (3.10) that $Y(x, t) \geqq \gamma/2$ for all $x \in \Omega$, if t is small enough.

Meanwhile, by the continuity of $W_0(t)$ in t, there exist positive constants t_1 and $\alpha_1 > T_I$, and a subset E_1 of E_0 with positive measure, such that $(W_0(t) T_0)(x) \geqq \alpha_1$ for $x \in E_1$ whenever $0 \leqq t \leqq t_1$. Set $\beta_1 = f(\alpha_1)$; then if $0 \leqq t \leqq \min\{t_0, t_1\}$ we have that

$$W_0(t-s)\, Y(s) f(T(s)) \geqq W_0(t-s)[(\gamma/2)\beta_1 \aleph_{E_1}] \tag{3.11}$$

on Ω for all $s \in [0, t]$, where $\aleph_{E_1}$ is the indicator function for E_1. Similarly to the choice of t_1, α_1, and E_1, we can find positive constants $t_2 \leqq \min\{t_0, t_1\}$ and β_2, and a subset E_2 of E_1 with positive measure, such that

$$(W_0(t-s)[(\gamma/2)\beta_1 \aleph_{E_1}])(x) \geqq \beta_2 \tag{3.12}$$

whenever $0 \leqq s \leqq t \leqq t_2$ and $x \in E_2$. Then for all $t \in [0, t_2]$ and all $x \in E_2$ we have from (2.1a), (3.11), (3.12), and the choice of α_1 that

$$T(x, t) \geqq \alpha_1 + Q\beta_2 t. \tag{3.13}$$

Since $T(x, t)$ is bounded below by zero on Ω, we see from (3.13) that by selecting $t = t_2$ and Q large enough we can arrange that

$$\frac{1}{|\Omega|}\int_{\Omega} T(x, t_2)\, dx > T_I. \tag{3.14}$$

Considering (1.1) with initial data $T(x, t_2)$ and $Y(x, t_2)$, we then obtain complete asymptotic burning by Theorem 3.1, which proves the proposition. □

Theorem 3.1 and Propositions 3.3 and 3.4 show that T_{AV} is a "semithreshold parameter." We use this term to describe the fact that when T_{AV} is above the critical value T_I, complete asymptotic burning occurs regardless of all other parameters; when T_{AV} is less than T_I, complete asymptotic burning or eventual flame quenching occurs depending on the other parameters, in particular, depending on Q. In fact, the next result shows that under the above conditions there exists a closed interval $[Q_*, Q^*]$ such that if $Q > Q^*$ we are guaranteed complete asymptotic burning and if $Q < Q_*$ we are guaranteed eventual flame quenching.

THEOREM 3.3. *Under all conditions of Propositions* 3.3 *and* 3.4, *there exists a positive constant* Q^* *such that for every* $Q > Q^*$ *complete asymptotic burning occurs and there exists a positive constant* Q_* *such that for every* $Q < Q_*$ *eventual flame quenching occurs.*

Proof. Set $X = T + QY$; then as in § 7 of [4] we assume, for simplicity, that $d_0 = d_1 = 1$ and rewrite (1.1a), (1.1b) as

$$T_t = T_{xx} + (X - T)f(T), \tag{3.15a}$$

$$X_t = X_{xx}. \tag{3.15b}$$

Here $X(0) = T_0 + QY_0 \equiv X_0$ so that $X(t) = e^{t\Delta}X_0$. Thus we can rewrite (3.15a), (3.15b) as the single equation

$$T_t = T_{xx} + (e^{t\Delta}X_0 - T)f(T). \tag{3.16}$$

Now suppose T' solves (3.15) with X_0 replaced by X_0', with

$$X_0' \geqq X_0. \tag{3.17}$$

Set $\bar{T} = T' - T$; then $\bar{T}$ satisfies

$$\bar{T}_t = \bar{T}_{xx} + V_1(t)\bar{T} + V_2(t) \tag{3.18}$$

where

$$\begin{aligned} V_1(t) &= (X(t) - T)f'(\varphi) - f(T'), \\ V_2(t) &= e^{t\Delta}(X_0' - X_0)f(T'). \end{aligned} \tag{3.19}$$

Here $\varphi = T + \theta(T' - T)$, where $\theta \in (0, 1)$ is chosen so that

$$f(T') - f(T) = f'(\varphi)(T' - T). \tag{3.20}$$

Let $U(t, s)$ be the fundamental solution for the operator $(\cdot)_{xx} + V_1(t)$, then from (3.18) we see that $\bar{T}$ satisfies

$$\bar{T}(t) = U(t, 0)\bar{T}_0 + \int_0^t U(t, s)V_2(s)\, ds \tag{3.21}$$

where $\bar{T}_0 = T_0' - T_0$.

Since $U(t, s)$ is positivity preserving and (3.17) holds, we see that $\bar{T}(t) \geqq 0$ for all $t \geqq 0$ provided $T_0' \geqq T_0$. In particular, if all other parameters remain fixed, replacing Q by Q' in W_0 with $Q' \geqq Q$ implies that $T'(t) \geqq T(t)$.

With all parameters fixed and $d_0 = d_1$, let S^* be the set of all Q such that eventual flame propagation occurs and let S_* be the set of all Q such that eventual flame

quenching occurs. Our two examples in Propositions 3.3 and 3.4 show that both sets are nonempty. Set $Q^* = \inf S^*$ and set $Q_* = \sup S_*$. Our comparison principle that we have just shown for Q then implies the conclusions of the theorem. □

The above proof is basically reproduced from parts of § 7 of [4]. We have included it here for completeness.

We now continue with the case where $d_0 = d_1$ and show that the above results provide a detailed picture of the large-time behavior of T and Y under this assumption. Our next result follows easily from Theorem 3.1 and the proof of Theorem 3.3.

COROLLARY 3.1. *Under the conditions of Theorem* 3.1 *we have that Y converges to zero uniformly in t and T converges uniformly in t to the average of $T_0 + QY_0$ over Ω.*

Proof. From the proof of Theorem 3.3 we have that $T + QY = \exp(t\Delta)(T_0 + QY_0)$. Hence $T + QY$ converges uniformly to the average of $T_0 + QY_0$ over Ω, i.e., the constant obtained by replacing T_0 by $T_0 + QY_0$ in the right-hand side of (1.3). But by Theorem 3.1, $Y(t)$ converges uniformly to zero in t, hence $T(t)$ converges uniformly to this constant, and the proof of the corollary is completed. □

Our final result of this section shows that, when $d_0 = d_1$ and T_0, Y_0 satisfy the (not too restrictive) assumptions of Theorem 3.3, we have a nearly complete portrait of the asymptotic dynamics of the model given by (1.1) if Q is not in the interval $[Q_*, Q^*]$.

THEOREM 3.4. *Let Q_*, Q^* be as in Theorem* 3.3; *then under the conditions of Theorem* 3.3 *we have that if $Q > Q^*$, Y converges uniformly to zero and T converges uniformly to the average of $T_0 + QY_0$ over Ω. If $Q < Q_*$, then there exists a constant $T_1 \leqq T_I$ and a constant Y_1 such that $T(t)$ converges uniformly to T_1 and $Y(t)$ converges uniformly to Y_1. We have that the averages of $T_1 + QY_1$ and $T_0 + QY_0$ over Ω are equal.*

Proof. If $Q > Q^*$, then the result follows by Proposition 3.4 and Corollary 3.1. If $Q < Q_*$ we have that there exists a positive constant t_0 such that $t \geqq t_0$ implies that $T(t) \leqq T_I$. Hence $Y(t)f(T(t)) = 0$ for $t \geqq t_0$. Considering (1.1) with initial data $T(t_0)$, $Y(t_0)$, we thus see that $T(t) = \exp(t\Delta)(T(t_0))$ and $Y(t) = \exp(t\Delta)(Y(t_0))$ for $t \geqq t_0$, and thus we can take T_1 and Y_1 to be the averages of $T(t_0)$ and $Y(t_0)$ over Ω, respectively. That the average of $T_1 + QY_1$ over Ω equals the average of $T_0 + QY_0$ over Ω follows from the opening remarks of the proof of Corollary 3.1. This concludes the proof of the theorem, as well as our discussion of (1.1) with boundary conditions (3.1). □

4. Dirichlet boundary conditions for T. In this section we assume that B_0 is the Dirichlet operator on Γ. Let g and h be nonnegative continuous functions on Γ which are smooth enough so that the desired regularity of Theorem 2.4 holds. We assume throughout this section that the following boundary conditions hold for T and Y:

$$T(x) = g(x), \qquad x \in \Gamma, \tag{4.1a}$$

$$B_1 Y(x) = h(x), \qquad x \in \Gamma \tag{4.1b}$$

where B_1 is either the Dirichlet or Neumann boundary operator; in the latter case we set $h \equiv 0$. Our first task is to show that $T(t)$ is bounded uniformly in t. Recall from Theorem 2.5 that we already have that Y is bounded. Recall from § 2 that A_0 is the closure in L^1 of $d_0\Delta$ on $C^2_{B_0}$, and $A_{0,p}$ is A_0 restricted to $D(A_0) \cap L^p$. Let λ_1 be the first eigenvalue of $-A_{0,2}$. Then $\lambda_1 > 0$ and it is shown in [3] that there exists a constant M such that for all $u \in C(\bar{\Omega})$ and all $t \geqq 0$

$$\| W_0(t)u \|_\infty \leqq M \| u \|_\infty e^{-\lambda_1 t}. \tag{4.2}$$

The next result then follows easily.

THEOREM 4.1. *Let T and Y satisfy* (4.1) *on* Γ *under the conditions of Theorem* 2.5. *In particular, assume that T_0 and Y_0 are in $C(\bar{\Omega})$; then there exists a constant K such that $\|T(t)\|_\infty \leqq K$ for all $t \geqq 0$.*

Proof. From (1.2), (2.3), (2.4), and (4.2) we have that

$$\begin{aligned}\|T(t)\|_\infty &\leqq \|T_0\|_\infty + \sup_\Gamma g + Q\int_0^t \|W_0(t-s)Y(s)f(T(s))\|_\infty\, ds \\ &\leqq \|T_0\|_\infty + \sup_\Gamma g + Q\int_0^t M\|Y(s)f(T(s))\|_\infty\, e^{-\lambda_1(t-s)}\, ds \\ &\leqq \|T_0\|_\infty + \sup_\Gamma g + QMBK_1\int_0^\infty e^{-\lambda_1(t-s)}\, ds \\ &\leqq \|T_0\|_\infty + \sup_\Gamma g + QMBK_1(1/\lambda_1)\end{aligned} \tag{4.3}$$

where K_1 equals the right-hand side of (2.4). The proof is thus established with K as the right-hand side of (4.3). □

THEOREM 4.2. *Under the conditions of Theorem* 4.1 *assume, in addition, that $g(x) > T_I$ for all $x \in \Gamma$, and that B_1 is the Neumann operator; then complete asymptotic burning occurs at a rate which is eventually exponential.*

Proof. As Γ is compact, there exists an $x_0 \in \Gamma$ such that $g(x) \geqq g(x_0) > T_I$ for all $x \in \Gamma$. Set $\alpha_0 = g(x_0)$; then by the maximum principle (see, e.g., [7] or [16]) $w_0(x) \geqq \alpha_0$ for all $x \in \Omega$. Now while $W_0(t)(T_0 - w_0)$ may not always be nonnegative, it goes to zero uniformly by (4.2), thus for any ε_1 such that $0 < \varepsilon_1 < \alpha_0 - T_I$, there exists a t_1 such that $t \geqq t_1$ implies that $\|W_0(t)(T_0 - w_0)\|_\infty < \varepsilon_1$. Set $\alpha_1 = \alpha_0 - \varepsilon_1$; then since $W_0(t)$ is positivity preserving and both Y and $f(T)$ are nonnegative, it follows from (2.3a) that $T(x, t) \geqq \alpha_1 > T_I$ for all $x \in \Omega$ and all $t \geqq t_1$. Set $\beta_1 = f(\alpha_1)$; then as in the proof of Theorem 3.1 it follows that

$$\|Y(t)\|_\infty \leqq \|Y_0\|_\infty\, e^{-\beta_1(t-t_1)} \tag{4.4}$$

for all $t \geqq t_1$. This completes the proof of the theorem. □

To underscore the advantage gain when T is kept above ignition temperature on the boundary, for the last result of this section we consider cases where T is below T_I on some (not necessarily connected) open subset of Γ. Recall that Q is proportional to the chemical heat release; the next theorem shows that eventually no combustion can occur in subsets of Ω of positive measure if Q is small enough.

THEOREM 4.3. *Under the conditions of Theorem* 4.1 *assume, in addition, that $g(x) < T_I$ for x in an open nonempty subset of Γ. Then there exists a set E of positive measure in Ω and a $t_1 > 0$ such that $T(x, t) \leqq T_I$ for all $x \in E$ and $t \geqq t_1$ if Q is small enough.*

Proof. Let K_1 equal the right-hand side of (2.4); then from the proof of Theorem 4.1 and the fact that $W_0(t)(T_0 - w_0) \leqq W_0(t)T_0$, we have that

$$T(x, t) \leqq (W_0(t)T_0)(x) + w_0(x) + QMBK_1(1/\lambda_1) \tag{4.5}$$

where, as before, λ_1 is the first eigenvalue of $-A_{0,2}$. By the continuity of $w_0(x)$ there exists a set of positive measure E_0 in Ω such that $w_0(x) < T_I$ for $x \in E_0$. In particular, there exists a $\gamma < T_I$ and a subset E of E_0 with positive measure such that $w_0(x) \leqq \gamma$ for $x \in E$. For positive ε such that $\varepsilon < T_I - \gamma$, choose t_1 such that $\|W_0(t)T_0\| < \varepsilon$ for all $t \geqq t_1$; then from (4.5) we have that for all $x \in E$ and all $t \geqq t_1$

$$T(x, t) \leqq \gamma + \varepsilon + QMBK_1(1/\lambda_1). \tag{4.6}$$

As $\gamma + \varepsilon < T_I$, it is now clear that if Q is small enough, $T(x, t) \leqq T_I$ for all $x \in E$ and all $t \geqq t_1$, thus completing the proof. □

We note that Theorem 4.3 holds regardless of whether B_1 represents the Dirichlet or Neumann operator, and regardless of the initial conditions T_0 and Y_0. In particular, even if combustion occurs initially throughout Ω, i.e., $T_0(x) > T_I$ for all $x \in \Omega$, we have eventual flame quenching on portions of Ω with positive measure.

5. Examples of maintained concentration levels in Ω. We allow in this section for B_0 to be either the Dirichlet or Neumann operator, and set B_1 to be the Dirichlet operator. Let g and h be nonnegative continuous functions on Γ with enough smoothness for the desired regularity in Theorem 2.4. Thus the boundary conditions assumed in this section are:

$$B_0 T = g \quad \text{on } \Gamma, \tag{5.1a}$$

$$Y = h \quad \text{on } \Gamma. \tag{5.1b}$$

Let λ_1 be the first eigenvalue of $-A_{1,2}$; then as with $A_{0,2}$ in § 4, $\lambda_1 > 0$ and there exists a constant M such that for all u in $C(\bar{\Omega})$

$$\| W_1(t)u \|_\infty \leqq M \| u \|_\infty e^{-\lambda_1 t} \tag{5.2}$$

for all $t \geqq 0$. We now use techniques similar to those in previous sections in obtaining the next result.

THEOREM 5.1. *Under the conditions of Theorem* 2.5 *assume that T and Y satisfy* (5.1) *on Γ. Then if h is strictly positive on a nonempty open subset of Γ, there exists a subset E of Ω with positive measure and constants $\gamma > 0$ and $t_1 > 0$ such that if the preexponential factor B is small enough we have that $Y(x, t) \geqq \gamma$ for all $x \in E$ and all $t \geqq t_1$.*

Proof. Let w_1 be as in Theorem 2.4; then by continuity there exists a subset E_0 of Ω with positive measure such that w_1 is strictly positive on E_0. Hence there exists a subset of E_0 with positive measure and a constant $\gamma_1 > 0$ such that $w_1(x) \geqq \gamma_1$ for all $x \in E$. Given ε such that $0 < \varepsilon < \gamma_1$, there exists by (5.2) a $t_1 > 0$ such that $\| W_1(t) w_1 \|_\infty < \varepsilon$ if $t \geqq t_1$. Then from (2.3b) and (5.2) we have for all $x \in E$ and $t \geqq t_1$ that

$$\begin{aligned} Y(x, t) &\geqq (W_1(t) Y_0)(x) + (\gamma_1 - \varepsilon) - \int_0^t \| W_1(t-s) Y(s) f(T(s)) \|_\infty \, ds \\ &\geqq (W_1(t) Y_0)(x) + (\gamma_1 - \varepsilon) - \int_0^t M \| Y(s) f(T(s)) \|_\infty e^{-\lambda_1 (t-s)} \, ds \\ &\geqq (W_1(t) Y_0)(x) + (\gamma_1 - \varepsilon) - M \left(\| Y_0 \|_\infty + \sup_\Gamma h \right) B \int_0^\infty e^{-\lambda_1 (t-s)} \, ds \\ &= (W_1(t) Y_0)(x) + (\gamma_1 - \varepsilon) - BM \left(\| Y_0 \|_\infty + \sup_\Gamma h \right) (1/\lambda_1). \end{aligned} \tag{5.3}$$

Note that $(\gamma_1 - \varepsilon) > 0$ and $(W_1(t) Y_0)(x) \geqq 0$, so that if we set

$$\gamma = (\gamma_1 - \varepsilon) - BM \left(\| Y_0 \|_\infty + \sup_\Gamma h \right) (1/\lambda_1), \tag{5.4}$$

then if B is small enough we have that $\gamma > 0$ and $Y(x, t) \geqq \gamma$ for all $x \in E$ and all $t \geqq t_1$, as desired. This completes the proof of the theorem. □

Fixed Dirichlet boundary conditions for Y occur, for example, in the case of a permeable pellet, as studied in [10], and in models of a combustible fluid where the burning occurs in a thin layer at the surface of the fluid. Theorem 5.1 shows that if B is small enough then in portions of Ω the rate of burning does not exceed the rate at which the concentration is diffusing in from the boundary.

6. Remarks. We note that the averaging condition $T_{AV} > T_I$ imposed on the initial temperature T_0 in Theorem 3.1 is a reasonable one since in practice T_I is a small number while the burnt temperature is typically very large. The importance of this condition is underscored by Proposition 3.3, in which an example of eventual flame quenching is produced when the averaging condition is not met. T_{AV} is in some sense the analogue of the quantity $(T_0(-\infty)+T_0(+\infty))/2$ considered in [4] with $\Omega=\mathbb{R}$. Note that in both cases imposing the condition that the respective average involving T_0 is above ignition temperature guarantees that eventually $T(x, t) > T_I$ and $Y(x, t) \to 0$ as $t \to \infty$ regardless of the values of all other parameters.

Meanwhile, we note that global existence and uniqueness results have been established for the more general Cauchy problem in $\mathbb{R}$ which couples T and Y with pressure, velocity, and density terms under certain conditions placed on these additional terms [12], [17].

The system (1.1) is a special case of a general class of equations for which boundedness results are already known for Dirichlet and Neumann conditions imposed on both components (see, e.g., [8], [9], [11] and the references contained therein). These results could have been substituted for Theorem 3.2, and for Theorem 4.1 when both T and Y satisfy fixed Dirichlet boundary conditions. Note that Theorem 4.1 holds with mixed conditions, although the techniques in the aforementioned papers can probably be modified to handle this case as well. Our purpose in developing self-contained boundedness results here has been to obtain bounds depending directly on the special structure of (1.1) and (1.2), and to preserve a common thread of argument.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] S. AGMON, A. DOUGLIS, AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* II, Comm. Pure Appl. Math., 17 (1989), pp. 35-92.

[3] H. AMANN, *Dual semigroups and second-order linear elliptic boundary value problems*, Israel J. Math., 45 (1983), pp. 225-254.

[4] J. D. AVRIN, *Qualitative theory for a model of laminar flames with arbitrary nonnegative initial data*, J. Differential Equations, 84 (1990), pp. 290-308.

[5] H. BERESTYCKI, B. NICOLAENKO, AND B. SCHEURER, *Traveling wave solutions to combustion models and their singular limits*, SIAM J. Math. Anal., 16 (1985), pp. 1207-1242.

[6] P. CLAVIN, *Dynamical behavior of premixed fronts in laminar and turbulent flows*, Progr. Energy Comb. Sci., 11 (1985), pp. 1-59.

[7] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Second edition, Springer-Verlag, Berlin, Heidelberg, 1983.

[8] S. L. HOLLIS, R. H. MARTIN, AND M. PIERRE, *Global existence and boundedness in reaction-diffusion systems*, SIAM J. Math. Anal., 18 (1987), pp. 744-761.

[9] S. L. HOLLIS AND J. MORGAN, *Global existence and asymptotic decay for systems of convective reaction-diffusion equations*, preprint.

[10] A. K. KAPILA AND B. J. MATKOWSKY, *Reaction-diffusion system with Arrhenius kinetics: multiple solutions, ignition, and extinction*, SIAM J. Appl. Math., 36 (1979), pp. 373-389.

[11] M. KIRANE, *Global bounds and asymptotics for a system of reaction-diffusion equations*, J. Math. Anal. Appl., 138 (1989), pp. 328-342.

[12] B. LARROUTUROU, *The equations of one-dimensional unsteady flame propagation: existence and uniqueness*, SIAM J. Math. Anal., 19 (1988), pp. 32-59.

[13] M. MARION, *Etude mathématique d'un modèle de flamme laminaire sans température d'ignition:* I-*Cas scalaire*, Ann. Fac. Sci. Toulouse, 6 (1984), pp. 215-255.

[14] ———, *Qualitative properties of a nonlinear system for laminar flames without ignition temperature*, Nonlinear Anal. TMA, 9 (1985), pp. 1269-1292.

[15] ———, *Attractors for reaction-diffusion equations: existence and estimate of their dimension*, Appl. Anal., 25 (1987), pp. 101-147.

[16] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[17] J. M. ROQUEJOFFRE, *Etude Mathématique et numérique d'un problème en théorie de la combustion*, thesis, INRIA Sophia Antipolis, June 1988.

[18] G. I. SIVASHINSKY, *Instabilities, pattern formation, and turbulence in flames*, Ann. Rev. Fluid Mech., 15 (1988), pp. 179-199.

[19] G. SOD, private communication, 1989.

[20] D. TERMAN, *Connection problems arising from nonlinear diffusion equations*, in Proc. Microconference on Nonlinear Diffusion, J. Serrin, L. Peletier, and W.-M. Ni, eds., Berkeley, CA, 1986.

[21] ———, *Stability of planar wave solutions to a combustion model*, SIMA J. Math. Anal., 21 (1990), pp. 1139-1171.

[22] F. WILLIAMS, *Combustion Theory*, second ed., Addison-Wesley, Reading, MA, 1985.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 392-399, March 1991

006

STABILITY ANALYSIS FOR THE SLOW TRAVELING PULSE OF THE FITZHUGH–NAGUMO SYSTEM*

GILBERTO FLORES†

Abstract. This paper is concerned with the existence and stability of the slow traveling pulse for the FitzHugh–Nagumo system $u_t = u_{xx} + u(1-u)(u-a) - w$, $w_t = \varepsilon(u - \gamma w)$. This traveling wave is obtained as a perturbation of the standing wave of the Nagumo equation $u_t = u_{xx} + u(1-u)(u-a)$. Its existence is established by analyzing how the unstable manifold of the origin exits a suitable block. This geometric proof is an alternative approach to the singular perturbation expansion proposed by Casten, Cohen, and Lagerstrom [*Quart. Appl. Math.*, 32 (1975), pp. 335-367], as well as to the existence proof of Hastings [*SIAM J. Appl. Math.*, 42 (1982), pp. 247-260].

The method also allows use of the techniques developed by Evans [*Indiana Univ. Math. J.*, 24 (1985), pp. 193-226] to analyze the spectrum of the variational equation around the traveling wave. It is shown that there is exactly one unstable mode.

Key words. nerve conduction, traveling wave, linearized equation, spectrum, stability

AMS(MOS) subject classifications. 34C05, 35K55, 35Q99

Introduction. The FitzHugh–Nagumo system

$$\text{(A)} \qquad \begin{aligned} u_t &= u_{xx} + u(1-u)(u-a) - w, \qquad -\infty < x < \infty, \\ w_t &= \varepsilon(u - \gamma w), \end{aligned}$$

where $0 < a < \frac{1}{2}$, $\varepsilon \geqq 0$, $\gamma \geqq 0$, is a simplified version of the model proposed by Hodgkin and Huxley [12a, b] to describe the conduction of electrical impulses along a nerve axon. The experimental and numerical evidence shows that the voltage travels along the nerve with little distortion of the shape and amplitude, and at nearly constant speed. The mathematical problem is to classify the waveforms and to determine their stability.

This program has been carried out for the Nagumo equation $u_t = u_{xx} + f(u)$, $f(u) = u(1-u)(u-a)$, which corresponds to $\varepsilon = 0$, $w \equiv 0$ in (A). The equation has a front, that is, a traveling wave $\varphi(\xi)$, $\xi = x + ct$, with $\varphi(-\infty) = 0$, $\varphi(+\infty) = 1$. The speed is $c = \sqrt{2}(\frac{1}{2} - a)$. Fife and McLeod [7] proved that this front is globally stable: If the initial datum $u_0(x)$ satisfies $u_0(-\infty) < a < u_0(+\infty)$, then the corresponding solution approaches a translate of the front. The Nagumo equation also has a solitary wave, that is, a wave solution with a single hump and vanishing at $\pm\infty$. Its speed is $c = 0$, so it is a standing wave. It has been established in [9] that the standing wave is a saddle of codimension 1, and that it is a threshold for the steady state $\bar{u} = 1$. The status of the problem for the FitzHugh–Nagumo system and related models will be discussed at the end of the introduction.

In terms of a moving frame $\xi = x + ct$, (A) becomes

$$\text{(B)} \qquad \begin{aligned} u_t &= u_{\xi\xi} - cu_\xi + f(u) - w, \\ w_t &= -cw_\xi + \varepsilon(u - \gamma w). \end{aligned}$$

Traveling waves of (A) correspond to stationary solutions (B). Carpenter [1], Hastings [10], and Langer [14] have proved the existence of a fast traveling pulse, $((u(\xi), w(\xi)) \to (0, 0)$ as $|\xi| \to \infty)$ with speed $c(\varepsilon) = \sqrt{2}(\frac{1}{2} - a) + o(1)$. The construction in Carpenter [1]

* Received by the editors October 25, 1989; accepted for publication (in revised form) April 4, 1990.
† Departamento de Matemáticas y Mecánica, Instituto de Investigationes en Matematicas Aplicada, Universidad Nacional Autonoma de Mexico, Apdo. Postal 20-726, 01000 México, D.F., Mexico.

and Langer [14] is explicit: for small $\varepsilon > 0$, the fast pulse is a perturbation of a singular orbit formed by piecing together the front of the Nagumo equation with the graph of $w = f(u)$. Jones [13] has proved that this wave is stable under small perturbations. He has shown that the continuous spectrum of the linearization of (B) around the fast traveling wave is contained in a half-plane $\operatorname{Re}(\lambda) < -b^2$, that $\lambda = 0$ is a simple eigenvalue, and that there are no eigenvalues with positive real part.

Using singular perturbations, Casten, Cohen, and Lagerstrom [2], have constructed a series expansion of a slow traveling pulse with $c(\varepsilon) = O(\sqrt{\varepsilon})$, the first term being the standing wave of the Nagumo equation. Hastings [7] has proved the existence of a pulse with $c(\varepsilon) = O(\sqrt{\varepsilon})$. The numerical evidence for this kind of problems suggests that slow waves are unstable.

In this work, the existence of a slow traveling wave ($c(\varepsilon) = O(\sqrt{\varepsilon})$) is proved by the method of isolating blocks, which yields the slow wave as a perturbation of the standing wave for the scalar equation. It is also shown that for small $\varepsilon > 0$, the continuous spectrum of the linearization of (B) around the slow wave lies in a half-plane $\operatorname{Re}(\lambda) < -b^2$, and that there is exactly one positive eigenvalue. The stability analysis in this case is simplified by the fact that the slow pulse is a perturbation of a homoclinic orbit, and by the convergence of the analytic functions determining the point spectrum of the linearization of (B) at the wave to the corresponding function of the reduced equation.

The limit of the fast pulses is a "singular orbit" consisting of heteroclinic orbits and critical points. The behavior of the analytic functions is more complicated (see Jones [13]).

The analogy with the scalar equation suggests that the stable manifold of the slow pulse is a threshold for the nerve impulse: solutions below this manifold should decay, while solutions above it should propagate and approach the fast pulse. The decay of solutions of initial-boundary problems corresponding to small initial data has been proved by Schoenbeck [18] and Weixi [20].

The existence of multiple pulses and infinite trans of the FitzHugh-Nagumo system has been proved by Hastings [10], [11] and Carpenter [1], but there are no results concerning their stability. The situation for the Hodgkin-Huxley equation is the same.

A little more is known about the piecewise linear model introduced by McKean [15], where the cubic is replaced by the broken line $-u + H(u-a)$, H being the Heaviside step function. Rinzel and Keller [17] have proved the existence of solitary waves and some infinite trains. They have also made numerical studies to establish the instability of the slow waves. The existence of multiple pulses was proved by Evans, Fenichel, and Feroe [4]. Feroe [5] proved that the fast pulse is stable, and he has also made numerical studies on the stability of multiple impulses [6]: the fastest is stable, and the others are unstable. The interesting fact is that some unstable modes correspond to growth in the amplitude, and some correspond to the spacing between the pulses. The only analytic results in this direction are bounds on the number of unstable modes of n pulse solutions, which were obtained by Wang [19a, b]: fast n-pulses have at most n unstable modes, slow n-pulses have at least n and at most $2n-1$ unstable modes.

1. Existence of the slow traveling wave. It will be shown that the system for traveling wave solutions of the FitzHugh-Nagumo equation

$$
\begin{aligned}
u' &= v, \\
v' &= cv - f(u) + w, \qquad ' = d/d\xi, \\
w' &= \varepsilon / c(u - \gamma w),
\end{aligned}
\tag{1}
$$

has a homoclinic orbit for $c = O(\sqrt{\varepsilon})$, which is a perturbation of the standing wave of the Nagumo equation.

Let $M = \max\{f'(u): 0 \leqq u \leqq 1\}$; then $\frac{1}{6} < M < \frac{1}{3}$. Choose $\gamma < 1/M$ (so that the origin is the only critical point of (1)), and $\delta > 0$ such that $M + \delta < 1/\gamma$. We will consider parameter values ε, c in the region $\varepsilon \leqq (M+\delta)c^2$. Since $\varepsilon/c \leqq (M+\delta)c$, the (one-dimensional) unstable manifold of the origin for (1), denoted by $U_{\varepsilon,c}(\xi)$, depends continuously on (ε, c) in the region under consideration.

We use the notion of an isolating block as defined in § 1 of Carpenter [1]: equation (1) is the autonomous system $X' = G(X)$, where $X = (u, v, w)$.

DEFINITION. A set B is *block* for (1) if it is homeomorphic to $[0,1]^3$ and there exist C^1 functions $f_1, \cdots, f_6 : \mathbb{R}^3 \to \mathbb{R}$ such that $B = \bigcap_{i=1}^{6} f_i^{-1}([0,\infty])$ and $\langle \nabla f_i, G\rangle \neq 0$ on $f_i = 0$.

Thus a block is a set homeomorphic to a closed ball, such that the trajectories of the flow are transversal to its boundary.

Take a block $\tilde{B}$ around $(0,0)$ for the reduced system $u' = v, v' = -f(u)$, with sides parallel to the lines $v = \pm u$, in such a way that the standing wave leaves $\tilde{B}$ through b_1^- and comes back through b_2^+ (see Fig. 1).

Similarly, we can construct blocks for $u' = v, v' = -f(u) + w$, for $|w|$ small, around the smallest root $u_1 = u_1(w)$ of the equation $w = f(u)$ (see Fig. 2). The three-dimensional

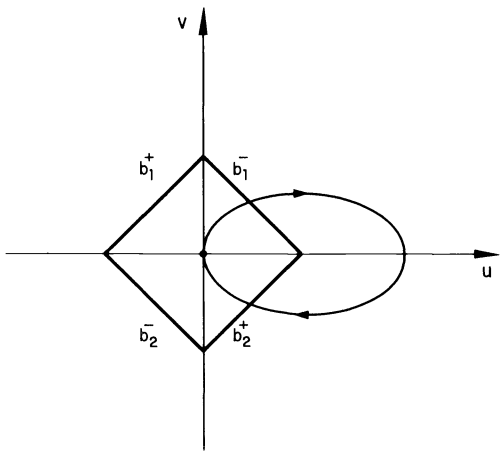


FIG. 1. *Exit set* $\tilde{B}^- = b_1^- \cup b_2^-$. *Entrance set* $\tilde{B}^+ = b_1^+ \cup b_2^+$.

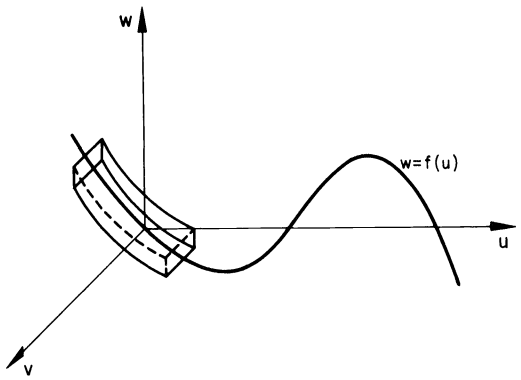


FIG. 2. *Exit set* $B^- = b_1^- \cup b_2^-$ *front and back.*

box constructed in this manner is a block for (1) because the top and bottom, being parallel to the u, v plane, are transversal to the vector field for $\varepsilon > 0$.

We will concentrate on the behavior of the unstable manifold $U_{\varepsilon,c}(\xi)$ after it returns to B through b_2^+. $U_{\varepsilon,c}(\xi)$ is normalized by the condition $U_{\varepsilon,c}(\xi) = a$ for the first time at $\xi = 0$. The asymptotic behavior of $U_{\varepsilon,c}(\xi)$ has been analyzed by Hastings [11]: the sets $E^- = \{(u, v, w)\colon u < 0,\ v < 0,\ v' < 0,\ w' < 0\}$ and $E^+ = \{(u, v, w)\colon u > 1, v > 0, v' > 0, w' > 0\}$ are invariant regions for (1). Moreover, $U_{\varepsilon,c}(\xi)$ enters $E^+(E^-)$ if and only if u and v tend to $+\infty(-\infty)$.

Let $\Omega = \{(c, \varepsilon)\colon c > 0, \varepsilon \geqq 0\}$, $\Omega_1 = \{(c, \varepsilon)\colon U_{\varepsilon,c}(\xi)$ is bounded$\}$, $\Omega_2 = \{(c, \varepsilon)\colon U_{\varepsilon,c}(\xi)$ enters $E^+\}$, $\Omega_3 = \{(c, \varepsilon)\colon U_{\varepsilon,c}(\xi)$ enters $E^-\}$; then Ω_2 and Ω_3 are disjoint open sets, and $\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3$. The region between the parabolas $\varepsilon = Mc^2$ and $\varepsilon = (1/\gamma)c^2$ is contained in Ω_2. Let C_3 be the component of Ω_3 containing the segment $(0, c_0) \times \{0\}$, where $c_0 = \sqrt{2}(\frac{1}{2} - a)$ (see Fig. 3). It will be shown that C_3 contains a portion of a parabola. For $0 < c < c_0$, $U_{\varepsilon,c}(\xi)$ returns to B through b_2^+. For $0 < c < c_0$ and $\varepsilon = 0$, $U_{0,c}(\xi)$ leaves B for the second time through b_2^-, and $u < 0$, $v' < 0$ at the exit point. By continuity, the same is true for $0 \leqq \varepsilon \leqq \varepsilon_0$, $0 < c \leqq c_0$. The claim is that $U_{\varepsilon,c}(\xi)$ enters E^- if it leaves B through b_2^- for $c \leqq c_0$ and $\varepsilon \leqq (M + \delta)c^2$. This is the content of the following lemma.

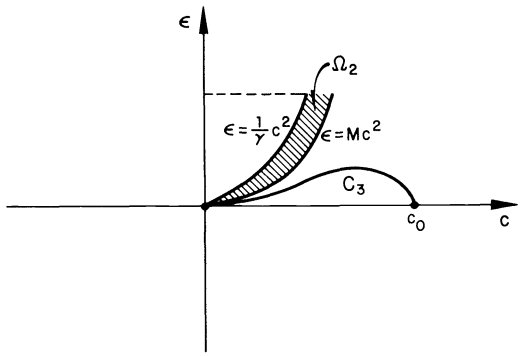


FIG. 3

LEMMA 1. *If $c \leqq c_0$, $\varepsilon \leqq (M + \delta)c^2$ and $U_{\varepsilon,c}(\xi)$ leaves B through b_2^-, then $U_{\varepsilon,c}(\xi_0) \in E^-$ for some $\xi_0 > 0$.*

Proof. Assume that $U_{\varepsilon,c}(\xi)$ leaves B (after returning through b_2^+) for the first time at $\xi = T(\varepsilon, c)$, and that $U_{\varepsilon,c}(T) \in b_2^-$, then $u(T) < 0$, $v(T) < 0$ and $v'(T) < 0$. There are two possibilities:

(i) $w(T) \geqq 0$, in which case $w'(T) = (\varepsilon/c)(u(T) - \gamma w(T)) < 0$, so that $U_{\varepsilon,c}(T) \in E^-$.

(ii) $w(T) < 0$. Only the case $w'(T) \geqq 0$ needs to be considered. There are two subcases:

(a) $U_{\varepsilon,c}(\xi)$ leaves the region $u < 0$, $v < 0$, $v' < 0$ for the first time at $T_1 > T$, and $w'(\xi) \geqq 0$ at $T \leqq \xi \leqq T_1$. In this situation, $v'(T_1) = 0$, which implies $w(T_1) > 0$. For small $\delta > 0$, the inequalities $u(T_1 - \delta) < 0$, $v(T_1 - \delta) < 0$, $v'(T_1 - \delta) < 0$, and $w(T_1 - \delta) > 0$ hold, so that $w'(T_1 - \delta) < 0$, a contradiction.

(b) $U_{\varepsilon,c}(\xi)$ remains in the region $u < 0$, $v < 0$, $v' < 0$ as long as it exists, and $w'(\xi) \geqq 0$ for $T \leqq \xi < T^*(\varepsilon, c)$ (T^* is the largest positive time for which $U_{\varepsilon,c}(\xi)$ is defined). If $T^* < \infty$, then $u(\xi) \to -\infty$, $v(\xi) \to -\infty$ as $\xi \to T^*$, from which it follows that $w'(\xi) = (\varepsilon/c)(u(\xi) - \gamma w(\xi)) \to -\infty$ as $\xi \to T^*$, so that eventually $w'(\xi) < 0$, a contradiction.

If $T^* = \infty$, the claim is that $w(\xi) \to 0$ as $\xi \to \infty$. If this were not true, there would exist $\alpha > 0$ such that $w(\xi) \leqq -\alpha$ for $\xi \geqq T$, then $v'(\xi) \leqq -\alpha$, so that $v(\xi) \to -\infty$ and $u(\xi) \to -\infty$ as $\xi \to \infty$. This implies that $w'(\xi) \to -\infty$, a contradiction. But if $w(\xi) \to 0$, then $w'(\xi) \to -\infty$, again a contradiction.

It follows that $U_{\varepsilon,c}(\xi)$ must enter E^- at some $\xi > T$. The proof is finished.

Now let $\tilde{w} = (1/\sqrt{\varepsilon})w$ and $\sqrt{\varepsilon} = \alpha c$. Then (1) becomes

$$u' = v, \quad v' = cv - f(u) + \alpha c\tilde{w}, \quad \tilde{w}' = \alpha(u - \gamma\alpha c\tilde{w}). \tag{2}$$

When $\alpha = 0$, we get $u' = v$, $v' = cv - f(u)$, $\tilde{w}' = 0$, so that $\tilde{U}_{0,c}(\xi)$ leaves B through b_2^- if $c \leqq c_0$, and by continuity the same is true for $\alpha \leqq \sqrt{m}$. By Lemma 1, $U_{\alpha,c}(\xi)$ enters E^- if $c \leqq c_0$, $\alpha \leqq \sqrt{m}$, so that $(\varepsilon, c) \in \Omega_3$ if $c \leqq c_0$ and $\varepsilon \leqq mc^2$.

Restrict the values of (ε, c) to the region $c \leqq c_0$, $\varepsilon \leqq \min\{(M+\delta)c^2, \varepsilon_0\}$, where $\varepsilon_0 = mc_0^2$. Take $0 < \varepsilon < \varepsilon_0$, and define $c_2 = \sqrt{\varepsilon/(M+\delta)}$, $c_3 = \sqrt{\varepsilon/m}$. Since B is a block, the exit point of $U_{\varepsilon,c}(\xi)$ depends continuously on (ε, c). Therefore, $U_{\varepsilon,c}(\xi)$ leaves B through b_2^- if $c_3 < c < c_0$. Since $(\varepsilon, c_2) \in \Omega_2$, it follows from Lemma 1 that $U_{\varepsilon,c}(\xi)$ leaves B through b_1^-. By continuity, there exists $\bar{c}$, such that $\sqrt{\varepsilon}/\sqrt{M+\delta} < \bar{c} < \sqrt{\varepsilon}/\sqrt{m}$ and $U_{\varepsilon,\bar{c}}(\xi)$ remains in B after returning through b_2^+. By choosing B small enough we can guarantee that $U_{\varepsilon,\bar{c}}(\xi) \to (0,0,0)$ as $\xi \to \infty$, and therefore $U_{\varepsilon,\bar{c}}(\xi)$ is the slow traveling pulse.

2. Stability analysis. To simplify the notation, we make the change $c \to -c$. The system for traveling waves is

$$u' = v, \quad v' = -cv - f(u) + w, \quad w' = -\frac{\varepsilon}{c}(u - \gamma w). \tag{3}$$

If $U_\varepsilon = (u_\varepsilon, v_\varepsilon, w_\varepsilon)$ is the travelling wave found in the previous section, and L_ε is the linearization around U_ε of

$$u_t = u_{\xi\xi} + cu_\xi + f(u) - w, \qquad w_t = cw_\xi + \varepsilon(u - \gamma w), \tag{4}$$

then

$$L_\varepsilon \begin{pmatrix} p \\ r \end{pmatrix} = \begin{pmatrix} \dfrac{d^2p}{d\xi^2} + c\dfrac{dp}{d\xi} + f'(u_\varepsilon)p - r \\ c\dfrac{dr}{c\xi} + \varepsilon(p - \gamma r) \end{pmatrix}.$$

The spectrum of L_ε, denoted by $\sigma(L_\varepsilon)$, consists of the values of $\lambda \in \mathbb{C}$ for which $L_\varepsilon \binom{p}{r} = \lambda \binom{p}{r}$ has a bounded solution. The spectrum can be divided into two parts: the point spectrum $\sigma_p(L_\varepsilon)$, which consists of isolated eigenvalues of finite multiplicity, and the essential spectrum $\sigma_e(L_\varepsilon) = \sigma(L_\varepsilon) \setminus \sigma_p(L_\varepsilon)$.

In the limit case $\varepsilon = 0$ (and $c = 0$), the corresponding linear operator is $L_0(p) = (d^2p)/dx^2 + f'(u_0)p$, where $u_0(x)$ is the standing wave of the Nagumo equation. The spectrum of L_0 can be computed: $\sigma_e(L_0) = (-\infty, -a]$, and there is exactly one positive eigenvalue (see Flores [9] and McKean and Moll [16]). It will be shown that for small $\varepsilon > 0$, $\sigma_e(L\varepsilon)$ lies in the left half-plane, and that the point spectrum has one positive eigenvalue, so that U_ε is a saddle of codimension 1.

The eigenvalue equation $L_\varepsilon \binom{p}{r} = \lambda \binom{p}{r}$ can be written as the system

$$\begin{aligned} p' &= q, \\ q' &= cq + [\lambda - f'(u_\varepsilon)]p + r, \\ r' &= -\frac{\varepsilon}{c}p + \left(\frac{\lambda + \varepsilon\gamma}{c}\right)r. \end{aligned} \tag{5}$$

If $z=(p, q, r)$, then (5) is the linear system in $\mathbb{C}^3$: $z'=Az$, where

$$A=\begin{pmatrix} 0 & 1 & 0 \\ \lambda - f'(u_\varepsilon) & -c & 1 \\ -\varepsilon/c & 0 & (\lambda+\varepsilon\gamma)/c \end{pmatrix}.$$

System (5) is asymptotically autonomous, the asymptotic equation being given by the matrix

$$A_0=\begin{pmatrix} 0 & 1 & 0 \\ \lambda + a & -c & 1 \\ -\varepsilon/c & 0 & (\lambda+\varepsilon\gamma)/c \end{pmatrix}.$$

Let L_1 be the linearization of (4) around $(0, 0, 0)$, that is,

$$L_1\begin{pmatrix} p \\ r \end{pmatrix}=\begin{pmatrix} \dfrac{d^2p}{d\xi^2}+c\dfrac{dp}{d\xi}-ap-r \\ c\dfrac{dr}{c\xi}+\varepsilon(p-\gamma r) \end{pmatrix};$$

then $\sigma_e(L_1)=S=\{\lambda\in\mathbb{C}: A_0(\lambda)$ has a pure imaginary eigenvalue$\}$.

Since

$$L_\varepsilon = L_1+\begin{pmatrix} [f'(u_{\varepsilon_0})+a]p \\ 0 \end{pmatrix},$$

and the second term in the right-hand side is a relatively compact perturbation, it follows that any connected component of $\mathbb{C}\backslash S$ is entirely the essential spectrum, or the only spectrum in it is discrete.

The claim is that S lies in the left half-plane for small $\varepsilon>0$. Let

$$P(\alpha, \varepsilon, \lambda)=\det(A_0(\varepsilon, \lambda)-\alpha I)=[\alpha^2+c\alpha-(\lambda+a)]\left[\frac{\lambda+\varepsilon\gamma}{c}-\alpha\right]-\frac{\varepsilon}{c}.$$

Since $c(\varepsilon)<0$, $S=\{\lambda\in\mathbb{C}: c(\varepsilon)P(\alpha, \varepsilon, \lambda)=0$ for some $\alpha=i\tau, \tau\in\mathbb{R}\}$. When $\varepsilon=0$, the equation is $[\alpha^2-(\lambda+a)]\lambda=0$, and this has solutions $\alpha=i\tau$, $\tau\in\mathbb{R}$ if and only if $\lambda\leqq -a$, so that $S_0=(-\infty, -a]$. By continuity, S lies in the left half-plane for small $\varepsilon>0$.

Let G be the connected component of $\mathbb{C}\backslash S$ which contains a half-plane of the form $\{\lambda: \mathrm{Re}\,(\lambda)>b\}$, $b<0$. From Theorem 1 and Corollary 3 of Evans [3c], $\lambda\in\rho(L_\varepsilon)$ if $|\lambda|$ is large. Therefore, the only part of the spectrum in G is $\sigma_p(L_\varepsilon)$. Moreover, the number of eigenvalues of $A_0(\lambda)$ in $\mathrm{Re}\,(\lambda)>0$ is constant in G. To compute this number, note that $c(\varepsilon)P(\alpha, \varepsilon, \lambda)=[a^2+c\alpha-(\lambda+a)][\lambda+\varepsilon\gamma-c\alpha]-\varepsilon=0$. At $\varepsilon=0$ we get $[\alpha^2-(\lambda+a)]\lambda=0$. For $\lambda\neq 0$ there are two eigenvaues: $\alpha_1(0, \lambda)=\sqrt{\lambda+a}$, and $\alpha_2(0, \lambda)=-\sqrt{\lambda+a}$ with $\mathrm{Re}\,(\alpha_2(0, \lambda))<0<\mathrm{Re}\,(\alpha_1(0, \lambda))$. By continuity there are two eigenvalues $\alpha_1(\varepsilon, \lambda)$, $\alpha_2(\varepsilon, \lambda)$ with $\mathrm{Re}\,(\alpha_2(\varepsilon, \lambda))<0<\mathrm{Re}\,(\alpha_1(\varepsilon, \lambda))$ for small $\varepsilon>0$. Since $\alpha_1+\alpha_2+\alpha_3=(\lambda+\varepsilon\gamma)/c-c$, it follows that $\alpha_3(\varepsilon, \lambda)\sim\lambda/c$. In particular $\mathrm{Re}\,(\alpha_3)\sim\mathrm{Re}\,(\lambda/c)\sim -k^2/\sqrt{\varepsilon}$.

Therefore, the eigenvalues of A_0 satisfy $\mathrm{Re}\,(\alpha_3(\varepsilon, \lambda))<\mathrm{Re}\,(\alpha_2(\varepsilon, \lambda))<0<\mathrm{Re}\,(\alpha_1(\varepsilon, \lambda))$ if ε is small enough. Let X_1, X_2, X_3 be the eigenvalues of A_0 corresponding to $\alpha_1, \alpha_2, \alpha_3$, respectively, then $X_1=(1, \alpha_1, -\varepsilon(c\alpha_1-(\lambda+\varepsilon\gamma))^{-1})$. System (5) has a solution $\varphi(\xi; \varepsilon, \lambda)=X_1e^{\alpha_1(\varepsilon, \lambda)\xi}+o(e^{\alpha_1(\varepsilon, \lambda)\xi})$ as $\xi\to-\infty$. This is a candidate for being an eigenfunction. To see if $\varphi(\xi; \varepsilon, \lambda)$ is bounded as $\xi\to+\infty$, consider the adjoint system $z^{*\prime}=Bz^*$, where $B=-A^*$. This system is asymptotic to $z^{*\prime}=B_0z^*$, where $B_0=-A_0^*$.

The eigenvalues of B_0 are $\{\beta_1, \beta_2, \beta_3\} = \{-\bar{\alpha}_1, -\bar{\alpha}_2, -\bar{\alpha}_3\}$. Let Y_1, Y_2, Y_3 be the corresponding eigenvectors. Y_1 is given by

$$Y_1 = \left(1, (c-\beta_1)^{-1}, \left[(\beta_1 - c)\left(\beta_1 + \frac{\bar{\lambda} + \varepsilon\gamma}{c}\right)\right]^{-1}\right).$$

The adjoint system has a solution $\psi(\xi; \varepsilon, \lambda) = Y_1 e^{\beta_1 \xi} + o(e^{\beta_1 \xi})$ as $\xi \to \infty$. The function $D_\varepsilon(\lambda) = \varphi(\xi; \varepsilon, \lambda) \cdot \psi(\xi; \varepsilon, \lambda)$, originally introduced by Evans [3a-d], containing the information relative to the stability. Indeed, λ is an eigenvalue of L_ε if and only if $D_\varepsilon(\lambda) = 0$. To see this take $\{\varphi_1, \varphi_2, \varphi_3\}$, $\{\psi_1, \psi_2, \psi_3\}$ as fundamental systems of $z' = Az$ and $z^{*\prime} = Bz^*$, respectively, such that $\varphi_i \sim X_i e^{\alpha_i \xi}$ as $\xi \to \infty$, $\psi_i \sim Y_i e^{\beta_i \xi}$ as $\xi \to \infty$. Since $X_2 \cdot Y_1 = 0$, $X_3 \cdot Y_1 = 0$, and $\varphi(\xi; \varepsilon, \lambda) = \sum_{i=1}^{3} c_i(\lambda)\varphi_i(\xi, \varepsilon, \lambda)$ it follows that

$$D_\varepsilon(\lambda) = \varphi(\xi; \varepsilon, \lambda) \cdot \psi(\xi; \varepsilon, \lambda) = \sum_{i=1}^{3} c_i(\lambda)\varphi_i(\xi; \varepsilon, \lambda) \cdot \psi(\xi; \varepsilon, \lambda) \sim c_1(\lambda) X_1 \cdot Y_1$$

as $\xi \to \infty$ so that $D_\varepsilon(\lambda) = 0$ if and only if $c_1(\lambda) = 0$, in which case $\varphi(\xi; \varepsilon, \lambda)$ is bounded as $\xi \to \infty$.

There are solutions $\psi_0(\xi; \lambda)$, $\psi_0(\xi; \lambda)$ corresponding to the reduced system

$$p' = q, \qquad q' = [\lambda - f'(u_0)]p. \tag{6}$$

The function $D_0(\lambda) = \varphi_0(\xi; \lambda) \cdot \psi_0(\xi; \lambda)$ has roots at $\lambda_0 = 0$ and $\lambda_1 > 0$. These are the only roots in the half-plane $\mathrm{Re}\,(\lambda) \geqq 0$. It will be shown that $D_\varepsilon(\lambda) = 0$ has exactly one positive root for small $\varepsilon > 0$. This is due to the fact that $D_\varepsilon(\lambda) \to D_0(\lambda)$ as $\varepsilon \downarrow 0$, uniformly on compact subsets of G.

Let $\varphi(\xi; \varepsilon, \lambda) = (p(\xi; \varepsilon, \lambda), q(\xi; \varepsilon, \lambda), r(\xi; \varepsilon, \lambda))$ be the solution of (5) with $\varphi(\xi; \varepsilon, \lambda) \sim X_1 e^{\alpha_1 \xi}$ as $\xi \to -\infty$; then

$$r(\xi; \varepsilon, \lambda) = e^{(\lambda/c)\xi} r(0; \varepsilon, \lambda) + \frac{\varepsilon}{c} \int_0^\xi e^{(\lambda/c)(\xi - \sigma)} [(p(\sigma) - \gamma r(\sigma)]\, d\sigma.$$

Since $\alpha_1(\varepsilon, \lambda) = \sqrt{\lambda + a} + o(1)$ as $\varepsilon \to 0$, there are bounds of the form $|\varphi(\xi; \varepsilon, \lambda)| \leqq M e^{\alpha \xi}$, $\xi \geqq 0$, ε small, where the constants M and α are uniform for λ in compact subsets of G. Moreover, $\mathrm{Re}\,(\lambda/c) \to -\infty$ as $\varepsilon \downarrow 0$, uniformly for $0 \leqq \xi \leqq \xi_0$ and λ in compact subsets of G. From the construction of the slow wave $U_\varepsilon(\xi)$, it follows that $|u_\varepsilon(\xi) - u_0(\xi)| \to 0$ as $\varepsilon \downarrow 0$, and therefore $D_\varepsilon(\lambda) \to D_0(\lambda)$ as $\varepsilon \downarrow 0$ uniformly on compacts subsets of G. By Rouché's theorem, $D_\varepsilon(\lambda) = 0$ has exactly one root $\lambda_1(\varepsilon)$ near λ_1, the positive root of $D_0(\lambda) = 0$, and since $D_\varepsilon(\lambda)$ is real for λ real, it follows that $\lambda_1(\varepsilon) > 0$.

REFERENCES

[1] G. CARPENTER, *A geometric approach to singular perturbation problems with applications to nerve impulse equations*, J. Differential Equations, 23 (1977), pp. 152-173.

[2] R. CASTEN, H. COHEN, AND P. LAGERSTROM, *Perturbation analysis of an approximation to Hodgkin-Huxley theory*, Quart. Appl. Math., 32 (1975), pp. 335-367.

[3a] J. EVANS, *Nerve axon equations*: I. *Linear approximations*, Indiana Univ. Math. J., 21 (1972), pp. 877-955.

[3b] ———, *Nerve axon equations*: II. *Stability at rest*, Indiana Univ. Math. J., 22 (1972), pp. 75-90.

[3c] ———, *Nerve axon equations*: III. *Stability of the nerve impulse*, Indiana Univ. Math. J., 22 (1972), pp. 577-594.

[3d] ———, *Nerve axon equations*: IV. *The stable and unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 193-226.

[4] J. EVANS, N. FENICHEL, AND J. FEROE, *Double impulse solutions in nerve axon equations*, SIAM J. Appl. Math., 42 (1982), pp. 219-234.

[5] J. FEROE, *Temporal stability of solitary impulse solutions of a nerve axon equation*, Biophys. J., 21, 1978, pp. 103-110.

[6] ———, *Existence and stability of multiple impulse solutions of a nerve axon equation*, SIAM J. Appl. Math., 42 (1982), pp. 235-246.

[7] P. FIFE AND J. B. MCLEOD, *The approach of solutions of nonlinear diffusion equations to traveling front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335-361.

[8] R. FITZHUGH, *Impulses and physiological states in theoretical models of nerve membranes*, Biophys. J., 1 (1961), pp. 445-466.

[9] G. FLORES, *The stable manifold of the standing wave of the Nagumo equation*, J. Differential Equations, 80 (1989), pp. 306-314.

[10] S. HASTINGS, *On the existence of homoclinic and periodic orbits for the FitzHugh-Nagumo equations*, Quart. J. Math. Oxford Ser. (2), 27 (1976), pp. 123-134.

[11] S. HASTINGS, *Single and multiple pulse waves for the FitzHugh-Nagumo equations*, SIAM J. Appl. Math., 42 (1982), pp. 247-260.

[12a] A. F. HODGKIN AND G. F. HUXLEY, J. Physiol., 116 (1952), pp. 449-506.

[12b] ———, J. Physiol., 117 (1952), pp. 500-544.

[13] C. JONES, *Stability of the traveling wave solution of the FitzHugh-Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431-469.

[14] R. LANGER, *Existence of homoclinic travelling wave solutions to the FitzHugh-Nagumo equations*, Ph.D. thesis, Northeastern University, Evanston, IL, 1980.

[15] H. MCKEAN, *Nagumo's equation*, Adv. in Math, 4 (1970), pp. 209-223.

[16] H. MCKEAN AND V. MOLL, *Stabilization to the standing wave in a caricature of the nerve equation*, Comm. Pure Appl. Math, 39 (1986), pp. 485-529.

[17] J. RINZEL AND J. KELLER, *Traveling wave solutions of a nerve conduction equation*, Biophys. J., 13 (1973), pp. 1313-1337.

[18] M. E. SCHOENBECK, *Boundary value problems for the FitzHugh-Nagumo equations*, J. Differential Equations, 30 (1978), pp. 119-147.

[19a] W. P. WANG, *Multiple impulse solutions to McKean's caricature of the nerve equation.* I: *Existence*, Comm. Pure Appl. Math., 41 (1988), pp. 71-103.

[19b] ———, *Multiple impulse solutions to McKean's caricature of the nerve equation.* II: *Stability*, Comm. Pure Appl. Math., 41 (1988), pp. 997-1025.

[20] S. WEIXI, *Threshold phenomena for the first initial boundary value problem of the FitzHugh-Nagumo equations*, Math. Methods Appl. Sci., 11 (1989), pp. 587-598.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 400-413, March 1991

007

ON THE BIFURCATION OF RADIALLY SYMMETRIC STEADY-STATE SOLUTIONS ARISING IN POPULATION GENETICS*

K. J. BROWN† AND A. TERTIKAS‡

Abstract. This paper considers a semilinear elliptic equation which arises in a selection-migration model in population genetics, involving two alleles A_1 and A_2 such that A_1 is at an advantage over A_2 in certain subregions and at a disadvantage in others. The system is studied on all of R^n and is assumed to possess radial symmetry. Existence and asymptotic properties of solutions of the corresponding ordinary differential equation are investigated and, by using shooting method type arguments, results are obtained on the bifurcation of solutions from the trivial solutions corresponding to the cases where A_1 or A_2 is extinct. The nature of the results obtained varies according to whether A_1 or A_2 has an overall advantage.

Key words. population genetics, bifurcation theory, indefinite weight functions, radial symmetry

AMS(MOS) subject classifications. 34B15, 35J65, 92A10

1. Introduction. In this paper we discuss the steady-state solutions of the semilinear parabolic problem

$$u_t(x,t)=\Delta u+\lambda g(|x|)f(u) \quad \text{for } x \text{ in } R^n, \quad t\geqq 0 \tag{1.1}$$

where Δ denotes the Laplacian, g assumes both positive and negative values on R^n, and $f:[0,1]\to R$ is a nonnegative concave function such that $f(0)=0=f(1)$.

Such problems arise in population genetics. Consider a model with two alleles A_1 and A_2 corresponding to three possible genotypes: A_1A_1, A_1A_2, and A_2A_2. Let $u(x,t)$ denote the frequency of the allele A_1 at time t at the point x in R^n. Changes in gene frequency are assumed to be caused only by the flow of genes and by selective advantages for certain genotypes in certain subregions of D. Then u satisfies (1.1); the term $\Delta u/\lambda$ represents the effect of gene flow; the term $g(x)f(u)$ with $f(u)=u(1-u)[h(1-u)+(1-h)u]$ for some constant h, $0<h<1$, represents the effect of natural selection where the fitness coefficients of the genotypes A_1A_2 and A_2A_2 relative to A_1A_1 are, respectively, $1-hg(x)$ and $1-g(x)$. The fact that g changes sign on R^n corresponds to the allele A_1 being at an advantage in some parts of R^n and at a disadvantage in others. A more detailed account of this model can be found in Fleming [3].

Since u represents a population frequency, the physically meaningful steady-state solutions of (1.1) satisfy

$$-\Delta u=\lambda g(|x|)f(u) \quad \text{on } R^n\colon 0\leqq u\leqq 1. \tag{1.2}_\lambda$$

It is clear that $(1.2)_\lambda$ has the trivial constant solutions $u\equiv 0$ and $u\equiv 1$ corresponding to the extinction of one of the alleles. We are interested in the existence of nontrivial solutions.

Nontrivial radially symmetric solutions of $(1.2)_\lambda$ satisfy the ordinary differential equation (ODE)

$$\begin{aligned} &u''(r)+\frac{n-1}{r}u'(r)+\lambda g(r)f(u(r))=0 \quad \text{for } r>0,\\ &u'(0)=0, \qquad 0<u(r)<1 \quad \text{for } r>0. \end{aligned} \tag{1.3}_\lambda$$

* Received by the editors November 2, 1988; accepted for publication (in revised form) February 21, 1990.
† Department of Mathematics, Heriot-Watt University, Riccarton, Edinburgh EH14 4AS, Scotland.
‡ Department of Mathematics, University of Crete, Greece.

We have already studied various aspects of $(1.2)_\lambda$ in [2] and [7]. As in [2] and [7] we shall assume throughout that g satisfies

(G_0) $g:[0,\infty)\to R$ is locally Holder continuous and there exists r_0, $R_0>0$ such that $g(r_0)>0$ and $g(r)<0$ for $r>R_0$.

The uniqueness of radially symmetric solutions of $(1.3)_\lambda$ is discussed by Tertikas in [7] under the additional hypothesis that f is concave (this occurs whenever $\frac{1}{3}<h<\frac{2}{3}$). In [2] equation $(1.2)_\lambda$ was studied without assuming that g is radially symmetric or that f is concave and results on the existence and nonexistence of solutions were obtained. These results correspond closely to what might be expected from the model. It was shown that, if g is negative and bounded away from zero at infinity and λ is sufficiently small, then there exist no solutions of $(1.2)_\lambda$; this corresponds to the fact that if the allele A_1 suffers a significant overall disadvantage and the rate of gene flow is sufficiently large, then A_1 and A_2 cannot coexist. On the other hand, if the alleles both hold some kind of advantage in the sense that $\int_{R^n} g\,dx>0$, although g is negative at infinity, then it was shown that, when $n=1, 2$, solutions of $(1.2)_\lambda$ can exist no matter how great the rate of gene flow. The purpose of this paper is to demonstrate how the results of [2] can be considerably sharpened in the case where g is assumed to be radially symmetric so that the problem may be attacked by using ODE methods. Because of the radial symmetry of g we are able to replace the bare existence results of [2] with precise descriptions of the bifurcation diagrams in the (λ, u) plane for $(1.2)_\lambda$. Properties of bifurcation diagrams for $(1.2)_\lambda$ on bounded regions have been studied by Fleming [3] in the case of Neumann boundary conditions and by Hess and Kato [5] in the case of Dirichlet boundary conditions. In general, it is difficult to obtain bifurcation results for problems on unbounded regions because the linearized problem can no longer be formulated as a problem involving compact operators and so it is no longer obvious that the linearized problem has eigenvalues which might correspond to bifurcation points. However, in this special case, using ODE techniques we are able to describe how bifurcation occurs.

We now describe how the paper is organized. In § 2 we prove the existence of radially symmetric solutions of $(1.2)_\lambda$ for sufficiently large λ by constructing appropriate weak sub- and supersolutions. Similar existence theorems are proved in [2] and [7] but the sub- and supersolutions used here are simpler than those used previously. In order to obtain our main results an understanding of the asymptotic properties of solutions as $r\to\infty$ is necessary; in § 3 we prove the required asymptotic results, obtaining conditions in terms of integrals involving g for all solutions to tend to zero at infinity. In § 4 we study $(1.3)_\lambda$ by using the shooting method; the uniqueness results of [7] make possible a very clear and simple description of what happens to solutions as the initial value changes and this description is the principal tool in the proofs of our main existence results. In addition at the end of § 4 we deduce that all solutions of $(1.2)_\lambda$ are radially symmetric; this result is of interest as the radial symmetry of all solutions is proved without the assumption that g is a decreasing function of r (cf. Gidas, Ni, and Nirenberg [4]). Finally, in §§ 5 and 6 we describe our main existence results. In § 5 we show that bifurcation occurs from the zero solution when g is bounded away from zero at infinity and that the bifurcation diagram is as shown in Fig. 1. In § 6 we consider the case where $\int_{R^n} g\,dx>0$ and show that the bifurcation diagram is as shown in Fig. 2, i.e., a nontrivial branch of solutions bifurcates from the trivial branch $u\equiv 1$ at $\lambda=0$.

2. Existence of solutions for large λ. We shall prove the existence of solutions by constructing appropriate sub- and supersolutions.

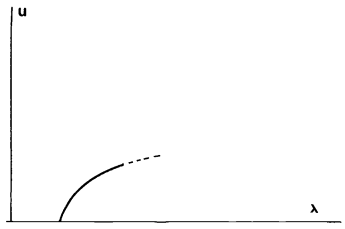


FIG. 1. *Bifurcation diagram when* $\int g < 0$.

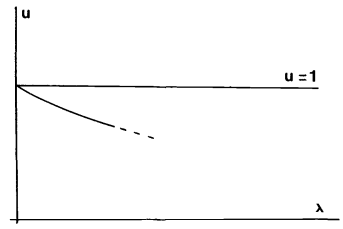


FIG. 2. *Bifurcation diagram when* $\int g > 0$.

Let $D \subset R^n$ be a domain with smooth boundary, and let D_1 be a subregion of D such that ∂D_1 is smooth and $\bar{D}_1 \subset D$. Let $D_2 = D - D_1$, and let $n(x)$ denote the outward normal to D_1 at $x \in \partial D_1$. Suppose that u_1 and u_2 are smooth subsolutions of $-\Delta u = \lambda g(x) f(u)$ on D_1 and D_2, respectively, and that on ∂D_1, $u_1 = u_2$ and $\partial u_1/\partial n \leqq \partial u_2/\partial n$. Then, if we define u by

$$u = u_1 \quad \text{on } D_1, \qquad u = u_2 \quad \text{on } D_2,$$

we say that u is a weak subsolution of $-\Delta u = \lambda g(x) f(u)$ on D. Roughly speaking, u can be regarded as the supremum of the two subsolutions u_1 and u_2. Weak supersolutions are defined similarly. Such weak sub- and supersolutions are discussed by Berestycki and Lions in [1] in the case of bounded regions with Dirichlet boundary conditions and the existence of solutions lying between weak sub- and supersolutions is obtained. Ni in [6] proves the existence of a solution lying between smooth sub- and supersolutions for semilinear elliptic equations on all of R^n; the solution is obtained as the limit of solutions on bounded regions and so the proof in [6] can be adapted to apply to the case of weak sub- and supersolutions.

LEMMA 2.1. *Suppose* $f: [0, 1] \to R$ *satisfies*

(F_0) $\quad f(0) = f(1) = 0, \quad f(u) > 0$ *for* $0 < u < 1, \quad f'(0) > 0, \quad f'(1) < 0.$

There exists $\Lambda_0 > 0$ *such that* $(1.2)_\lambda$ *has arbitrarily small nontrivial radially symmetric weak subsolutions for all* $\lambda > \Lambda_0$.

Proof. By our assumption (G_0) on g there exists an annulus or ball Ω centred at r_0 and a constant $m > 0$ such that $g(|x|) \geqq m$ for all $x \in \Omega$. Let λ_1 denote the principal eigenvalue and φ_1 the corresponding positive radially symmetric eigenfunction such that $\sup_\Omega \varphi_1(x) = 1$ of

$$-\Delta \varphi = \lambda \varphi \quad \text{on } \Omega, \qquad \varphi = 0 \quad \text{on } \partial\Omega.$$

Let $f_1(u) = \lambda m f(u) - \lambda_1 u$. Then $f_1(0) = 0$, and $f_1'(0) = \lambda m f'(0) - \lambda_1$. Hence $f_1'(0) > 0$ provided $\lambda > \lambda_1 / m f'(0) = \Lambda_0$, say. Thus

$$\tau(\lambda) = \sup \{t \in [0, 1];\ \lambda m f(u) > \lambda_1 u \text{ for } 0 \leqq u \leqq t\} > 0 \tag{2.1}$$

for any $\lambda > \Lambda_0$.

Suppose $\lambda > \Lambda_0$ and $0 < \varepsilon < \tau(\lambda)$. Then

$$\Delta(\varepsilon\varphi_1) + \lambda g(|x|) f(\varepsilon\varphi_1) \geqq -\lambda_1 \varepsilon_1 \varphi_1 + \lambda m f(\varepsilon\varphi_1) \geqq 0$$

for all $x \in \Omega$ and so the function

$$\underline{u}_\varepsilon(x) = \begin{cases} \varepsilon\varphi_1(x) & \text{for } x \in \Omega, \\ 0 & \text{for } x \in R^n - \Omega \end{cases}$$

is a weak subsolution of $(1.2)_\lambda$.

LEMMA 2.2. *Let f satisfy the hypotheses of Lemma* 2.1. *Then there exists $\Lambda_1 > 0$ such that $(1.2)_\lambda$ has radially symmetric weak supersolutions arbitrarily close to one for all $\lambda > \Lambda_1$.*

Proof. Let Ω be an annulus or ball such that $g(|x|) \leqq -m$ for $x \in \Omega$ where m is a positive constant. Now u is a solution of

$$-\Delta u = \lambda g(|x|) f(u) \quad \text{in } \Omega, \qquad u = 1 \quad \text{on } \partial\Omega \tag{2.2}$$

if and only if $v = 1 - u$ is a solution of

$$-\Delta v = \lambda \hat{g}(|x|) \hat{f}(v) \quad \text{in } \Omega, \qquad v = 0 \quad \text{on } \partial\Omega \tag{2.3}$$

where $\hat{g}(r) = -g(r)$ and $\hat{f}(v) = f(1 - v)$. By Lemma 2.1 there exist arbitrarily small subsolutions of the form $\varepsilon\varphi_1$ of (2.3) and so there exist supersolutions of (2.2) of the form $1 - \varepsilon\varphi_1$ arbitrarily close to 1. Therefore a radially symmetric weak supersolution of $(1.2)_\lambda$ is given by

$$\bar{u}_\varepsilon(x) = \begin{cases} 1 - \varepsilon\varphi_1(x) & \text{for } x \in \Omega, \\ 1 & \text{for } x \in R^n - \Omega \end{cases}$$

and so the proof is complete.

THEOREM 2.3. *Let f satisfy the hypotheses of Lemma* 2.1. *Then for all $\lambda > \max\{\Lambda_0, \Lambda_1\}$, problem $(1.2)_\lambda$ has a nontrivial radially symmetric solution u_λ such that $\sup_{x \in R^n} u_\lambda(x) \to 1$ and $\inf_{x \in R^n} u_\lambda(x) \to 0$ as $\lambda \to \infty$.*

Proof. By Lemmas 2.1 and 2.2 we have that for each $\lambda > \max\{\Lambda_0, \Lambda_1\}$ there exist weak sub- and supersolutions $\underline{u}_\varepsilon$ and $\bar{u}_\varepsilon$. By choosing ε sufficiently small we can ensure that $\underline{u}_\varepsilon < \bar{u}_\varepsilon$. Hence by Ni's existence theorem in [6] there exists a solution to $(1.2)_\lambda$. Moreover, since g, $\underline{u}_\varepsilon$, and $\bar{u}_\varepsilon$ are radially symmetric, the iteration scheme of [6] will give us a radially symmetric solution.

The proof of the second part of the theorem depends on the construction of the weak sub- and supersolutions. In the notation of Lemma 2.1 we have that $u_\lambda(x) \geqq \tau(\lambda)\varphi_1(x)$ for all $x \in \Omega$. It is clear from (2.1) that $\lim_{\lambda \to \infty} \tau(\lambda) = 1$ and so $\limsup_{\lambda \to \infty} u_\lambda(x) = 1$. Similarly, in the notation of Lemma 2.2 we have that $u_\lambda(x) \leqq 1 - \tau(\lambda)\varphi_1(x)$, and so $\liminf_{\lambda \to \infty} u_\lambda(x) = 0$.

3. Asymptotic behaviour of solutions. The main objective of this section is to obtain conditions which ensure that all solutions of $(1.3)_\lambda$ tend to zero as $r \to \infty$. We also obtain information on rates of decay by investigating the asymptotic behavior of $u'(r)/u(r)$. We shall later require asymptotic results on solutions of the linearization of $(1.3)_\lambda$ about zero. So that we may deal with $(1.3)_\lambda$ where $f(u) = u(1-u)((1-h)u + h(1-u))$ and its linearization simultaneously we assume throughout this section only that f satisfies

$$\text{(F)} \qquad f : [0, 1] \to R^+, \qquad f(0) = 0, \qquad f'(0) > 0, \qquad 0 < f(u) \leqq u \quad \text{for } 0 < u < 1.$$

The assumption that $g(r)<0$ for r sufficiently large is essential for our results of this section. If g is allowed to oscillate arbitrarily at ∞, then $\lim_{r\to\infty} u(r)$ need not exist. For example, $u(x)=(1+e^{\cos x})^{-1}$ satisfies the differential equation

$$u''(x)-[\cos x+(e^{\cos x}-1)(e^{\cos x}+1)^{-1}]u(1-u)=0 \quad \text{for } x\in R, \qquad u'(0)=0.$$

First we discuss the case $n=1, 2$. Our main result is the following.

THEOREM 3.1. *Let f satisfy* (F), *and let u be a solution of* $(1.3)_\lambda$.

(i) *Suppose $n=1$. Then $\lim_{r\to\infty} u(r)=0$ if and only if $\int^\infty rg(r)\,dr=-\infty$.*

(ii) *Suppose $n=2$. Then $\lim_{r\to\infty} u(r)=0$ if and only if $\int^\infty r\ln rg(r)\,dr=-\infty$.*

For $n=2$ the classical transformation

$$s=\ln r, \qquad v(s)=u(r)$$

transforms $(1.3)_\lambda$ into

$$v_{ss}+\lambda\, e^{2s}g(e^s)f(v(s))=0.$$

Thus we can investigate the asymptotics of $(1.3)_\lambda$ for $n=1, 2$ by studying the asymptotics of solutions of

$$w''(x)+G(x)f(w(x))=0 \quad \text{for } x>0, \qquad 0<w<1 \tag{3.1}$$

where $G(x)=\lambda g(x)$ when $n=1$, and $G(x)=\lambda\, e^{2x}g(e^x)$ when $n=2$. Clearly, $G(x)<0$ for $x>R_0$.

Suppose w is a solution of (3.1). Since w is bounded and eventually convex, it is easy to see that w is eventually decreasing, $\lim_{x\to\infty} w'(x)=0$, and $\lim_{x\to\infty} w(x)$ exists.

LEMMA 3.2. *If $\int^\infty xG(x)\,dx=-\infty$, then $\lim_{x\to\infty} w(x)=0$.*

Proof. Multiplying (3.1) by x and integrating gives

$$xw'(x)-R_0w'(R_0)-w(x)+w(R_0)+\int_{R_0}^{x} tG(t)f(w(t))\,dt=0.$$

Suppose that $\lim_{x\to\infty} w(x)=\beta>0$. Then $\lim_{x\to\infty} f(w(x))=f(\beta)>0$ and so $\int^\infty tG(t)f(w(t))\,dt=-\infty$. Hence $xw'(x)\to\infty$ as $x\to\infty$, which is impossible as $w'(x)\leqq 0$ for $x>R_0$. Thus $\lim_{x\to\infty} w(x)=0$.

LEMMA 3.3. *If $\int^\infty xG(x)\,dx>-\infty$, then $\lim_{x\to\infty} w(x)>0$.*

Proof. Suppose the contrary, i.e., $\lim_{x\to\infty} w(x)=0$. Since $f'(0)>0$, there exists $\alpha>R_0$ such that $f(w(\cdot))$ is a decreasing function for $x>\alpha$. Integrating (3.1), we obtain

$$w'(X)-w'(x)+\int_x^X G(t)f(w(t))\,dt=0$$

where $\alpha\leqq x\leqq X$. Since, for $t>x$, $f(w(t))\leqq f(w(x))\leqq w(x)$, it follows that

$$w'(X)-w'(x)+w(x)\int_x^X G(t)\,dt\leqq 0,$$

and so, letting $x\to\infty$,

$$\frac{w'(x)}{w(x)}\geqq\int_x^\infty G(t)\,dt. \tag{3.2}$$

Integrating from $x=\alpha$ to $x=\infty$, we obtain

$$\int_\alpha^\infty\int_x^\infty G(t)\,dt\,dx=\int_\alpha^\infty (t-\alpha)G(t)\,dt>-\infty$$

whereas

$$\int_{\alpha}^{\infty} \frac{w'(x)}{w(x)}\,dx = \lim_{x\to\infty} [\ln w(x) - \ln w(\alpha)] = -\infty,$$

which contradicts (3.2). Hence $\lim_{x\to\infty} w(x) > 0$.

It is clear that Theorem 3.1 follows directly from the two previous lemmas.

We now study the case $n \geqq 3$. Suppose that u is a solution of $(1.3)_\lambda$. Since

$$(r^{n-1}u')' = -\lambda r^{n-1} g(r) f(u),$$

$r^{n-1}u'$ is eventually increasing and so is eventually of one sign. Hence u is eventually monotone, and so $\lim_{r\to\infty} u(r)$ exists.

The following result, which can be proved by an argument similar to that used in the proof of Lemma 3.2, enables us to evaluate this limit.

LEMMA 3.4. *Suppose f satisfies (F) and $\int^{\infty} rg(r)\,dr = -\infty$. Let u be a solution of $(1.3)_\lambda$ such that $\lim_{r\to\infty} u(r) = \beta$. Then $f(\beta) = 0$.*

We can now give our main asymptotic result for the case $n \geqq 3$.

THEOREM 3.5. *Suppose $n \geqq 3$ and u is a solution of $(1.3)_\lambda$. Then $\lim_{r\to\infty} u(r) = 0$ when either*

(i) $\int^{\infty} rg(r)\,dr = -\infty$ *and* $f(u) = u$, *or*

(ii) $\lim_{r\to\infty} r^2 g(r) = -\infty$ *and* $f:[0,1] \to R$ *satisfies* (F_0).

Proof. Part (i) is immediate from Lemma 3.4.

(ii) It follows from Lemma 3.4 that either $\lim_{r\to\infty} u(r) = 0$ or $\lim_{r\to\infty} u(r) = 1$. Suppose that $\lim_{r\to\infty} u(r) = 1$. Let $v(r) = r^{1/2(n-1)}(u(r) - 1)$. Then $v(r) < 0$ for all r and

$$v''(r) + \frac{1}{4} r^{-2} \left\{ 4\lambda r^2 g(r) \frac{f(u(r))}{u(r)-1} - (n-1)(n-3) \right\} v = 0,$$

i.e.,

$$v''(r) + \tfrac{1}{4} q(r) r^{-2} v = 0. \tag{3.3}$$

Since $\lim_{r\to\infty} u(r) = 1$, $\lim_{r\to\infty} f(u(r))/(u(r)-1) = f'(1)$, and so $\underline{\lim}_{r\to\infty} q(r) > 1$. If $k > 1$ every solution of the Euler equation

$$w''(r) + \tfrac{1}{4} k r^{-2} w(r) = 0$$

is oscillatory. Hence it follows from the Sturmian comparison theorem that every solution of (3.3) is oscillatory, i.e., that v is oscillatory. This is impossible and so $\lim_{r\to\infty} u(r) = 0$.

Finally, in this section we investigate the asymptotic behavior of u'/u. In the next two results we deal only with the case where $n \geqq 2$; the same results hold when $n = 1$ and the corresponding proofs are simpler.

LEMMA 3.6. *Suppose f satisfies* (F) *and that g is bounded below. If u is a solution of $(1.3)_\lambda$ such that $\lim_{r\to\infty} u(r) = 0$, then $u'(r)/u(r)$ is uniformly bounded for $r > 0$.*

Proof. Let $z(r) = u'(r)/u(r)$. Then $z' = u''/u - z^2$, and so it follows from $(1.3)_\lambda$ that

$$z' + z^2 + \frac{n-1}{r} z + \lambda g(r) \frac{f(u)}{u} = 0.$$

Hence

$$z' + \frac{1}{2} z^2 = -\lambda g(r) \frac{f(u)}{u} - \frac{n-1}{r} z - \frac{1}{2} z^2.$$

Since $f(u)/u \leqq 1$ and $g(r)$ is bounded, there exists $K>0$ such that $z'+\frac{1}{2}z^2<0$ whenever $z<-K$. Hence, when $z<-K$, $z'/z^2<-\frac{1}{2}$, and so $d/dz(1/z)>\frac{1}{2}$. Thus, if $z(x_0)<-K$, $1/z$ is an increasing function, and so $z(x)<-K$ for all $x>x_0$. Since, however, $d/dz(1/z)>\frac{1}{2}$, $1/z$ must eventually become positive and this is impossible. Hence $z(x)\geqq -K$ for all x and the proof is complete.

THEOREM 3.7. *Suppose f satisfies* (F), *g is bounded below, and g is bounded away from zero at ∞, i.e., there exist $k, R_0>0$ such that $g(r)<-k$ for $r\geqq R_0$. If u is a solution of* $(1.3)_\lambda$, *then* $\lim_{r\to\infty} e^{\beta r}u(r)=0$ *for any* $\beta<\sqrt{\lambda k f'(0)}$.

Proof. It follows from Theorems 3.1 and 3.5 that $\lim_{r\to\infty} u(r)=0$. Since

$$\frac{u''}{u}=-\frac{n-1}{r}\frac{u'}{u}-\lambda g(r)\frac{f(u)}{u}$$

and $\lim_{r\to\infty} f(u)/u=f'(0)$, we have that

$$u''(r)\geqq \alpha u(r) \quad \text{for } r \text{ sufficiently large}$$

for any constant $\alpha<\lambda k f'(0)$.

It is now easy to complete the proof by using standard differential inequality arguments.

The previous theorem shows that solutions decay exponentially when g is bounded away from zero at infinity. The final two results in this section show that exponential decay does not occur when $|g|$ is small at infinity.

THEOREM 3.8. *Suppose $n=1$, f satisfies* (F), *and $\int_0^\infty g(r)\,dr$ converges. If u is a solution of* $(1.3)_\lambda$, *then* $\lim_{r\to\infty} u'(r)/u(r)=0$.

Proof. Dividing $(1.3)_\lambda$ by u and integrating, we obtain

$$\int_0^r \frac{u''}{u}\,ds+\lambda\int_0^r g(s)\frac{f(u(s))}{u(s)}\,ds=0.$$

Since $f(u)/u$ is bounded, the second integral converges and so the first integral must converge. Hence

$$\lim_{r\to\infty}\left\{\frac{u'(r)}{u(r)}+\int_0^r\left[\frac{u'(s)}{u(s)}\right]^2 ds\right\} \text{ exists.}$$

If $\int_0^\infty [u'(s)/u(s)]^2\,ds$ converges the conclusion is obvious. Suppose $\int_0^\infty [u'(s)/u(s)]^2\,ds$ diverges. Then there exists $R>0$ such that

$$-\frac{u'(r)}{u(r)}>\frac{1}{2}\int_0^r\left[\frac{u'(s)}{u(s)}\right]^2 ds \quad \text{for } r\geqq R. \tag{3.4}$$

Let $v(r)=\int_0^r [u'(s)/u(s)]^2\,ds$. Then $\lim_{r\to\infty} v(r)=\infty$. However, (3.4) implies that $v'(r)>\frac{1}{4}[v(r)]^2$, and so $d/dr(1/v)\leqq -\frac{1}{4}$ for $r\geqq R$. Hence $1/v(r)\leqq -\frac{1}{4}r+C$ for some constant C and this is a contradiction.

COROLLARY 3.9. *Suppose $n=2$, f satisfies* (F) *and $\int_0^\infty rg(r)\,dr$ converges. If u is a solution of* $(1.3)_\lambda$, *then* $\lim_{r\to\infty} ru'(r)/u(r)=0$.

Proof. By using the transformation $s=\ln r$ and $v(s)=u(r)$, equation $(1.3)_\lambda$ can be transformed into

$$v_{ss}+\lambda G(s)f(v)=0$$

where $G(s)=e^{2s}g(e^s)$. Then $\int_0^\infty G(s)\,ds=\int^\infty rg(r)\,dr$ and so $\lim_{r\to\infty} v_s(s)/v(s)=0$ by Theorem 3.8. But $v_s=ru_r$ and the proof is complete.

4. The shooting method. Uniqueness of solutions of boundary value problems plays a vital role in the development of the shooting method for the study of $(1.3)_\lambda$. We require the following results which are proved in [7] and [8], respectively.

THEOREM 4.1. *Suppose f satisfies* (F_0) *of Lemma* 2.1 *and f is concave, i.e., $f''(u) \leqq 0$ for $0 \leqq u \leqq 1$.*

(i) *Equation* $(1.2)_\lambda$ *has at most one nontrivial radially symmetric solution u such that* $\lim_{|x|\to\infty} u(x) = 0$.

(ii) *Let $R > 0$, $k \geqq 0$, and let $B_R = \{x \in R^n : |x| \leqq R\}$. Then the boundary value problem*

$$\begin{aligned} &-\Delta u(x) = \lambda g(x) f(u) \quad \text{in } B_R, \\ &0 < u(x) < 1 \quad \text{in } B_R, \qquad u = k \quad \text{on } \partial B_R \end{aligned} \tag{4.1}$$

has at most one radially symmetric solution.

For the remainder of this section we assume that f satisfies (F_0) and is concave. In addition we assume that

$$\begin{aligned} &\int^{\infty} r g(r)\, dr = -\infty && \text{if } n = 1, \\ &\int r \ln r\, g(r)\, dr = -\infty && \text{if } n = 2, \\ &\lim_{r\to\infty} r^2 g(r) = -\infty && \text{if } n \geqq 3. \end{aligned} \tag{G_1}$$

It follows from Theorems 3.1 and 3.5 that all solutions of $(1.3)_\lambda \to 0$ as $|x| \to \infty$ provided f satisfies (F_0) and g satisfies (G_0) and (G_1).

Consider the initial value problem

$$\begin{aligned} &u''(r) + \frac{n-1}{r} u'(r) + \lambda g(r) f(u) = 0 \quad \text{for } r > 0, \\ &u'(0) = 0, \qquad u(0) = p \quad \text{where } 0 < p < 1. \end{aligned} \tag{$4.2)_\lambda$}$$

We denote a solution of $(4.2)_\lambda$ by $u(\cdot, p, \lambda)$. It follows from standard theorems on the continuous dependence of solutions on parameters and on initial data that $(p, \lambda) \to u(\cdot, p, \lambda)$ is a continuous function from $[0, 1] \times [0, \infty)$ to $C[0, R]$ for any $R > 0$.

Let $A(\lambda) = \{p \in (0, 1)$: there exists $R > 0$ such that $0 < u(r, p, \lambda) < 1$ for $0 \leqq r < R$ and $u(R, p, \lambda) = 0\}$ and $B(\lambda) = \{p \in (0, 1)$: there exists $R > 0$ such that $0 < u(r, p, \lambda) < 1$ for $0 \leqq r < R$ and $u(R, p, \lambda) = 1\}$.

It is straightforward to prove from the continuous dependence of solutions on parameters that $A(\lambda)$ and $B(\lambda)$ are open disjoint subsets of $(0, 1)$. Thus the following result is obvious.

LEMMA 4.2. (i) *Let $p \in (0, 1)$. Then $u(\cdot, p, \lambda)$ is a solution of* $(1.3)_\lambda$ *if and only if $p \notin A(\lambda) \cup B(\lambda)$.*

(ii) *If $A(\lambda)$ and $B(\lambda)$ are both nonempty, then there exists a solution of* $(1.3)_\lambda$.

We next use the uniqueness result of Theorem 4.1(ii) to obtain further information about $A(\lambda)$ and $B(\lambda)$.

LEMMA 4.3. (i) *If $A(\lambda) \neq \emptyset$, then there exists $p(\lambda)$, $0 < p(\lambda) \leqq 1$, such that $A(\lambda) = (0, p(\lambda))$.*

(ii) *Suppose* $(1.3)_\lambda$ *has a solution. Then there exists $p(\lambda)$, $0 < p(\lambda) < 1$ such that $A(\lambda) = (0, p(\lambda))$, $B(\lambda) = (p(\lambda), 1)$ and $u(\cdot, p(\lambda), \lambda)$ is the solution of* $(1.3)_\lambda$.

Proof. If $p \in A(\lambda)$, there exists $R > 0$ such that $0 < u(r, p, \lambda) < 1$ for $0 < r < R$ and $u(R, p, \lambda) = 0$. Assume that $q < p$ and $q \notin A(\lambda)$. Then either $u(r_0, q, \lambda) = 1$ for some r_0, $0 < r_0 \leqq R$ or $u(r, q, \lambda) > 0$ for $0 < r \leqq R$. In either case there exists r_1, $0 < r_1 < R$, such that $u(r_1, q, \lambda) = u(r_1, p, \lambda) (= k$, say). Thus $u(\cdot, p, \lambda)$ and $u(\cdot, q, \lambda)$ correspond to

distinct radially symmetric solutions of (4.1), which is impossible. Hence there exists $p(\lambda) \in (0, 1)$ such that $A(\lambda) = (0, p(\lambda))$. Part (ii) follows from the fact that under our hypotheses on f and g, $(1.3)_\lambda$ has at most one solution.

We shall use the preceding result to investigate the bifurcation of solutions. To this end let

$$\Lambda = \{\lambda > 0: \text{there exists a solution of } (1.3)_\lambda\}.$$

It follows from straightforward continuous dependence arguments that Λ is an open set. The results of § 2 show that $\lambda \in \Lambda$ provided that λ is sufficiently large. If $\lambda \in \Lambda$, by Theorem 4.1 equation $(1.3)_\lambda$ has a unique solution, which we denote by u_λ. The next result shows how $p(\lambda)$ and u_λ behave as λ approaches the boundary of Λ.

LEMMA 4.4. *Suppose that $\mu > 0$ is such that $(\mu, \mu + \varepsilon) \subset \Lambda$ for some $\varepsilon > 0$ but $\mu \notin \Lambda$. Then, as $\lambda \to \mu^+$,*

(i) *Either $p(\lambda) \to 1$ and $u_\lambda \to 1$ uniformly on compact subintervals of $[0, \infty)$, or*

(ii) *$p(\lambda) \to 0$ and $u_\lambda \to 0$ uniformly on compact subintervals of $[0, \infty)$.*

Proof. Since $\mu \notin \Lambda$, either $A(\mu) = (0, 1)$ or $B(\mu) = (0, 1)$. Suppose $A(\mu) = (0, 1)$. Let $q \in (0, 1)$. Then $q \in A(\mu)$. Using continuous dependence of solutions it is easy to show that $q \in A(\lambda)$, provided that λ is sufficiently close to μ. But, if $q \in A(\lambda)$, then by Lemma 4.3, $p(\lambda) > q$. Since q can be chosen arbitrarily close to 1, it follows that $p(\lambda) \to 1$ as $\lambda \to \mu^+$. By the continuous dependence of solutions on initial data $u_\lambda(r) = u(r, p(\lambda), \lambda) \to 1$ uniformly on compact subintervals of $[0, \infty)$.

In the case where $B(\mu) = (0, 1)$ a similar argument shows that (ii) holds.

Finally in this section we make use of our previous results on the shooting method to prove that under certain hypotheses all solutions of $(1.2)_\lambda$ are radially symmetric.

LEMMA 4.5. (i) *If $A(\lambda) \neq \emptyset$, then there are arbitrarily small radially symmetric weak subsolutions of $(1.2)_\lambda$.*

(ii) *If $B(\lambda) \neq \emptyset$, then there are radially symmetric weak supersolutions of $(1.2)_\lambda$ which are arbitrarily close to 1 on any compact subinterval of $[0, \infty)$.*

Proof. If $A(\lambda) \neq \emptyset$, then $A(\lambda) = (0, p(\lambda))$. For all $p < p(\lambda)$ let $R(p)$ denote the first positive zero of $u(r, p, \lambda)$. Let

$$v_p(r) = \begin{cases} u(r, p, \lambda) & \text{for } 0 \leqq r \leqq R(p), \\ 0 & \text{for } r > R(p). \end{cases}$$

Clearly, $v_p(r)$ is a radially symmetric weak subsolution of $(1.2)_\lambda$. Let $\varepsilon > 0$. There exists $p_1 \leqq p(\lambda)$ such that $|u(r, p, \lambda)| < \varepsilon$ for all $p \leqq p_1$ and $r \leqq R_0$. If $R(p) \leqq R_0$, clearly $v_p(r) < \varepsilon$ for all $r > 0$. If $R(p) > R_0$, then since

$$u'' + \frac{n-1}{r} u' = -\lambda g(r) f(u) > 0 \quad \text{for } R_0 < r < R(p),$$

u has no local maximum turning points for $R_0 < r < R(p)$, and so $u(r, p, \lambda)$ is a decreasing function on $(R_0, R(p))$; hence $v_p(r) < \varepsilon$ for all $r > 0$. Hence arbitrarily small radially symmetric weak subsolutions v_p can be obtained by making p sufficiently small.

Part (ii) has a similar proof.

LEMMA 4.6. *Suppose that $(1.2)_\lambda$ has a nonradially symmetric solution u. Then $(1.2)_\lambda$ also has a radially symmetric solution.*

Proof. Assume the contrary. Then $A(\lambda) = (0, 1)$ or $B(\lambda) = (0, 1)$. Suppose $A(\lambda) = (0, 1)$. Then by Lemma 4.5 there exist arbitrarily small radially symmetric subsolutions, with compact support. It follows from the maximum principle that $0 < u(x) < 1$ for all x in R^n. Hence there exists a radially symmetric weak subsolution $\underline{u}$ such that

$0 \leqq \underline{u}(|x|) \leqq u(x)$ for all $x \in R^n$. Using Ni's results in [6], we have that the iteration starting from the radially symmetric subsolution $\underline{u}$ converges to a radially symmetric solution lying below the (super) solution u. This is a contradiction, and so the proof is complete. A similar contradiction arises if $B(\lambda) = (0, 1)$.

THEOREM 4.7. *Suppose f satisfies* (F_0) *and is concave and g satisfies* (G_1). *Then every solution of* $(1.2)_\lambda$ *is radially symmetric.*

Proof. Suppose u is a nonradially symmetric solution of $(1.2)_\lambda$. Then there exists a radially symmetric solution of $(1.2)_\lambda$, and so by Lemma 4.3 $A(\lambda)$ and $B(\lambda)$ are nonempty. Hence by Lemma 4.5 there exist radially symmetric weak sub- and supersolutions $\underline{u}$ and $\bar{u}$ such that $\underline{u} \leqq u \leqq \bar{u}$. Ni's results in [6] imply the existence of radially symmetric solutions u_1 and u_2 such that $\underline{u} \leqq u_1 \leqq u \leqq u_2 \leqq \bar{u}$. Since u is nonradially symmetric, $u_1 \neq u$ and $u_2 \neq u$, and so $u_1 \neq u_2$. But this is impossible because of Theorem 4.1, and so every solution of $(1.2)_\lambda$ is radially symmetric.

5. The case $\int_{R^n} g(x)\,dx < 0$. In this section we shall discuss the equation

$$-\Delta u = \lambda g(|x|)u(1-u)(h(1-u)+(1-h)u) \quad \text{on } R^n: 0 < u < 1 \tag{5.1}_\lambda$$

where $\frac{1}{3} < h < \frac{2}{3}$. Then $f(u) = u(1-u)(h(1-u)+(1-h)u)$ satisfies (F_0) and is concave. We shall assume throughout the section that g satisfies (G_1). Thus the results of § 4 apply to equation $(5.1)_\lambda$.

We first show that $(1.3)_\lambda$ has no solutions when λ is sufficiently small.

LEMMA 5.1. *If $-\infty \leqq \int_{R^n} g(x)\,dx < 0$, then there exists $\lambda_0 > 0$ such that*

$$\int_{R^n} |\nabla u|^2\,dx \geqq \lambda_0 \int_{R^n} g(x)f(u)u\,dx$$

for all $u \in H^1(R^n)$ such that $0 \leqq u \leqq 1$ and $\int_{R^n} gu^2\,dx > 0$.

Proof. Choose $R > R_0$ such that $\int_B g\,dx < 0$ where $B = \{x \in R^n: |x| \leqq R\}$. It is proved in [2] that there exists $\lambda_1 > 0$ such that

$$\int_B |\nabla u|^2\,dx \geqq \lambda_1 \int_B u^2\,dx$$

for all $u \in H^1(B)$ such that $\int_B gu^2\,dx > 0$.

Let $K = \sup\{|g(x)|: x \in B\}$ and $M = \sup\{f(u)/u: 0 \leqq u \leqq 1\}$. Let $u \in H^1(R^n)$ such that $0 \leqq u \leqq 1$ and $\int_{R^n} gu^2\,dx > 0$. Then $\int_B gu^2\,dx > 0$ and so

$$\begin{aligned}\int_{R^n} |\nabla u|^2\,dx &\geqq \int_B |\nabla u|^2\,dx \geqq \lambda_1 \int_B u^2\,dx \\ &\geqq (KM)^{-1}\lambda_1 \int_B g(x)f(u)u\,dx \\ &\geqq (KM)^{-1}\lambda_1 \int_{R^n} g(x)f(u)u\,dx.\end{aligned}$$

THEOREM 5.2. *Suppose $-\infty \leqq \int_{R^n} g(x)\,dx < 0$, λ_0 is as in Lemma* 5.1, *and $\lambda < \lambda_0$. Then there does not exist a solution of* $(5.1)_\lambda$.

Proof. Assume the contrary, that $(5.1)_\lambda$ has a solution u. Then by Lemma 4.3, $A(\lambda)$ is nonempty, and so there exists $p \in (0, 1)$ and $R > 0$ such that $0 < u(r, p, \lambda) < 1$ for $r \in (0, R)$ and $u(R, p, \lambda) = 0$. Define v by $v(r) = u(r, p, \lambda)$ for $0 \leqq r \leqq R$ and $v(r) = 0$ for $r > R$. Clearly, $0 \leqq v < 1$ and $v \in H^1(R^n)$. We now show that $\int_{R^n} gv^2\,dx > 0$. Let

$w = v^2/f(v)$. Since v is bounded away from 1, $w \in L^2(R^n)$. Moreover,

$$\nabla w = v/f^2(v)[2f(v) - vf'(v)]\nabla v,$$

and so $w \in H^1(R^n)$. Since v satisfies

$$-\Delta v = \lambda g(r)f(v) \quad \text{on } B = \{x \in R^n : |x| \leqq R\}, \qquad v(x) = 0 \quad \text{on } \partial B \tag{5.2}$$

multiplying by w and integrating, we obtain

$$\lambda \int_B gv^2\,dx = -\int_B \Delta v\, w = \int_B \nabla v \cdot \nabla w\,dx$$

$$= \int_B v/f^2(v)[2f(v) - vf'(v)]|\nabla v|^2\,dx > 0$$

as $f(v) = v(1-v)(h(1-v) + (1-h)v)$. Hence $\int_{R^n} gv^2\,dx > 0$. Therefore by Lemma 5.1

$$\int_{R^n} |\nabla v|^2\,dx \geqq \lambda_0 \int_{R^n} g(x)f(v)v\,dx.$$

However, multiplying (5.2) by v and integrating gives

$$\int_{R^n} |\nabla v|^2\,dx = \lambda \int_{R^n} g(x)f(v)v\,dx,$$

and this is a contradiction.

The previous theorem shows that Λ is bounded away from zero. The following results show that bifurcation from the zero solution occurs at points on the boundary of Λ.

THEOREM 5.3. *Suppose* $-\infty \leqq \int_{R^n} g(x)\,dx < 0$. *Suppose* $\mu \notin \Lambda$ *but* $(\mu, \mu + \varepsilon) \subset \Lambda$ *for some* $\varepsilon > 0$. *Then*

(i) $u_\lambda \to 0$ *as* $\lambda \to \mu^+$,

(ii) μ *is an eigenvalue for the linear problem*

$$-\Delta u = \lambda g(r)f'(0)u \quad \text{for } x \text{ in } R^n, \qquad \lim_{x\to\infty} u(x) = 0 \tag{5.3}$$

and there exists a corresponding nonnegative radially symmetric eigenfunction.

Proof (i). Suppose the contrary. Then by Lemma 4.4 $u_\lambda \to 1$ uniformly on compact subintervals of $[0, \infty)$ as $\lambda \to \mu^+$. Choose $R > R_0$ such that $\int_{|x|\leqq R} g(x)\,dx < 0$. Now $u = u_\lambda$ satisfies

$$\frac{(r^{n-1}u')'}{1-u} + \frac{\lambda r^{n-1}g(r)f(u)}{1-u} = 0 \quad \text{for } r > 0.$$

Integrating from 0 to R, we obtain

$$\frac{R^{n-1}u'(R)}{1-u(R)} - \int_0^R \frac{s^{n-1}[u'(s)]^2}{(1-u(s))^2}\,ds + \lambda \int_0^R \frac{s^{n-1}g(s)f(u(s))}{1-u(s)}\,ds = 0. \tag{5.4}$$

Since $\lim_{r\to\infty} u(r) = 0$ and $u''(r) > 0$ at any critical point of u when $r > R_0$, it follows that $u'(R) \leqq 0$. Hence by (5.4)

$$\int_0^R \frac{s^{n-1}g(s)f(u(s))}{1-u(s)}\,ds > 0.$$

As $\lambda \to \mu^+$, $u_\lambda \to 1$ and so $f(u_\lambda(s))/(1-u_\lambda(s)) \to -f'(1)$ uniformly on $[0, R]$. Hence $f'(1)\int_0^R s^{n-1}g(s)\,ds \leqq 0$, and this is a contradiction.

(ii) Suppose $\lambda_k \to \mu^+$. Let u_k denote u_{λ_k} and let r_k denote the point where u_k attains its maximum value. Since u_k is decreasing on $[R_0, \infty)$, $r_k \leqq R_0$. Let $v_k = u_k/u_k(r_k)$. Then

$$(r^{n-1}v_k')' + \lambda_k r^{n-1} g(r)[f(u_k)/u_k]v_k = 0 \quad \text{for } r > 0, \qquad v_k'(0) = 0.$$

Hence

$$r^{n-1}v_k'(r) = -\lambda_k \int_0^r s^{n-1} g(s) \left[\frac{f(u_k)}{u_k}\right] v_k \, ds. \tag{5.5}$$

Let $R > R_0$. It follows from (5.5) that $\{v_k\}$ is a bounded, equicontinuous sequence of functions on $[0, R]$. By considering the integral equation satisfied by v_k and letting $k \to \infty$, it is straightforward to show that $v'(0) = 0$ and that v must satisfy $-\Delta v = \lambda g(r) f'(0) v$ on $[0, R]$. Using a standard diagonalization procedure we can find a subsequence $\{v_k\}$ converging to v on $[0, \infty)$ such that v satisfies $-\Delta v = \lambda g(r) f'(0) v$ on $[0, \infty)$. Since $v_k \geqq 0$ for all k, $v \geqq 0$. As v_k attains the value one at some point of $[0, R_0)$ for all k, $v \not\equiv 0$. Finally, as g satisfies (G_1), it follows from the results of § 3 that $\lim_{r\to\infty} v(r) = 0$.

By placing further restrictions on g we can ensure that bifurcation occurs from the zero solution at only a single point.

LEMMA 5.4. *Suppose that g is bounded below and that g is bounded away from zero as $r \to \infty$. If μ and ν are eigenvalues of* (5.3) *corresponding to positive eigenfunctions u and v, then $\mu = \nu$.*

Proof. We may suppose without loss of generality that $\mu \leqq \nu$. Multiplying the u-equation by u and integrating, we obtain

$$r^{n-1}u'(r)u(r) - \int_0^r s^{n-1}(u')^2 \, ds + \mu f'(0) \int_0^r s^{n-1} g(s) u^2 \, ds = 0. \tag{5.6}$$

Multiplying the v-equation by u^2/v and integrating we obtain

$$\begin{aligned} &r^{n-1}\frac{v'(r)}{v(r)}[u(r)]^2 - \int_0^r s^{n-1} v' \left[\frac{2uu'}{v} - \frac{u^2 v'}{v^2}\right] ds \\ &\quad + \nu f'(0) \int_0^r s^{n-1} g(s) u^2 \, ds = 0. \end{aligned} \tag{5.7}$$

By Lemma 3.6, u'/u and v'/v are bounded, and by Theorem 3.7, u, v, u', v' decay exponentially. Hence letting $r \to \infty$ and then combining appropriate multiples of (5.6) and (5.7), we obtain

$$\nu \int_0^\infty s^{n-1}(u')^2 \, ds - \mu \int_0^\infty s^{n-1} v' \left[\frac{2uu'}{v} - \frac{u^2 v'}{v^2}\right] ds = 0,$$

i.e.,

$$\int_0^\infty s^{n-1} \left[(\nu - \mu)(u')^2 + \mu \left(\frac{uv'}{v} - u'\right)^2\right] ds = 0.$$

Hence $\nu = \mu$ and the proof is complete.

It follows from the preceding results that the boundary of Λ must coincide with the unique eigenvalue of (5.3) corresponding to a positive eigenfunction. Thus we have Theorem 5.5.

THEOREM 5.5. *Suppose that g is bounded below, and that g is bounded away from zero as $r \to \infty$. Then there exists $\mu_0 > 0$ such that for problem $(5.1)_\lambda$, $\Lambda = (\mu_0, \infty)$ and $u_\lambda \to 0$ as $\lambda \to \mu_0^+$.*

6. Bifurcation from $u \equiv 1$. We again discuss $(5.1)_\lambda$ with the assumption that $\frac{1}{3} < h < \frac{2}{3}$ (so that f is concave). We restrict attention to the cases $n = 1$, 2. We shall assume that g satisfies

$$(G_2) \qquad \begin{aligned} &\lim_{r\to\infty} r^2 g(r) = -\infty \quad \text{when } n = 1, \\ &\lim_{r\to\infty} r^2 (\ln r)^2 g(r) = -\infty \quad \text{when } n = 2. \end{aligned}$$

Clearly, (G_2) implies (G_1), and so, if g satisfies (G_2), equation $(5.1)_\lambda$ has at most one solution u, and u must be radially symmetric and $\lim_{r\to\infty} u(r) = 0$.

We shall also assume that g satisfies

$$(G_3) \qquad \int_{R^n} g(|x|)\, dx > 0.$$

Hypotheses (G_2) and (G_3) are satisfied by functions g which have the following asymptotic behavior as $r \to \infty$:

$$g(r) \sim -r^{-\alpha} \quad \text{where } 1 < \alpha < 2 \quad \text{when } n = 1,$$

$$g(r) \sim -r^{-2}(\ln r)^{-\alpha} \quad \text{where } 1 < \alpha < 2 \quad \text{when } n = 2.$$

We shall prove results only for the case $n = 2$; identical results with somewhat simpler proofs hold when $n = 1$. Our main result (Theorem 6.4) shows that bifurcation occurs from the branch of trivial solutions $u \equiv 1$ when $\lambda = 0$.

It is shown in [2] that, if g satisfies (G_3), then $(5.1)_\lambda$ has subsolution with compact support for all $\lambda > 0$ and that, if g satisfies (G_2), then $(5.1)_\lambda$ has a supersolution which is identically equal to 1 on a big interval of the form $[0, R]$ for all $\lambda > 0$. Thus we have the following result.

LEMMA 6.1. *Suppose g satisfies (G_2) and (G_3). Then equation $(5.1)_\lambda$ has a unique solution u_λ for all $\lambda > 0$.*

We now investigate the behavior of u_λ as $\lambda \to 0$. The following identity enables us to exclude the possiblity that $u_\lambda \to 0$.

LEMMA 6.2. *Suppose g satisfies (G_2) and (G_3). If u is a solution of $(5.1)_\lambda$ then*

$$\int_0^\infty r\left[\frac{u'}{f(u)}\right]^2 f'(u)\, dr + \lambda \int_0^\infty r g(r)\, dr = 0.$$

Proof. Since u is a solution of $(5.1)_\lambda$, we have

$$(ru')'/f(u) + \lambda r g(r) = 0 \quad \text{for } r > 0.$$

Hence

$$(6.1) \qquad \frac{ru'}{f(u)} + \int_0^r s\left(\frac{u'}{f(u)}\right)^2 f'(u)\, ds + \lambda \int_0^r s g(s)\, ds = 0.$$

By Corollary 3.9, $\lim_{r\to\infty} ru'/u = 0$, and so $\lim_{r\to\infty} ru'/f(u) = 0$. Since $\int_0^\infty rg(r)\, dr$ converges, the required identity follows by letting $r \to \infty$ in (6.1).

COROLLARY 6.3. *Suppose g satisfies (G_2) and (G_3). Let u be a solution of $(5.1)_\lambda$. Then $\max\{u(r): r > 0\} \geqq \sigma$ where $\sigma = \inf\{u \in (0, 1): f'(u) = 0\}$.*

Proof. Assume the contrary. Then $u(r) < \sigma$, and so $f'(u(r)) > 0$ for all $r > 0$. This is clearly impossible because of Lemma 6.2.

THEOREM 6.4. *Suppose g satisfies* (G_2) *and* (G_3). *Suppose* $\lambda_k \to 0$ *and* $u_k = u_{\lambda_k}$. *Then* $u_k \to 1$ *uniformly on compact subintervals of* $[0, \infty)$.

Proof. Because of the continuous dependence of $u(\cdot, p, \lambda)$ on p and λ it suffices to prove that $\lim_{k\to\infty} u_k(0) = 1$. Assume the contrary. Then there exist $l \in [0, 1)$ and subsequences which we again denote by λ_k and u_k such that $\lim_{k\to\infty} u_k(0) = l$.

Suppose $l = 0$. It follows from continuous dependence arguments that u_k converges uniformly to zero on $[0, R_0]$, and so, since u_k is decreasing on $[R_0, \infty)$, that u_k converges uniformly to zero on R. This is impossible because of Corollary 6.3.

Suppose $0 < l < 1$. Since

$$(ru_k')' = \lambda_k r g(r) f(u_k) \quad \text{for } r > 0, \qquad u_k'(0) = 0 \tag{6.2}$$

it follows that

$$u_k(r) = u_k(0) + \lambda_k \int_0^r s \ln\left(\frac{r}{s}\right) g(s) f(u_k(s))\, ds \quad \text{for } r > 0. \tag{6.3}$$

Since $\lim_{k\to\infty} \lambda_k \int_0^r s \ln(r/s) g(s) f(u_k(s))\, ds = 0$ for all r, $0 \leqq r \leqq R$, letting $k \to \infty$ in (6.3), we have that u_k converges uniformly to l on $[0, R]$. Hence

$$\lim_{k\to\infty} \int_0^\infty r g(r) f(u_k)\, dr = \int_0^\infty r g(r) f(l)\, dr > 0. \tag{6.4}$$

However, integrating (6.2) from zero to infinity and using Corollary 3.9, shows that $\int_0^\infty r g(r) f(u_k)\, dr = 0$ for all k, and this is a contradiction.

Hence $\lim_{k\to\infty} u_k(0) = 1$ and the proof is complete.

REFERENCES

[1] H. BERESTYCKI AND P. L. LIONS, *Some Applications of the Method of Super and Subsolutions in Bifurcation and Nonlinear Eigenvalue Problems*, Springer-Verlag, Berlin, New York, 1980, pp. 16–41.

[2] K. J. BROWN, S. S. LIN, AND A. TERTIKAS, *Existence and nonexistence of steady state solutions for a selection migration model in population genetics*, J. Math. Biol., 27 (1989), pp. 91–104.

[3] W. H. FLEMING, *A selection-migration model in population genetics*, J. Math. Biol., 2 (1975), pp. 219–233.

[4] B. GIDAS, W. M. NI, AND L. NIRENBERG, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 210–243.

[5] P. HESS AND T. KATO, *On some linear and nonlinear eigenvalue problems with an indefinite weight function*, Comm. Partial Differential Equations, 5 (1980), pp. 999–1030.

[6] W. M. NI, *On the elliptic equation* $\Delta u + k(x)u^{(n+2)/n-2} = 0$, *its generalizations and applications in geometry*, Indiana Univ. Math. J., 31 (1982), pp. 493–529.

[7] A. TERTIKAS, *Existence and uniqueness of solutions for a nonlinear diffusion problem arising in population genetics*, Arch. Rational Mech. Anal., 103 (1988), pp. 289–319.

[8] ———, *Semilinear elliptic equations on* R^n, Ph.D. thesis, Heriot-Watt University, Edinburgh, Scotland, 1987.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 414-423, March 1991

008

ON THE REPRESENTATION OF STOKES FLOWS*

WERNER KRATZ†

Abstract. In this paper representations of Stokes flows in dimensions 2 and 3, which reduce the Stokes equations to the Laplace equation for an auxiliary function, are given. While it is known that two- and three-dimensional Stokes flows may be reduced to biharmonic problems, the representations here are new. The main result of this paper reads as follows: Given a domain $G \subseteq \mathbb{R}^3$, which is star-shaped with respect to the origin, and functions $\vec{v}: G \to \mathbb{R}^3$, $p: G \to \mathbb{R}$, then $\vec{v}$ and p represent a Stokes flow with velocity field $\vec{v}$ and pressure p in G (i.e., $\Delta \vec{v} = \operatorname{grad} p$ and $\operatorname{div} \vec{v} = 0$ in G) if and only if $\vec{v}$ and p are of the form

$$\vec{v}(\vec{x}) = \tilde{\vec{v}}(\vec{x}) - \tfrac{1}{3}\{\operatorname{div} \tilde{\vec{v}}(\vec{x}) \cdot \vec{x} + \vec{x} \times \operatorname{curl} \tilde{\vec{v}}(\vec{x})\}, \qquad p(\vec{x}) = -\tfrac{4}{3} \operatorname{div} \tilde{\vec{v}}(\vec{x}),$$

where $\tilde{\vec{v}}$ is a harmonic function (i.e., $\Delta \tilde{\vec{v}} = 0$) in G.

Key words. Stokes flows, Stokes equations, harmonic functions

AMS(MOS) subject classifications. 76D07, 35Q10, 31A10

1. Introduction. In this paper we derive formulae for the velocity field $\vec{v}$ and the pressure p of two- and three-dimensional *Stokes flows*, which represent $\vec{v}$ and p via a certain (uniquely determined) auxiliary *harmonic function* $\tilde{\vec{v}}$. For dimension 2 our Theorem 1 extends a result of [3, Thm. 1 and Lemma 2], where $\vec{v}$ and p of a Stokes flow are expressed by two holomorphic functions provided the corresponding domain $G \subseteq \mathbb{R}^2$ is simply connected. Our two-dimensional representation formula here is valid for any domain $G \subseteq \mathbb{R}^2$, and, moreover, we can express the auxiliary function $\tilde{\vec{v}}$ explicitly in terms of $\vec{v}$ and p. Besides the result in [3] there are well-known reductions of the Stokes problem in dimensions 2 and 3 to biharmonic problems (see, e.g., [1], [9], and [6]).

The *main result* of this paper is a representation theorem (Theorem 2 below) for *three-dimensional Stokes flows*. This theorem is quite similar to the two-dimensional Theorem 1, and its essential part reads as follows. Given a domain $G \subseteq \mathbb{R}^3$, which is *star-shaped* with respect to the origin, and given functions $\vec{v}: G \to \mathbb{R}^3$, $p: G \to \mathbb{R}$, then $\vec{v}$ and p represent a Stokes flow in G, i.e.,

$$\Delta \vec{v} = \operatorname{grad} p, \qquad \operatorname{div} \vec{v} = 0 \quad \text{in } G,$$

if and only if there exists a harmonic function $\tilde{\vec{v}}$ (i.e., $\Delta \tilde{\vec{v}} = 0$) in G such that

$$\vec{v}(\vec{x}) = \tilde{\vec{v}}(\vec{x}) - \tfrac{1}{3}\{\operatorname{div} \tilde{\vec{v}}(\vec{x}) \cdot \vec{x} + \vec{x} \times \operatorname{curl} \tilde{\vec{v}}(\vec{x})\},$$
$$p(\vec{x}) = -\tfrac{4}{3} \operatorname{div} \tilde{\vec{v}}(\vec{x}) \quad \text{for } \vec{x} \in G.$$

Moreover, $\tilde{\vec{v}}$ is uniquely determined by $\vec{v}$ and p, and we give an explicit formula expressing $\tilde{\vec{v}}$ in terms of $\vec{v}$ and p (see Theorem 2 below). In the case of so-called *potential flows* (i.e., $\vec{v} = \operatorname{grad} \phi$ with $\Delta \phi = 0$) our representation formula becomes trivial, i.e., $\vec{v} = \tilde{\vec{v}}$ (since $\operatorname{div} \vec{v} = 0$, $\operatorname{curl} \vec{v} = 0$ in that case). Moreover, it may be interesting to ask whether the auxiliary harmonic function $\tilde{\vec{v}}$ has any "concrete" physical interpretation with respect to the Stokes flow.

Let us shortly summarize the setup of this paper. In § 2 we state our results on two- and three-dimensional Stokes equations (Theorems 1 and 2). Some auxiliary

* Received by the editors October 25, 1989; accepted for publication (in revised form) April 2, 1990.
† Abteilung Mathematik, Universität Ulm, Albert-Einstein-Allee 11, D-7900 Ulm, Federal Republic of Germany.

formulae from the vector analysis are given in § 3. Section 4 is devoted to the proof of the two-dimensional result. The main result, i.e., the three-dimensional Theorem 2, is shown in § 5, where we also discuss the assumption on the domain (i.e., *star-shaped* with respect to the origin) in detail. But it remains an open question whether Theorem 2 can be extended to more general domains (e.g., simply connected domains).

Finally, we want to point out that these representation formulae can serve as the basis of numerical algorithms (see [3] for such an algorithm in dimension two via conformal mappings) to solve Stokes boundary value problems. In view of our results these problems reduce to solving Laplace's equation $\Delta\tilde{\vec{v}} = 0$ with "suitably adapted" boundary conditions.

2. Main results. To formulate our results we need some *notation* and *notions*. We will use the usual *inner product* in $\mathbb{R}^2$ or $\mathbb{R}^3$ and *vector product* in $\mathbb{R}^3$. The differential operators grad, div, and Δ denote the *gradient*, *divergence*, and *Laplacian* of scalar functions, respectively, vector fields in $\mathbb{R}^2$ or $\mathbb{R}^3$. While curl is defined as usual for vector fields in $\mathbb{R}^3$, we introduce the following additional notation in $\mathbb{R}^2$:

$$\operatorname{curl}\vec{v} := \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \quad \text{for a vector field } \vec{v} = \begin{pmatrix} u \\ v \end{pmatrix} \quad \text{in } \mathbb{R}^2;$$

$$\vec{x}^{\perp}\alpha := \begin{pmatrix} \alpha y \\ -\alpha x \end{pmatrix}, \quad \text{and} \quad \vec{x}^{\perp} := \begin{pmatrix} y \\ -x \end{pmatrix} \quad \text{for } \vec{x} = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2, \quad \alpha \in \mathbb{R}.$$

Moreover, we need the following definition.

DEFINITION 1. Let $G \subseteq \mathbb{R}^n$ be a domain. A function $\vec{v}: G \to \mathbb{R}^n$ is called

(i) *harmonic* in G, if $\vec{v} \in C_2(G)$ and if $\Delta\vec{v}(\vec{x}) = 0$ for $\vec{x} \in G$; and it is called a

(ii) *Stokes function* in G, if $\vec{v} \in C_2(G)$ and if there exists a (pressure-) function $p: G \to \mathbb{R}$ with $p \in C_1(G)$ and

$$\Delta\vec{v}(\vec{x}) = \operatorname{grad} p(\vec{x}), \qquad \operatorname{div}\vec{v}(\vec{x}) = 0 \quad \text{for } \vec{x} \in G. \tag{1}$$

Remark 1. Obviously, for a given Stokes function the corresponding pressure p is unique up to an additive constant. Next, every function that is harmonic in a domain G belongs to $C_\infty(G)$, even every continuous solution q of $\Delta q = h$ if $h \in C_\infty(G)$. Moreover, by a result of Weyl [10], every Stokes function $\vec{v}$ in a domain G is in $C_\infty(G)$, and a corresponding pressure-function p is harmonic in G, since $\Delta p = \operatorname{div}\{\operatorname{grad} p\} = \operatorname{div}\{\Delta\vec{v}\} = 0$ by (1). For this reason, we will not discuss differentiability conditions from now on.

THEOREM 1. *Let $G \subseteq \mathbb{R}^2$ be a domain. A function $\vec{v}: G \to \mathbb{R}^2$ is a Stokes function with corresponding pressure p in G, if and only if there exists a harmonic function $\tilde{\vec{v}}$ in G such that*

$$\begin{aligned} \vec{v}(\vec{x}) &= \tilde{\vec{v}}(\vec{x}) - \tfrac{1}{2}\{\operatorname{div}\tilde{\vec{v}}(\vec{x})\cdot\vec{x} + \vec{x}^{\perp}\operatorname{curl}\tilde{\vec{v}}(\vec{x})\}, \\ p(\vec{x}) &= -2\operatorname{div}\tilde{\vec{v}}(\vec{x}) \quad \textit{for } \vec{x} \in G; \end{aligned} \tag{2}$$

this harmonic function $\tilde{\vec{v}}$ is unique, and we have

$$\tilde{\vec{v}}(\vec{x}) = \vec{v}(\vec{x}) - \tfrac{1}{4}\{p(\vec{x})\cdot\vec{x} - \vec{x}^{\perp}\operatorname{curl}\vec{v}(\vec{x})\} \quad \text{for } \vec{x} \in G. \tag{3}$$

This theorem is a quite statisfactory completion of [3, Thm. 1], where G was assumed to be simply connected. But our main objective here is the following representation theorem in the most interesting dimension 3, where no such result seems to exist. The assumption on the domain G is not as satisfactory as in dimension 2 above. We have to assume that G is *star-shaped with respect to the origin*, i.e., $t\vec{x} \in G$ for all $t \in [0, 1]$ whenever $\vec{x} \in G$.

THEOREM 2. *Assume that $G \subseteq \mathbb{R}^3$ is a star-shaped domain with respect to the origin. Then a function $\vec{v}: G \to \mathbb{R}^3$ is a Stokes function with corresponding pressure p in G, if and only if there exists a harmonic function $\tilde{\vec{v}}$ in G such that*

$$\begin{aligned} \vec{v}(\vec{x}) &= \tilde{\vec{v}}(\vec{x}) - \tfrac{1}{3}\{\operatorname{div} \tilde{\vec{v}}(\vec{x}) \cdot \vec{x} + \vec{x} \times \operatorname{curl} \tilde{\vec{v}}(\vec{x})\}, \\ p(\vec{x}) &= -\tfrac{4}{3} \operatorname{div} \tilde{\vec{v}}(\vec{x}) \quad \text{for } \vec{x} \in G; \end{aligned} \tag{4}$$

this harmonic function $\tilde{\vec{v}}$ is unique, and we have

$$\tilde{\vec{v}}(\vec{x}) = \vec{v}(\vec{x}) - \tfrac{1}{4}\{p(\vec{x}) \cdot \vec{x} - \vec{x} \times \operatorname{curl} \vec{v}(\vec{x})\} + \vec{x} \times \operatorname{grad} \phi(\vec{x}) \quad \text{for } \vec{x} \in G, \tag{5}$$

where the function ϕ is harmonic in G and given by

$$\phi(\vec{x}) = -\frac{1}{4} \int_0^1 t^4 \vec{x} \cdot \operatorname{curl} \vec{v}(t\vec{x})\, dt \quad \text{for } \vec{x} \in G. \tag{6}$$

We will prove this theorem in § 5, in particular the assumption on G (i.e., star-shaped with respect to the origin) will be discussed there in detail.

Remark 2. Of course, two-dimensional Stokes functions $\vec{v} = \binom{u}{v}$ are also Stokes functions in dimension 3 (with $u_z = v_z \equiv 0$) by setting the third component $w \equiv 0$. But in that case the three-dimensional representation formula (4) above does *not* reduce to the corresponding two-dimensional formula (2) (e.g., we have the factor $\frac{1}{3}$ in (4) instead of $\frac{1}{2}$ in (2)). Moreover, it is not clear at all, how the representation (2) can be derived from formula (4) (or vice versa) in the case of two-dimensional flows.

3. Auxiliary formulae from the vector analysis. For the proofs of Theorems 1 and 2 we need several formulae from the vector analysis, which are, at least partly, well known (see, e.g., [8, § 124, p. 384]). These formulae can be verified directly, and we assume, of course, the existence and/or continuity of the partial derivatives involved.

LEMMA 1. *Let $\vec{v}: G \to \mathbb{R}^n$, $\phi: G \to \mathbb{R}$, $G \subseteq \mathbb{R}^n$. Then*

$$\operatorname{curl}\{\phi(\vec{x}) \cdot \vec{x}\} = \operatorname{grad} \phi(\vec{x}) \cdot \vec{x}^{\perp} \quad \text{for } n = 2; \tag{7}$$

$$\operatorname{curl}\{\vec{x}^{\perp} \operatorname{curl} \vec{v}(\vec{x})\} = [\Delta \vec{v}(\vec{x}) - \operatorname{grad}\{\operatorname{div} \vec{v}(\vec{x})\}] \cdot \vec{x}^{\perp} - 2 \operatorname{curl} \vec{v}(\vec{x}) \quad \text{for } n = 2; \tag{8}$$

$$\begin{aligned} \operatorname{div}\{\phi(\vec{x}) \cdot \vec{x}\} &= 2\phi(\vec{x}) + \vec{x} \cdot \operatorname{grad} \phi(\vec{x}) \text{ for } n = 2, \textit{ respectively}, \\ \operatorname{div}\{\phi(\vec{x}) \cdot \vec{x}\} &= 3\phi(\vec{x}) + \vec{x} \cdot \operatorname{grad} \phi(\vec{x}) \text{ for } n = 3; \end{aligned} \tag{9}$$

$$\operatorname{curl}\{\phi(\vec{x}) \cdot \vec{x}\} = -\vec{x} \times \operatorname{grad} \phi(\vec{x}) \quad \text{for } n = 3; \tag{10}$$

$$\Delta\{\phi(\vec{x}) \cdot \vec{x}\} = 2 \operatorname{grad} \phi(\vec{x}) + \Delta\phi(\vec{x}) \cdot \vec{x} \quad \text{for } n = 2 \text{ or } n = 3; \tag{11}$$

$$\begin{aligned} \operatorname{curl}\{\operatorname{grad} \phi(\vec{x})\} &= 0 \quad \text{for } n = 2 \text{ or } n = 3, \\ \operatorname{div}\{\operatorname{curl} \vec{v}(\vec{x})\} &= 0, \qquad \operatorname{div}\{\vec{x} \times \operatorname{grad} \phi(\vec{x})\} = 0 \quad \text{for } n = 3; \end{aligned} \tag{12}$$

$$\begin{aligned} \operatorname{div}\{\vec{x}^{\perp} \operatorname{curl} \vec{v}(\vec{x})\} &= \vec{x} \cdot \Delta \vec{v}(\vec{x}) - \vec{x} \cdot \operatorname{grad}\{\operatorname{div} \vec{v}(\vec{x})\} \quad \text{for } n = 2, \textit{ respectively}, \\ \operatorname{div}\{\vec{x} \times \operatorname{curl} \vec{v}(\vec{x})\} &= \vec{x} \cdot \Delta \vec{v}(\vec{x}) - \vec{x} \cdot \operatorname{grad}\{\operatorname{div} \vec{v}(\vec{x})\} \quad \text{for } n = 3; \end{aligned} \tag{13}$$

$$\begin{aligned} \Delta\{\vec{x}^{\perp} \operatorname{curl} \vec{v}(\vec{x})\} &= \vec{x}^{\perp} \operatorname{curl}\{\Delta \vec{v}(\vec{x})\} - 2\Delta \vec{v}(\vec{x}) \\ &\quad + 2 \operatorname{grad}\{\operatorname{div} \vec{v}(\vec{x})\} \quad \text{for } n = 2, \textit{ respectively}, \\ \Delta\{\vec{x} \times \operatorname{curl} \vec{v}(\vec{x})\} &= \vec{x} \times \operatorname{curl}\{\Delta \vec{v}(\vec{x})\} - 2\Delta \vec{v}(\vec{x}) \\ &\quad + 2 \operatorname{grad}\{\operatorname{div} \vec{v}(\vec{x})\} \quad \text{for } n = 3; \end{aligned} \tag{14}$$

$$\text{curl}\,\{\text{curl}\,\vec{v}(\vec{x})\} = \text{grad}\,\{\text{div}\,\vec{v}(\vec{x})\} - \Delta\vec{v}(\vec{x}) \quad \textit{for } n = 3; \tag{15}$$

$$\Delta\{\vec{x}\cdot\text{curl}\,\vec{v}(\vec{x})\} = \vec{x}\cdot\text{curl}\,\{\Delta\vec{v}(\vec{x})\} \quad \textit{for } n = 3; \tag{16}$$

$$\begin{aligned}\text{curl}\,\{\vec{x}\times\text{curl}\,\vec{v}(\vec{x})\} &= \vec{x}\times\text{grad}\,\{\text{div}\,\vec{v}(\vec{x})\} - \text{grad}\,\{\vec{x}\cdot\text{curl}\,\vec{v}(\vec{x})\}\\ &\quad - \text{curl}\,\vec{v}(\vec{x}) - \vec{x}\times\Delta\vec{v}(\vec{x}) \quad \textit{for } n = 3;\end{aligned} \tag{17}$$

$$\Delta\{\vec{x}\times\text{grad}\,\phi(\vec{x})\} = \vec{x}\times\text{grad}\,\{\Delta\phi(\vec{x})\} \quad \textit{for } n = 3; \tag{18}$$

$$\text{curl}\,\{\vec{x}\times\text{grad}\,\phi(\vec{x})\} = \Delta\phi(\vec{x})\cdot\vec{x} - \text{grad}\,\{\phi(\vec{x}) + \vec{x}\cdot\text{grad}\,\phi(\vec{x})\} \quad \textit{for } n = 3. \tag{19}$$

4. Derivation and proof of the two-dimensional theorem. Using essentially the formulae of Lemma 1 for dimension 2 we can now derive Theorem 1.

Proof of Theorem 1. (i) First, assume that $\tilde{\vec{v}}$ is harmonic in G and that $\vec{v}$ and p are given by (2). Then the representation (2) and the formulae (9), (11), (13), and (14) (with $\tilde{\vec{v}}$ instead of $\vec{v}$ and $\phi = \text{div}\,\tilde{\vec{v}}$) yield:

$$\begin{aligned}&\text{div}\,\vec{v} = \text{div}\,\tilde{\vec{v}} - \tfrac{1}{2}\{2\,\text{div}\,\tilde{\vec{v}} + \vec{x}\cdot\text{grad}\,(\text{div}\,\tilde{\vec{v}}) - \vec{x}\cdot\text{grad}\,(\text{div}\,\tilde{\vec{v}})\} \equiv 0,\\ &\Delta\vec{v} = -\tfrac{1}{2}\{2\,\text{grad}\,(\text{div}\,\tilde{\vec{v}}) + 2\,\text{grad}\,(\text{div}\,\tilde{\vec{v}})\} = \text{grad}\,p \text{ in } G.\end{aligned}$$

Hence, $\vec{v}$ is a Stokes function with corresponding pressure $p = -2\,\text{div}\,\tilde{\vec{v}}$ in G according to Definition 1.

(ii) Now, assume that $\vec{v}$ is a Stokes function with corresponding pressure p and that $\tilde{\vec{v}}$ is given by (3). Then the formulae (1), (7)-(9), and (11)-(14) (use also that $\Delta p \equiv 0$ by Remark 1) show that

$$\begin{aligned}\Delta\tilde{\vec{v}} &= \Delta\vec{v} - \tfrac{1}{4}\{2\,\text{grad}\,p + \Delta p\cdot\vec{x} - \vec{x}^{\perp}\,\text{curl}\,(\Delta\vec{v})\\ &\quad + 2\Delta\vec{v} - 2\,\text{grad}\,(\text{div}\,\vec{v})\} \equiv 0,\\ \text{div}\,\tilde{\vec{v}} &= \text{div}\,\vec{v} - \tfrac{1}{4}\{2p + \vec{x}\cdot\text{grad}\,p - \vec{x}\cdot\Delta\vec{v} + \vec{x}\cdot\text{grad}\,(\text{div}\,\vec{v})\} = -\tfrac{1}{2}p,\\ \vec{x}^{\perp}\,\text{curl}\,\tilde{\vec{v}} &= \vec{x}^{\perp}\{\text{curl}\,\vec{v} - \tfrac{1}{4}[(\text{grad}\,p - \Delta\vec{v} + \text{grad}\,(\text{div}\,\vec{v}))\cdot\vec{x}^{\perp} + 2\,\text{curl}\,\vec{v}]\}\\ &= \tfrac{1}{2}\vec{x}^{\perp}\,\text{curl}\,\vec{v},\\ \vec{v} &= \tilde{\vec{v}} + \tfrac{1}{4}\{-2\,\text{div}\,\tilde{\vec{v}}\cdot\vec{x} - 2\vec{x}^{\perp}\,\text{curl}\,\tilde{\vec{v}}\},\end{aligned}$$

which is (3). This completes the proof of Theorem 1 up to the uniqueness of $\tilde{\vec{v}}$, which is a consequence of the following proposition. □

PROPOSITION 1. *Assume that $\vec{v}_1$ and $\vec{v}_2$ are harmonic in G such that*

$$\begin{aligned}&\vec{v}_1(\vec{x}) - \tfrac{1}{2}\{\text{div}\,\vec{v}_1(\vec{x})\cdot\vec{x} + \vec{x}^{\perp}\,\text{curl}\,\vec{v}_1(\vec{x})\}\\ &\quad = \vec{v}_2(\vec{x}) - \tfrac{1}{2}\{\text{div}\,\vec{v}_2(\vec{x})\cdot\vec{x} + \vec{x}^{\perp}\,\text{curl}\,\vec{v}_2(\vec{x})\} \quad \textit{for } \vec{x}\in G.\end{aligned}$$

Then there exists $\alpha\in\mathbb{R}$ with

$$\vec{v}_1(\vec{x}) - \vec{v}_2(\vec{x}) = \alpha\vec{x} \quad \textit{for } \vec{x}\in G.$$

Proof. Consider $\vec{v} = \vec{v}_1 - \vec{v}_2$, so that $\vec{v} = \tfrac{1}{2}\{\text{div}\,\vec{v}\cdot\vec{x} + \vec{x}^{\perp}\,\text{curl}\,\vec{v}\}$ and $\Delta\vec{v}\equiv 0$. Then the formulae (2) and (13) yield that $\text{curl}\,\vec{v} = -\text{curl}\,\vec{v}$, i.e., $\text{curl}\,\vec{v}\equiv 0$, and $\text{grad}\,(\text{div}\,\vec{v})\equiv 0$ (use $\Delta\vec{v}\equiv 0$ again). Hence, $\text{curl}\,\vec{v}\equiv 0$ and $\text{div}\,\vec{v}\equiv 2\alpha\in\mathbb{R}$ in G, and it follows that $\vec{v} = \tfrac{1}{2}\{\text{div}\,\vec{v}\cdot\vec{x} + \vec{x}^{\perp}\,\text{curl}\,\vec{v}\} = \alpha\vec{x}$ for $\vec{x}\in G$. □

Remark 3. Since $\text{div}\,(\alpha\vec{x}) = 2\alpha$ it follows that $\tilde{\vec{v}}$ is uniquely determined by (2) when $\vec{v}$ and p are given; and this constant α is just the free additive constant in the pressure p (compare Remark 1) for a given Stokes function $\vec{v}$.

5. Derivation and proof of the three-dimensional theorem. The main Theorem 2 of this paper follows from Propositions 2-4, and Lemma 2 below.

PROPOSITION 2. *Let $G \subseteq \mathbb{R}^3$ be a domain, and assume that $\vec{v}$ and p are given by* (4), *where $\tilde{\vec{v}}$ is harmonic in G. Then $\vec{v}$ is a Stokes function with corresponding pressure p in G.*

Proof. The representation (4) and the formulae (9), (11), (13), and (14) (with $\tilde{\vec{v}}$ instead of $\vec{v}$ and $\phi = \operatorname{div} \tilde{\vec{v}}$) yield:

$$\operatorname{div} \vec{v} = \operatorname{div} \tilde{\vec{v}} - \tfrac{1}{3}\{3 \operatorname{div} \tilde{\vec{v}} + \vec{x} \cdot \operatorname{grad} (\operatorname{div} \tilde{\vec{v}}) - \vec{x} \cdot \operatorname{grad} (\operatorname{div} \tilde{\vec{v}})\} \equiv 0,$$

$$\Delta \vec{v} = -\tfrac{1}{3}\{2 \operatorname{grad} (\operatorname{div} \tilde{\vec{v}}) + 2 \operatorname{grad} (\operatorname{div} \tilde{\vec{v}})\} = \operatorname{grad} p. \qquad \square$$

Remark 4. For *fixed* $\vec{x}_0 \in \mathbb{R}^3$ we have obviously that

$$\vec{v}(\vec{x}) := \tilde{\vec{v}}(\vec{x}) - \tfrac{1}{3}\{\operatorname{div} \tilde{\vec{v}}(\vec{x}) \cdot (\vec{x} - \vec{x}_0) + (\vec{x} - \vec{x}_0) \times \operatorname{curl} \tilde{\vec{v}}(\vec{x})\}$$

is also a Stokes function in G with corresponding pressure

$$p(\vec{x}) := -\tfrac{4}{3} \operatorname{div} \tilde{\vec{v}}(\vec{x}), \quad \text{whenever } \tilde{\vec{v}} \text{ is harmonic in } G.$$

The next proposition deals with the question whether a given Stokes function *can be represented by* (4) with a harmonic $\tilde{\vec{v}}$.

PROPOSITION 3. *Assume that $\vec{v}$ is a Stokes function with corresponding pressure p in a domain $G \subseteq \mathbb{R}^3$. Then the representation formula* (4) *holds with a function $\tilde{\vec{v}}$, which is harmonic in G, if and only if this function $\tilde{\vec{v}}$ satisfies* (5), *where ϕ is harmonic in G such that*

$$4\phi(\vec{x}) + \vec{x} \cdot \operatorname{grad} \phi(\vec{x}) + \tfrac{1}{4}\vec{x} \cdot \operatorname{curl} \vec{v}(\vec{x}) = 0 \quad \text{for } \vec{x} \in G. \tag{20}$$

SUPPLEMENT. *Assume that formula* (4) *holds with a function $\tilde{\vec{v}}$, which is harmonic in G. Then, it follows that formula* (5) *holds for the function*

$$\phi(\vec{x}) := -\tfrac{1}{12}\vec{x} \cdot \operatorname{curl} \tilde{\vec{v}}(\vec{x}), \tag{21}$$

and this ϕ is harmonic in G and satisfies (20).

Proof. First, it follows from Definition (1) and (12) that

$$\operatorname{curl} \{\Delta \vec{v}(\vec{x})\} = 0 \quad \text{for } \vec{x} \in G. \tag{22}$$

(i) Assume that (5) holds with a function ϕ, which is harmonic and which satisfies (20). Then the formulae (5), (11), (14), (18), (22) and (1) (observe that $\Delta\phi = 0$ and that $\Delta p = \operatorname{div} (\operatorname{grad} p) = 0$ by (1)) yield

$$\Delta \tilde{\vec{v}} = \Delta \vec{v} - \tfrac{1}{4}\{2 \operatorname{grad} p + 2 \operatorname{grad} p\} = 0.$$

Hence, $\tilde{\vec{v}}$ is harmonic in G. Moreover, by (1), (9), (13), (12), and (10), (17), (19), (20) we obtain that

$$\begin{aligned}
\operatorname{div} \tilde{\vec{v}} &= -\tfrac{3}{4} p, \\
\operatorname{curl} \tilde{\vec{v}} &= \operatorname{curl} \vec{v} - \tfrac{1}{4}\{-\vec{x} \times \operatorname{grad} p + \operatorname{grad} \{\vec{x} \cdot \operatorname{curl} \vec{v}\} \\
&\qquad + \operatorname{curl} \vec{v} + \vec{x} \times \operatorname{grad} p\} - \operatorname{grad} \{\phi + \vec{x} \cdot \operatorname{grad} \phi\} \\
&= \tfrac{3}{4} \operatorname{curl} \vec{v} + 3 \operatorname{grad} \phi.
\end{aligned}$$

By putting these identities into (5), it follows that

$$\begin{aligned}
&\tilde{\vec{v}} - \tfrac{1}{3}\{\operatorname{div} \tilde{\vec{v}} \cdot \vec{x} + \vec{x} \times \operatorname{curl} \tilde{\vec{v}}\} \\
&\quad = \vec{v} - \tfrac{1}{4}\{p\vec{x} - \vec{x} \times \operatorname{curl} \vec{v}\} + \vec{x} \times \operatorname{grad} \phi \\
&\qquad - \tfrac{1}{3}\{-\tfrac{3}{4} p\vec{x} + \tfrac{3}{4}\vec{x} \times \operatorname{curl} \vec{v} + 3\vec{x} \times \operatorname{grad} \phi\} \\
&\quad = \vec{v},
\end{aligned}$$

and

$$p = -\tfrac{4}{3} \operatorname{div} \tilde{\vec{v}};\ \text{i.e., (4) holds.}$$

(ii) Now, assume that (4) holds with a harmonic function $\tilde{\vec{v}}$, and define $\phi(\vec{x}) := -\frac{1}{12}\vec{x} \cdot \operatorname{curl} \tilde{\vec{v}}(\vec{x})$ for $\vec{x} \in G$ according to (21). Then ϕ is harmonic in G by (16). By (4), (10), (17), and (21) (observe that $\Delta\tilde{\vec{v}} = 0$) we get $\operatorname{div} \tilde{\vec{v}} = -\frac{3}{4}p$, and

$$\begin{aligned}\operatorname{curl} \vec{v} = \operatorname{curl} \tilde{\vec{v}} - \tfrac{1}{3}\{&-\vec{x} \times \operatorname{grad}(\operatorname{div} \tilde{\vec{v}}) + \vec{x} \times \operatorname{grad}(\operatorname{div} \tilde{\vec{v}}) \\ &- \operatorname{curl} \tilde{\vec{v}} + 12 \operatorname{grad} \phi\},\end{aligned}$$

thus

$$\operatorname{curl} \tilde{\vec{v}} = \tfrac{3}{4} \operatorname{curl} \vec{v} + 3 \operatorname{grad} \phi. \tag{23}$$

Putting these formulae for $\operatorname{div} \tilde{\vec{v}}$ and $\operatorname{curl} \tilde{\vec{v}}$ into (4) yields

$$\begin{aligned}\tilde{\vec{v}} &= \vec{v} + \tfrac{1}{3}\{-\tfrac{3}{4}p\vec{x} + \tfrac{3}{4}\vec{x} \times \operatorname{curl} \vec{v} + 3\vec{x} \times \operatorname{grad} \phi\} \\ &= \vec{v} - \tfrac{1}{4}\{p\vec{x} - \vec{x} \times \operatorname{curl} \vec{v}\} + \vec{x} \times \operatorname{grad} \phi;\end{aligned}$$

i.e., (5) holds. Finally, the definition of ϕ by (21) and formula (23) imply that

$$\begin{aligned}4\phi + \vec{x} \cdot \operatorname{grad} \phi &= -\tfrac{1}{3}\vec{x} \cdot \operatorname{curl} \tilde{\vec{v}} + \tfrac{1}{3}\vec{x} \cdot \{\operatorname{curl} \tilde{\vec{v}} - \tfrac{3}{4} \operatorname{curl} \vec{v}\} \\ &= -\tfrac{1}{4}\vec{x} \cdot \operatorname{curl} \vec{v};\end{aligned}$$

i.e., (20) holds. This completes the proof of Proposition 3 and of the supplement. □

The next lemma shows that any Stokes function $\vec{v}$ can be represented by (4), where the auxiliary harmonic function $\tilde{\vec{v}}$ is given by (5) and (6). But for this lemma it is crucial to assume that the domain G is *star-shaped with respect to the origin* (and this assumption of Theorem 2 is needed only for this lemma).

LEMMA 2. *Assume that $G \subseteq \mathbb{R}^3$ is a domain which is star-shaped with respect to the origin, and that a function $\vec{v}: G \to \mathbb{R}^3$ satisfies $\vec{v} \in C_3(G)$ and* (22), *i.e.*, $\operatorname{curl}\{\Delta\vec{v}(\vec{x})\} \equiv 0$. *Then the function $\phi(\vec{x})$ given by* (6) *is harmonic in G and satisfies the differential equation* (20).

Proof. Consider the function $h(\vec{x}) := \vec{x} \cdot \operatorname{curl} \vec{v}(\vec{x})$. Then h is harmonic in G by (16) and (22). Just because G is star-shaped with respect to the origin the *function ϕ is well defined* by (6), and it follows that

$$\Delta\phi(\vec{x}) = \Delta\left\{-\frac{1}{4}\int_0^1 t^3 h(t\vec{x})\, dt\right\} = -\frac{1}{4}\int_0^1 t^5 (\Delta h)(t\vec{x})\, dt = 0.$$

Hence, ϕ is harmonic in G; and partial integration yields

$$\begin{aligned}\vec{x} \cdot \operatorname{grad} \phi(\vec{x}) &= -\frac{1}{4}\vec{x} \cdot \int_0^1 t^4 (\operatorname{grad} h)(t\vec{x})\, dt = -\frac{1}{4}\int_0^1 t^4 \frac{d}{dt}\{h(t\vec{x})\}\, dt \\ &= -\frac{1}{4}h(\vec{x}) + \int_0^1 t^3 h(t\vec{x})\, dt \\ &= -\frac{1}{4}\vec{x} \cdot \operatorname{curl} \vec{v}(\vec{x}) - 4\phi(\vec{x}) \quad \text{for } \vec{x} \in G.\end{aligned}$$

Thus, (20) holds. □

The *uniqueness* of $\tilde{\vec{v}}$ (compare Proposition 1 and Remark 3) in Theorem 2 is a consequence of the following proposition.

PROPOSITION 4. *Let $G \subseteq \mathbb{R}^3$ be a domain with $0 \in G$, and assume that $\vec{v}_1$ and $\vec{v}_2$ are harmonic functions in G such that*

$$\begin{aligned} &\vec{v}_1(\vec{x}) - \tfrac{1}{3}\{\operatorname{div} \vec{v}_1(\vec{x}) \cdot \vec{x} + \vec{x} \times \operatorname{curl} \vec{v}_1(\vec{x})\} \\ &\quad = \vec{v}_2(\vec{x}) - \tfrac{1}{3}\{\operatorname{div} \vec{v}_2(\vec{x}) \cdot \vec{x} + \vec{x} \times \operatorname{curl} \vec{v}_2(\vec{x})\} \quad \text{for } \vec{x} \in G. \end{aligned} \tag{24}$$

Then there exists $\alpha \in \mathbb{R}$ with

$$\vec{v}_1(\vec{x}) - \vec{v}_2(\vec{x}) = \alpha \vec{x} \quad \text{for } \vec{x} \in G. \tag{25}$$

Proof. Proposition 2 and equation (24) yield $0 = \operatorname{grad}\{\operatorname{div} \vec{v}_1 - \operatorname{div} \vec{v}_2\}$, thus

$$\operatorname{div}(\vec{v}_1 - \vec{v}_2) \equiv 3\alpha \quad \text{in } G \quad \text{for some } \alpha \in \mathbb{R}.$$

Then the function $\vec{v}_0(\vec{x}) := \vec{v}_1(\vec{x}) - \vec{v}_2(\vec{x}) - \alpha \vec{x}$ satisfies (observe that $\operatorname{div} \vec{x} = 3$, $\operatorname{curl} \vec{x} = 0$):

$$\operatorname{div} \vec{v}_0(\vec{x}) = 0, \quad \Delta \vec{v}_0(\vec{x}) = 0, \quad \vec{v}_0(\vec{x}) = \tfrac{1}{3}\vec{x} \times \operatorname{curl} \vec{v}_0(\vec{x}) \quad \text{for } \vec{x} \in G. \tag{26}$$

Setting $\psi(\vec{x}) := \vec{x} \cdot \operatorname{curl} \vec{v}_0(\vec{x})$, we obtain from (26), (16), and (17) the formulae

$$\Delta\psi(\vec{x}) = 0, \quad 4\psi(\vec{x}) + \vec{x} \cdot \operatorname{grad} \psi(\vec{x}) = 0, \quad \operatorname{grad} \psi(\vec{x}) = -4 \operatorname{curl} \vec{v}_0(\vec{x}). \tag{27}$$

Since $0 \in G$ we have that $U_\varepsilon(0) \subseteq G$ for some $\varepsilon > 0$. We consider the function $f(t) := \psi(t\vec{x})$ on $[0, 1]$ for a fixed $\vec{x} \in U_\varepsilon(0)$. Formulae (27) imply that $f(0) = 0$ and $4f(t) + tf'(t) = 0$ on $[0, 1]$. Hence, $f(t) \equiv 0$ on $[0, 1]$, in particular $\psi(\vec{x}) = f(1) = 0$; and it follows from (17) and (26) that $\operatorname{curl} \vec{v}_0(\vec{x}) = \frac{1}{3} \operatorname{curl}\{\vec{x} \times \operatorname{curl} \vec{v}_0(\vec{x})\} = -\frac{1}{3} \operatorname{curl} \vec{v}_0(\vec{x})$. Therefore, $\operatorname{curl} \vec{v}_0(\vec{x})$ and then also $\vec{v}_0(\vec{x})$ are identically zero in $U_\varepsilon(0)$. Since $\vec{v}_0$ is harmonic in G, it follows that $\vec{v}_0(\vec{x}) = 0$ for all $\vec{x} \in G$ by [2, Chap. X, Thm. IV], i.e., (25) holds. □

Remark 5. (i) Without the assumption $0 \in G$ the foregoing proof shows that

$$\vec{v}_1(\vec{x}) - \vec{v}_2(\vec{x}) = \alpha \vec{x} + \vec{v}_0(\vec{x}) \quad \text{for } \vec{x} \in G, \tag{25'}$$

where

$$\vec{v}_0(\vec{x}) = -\tfrac{1}{12}\vec{x} \times \operatorname{grad} \psi(\vec{x}) \quad \text{with } \psi(\vec{x}) = \vec{x} \cdot \operatorname{curl} \vec{v}_0(\vec{x}),$$

and where the function $\psi(\vec{x})$ satisfies

$$\Delta\psi(\vec{x}) = 0 \quad \text{and} \quad 4\psi(\vec{x}) + \vec{x} \cdot \operatorname{grad} \psi(\vec{x}) = 0 \quad \text{for } \vec{x} \in G; \tag{28}$$

i.e., $\psi(\vec{x})$ is harmonic in G with $\psi(\vec{x}) = r^{-4}\tilde{\psi}(\vartheta, \varphi)$, where (r, ϑ, φ) denote the usual spherical coordinates of $\vec{x}$. Any (separated) spherical harmonic of degree -4 satisfies (28). This discussion describes in general (i.e., if $0 \notin G$) the *exact freedom in the representation formula* (4). To prove this assume that $\psi \in C_2(G)$ satisfies (28); and define

$$\vec{v}_0(\vec{x}) := -\tfrac{1}{12}\vec{x} \times \operatorname{grad} \psi(\vec{x}).$$

Then we have (use (19))

$$\Delta \vec{v}_0(\vec{x}) = 0, \quad \operatorname{div} \vec{v}_0(\vec{x}) = 0, \quad \operatorname{curl} \vec{v}_0(\vec{x}) = -\tfrac{1}{4} \operatorname{grad} \psi(\vec{x});$$

thus,

$$\vec{v}_0(\vec{x}) - \tfrac{1}{3}\{\operatorname{div} \vec{v}_0(\vec{x}) \cdot \vec{x} + \vec{x} \times \operatorname{curl} \vec{v}_0(\vec{x})\} = 0$$

(compare formula (4) with $\vec{v} \equiv 0$, $p \equiv 0$, and $\tilde{\vec{v}} = \vec{v}_0$). In particular, the function $\tilde{\vec{v}}$ in (4) is *not* uniquely determined by $\vec{v}$ and p if $0 \notin G$.

(ii) The question whether Theorem 2 does not hold for more general domains is, by the preceding results, equivalent with the problem whether the differential equation (20) has a harmonic solution ϕ in G for any given function $\vec{v} \in C_3(G)$ that satisfies (22). In case that G is star-shaped with respect to the origin, Lemma 2 above answers this question affirmatively, and it even gives an explicit formula for ϕ (which is unique in this case by Proposition 4), i.e., (6). But the *question concerning more general domains remains open.* Concerning the solvability of (20) we have besides Lemma 2 only the following *local result* (where ε is certainly not the best possible).

PROPOSITION 5. *Given* $\vec{x}_0 \in \mathbb{R}^3$, $G := U_R(\vec{x}_0) = \{\vec{x}_0 \in \mathbb{R}^3 \,|\, |\vec{x} - \vec{x}_0| < R\}$ *with* $R > 0$, *and a function* $\vec{v} \in C_3(G)$ *that satisfies* (22). *Then there exists a function* ϕ, *which is harmonic and which satisfies the differential equation* (20) *in* $G' = U_\varepsilon(\vec{x}_0)$ *with* $\varepsilon := R/8e$.

Proof. Let $\gamma := |\vec{x}_0| \geqq R$, since otherwise $0 \in G$ and then the assertion follows from Lemma 2 (even with $\varepsilon = R$). Now there exists an orthogonal matrix U such that

$$U^T \vec{x}_0 = \begin{pmatrix} \gamma \\ 0 \\ 0 \end{pmatrix}.$$

For given functions $\vec{v}$ and ϕ consider the functions

$$h(\vec{x}) := \tfrac{1}{4}(U\vec{x} + \vec{x}_0) \operatorname{curl} \vec{v}(U\vec{x} + \vec{x}_0), \qquad \psi(\vec{x}) := \phi(U\vec{x} + \vec{x}_0). \tag{29}$$

Then $\Delta\psi(\vec{x}) = (\Delta\phi)(U\vec{x} + \vec{x}_0)$, since U is orthogonal, and $\operatorname{grad} \psi(\vec{x}) = (\operatorname{grad} \phi)(U\vec{x} + \vec{x}_0) \cdot U$ (where grad is a row-vector).

By (16) and (22) the function $\vec{x} \cdot \operatorname{curl} \vec{v}(\vec{x})$ is harmonic in G, thus

$$h(\vec{x}) \quad \text{is harmonic in} \quad U_R(0); \tag{30}$$

and our assertion is now equivalent with the existence of a function ψ satisfying

$$\begin{aligned} &\psi(\vec{x}) \text{ is harmonic in } U_\varepsilon(0), \quad \varepsilon = \frac{R}{8e} > 0, \quad \gamma \geqq R, \quad \text{with} \\ &4\psi(\vec{x}) + \vec{x} \cdot \operatorname{grad} \psi(\vec{x}) + \gamma\psi_z(\vec{x}) + h(\vec{x}) = 0 \quad \text{for } |\vec{x}| < \varepsilon. \end{aligned} \tag{31}$$

We construct $\psi(\vec{x}) = \psi(r, \vartheta, \varphi)$ by expansion with respect to spherical harmonics, where (r, ϑ, φ) denote the spherical coordinates of $\vec{x}$, i.e., $x = r \cos\varphi \sin\vartheta$, $y = r \sin\varphi \sin\vartheta$, $z = r \cos\vartheta$. By $Y_{nk}(\vartheta, \varphi)$, $-n \leqq k \leqq n$, we denote a complete, orthonormal system of spherical harmonics of degree n; more precisely:

$$\begin{aligned} Y_{nk}(\vartheta, \varphi) &:= c_{nk} \cos(k\varphi) P_n^k(\cos\vartheta) \qquad \text{for } 0 \leqq k \leqq n, \\ Y_{n,-k}(\vartheta, \varphi) &:= c_{nk} \sin(k\varphi) P_n^k(\cos\vartheta) \qquad \text{for } 1 \leqq k \leqq n, \end{aligned}$$

where

$$P_n^k(t) := \frac{1}{2^n \cdot n!} (1 - t^2)^{k/2} \left(\frac{d}{dt}\right)^{k+n} (t^2 - 1)^n$$

denote the Legendre functions (see [4, Chap. IV]) and where

$$c_{nk} := \left\{ \frac{2n+1}{2\pi} \frac{(n-k)!}{(n+k)!} \right\}^{1/2}.$$

Then we have

$$\int_{-\pi}^{\pi} \int_0^{\pi} Y_{nk}(\vartheta, \varphi) Y_{ml}(\vartheta, \varphi) \sin\vartheta \, d\vartheta \, d\varphi = \delta_{nm}\delta_{kl},$$

$$|Y_{nk}(\vartheta, \varphi)| \leqq \left\{ \frac{2n+1}{2\pi} \right\}^{1/2} \quad \text{for } (\vartheta, \varphi) \in [0, \pi] \times [-\pi, \pi] \tag{32}$$

by [5, Lemmas 8 and 14]. The recursion

$$ntP_n^k(t)+(1-t^2)\frac{d}{dt}P_n^k(t)=(n+k)P_{n-1}^k(t)$$

[4, p. 171] and $\partial r/\partial z=\cos\vartheta$, $\partial\varphi/\partial z=0$, $\partial\vartheta/\partial z=-\sin\vartheta/r$ yield the formula

$$\frac{\partial}{\partial z}\{r^nY_{nk}\}=\left\{\frac{2n+1}{2n-1}(n+k)(n-k)\right\}^{1/2}r^{n-1}Y_{n-1,k}\quad\text{for } |k|\leqq n-1, \tag{33}$$

$$=0\quad\text{for } k=\pm n,\quad n\geqq 0.$$

Since $h(\vec{x})$ is harmonic in $U_R(0)$ by (30) it can be expanded with respect to spherical harmonics (see [7] or [2]), i.e.,

$$h(\vec{x})=\sum_{n=0}^{\infty}\sum_{k=-n}^{n}b_{nk}r^nY_{nk}(\vartheta,\varphi)\quad\text{for } |\vec{x}|=r<R, \tag{34}$$

where the coefficients satisfy $b_{nk}=o(\rho^{-n})$ as $n\to\infty$ for all $\rho<R$. Hence, for *any fixed* $0<\rho<R$ we have

$$|b_{nk}|\leqq c\rho^{-n}\quad\text{for } |k|\leqq n,\ n\geqq 0\quad\text{with a constant } c=c(\rho)>0. \tag{35}$$

Now we construct ψ via the setting

$$\psi(\vec{x})=\sum_{n=0}^{\infty}\sum_{k=-n}^{n}a_{nk}r^nY_{nk}(\vartheta,\varphi).$$

Then ψ is automatically harmonic in $U_\varepsilon(0)$, provided we can show appropriate estimates of the coefficients. Now, by (33),

$$4\psi(\vec{x})+\vec{x}\cdot\operatorname{grad}\psi(\vec{x})+\gamma\psi_z(\vec{x})$$

$$=\sum_{n=0}^{\infty}\sum_{k=-n}^{n}\{(4+n)a_{nk}+\gamma\alpha_{nk}a_{n+1,k}\}r^nY_{nk}(\vartheta,\varphi)$$

with

$$\alpha_{nk}:=\left\{\frac{2n+3}{2n+1}(n+1+k)(n+1-k)\right\}^{1/2}.$$

Comparing coefficients, we find the differential equation (31) for ψ is equivalent with

$$(4+n)a_{nk}+\gamma\alpha_{nk}a_{n+1,k}+b_{nk}=0\quad\text{for } |k|\leqq n,\quad n\geqq 0. \tag{36}$$

These equations are fulfilled if we define the a_{nk} by

$$a_{n+1,k}=\frac{1}{\gamma\alpha_{nk}}\{-b_{nk}-(4+n)a_{nk}\}\quad\text{for } -n\leqq k\leqq n,\quad n\geqq 0,$$

$$a_{nn}=\frac{-1}{4+n}b_{nn},\quad a_{n,-n}=\frac{-1}{4+n}b_{n,-n}\quad\text{for } n\geqq 0. \tag{37}$$

We show recursively that

$$(4+n)|a_{n,\pm(n-k)}|\leqq(8/\rho)^nc\left\{\frac{n^k}{k!}\right\}^{1/2} \tag{38}$$

holds for $0 \leqq k \leqq n$, $n \geqq 0$. This is clear for $n = 0$, and for $k = 0$, $n \geqq 0$. For $1 \leqq k \leqq n+1$ it follows inductively from (35), (37), and (38) that

$$\begin{aligned}(4+n+1)|a_{n+1,n+1-k}| &\leqq \frac{4+n+1}{\gamma}\left\{\frac{2n+1}{2n+3}\cdot\frac{1}{k(2n+2-k)}\right\}^{1/2} \\ &\quad\cdot c\rho^{-n}\left\{1+8^n\left(\frac{n^{k-1}}{(k-1)!}\right)^{1/2}\right\} \\ &\leqq c\cdot\rho^{-n-1}\left\{\frac{n^{k-1}}{k!}\,\frac{(2n+1)(4+n+1)^2}{(2n+3)(n+1)}\right\}^{1/2} \\ &\quad\cdot\left\{8^n+\left(\frac{n^{k-1}}{(k-1)!}\right)^{-1/2}\right\} \\ &\leqq c(8/\rho)^{n+1}\left\{\frac{(n+1)^k}{k!}\right\}^{1/2} \qquad \text{(observe that } \gamma \geqq R > \rho\text{)}.\end{aligned}$$

Now, (38) yields

$$|a_{nk}| \leqq c(8/\rho)^n\left\{\frac{n^n}{n!}\right\}^{1/2}\frac{1}{4+n} \leqq c(8e/\rho)^n.$$

This estimate and (32) show that the series defining $\psi(\vec{x})$ converges absolutely (even after termwise differentiation by (33) and by corresponding formulae for the partial derivatives with respect to x and y) for $|\vec{x}| < R/8e$, and this completes the proof. □

Acknowledgment. The author records his thanks to the referees for several very helpful comments.

REFERENCES

[1] P. R. GARABEDIAN, *Free boundary flows of a viscous liquid*, Comm. Pure Appl. Math., 19 (1966), pp. 421-434.
[2] O. D. KELLOG, *Foundations of Potential Theory*, Dover, New York, 1953.
[3] W. KRATZ AND A. PEYERIMHOFF, *A numerical algorithm for the Stokes problem based on an integral equation for the pressure via conformal mappings*, Numer. Math., to appear (1990/91).
[4] W. MAGNUS, F. OBERHETTINGER, AND R. P. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, Berlin, 1966.
[5] C. MÜLLER, *Spherical Harmonics*, Lecture Notes 17, Springer-Verlag, Berlin, 1966.
[6] J. C. NEDELEC, *Eléments finis mixtes incompressibles pour l'équation de Stokes dans* $\mathbb{R}^3$, Numer. Math., 39 (1982), pp. 97-112.
[7] G. SANSONE, *Orthogonal Functions*, Wiley, New York, 1959.
[8] W. I. SMIRNOV, *Lehrgang der Höheren Mathematik* II, VEB Deutscher Verlag der Wissenschaften, Berlin, 1979.
[9] R. TEMAN, *Navier-Stokes Equations*, North-Holland, Berlin, 1984.
[10] H. WEYL, *The method of orthogonal projection in potential theory*, Duke Math. J., 7 (1940), pp. 411-444.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 424–439, March 1991

009

ISOLATED SINGULARITIES OF p-HARMONIC FUNCTIONS IN THE PLANE*

JUAN J. MANFREDI†

Dedicated to the memory of José Luis Rubio de Francia.

Abstract. A classification theorem for isolated singularities of p-harmonic functions is given and a precise asymptotic representation near the singularity is obtained. The approach is based on using the theory of quasiregular mappings to linearize the p-Laplacian by means of a hodographic transformation.

Key words. isolated singularities, p-harmonic functions, nonlinear elliptic equations

AMS(MOS) subject classifications. 35C20, 35J60

1. Introduction and statements of results.

Let Ω be a domain in $\mathbf{R}^n$, $n \geq 2$, and fix $p \in (1, \infty)$. A real function u defined in Ω is called p-harmonic if $u \in W^{1,p}_{\text{loc}}(\Omega)$ and

$$\int_{\Omega} |\nabla u|^{p-2} \langle \nabla u, \nabla \phi \rangle dx = 0 \tag{1.1}$$

for all $\phi \in W^{1,p}(\Omega)$ with compact support contained in Ω; that is, u is a weak solution of the p-Laplace equation

$$\operatorname{div}(|\nabla u|^{p-2} \nabla u) = 0. \tag{1.2}$$

It is well known [DB], [L] that $u \in C^{1,\alpha}_{\text{loc}}$, where $\alpha = \alpha(p, n) \in (0, 1]$. Indeed, in dimension two ($n = 2$) we have that

$$u \in C^{k,\alpha}_{\text{loc}}(\Omega) \cap W^{k+2,q}_{\text{loc}}(\Omega), \tag{1.3}$$

where the integer $k > 1$ and the exponent $\alpha \in (0, 1]$ are determined by the equation

$$k + \alpha = \frac{1}{6} \left(7 + \frac{1}{p-1} + \sqrt{1 + \frac{14}{p-1} + \frac{1}{(p-1)^2}} \right).$$

The integrability exponent q is any number such that

$$1 \leq q < \frac{2}{2 - \alpha} \leq 2.$$

Moreover, for $p \neq 2$ the regularity class in (1.3) is optimal. See [IM] for these results. To prove (1.3) we use complex analysis (quasiregular mappings) to linearize the p-Laplacian by means of a hodograph transformation [BI], [IM], [Ar].

*Received by the editors April 28, 1989; accepted for publication (in revised form) January 5, 1990. This research was partially supported by National Science Foundantion grant DMS-8901524.

†Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

The purpose of this paper is to apply the ideas of [IM] to the study of isolated singularities of p-harmonic functions in two dimensions.

Let us begin by reviewing some known results. We will assume that u is a *singular p-harmonic* function; that is, u is p-harmonic in $(\mathbf{B}^n)' = \mathbf{B}^n\backslash\{0\} = \{x \in \mathbf{R}^n, 0 < |x| < 1\}$. An example is given by

$$u_1(x) = \begin{cases} |x|^{(p-n)/(p-1)} & \text{if } p \neq n \\ \log \frac{1}{|x|} & \text{if } p = n. \end{cases} \tag{1.4}$$

The function u_1 satisfies

$$\operatorname{div}(|\nabla u_1|^{p-2}\nabla u_1) = c\delta_0 \tag{1.5}$$

in the sense of distributions, where $c = c(p, n)$ is an appropriate constant. A classical result of Serrin [S] states that if $1 < p \leq n$ and u is a *nonnegative* singular p-harmonic function, then there exists a constant $c_1 > 0$ such that

$$\frac{1}{c_1} < \frac{u}{u_1} < c_1 \qquad \text{in } (\mathbf{B}^n)'.$$

Moreover, u satisfies (1.5) for a certain c.

Kichenassamy and Veron [KV] extended Serrin's result to cover the case $u/u_1 \in L^\infty$, in which case they proved the existence of a number $\gamma \in \mathbf{R}$ such that

$$\begin{aligned} &\text{(i)} \quad u - \gamma u_1 \in L^\infty, \\ &\text{(ii)} \quad |\nabla(u - \gamma u_1)| = o\left(|x|^{(1-n)/(p-1)}\right), \qquad \text{as } x \to 0, \text{ and} \\ &\text{(iii)} \quad -\operatorname{div}(|\nabla u|^{p-2}\nabla u) = \gamma|\gamma|^{p-2}\delta_0. \end{aligned} \tag{1.6}$$

A general question is whether these results, conveniently modified, continue to hold when u_1 is replaced by an *anisotropic* singularity v; that is, v is p-harmonic in $\mathbf{R}^n\backslash\{0\}$ of the form:

$$v(x) = |x|^\beta w\left(\frac{x}{|x|}\right), \qquad \beta < 0. \tag{1.7}$$

Veron [V] has shown the existence of plenty of anisotropic singularities in n-dimensions, where this general problem is posed.

In the plane all solutions of the form (1.6), now called *quasiradial* are known. For $N = 1, 2, \cdots$ there exists a real analytic 2π-periodic function Φ_N with $\Phi_N(0) = 1$ (a normalization) and a number μ_N (μ_N decreasing to $-\infty$ as $N \uparrow \infty$) such that

$$u_N(se^{i\alpha}) = \Phi_N(\alpha)s^{\mu_N} \tag{1.8}$$

is a singular p-harmonic function. Assuming $p \neq 2$, we find $\Phi_1(\alpha) \equiv 1$ and $\mu_1 = (p-2)/(p-1)$ so that u_1 coincides with the "fundamental solution" in (1.4). While $\mu_1 > 0$ for $p > 2$ (case of bounded singularities)

$$\mu_2 = \frac{p - (3 + \sqrt{p^2 - 3p + 3})}{3(p-1)} < 0 \tag{1.9}$$

for all p. These facts follow easily from the results of §5. See also [KV].

We can think of u_N as a nonlinear analogue of $Re(z^{-N})$. Our main result, Theorem 1, is a nonlinear version of the Laurent series development for harmonic functions.

THEOREM 1. *Let u be a singular p-harmonic function in $(\mathbf{B}^2)'$ such that 0 is not a removable singularity and for some $\beta \le (p-2)/(p-1)$ we have*

$$|u(x)| \le c|x|^{\beta} \qquad \text{for } |x| < \tfrac{1}{2}. \tag{1.10}$$

Then there exists $N \ge 1$ (N is the largest integer such that $\beta \le \mu_N$) and a constant $\gamma_N \ne 0$ such that

$$\sum_{|\nu|=m} |D^{\nu}u - \gamma_N D^{\nu}u_N|(x) \le c_m |x|^{\mu_N - m + \delta_N} \tag{1.11}$$

for $m = 1, 2, \cdots$. Here c_m are constants, $\delta_N > 0$ and the inequality holds for $0 < |x| < \delta = \delta(m)$.

Remark. If u satisfies (1.10) for some $\beta > (p-2)/(p-1)$ we will see in §6 that 0 is a removable singularity. Thus, we are in the situation treated in detail in [IM].

Note that $|D^{\nu}u_N(x)| \sim |x|^{\mu_N - m}$. The theorem is relevant because $\delta_N > 0$. The estimate (1.11) is an extension of (1.6–ii).

In §2 we will compute the explicit values of μ_N and δ_N. It turns out that

$$\begin{aligned} &\mu_N + \delta_N < 0 && \text{for } N \ge 3, \\ &\mu_N + \delta_N = 0 && \text{for } \begin{cases} N = 2, & p > 2, \\ N = 1, & 1 < p < 2, \end{cases} \\ &\mu_N + \delta_N = 1 && \text{for } N = 1, \quad p > 2. \end{aligned} \tag{1.12}$$

We can now integrate (1.10) to obtain the following.

THEOREM 2. *In the situation of Theorem 1, in a small ball centered at 0, we have:*

If $N \ge 3$,

$$|u(x) - \gamma_N u_N(x)| \le C|x|^{\mu_N + \delta_N}, \tag{1.13}$$

if $N = 2$, $p > 2$ or $N = 1$, $1 < p < 2$,

$$|u(x) - \gamma_N u_N(x)| \le C \log \frac{1}{|x|}, \tag{1.14}$$

and if $N = 1$, $p > 2$, $\lim_{x \to 0} u(x) = u(0)$ exists, $u - \gamma_1 u_1$ is Lipschitz in a neighborhood of 0 and

$$|u(x) - u(0) - \gamma_1 u_1(x)| \le C|x|. \tag{1.15}$$

When $p > 2$ we may have bounded singularities corresponding to the case of nonnegative β in (1.10). Theorem 2 says in this case that $u - \gamma_1 u_1$ is Lipschitz, that is, u is "as good as u_1" in terms of smoothness. This is an extension of Serrin and Kichenassamy–Veron results when $p > n = 2$. It would be very interesting to see whether some version of (1.15) holds in the case $p > n \ge 3$.

The next corollary answers the two-dimensional case of a question posed by Veron in [V].

COROLLARY 1. *With the hypothesis of Theorem* 1, *if* u *is unbounded at zero, for some* $N = 1, 2, \cdots$

$$\lim_{x\to 0} \frac{u(x)}{|x|^{\mu_N}} = \gamma_N \Phi_N\left(\frac{x}{|x|}\right) \tag{1.16}$$

and if u *is bounded at zero,* $\lim_{x\to 0} u(x) = u(0)$ *exists,* $p > 2$, *and*

$$\lim_{x\to 0} \frac{(u(x) - u(0))}{|x|^{(p-2)/(p-1)}} = \gamma_1. \tag{1.17}$$

This article is organized as follows. In §2 we linearize the p-Laplacian in the hodograph plane following [IM]. Next in §3 we find a series expansion of the solutions of the linearized p-Laplacian and use it to obtain several estimates. In §4 we study how to recover p-harmonic functions from their "hodograph transforms." This is used in §5 to construct the family u_N in (1.8) and find the μ_N's. The details of the proofs of Theorems 1 and 2 are presented in §6. Finally, in §7 we present an application of Theorem 2 to a problem of uniqueness.

2. The hodograph transformation. Let $u : (\mathbf{B}^2)' \to \mathbf{R}$ be a p-harmonic function. The complex gradient of u, $f = \partial u/\partial z$, is

$$f = \tfrac{1}{2}(u_x, -iu_y)$$

and it is a fundamental fact that f is a K-quasiregular mapping in $(\mathbf{B}^2)'$, where

$$K = \max\left\{p - 1, \frac{1}{p-1}\right\}.$$

See [IM], [M].

It follows from the existence theorem for the Beltrami equation [A, Chap. 5] and the chain rule for the complex dilatation [A, p. 9] that there exists $g_1 = \overline{\mathbf{R}^2} \to \overline{\mathbf{R}^2}$ K-quasiconformal fixing 0, 1, and ∞ and an analytic function h defined on $g_1((\mathbf{B}^2)')$ (itself a punctured neighborhood of 0) such that

$$f(z) = (h \circ g_1)(z) \tag{2.1}$$

for $z \in (\mathbf{B}^2)'$.

LEMMA 1. *Suppose* h *has a pole of order* N *at* 0, $N \geq 1$. *Then there exists a neighborhood of zero* $U \subset \mathbf{B}^2$ *and* $g : U \to U$ K*-quasiconformal, satisfying* $g(0) = 0$, *such that*

$$f(z) = (g(z))^{-N}$$

for $z \in U' = U\backslash\{0\}$.

In other words, near the singularity we may assume $h(\zeta) = \zeta^{-N}$.

Proof. This is an exercise in complex analysis. The function $\frac{1}{h}$ has a zero of order N at 0. So it has an N-rooth l, which is invertible in a neighborhood of 0. It is easily checked that $h \circ l^{-1}(\zeta) = \zeta^{-N}$ and $g = l \circ g_1$. □

As it is shown in [BI] and [IM], $f \in W^{1,2}_{\text{loc}}((\mathbf{B}^2)')$ and (1.2) translates into

$$\frac{\partial f}{\partial \bar{z}} = \left(\frac{1}{p} - \frac{1}{2}\right)\left\{\frac{\bar{f}}{f}\frac{\partial f}{\partial z} + \frac{f}{\bar{f}}\frac{\overline{\partial f}}{\partial z}\right\}. \tag{2.2}$$

Therefore, in a neighborhood of 0, g satisfies

$$\frac{\partial g}{\partial \bar{z}} = \left(\frac{1}{p} - \frac{1}{2}\right)\left\{\frac{g}{\bar{g}}\frac{\overline{\partial g}}{\partial z} + \frac{g^N}{\bar{g}^N}\frac{\partial g}{\partial z}\right\}.$$

We now apply the hodograph transformation; i.e., we set H the inverse of g, $z = H(g(z))$, $\xi = g(H(\xi))$ and $H(0) = 0$. We then obtain [IM] the following *linear* equation for H:

$$H_{\bar{\xi}} = \left(\frac{1}{2} - \frac{1}{p}\right)\left[\frac{\xi}{\bar{\xi}}H_\xi + \frac{\xi^N}{\bar{\xi}^N}\overline{H_\xi}\right]. \tag{2.3}$$

This equation is similar, but different from (17) in [IM]. The exponent of $\xi/\bar{\xi}$ has the opposite sign. This introduces some technical complications in the next section.

We shall call H the *hodograph transform* of u.

3. Series expansion for $H(\xi)$. We look for solutions of (2.3) in $W^{1,2}(U)$. Any such solution is a quasiregular mapping and is in $C^\infty(U')$. Write

$$H(r,\theta) = \sum_{k=-\infty}^{\infty} a_k(r)e^{i(k+N)\theta},$$

where

$$a_k(r) = \frac{1}{2\pi}\int_0^{2\pi} H(r,\theta)e^{-i(k+N)\theta}d\theta.$$

We have that $a_k \in C[0,\delta) \cap C^\infty(0,\delta)$ for some $\delta > 0$. Given $\xi = re^{i\theta}$, (2.3) becomes

$$2rH_r = -ipH_\theta + (p-2)ie^{2iN\theta}\bar{H}_\theta. \tag{3.1}$$

Compute Hr, H_θ, and substitute in (3.1) to get

$$2ra_k'(r) = p(k+N)a_k(r) - (p-2)(k-N)\overline{a_{-k}}(r) \tag{3.2}$$

for $k = 0, 1, 2, \cdots$. Proceed as in [IM] to obtain

$$r(ra_k')' - pNra_k' + (N^2 - k^2)(p-1)a_k = 0. \tag{3.3}$$

For $k \in \mathbf{Z}$ the solutions of (3.3) are

$$a_k(r) = A_k^+ r^{\lambda_k^+} + A_k^- r^{\lambda_k^-},$$

where

$$\lambda_k^\pm = \frac{pN \pm \sqrt{4k^2(p-1) + N^2(p-2)^2}}{2} \tag{3.4}$$

and A_k^+, A_k^- are arbitrary complex numbers. Note that λ_k^+ and λ_k^- are even functions of k, $\lambda_k^- < \lambda_k^+$, $\lambda_N^- = 0$, and $\lambda_N^+ = pN$.

If $|k| > N$, $\lambda_k^- < 0$ forcing $A_k^- = 0$ in this case. Furthermore, from (3.2) we obtain $A_N^- = 0$ and

$$\begin{aligned} A_{-k}^+ &= \varepsilon_k^+\overline{A_k^+}, \qquad k = 0, 1, 2, \cdots, \\ A_{-k}^- &= \varepsilon_k^-\overline{A_k^-}, \qquad k = 0, 1, 2, \cdots, N-1, \end{aligned} \tag{3.5}$$

where

$$\varepsilon_k^{\pm} = \frac{p(N+k) - 2\lambda_k^{\pm}}{(p-2)(k-N)} \quad \text{for } k \neq N, \tag{3.6}$$

$$\varepsilon_N^{+} = \frac{p-2}{p}, \tag{3.7}$$

and ε_N^{-} is not defined. Finally, $H(0) = 0$ implies that $A_N^{-} = 0$. We have the following variant of Theorem 2 in [IM], proved in exactly the same way.

THEOREM 3. *Suppose $H \in W^{1,2}(\mathbf{B}^2)$ is a solution of* (2.3) *satisfying $H(0) = 0$. Then*

$$\begin{aligned} H(r,\theta) = \sum_{k=0}^{N-1} & \left[\left\{A_k^{-} e^{i(k+N)\theta} + \varepsilon_k^{-} \bar{A}_k^{-} e^{-i(k-N)\theta}\right\} r^{\lambda_k^{-}}\right. \\ & \left. + \left\{A_k^{+} e^{i(k+N)\theta} + \varepsilon_k^{+} \bar{A}_k^{+} e^{-i(k-N)\theta}\right\} r^{\lambda_k^{+}}\right] \\ & + \sum_{K=N}^{\infty} \left\{A_k^{+} e^{i(k+N)\theta} + \varepsilon_k^{+} \bar{A}_k^{+} e^{-i(k-N)\theta}\right\} r^{\lambda_k^{+}}, \end{aligned} \tag{3.8}$$

where $\lambda_k^{\pm}$ and $\varepsilon_k^{\pm}$ are given by (3.4), (3.6), *and* (3.7), *$A_0^{\pm}, \cdots, A_{N-1}^{\pm}$ are arbitrary complex numbers, and $A_N^{+}, A_{N+1}^{+}, \cdots$ are arbitrary complex numbers satisfying*

$$\sum_{k=N}^{\infty} k|A_k^{+}|^2 < +\infty. \tag{3.9}$$

Moreover, the series (3.8) *is convergent in $W^{1,2}(\mathbf{B}^2)$. Conversely any series expansion like* (3.8) *with coefficients satisfying* (3.9) *gives a $W^{1,2}(\mathbf{B}^2)$ solution of* (2.3) *vanishing at the origin.*

Remarks. 1. Theorem 3 remains true if we replace $\mathbf{B}^2$ by another ball $B_R(0)$ for any $R > 0$ since (2.3) is invariant under $\xi \to \lambda\xi$, $\lambda > 0$.

2. A computation gives, for $k \neq N$,

$$\varepsilon_k^{\pm} = \frac{\lambda_k^{\pm} - N - k}{\lambda_k^{\pm} - N + k}. \tag{3.10}$$

Thus for $k > N$, $|\varepsilon_k^{+}| < 1$. Since $\lambda_k \sim k$ as $k \to \infty$ it follows that (3.9) is equivalent to $H \in W^{1,2}(\mathbf{B}^2)$.

3. It will be useful to rearrange the exponents $\lambda_k^{\pm}$ in increasing order. We obtain

$$0 < \lambda_{N-1}^{-} < \lambda_{N-2}^{-} < \cdots < \lambda_0^{-} < \lambda_0^{+} < \cdots < \lambda_N^{+} < \cdots. \tag{3.11}$$

4. Inverting the hodograph transform. We will now study the possibility of reversing the process

$$u \to f \to g \to H. \tag{4.1}$$

Let us start with a quasiconformal solution of (2.3) such that $H(0) = 0$. By inverting H in a neighborhood of zero we obtain g and $f = g^{-N}$. Since $f_{\bar{z}}$ is real, f is an

exact complex differential. Therefore, in any given simply connected region, we can find a function u with complex gradient f which will be p-harmonic. Since we allow singularities at zero, our domain is not simply connected and we have to be a little careful.

From the invertibility of H we deduce the invertibility of the leading term in the expansion (3.8). The building block in (3.8) is

$$H_k^{\pm} = \{A_k^{\pm} e^{i(k+N)\theta} + \varepsilon_k^{\pm} \bar{A}_k^{\pm} e^{-i(k-N)\theta}\} r^{\lambda_k^{\pm}}, \tag{4.2}$$

which is itself a quasiregular mapping. By using (3.10) and the argument principle it follows that $H_k^{\pm}$ is invertible only when $k = N-1$. Thus, if $N > 1$ we must have

(4.3)
$$A_{N-1}^- \neq 0 \text{ or } A_{N-1}^- = A_{N-2}^- = \cdots = A_0^- = A_0^+ = \cdots = A_{N-2}^+ = 0, \qquad A_{N-1}^+ \neq 0$$

and

$$|A_0^- + \varepsilon_0^- \overline{A_0^-}| + |A_0^+ + \varepsilon_0^+ \overline{A_0^+}| > 0 \tag{4.4}$$

when $N = 1$. We set for $N > 1$

$$H_{N-1}(r,\theta) = \{A_{N-1}^- e^{i(2N-1)\theta} + \varepsilon_{N-1}^- \bar{A}_{N-1}^- e^{i\theta}\} r^{\lambda_{N-1}^-}$$

and

$$H_0(r,\theta) = (A_0^- + \varepsilon_0^- \bar{A}_0^-) e^{i\theta} r^{\lambda_0^-} + (A_0^+ + \varepsilon_0^+ \bar{A}_0^+) e^{i\theta} r^{\lambda_0^+}.$$

Let us call α_N and β_N the first and second powers of r in the expansion (3.8). Observe that

$$\text{if } N \geq 2 \quad \begin{cases} \alpha_N = \lambda_{N-1}^+, & \beta_N = \lambda_N^+, \quad \text{in the + case,} \\ \alpha_N = \lambda_{N-1}^-, & \beta_N = \lambda_{N-2}^-, \quad \text{in the − case, and} \end{cases}$$

$$\text{if } N = 1 \quad \begin{cases} A_0^+ + \varepsilon_0^- A_0^- \neq 0, & \alpha_1 = \lambda_0^-, \beta_1 = \lambda_0^+, \\ A_0^- + \varepsilon_0^- A_0^- = 0, & \alpha_1 = \lambda_0^+, \beta_1 = \lambda_1^+. \end{cases}$$

In any case, we always have

$$|D^\nu H(r,\theta) - D^\nu H_{N-1}(r,\theta)| \leq C_m r^{\beta_N - m}, \tag{4.5}$$

where $|\nu| = m \geq 0$, $r \leq \delta(m)$ and C_m is a constant.

Denote by g_{N-1} the inverse of H_{N-1} and set

$$f_{N-1} = (g_{N-1})^{-N}.$$

Since H is close to H_{N-1} near zero, we expect g to be close to g_{N-1} and f to be close to f_{N-1} near zero. The next lemma corroborates this suspicion and provides us with the main estimate needed in the proof of Theorem 1.

LEMMA 2. *Let m be a nonnegative integer. There exists $\delta = \delta(m) > 0$ and a constant C_m such that*

$$\sum_{|\nu|=m} |D^\nu f(z) - D^\nu f_{N-1}(z)| \leq C_m |z|^{-(N/\alpha_N)-m+(\beta_N/\alpha_N)-1} \tag{4.6}$$

for $0 < |z| < \delta$.

Proof. The proof of this lemma is somewhat long and it will be divided into a series of steps:

Step 1 (Estimates for H and H_{N-1}). There exists $\rho = \rho(m) > 0$ and constants $C > 0$ and C_m such that

$$\begin{aligned} &\text{(i)} \quad (1/C)|\xi|^{\alpha_N} \le H(\xi) \le C|\xi|^{\alpha_N}, \\ &\text{(ii)} \quad (1/C)|\xi|^{2\alpha_N - 2} \le J_H(\xi) = |H_\xi|^2 - |H_{\bar\xi}|^2 \le C|\xi|^{2\alpha_N - 2}, \\ &\text{(iii)} \quad \textstyle\sum_{|\nu|=m} |D^\nu H(\xi)| \le C_m |\xi|^{\alpha_N - m} \end{aligned}$$

for $|\xi| \le \rho$. Moreover, H_{N-1} satisfies the same estimates.

Proof. It follows immediately from Corollary 1 in [IM].

Step 2 (Estimates for g and g_{N-1}). There exist $\delta = \delta(m) > 0$ and constants $C > 0$ and C_m such that

$$\begin{aligned} &\text{(i)} \quad \tfrac{1}{C}\,|z|^{1/\alpha_N} \le |g(z)| \le C|z|^{1/\alpha_N}, \\ &\text{(ii)} \quad C|z|^{-2+2/\alpha_N} \le J_g(z) = |g_z|^2 - |g_{\bar z}|^2 \le C|z|^{-2+2/\alpha_N}, \\ &\text{(iii)} \quad \textstyle\sum_{|\nu|=m} |D^\nu g(z)| \le C_m |z|^{-m+1/\alpha_N} \end{aligned}$$

for $|z| \le \delta$. Moreover g_{N-1} satisfies the same estimates.

Proof. The estimation of higher derivatives of an inverse function is done with a device introduced in [IM, Lemma 1] that we shall use below (Steps 4 and 5). As stated, Step 2 follows from Corollary 2 in [IM].

Step 3 (Estimates for f and f_{N-1}). There exists $\delta = \delta(m) > 0$ and constants $C > 0$ and C_m such that

$$\begin{aligned} &\text{(i)} \quad \tfrac{1}{C}\,|z|^{-N/\alpha_N} \le |f(z)| \le C|z|^{-N/\alpha_N}, \\ &\text{(ii)} \quad \textstyle\sum_{|\nu|=m} |D^\nu f(z)| \le C_m |z|^{-(N/\alpha_N)-m} \end{aligned}$$

for $0 < |z| < \delta$. Moreover, f_{N-1} satisfies the same estimates.

Proof. Since $f = g^{-N}$ (i) is clear and (ii) follows by induction.

Step 4 (Estimate for $g - g_{N-1}$). For $|z| \le \delta = \delta(m)$ and a constant C_m, we have

$$\sum_{|\nu|=m} |D^\nu g(z) - D^\nu g_{N-1}(z)| \le C_m |z|^{-m+1/\alpha_N + (\beta_N/\alpha_N) - 1}.$$

Proof. Given a positive integer s, let H^s denote one of the partials

$$\frac{\partial^s H}{\partial \xi^i \partial \bar\xi^j} \qquad \text{or} \qquad \frac{\overline{\partial^s H}}{\partial \xi^i \partial \bar\xi^j},$$

where $i+j = s$. Any expression of the form $H^{s_1} H^{s_2} \cdots H^{s_k}$, where $s_1+s_2+\cdots+s_k = s$, will be called a monomial of type (s,k). The set of all finite linear combinations of monomials of type (s,k) is denoted $\mathbf{P}(s,k)$ or $\mathbf{P}_H(s,k)$ if we want to emphasize the dependence on H. According to Lemma 1 in [IM] if $|\nu| = m \ge 1$, there exists $P^\nu \in \mathbf{P}(4m-3, 3m-2)$ such that

$$D^\nu g(H(\xi)) = (J_H(\xi))^{1-2m} P^\nu_H(\xi) \tag{4.7}$$

and

$$D^\nu g_{N-1}(H_{N-1}(\xi)) = (J_{H_{N-1}}(\xi))^{1-2m} P^\nu_{H_{N-1}}(\xi). \tag{4.8}$$

Put $z = H(\xi)$. We have

$$\begin{aligned}|D^\nu g(z) - D^\nu g_{N-1}(z)| &\le |D^\nu g(H(\xi)) - D^\nu g_{N-1}(H_{N-1}(\xi))| \\ &\quad + |D^\nu g_{N-1}(H_{N-1}(\xi)) - D^\nu g_{N-1}(H(\xi))| \\ &\equiv \mathrm{I} + \mathrm{II}.\end{aligned}$$

Estimate of I:

$$\begin{aligned}|\mathrm{I}| \le (J_H(\xi) J_{H_{N-1}}(\xi))^{1-2m} \big[&(J_{H_{N-1}}(\xi))^{2m-1} (P_H^\nu(\xi) - P_{H_{N-1}}^\nu(\xi)) \\ &+ P_{H_{N-1}}^\nu(\xi)(J_{H_{N-1}}^{2m-1}(\xi) - J_H^{2m-1}(\xi))\big].\end{aligned}$$

To estimate $P_H^\nu(\xi) - P_{H_{N-1}}^\nu(\xi)$ note that

$$|H^{s_1} H^{s_2} \cdots H^{s_k}(\xi) - H_{N-1}^{s_1} H_{N-1}^{s_2} \cdots H_{N-1}^{s_k}(\xi)| \le C|\xi|^{(k-1)\alpha_N + \beta_N - s}$$

for all monomials of type (s, k). Therefore

$$|P_H^\nu(\xi) - P_{H_{N-1}}^\nu(\xi)| \le C|\xi|^{(3m-3)\alpha_N + \beta_N - 4m + 3} \tag{4.9}$$

and

$$|P_{H_{N-1}}^\nu(\xi)| \le C|\xi|^{(3m-2)\alpha_N - 4m + 3}. \tag{4.10}$$

Now using Step 1, (4.9), and (4.10) we obtain

$$|\mathrm{I}| \le C|\xi|^{-(m+1)\alpha_N + \beta_N + 1}. \tag{4.11}$$

Estimate of II:

$$|\mathrm{II}| = |D^\nu g_{N-1}(H(\xi)) - D^\nu g_{N-1}(H(\xi) + 0(|\xi|^{\beta_N}))|.$$

Since $\beta_N > \alpha_N$ and $|\xi|$ is small, the mean value theorem gives

$$|\mathrm{II}| \le C \sum_{|\nu| = m+1} |D^\nu g_{N-1}(H(\xi) + 0(|\xi|^{\beta_N}))|\, |\xi|^{\beta_N}.$$

By Step 2,

$$|D^\nu g_{N-1}(z)| \le C|z|^{-(m+1)+1/\alpha_N}$$

for $|\nu| = m + 1$. Thus

$$|\mathrm{II}| \le C|H(\xi)|^{-(m+1)+1/\alpha_N} |\xi|^{\beta_N} \le C|\xi|^{-(m+1)\alpha_N + 1 + \beta_N}. \tag{4.12}$$

From (4.11), (4.12), and another application of Step 2 we obtain the estimate in Step 4.

Step 5 (Estimate for $f - f_{N-1}$). This is the final step in which we will prove (4.6).

Proof. Let us begin by observing that $D^\nu(g^{-N})$ is a linear combination of terms of the form

$$g^{-N-k} P_g(m, k), \tag{4.13}$$

where $k = 1, 2, \cdots, m$ and $P_g(m,k)$ is an element of $\mathbf{P}_g(m,k)$. Thus, it suffices to establish (4.6) for the difference of corresponding terms of the form (4.13). Writing

$$\begin{aligned} g^{-N-k}P_g(m,k) - g_{N-1}^{-N-k}P_{g_{N-1}}(m,k) \\ = (gg_{N-1})^{-N-k}\big[g_{N-1}^{N+k}(P_g(m,k) - P_{g_{N-1}}(m,k)) \\ + P_{g_{N-1}}(m,k)(g_{N-1}^{N+k} - g^{N+k})\big] \end{aligned}$$

and using

$$|Pg(m,k) - P_{g_{N-1}}(m,k)| \le C_m|z|^{-m+(k/\alpha_N)+(\beta_N/\alpha_N)-1}$$

and

$$|g^{N+k} - g_{N-1}^{N+k}| \le C_m|z|^{((N+k)/\alpha_N)+(\beta_N/\alpha_N)-1}$$

we can easily finish the proof. □

5. Quasiradial singular p-harmonic functions. Consider first the case $N \ge 2$. Then $H_{N-1}(r,\theta)$ is quasiradial; i.e.,

$$H_{N-1}(r,\theta) = \psi_{N-1}(\theta)r^{\alpha_N} \tag{5.1}$$

and so is its inverse g_{N-1}, and f_{N-1}. Indeed,

$$f_{N-1}(s,\alpha) = \varphi_{N-1}(\alpha)s^{-N/\alpha_N}. \tag{5.2}$$

In (5.1) and (5.2) ψ_{N-1} and φ_{N-1} are certain nontrivial real analytic functions of period 2π.

LEMMA 3A. *If $N \ge 2$ and $\alpha_N = \lambda_{N-1}^-$ the function f_{N-1} is the complex gradient of a p-harmonic function v_N with an isolated singularity at zero. Moreover, v_N is quasiradial and there exists a constant γ_N such that*

$$v_N(se^{i\alpha}) = \gamma_N u_N(se^{i\alpha}),$$

where

$$u_N(se^{i\alpha}) = \Phi_N(\alpha)s^{\mu_N}$$

is as in (1.8) *and $\mu_N = (-N/\alpha_N) + 1$.*

Proof. Write $f_{N-1} = p - iq$. It is enough to check that

$$\int_{|z|=\varepsilon} p\,dx + q\,dy = 0,$$

which follows since p and q are homogeneous functions of degree $-N/\alpha_N \ne -1$.

From the expression of μ_N it is clear that it decreases to $-\infty$ as $N \to \infty$.

LEMMA 3B. *If $N \ge 2$ and $\alpha_N = \lambda_{N-1}^+$ the function f_{N-1} is the complex gradient of a p-harmonic function $\bar{v}_N$ in a punctured neighborhood of zero. Moreover $\bar{v}_N$ is quasiradial and there exists a constant $\bar{\gamma}_N$ such that*

$$\bar{v}_N(se^{i\alpha}) = \bar{\gamma}_N\bar{u}_N(se^{i\alpha}),$$

where

$$\bar{u}_N(se^{i\alpha}) = \bar{\Phi}_N(\alpha)s^{\bar{\mu}_N}$$

and $\bar{\mu}_N = \frac{-N}{\alpha_N} + 1 > (p-2)^+/(p-1)$.

Proof. Again it is easy to check that $-N/\alpha_N \neq -1$. Therefore the existence of $\bar{v}_N$ follows as in Lemma 3A. From (3.4) we have

$$\bar{\mu}_N = \frac{-N}{\lambda_{N-1}^+} + 1 = \frac{-2}{p + \sqrt{4(1-\frac{1}{N})^2 + (p-2)^2}} + 1$$
$$> \frac{-2}{p+|p-2|} + 1 \geq \frac{(p-2)^+}{p-1}. \qquad \square$$

Remark. We will see in §6 that $\bar{v}_N$ are indeed p-harmonic in a neigborhood of zero. These functions are considered also in [Ar] and in [Kv].

We now consider the case $N = 1$. The previous argument does not work since we may have $\lambda_0^- = 1$. It is convenient to distinguish whether $1 < p < 2$ or $p > 2$.

Case $1 < p < 2$**.** For certain $\alpha, \beta \in \mathbf{R}$ we may write

$$H_0(r,\theta) = \alpha e^{i\theta} r^{p-1} + i\beta e^{i\theta} r. \tag{5.3}$$

If $\alpha = 0$, then $H_0(\xi) = i\beta\xi$, for $\xi \in \mathbf{C}$ and it is not hard to see that $f_0 = (g_0)^{-1}$ is not the complex gradient of a function in a neighborhood of zero.

If $\alpha \neq 0$, we redefine

$$H_0(r,\theta) = \alpha e^{i\theta} r^{p-1} \tag{5.4}$$

and check easily that H_0 is the hodograph transform of a multiple of

$$u_1(se^{i\alpha}) = s^{\mu_1},$$

where $\mu_1 = (p-2)/(p-1)$.

Case $p > 2$**.** In this case we write

$$H_0(r,\theta) = i\alpha e^{i\theta} r + \beta e^{i\theta} r^{p-1} \tag{5.5}$$

for certain $\alpha, \beta \in \mathbf{R}$. If $\alpha = 0$, H_0 is the hodograph transform of a multiple of u_1 and if $\alpha \neq 0$, we redefine

$$H_0(r,\theta) = i\alpha e^{i\theta} r, \tag{5.6}$$

which is not the hodograph transform of any function in a neighborhood of zero.

LEMMA 4. *Let* $N = 1$. *Then* H_0 (*redefined according to* (5.4) *and* (5.6)) *is the hodograph transform of a p-harmonic function if and only if* H *is the hodograph transform of a p-harmonic function in a neighborhood of zero.*

Proof. It is enough to show that $f - f_0$ is the gradient of a function in a punctured neighborhood of zero. From Lemma 2 we have

$$\int_{|z|=\varepsilon} |f - f_0| d|z| \leq C\varepsilon^{(-1+\beta_1)/\alpha_1}.$$

Since $\beta_1 = \lambda_0^+$ or $\beta_1 = \lambda_1^+$, it follows that $\beta_1 > 1$. Therefore

$$\int_{|z|=\varepsilon} (p - p_0)dx + (q - q_0)dy = 0. \qquad \square$$

From Lemma 4 we conclude that if H is the hodograph transform of some u, we must have

$$\begin{aligned} &\text{If } 1 < p < 2, \quad \operatorname{Re} A_0^- \neq 0, \quad \text{and} \\ &\text{if } p > 2, \quad \operatorname{Im} A_0^- = 0 \quad \text{and} \quad \operatorname{Re} A_0^+ \neq 0. \end{aligned} \tag{5.7}$$

In particular $\alpha_1 = p - 1$ and we also have $\mu_1 = (-1/\alpha_1) + 1$.

6. Proof of Theorems 1 and 2. Let u be a singular p-harmonic function in $(\mathbf{B}^2)'$ satisfying (1.10).

LEMMA 5. *For $0 < |x| < \frac{1}{2}$ we have*

$$|\nabla u(x)| \leq C|x|^{\beta-1}.$$

Proof. If a ball of radius $2R$, $B_{2R} \subset (\mathbf{B}^2)'$ Cacciopoli's inequality gives

$$\left(\fint_{B_R} |\nabla u|^p dx \right)^{1/p} \leq \frac{C}{R} \left(\fint_{B_{2R}} |u|^p \right)^{1/p}.$$

Fix $x, |x| < \frac{1}{2}$. For $y \in B_{|x|/2}(x)$ we have $|u(y)| \leq C|y|^\beta$, $|x|/2 \leq |y| \leq 3|x|/2$ and

$$\left(\fint_{B_{(|x|/4)}(x)} |\nabla u|^p dx \right)^{1/p} \leq \frac{C}{|x|} |x|^\beta = C|x|^{\beta-1}. \tag{6.1}$$

The lemma now follows from (6.1) and the L^∞-estimate for ∇u ([DB], [L]):

$$\sup_{B_{|x|/8}(x)} |\nabla u| \leq C \left(\int_{B_{|x|/4}(x)} |\nabla u|^p dx \right)^{1/p}. \qquad \square$$

Let $f = \frac{1}{2}(u_x - iu_y)$ and write $f = h \circ g_1$ as in (2.1). From Lemma 5 and the Hölder continuity of quasiconformal mappings ([LV], [A]) we infer that h has a pole at 0 of order $N \geq 0$.

If $N = 0$, f is indeed quasiregular in a neighborhood of the origin. Thus u is smooth at 0 and the singularity is removable.

Consider now the case where $N \geq 2$ and $\alpha_N = \lambda_{N-2}^+$. Applying Lemma 2 for $m = 0$ and Lemma 3B, we obtain

$$|\nabla u - \bar{\gamma}_N \nabla \bar{u}_N| \leq C_o |x|^{\bar{\mu}_N - 1 + ((\beta_N/\alpha_N)-1)}.$$

Therefore we have

$$|\nabla u(x)| \leq C|x|^{\mu_N - 1}.$$

Since $\bar{\mu}_N > (p-2)^+/(p-1) > (p-2)/p$, it follows that

$$\int_{|x|<\delta} |\nabla u|^p \, dx < \infty$$

for small $\delta > 0$. Let now $\phi \in C_0^\infty(|x| < \delta)$,

$$\int_{|x|<\delta} |\nabla u|^{p-2}\langle \nabla u, \nabla \phi\rangle \, dx = \int_{|x|<\epsilon} |\nabla u|^{p-2}\langle \nabla u, \nabla \phi\rangle \, dx + \int_{\epsilon<|x|<\delta} |\nabla u|^{p-2}\langle \nabla u, \nabla \phi\rangle \, dx.$$

By Hölder's inequality and the integrability of $|\nabla u|^p$ in $|x| < \delta$ the first integral tends to zero as $\epsilon \to 0$. Transform the second integral using the divergence theorem,

$$\int_{\epsilon<|x|<\delta} |\nabla u|^{p-2}\langle \nabla u, \nabla \phi\rangle \, dx = \int_{|x|=\epsilon} |\nabla u|^{p-2}\frac{\partial u}{\partial r}\phi \, dx.$$

Thus

$$\left| \int_{\epsilon<|x|<\delta} |\nabla u|^{p-2}\langle \nabla u, \nabla \phi\rangle \, dx \right| \le C\epsilon\epsilon^{(p-1)(\bar{\mu}_N - 1)}\|\phi\|_\infty,$$

which tends to zero as $\epsilon \to 0$ since $\bar{\mu}_N > (p-2)^+/(p-1) \ge (p-2)/(p-1)$. Therefore u is indeed p-harmonic in a neigborhood of zero, and the singularity is again removable.

Thus $N \ge 1$ of $N \ge 2$ and $\alpha_N = \lambda_{N-1}^-$ for a true singularity.

From Lemma 1 we write $f = g^{-N}$ in a neighborhood of zero, g quasiconformal satisfying $g(0) = 0$ and apply the results of §§3, 4, and 5. From Lemma 2 we obtain

$$\sum_{|\nu|=m+1} |D^\nu u - \gamma_N D^\nu u_N| \le C_m |x|^{-(N/\alpha_N)-m+((\beta_N/\alpha_N)-1)} \tag{6.2}$$

for m nonnegative integer. Set

$$\delta_N = \frac{\beta_N}{\alpha_N} - 1 > 0$$

and we have proved (1.11).

Combining Lemma 5 and the estimates in Step 3 of the proof of Lemma 2 we see that

$$|x|^{-N/\alpha_N} \le C|x|^{\beta-1}.$$

Therefore

$$\beta \le 1 - \frac{N}{\alpha_N} = \mu_N. \tag{6.3}$$

Indeed, N is the largest integer such that (6.3) holds. If $\beta > (p-2)/(p-1)$, N must be zero and so 0 is a removable singularity of u. This justifies the remark after Theorem 1 and finishes its proof.

To prove Theorem 2 all we have to do is to integrate (6.2) when $m = 0$. From the definitions of δ_N, μ_N, α_N, β_N, and (3.4) we can easily check (1.12).

Write (6.2) for $m = 0$,

$$|\nabla u(x) - \gamma_N \nabla u_N(x)| \le C|x|^{\mu_N - 1 + \delta_N}. \tag{6.4}$$

In particular,

$$|\nabla u(x)| \le C|x|^{\mu_N - 1}. \tag{6.5}$$

We now integrate (6.5). Let s be a small positive number and $|x| < s$,

$$\left|u(x) - u\left(s\frac{x}{|x|}\right)\right| \leq C\int_{|x|}^{s} t^{\mu_N - 1}dt \leq C|s^{\mu_N} - |x|^{\mu_N}| \tag{6.6}$$

since $\mu_N \neq 0$.

If $\mu_N < 0$, (6.6) implies

$$|u(x)| \leq C|x|^{\mu_N} \tag{6.7}$$

for small $|x|$. On the other hand, if $\mu_N > 0$ (Case $N = 1$, $p > 2$) we obtain

$$|u(x)| \leq C. \tag{6.8}$$

Similarly we integrate (6.4) to get

$$\left|u(x) - \gamma_N u_N(x) - u\left(s\frac{x}{|x|}\right) + \gamma_N u_N\left(s\frac{x}{|x|}\right)\right| \leq C\int_{|x|}^{s} t^{\mu_N - 1 + \delta_N}dt \tag{6.9}$$

for $|x| < \delta$. We will consider several cases:

(a) $N \geq 3$. Since $\mu_N + \delta_N < 0$, (1.13) follows from (6.9).

(b) $N = 2$, $p > 2$ and $N = 1$, $1 < p < 2$. Again (1.14) follows from (6.9).

(c) $N = 1$, $p > 2$. Since $\mu_N + \delta_N = 1$, (6.4) implies that $u - \gamma_1 u_1$ is Lipschitz across the origin. Thus (1.15) follows easily.

7. Remark on a uniqueness problem. Suppose $u \in W^{1,p-1}(\mathbf{B}^n)$ and u is p-harmonic in the following "ultra-weak" sense: (1.1) is required to hold only for test functions $\phi \in C_0^\infty(\mathbf{B}^n)$. Does it follow that u is in $W^{1,p}(\mathbf{B}^n)$ (and so that u is a classical p-harmonic function)?

Theorem 1 allows us to answer a particular case of this question in two dimensions.

THEOREM 4. *Suppose u is p-harmonic in $(\mathbf{B}^2)' = \mathbf{B}^2 - \{0\}$, $\nabla u \in L^{p-1}(\mathbf{B}^2)$ and*

$$\int_{\mathbf{B}^2} |\nabla u|^{p-2}\langle \nabla u, \nabla \phi\rangle dx = 0 \tag{7.1}$$

for all $\phi \in C_0^\infty(\mathbf{B}^2)$. Then u is p-harmonic in $\mathbf{B}^2$.

First we need the following lemma.

LEMMA 6. *Let u be a p-harmonic function in $(\mathbf{B}^2)'$ such that $\nabla u \in L^\epsilon(\mathbf{B}^2)$ for some $\epsilon > 0$. Then, there exists $\lambda > 0$ and a constant C such that*

$$|u(x)| \leq C|x|^{-\lambda} \tag{7.2}$$

for $|x| < \frac{1}{2}$.

Proof. Set $f = \partial u/\partial z$ as in §2 and write $f = h \circ g_1$ as in (2.1). Since g_1 is K-quasiconformal, where $K = \max\{p-1, 1/(p-1)\}$, there exists $\delta_p > 0$ such that

$$\int_{|x|<1/2} (J_{g_1}(x))^{1+\delta_p}\, dx < \infty. \tag{7.3}$$

Here J_{g_1} is the Jacobian of g_1 and (7.3) follows from the higher integrability of Jacobians of quasiconformal mappings [LV].

Let $\eta > 0$ be a small number such that $g_1^{-1}(\{w : |w| < \eta\})$ is contained in $\{x : |x| < \frac{1}{2}\}$, and set $q = 1+\delta_p$, $r = (1+\delta_p)/\delta_p$, $s = \epsilon\delta_p/(1+\delta_p)$, and $t = \delta_p/(1+\delta_p)$. Then

$$\begin{aligned}\int_{|w|<\eta} |h(w)|^s\, dw &= \int_{|w|<\eta} |h(w)|^s J^t_{g_1^{-1}}(w) J^{-t}_{g_1^{-1}}(w)\, dw \\ &\le \left(\int_{|w|<\eta} |h(w)|^{sr} J^{tr}_{g_1^{-1}}(w)\, dw\right)^{1/r} \left(\int_{|w|<\eta} \left(J_{g_1^{-1}}(w)\right)^{-tq}\, dw\right)^{1/q} \\ &= A^{1/r} B^{1/q}.\end{aligned}$$

To estimate A we change variables and use $f \in L^\epsilon(\mathbf{B}^2)$,

$$A = \int_{g_1^{-1}(|w|<\eta)} |h \circ g_1(x)|^\epsilon\, dx < \infty.$$

Similarly, after changing variables and using (7.3)

$$\begin{aligned}B &= \int_{g_1^{-1}(|w|<\eta)} \left[J_{g_1^{-1}}(g(x))\right]^{-\delta_p} J_{g_1}(x)\, dx \\ &= \int_{g_1^{-1}(|w|<\eta)} [J_{g_1}(x)]^{1+\delta_p}\, dx < \infty.\end{aligned}$$

Therefore h is a holomorphic function in a punctured neighborhood of zero satisfying

$$\int_{|w|<h} |h(w)|^{\epsilon\delta_p/(1+\delta_p)}\, dw < \infty. \tag{7.4}$$

For $|w| < \eta/2$ the subharmonicity of $|h|^{\epsilon\delta_p/(1+\delta_p)}$ gives

$$|h(w)|^{\epsilon\delta_p/(1+\delta_p)} \le \frac{4}{\pi|w|^2} \int\int_{|w-\eta|<|w|/2} |h(\eta)|^{\epsilon\delta_p/(1+\delta_p)}\, d\eta.$$

Thus $|h(w)| \le C|w|^\nu$ for some $\nu > 0$ and $|w| < \eta/2$. Since g_1 is Hölder continuous, it follows that

$$|\nabla u(x)| = |f(x)| \le C|x|^{-\lambda-1}$$

for some $\lambda > 0$ and $|x| < 1/2$. Upon integration we obtain the estimate (7.2). □

Proof of Theorem 4. From (7.1) and the divergence theorem we obtain

$$\lim_{\varepsilon\to 0} \int_{|x|=\varepsilon} |\nabla u|^{p-2} \frac{\partial u}{\partial r}\, d\sigma = 0. \tag{7.5}$$

If u is singular p-harmonic, from (1.11) we deduce

$$\nabla u(x) = \gamma_N \nabla u_N(x) + 0(|x|^{\mu_N - 1 + \delta_N})$$

for some constant $\gamma_N \neq 0$ and $N \geq 1$. It follows then that

$$\int_{|x|=\varepsilon} |\nabla u|^{p-2} \frac{\partial u}{\partial r}\, d\sigma = |\gamma_N|^{p-2}\gamma_N \int_{|x|=\varepsilon} |\nabla u_N|^{p-2} \frac{\partial u_N}{\partial r} + O(\varepsilon^{(\mu_N-1)(p-1)+\delta_N+1}).$$

Since

$$\int_{|x|=\varepsilon} |\nabla u_N|^{p-2} \frac{\partial u_N}{\partial r} = C\varepsilon^{(\mu_N-1)(p-1)+1}$$

and $\gamma_N \neq 0$, $C \neq 0$ we must have $(\mu_N-1)(p-1)+1 > 0$. Therefore $\mu_N > (p-2)/(p-1)$, which is impossible. Thus we have arrived at a contradiction if we assume that u is singular p-harmonic. So it must be that u is p-harmonic in $\mathbf{B}^2$. □

Acknowledgments. This paper was completed while the author was visiting the Department of Mathematics at Northwestern University. The author wishes to thank the members of this Department for the pleasant mathematical atmosphere they offered him. Especial thanks are due to Emmanuele DiBenedetto for enlightening conversations about p-harmonic functions and for suggesting the question treated in §7.

The author would like also to thank the referee who pointed out the need for Lemma 3B and Lemma 6.

REFERENCES

[A] L. Ahlfors, *Lectures on Quasiconformal Mappings*, Wadsworth & Brooks, Belmont, CA, 1987.

[Ar] G. Aronsson, *Representation of a p-harmonic function near a critical point in the plane,* Linköping University, Sweden, preprint LiTh-MAT-R-88-06.

[BI] B. Bojarski and T. Iwaniec, *p-harmonic functions and quasiregular mappings,* preprint, Universität Bonn, Bonn, FRG.

[DB] E. DiBenedetto, *$C^{1+\alpha}$-local regularity of weak solutions of degenerate elliptic equations,* Nonlinear Anal., Theory, Methods, Appl., 7 (1983), pp. 827–850.

[IM] T. Iwaniec and J. Manfredi, *Regularity of p-harmonic functions in the plane,* Rev. Mat. Iberoamericana, 5 (1989), pp. 1–19.

[KV] S. Kichenassamy and L. Veron, *Singular solutions of the p-Laplace equation,* Math. Ann., 273 (1986), pp. 599–615.

[L] J. Lewis, *Regularity of the derivatives of solutions to certain elliptic equations,* Indiana Univ. Math. J., 32 (1983), pp. 849–858.

[LV] O. Lehto and K. Virtanen, *Quasiconformal Mappings in the Plane*, Springer-Verlag, Berlin, 1973.

[M] J. Manfredi, *p-harmonic functions on the plane,* Proc. Amer. Math. Soc., 103(2) (1988), pp. 473–479.

[S] J. Serrin, *Local behaviour of solutions of quasilinear equations,* Acta Math., 111 (1964), 247–302.

[V] L. Veron, *Singularities of some quasilinear equations,* in *Nonlinear Equations and their Equilibrium States,* W. N. Ni, L. Peletier, and J. Serrin, eds., Math. Sci. Res. Inst. Publ., Springer, Berlin, 1988.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 440-462, March 1991

010

APPROXIMATION OF SOLUTIONS OF SINGULAR SECOND-ORDER BOUNDARY VALUE PROBLEMS*

A. M. FINK†, JUAN A. GATICA‡, GASTON E. HERNANDEZ§, AND PAUL WALTMAN¶

Abstract. The boundary value problems

$$(P_1)\qquad y''+\frac{n-1}{x}y'+f(x,y)=0,\qquad y'(0)=y(1)=0,$$

and

$$(P_2)\qquad y''+f(x,y)=0,\quad \alpha y(0)-\beta y'(1)=0,\quad \gamma y(1)+\delta y'(1)=0$$

with the function $f(x, y)$ satisfying conditions that allow for singularities to be present are studied with the view of obtaining general existence, uniqueness, and approximation of positive solutions. Furthermore, the behavior of solutions near $x=1$ is described for the special case $f(x, y)=a(x)y^{-p}$.

Key words. singular nonlinear boundary value problems, radially symmetric solutions of nonlinear elliptic partial differential equations, order reversing operators, sequence of iterates, existence and uniqueness

AMS(MOS) subject classifications. primary 34B15; secondary 34A45, 35J65

1. Introduction. This paper deals with the existence, uniqueness, and approximation of solutions of boundary value problems of the following types:

$$(P_1)\qquad y''+\frac{n-1}{x}y'+f(x,y)=0,\qquad y'(0)=y(1)=0,$$

and

$$(P_2)\qquad y''+f(x,y)=0,\quad \alpha y(0)-\beta y'(1)=0,\quad \gamma y(1)+\delta y'(1)=0.$$

In both problems it is desired to include the case when $f(x, y)$ is singular at $y=0$; the conditions that are imposed on f are such that the special case $f(x, y)=a(x)y^{-p}$, $p>0$, $a(x)$ continuous is included in the results and the behavior of the solution of (P_1) in this special case is completely described for x near 1, i.e., near the point where the singularity occurs.

The paper is organized as follows. Section 2 deals with transforming these problems into fixed-point problems by using the appropriate Green's function integral operator and showing the equivalence of the problems. Once this is done the monotonicity introduced by the singularity in the dependent variable is used to determine a region

* Received by the editors May 1, 1989; accepted for publication (in revised form) March 8, 1990.

† Department of Mathematics, Iowa State University, Ames, Iowa 50011 and Department of Mathematics, University of Iowa, Iowa City, Iowa 52242.

‡ Department of Mathematics, University of Iowa, Iowa City, Iowa 52242.

§ Department of Mathematics, University of Connecticut, Storrs, Connecticut 06268.

¶ Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia 30322.

bounded between the graphs of two functions constructed iteratively where the graph of any solution must lie provided that certain initial inequalities hold; furthermore, it is shown that under slightly more restrictive conditions there can exist at most one solution.

Section 3 is devoted to finding a general approximation scheme, by regularizing the problems and showing that the conditions required in § 2 hold for the regularized problems. A general uniqueness result is obtained and general conditions are found that imply the existence of solutions to the problems and also some of their regularity properties.

Section 4 is devoted exclusively to the discussion of the special case $f(x, y) = a(x)y^{-p}$ for problem (P_1). The behavior of solutions when x is near 1 is described.

Section 5 simply shows how the approximation method used in the previous sections works for some specific examples. We describe the method used to compute the iterations and show graphs of the approximations for various specific examples.

The existence of positive solutions to (P_1) belonging to $C^1([0, 1]) \cap C^2((0, 1))$ was dealt with in [5] while those to (P_2), in the same class of functions, were studied in [4]: uniqueness results were given in both papers. An iterative procedure to approximate some positive solutions to (P_2) when $f(x, y) = a(x)y^{-p}$, $0 < p \leqq 1$, was described in [6].

Problem (P_1) comes from the study of existence of positive radially symmetric solutions to

$$\Delta u + f(x, u) = 0 \quad \text{in } \Omega,$$
$$u \,|\, \partial\Omega = 0,$$

where Ω is the open unit ball centered at the origin in $\mathbb{R}^n$; this partial differential equation with $f(x, u) = a(x)u^{-p}$, in more general domains but without radial symmetry, was studied in [13]. Problems with similar singularities were studied in [2].

Problem (P_2) is a generalized Emden-Fowler equation and has been extensively studied; [9] provides an excellent survey of the literature concerning Emden-Fowler equations. Equations with similar types of singularities were studied in [1], [3], [9], and [10].

2. Equivalent integral equation. In what follows it will be assumed, depending on the problem being considered, that the function f satisfies some of the following hypotheses:

- H_1 $f:[0, 1) \times (0, \infty) \to (0, \infty)$ is continuous.
- H_1' $f:(0, 1) \times (0, \infty) \to (0, \infty)$ is continuous.
- H_2 $f(x, y)$ is strictly decreasing in y for $x \in (0, 1)$, and integrable over $[0, 1]$ for each fixed $y > 0$.
- H_3 $\alpha, \beta, \gamma, \delta$ are nonnegative and $\rho = \gamma\beta + \alpha\gamma + \alpha\delta > 0$.

Note that none of these hypotheses implies the existence of a singularity.

It can be easily verified that the Green's function for the problem

$$y'' + \frac{n-1}{x} y' = 0, \qquad y'(0) = y(1) = 0,$$

is the function $G:(0, 1] \times (0, 1] \to [0, \infty)$ given by

$$G(x, t) = \begin{cases} \dfrac{1}{n-2} t^{n-1}(t^{2-n} - 1), & 0 < x \leqq t \leqq 1, \\ \dfrac{1}{n-2} t^{n-1}(x^{2-n} - 1), & 0 < t \leqq x \leqq 1, \end{cases} \tag{1}$$

if $n>2$ and

$$G(x,t)=\begin{cases}-t\ln(t), & 0<x\leqq t\leqq 1,\\ -t\ln(x), & 0<t\leqq x\leqq 1,\end{cases} \tag{2}$$

if $n=2$. It is clear that in either case $G(x,t)>0$ if $(x,t)\in(0,1)\times(0,1)$.

The Green's function for

$$\begin{aligned} &y''=0,\\ &\alpha y(0)-\beta y'(0)=0,\\ &\gamma y(1)+\delta y'(1)=0,\end{aligned} \tag{3}$$

assuming that $\rho>0$, is given by

$$G(x,t)=\begin{cases}\dfrac{1}{\rho}(\gamma+\delta-\gamma x)(\beta+\alpha t), & 0\leqq t\leqq x\leqq 1,\\ \dfrac{1}{\rho}(\beta+\alpha x)(\gamma+\delta-\gamma t), & 0\leqq x\leqq t\leqq 1,\end{cases}$$

and if H_3 is satisfied then $G(x,t)>0$ for $(x,t)\in(0,1)\times(0,1)$.

In what follows it will often be necessary to have that positive solutions of (P_1) or (P_2) are also solutions of the integral equation:

$$y(x)=\int_0^1 G(x,t)f(t,y(t))\,dt. \tag{4}$$

This may follow from the particular assumptions of the cases considered.

Before entering a detailed discussion of the particular results obtained, it will be convenient to establish a general principle which is the basis for what is to come. To avoid introducing excessive notation, X will stand for the space of real-valued continuous functions defined on $[0,1]$, with supremum norm, K will be the cone, in X, of nonnegative functions, $\leqq$ will denote the order induced in X by K, and D will be either the set

$$D=\{\phi\in K:\exists\theta>0 \text{ such that } \phi(x)\geqq\theta(1-x),\ x\in[0,1]\}, \tag{5}$$

when discussing (P_1), or the set

$$D=\{\phi\in K:\exists\theta>0 \text{ such that } \phi(x)\geqq g_\theta(x),\ x\in[0,1]\}, \tag{6}$$

where

$$g_\theta(x)=\begin{cases}\theta x, & 0\leqq x\leqq\frac{1}{2},\\ \theta(1-x), & \frac{1}{2}\leqq x\leqq 1,\end{cases}$$

when discussing (P_2). T will denote the Green's function operator

$$T\phi(x)=\int_0^1 G(x,t)f(t,\phi(t))\,dt \tag{7}$$

considered over its "natural domain," i.e., the set of all functions in K for which this formula defines a continuous function on $[0,1]$. Note that if f satisfies H_2 and if ϕ is in the natural domain of T, then any function $\psi\geqq\phi$ is also in the natural domain of T. If f is sufficiently well behaved, then D is contained in the natural domain of T and the restriction of T to D is easily seen to be such that $T:D\to D$; see [5].

It is important to note, however, that under very mild conditions any positive solution of (P_1) or (P_2) is a solution of the corresponding version of (4).

LEMMA 2.1. *If* H_1 *and* H_2 *hold, any positive solution of* (P_1) *is also a solution of* (4). *If* H_1', H_2 *and* H_3 *hold, any positive solution of* (P_2) *is also a solution of* (4).

Proof. If H_1 and H_2 hold and $\phi:[0,1]\to[0,\infty)$ is a solution of (P_1), then

$$(x^{n-1}\phi'(x))' = -x^{n-1}f(x,\phi(x)), \qquad x\in(0,1),$$

and $\phi'(0)=\phi(1)=0$. Thus, if $x\in(0,1]$,

$$x^{n-1}\phi'(x) = -\int_0^x t^{n-1}f(t,\phi(t))\,dt,$$

$$\phi'(x) = -\frac{1}{x^{n-1}}\int_0^x t^{n-1}f(t,\phi(t))\,dt$$

and, integrating again:

$$\phi(x)-\phi(0) = -\int_0^x\int_0^t \left(\frac{u}{t}\right)^{n-1} f(u,\phi(u))\,du\,dt$$

$$= -\int_0^x\int_u^x \left(\frac{u}{t}\right)^{n-1} f(u,\phi(u))\,dt\,du.$$

The cases $n=2$ and $n>2$ must be discussed separately. If $n=2$,

$$\phi(x)-\phi(0) = -\int_0^x\int_u^x \frac{u}{t}f(u,\phi(u))\,dt\,du$$

$$= -\int_0^x (\ln(x)-\ln(u))uf(u,\phi(u))\,du$$

$$= -\ln(x)\int_0^x uf(u,\phi(u))\,du + \int_0^x u\ln(u)f(u,\phi(u))\,du$$

and, in particular,

$$\phi(1)-\phi(0) = -\phi(0) = \int_0^1 u\ln(u)f(u,\phi(u))\,du.$$

This implies that

$$\int_0^1 |u\ln(u)|f(u,\phi(u))\,du<\infty$$

and the Green's function operator T may be applied to ϕ, obtaining

$$T\phi(x) = -\int_0^x t\ln(x)f(t,\phi(t))\,dt - \int_x^1 t\ln(t)f(t,\phi(t))\,dt.$$

This implies that $T\phi(1)=0$ and

$$(T\phi)'(x) = -\frac{1}{x}\int_0^x tf(t,\phi(t))\,dt,$$

$$(T\phi)''(x) = \frac{1}{x^2}\int_0^x tf(t,\phi(t))\,dt - f(x,\phi(x)), \qquad x\in(0,1),$$

so that

$$(T\phi)''(x) = -\frac{1}{x}(T\phi)'(x) - f(x, \phi(x)), \qquad x \in (0, 1),$$

which in turn implies that

$$(x(T\phi)'(x))' = -f(x, \phi(x)) = (x\phi'(x))', \qquad x \in (0, 1).$$

Since $T\phi(1) = \phi(1) = 0$, it must be the case that

$$x(T\phi)'(x) = x\phi'(x), \qquad x \in (0, 1),$$

and so

$$(T\phi)'(x) = \phi'(x), \qquad x \in (0, 1).$$

Since $T\phi(0) = \phi(0)$, the conclusion is that

$$T\phi(x) = \phi(x), \qquad x \in [0, 1].$$

The case $n > 2$ follows in identical fashion, the only difference being that

$$\int_u^x t^{-n+1}\, dt = \frac{1}{-n+2}[x^{-n+2} - u^{-n+2}].$$

If H_1', H_2, and H_3 hold, the discussion is simpler since then for any ψ in the domain of the Green's function operator:

$$(T\psi)''(x) = -f(x, \psi(x)), \qquad x \in (0, 1).$$

THEOREM 2.2. *If f satisfies H_1 or H_1', H_2 (and H_3 if considering (P_2)), and if there exists $\phi \in K$ such that $T\phi$, $T^2\phi$ are defined and either $T\phi \leqq \phi$, $T^2\phi \leqq \phi$ or $T\phi \geqq \phi$, $T^2\phi \geqq \phi$, then the sequence of iterates $\{T^n\phi\}_{n=0}^{\infty}$ is defined, the subsequences $\{T^{2n}\phi\}_{n=0}^{\infty}$, $\{T^{2n+1}\phi\}_{n=0}^{\infty}$ are monotone and uniformly convergent, with limits*

$$\psi_0 = \lim_{n\to\infty} T^{2n}\phi, \qquad \psi_1 = \lim_{n\to\infty} T^{2n+1}\phi.$$

Furthermore, if the function $yf(x, y)$ is strictly monotone for each x, then $\psi_0 = \psi_1$.

Proof. Consider the case when $T\phi \geqq \phi$ and $T^2\phi \geqq \phi$; the proof for the other case is analogous.

Since T is decreasing, it follows that

$$\phi \leqq T^2\phi \leqq T\phi$$

and hence

$$\phi \leqq T^2\phi \leqq T^3\phi \leqq T\phi.$$

It follows by induction that $\phi \leqq T^n\phi \leqq T\phi$ for all $n \geqq 1$ and that the subsequences $\{T^{2n}\phi\}_{n=0}^{\infty}$, $\{T^{2n+1}\phi\}_{n=0}^{\infty}$ are monotone increasing and monotone decreasing, respectively. Furthermore,

$$T^{2n+1}\phi \geqq T^{2n}\phi, \qquad n \geqq 0. \tag{8}$$

The monotone convergence theorem coupled with Dini's theorem imply that both subsequences are uniformly convergent on $[0, 1]$, so that ψ_0, ψ_1 exist and, by (8),

$$\psi_0 \leqq \psi_1.$$

It is immediate that $T\psi_0 = \psi_1$ and $T\psi_1 = \psi_0$. By definition of T it is the case that whenever φ is in the domain of T, then, in the case of problem (P_1),

$$(T\varphi)'(x) = -\int_0^x \left(\frac{t}{x}\right)^{n-1} f(t, \varphi(t))\, dt, \qquad 0 < x \leqq 1,$$

$$(T\varphi)'(0) = 0,$$

$$(T\varphi)''(x) = -\left(\frac{n-1}{x}\right)(T\varphi)'(x) - f(x, \varphi(x)), \qquad x \in (0, 1),$$

and for problem (P_2) an easy computation shows that

$$(T\varphi)''(x) = -f(x, \varphi(x)), \qquad x \in (0, 1).$$

To prove that $\psi_0 = \psi_1$ when $yf(x, y)$ is strictly increasing, recall that $\psi_0 \leqq \psi_1$ and, to proceed by contradiction, it is assumed that there exists $x \in (0, 1)$ such that $\psi_0(x) < \psi_1(x)$.

Define $\omega : [0, 1] \to \mathbb{R}$ by

$$\omega(x) = \psi_0(x)\psi_1'(x) - \psi_0'(x)\psi_1(x)$$

and observe that if the problem is (P_1) then $\omega(0) = \omega(1) = 0$; if the problem is (P_2) then it is the case that

$$\alpha\psi_i(0) - \beta\psi_i'(0) = 0, \qquad i = 0, 1,$$

$$\gamma\psi_i(1) + \delta\psi_i'(1) = 0,$$

and therefore, using the fact that $\rho = \gamma\beta + \gamma\alpha + \alpha\delta > 0$, it follows again that $\omega(0) = \omega(1) = 0$.

Also,

$$\omega'(x) = \psi_0'(x)\psi_1'(x) + \psi_0(x)\psi_1''(x) - \psi_0''((x)\psi_1(x) - \psi_0'(x)\psi_1'(x),$$

$$\omega'(x) = \psi_0(x)\psi_1''(x) - \psi_0''(x)\psi_1(x), \qquad x \in (0, 1),$$

and recalling that $\psi_0 = T\psi_1$, $\psi_1 = T\psi_0$, for (P_1):

$$\omega'(x) = \psi_0(x)\left[-\frac{(n-1)}{x}(T\psi_0)'(x) - f(x, \psi_0(x))\right]$$

$$- \psi_1(x)\left[-\frac{(n-1)}{x}(T\psi_1)'(x) - f(x, \psi_1(x))\right]$$

$$= -\frac{(n-1)}{x}\psi_0(x)\psi_1'(x) - \psi_0(x)f(x, \psi_0(x)) + \frac{(n-1)}{x}\psi_0'(x)\psi_1(x) + \psi_1(x)f(x, \psi_1(x))$$

$$= -\frac{(n-1)}{x}[\psi_0(x)\psi_1'(x) - \psi_0'(x)\psi_1(x)] + \psi_1(x)f(x, \psi_1(x)) - \psi_0(x)f(x, \psi_0(x)).$$

Thus, in this case:

$$\omega'(x) + \frac{n-1}{x}\omega(x) = \psi_1(x)f(x, \psi_1(x)) - \psi_0(x)f(x, \psi_0(x)), \qquad x \in (0, 1),$$

and the right-hand side is of one sign on $(0, 1)$. This is impossible since $\omega(0) = \omega(1) = 0$.

For (P_2), the discussion is even simpler since then

$$\omega'(x) = \psi_1(x)f(x, \psi_1(x)) - \psi_0(x)f(x, \psi_0(x)),$$

so ω' is of one sign on $(0, 1)$ and $\omega(0) = \omega(1) = 0$.

It should be pointed out that the convergence of the sequence of iterates to the solution of the integral equation in the case that $f(x, y) = a(x)y^{-p}$ with $p \in (0, 1)$ follows from a result in [12].

3. A general approximation result. This section is devoted to finding a large class of functions f for which it is easy to implement an iterative scheme to approximate the solutions of (P_1) or (P_2). The first step is to show that in many cases these problems have unique solutions; the following result generalizes the uniqueness criteria found in [3] and in [4].

THEOREM 3.1. *If* H_1 *and* H_2 *hold, then* (P_1) *has at most one positive solution. If* H_1', H_2, *and* H_3 *hold, then* (P_2) *has at most one positive solution.*

Proof. The monotonicity of the operator precludes the existence of two ordered solutions. If the problem has at least two distinct solutions ϕ_1, ϕ_2, it must be the case that there exist a, $b \in [0, 1]$, $a < b$, such that one of the solutions, say ϕ_1, is strictly greater than the other solution, ϕ_2, over $[a, b]$. Define $\omega : [0, 1] \to \mathbb{R}$ by

$$\omega(x) = \phi_1(x) - \phi_2(x).$$

It is the case that $\omega(x) > 0$ if $x \in [a, b]$; the claim is that $\omega(x) \geqq 0$ for $x \in [0, 1]$ for, if this were not the case, ω must have a point of minimum in $[0, 1]$, say x_0, with $\omega(x_0) < 0$.

The discussion now must be separated for the two problems. For problem (P_1), since $\omega(1) = 0$, it must be the case that $x_0 \in [0, 1)$; if $x_0 \in (0, 1)$ then $\omega(x_0) < 0$, $\omega'(x_0) = 0$, $\omega''(x_0) \geqq 0$. But

$$\omega''(x_0) = f(x_0, \phi_2(x_0)) - f(x_0, \phi_1(x_0)) < 0,$$

which is a contradiction. Thus the minimum must occur at $x = 0$ so there must exist $\varepsilon > 0$ such that if $x \in [0, \varepsilon)$, then $\phi_1(x) < \phi_2(x)$, and

$$\omega'(x) = \phi_1'(x) - \phi_2'(x) = \int_0^x \left(\frac{t}{x}\right)^{n-1} (f(t, \phi_2(t)) - f(t, \phi_1(t)))\, dt < 0,$$

which contradicts the fact that the minimum of ω occurs at $x = 0$.

For problem (P_2), we reason in the same way to eliminate the possibility of $x_0 \in (0, 1)$. If the minimum occurs at $x = 0$, then ω must be negative in a neighborhood $[0, \delta)$ of 0, and from the boundary conditions it follows that if $\beta = 0$ then $\omega(0) = 0$, so it may be assumed, in this case, that $\beta > 0$, obtaining:

$$\omega'(0) = \phi_1'(0) - \phi_2'(0) = \frac{\alpha}{\beta}(\phi_1(0) - \phi_2(0)) = \frac{\alpha}{\beta}\omega(0)$$

and

$$\omega'(x) - \omega'(0) = \int_0^x \omega''(t)\, dt = \int_0^x (f(t, \phi_2(t)) - f(t, \phi_1(t)))\, dt < 0,$$

so it follows that $\omega'(x) \leqq \omega'(0) = (\alpha/\beta)\omega(0) < 0$ on $[0, \delta)$, contradicting the fact that ω has its minimum at $x = 0$. Finally, if the minimum occurs at $x = 1$, ω must be negative in some neighborhood $(1 - \varepsilon, 1]$ of 1. For $x \in (1 - \varepsilon, 1]$,

$$\omega'(1) - \omega'(x) = \int_x^1 \omega''(t)\, dt = \int_x^1 (f(t, \phi_2(t)) - f(t, \phi_1(t)))\, dt < 0,$$

so that $\omega'(x) > \omega'(1)$ on $(1 - \varepsilon, 1)$. If $\delta = 0$ then the boundary conditions imply that $\omega(1) = 0$, so it must be the case that $\delta \neq 0$. In this situation $\omega'(1) = -(\gamma/\delta)\omega(1) > 0$ and $\omega'(x) > 0$ in $(1 - \varepsilon, 1)$, contradicting the fact that $x = 1$ is a minimum for ω.

For the rest of this section the focus will be the study of how good an approximation to the positive solution of (P_1) or (P_2) is the positive solution of the new problem

$$(P_1)_\varepsilon \qquad \begin{aligned} &y''+\frac{n-1}{x}y'+f(x, y+\varepsilon)=0,\\ &y'(0)=y(1)=0, \end{aligned}$$

or

$$(P_2)_\varepsilon \qquad \begin{aligned} &y''+f(x, y+\varepsilon)=0,\\ &\alpha y(0)-\beta y'(0)=0,\\ &\gamma y(1)+\delta y'(1)=0, \end{aligned}$$

respectively, for $\varepsilon>0$ and small.

Denoting $f_\varepsilon(x, y)=f(x, y+\varepsilon)$ and

$$T_\varepsilon\phi(x)=\int_0^1 G(x, t)f_\varepsilon(t, \phi(t))\,dt,$$

it is clear that if f satisfies H_1 then f_ε is continuous on $[0, 1)\times[0, \infty)$, if f satisfies H_1', then f_ε is continuous on $(0, 1)\times[0, \infty)$, and in both cases T_ε is defined on all of K and is monotone decreasing there if f satisfies H_1, and if f is continuous then f_ε satisfies appropriate integrability conditions (see [5]) for the problems to have solutions. Thus, problems $(P_1)_\varepsilon$ and $(P_2)_\varepsilon$ have a unique positive solution, denoted by ϕ_ε and if $yf_\varepsilon(x, y)$ is (strictly) monotone in y, then by Theorem 2.2 the sequence of iterates $\{T_\varepsilon^k 0\}_{k=1}^\infty$ converges uniformly to ϕ_ε, having in addition that the subsequence of even iterates is monotone decreasing and the subsequence of odd iterates is monotone increasing. It should be noted at this point that if $f(x, y)=a(x)y^{-p}$. $0<p\leqq 1$, and if $a(x)$ is continuous in its required domain, then $yf_\varepsilon(x, y)$ is strictly increasing in y for all $\varepsilon>0$, and therefore the convergence of the iterates of $T_\varepsilon 0$ to ϕ_ε is assured in this particular case.

THEOREM 3.2. *If* $0<\mu<\varepsilon$, *then* $0\leqq\phi_\mu(x)-\phi_\varepsilon(x)<\varepsilon-\mu$, $x\in[0, 1]$.

Proof. Let $\omega(x)=(\phi_\varepsilon(x)+\varepsilon)-(\phi_\mu(x)+\mu)$. It will be shown that $\omega(x)\geqq 0$ on $[0, 1]$.

Consider first problem (P_1). Define

$$h(x)=\frac{f(x, \phi_\mu(x)+\mu)-f(x, \phi_\varepsilon(x)+\varepsilon)}{(\phi_\mu(x)+\varepsilon)-(\phi_\mu(x)+\mu)}$$

when $\omega(x)\neq 0$, and $h(x)=0$ otherwise. Then $h(x)\geqq 0$, $x\in[0, 1]$, and ω satisfies

$$\omega''(x)+\frac{n-1}{x}\omega'(x)-h(x)\omega(x)=0.$$

It follows that ω cannot have an interior negative minimum (it would have to be the case that $x_0\in(0, 1)$, $\omega(x_0)<0$, $\omega'(x_0)=0$, $\omega''(x_0)\geqq 0$, but then $\omega''(x_0)=h(x_0)\omega(x_0)<0$, a contradiction). If $x=0$ is a point of minimum, then ω must be negative in a neighborhood $[0, \varepsilon)$ of 0, and proceeding as in the proof of Theorem 3.1, a contradiction results; the same argument as in the proof of Theorem 3.1 also shows that $x=1$ cannot be the point of minimum for ω. Thus, in the case of problem (P_1) it must be the case that $\omega(x)\geqq 0$ for $x\in[0, 1]$. The argument for (P_2) follows the same line of reasoning as in the proof of Theorem 3.1 to show that $\omega(x)\geqq 0$ for $x\in[0, 1]$.

Thus we have that if $0<\mu<\varepsilon$, then

$$\phi_\mu(x)+\mu \leqq \phi_\varepsilon(x)+\varepsilon, \qquad x\in[0,1],$$

and therefore

$$\phi_\mu(x)=\int_0^1 G(x,t)f(t,\phi_\mu(t)+\mu)\,dt \geqq \int_0^1 G(x,t)f(t,\phi_\varepsilon(t)+\varepsilon)=\phi_\varepsilon(x),$$

concluding that

$$0\leqq \phi_\mu(x)-\phi_\varepsilon(x)\leqq \varepsilon-\mu.$$

THEOREM 3.3.

(I) *If f satisfies* H_1 *and* H_2 *then* $\lim_{\varepsilon\to 0}\phi_\varepsilon=\phi$ *exists uniformly on* $[0,1]$ *and* ϕ *is a positive solution of* (P_1) *belonging to* $C^1([0,1))\cap C^2((0,1))$.

(II) *If f satisfies* H_1', H_2 *and* H_3 *holds, then* $\lim_{\varepsilon\to 0}\phi_\varepsilon=\phi$ *exists uniformly on* $[0,1]$ *and* ϕ *is a positive solution of* (P_2) *in*:

(a) $C([0,1])\cap C^2((0,1))$ *if* $\beta=\delta=0$,
(b) $C^1([0,1))\cap C^2((0,1))$ *if* $\alpha=0$,
(c) $C^1((0,1])\cap C^2((0,1))$ *if* $\gamma=0$,
(d) $C^1([0,1))\cap C^2((0,1))$ *if* $\alpha\neq 0$ *and* $\beta\neq 0$,
(e) $C^1((0,1])\cap C^2((0,1))$ *if* $\gamma\neq 0$ *and* $\delta\neq 0$,
(f) $C^1([0,1])\cap C^2((0,1))$ *if* α, β, γ, δ *are all positive.*

Proof. The existence of $\phi=\lim_{\varepsilon\to 0}\phi_\varepsilon$ uniformly on $[0,1]$ is an immediate consequence of Theorem 2.2, which also implies that if $0<\mu<\varepsilon$ then $\phi_\varepsilon\leqq\phi_\mu$ and $\phi_\mu+\mu<\phi_\varepsilon+\varepsilon$.

(I) For each $\varepsilon>0$, it is the case that

$$\phi_\varepsilon'(x)=-\int_0^x\left[\frac{t}{x}\right]^{n-1} f(t,\phi_\varepsilon(t)+\varepsilon)\,dt, \qquad x\in(0,1],$$

$$\phi_\varepsilon'(0)=0.$$

Let $\eta>0$, $\eta<1$ and consider $\{\phi_\varepsilon\}_{\varepsilon>0}$ on the interval $[0,1-\eta]$. Since $\phi_\varepsilon\leqq\phi_\mu$ if $0<\mu<\varepsilon$, it follows that fixing any $\varepsilon>0$, $\phi(x)\geqq\phi_\varepsilon(x)>0$ on $[0,1-\eta]$ and thus it is immediate that

$$\lim_{\varepsilon\to 0}\phi_\varepsilon'(x)=\phi'(x)$$

exists uniformly on $[0,1-\eta]$. This implies that $\phi\in C^1([0,1))$, $\phi'(0)=0$, and the same line of reasoning implies that $\phi\in C^2((0,1))$ with

$$\phi''(x)=-\frac{n-1}{x}\phi'(x)-f(x,\phi(x)), \qquad x\in(0,1).$$

Since $\phi(1)=0$, (I) follows.

(II) The existence of the uniform limit $\lim_{\varepsilon\to 0}\phi_\varepsilon(x)=\phi(x)$ on $[0,1]$ implies that if $0<\eta<\frac{1}{2}$, then

$$\lim_{\varepsilon\to 0} f(x,\phi_\varepsilon(x)+\varepsilon)=f(x,\phi(x))$$

uniformly on $[\eta,1-\eta]$, and $\phi(x)>0$ for $x\in[\eta,1-\eta]$.

Thus

$$-\int_{\eta}^{1-\eta}\int_{\eta}^{t} f(u,\phi(u))\,du\,dt = \lim_{\varepsilon\to 0}\left(-\int_{\eta}^{1-\eta}\int_{\eta}^{t} f(u,\phi_\varepsilon(u)+\varepsilon)\,du\,dt\right)$$

$$= \lim_{\varepsilon\to 0} -\int_{\eta}^{1-\eta}\int_{\eta}^{t} \phi_\varepsilon''\,du\,dt$$

$$= \lim_{\varepsilon\to 0} -\int_{\eta}^{1-\eta} (\phi_\varepsilon'(t)-\phi_\varepsilon'(\eta))\,dt$$

$$= \lim_{\varepsilon\to 0} [\phi_\varepsilon(1-\eta)-\phi_\varepsilon(\eta)-\phi_\varepsilon'(\eta)(1-2\eta)].$$

Since ϕ_ε converges uniformly on $[0,1]$, this implies that $\lim_{\varepsilon\to 0}\phi_\varepsilon'(\eta)$ exists. This, together with the fact that $\phi_\varepsilon''(x) = -f(x,\phi_\varepsilon(x)+\varepsilon)$, $x\in[0,1]$, and therefore $\lim_{\varepsilon\to 0}\phi_\varepsilon''(x)$ exists uniformly on compact subsets of $(0,1)$, implies that $\{\phi_\varepsilon'\}_{\varepsilon>0}$ converges uniformly on compact subsets of $(0,1)$ to a differentiable function with derivative $-f(x,\phi(x))$ on $[\eta,1-\eta]$. From this it follows that ϕ is twice differentiable on $[\eta,1-\eta]$ and that

$$\phi''(x) = -f(x,\phi(x)), \qquad x\in[\eta,1-\eta].$$

Thus $\phi\in C([0,1])\cap C^2((0,1))$ and is a positive solution of

$$y''+f(x,y)=0.$$

Now the validity of II(a) is obvious.

To prove II(b) it suffices to observe that since

$$\phi_\varepsilon'(x) = \frac{1}{\rho}\left[-\gamma\int_0^x (\beta+\alpha t)f(t,\phi_\varepsilon(t)+\varepsilon)\,dt+\alpha\int_x^1 (\gamma+\delta-\gamma t)f(t,\phi_\varepsilon(t)+\varepsilon)\,dt\right],$$

if $\alpha=0$ then $\phi_\varepsilon'(0)=0$ for all $\varepsilon>0$ and again by restricting attention to intervals of the form $[0,1-\eta]$ where $0<\eta<1$, it becomes apparent that $\{\phi_\varepsilon'\}_{\varepsilon>0}$ converges uniformly in this interval, and this in turn implies that $\phi\in C^1([0,1))\cap C^2((0,1))$ and $\phi'(0)=0$.

The argument for II(c) is entirely similar.

To prove II(d), observe that it cannot be the case, for any $\varepsilon>0$, that $\phi_\varepsilon(0)=0$ since this would imply that $\phi_\varepsilon'(0)=0$ and then, since

$$\phi_\varepsilon'(0) = \alpha\int_0^1 (\gamma+\delta-\gamma t)f(t,\phi_\varepsilon(t)+\varepsilon)\,dt,$$

it would have to be the case that $\gamma=\delta=0$, and this contradicts the assumption that $\rho>0$. Thus, it must be the case that $\phi_\varepsilon(0)>0$ for all $\varepsilon>0$, and hence $\phi(0)>0$, implying that the convergence of $f(x,\phi_\varepsilon(x)+\varepsilon)$ is uniform on compact subsets of $[0,1)$, yielding the desired result.

The proof of II(e) is analogous to that of II(d), and II(f) follows from II(d) and II(e).

Observation. Theorem 2.3 is a generalization of the existence results in [5], [6], and [11]. It also provides a good tool to approximate the solution of (P_1) or (P_2).

4. The case $f(x,y)=a(x)y^{-p}$. This section is devoted to understanding the behavior near $x=1$, that is to say, near the singularity, of the positive solution of (P_1)

in the special case when $f(x, y) = a(x)y^{-p}$, $p > 0$. Thus, the focus is on the positive solution ϕ of

$$y'' + \frac{n-1}{x} y' + a(x)y^{-p} = 0, \tag{9}$$
$$y'(0) = y(1) = 0,$$

with the aim of understanding its behavior near $x = 1$.

For the rest of this section it will be necessary to assume that $a : [0, 1] \to [0, \infty)$ is continuous and that $a(x) > 0$ if $x \in [0, 1)$.

THEOREM 4.1. *Let ϕ be the solution of* (9).

(I) *If $0 < p < 1$, there exist positive real numbers $0 < A < B$ such that*

$$A(1-x) \leqq \phi(x) \leqq B(1-x), \qquad x \in [0, 1].$$

(II) *If $p > 1$, there exist positive real numbers $0 < A < B$ such that*

$$A(1-x) \leqq \phi(x) \leqq B(1-x)^{2/p+1}, \qquad x \in [0, 1].$$

If in addition $a(1) > 0$, then there exist positive real numbers $A < B$ such that

$$A(1-x)^{2/p+1} \leqq \phi(x) \leqq B(1-x)^{2/p+1}, \qquad x \in [0, 1].$$

(III) *If $p = 1$, then given $q > 1$ there exist positive real numbers $A < B$ such that*

$$-Ax \ln x \leqq \phi(x) \leqq B(1-x)^{2/q+1}, \qquad x \in [0, 1].$$

If furthermore, $\int_0^1 a(x)/(1-x)\, dx < \infty$, then there exist positive real numbers $A < B$ such that

$$-Ax \ln x \leqq \phi(x) \leqq -Bx \ln (x), \qquad x \in [\tfrac{1}{2}, 1].$$

Proof. (I) It will first be established that, defining $\psi(x) = 1 - x$ for $x \in [0, 1]$, there exist constants $0 < C < D$ such that

$$C\psi(x) \leqq T\psi(x) \leqq D\psi(x), \qquad x \in [0, 1].$$

To see this, it is necessary to separate the discussion into two cases, namely, $n = 2$ and $n > 2$.

If $n = 2$, the integral operator is defined by

$$T\psi(x) = -\ln (x) \int_0^x ta(t)(\psi(t))^{-p}\, dt - \int_x^1 t \ln (t) a(t)(\psi(t))^{-p}\, dt.$$

The continuity of $a(x)$ implies that ψ belongs to the domain of T, and furthermore, it is clear that if $x \in [0, 1)$,

$$\frac{T\psi(x)}{\psi(x)} = -\frac{\ln (x)}{1-x} \int_0^x ta(t)(1-t)^{-p}\, dt - \frac{1}{1-x} \int_x^1 t \ln (t) a(t)(1-t)^{-p}\, dt,$$

and it is obvious that $T\psi(x)/\psi(x) > 0$ when $x \in [0, 1]$. Furthermore, by l'Hôpital's rule:

$$\lim_{x \to 1} \frac{\int_x^1 t \ln (t) a(t)(1-t)^{-p}\, dt}{1-x} = \lim_{x \to 1} \frac{x \ln (x) a(x)}{(1-x)^p}.$$

But

$$\lim_{x \to 1} \frac{x \ln (x)}{(1-x)^p} = \lim_{x \to 1} \frac{\ln (x) + 1}{p(1-x)^{p-1}} = 0,$$

so that

$$\lim_{x\to 1}\frac{\int_x^1 t\ln(t)a(t)(1-t)^{-p}\,dt}{1-x}=0.$$

Furthermore,

$$\lim_{x\to 1}\frac{\ln(x)}{1-x}=\lim_{x\to 1}-\frac{1}{x}=-1,$$

so, since $\int_0^1 ta(t)(1-t)^{-p}\,dt>0$,

$$\lim_{x\to 1}\frac{T\psi(x)}{\psi(x)}=\int_0^1 ta(t)(1-t)^{-p}\,dt>0,$$

and it follows that $T\psi/\psi$ can be considered as a continuous, strictly positive, function on $[0,1]$, and so there must exist $0<C<D$ such that

$$C\leqq\frac{T\psi(x)}{\psi(x)}\leqq D,$$

from which the assertion follows when $n=2$.

If $n>2$, then the integral operator is:

$$T\psi(x)=\frac{1}{n-2}\Big[(x^{2-n}-1)\int_0^x t^{n-1}a(t)(\psi(t))^{-p}\,dt + \int_x^1 t^{n-1}(t^{2-n}-1)a(t)(\psi(t))^{-p}\,dt\Big],$$

and again ψ is in the domain of T, with $T\psi(x)>0$ if $x\in[0,1)$. But

$$\begin{aligned}\frac{T\psi(x)}{\psi(x)}&=\frac{1}{n-2}\Big[\frac{(x^{2-n}-1)}{1-x}\int_0^x t^{n-1}a(t)(1-t)^{-p}\,dt - \frac{1}{1-x}\int_x^1 t^{n-1}(t^{2-n}-1)a(t)(1-t)^{-p}\,dt\Big]\\ &=\frac{1}{n-2}\Big[\frac{(1-x^{n-2})}{x^{n-2}(1-x)}\int_0^x t^{n-1}a(t)(1-t)^{-p}\,dt + \frac{1}{1-x}\int_x^1 t(1-t^{n-2})a(t)(1-t)^{-p}\,dt\Big]\\ &=\frac{1}{n-2}\Big[\frac{1+x+\cdots+x^{n-3}}{x^{n-2}}\int_0^x t^{n-1}a(t)(1-t)^{-p}\,dt + \frac{1}{1-x}\int_x^t t(1-t^{n-2})a(t)(1-t)^{-p}\,dt\Big],\end{aligned}$$

and, again by l'Hôpital's rule:

$$\lim_{x\to 1}\frac{T\psi(x)}{\psi(x)}=\int_0^x t^{n-1}a(t)(1-t)^{-p}\,dt<\infty,$$

and the existence of the positive numbers C and D is established in the same way as before.

Thus, in either case,

$$C\psi(x) \leqq T\psi(x) \leqq D\psi(x), \qquad x \in [0, 1].$$

For $\lambda > 0$, observe that

$$T(\lambda\psi) = \lambda^{-p} T\psi,$$

and setting $\psi_0 = \lambda\psi$, it follows that if $D/\lambda^{p+1} \leqq 1$ then $T\psi_0 \leqq \psi_0$. Furthermore,

$$T\psi_0 = \lambda^{-p} T\psi \geqq C\lambda^{-p}\psi,$$

and it follows that

$$T^2\psi_0 \leqq C^{-p}\lambda^{(p)^2} T\psi \leqq C^{-p}\lambda^{(p)^2} D\psi,$$

so that

$$T^2\psi_0 \leqq C^{-p}\lambda^{(p^2-1)} D\psi_0.$$

Thus, choosing λ large enough so as to have

$$\frac{D}{\lambda^{p+1}} \leqq 1 \quad \text{and} \quad C^{-p} D\lambda^{(p^2-1)} \leqq 1,$$

which is possible since $0 < p < 1$, it becomes the case that

$$T\psi_0 \leqq \psi_0 \quad \text{and} \quad T^2\psi_0 \leqq \psi_0.$$

This implies that the sequence of even iterates is monotone increasing and so, by Theorem 1.2, they converge to the solution ϕ of (9). This yields (I).

(II) The first thing to be shown is the existence of $B > 0$ such that

$$\phi(x) \leqq B(1-x)^{2/p+1}, \qquad x \in [0, 1].$$

To do this, for $\varepsilon > 0$ let ϕ_ε be the solution to

$$y'' + \frac{n-1}{x} y' + a(x)(y+\varepsilon)^{-p} = 0,$$

$$y'(0) = y(1) = 0,$$

and for $\lambda > 1$ define $\chi(x) = \lambda\phi(0)(1-x)^{2/(p+1)}$. Recall that $\phi(0) > \phi_\varepsilon(0)$ by Theorem 2.2, so that $\chi(0) > \phi_\varepsilon(0)$.

Define $\omega(x) = \chi(x) - \phi_\varepsilon(x)$ and observe that $\omega(0) > 0$, $\omega(1) = 0$. It will be shown, by contradiction, that $\omega(x) \geqq 0$ for $x \in [0, 1]$. If this were not the case, there would exist $x_0 \in (0, 1)$ such that $\omega(x_0) < 0$, $\omega'(x_0) = 0$, and $\omega'(x_0) \geqq 0$. But

$$\chi'(x) = -\frac{2}{p+1}\lambda\phi(0)(1-x)^{(1-p)/(p+1)},$$

$$\chi''(x) = \frac{2(1-p)}{(p+1)^2}\lambda\phi(0)(1-x)^{-(2p/(p+1))},$$

and the fact that $\omega'(x_0) = 0$ implies that

$$\chi'(x_0) = \phi'_\varepsilon(x_0),$$

so

$$\phi'_\varepsilon(x_0) = -\frac{2}{p+1}\lambda\phi(0)(1-x_0)^{(1-p)/(p+1)},$$

and this implies that

$$\phi_\varepsilon''(x_0) = -\frac{n-1}{x_0}\left(-\frac{2}{p+1}\lambda\phi(0)(1-x_0)^{(1-p)/(p+1)}\right) - \frac{a(x_0)}{(\phi_\varepsilon(x_0)+\varepsilon)^p}.$$

Thus

$$\begin{aligned}\omega''(x_0) &= -\frac{2(p-1)\lambda\phi(0)}{(p+1)^2}(1-x_0)^{-(2p/(p+1))} \\ &\quad -\frac{2(n-1)}{(p+1)x_0}\lambda\phi(0)(1-x_0)^{(1-p)/(1+p)} + \frac{a(x_0)}{(\phi_\varepsilon(x_0)+\varepsilon)^p} \\ &= -\frac{2\lambda\phi(0)(1-x_0)^{-(2p/(p+1))}}{p+1}\left[\frac{p-1}{p+1} + \frac{(n-1)}{x_0}(1-x_0)\right] + \frac{a(x_0)}{(\phi_\varepsilon(x_0)+\varepsilon)^p} \\ &\leqq -\frac{2(p-1)}{(p+1)^2}\lambda\phi(0)(1-x_0)^{-(2p/(p+1))} + \frac{\|a\|}{(\phi_\varepsilon(x_0))^p},\end{aligned}$$

where $\|a\| = \sup_{x\in[0,1]} a(x)$.

Since $0 < \chi(x_0) < \phi_\varepsilon(x_0)$, it follows that

$$\begin{aligned}\omega''(x_0) &\leqq -\frac{2(p-1)}{(p+1)^2}\lambda\phi(0)(1-x_0)^{-(2p/(p+1))} + \frac{\|a\|}{\lambda^p(\phi(0))^p(1-x_0)^{(2p/(p+1))}} \\ &= (1-x_0)^{-(2p/(p+1))}\left[\frac{\|a\|}{\lambda^p(\phi(0))^p} - \frac{2(p-1)}{(p+1)^2}\phi(0)\lambda\right].\end{aligned}$$

Hence, if λ is chosen so that

$$\frac{\|a\|}{\lambda^p(\phi(0))^p} < \frac{2(p-1)}{(p+1)^2}\phi(0)\lambda,$$

i.e., so that

$$\frac{\|a\|(p+1)^2}{2(p-1)(\phi(0))^{p+1}} < \lambda^{p+1},$$

it follows that $\omega''(x_0) < 0$, which is a contradiction. Since the selection of λ does not depend on ε, we have shown that for B sufficiently large

$$\phi_\varepsilon(x) \leqq B(1-x)^{2/p+1}, \qquad x\in[0,1], \quad \varepsilon > 0$$

and since $\phi(x) = \lim_{\varepsilon\to 0}\phi_\varepsilon(x)$, $x\in[0,1]$, the inequality

$$\phi(x) \leqq B(1-x)^{2/p+1}, \qquad x\in[0,1],$$

has been shown.

To complete the proof of the first part of (II), it suffices to observe that if $\varepsilon > 0$, then if $n > 2$ and $x\in[0,1)$,

$$\begin{aligned}\frac{\phi_\varepsilon(x)}{1-x} &= \frac{1}{n-2}\left[\frac{1-x^{n-2}}{x^{n-2}(1-x)}\int_0^x t^{n-1}a(t)(\phi_\varepsilon(t)+\varepsilon)^{-p}\,dt\right. \\ &\quad \left. + \frac{1}{1-x}\int_x^1 t(1-t^{n-2})a(t))(\phi_\varepsilon(t)+\varepsilon)^{-p}\,dt\right]\end{aligned}$$

and so

$$\frac{\phi_\varepsilon(x)}{1-x}=\frac{1}{n-2}\left[\frac{1+x+\cdots+x^{n-3}}{x^{n-2}}\int_0^x t^{n-1}a(t)(\phi_\varepsilon(t)+\varepsilon)^{-p}\,dt\right.$$
$$\left.+\frac{1}{1-x}\int_x^1 t(1-t^{n-2})a(t)(\phi_\varepsilon(t)+\varepsilon)^{-p}\,dt\right],$$

and hence

$$\lim_{x\to 1}\frac{\phi_\varepsilon(x)}{1-x}=\int_0^1 t^{n-1}a(t)(\phi_\varepsilon(t)+\varepsilon)^{-p}\,dt>0,$$

and thus, fixing $\varepsilon>0$, there exists $A>0$ such that

$$\frac{\phi_\varepsilon(x)}{1-x}\geqq A, \qquad x\in[0,1].$$

Thus, $\phi_\varepsilon(x)\geqq A(1-x)$, $x\in[0,1]$, and since $\phi_\varepsilon(x)\leqq\phi(x)$, $x\in[0,1]$, we have that

$$A(1-x)\leqq\phi(x)\leqq B(1-x)^{2/p+1}, \qquad x\in[0,1],$$

if $n>2$.

If $n=2$ and $x\in(0,1)$,

$$\frac{\phi_\varepsilon(x)}{1-x}=-\frac{\ln(x)}{1-x}\int_0^x ta(t)(\phi_\varepsilon(t)+\varepsilon)^{-p}\,dt-\frac{1}{1-x}\int_x^1 t\ln(t)a(t)(\phi_\varepsilon(t)+\varepsilon)^{-p}\,dt$$

and

$$\lim_{x\to 1}\frac{\phi_\varepsilon(x)}{1-x}=\int_0^1 ta(t)(\phi_\varepsilon(t)+\varepsilon)^{-p}\,dt>0,$$

and the same argument given above applies.

If it is known in addition that $a(1)>0$, then the proof of the existence of A such that

$$A(1-x)^{2/p+1}\leqq\phi(x), \qquad x\in[0,1]$$

goes as follows.

Define $\psi(x)=(1-x)^{2/p+1}$. The fact that there exists $B>0$ such that

$$\phi(x)\leqq B\psi(x), \qquad x\in[0,1],$$

implies that ψ is in the domain of the integral operator T, and that

$$B^{-p}T\psi(x)\leqq\phi(x), \qquad x\in[0,1].$$

Now, for $x\in(0,1)$, if $n>2$,

$$\frac{T\psi(x)}{\psi(x)}=\frac{1}{n-2}\left[\frac{1-x^{n-2}}{x^{n-2}(1-x)^{2/p+1}}\int_0^x t^{n-1}a(t)(1-t)^{-(2p/(p+1))}\,dt\right.$$
$$\left.+\frac{1}{(1-x)^{2/p+1}}\int_x^1 t(1-t^{n-2})a(t)(1-t)^{-(2p/(p+1))}\,dt\right],$$

so that

$$\frac{T\psi(x)}{\psi(x)}=\frac{1}{n-2}\left[\frac{(1-x)^{(p-1)/(p+1)}(1+x+\cdots+x^{n-3})}{x^{n-2}}\int_0^x t^{n-1}a(t)(1-t)^{-(2p/(p+1))}\,dt\right.$$
$$\left.+\frac{1}{(1-x)^{2/p+1}}\int_x^1 t(1-t)^{(p-1)/(p+1)}(1+t+\cdots+t^{n-3})a(t)\,dt\right].$$

Now observe that, by l'Hôpital's rule:

$$\lim_{x\to 1}\frac{\int_x^1 t(1-t)^{(1-p)/(1+p)}(1+t+\cdots+t^{n-3})a(t)\,dt}{(1-x)^{2/p+1}}$$

$$=\lim_{x\to 1}\frac{x(1-x)^{(p-1)/(p+1)}(1+\cdots+x^{n-3})a(x)}{(2/p+1)(1-x)^{(1-p)/(1+p)}}$$

$$=\lim_{x\to 1}\frac{(p+1)xa(x)(1+\cdots+x^{n-3})}{2}$$

$$=\frac{(n-2)(p+1)a(1)}{2},$$

and that if $\int_0^1 t^{n-1}a(t)(1-t)^{-(2p/(p+1))}\,dt<\infty$, then

$$\lim_{x\to 1}\frac{(1-x)^{(p-1)/(p+1)}(1+\cdots+x^{n-3})}{x^{n-2}}\int_0^x t^{n-1}a(t)(1-t)^{-(2p/(p+1))}\,dt=0,$$

and if the integral diverges, then by using l'Hôpital's rule again, it follows that

$$\lim_{x\to 1}\frac{\int_0^x t^{n-1}a(t)(1-t)^{-(2p/(1+p))}\,dt}{(x-1)^{(1-p)/(p+1)}}=\lim_{x\to 1}\frac{x^{n-1}a(x)(1-x)^{-(2p/(p+1))}}{-((1-p)/(p+1))(1-x)^{-(2p/(p+1))}}=\frac{a(1)}{p-1}.$$

Thus, if $a(1)>0$, it is the case that $\lim_{x\to 1}(T\psi(x)/\psi(x))>0$, and it follows that there exists $D>0$ for which

$$\frac{T\psi(x)}{\psi(x)}\geqq D,$$

so that $T\psi(x)\geqq D\psi(x)$.

Combining this with the fact that $B^{-p}T\psi(x)\leqq\phi(x)$, $x\in[0,1]$, the result follows if $n>2$.

If $n=2$, then

$$\frac{T\psi(x)}{\psi(x)}=-\frac{\ln(x)}{(1-x)^{2/p+1}}\int_0^x ta(t)(1-t)^{-(2p/(p+1))}\,dt$$

$$-\frac{1}{(1-x)^{2/p+1}}\int_x^1 t\ln(t)a(t)(1-t)^{-(2p/(p+1))}\,dt$$

and hence, since

$$\lim_{x\to 1}\frac{\int_x^1 t\ln(t)a(t)(1-t)^{-(2p/(1+p))}\,dt}{(1-x)^{2/p+1}}=\lim_{x\to 1}\frac{-x\ln(x)a(x)(1-x)^{-(2p/(p+1))}}{-(2/p+1)(1-x)^{-(1-p)/(p+1)}}$$

$$=\lim_{x\to 1}\frac{x\ln(x)a(x)}{(2/p+1)(1-x)}$$

and since $\lim_{x\to 1}(\ln(x)/(1-x))=-1$, we get that

$$\lim_{x\to 1}\frac{-\int_x^1 t\ln(t)a(t)(1-t)^{-(2p/(1+p))}\,dt}{(1-x)^{2/p+1}}=\frac{p+1}{2}a(1).$$

Similar computations show that

$$\lim_{x\to 1}\left[-\frac{\ln(x)}{(1-x)^{2/p+1}}\int_0^x ta(t)(1-t)^{-(2p/(p+1))}\,dt\right)=0$$

or

$$\lim_{x\to 1}\left(-\frac{\ln(x)}{(1-x)^{2/p+1}}\int_0^x ta(t)(1-t)^{-(2p/(p+1))}\,dt\right)=\frac{a(1)}{1-p},$$

and in either case the result now follows in the same way as for $n>2$.

Proof of (III). Let $\lambda>1$, ϕ_ε the solution

$$(9)_\varepsilon \qquad y''+\frac{n-1}{x}y'+a(x)(y+\varepsilon)^{-1}=0,$$
$$y'(0)=y(1)=0,$$

and define $\chi:[0,1]\to[0,\infty)$, $\omega:[0,1]\to\mathbb{R}$, by

$$\chi(x)=\lambda\phi(0)(1-x)^{2/q+1}, \qquad x\in[0,1]$$

and

$$\omega(x)=\chi(x)-\phi_\varepsilon(x), \qquad x\in[0,1].$$

Since $\phi(0)\geqq\phi_\varepsilon(0)>0$, it follows that $\omega(x_0)>0$, and it is obvious that $\omega(1)=0$. It will be proved, by contradiction, that $\omega(x)\geqq 0$ for all $x\in[0,1]$. If this were not true, there would exist $x_0\in(0,1)$ such that

$$\omega(x_0)<0, \quad \omega'(x_0)=0, \quad \omega''(x_0)\geqq 0.$$

Now

$$\chi'(x)=-\frac{2\lambda\phi(0)}{q+1}(1-x)^{(1-q)/(1+q)}$$

and

$$\chi''(x)=-\frac{2\lambda(q-1)}{(q+1)^2}\phi(0)(1-x)^{-(2q/(1+q))},$$

and therefore

$$\phi_\varepsilon'(x_0)=-\frac{2\lambda}{q+1}\phi(0)(1-x_0)^{(1-q)/(1+q)},$$
$$\phi_\varepsilon''(x_0)=-\frac{2(n-1)\lambda\phi(0)}{x_0(q+1)}(1-x_0)^{(1-q)/(1+q)}-\frac{a(x_0)}{\phi_\varepsilon(x_0)+\varepsilon},$$

so that

$$\begin{aligned}\omega''(x_0)&=-\frac{2\lambda(q-1)\phi(0)}{(q+1)^2}(1+x_0)^{-(2q/(q+1))}\\&\quad-\frac{2\lambda\phi(0)(n-1)}{(q+1)x_0}(1-x_0)^{(1-q)/(1+q)}+\frac{a(x_0)}{\phi_\varepsilon(x_0)+\varepsilon}\\&\leqq-\frac{2\lambda\phi(0)}{q+1}(1-x_0)^{-(2q/(q+1))}\left[\frac{q-1}{q+1}+\frac{n-1}{x_0}(1-x_0)\right]+\frac{\|a\|}{\phi_\varepsilon(x_0)}\\&\leqq-\frac{2\lambda\phi(0)}{q+1}(1-x_0)^{-(2q/(q+1))}\cdot\frac{q-1}{q+1}+\frac{\|a\|}{\lambda\phi(0)(1-x_0)^{2/q+1}}\\&=(1-x_0)^{-(2q/(q+1))}\left[\frac{\|a\|}{\lambda\phi(0)}-\frac{2\lambda\phi(0)(q-1)}{(q+1)^2}(1-x_0)^{(2(1-q)/(1+q))}\right]\\&\leqq(1-x_0)^{-(2q/(q+1))}\left[\frac{\|a\|}{\lambda\phi(0)}-\frac{2\lambda(q-1)\phi(0)}{(q+1)^2}\right],\end{aligned}$$

and therefore, if λ is chosen so that

$$\frac{\|a\|}{\lambda\phi(0)}-\frac{2\lambda(q-1)\phi(0)}{(q+1)^2}\leqq 0,$$

i.e.,

$$\lambda^2\geqq\frac{(q+1)^2\|a\|}{2(q-1)(\phi(0))^2},$$

then $\omega''(x_0)<0$, and this is a contradiction. Since the selection of λ does not depend on ε, it follows that

$$\phi_\varepsilon(x)\leqq B(1-x)^{(2/(q+1))},\qquad x\in[0,1],\quad \varepsilon>0,$$

and this implies that

$$\phi(x)\leqq B(1-x)^{(2/(q+1))},\qquad x\in[0,1].$$

To prove the other side of the inequality, recall that $\phi_\varepsilon\leqq\phi$, $\varepsilon>0$. Pick $\varepsilon>0$ and observe that if $n>2$ and $x\in[\frac{1}{2},1)$, then

$$\frac{\phi_\varepsilon(x)}{-x\ln(x)}=\frac{1}{n-2}\left[-\frac{1-x^{n-2}}{x^{n-2}x\ln(x)}\int_0^x t^{n-1}\frac{a(t)}{\phi_\varepsilon(t)+\varepsilon}\,dt\right.$$
$$\left.-\frac{1}{x\ln x}\int_x^1 t(1-t^{n-2})\frac{a(t)}{\phi_\varepsilon(t)+\varepsilon}\,dt\right]$$

and

$$\lim_{x\to 1}\frac{\phi_\varepsilon(x)}{-x\ln(x)}=\int_0^1 t^{n-1}\frac{a(t)}{\phi_\varepsilon(t)+\varepsilon}\,dt>0.$$

Thus, there must exist $A>0$ such that $\phi_\varepsilon(x)\geqq -Ax\ln(x)$, $x\in[\frac{1}{2},1]$ and therefore

$$\phi(x)\geqq -Ax\ln(x),\qquad x\in[\tfrac{1}{2},1].$$

If $n=2$, and $x\in[\frac{1}{2},1)$,

$$\frac{\phi_\varepsilon(x)}{-x\ln(x)}=\frac{1}{x}\int_0^x\frac{ta(t)}{\phi_\varepsilon(t)+\varepsilon}\,dt+\frac{1}{x\ln(x)}\int_x^1\frac{t\ln(t)a(t)}{\phi_\varepsilon(t)+\varepsilon}\,dt,$$

and the same argument applies.

For the proof of the last statement of (III), assume that

$$\int_0^1\frac{a(x)}{1-x}<\infty.$$

Under this hypothesis it has been shown in [5] that the unique solution ϕ of (9) is such that there exists $\theta>0$ for which

$$\theta(1-x)\leqq\phi(x),\qquad x\in[0,1].$$

Define $\psi:[0,1]\to[0,\infty)$ by

$$\psi(x)=\tfrac{1}{2}\ln(2),\qquad x\in[0,\tfrac{1}{2}],$$
$$\psi(x)=-x\ln(x),\qquad x\in[\tfrac{1}{2},1].$$

Consider first the case $n=2$. If $x\in[\frac{1}{2},1)$,

$$\frac{\phi(x)}{\psi(x)}=\frac{1}{x}\int_0^x\frac{ta(t)}{\phi(t)}\,dt+\frac{1}{x\ln(x)}\int_x^1\frac{t\ln(t)a(t)}{\phi(t)}\,dt$$

and

$$\lim_{x\to 1}\frac{\int_x^1 (t\ln(t)a(t)/\phi(t))\,dt}{x\ln(x)}=\lim_{x\to 1}\frac{(-x\ln(x)a(x)/\phi(x))}{1+\ln(x)}.$$

But

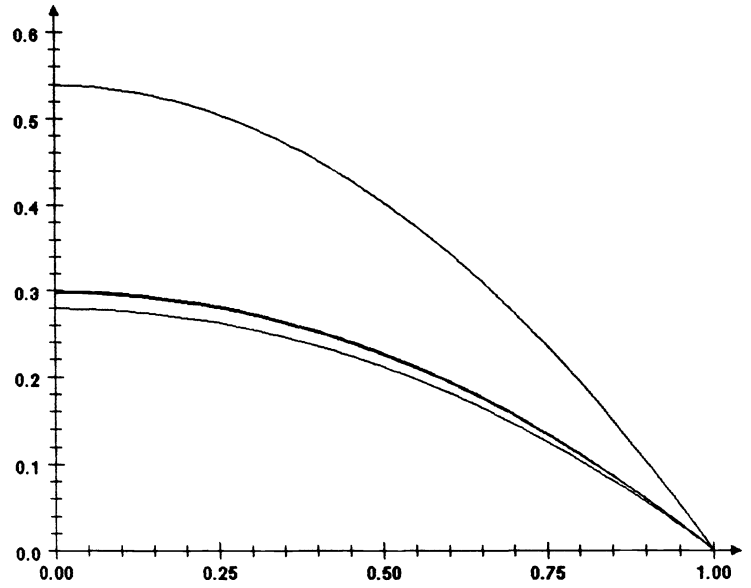
$$\lim_{x\to 1}\frac{-\ln(x)}{\phi(x)}=\lim_{x\to 1}\frac{1/x}{\int_0^x (t/x)(a(t)/\phi(t))\,dt}$$

$$=\frac{1}{\int_0^1 (ta(t)/\phi(t))\,dt},$$

and the conclusion is that

$$\lim\frac{\phi(x)}{\psi(x)}=\int_0^1\frac{ta(t)}{\phi(t)}\,dt+\frac{a(1)}{\int_0^1 (ta(t)/\phi(t))\,dt},$$

and hence the function ϕ/ψ can be defined to be continuous and positive on $[0, 1]$,



X	0.0000	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000	1.0000
FI 0	0.5386	0.5332	0.5171	0.4901	0.4524	0.4040	0.3447	0.2747	0.1939	0.1023	0.0000
FI 1	0.2788	0.2761	0.2681	0.2546	0.2358	0.2114	0.1813	0.1455	0.1037	0.0555	0.0000
FI 2	0.2994	0.2966	0.2879	0.2735	0.2532	0.2269	0.1947	0.1562	0.1113	0.0595	0.0000
FI 3	0.2971	0.2942	0.2857	0.2713	0.2512	0.2252	0.1931	0.1550	0.1104	0.0591	0.0000
FI 4	0.2974	0.2945	0.2859	0.2716	0.2514	0.2254	0.1933	0.1551	0.1105	0.0591	0.0000
FI 5	0.2973	0.2945	0.2859	0.2715	0.2514	0.2253	0.1933	0.1551	0.1105	0.0591	0.0000
FI 6	0.2973	0.2945	0.2859	0.2715	0.2514	0.2253	0.1933	0.1551	0.1105	0.0591	0.0000

FIG. 1

so there exist positive real numbers $A < B$ such that

$$A \leqq \frac{\phi(x)}{\psi(x)} \leqq B, \qquad x \in [\tfrac{1}{2}, 1],$$

and therefore $-Ax \ln x \leqq \phi(x) \leqq -Bx \ln x$, $x \in [\frac{1}{2}, 1]$.

If $n > 2$, the argument is entirely similar, the change being in the Green's function, but the fact that

$$\lim \frac{1-x}{-x \ln (x)} = 1$$

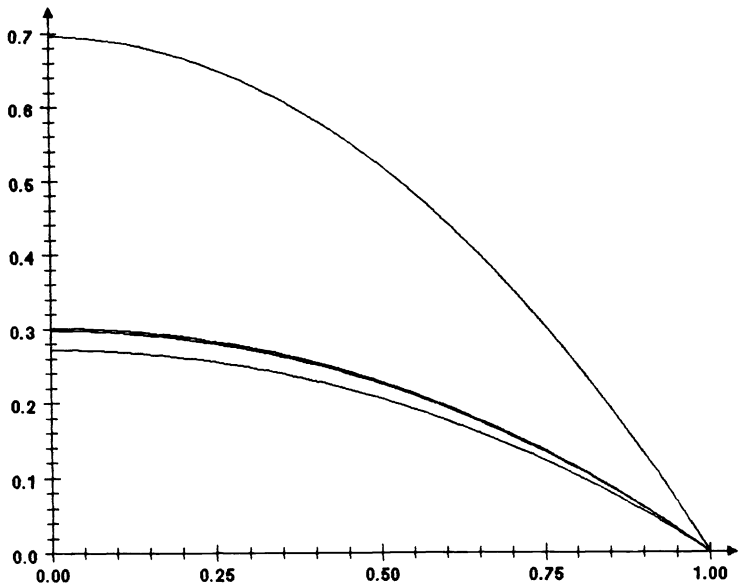
is all that is needed to carry out the required computations.

Observation. Similar results for the case $p > 1$, but with a somewhat more restrictive assumption on the function $a(x)$, have been recently obtained in [7].

5. Numerical examples. Four figures will be presented that correspond to the graphs of several iterates of $T_\varepsilon(0)$, where T_ε is the operator arising from

$$y'' + \frac{1}{x} y' + (y+\varepsilon)^{-p} = 0,$$

$$y'(0) = y(1) = 0.$$



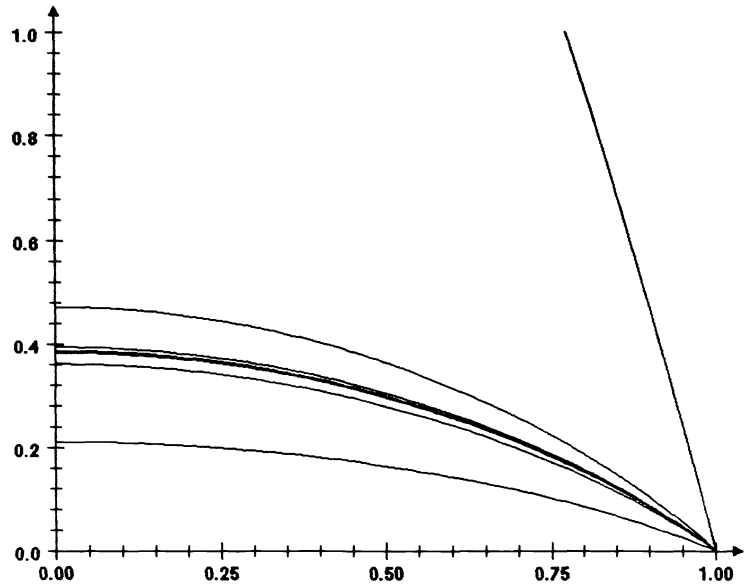
X	0.0000	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000	1.0000
FI 0	0.6956	0.6887	0.6678	0.6330	0.5843	0.5217	0.4452	0.3548	0.2504	0.1322	0.0000
FI 1	0.2711	0.2685	0.2606	0.2476	0.2292	0.2055	0.1763	0.1415	0.1009	0.0540	0.0000
FI 2	0.3005	0.2977	0.2890	0.2745	0.2541	0.2278	0.1954	0.1568	0.1117	0.0598	0.0000
FI 3	0.2971	0.2943	0.2857	0.2714	0.2512	0.2252	0.1932	0.1550	0.1105	0.0591	0.0000
FI 4	0.2975	0.2947	0.2861	0.2717	0.2515	0.2255	0.1934	0.1552	0.1106	0.0592	0.0000
FI 5	0.2975	0.2946	0.2860	0.2717	0.2515	0.2255	0.1934	0.1552	0.1106	0.0592	0.0000
FI 6	0.2975	0.2946	0.2860	0.2717	0.2515	0.2255	0.1934	0.1552	0.1106	0.0592	0.0000

FIG. 2

These iterations were computed using a parallel program running on an Encore Multimax computer with 10 processors. Although parallel programs are not essential for these computations, we have larger problems in mind for the future.

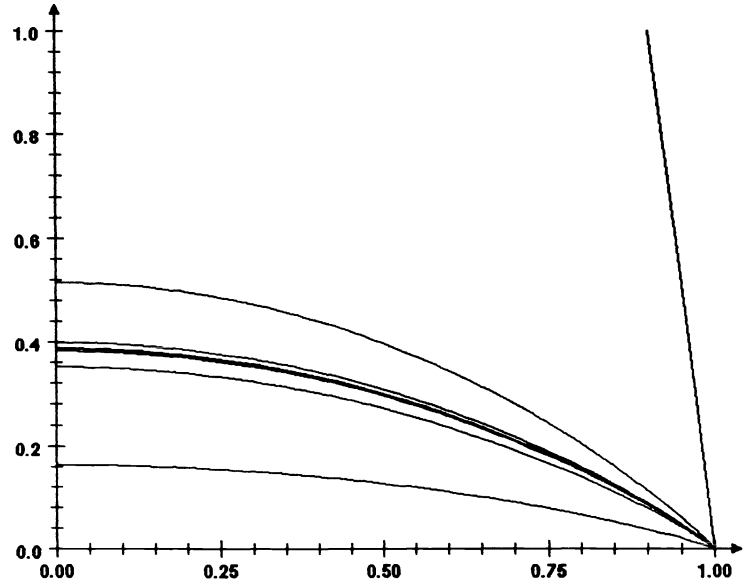
The program uses Simpson's rule with 100 points in the interval [0, 1], and it is very easy to add more points to the subdivision. Two arrrays are used (both are shared memory), one of them to store the past iteration and one to store the current iteration. Each processor works one point at a time, computing the integral corresponding to that point. The shared memory allows all processors to access the values of the previous iteration.

Figure 1 corresponds to $p = 1/9$, $\varepsilon = 10^{-3}$, and the table that follows corresponds to the values of $T^i(0)$ evaluated at the indicated points of [0, 1], $i = 1, \cdots, 7$, where, for convenience, $T^i(0)$ has been denoted as Fl $(i-1)$. Figure 2 corresponds to $p = 1/9$,



X	0.0000	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000	1.0000
Fl 0	2.5000	2.4750	2.4000	2.2750	2.1000	1.8750	1.6000	1.2750	0.9000	0.4750	0.0000
Fl 1	0.2096	0.2078	0.2023	0.1929	0.1798	0.1626	0.1411	0.1150	0.0838	0.0463	0.0000
Fl 2	0.4709	0.4667	0.4540	0.4328	0.4028	0.3638	0.3152	0.2563	0.1860	0.1024	0.0000
Fl 3	0.3607	0.3575	0.3478	0.3316	0.3087	0.2788	0.2416	0.1965	0.1427	0.0786	0.0000
Fl 4	0.3939	0.3904	0.3799	0.3622	0.3371	0.3045	0.2638	0.2146	0.1559	0.0858	0.0000
Fl 5	0.3826	0.3792	0.3689	0.3517	0.3274	0.2957	0.2563	0.2084	0.1514	0.0834	0.0000
Fl 6	0.3863	0.3829	0.3725	0.3551	0.3306	0.2986	0.2587	0.2105	0.1528	0.0842	0.0000
Fl 7	0.3851	0.3817	0.3713	0.3540	0.3295	0.2976	0.2579	0.2098	0.1524	0.0839	0.0000
Fl 8	0.3855	0.3821	0.3717	0.3544	0.3299	0.2979	0.2582	0.2100	0.1525	0.0840	0.0000
Fl 9	0.3854	0.3819	0.3716	0.3543	0.3298	0.2978	0.2581	0.2099	0.1525	0.0840	0.0000
Fl 10	0.3854	0.3820	0.3716	0.3543	0.3298	0.2979	0.2581	0.2100	0.1525	0.0840	0.0000

FIG. 3



X	0.0000	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000	1.0000
Fl 0	5.3861	5.3322	5.1706	4.9013	4.5243	4.0396	3.4471	2.7469	1.9390	1.0234	0.0000
Fl 1	0.1624	0.1609	0.1566	0.1494	0.1392	0.1259	0.1093	0.0891	0.0649	0.0359	0.0000
Fl 2	0.5140	0.5094	0.4956	0.4725	0.4398	0.3972	0.3441	0.2799	0.2033	0.1120	0.0000
Fl 3	0.3506	0.3475	0.3381	0.3224	0.3001	0.2711	0.2349	0.1911	0.1388	0.0765	0.0000
Fl 4	0.3982	0.3947	0.3840	0.3661	0.3408	0.3078	0.2668	0.2170	0.1576	0.0869	0.0000
Fl 5	0.3817	0.3783	0.3681	0.3509	0.3267	0.2951	0.2557	0.2080	0.1511	0.0833	0.0000
Fl 6	0.3871	0.3837	0.3733	0.3559	0.3313	0.2992	0.2593	0.2110	0.1532	0.0845	0.0000
Fl 7	0.3853	0.3819	0.3716	0.3542	0.3297	0.2978	0.2581	0.2100	0.1525	0.0841	0.0000
Fl 8	0.3859	0.3825	0.3721	0.3548	0.3303	0.2983	0.2585	0.2103	0.1528	0.0842	0.0000
Fl 9	0.3857	0.3823	0.3719	0.3546	0.3301	0.2982	0.2584	0.2102	0.1527	0.0841	0.0000
Fl 10	0.3858	0.3823	0.3720	0.3547	0.3302	0.2982	0.2584	0.2102	0.1527	0.0842	0.0000

FIG. 4

$\varepsilon = 10^{-4}$ and the same number of iterates, while Figs. 3 and 4 correspond to $p = \frac{1}{3}$ and $\varepsilon = 10^{-3}$, $\varepsilon = 10^{-4}$, respectively, and the number of iterates shown is 11.

Acknowledgment. The authors thank Juan C. Gatica for writing the program that computes and graphs the iterates found in the figures.

REFERENCES

[1] J. V. BAXLEY, *A singular nonlinear boundary values problem: membrane response of a spherical cap*, SIAM J. Appl. Math., 48 (1988), pp. 497-505.

[2] M. G. CRANDALL, P. H. RABINOWITZ, AND L. TARTAR, *On a Dirichlet problem with a singular nonlinearity*, Comm. Partial Differential Equations, 2 (1977), pp. 193-222.

[3] A. CALLEGARY AND A. NACHMAN, *Some singular, nonlinear differential equations arising in boundary layer theory*, J. Math. Anal. Appl., 64 (1978), pp. 96-105.

[4] J. K. DIAZ, J. M. MOREL, AND L. OSWALD, *An elliptic equation with singular nonlinearity*, Comm. Partial Differential Equations, 12 (1987), pp. 1333–1344.
[5] J. A. GATICA, G. E. HERNANDEZ, AND P. WALTMAN, *Radially symmetric solutions of a class of singular elliptic equations*, Proc. Edinburgh Math. Soc., 33 (1990), pp. 169–180.
[6] J. A. GATICA, V. OLIKER, AND P. WALTMAN, *Singular nonlinear boundary value problems for second order differential equations*, J. Differential Equations, 79 (1989), pp. 62–78.
[7] A. C. LAZER AND P. J. MCKENNA, *On a singular nonlinear elliptic boundary value problem*, preprint.
[8] W. L. PERRY AND C. D. LUNING, *Positive solutions of negative exponent generalized Emden–Fowler boundary value problems*, SIAM J. Math. Anal., 12 (1981), pp. 874–879.
[9] A. NACHMAN AND A. CALLEGARY, *A nonlinear singular boundary value problem in the theory of pseudo plastic fluids*, SIAM J. Appl. Math., 38 (1980), pp. 275–281.
[10] A. NACHMAN AND S. TALIAFERRO, *Mass transfer in boundary layers for power law fluids*, Proc. Roy. Soc. London Ser. A, 365 (1979), pp. 313–326.
[11] S. D. TALIAFERRO, *A nonlinear singular boundary value problem*, Nonlinear Anal. TMA, 3 (1979), pp. 897–904.
[12] H. THIEME, *On a class of Hammerstein integral equations*, Manuscripta Math., 29 (1979), pp. 49–89.
[13] J. S. W. WONG, *On the generalized Emden–Fowler equation*, SIAM Rev., 17 (1975), pp. 339–360.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 463-479, March 1991

011

SOME SINGULAR NONLINEAR BOUNDARY VALUE PROBLEMS*

JOHN V. BAXLEY†

Abstract. Two-point boundary value problems associated with the (possibly) singular nonlinear ordinary differential equation $y''+g(x, y')+f(x, y)=0$, $a \leqq x \leqq b$, are considered. The goal is to obtain rather general existence and uniqueness theorems for positive solutions. In the case of general separated linear boundary conditions, the results allow $f(x, y)$ to be singular as $y \to 0^+$ and at the endpoints, with significant nonlinearity in both f and g. For the special condition $y'(a)=0$, the results also allow g to be singular as $x \to a^+$. In this way, the case $g(x, y')=((N-1)/x)y'$, which arises when seeking radial solutions of $\nabla^2 y=f(x, y)$, is included. The results extend previous theorems of Taliaferro and more recent theorems of Gatica, Waltman, et al.

Key words. singular boundary value problems, existence and uniqueness

AMS(MOS) subject classifications. 34B15, 35J65

1. Introduction.

Two-point boundary value problems associated with the second-order equation

$$y''=F(x, y, y'), \qquad a<x<b, \tag{1.1}$$

can be singular in a variety of ways: a and/or b may be infinite, F may be unbounded near some $x_0 \in [a, b]$, or F may be unbounded near some particular value of y (or y'). This last possibility does not occur in linear problems and is the source of some added excitement in nonlinear problems.

Problems on infinite intervals have been studied by Granas et al. [8], Erbe and Schmitt [5], Berestycki, Lions, and Peletier (e.g., [4]), and by Baxley [1] and will not be considered here. Problems on finite intervals in which the singularity is caused by F being unbounded near a particular value of y have been studied by Luning and Perry [9], Nachman and Callegari [11], Taliaferro [13], Gatica, Oliker, and Waltman [7], and Baxley [2]. Recently, Gatica, Hernandez, and Waltman [6] have considered a finite interval problem which has singularities arising from F becoming infinite as $x \to a$ and also as $y \to 0$. We should also note that the problem considered by Erbe and Schmitt [5] on $0<x<\infty$ is singular at $x=0$.

It should be realized that a singularity of (1.1) at some finite value of y (or y') may only be bluffing: a solution of an associated boundary value problem might not come close enough to the singular value to be affected. Here are two such examples. A problem involving heat and mass transfer in a porous catalyst leads to the problem

$$y''=\alpha y \exp\left[\frac{\gamma\beta(1-y)}{1+\beta(1-y)}\right], \qquad 0<x<1,$$

$$y'(0)=0, \qquad y(1)=1,$$

where α, β, γ are positive constants. The apparent singularity at $y=1+1/\beta$ is only an idle threat because the solution actually satisfies $y(x) \leqq 1$ on $0 \leqq x \leqq 1$. Similar

* Received by the editors October 2, 1989; accepted for publication (in revised form) March 5, 1990.
† Department of Mathematics and Computer Science, Wake Forest University, Winston-Salem, North Carolina 27109. This research was done while the author was visiting the Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia 30322, and was supported in part by National Science Foundation grant 86-01398.

behavior is encountered in this problem involving diffusion in a chemical catalytic converter:

$$y''=\frac{1}{y-2}(y')^2, \qquad 0<x<1,$$

$$y(0)=A, \qquad y(1)=B,$$

where $B<A<2$. The apparent singularity at $y=2$ does no real harm because the solution satisfies $B\leqq y(x)\leqq A$ on $0\leqq x\leqq 1$. These two examples from the engineering literature are discussed by Na [10] and Baxley [3]. The second example can actually be integrated by elementary methods after dividing by y'.

In almost all of the articles referenced above, attention has been confined to problems where $F(x, y, y')$ is independent of y', or at worst linear in y'. Motivation has come from interest in the problem

$$y''+a(x)y^{-p}=0, \qquad 0<x<1, \tag{1.2}$$

where $p>0$ and $a(x)>0$ and continuous on $[0, 1]$, and the problem

$$y''+\frac{N-1}{x}y'+a(x)y^{-p}=0, \qquad 0<x<1, \tag{1.3}$$

where again $p>0$ and $a(x)>0$ on $[0, 1]$, which is encountered in the search (see, e.g., [4]) for radial solutions of elliptic partial differential equations in R^N which have the form

$$\nabla^2 u+a(r)u^{-p}=0,$$

where $r=|x|$ is the radial coordinate and corresponds to the variable x in (1.3).

To facilitate comparison of our results, we state the existence and uniqueness theorems of Taliaferro [13]. He studied the equation (1.2) with the boundary conditions

$$y(0)=0, \qquad y(1)=0, \tag{1.4}$$

and proved that the problem (1.2), (1.4) has a positive solution $\phi\in C[0, 1]\cap C^2(0, 1)$ if and only if

$$\int_0^1 x(1-x)a(x)\,dx<\infty, \tag{1.5}$$

in which case such a positive solution is unique. Furthermore,

$$\lim_{x\to 0^+}\phi'(x) \quad \left(\text{respectively, } \lim_{x\to 1^-}\phi'(x)\right)$$

exists if and only if

$$\int_0^{1/2}\frac{a(x)}{x^p}\,dx \quad \left(\text{respectively, } \int_{1/2}^1\frac{a(x)}{(1-x)^p}\,dx\right)$$

is finite. Note that for existence of a continuous solution, the value of $p>0$ is otherwise unrestricted, but in order to have $\phi'(x)$ continuous at one of the endpoints, either p must not be too large, or $a(x)$ must tend to zero rather rapidly as x approaches that endpoint. In particular, (1.2), (1.4) has a positive solution $\phi\in C^1[0, 1]\cap C^2(0, 1)$ if and only if

$$\int_0^1\frac{a(x)}{(s(x))^p}\,dx<\infty \tag{1.6}$$

where $s(x) = x$ for $0 \leqq x \leqq \frac{1}{2}$ and $s(x) = 1 - x$ for $\frac{1}{2} < x \leqq 1$. Gatica, Oliker, and Waltman [7] extended the part of the above result regarding the existence of a $C^1[0, 1]$ solution in two ways: they replaced (1.2) by the more general equation

$$y'' + f(x, y) = 0, \tag{1.7}$$

where $f: (0, 1) \times (0, \infty) \to (0, \infty)$ continuously and $f(x, y)$ is nonincreasing in y for fixed $x \in (0, 1)$, and they dealt with the more general two-point boundary conditions

$$a_0 y(0) - a_1 y'(0) = 0, \qquad b_0 y(1) + b_1 y'(1) = 0, \tag{1.8}$$

where a_0, a_1, $b_1 \geqq 0$ and $a_0 b_0 + a_1 b_0 + a_0 b_1 > 0$. Assuming that $f(x, y) \to +\infty$ as $y \to 0^+$ uniformly on compact subsets of $(0, 1)$, and that

$$\int_0^1 f(x, \theta s(x))\, dx < \infty \tag{1.9}$$

for each $\theta > 0$, then Gatica, Oliker, and Waltman [7] prove existence of a positive solution $\phi \in C^1[0, 1] \cap C^2(0, 1)$ of (1.7), (1.8). They also provide a uniqueness theorem generalizing the corresponding statement of Taliaferro. Note that for $f(x, y) = a(x)/y^p$, (1.9) holds if and only if (1.6) holds.

In contrast to the shooting method used by Taliaferro, Gatica, Oliker, and Waltman [7] proved first an interesting fixed point theorem for decreasing maps on a Banach space and then applied it to the boundary value problem.

In [6], Gatica, Hernandez, and Waltman continued to exploit their fixed point theorem and attacked the more difficult problem (1.3) with the boundary conditions

$$y'(0) = 0, \qquad y(1) = 0. \tag{1.10}$$

Replacing $a(x)y^{-p}$ with $f: [0, 1) \times (0, \infty) \to (0, \infty)$ continuous (note the half-open interval), they assume f is decreasing in y for each x, $\int_0^1 f(x, y)\, dx < \infty$ for each $y > 0$, $f(x, y) \to +\infty$ as $y \to 0^+$ and $f(x, y) \to 0$ as $y \to +\infty$, both limits being uniform on compact subsets of $(0, 1)$, and finally that $\int_0^1 f(x, \theta(1-x))\, dx < \infty$ for each $\theta > 0$. They then prove existence of a positive solution $\phi \in C^1[0, 1] \cap C^2(0, 1)$ of (1.3), (1.10), and again provide a uniqueness theorem.

In this paper, we shall unify and extend these results in a two-stage procedure. Throughout, we assume our equation has the form

$$y'' + g(x, y') + f(x, y) = 0, \qquad a < x < b, \tag{1.11}$$

and we first study the problem (1.11) with the boundary conditions

$$a_0 y(a) - a_1 y'(a) = A, \qquad b_0 y(b) + b_1 y'(b) = B, \tag{1.12}$$

where a_0, a_1, b_0, b_1, A, $B \geqq 0$ and $a_0 b_0 + a_1 b_0 + a_0 b_1 > 0$. We extend the basic result of Taliaferro regarding the sufficiency of the condition (1.5) or (1.6) for existence to the equation (1.11) in which significant nonlinearity is allowed in the term $g(x, y')$, but no singularity. The result in [7] is also a special case. Our theorem does not require that the problem be either singular or nonlinear; after all, we would not expect nonsingular or linear problems to be worse than singular or nonlinear ones. We also include a uniqueness theorem which improves the one in [7].

Finally, we consider the problem studied in [6]. Modeling (1.3), we allow $g(x, y')$ in (1.11) to be singular at $x = a$. Our theorem allows, for example,

$$g(x, z) = -\frac{|z|^p}{(x-a)^q},$$

where $p \geqq 1$ and $q > 0$. We allow boundary conditions of the form

$$y'(a) = 0, \qquad b_0 y(b) + b_1 y'(b) = B, \tag{1.13}$$

where b_1, $B \geqq 0$ and $b_0 > 0$. Again we provide significant extensions of the existence and uniqueness theorems in [6].

Our strategy is to construct a sequence of problems which appproximate (1.11), (1.12), which are not singular, but which "converge" to the singular problem. We apply a previous result [3] to obtain existence of a positive solution ϕ_n of the nth approximating problem and use a priori estimates to justify application of Ascoli's theorem.

The approximating problems are constructed in § 2; § 3 contains basic information on qualitative behavior and a priori bounds for the approximating solutions. Existence and uniqueness for the problem (1.11), (1.12) with $g(x, y')$ nonsingular appear in §§ 4 and 5, while the results for $g(x, y')$ singular as $x \to a-$ appear in § 6.

This effort was motivated by the desire to understand the underlying differences in the problems studied in [6], [7] by Waltman and his colleagues, who very kindly placed preprints of their work at my disposal.

2. Basic assumptions and approximating problems. The context for our study of the problem (1.11), (1.12) will allow singular behavior as $x \to a+$, as $x \to b-$, and as $y \to 0+$, and we shall seek positive solutions of (1.11), (1.12). From the perspective of the singularity at $y = 0$, a solution $\phi(x)$ may keep its distance from zero and thereby render the singularity powerless, or the solution may approach zero as $x \to a+$ or as $x \to b-$, thus giving the singularity at $y = 0$ opportunity to cause havoc. If (1.11), (1.12) has a positive solution $\phi(x)$ which is bounded away from zero as $x \to a+$ (respectively, $x \to b-$), we shall say that the singularity at $y = 0$ is irrelevant at the endpoint $x = a$ (respectively, $x = b$).

Here are our basic assumptions regarding the differential equation (1.11):

(HF1) $f: (a, b) \times (0, \infty) \to (0, \infty)$ is continuous;

(HF2) $f(x, y)$ is nonincreasing in y for each fixed $x \in (a, b)$;

(HG1) $g: [a, b] \times (-\infty, \infty) \to (-\infty, \infty)$ is continuous;

(HG2) $zg(x, z) \geqq 0$ for all $(x, z) \in [a, b] \times (-\infty, \infty)$;

If, in (1.12), $b_0 > 0$, put $I = [a, b)$; if $b_0 = 0$ (Neumann data), put $I = [a, b]$.

(HG3) g satisfies a uniform Lipschitz condition in z on each compact subset $S \subset I \times (-\infty, \infty)$. That is, given such a set S, there exists $K > 0$ such that

$$|g(x, z_2) - g(x, z_1)| \leqq K|z_2 - z_1|$$

whenever (x, z_2), $(x, z_1) \in S$;

(HG4) If $a_1 = 0$ and $b_0 > 0$, then $g(x, z) = O(z^2)$, as $z \to +\infty$, uniformly for $x \in [a, b]$;

(HG5) If $b_0 = 0$, then $g(x, z) = O(z \log z)$ as $z \to +\infty$, uniformly for $x \in [a, b]$.

These conditions allow singular behavior of $f(x, y)$ near both endpoints of (a, b) and as $y \to 0^+$. The term $g(x, y')$ allows significant nonlinearity in y', but essentially no singularity; later we will allow $g(x, z)$ to be singular as $x \to a+$.

Our strategy for finding a solution of (1.11), (1.12) is to construct a sequence of approximate problems to which we can apply a previous existence theorem and then use Ascoli's theorem on the corresponding sequence of solutions.

We will encounter a technical problem in the course of our work which is easily dispatched if $f(x, y)$ satisfies the following more restrictive condition than (HF1).

(HF0) $f: (a, b] \times (0, \infty) \to (0, \infty)$ is continuous.

Our work will be facilitated if we keep this possibility in mind. We do this from the outset at follows. Define, for integers $n > 2/(b-a)$,

$$h_n(x) = \begin{cases} n^2(x-a) & \text{for } a \leqq x < a + \dfrac{1}{n}, \\ n & \text{for } a + \dfrac{1}{n} \leqq x \leqq b - \dfrac{1}{n}, \\ n^2(b-x) & \text{for } b - \dfrac{1}{n} < x \leqq b, \end{cases}$$

$$h_n^*(x) = \begin{cases} n^2(x-a) & \text{for } a \leqq x < a + \dfrac{1}{n}, \\ n & \text{for } a + \dfrac{1}{n} \leqq x \leqq b. \end{cases}$$

So $h_n(x)$, $h_n^*(x)$ are continuous (piecewise linear) functions on $a \leqq x \leqq b$. We shall use these functions to approximate $f(x, y)$ by bounded, continuous functions on $[a, b] \times (-\infty, \infty)$. If $f(x, y)$ satisfies the more restrictive condition (HF0), we put

$$f_n(x, y) = \min [h_n^*(x), f(x, y)] \quad \text{for } (x, y) \in (a, b] \times (0, \infty);$$

if f satisfies (HF1) but not (HF0), we put

$$f_n(x, y) = \min [h_n(x), f(x, y)] \quad \text{for } (x, y) \in (a, b) \times (0, \infty).$$

In either case, $f_n(x, y)$ has a unique continuous extension to $[a, b] \times (0, \infty)$, and since $f(x, y)$ is nonincreasing in y, then we may define

$$f_n(x, 0) \equiv \lim_{y \to 0^+} f_n(x, y) \quad \text{for } a \leqq x \leqq b$$

and then extend $f_n(x, y)$ to $[a, b] \times (-\infty, \infty)$ by

$$f_n(x, y) \equiv f_n(x, 0) \quad \text{if } y < 0.$$

It is clear that now

$$f_n : [a, b] \times (-\infty, \infty) \to (0, \infty) \text{ is continuous}$$

and $f_n(x, y)$ is nonincreasing in y for each fixed $x \in [a, b]$. It is important to note also that $f_n \leqq f_{n+1} \leqq f$.

For each integer $n > 2/(b-a)$, we view

$$\text{(ODE(n))} \qquad y'' + g(x, y') + f_n(x, y) = 0$$

as an approximation for (1.11). The existence of a solution $\phi_n \in C^2[a, b]$ of the boundary value problem consisting of (ODE(n)) and the boundary conditions (1.12) follows from existence theorems in [3]. In the case that $b_0 = 0$, then necessarily $a_0 b_1 > 0$ and existence follows from (HG5) and Theorem 2.1 of [3]. To conclude existence in case $b_0 > 0$, we first change variables, replacing x by $a + b - x$, to get the transformed problem

$$y'' + g(a + b - x, -y') + f_n(a + b - x, y) = 0, \qquad a < x < b,$$

$$b_0 y(a) - b_1 y'(a) = B, \qquad a_0 y(b) + a_1 y'(b) = A.$$

Note that $zg(a + b - x, -z) \leqq 0$. If $b_0 a_1 > 0$, existence is a simple consequence of Theorem 2.1 of [3]. If $b_0 > 0$ and $a_1 = 0$, then $b_0 a_0 > 0$ and existence follows from (HG4) and Theorem 2.2 of [3].

Before collecting some qualitative information about the graphs of these solutions $\phi_n(x)$, we state the following consequence of the maximum principle.

LEMMA 2.1. *Let* $T_n(y) = y'' + g(x, y') + f_n(x, y)$. *Suppose that* $T_n\phi(x) \geqq 0$, $T_n\psi(x) \leqq 0$, and $\phi(x) > \psi(x)$ *for* $x \in [c, d] \subset I$. *If* $\phi - \psi$ *is not constant on* $[c, d]$, *then* $\phi - \psi$ *cannot attain its maximum on* $[c, d]$ *at an interior point of* $[c, d]$. *This maximum can only occur at* c *with* $\phi'(c) < \psi'(c)$ *or at* d *with* $\phi'(d) > \psi'(d)$.

Proof. We use the standard procedure. Since $T_n\phi(x) \geqq 0$, $T_n\psi(x) \leqq 0$ *on* $[c, d]$, then $u = \phi - \psi$ satisfies

$$u'' + a(x)u' + b(x)u \geqq 0, \qquad c \leqq x \leqq d,$$

where

$$a(x) = \begin{cases} \dfrac{g(x, \phi'(x)) - g(x, \psi'(x))}{\phi'(x) - \psi'(x)} & \text{if } \phi'(x) \neq \psi'(x), \\ 0 & \text{if } \phi'(x) = \psi'(x). \end{cases}$$

and

$$b(x) = \frac{f_n(x, \phi(x)) - f_n(x, \psi(x))}{\phi(x) - \psi(x)}.$$

By (HG3), $a(x)$ is bounded on $[c, d]$ and since $f_n(x, y)$ is continuous and nonincreasing in y, then $b(x)$ is continuous and nonpositive on $[c, d]$. The maximum principle [12, pp. 6-7] therefore applies to complete the proof.

3. Approximating solutions: Behavior and estimates. We now describe qualitatively the graphs of $\phi_n(x)$.

LEMMA 3.1. $\phi_n(x)$ *has at most one critical point in* (a, b).

(a) *If such a critical point* $c_n \in (a, b)$ *exists, then* $a_0 b_0 > 0$ *and* ϕ_n *has an absolute maximum on* $[a, b]$ *at* c_n,

$$\phi_n'(x) < 0, \quad \phi_n(x) > (B - b_1\phi_n'(b))/b_0 \quad \text{for } c_n < x < b,$$

$$\phi_n'(x) > 0, \quad \phi_n(x) \geqq (a_1\phi_n'(a) + A)/a_0 \quad \text{for } a \leqq x < c_n.$$

(b) *If* ϕ_n *has no critical point in* (a, b), *then* ϕ_n *attains an absolute maximum on* $[a, b]$ *at exactly one endpoint* c_n *and either*

$$c_n = b, \quad a_0 > 0, \quad \phi_n'(x) > 0, \quad \phi_n(x) \geqq (a_1\phi_n'(a) + A)/a_0 \quad \text{for } a \leqq x < b,$$

or

$$c_n = a, \quad b_0 > 0, \quad \phi_n'(x) < 0, \quad \phi_n(x) > (B - b_1\phi_n'(b))/b_0 \quad \text{for } a < x < b.$$

Proof. If $\phi_n'(c) = 0$ for $c \in (a, b)$, then since (HG1), (HG2) imply $g(x, 0) = 0$, for $a \leqq x \leqq b$, it follows from (HF1) that

$$\phi_n''(c) = -f_n(c, \phi_n(c)) < 0$$

and thus c gives a strict local maximum. Clearly, at most one such c_n can exist. If such a c_n exists, then it certainly gives an absolute maximum and $\phi_n'(x) > 0$ on (a, c_n) and $\phi_n'(x) < 0$ on (c_n, b). If no such critical point in (a, b) exists, then $\phi_n'(x)$ has the same sign throughout (a, b). Suppose that $c_n > a$. By (HG2), we then have

$$\phi_n''(x) = -g(x, \phi_n'(x)) - f_n(x, \phi_n(x)) < 0 \quad \text{for } a < x < c_n,$$

and $\phi_n(x)$ is concave down on (a, c_n). Thus $\phi_n'(x)$ is decreasing on $[a, c_n]$ with $\phi_n'(a) > 0$. The boundary condition at $x = a$ gives $a_0\phi_n(a) = a_1\phi_n'(a) + A$. If $a_1 = 0$,

then $a_0 > 0$ by hypothesis; if $a_1 > 0$, then

$$a_0\phi_n(a) = a_1\phi_n'(a) + A \geq a_1\phi_n'(a) > 0$$

forces $a_0 > 0$. Thus for $a \leqq x \leqq c_n$,

$$\phi_n(x) \geqq \phi_n(a) = (a_1\phi_n'(a) + A)/a_0.$$

Suppose now that $c_n < b$ and focus attention on (c_n, b). It can no longer be concluded that $\phi_n''(x)$ is negative and ϕ_n may not be concave down on (c_n, b). (This loss of concavity is the root of technical difficulties with which (HF0) or some other alternative will assist later.) To show that $b_0 > 0$, suppose that $b_0 = 0$, $b_1 > 0$. Then the constant function $\phi(x) = \phi_n(c_n)$ satisfies $T_n(\phi) = f_n(x, \phi) \geqq 0$. We shall show that $\phi_n(x) \geqq \phi(x)$ on $[c_n, b]$, which contradicts the fact that c_n gives a strict maximum for $\phi_n(x)$. Assuming that $\phi_n(x) \geqq \phi(x)$ is false, we may suppose that $\phi - \phi_n$ attains a positive maximum at some point $x_0 \in [c_n, b]$. Clearly, $x_0 \neq c_n$ since $(\phi - \phi_n)(c_n) = 0$, and certainly $\phi - \phi_n$ is not constant on $(c_n, b]$. Choose $c \in (c_n, x_0)$ so that $\phi - \phi_n > 0$ on $[c, x_0]$. Since $b_1(\phi - \phi_n)'(b) = -B \leqq 0$, then surely $(\phi - \phi_n)'(x_0) \leqq 0$, contradicting Lemma 2.1. Thus $b_0 > 0$ and the boundary condition at $x = b$ gives $\phi_n(x) > \phi_n(b) = (B - b_1\phi_n'(b))/b_0$ for $c_n < x < b$, completing the proof.

Note that if the maximum of ϕ_n on $[a, b]$ occurs at $c_n < b$, we do not conclude that $\phi_n'(b) < 0$, in contrast to the fact that $\phi_n'(a) > 0$ if $c_n > a$. The possibility that $\phi_n'(b) = 0$ even though $c_n < b$ will be dealt with later.

LEMMA 3.2. *$\phi_n(x) \leqq \phi_{n+1}(x)$ for $a \leqq x \leqq b$ and integers $n \geqq 2/(b-a)$.*

Proof. Suppose on the contrary that for some n, the function $\phi_n - \phi_{n+1}$ attains a positive maximum at some point $x_0 \in [a, b]$. Since $\phi_n - \phi_{n+1}$ satisfies

$$a_0 y(a) - a_1 y'(a) = 0, \qquad b_0 y(b) + b_1 y'(b) = 0, \tag{3.1}$$

and $a_0 + b_0 > 0$, then $\phi_n - \phi_{n+1}$ is not constant on $[a, b]$. Note that $x_0 = b$ implies $b_0 = 0$ and thus $I = [a, b]$. Clearly, there exists an interval $[c, d] \subset [a, b]$ with $x_0 \in [c, d]$ and $\phi_n - \phi_{n+1}$ is positive and nonconstant on $[c, d]$. Since $T_n(\phi_n) = 0$ and $f_n(x, y) \leqq f_{n+1}(x, y)$ implies $T_n(\phi_{n+1}) \leqq T_{n+1}(\phi_{n+1}) = 0$, then Lemma 2.1 shows that $x_0 = a$ and $(\phi_n - \phi_{n+1})'(x_0) < 0$ or $x_0 = b$ and $(\phi_n - \phi_{n+1})'(x_0) > 0$. Each of these is impossible because $\phi_n - \phi_{n+1}$ satisfies (3.1).

We now obtain estimates on the derivatives ϕ_n' which will be central in allowing the use of Ascoli's theorem later.

LEMMA 3.3. *Suppose $c_n < b$ and $n > 2/(b-a)$. If $x_1 \in (a, b)$, $x_1 \geqq c_n$, then*

$$|\phi_n'(x)| \leqq |\phi_n'(x_1)| + 2\int_{x_1}^{x} f(s, \phi_n(s))\, ds \quad \text{for } x_1 \leqq x < b.$$

Proof. With x_1 as described, choose and fix $x \in (x_1, b)$. Then choose d so that $|\phi_n'(x)|$ assumes its maximum on $[x_1, x]$ at d. By Lemma 3.1, $\phi_n'(s) < 0$ on $[x_1, x]$ and (HG2) implies that

$$\phi_n''(s) = -g(s, \phi_n'(s)) - f_n(s, \phi_n(s)) \geqq -f_n(s, \phi_n(s)), \qquad x_1 \leqq s \leqq x.$$

Multiplying by $\phi_n'(s)$ and integrating over $[x_1, d]$, we obtain

$$|\phi_n'(d)|^2 \leqq |\phi_n'(x_1)|^2 - 2\int_{x_1}^{d} \phi_n'(s) f_n(s, \phi_n(s))\, ds.$$

Since $|\phi_n'(d)| = -\phi_n'(d)$ is the maximum value of $|\phi_n'(s)|$ on $[x_1, d]$, there follows

$$|\phi_n'(d)|^2 \leqq |\phi_n'(x_1)|^2 + 2|\phi_n'(d)| \int_{x_1}^{d} f_n(s, \phi_n(s))\, ds.$$

The desired result is obtained after division by $|\phi_n'(d)|$.

LEMMA 3.4. *Suppose* $c_n > a$.

(i) *If* $a_1 > 0$, *then* $0 \leqq \phi_n'(x) < a_0\phi_n(c_n)/a_1$, *for* $a \leqq x \leqq c_n$.

(ii) *If* $a_1 = 0$, $b_0 > 0$, *and* $x_2 \in (a, b)$, $x_2 \leqq c_n$, *then there exist positive numbers* K_1, z_1, *independent of* n, *so that*

$$0 \leqq \phi_n'(x) \leqq 2 \exp(2K_1\phi_n(c_n))\left\{z_1 + \phi_n'(x_2) + 2\int_x^{x_2} f(s, \phi_n(s))\, ds\right\}$$

for $a < x \leqq x_2$.

(iii) *If* $b_0 = 0$ *and* $x_2 \in (a, b)$, *then there exist positive numbers* K_2, z_2, *independent of* n, *so that*

$$\log\log \phi_n'(x) \leqq \log\log z_2 + K_2(b-a) + \frac{\int_x^{x_2} f(s, \phi_n(s))\, ds}{z_2 \log z_2}$$

for $a < x \leqq x_2$.

Proof. If $a_1 > 0$, the boundary condition at $x = a$ gives $\phi_n'(a) \leqq a_0\phi_n(a)/a_1$ and (i) follows from the facts that $\phi_n(a) < \phi_n(c_n)$ and ϕ_n is concave down on (a, c_n). We pass to (ii) and so may use (HG4). Thus there exist positive constants K_1, z_1 so that

$$g(x, z) \leqq K_1 z^2 \quad \text{for } z \geqq z_1.$$

Arguing by contradiction, suppose that there exists $x_0 \in (a, x_2)$ for which

$$\phi_n'(x_0) > 2 \exp(2K_1\phi_n(c_n))\left\{z_1 + \phi_n'(x_2) + 2\int_{x_0}^{x_2} f(s, \phi_n(s))\, ds\right\}.$$

Then we may choose $x_1 \in (x_0, x_2)$ so that

$$\phi_n'(x_1) = M \equiv \left\{z_1 + \phi_n'(x_2) + 2\int_{x_0}^{x_2} f(s, \phi_n(s))\, ds\right\},$$

and $\phi_n'(x) > M$ for $x_0 \leqq x < x_1$. The differential equation then gives

$$\phi_n''(x) = -g(x, \phi_n'(x)) - f_n(x, \phi_n(x)) \geqq -K_1(\phi_n'(x))^2 - f(x, \phi_n(x))$$

for $x_0 \leqq x \leqq x_1$. Dividing by $(\phi_n'(x))^2$ and integrating over $[x_0, x]$, where $x_0 < x \leqq x_1$, we get

$$\frac{1}{\phi_n'(x_0)} - \frac{1}{\phi_n'(x)} \geqq -K_1(x - x_0) - \frac{\int_{x_0}^{x} f(s, \phi_n(s))\, ds}{(\phi_n'(x))^2}$$

since $\phi_n'(x)$ is the minimum value of ϕ_n' on $[x_0, x]$. Then

$$\phi_n'(x) > 2\int_{x_0}^{x} f(s, \phi_n(s))\, ds$$

implies

$$\frac{1}{2\phi_n'(x)} \leqq \frac{1}{\phi_n'(x)}\left[1 - \frac{\int_{x_0}^{x} f(s, \phi_n(s))\, ds}{\phi_n'(x)}\right] \leqq K_1(x - x_0) + \frac{1}{\phi_n'(x_0)}.$$

Thus

$$2\phi_n'(x) \geqq \frac{\phi_n'(x_0)}{K_1\phi_n'(x_0)(x - x_0) + 1}, \qquad x_0 < x \leqq x_1. \tag{3.2}$$

Integrating (3.2) over $[x_0, x_1]$ and then using (3.2) again for the value $x = x_1$ produces

$$2(\phi_n(x_1) - \phi_n(x_0)) \geqq \frac{1}{K_1}\log(K_1\phi_n'(x_0)(x_1 - x_0) + 1)$$

and

$$2K_1\phi_n(c_n) \geqq \log\left[\frac{\phi_n'(x_0)}{2\phi_n'(x_1)}\right],$$

which leads to

$$\phi_n'(x_0) \leqq 2M \exp(2K_1\phi_n(c_n)),$$

a contradiction.

To prove (iii), we may use (HG5). So there exists $z_2 > 0$, $K_2 > 0$ such that

$$g(x, z) \leqq K_2 z \log z \quad \text{for } z \geqq z_2.$$

We may surely assume that $z_2 > B/b_1 = \phi_n'(b)$. If $\phi_n'(x) \leqq z_2$ on (a, x_2), there is nothing to prove. Otherwise there exists $x_1 \in (a, x_2)$ such that $\phi_n'(x_1) = z_2$ and $\phi_n'(x) > z_2$ for $a \leqq x < x_1$. Then

$$\phi_n''(s) \geqq -K_2\phi_n'(s) \log(\phi_n'(s)) - f(s, \phi_n(s)), \qquad a < s \leqq x_1.$$

After dividing by $\phi_n' \log \phi_n'$ and integrating from x to x_1, we obtain the desired conclusion.

4. Existence. We are now ready to formulate and prove our main result giving sufficient conditions for the existence of a solution ϕ of problem (1.11), (1.12). As might be expected, these conditions are less demanding if the singularity at $y = 0$ is irrelevant at one or both endpoints. Since we intend to construct our solution using Ascoli's theorem as the limit of the subsequence of the solutions $\{\phi_n\}$ of the approximating problems, we look to Lemmas 3.1 and 3.2 for guidance. Let N be the smallest integer larger than $2/(b-a)$ and suppose $n \geqq N$. If $A + a_1 > 0$, then Lemma 3.1 implies that $\phi_n(a) > 0$; from Lemma 3.2, it follows that $\phi(a)$ will also be positive and thus the singularity at $y = 0$ will be irrelevant at $x = a$. The other endpoint is more troublesome, because Lemma 3.1 does not guarantee that $\phi_n'(b)$ will be negative if $c_n < b$. If $B > 0$, then certainly $\phi_n(b) > 0$ and the singularity at $y = 0$ is also irrelevant at $x = b$. We immediately give simple conditions which force $\phi_n'(b) < 0$ when $c_n < b$. This can be accomplished via either the function $g(x, y')$ or the function $f(x, y)$. Here is the condition on g:

(HG6) There exists $\delta > 0$ so that $g(x, z) = O(|z|)$ as $z \to 0-$, uniformly for $x \in [b-\delta, b]$.

LEMMA 4.1. *Suppose $c_n < b$ and either that f satisfies the stronger condition* (HF0) *or that g satisfies* (HG6). *Then $\phi_n'(b) < 0$.*

Proof. By Lemma 3.1, ϕ_n attains its minimum on $[c_n, b]$ at b where $\phi_n'(b) \leqq 0$. Assuming that f satisfies (HF0) and $\phi_n'(b) = 0$, then

$$\phi_n''(b) = -f_n(b, \phi_n(b)) < 0,$$

which is clearly impossible. Thus $\phi_n'(b) < 0$. If $g(x, z)$ satisfies (HG6) and $\phi_n'(b) = 0$, then we may assume that $c_n < b - \delta$ and that there exists a constant k so that $|g(x, \phi_n'(x))| \leqq k|\phi_n'(x)|$ for $x \in [b-\delta, b)$. Thus

$$\phi_n''(x) + k\phi_n'(x) \leqq \phi_n''(x) + g(x, \phi_n'(x)) = -f(x, \phi_n(x)) < 0, \qquad b - \delta \leqq x < b.$$

Multiplying by $\exp(kx)$ and integrating, we get

$$\exp(kb)\phi_n'(b) < \exp(k(b-\delta))\phi_n'(b-\delta) < 0,$$

a contradiction. Thus $\phi_n'(b) < 0$.

If the hypotheses of Lemma 4.1 are satisfied, then Lemma 3.1 implies that the singularity at $y=0$ will be irrelevant at $x=b$ if $B=0$, $b_1>0$. In order to get a solution of (1.11), (1.12) which is sufficiently smooth at an endpoint of $[a, b]$, we shall need to impose certain integrability conditions on $f(x, y)$ near that endpoint. Let m be the midpoint of $[a, b]$. Here are the various conditions:

(a1) $\int_a^m f(x, \theta(x-a))\,dx<\infty$ for $\theta>0$,

(a2) $\int_a^m f(x, y)\,dx<\infty$ for $y>0$,

(a3) $\int_a^m (x-a)f(x, y)\,dx<\infty$ for $y>0$,

(b1) $\int_m^b f(x, \theta(b-x))\,dx<\infty$ for $\theta>0$,

(b2) $\int_m^b f(x, y)\,dx<\infty$ for $y>0$,

(b3) $\int_m^b (b-x)f(x, y)\,dx<\infty$ for $y>0$.

We say that $f(x, y)$ satisfies the strong integrability condition at $x=a$ if $f(x, y)$ satisfies (a2) in case $A+a_1>0$, and $f(x, y)$ satisfies (a1) in case $A+a_1=0$. Similarly, $f(x, y)$ satisfies the strong integrability condition at $x=b$ if $f(x, y)$ satisfies (b2) in the case where $B+b_1>0$, and $f(x, y)$ satisfies (b1) in the case where $B+b_1=0$; we also require that $f(x, y)$ satisfy HF0 if $B=0$, $b_0>0$, and $g(x, z)$ fails to satisfy (HG6). These are essentially the conditions used by Taliaferro in [13] and later in [7] in the case that $g(x, z)\equiv 0$.

To allow for the possibility that a solution's derivative is unbounded near a Dirichlet endpoint, we also consider "weak" integrability conditions analogous to those of Taliaferro [13]. We say that $f(x, y)$ satisfies the weak integrability condition at $x=b$ if $f(x, y)$ satisfies (b3). Similarly, $f(x, y)$ satisfies the weak integrability condition at $x=a$ if $f(x, y)$ satisfies (a3) when $g(x, z)=O(z)$ as $z\to+\infty$; otherwise, we must strengthen this condition to (a2). Since the weak conditions will only be considered at Dirichlet endpoints, it is clear that the weak condition at $x=a$ is no weaker than the strong condition if $A>0$ and $g(x, z)$ is not $O(z)$ for large z. We leave open the question of whether or not this strengthened condition is necessary.

For the purpose of easy reference and hopefully to make the proof transparent, we begin with several simple lemmas.

LEMMA 4.2. *If $f(x, y)$ satisfies the strong integrability condition at $x=a$, then $f(x, \phi_n(x))$ is integrable over the interval $[a, m]$. If $f(x, y)$ satisfies the strong integrability condition at $x=b$, then $f(x, \phi_n(x))$ is integrable over the interval $[m, b]$.*

Proof. If $A+a_1>0$, then Lemma 3.1 implies that there exists $y>0$ so that $\phi_n(x)\geqq y$ on $[a, m]$; thus by (HF2), $f(x, \phi_n(x))\leqq f(x, y)$ on $[a, m]$. If $A+a_1=0$, Lemma 3.1 implies that $\phi_n'(a)>0$ and so there exists $\theta>0$ for which $\phi_n(x)\geqq\theta(x-a)$ on $[a, m]$, so by (HF2) again, $f(x, \phi_n(x))\leqq f(x, \theta(x-a))$. The second statement follows similarly, but Lemma 4.1 must also be used.

LEMMA 4.3. *If $f(x, y)$ satisfies the weak integrability condition at $x=a$, then $\int_x^m f(s, y)\,ds$ is integrable over the interval $a\leqq x\leqq m$ for each fixed $y>0$. If $f(x, y)$ satisfies*

the weak integrability condition at $x = b$, then $\int_m^x f(s, y)\, ds$ is integrable over the interval $m \leqq x \leqq b$ for each fixed $y > 0$.

Proof. Fubini's theorem gives

$$\int_a^m \int_x^m f(s, y)\, ds\, dx = \int_a^m (s - a) f(s, y)\, ds,$$

implying that the first statement. The second statement is similar.

These integrability statements will be used in tandem with the next lemma to obtain equicontinuity.

LEMMA 4.4. *If $h(x)$ is integrable over some interval $[c, d]$, then given $\varepsilon > 0$, there exists $\delta > 0$ so that $c \leqq x_1 < x_2 \leqq d$ and $x_2 - x_1 < \delta$ implies $\int_{x_1}^{x_2} |h(x)|\, dx < \varepsilon$.*

Proof. This statement is the well-known fact from real analysis about absolute continuity of integrals.

Here are some intuitive remarks about our general strategy. To fix ideas, consider the case that c_n lies strictly between a and b. We want to use Lemma 3.3 to get uniform bounds on ϕ_n' to the right of c_n. If Lemma 4.2 applies, we expect uniform bounds up to b and we would like to integrate to get uniform bounds on $\phi_n(c_n)$; alternatively, we expect to integrate using Lemma 4.3 to get uniform bounds on $\phi_n(c_n)$. Then we want to use the estimates in Lemma 3.4 to obtain bounds to the left of c_n. But we have problems to overcome: if the c_n could accumulate at a and the strong integrability condition fails at a, the bound of Lemma 3.3 would fail us near c_n. Without an a priori bound on $\phi_n(c_n)$, the first two estimates in Lemma 3.4 are useless. Also, if the c_n could accumulate at b, the estimates of Lemma 3.4(ii) and (iii) would fail us near c_n. So an important part of the main proof is to show that the c_n cannot accumulate at a weak endpoint and the prevention of such accumulation at a is the technical problem to which we earlier referred. The extra assumption on $g(x, z)$ will be used via the following lemma.

LEMMA 4.5. *Suppose $c_n > a$ and $g(x, z) = O(z)$ as $z \to +\infty$. If $x_2 \in (a, b)$ and $x_2 \leqq c_n$, then there exist constants z_1 and K_1 so that*

$$\phi_n'(x) \leqq \exp\left(K_1(b - a)\right)\left\{ z_1 + \phi_n'(x_2) + \int_x^{x_2} f(s, \phi_n(s))\, ds \right\}$$

for $a < x < x_2$.

Proof. By hypothesis, there exist z_1 and K_1 so that $g(x, z) \leqq K_1 z$ for $z \geqq z_1$. If $\phi_n'(x) \leqq z_1 + \phi_n'(x_2)$ for $a \leqq x \leqq x_2$, there is nothing to prove. Otherwise, there exists $x_1 \in (a, x_2)$ such that $\phi_n'(x_1) = z_1 + \phi_n'(x_2)$. Then

$$\phi_n''(s) + K_1 \phi_n'(s) \geqq -f_n(s, \phi_n(s)) \quad \text{for } a < s < x_1,$$

and integration over $[x, x_1]$ leads quickly to the conclusion.

It will be convenient to have the following trivial extension of the usual Ascoli theorem.

LEMMA 4.6. *Suppose that $\{g_n\}$ is an equicontinuous sequence of functions on a finite interval $[c, d]$ and there exists a point x_0 in $[c, d]$ for which $\{g_n(x_0)\}$ is bounded. Then $\{g_n\}$ contains a uniformly convergent subsequence on $[c, d]$.*

Proof. The hypotheses imply that $\{g_n\}$ is uniformly bounded on $[c, d]$ so Ascoli's theorem applies.

Here is our main theorem.

THEOREM 4.1. *Suppose that f, g satisfy* (HF1)-(HF2) and (HG1)-(HG5). *At each endpoint of $[a, b]$, suppose either that $f(x, y)$ satisfies the strong integrability condition, or that the boundary condition at that endpoint is a Dirichlet condition and $f(x, y)$ satisfies*

the weak integrability condition. Let J denote the interval obtained by removing from $[a, b]$ any endpoint at which $f(x, y)$ fails to satisfy the strong integrability condition. Then the boundary value problem (1.11), (1.12) *has at least one solution* $\phi \in C^2(a, b) \cap C^1(J) \cap C[a, b]$. *Moreover, if* $A + a_0 = 0$, *then* ϕ' *satisfies the inequality*

$$|\phi'(x)| \leqq 2 \int_a^x f(s, \phi(s))\, ds < +\infty \quad \text{for } a \leqq x < b. \tag{4.1}$$

Proof. We shall show that some subsequence of $\{\phi_n\}$ satisfies the hypotheses of Lemma 4.6 and that the corresponding subsequence of $\{\phi_n'\}$ is uniformly bounded on each compact subset of J, and then use a slight modification of the familiar diagonalization argument with Ascoli's theorem. Because the argument differs if the point c_n where the maximum of ϕ_n occurs is an endpoint, we shall arrange our proof to deal conveniently with these special cases first. So we begin with the case that $c_n = b$ for infinitely many values of $n \geqq N$ and temporarily focus attention on these values of n.

Suppose $b_0 > 0$. Then the boundary condition at $x = b$ guarantees that $\phi_n(b) \leqq B/b_0$. If $a_1 > 0$, then $\{\phi_n'\}$ is uniformly bounded on $[a, b]$ by Lemma 3.4(i) and hence $\{\phi_n\}$ is equicontinuous on $[a, b]$. So we pass to the case that $a_1 = 0$. Here the concavity of ϕ_n on $[a, b]$ implies that

$$\phi_n'(m)(m - a) < \int_a^m \phi_n'(x)\, dx = \phi_n(m) - \phi_n(a) < \frac{B}{b_0} \tag{4.2}$$

so Lemma 3.4(ii) applies (with $x_2 = m$) to give constants C_1 and C_2 so that

$$\phi_n'(x) \leqq C_1 + C_2 \int_x^m f(s, \phi_n(s))\, ds \leqq C_1 + C_2 \int_x^m f(s, \phi_N(s))\, ds \tag{4.3}$$

for $a < x \leqq m$. Thus, using Lemma 4.2 if needed, we conclude from the concavity that $\{\phi_n'\}$ is uniformly bounded on each compact subset of J. To show the equicontinuity of $\{\phi_n\}$ on $[a, b]$, let $\varepsilon > 0$ be given. Define

$$h(x) = C_1 + C_2 \int_x^m f\left(s, \frac{\varepsilon}{2}\right) ds \quad \text{for } a < x \leqq m$$

and

$$h(x) = \frac{B}{(m - a)b_0} \quad \text{for } m < x \leqq b.$$

By Lemma 4.3, $h(x)$ is integrable over $[a, b]$. Let $\delta > 0$ be the number given by Lemma 4.4 (with ε replaced by $\varepsilon/2$ and $[c, d]$ replaced by $[a, b]$). We show that $a \leqq x_1 < x_2 \leqq b$ and $x_2 - x_1 < \delta$ implies

$$|\phi_n(x_2) - \phi_n(x_1)| < \varepsilon. \tag{4.4}$$

If $\phi_n(b) - \phi_n(a) < \varepsilon$, there is nothing to prove. Otherwise, choose $a_n \in (a, b)$ so that $\phi_n(a_n) = \phi_n(a) + \varepsilon/2$. Then if $x_1 \geqq a_n$, (4.2) and (4.3) imply $\phi_n'(x) \leqq h(x)$ on $[x_1, x_2]$ and integration gives (4.4) with ε replaced by $\varepsilon/2$. For $x_1 < a_n$, (4.4) follows immediately from the triangle inequality. If $b_0 = 0$, then $a_0 > 0$ and $b_1 > 0$. Thus from the boundary conditions, $\{\phi_n(a)\}$ is bounded by A/a_0 and $\{\phi_n'(b)\}$ is bounded by B/b_1. Also the strong integrability condition is satisfied at $x = b$. If $f(x, y)$ satisfies (a2), then Lemma 3.4(iii) applies (with x_2 replaced by b) to give a uniform bound for $\{\phi_n'\}$ on $[a, b]$ and equicontinuity of $\{\phi_n\}$ on $[a, b]$ is immediate. Otherwise, Lemma 4.5 may be applied to give (4.3) again and the argument in the case $b_0 > 0$ may be repeated. So we have the desired subsequence if $c_n = b$ for infinitely many n.

We now pass to the case that $c_n = a$ for infinitely many n. First suppose that $a_1 > 0$. Then $-A/a_1 < \phi_n'(a) \leqq 0$ and the strong integrability condition is satisfied at $x = a$. The bound for ϕ_n' of Lemma 3.3 may now be used for $h(x)$ with Lemmas 4.3 and 4.4 to show that $\{\phi_n\}$ is equicontinuous on $[a, b]$. If $x = b$ is a Dirichlet endpoint, then $\{\phi_n(b)\}$ is bounded and we are done; otherwise the strong integrability condition is also satisfied at $x = b$, Lemma 3.3 gives a uniform bound on $\{\phi_n'\}$ on $[a, b]$, and at least one of the boundary conditions then gives a bound on $\{\phi_n\}$ at the corresponding endpoint. Next suppose that $a_1 = 0$. Then $\phi_n(a) = A/a_0$ for every n of interest, providing a uniform bound for $\{\phi_n\}$ on $[a, b]$. With an eye to using Lemma 3.3, we shall show that if $x_1 \in (a, b)$, then $\{\phi_n'(x_1)\}$ is bounded. Choose x_0 as the midpoint of $[a, x_1]$. Applying Lemma 3.3 (with the roles of x and x_1 interchanged), we easily conclude that

$$\phi_n'(x) \leqq \phi_n'(x_1) + 2\int_{x_0}^{x_1} f(s, \phi_N(s))\, ds \quad \text{for } x_0 \leqq x \leqq x_1.$$

If $\{\phi_n'(x_1)\}$ was unbounded below, integration from x_0 to x_1 would force $\{\phi_n(x_0)\}$ to be unbounded, contradicting the uniform bound on $[a, b]$. To show the equicontinuity of $\{\phi_n\}$ on $[a, b]$, choose $\varepsilon > 0$. If $\phi_n(b) - \phi_n(a) < \varepsilon$, there is nothing to prove. Otherwise, choose x_1 so that $\phi_N(x_1) = A/a_0 - \varepsilon/3$ and b_n so that $\phi_n(b_n) < \phi_n(b) + \varepsilon$. With this choice of x_1, let

$$h(x) = \phi_n'(x_1) + \int_{x_1}^{x} f\left(s, \frac{\varepsilon}{3}\right) ds \quad \text{for } x_1 < x < b.$$

Then Lemma 3.3 shows that $h(x)$ is a bound for $\phi_n'(x)$ on $[x_1, b_n]$, and Lemmas 4.3 and 4.4 may be used as before to finish the proof. So now we have the desired subsequence in the case that $c_n = a$ for infinitely many n.

If neither of the earlier cases hold, then there exists $M > N$ for which $c_n \in (a, b)$ for all $n \geqq M$. It is necessary to know shortly that the sequence $\{c_n\}$ is bounded away from any weak endpoint, for $n \geqq M$. If b is a weak endpoint, then

$$\phi_n(c_n) - \phi_n(b) > \phi_M(c_M) - \phi_M(b) \equiv Q \quad \text{for } n > M.$$

Choose $b_n \in (c_n, b)$ so that $\phi_n(b_n) = \phi_n(b) + Q/2$. Integrating the bound of Lemma 3.3 (with x_1 replaced by c_n), we find that

$$\frac{Q}{2} < \phi_n(c_n) - \phi_n(b_n) \leqq 2\int_{c_n}^{b}\int_{c_n}^{x} f\left(s, \frac{Q}{2}\right) ds\, dx,$$

a clear contradiction of Lemma 4.4 if $\{c_n\}$ is not bounded away from $x = b$.

Suppose a is a weak endpoint. If $f(x, y)$ satisfies (a2), then the estimate of Lemma 3.3 (with $x_1 = c_n$) may be integrated to show that $\phi_n(c_n)$ is bounded and then the estimate of Lemma 3.4(ii) may be used (with $x_2 = c_n$) and integrated to contradict Lemma 4.4 if c_n is not bounded away from $x = a$; otherwise we may integrate the estimate of Lemma 4.5 and again contradict Lemma 4.4 if c_n is not bounded away from $x = a$.

If b is a strong endpoint, then Lemma 3.3 (with $x_1 = c_n$) implies that $\{\phi_n'(b)\}$ is bounded; here we use either that a is a strong endpoint or that $\{c_n\}$ is bounded away from a. Since $c_n < b$, Lemma 3.1 forces $b_0 > 0$. Whether or not b is a strong endpoint, the boundary condition at $x = b$ then gives a bound on $\{\phi_n(b)\}$ and integration of the estimate of Lemma 3.3 (with $x_1 = c_n$) over $[c_n, b]$ produces a bound on $\{\phi_n(c_n)\}$ for $n \geqq M$. The uniform boundedness of $\{\phi_n'\}$ on each compact subset of J now follows from Lemma 3.3 and either Lemma 3.4(i) or (ii) or Lemma 4.5.

To establish the equicontinuity of ϕ_n on $[a, b]$, let $\varepsilon > 0$ be given. It suffices to show that on each of $[a, c_n]$ and $[c_n, b]$, there exists $\delta > 0$ so that

$$|\phi_n(x_2) - \phi_n(x_1)| < \varepsilon/2$$

whenever $|x_2 - x_1| < \delta$. The argument is similar in each case. To illustrate, consider the interval $[a, c_n]$. If

$$\phi_n(c_n) - \phi_n(a) < \varepsilon/2$$

there is nothing to prove. Otherwise, choose $a_n \in [a, c_n]$ so that $\phi_n(a_n) = \phi_n(a) + \varepsilon/4$. Depending on the integrability hypothesis satisfied at $x = a$, either the estimate of Lemma 3.4(i) or (ii) or that of Lemma 4.5 may be integrated and Lemma 4.4 used to show that there exists $\delta > 0$ for which

$$\phi_n(x_2) - \phi_n(x_1) < \varepsilon/4$$

whenever $x_2 - x_1 < \delta$ on $[a_n, c_n]$, and the desired conclusion follows.

We thus conclude in all cases that there exists a subsequence of $\{\phi_n\}$ which satisfies the hypotheses of Lemma 4.6 and for which $\{\phi_n'\}$ is uniformly bounded on each compact subset of J. To apply Ascoli's theorem to $\{\phi_n'\}$, we need to know that $\{\phi_n'\}$ is equicontinuous on each compact subinterval J_1 of J. Since $\{\phi_n'\}$ is uniformly bounded on J_1 and $g(x, z)$ is continuous, $\{g(x, \phi_n'(x))\}$ is uniformly bounded on J_1. Using the differential equation, there then exists a constant C so that

$$|\phi_n''(x)| < C + f(x, \phi_N(x)) \quad \text{for } x \in J_1.$$

Using Lemmas 4.2 and 4.4, we may integrate this estimate to show that $\{\phi_n'\}$ is equicontinuous on J_1. The usual diagonalization argument now produces a further subsequence of $\{\phi_n\}$ which converges uniformly on $[a, b]$ and for which $\{\phi_n'\}$ converges uniformly on each compact subset of J. It is easy to see that the limit function ϕ satisfies all the desired properties, finally completing the proof.

5. Uniqueness. If $b_0 = 0$ in (1.12), then we encounter added technical problems in establishing uniqueness. So we begin with the case where $b_0 > 0$.

THEOREM 5.1. *Suppose that f satisfies* (HF1)-(HF2) *and that g satisfies* (HG1), (HG3), *and suppose that $g(x, z)$ is nondecreasing in z for each $x \in [a, b]$. If $b_0 > 0$, then the boundary value problem* (1.11), (1.12) *has at most one positive solution.*

Proof. Assume that (1.11), (1.12) has two distinct solutions ϕ, ψ and let $u = \phi - \psi$. We may suppose that u has a positive maximum at some $x_0 \in [a, b]$. Since u satisfies the boundary conditions

$$a_0 u(a) - a_1 u'(a) = 0, \qquad b_0 u(b) + b_1 u'(b) = 0, \tag{5.1}$$

then u cannot be a positive constant on $[a, b]$. The argument of Lemma 2.1 may be repeated to show that x_0 is not an interior point of $[a, b]$. Since $b_0 > 0$, the second of the conditions (5.1) shows that the positive maximum cannot occur at b. Thus $x_0 = a$ and from the first of conditions (5.1), we see that necessarily $a_0 = 0$ and $u'(a) = 0$. It is easy to see that $u(x) \geqq 0$ on $[a, b]$. Otherwise, $-u$ would have a positive maximum on $(a, b]$ and this is impossible by the same argument just used to show that u cannot have a positive maximum on $(a, b]$. Then $u'(x) \leqq 0$ on $[a, b]$ since otherwise we see that u has a positive maximum on $(a, b]$. Since $g(x, z)$ is nondecreasing in z, (HF2) implies

$$u''(x) = g(x, \psi'(x)) - g(x, \phi'(x)) + f(x, \psi(x)) - f(x, \phi(x)) \geqq 0$$

on (a, b), forcing u to be constant on $[a, b]$, which is impossible.

In our proof above, we used the assumption that $b_0 > 0$ to rule out a maximum at b. If $b_0 = 0$, we need some further hypothesis. Here are several possibilities, any one of which suffices:

(5.2) $f(x, y)$ satisfies (HF0),

(5.3) $g(x, z) \equiv 0$,

(5.4) $\liminf_{x \to b^-} [f(x, y_2) - f(x, y_1)] > 0, \qquad 0 < y_2 < y_1$.

THEOREM 5.2. *Suppose $f(x, y)$ satisfies* (HF1)-(HF2), *$g(x, z)$ satisfies* (HG1), (HG3), *and also $g(x, z)$ is nondecreasing in z for each $x \in [a, b]$. If $b_0 = 0$ and any one of* (5.2), (5.3), (5.4) *is satisfied, then the boundary value problem* (1.11), (1.12) *has at most one solution.*

Proof. The proof of Theorem 5.1 may be used if we can rule out a positive maximum for u at b. Assuming u has a positive maximum at b, the second condition in (5.1) now implies that $u'(b) = 0$. If (5.2) holds, the maximum principle applies as in the proof of Lemma 2.1 to show that $u'(b) > 0$, a contradiction. If (5.3) holds, we may assume that $u(x) > 0$ and nonconstant on some interval $[c, b]$ where $a < c < b$. Then

$$u''(x) = f(x, \psi(x)) - f(x, \phi(x)) \geqq 0 \quad \text{for } c \leqq x \leqq b,$$

and since $u'(x)$ must be positive somewhere in (c, b), we get $u'(b) > 0$, a contradiction. Finally if (5.4) is true, it is easy to see that $u''(x) > 0$ in some neighborhood of b and we again reach the contradiction that $u'(b) > 0$.

It seems likely that without some additional hypothesis, Theorem 5.1 does not remain true in the case that $b_0 = 0$.

6. Existence and uniqueness with $g(x, y')$ singular. We now turn attention to the problem (1.11), (1.13), allowing $g(x, z)$ to be singular at $x = a$. We continue to assume that f satisfies (HF1)-(HF2), but relax the hypotheses (HG1)-(HG3) to the following.

(HG1*) $g : (a, b] \times (-\infty, 0] \to (-\infty, 0]$ is continuous.

(HG2*) $g(x, 0) = 0$, for all $x \in (a, b]$.

(HG3*) g satisfies a uniform Lipschitz condition in z on each compact set $S \subset (a, b) \times (-\infty, 0]$.

THEOREM 6.1. *Suppose f, g satisfy* (HF1)-(HF2), (HG1*)-(HG3*), *and that f satisfies the strong integrability condition at $x = a$. Suppose that $f(x, y)$ satisfies the strong integrability condition at $x = b$ and let $J = [a, b]$, or that b is a Dirichlet endpoint and $f(x, y)$ satisfies the weak integrability condition at b and let $J = [a, b)$. Then the boundary value problem* (1.11), (1.13) *has at least one positive solution $\phi \in C^1(J) \cap C^2(a, b)$.*

Proof. For each integer $n > 1/(b - a)$, consider the boundary value problem

$$y'' + \hat{g}(x, y') + f(x, y) = 0, \qquad a + 1/n < x < b,$$

(BVP(n)) $$y'(a + 1/n) = 0,$$

$$b_0 y(b) + b_1 y'(b) = B,$$

where $\hat{g}(x, y) \equiv -g(x, -y)$, for $y > 0$, $\hat{g}(x, y) \equiv g(x, y)$, for $y \leqq 0$. We apply Theorem 4.1 with $A = 0$, $a_1 = 1$, $b_0 > 0$ to conclude the existence of a solution $\phi_n \in C^2(a + 1/n, b) \cap C^1([a + 1/n, b] \cap J) \cap C[a + 1/n, b]$ of (BVP(n)) for which

(6.1) $$|\phi_n'(x)| \leqq 2 \int_{a+1/n}^{x} f(s, \phi_n(s))\, ds < \infty$$

for $a+1/n \leqq x \leqq b$. Using the facts that $\phi_n'(a+1/n)=0$, $\phi_{n+1}'(a+1/n)<0$ and repeating the argument of Lemma 3.2, it is easy to see that $\phi_n(x) \leqq \phi_{n+1}(x)$ for $a+1/n \leqq x \leqq b$. Let N be the smallest integer larger than $1/(b-a)$ and put

$$\phi_0(x)=\begin{cases}\phi_N(x), & a+1/N \leqq x \leqq b,\\ \phi_N(a+1/N), & a \leqq x < a+1/N.\end{cases}$$

Since ϕ_n is decreasing on $[a+1/n, b]$, it follows that

$$\phi_n(x) \geqq \phi_0(x) \quad \text{for } a+1/n \leqq x \leqq b, \quad n \geqq N,$$

and (HF2) implies

$$|\phi_n'(x)| \leqq 2\int_a^x f(s, \phi_0(s))\, ds < \infty, \qquad a+\frac{1}{n} < x < b. \tag{6.2}$$

Thus we have a uniform bound for $\{\phi_n'\}$ on compact subsets of $J \cap (a, b]$ and integration of (6.2) produces a uniform bound for $\{\phi_n\}$ on $[a, b]$. Using Ascoli's theorem and the usual diagonalization argument, we complete the proof to get a solution ϕ of the differential equation on $(a, b]$ which satisfies the boundary condition at $x = b$. It remains to show that $\phi'(a)=0$. We may let $n \to \infty$ in (6.2) to get

$$|\phi'(x)| \leqq 2\int_a^x f(s, \phi_0(s))\, ds < \infty, \qquad a < x < b.$$

It follows that $\lim_{x \to a^+} \phi'(x)=0$. Since ϕ is monotone on $(a, b]$ and bounded, then $\phi(a) \equiv \lim_{x \to a^+} \phi(x)$ exists, and the mean value theorem shows that $\phi'(a)=0$.

The following uniqueness theorem is proved using the same argument as we used in Theorem 5.1.

THEOREM 6.2. *Suppose that f satisfies* (HF1)-(HF2) *and that g satisfies* (HG1*), (HG3*), *and suppose that $g(x, z)$ is nondecreasing in z for each $x \in (a, b]$. Then the boundary value problem* (1.11), (1.13) *has at most one positive solution.*

Acknowledgments. The author is indebted to Paul Waltman and the hospitality of his colleagues at Emory.

REFERENCES

[1] J. V. BAXLEY, *Nonlinear second order boundary value problems on* $[0, \infty)$, in Qualitative Properties of Differential Equations, W. Allegretto and G. J. Butler, eds., Proc. 1984 Edmonton Conference, University of Alberta Press, Edmonton, Alberta, Canada, 1986, pp. 50-58.

[2] ———, *A singular nonlinear boundary value problem: membrane response of a spherical cap*, SIAM J. Appl. Math., 48 (1988), pp. 497-505.

[3] ———, *Existence theorems for second order nonlinear boundary value problems*, J. Differential Equations, 85 (1990), pp. 125-150.

[4] H. BERESTYCKI, P. L. LIONS, AND L. A. PELETIER, *An* ODE *approach to the existence of positive solutions for semilinear problems in* R^N, Indiana Univ. Math. J., 30 (1981), pp. 141-157.

[5] L. ERBE AND K. SCHMITT, *On radial solutions of some semilinear elliptic equations*, Differential and Integral Equations, 1 (1988), pp. 71-78.

[6] J. GATICA, G. HERNANDEZ, AND P. WALTMAN, *Radially symmetric solutions of a class of singular elliptic equations*, Proc. Edinburgh Math. Soc., 33 (1990), pp. 169-180.

[7] J. GATICA, V. OLIKER, AND P. WALTMAN, *Singular nonlinear boundary value problems for second order ordinary differential equations*, J. Differential Equations, 79 (1989), pp. 62-78.

[8] A. GRANAS, R. B. GUENTHER, J. LEE, AND D. O'REGAN, *Boundary value problems on infinite intervals and semiconductor devices*, J. Math. Anal. Appl., 116 (1986), pp. 335-348.

[9] C. LUNING AND W. PERRY, *Positive solutions of negative exponent generalized Emden-Fowler boundary value problems*, SIAM J. Appl. Math., 12 (1981), pp. 874-879.

[10] T. Y. NA, *Computational Methods in Engineering Boundary Value Problems*, Academic Press, New York, 1979.

[11] A. NACHMAN AND A. CALLEGARI, *A nonlinear boundary value problem in the theory of pseudoplastic fluids*, SIAM J. Appl. Math., 11 (1980), pp. 275–281.

[12] M. PROTTER AND H. WEINBERGER, *Maximum principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1968.

[13] S. TALIAFERRO, *A nonlinear singular boundary value problem*, Nonlinear Anal. Theory Methods Appl., 3 (1979), pp. 897–904.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 480-499, March 1991

012

INVERSION OF DISCONTINUITIES FOR THE SCHRÖDINGER EQUATION IN THREE DIMENSIONS*

LASSI PÄIVÄRINTA† AND ERKKI SOMERSALO†

Abstract. This work deals with the inverse scattering problem for the Schrödinger operator in three dimensions. The following problem is studied: If the inverse scattering problem is solved approximately by using the linearizing Born approximation, what information about the true potential is obtained if the scatterer is not necessarily weak? It is shown that in a certain sense, the leading singularities of the potential are recovered exactly. More accurately, under certain a priori smoothness assumptions of the scattering potential, the difference of the potential and the one obtained by using the Born approximation is in a smoother class of functions. Especially, for bounded potentials the approximation agrees with the true potential up to a Lipschitz continuous function. For the a priori scale of function spaces we have chosen the Zygmund classes equipped with a suitable weight at infinity.

Key words. Schrödinger equation, Born approximation, inverse scattering

AMS(MOS) subject classifications. 35R30, 35J10, 81C05

Introduction. One of the most straightforward and widely used methods of approaching the multidimensional inverse scattering problems on various fields of physics is to use the Born approximation to linearize the typically nonlinear problem. The popularity of this approach is no doubt in the relative ease at which linear inverse problems are treated, both theoretically and numerically, as compared to nonlinear ones, where there are few standard methods available. Naturally, the problems then lie in the validation of the linearizing approximation. The general philosophy is that for weak potentials, or scatterers in general, the scattered field is well approximated by the Born approximation. Application to inverse scattering problems contains in principle a more difficult problem: Given a scattering amplitude of other scattering data, how do we decide whether the unknown scatterer is weak enough to justify the linearized scheme? This work is intended to contribute to this problem in the special case of three-dimensional inverse scattering of the Schrödinger equation. Explicitly, we ask the following question: Given the far field amplitude of the scattering solution of the Schrödinger equation, what information of the potential is retrieved if we blindly apply the Born inversion scheme to this data? The main result of the work is that, roughly, the Born inversion yields the singularities of the original potential exactly, up to a certain degree of smoothness. Similar results have been known for some time in other scattering configurations. One of the pioneering works that should be mentioned here is the important article by Beylkin [4], followed by a number of other works. The basic ideas of this work, however, come closer to those of [14]. In fact, it is not difficult to see that the results of [14] can be directly translated to the language of (one-dimensional) quantum scattering.

The plan of the present article is: In § 1 we discuss the background of the linearized inversion scheme and prove the main result (Theorem 1.8) of the work. In § 2 we show that in some cases it is possible to improve the general results of § 1. Especially, the calculus developed in § 1 is applied to show that the linearized inversion scheme is sufficient for inverting discontinuities of a bounded potential: The true potential is shown to differ from the approximation by a Lipschitz continuous function. There are two appendices. Appendix A is a quick reference on the results concerning certain

* Received by the editors July 17, 1989; accepted for publication (in revised form) April 2, 1990.
† Department of Mathematics, University of Helsinki, Hallituskatu 15, 00100 Helsinki, Finland.

function spaces which are used in the work. Finally, some of the most technical proofs are collected in Appendix B.

1. Asymptotic expansion of the linearized equations. To fix the notations, we will review some basic definitions and facts from the scattering theory of the Schrödinger equation in $\mathbf{R}^3$. The general reference on quantum scattering theory is, e.g., [9]. Let q be a real valued potential in $\mathbf{R}^3$ appearing in the Schrödinger operator

$$H = -\Delta + q(x). \tag{1.1}$$

Since the aim of this work is mainly to establish certain general ideas, we will avoid extra complications, at the cost of some generality, by assuming that the potential has no (infinite) singularities. More precisely, let w_δ be a weight function

$$w_\delta(x) = (1+|x|^2)^\delta.$$

The general assumption in this work is

$$w_\delta q \in L^\infty \tag{1.2}$$

for some $\delta > \frac{3}{2}$, abbreviated as $q \in L^\infty_\delta$. In most of this work, the assumption about the rate of decay at infinity could be released. Our choice of δ is made to gain some notational convenience (see Remark 1.10 below). Later, we will study the smoothing properties of certain operators, so we need a scale of function spaces generalizing (1.2) to measure the degree of smoothness in q. For $s > 0$ we define the weighted Zygmund space Λ^s_δ by the condition

$$f \in \Lambda^s_\delta \quad \text{if } w_\delta f \in \Lambda^s, \tag{1.3}$$

where again $\delta > \frac{3}{2}$, and Λ^s denotes the classical Zygmund space with smoothness index s. In Appendix A, we have collected the definitions of the Zygmund spaces and some other closely related function spaces that will be used in this work. Also, some important results that would take us too far off the main topic are left to this Appendix. The useful property of the spaces Λ^s, and the reason why we have chosen to work with these spaces, is that they are obtained via interpolation from the simple spaces C^k, the space of k times continuously differentiable functions. More precisely, we will prove in Appendix A the following interpolation property.

THEOREM 1.1. *Let k_0 and k_1 be integers, $0 \leqq k_0 < k_1$. Further, let $0 < \theta < 1$ and $s = (1-\theta)k_0 + \theta k_1$. Then, for any $\varepsilon > 0$,*

$$\Lambda^{s+\varepsilon} \subset [C^{k_0}, C^{k_1}]_\theta \subset \Lambda^s,$$

with continuous embeddings.

In the sequel, we will also need the weighted L^2 spaces L^2_δ, $\delta \in \mathbf{R}$, defined as

$$L^2_\delta = \left\{ f \in L^2_{\text{loc}} \,\middle|\, \|f\|_\delta = \left(\int_{\mathbf{R}^3} |f(x)|^2 (1+|x|^2)^\delta \, dx \right)^{1/2} < \infty \right\}.$$

The following fundamental result of the stationary scattering theory, usually referred to as the *limiting absorption principle*, can be found in [1], [9].

PROPOSITION 1.2. *Assume that the potential q satisfies* (1.2). *Then for all $k > 0$ there exist the limits*

$$\lim_{\varepsilon \to 0+} (H - (k^2 \pm i\varepsilon))^{-1} =: \mathcal{R}^\pm(k) \tag{1.4}$$

in the uniform operator topology from L^2_δ to $L^2_{-\delta}$ as $\delta > \frac{1}{2}$. Moreover, for large k,

$$\|\mathcal{R}^\pm(k)\| \leqq \frac{C}{k}. \tag{1.5}$$

Although not explicitly stated, the estimate (1.5) follows from the a priori estimate (A.2′) of reference [1] (see also [11], [13]).

It should be noted that in [1], the conditions on the potential q are less restrictive than (1.2). From (1.4) it is evident that the operators $\mathscr{R}^+(k)$ and $\mathscr{R}^-(k)$ are adjoint, i.e.,

$$(\mathscr{R}^+(k))^* = \mathscr{R}^-(k).$$

The family of generalized eigenfunctions $\psi^\pm$ of the Schrödinger operator H with outgoing (+) and incoming (−) radiation condition at infinity are now defined as

$$\psi^\pm(x, k, \theta) = \psi_0(x, k, \theta) - \mathscr{R}^\pm(k)(q\psi_0)(x, k, \theta), \tag{1.6}$$

where $k>0$, θ is a unit vector, i.e., $\theta \in S^2$, and ψ_0 is the incident plane wave,

$$\psi_0(x, k, \theta) = e^{ik\theta\cdot x}.$$

It is well known that the wave functions satisfy the Lippmann-Schwinger equation

$$\begin{aligned}\psi^\pm(x, k, \theta) &= e^{ik\theta\cdot x} - \frac{1}{4\pi}\int_{\mathbf{R}^3} \frac{e^{\pm ik|x-y|}}{|x-y|} q(y)\psi^\pm(y, k, \theta)\, dy \\ &= \psi_0(x, k, \theta) - \mathscr{R}_0^\pm(k) q\psi^\pm(x, k, \theta).\end{aligned} \tag{1.7}$$

Asymptotically, ψ^+ admit the expansion

$$\psi^+(x, k, \theta) = e^{ik\theta\cdot x} - \frac{e^{ik|x|}}{4\pi|x|} A(k, \hat{x}, \theta) + o\left(\frac{1}{|x|}\right),$$

as $|x|$ tends to infinity, and $\hat{x} = x/|x|$. The *scattering amplitude* A is obtained as

$$A(k, \hat{x}, \theta) = \int_{\mathbf{R}^3} e^{-ik\hat{x}\cdot y} q(y)\psi^+(y, k, \theta)\, dy. \tag{1.8}$$

In [7], an estimate of the decay rate of the residual term is given in terms of the decay rate of the potential at infinity (see also [8]). For reasons of purely technical nature, we define the wave functions $\psi^\pm(x, k, \theta)$ for negative values of k also by extending the Lippmann-Schwinger equation (1.7) to negative k's. Then

$$\psi^+(x, -k, \theta) = \overline{\psi^+(x, k, \theta)} = \psi^-(x, k, -\theta). \tag{1.9}$$

Hence we can also extend (1.8) to negative values of k, and

$$A(-k, \hat{x}, \theta) = \overline{A(k, \hat{x}, \theta)} = A(-k, -\theta, -\hat{x}). \tag{1.10}$$

The latter equality is simply the well-known reciprocity relation. The classical inverse scattering problem is to reconstruct the potential q from the knowledge of the far field data $A(k, \hat{x}, \theta)$, when k, $\hat{x}$, and θ are restricted to some given set.

To state the linearized inverse scattering equations in a compact form, we introduce the cylinders $M_0 = \mathbf{R}\times S^2$ and $M = M_0 \times S^2$, S^2 being the surface of the unit ball, and the measures μ_θ and μ on M_0 and M, respectively, as

$$d\mu_\theta(k, \theta') = \frac{1}{4} k^2\, dk |\theta - \theta'|^2\, d\theta',$$

$$d\mu(k, \theta', \theta) = \frac{1}{4\pi}\, d\theta\, d\mu_\theta(k, \theta').$$

Here, $d\theta$ and $d\theta'$ denote the usual Lebesgue measure on S^2. Since the spaces M_0 and M will be treated as the Fourier spaces of $\mathbf{R}^3$, we will define the equivalent of the

usual inverse Fourier transform on M_0 and M. If $\varphi: M_0 \to \mathbf{C}$ is a smooth, rapidly decreasing function (with respect to $k(\theta - \theta')$), henceforth denoted as $\varphi \in S(M_0)$, we set

$$\mathcal{F}_{M_0}^{-1}\varphi(x) = \left(\frac{1}{2\pi}\right)^3 \int_{M_0} e^{-ik(\theta-\theta')\cdot x}\varphi(k, \theta')\, d\mu_\theta(k, \theta'), \qquad x \in \mathbf{R}^3.$$

Obviously, $\mathcal{F}_{M_0}^{-1}$ maps $S(M_0)$ to $S(\mathbf{R}^3)$. Similarly, we define $\mathcal{F}_M^{-1}: S(M) \to S(\mathbf{R}^3)$ by

$$\mathcal{F}_M^{-1}\varphi(x) = \left(\frac{1}{2\pi}\right)^3 \int_M e^{-ik(\theta-\theta')\cdot x}\varphi(k, \theta', \theta)\, d\mu(k, \theta', \theta).$$

Let u_θ be the coordinate mapping $M_0 \to \mathbf{R}^3$ given as

$$u_\theta(k, \theta') = k(\theta - \theta'), \tag{1.11}$$

where θ is considered as a fixed parameter. If we write $\xi = k(\theta - \theta')$, then k and θ' are obtained back as

$$k = \frac{|\xi|}{2\theta \cdot \hat{\xi}}, \qquad \theta \cdot \hat{\xi} \neq 0, \tag{1.12}$$
$$\theta' = \theta - 2(\theta \cdot \hat{\xi})\hat{\xi},$$

where $\hat{\xi} = \xi/|\xi|$. It is a simple matter to check that the measures μ_0 and μ are chosen to be compatible with the coordinate mapping u_θ in the sense that if $\varphi \in S(\mathbf{R}^3)$, then

$$\int_{M_0} \varphi \circ u_\theta(k, \theta')\, d\mu_\theta(k, \theta') = \int_M \varphi \circ u_\theta(k, \theta')\, d\mu(k, \theta, \theta') = \int_{\mathbf{R}^3} \varphi(x)\, dx.$$

Especially, for φ as above,

$$\mathcal{F}_{M_0}^{-1}(\varphi \circ u_\theta) = \mathcal{F}_M^{-1}(\varphi \circ u_\theta) = \mathcal{F}^{-1}\varphi,$$

the usual inverse Fourier transform in $\mathbf{R}^3$.

By duality, the operators $\mathcal{F}_{M_0}^{-1}$ and $\mathcal{F}_M^{-1}$ are extended to the spaces of tempered distributions $S'(M_0)$ and $S'(M)$, i.e., if $\varphi \in S(\mathbf{R}^3)$ and $T \in S'(M)$, then

$$\langle \mathcal{F}_M^{-1}T, \varphi\rangle_{\mathbf{R}^3} = \langle T, \mathcal{F}^{-1}\varphi(k(\theta - \theta'))\rangle_M.$$

Before the next definition, which is central in the whole work, some comments are in order. The measures μ_θ and μ as well as their compatibility with the change of variables (1.11), (1.12) has been found by several authors (see, e.g., [10] and references therein). An interesting generalization of the measure μ in M has been given recently in [5].

DEFINITION 1.3. The linearized inverse scattering equations or inverse Born approximations of q are defined as

$$\tilde{q}_\theta(x) = \mathcal{F}_{M_0}^{-1}(A(k, \theta', \theta))(x), \qquad \tilde{q}(x) = \mathcal{F}_M^{-1}(A(k, \theta', \theta))(x)$$

interpreted in the sense of distributions, where A is given as in (1.8).

The distribution $\tilde{q}_\theta$ is considered as an approximate solution to the inverse scattering problem when $A(k, \theta', \theta)$ is known with θ fixed. Similarly, in $\tilde{q}$ it is assumed that the data consist of the whole of $A(k, \theta', \theta)$.

Note that by using (1.10), we can write the definitions of $\tilde{q}_\theta$ and $\tilde{q}$ as

$$\tilde{q}_\theta(x) = \frac{1}{32\pi^3}\int_0^\infty k^2\, dk \int_{S^2} |\theta - \theta'|^2\, e^{-ik(\theta-\theta')\cdot x}(A(k, \theta', \theta) + \overline{A(k, -\theta, -\theta')}),$$

$$\tilde{q}(x) = \frac{1}{64\pi^4}\operatorname{Re}\left(\int_0^\infty k^2\, dk \int_{S^2\times S^2} d\theta\, d\theta' |\theta - \theta'|^2\, e^{-ik(\theta-\theta')\cdot x}A(k, \theta', \theta)\right).$$

These representations show that in the Definition 1.3 we use only physical data, and moreover, $\tilde{q}$ is real.

The above definitions are by no means new. Heuristically, they can be arrived at in a number of different ways (cf. [15]). One very straightforward approach is to note that if the scattering potential q is weak, we have approximately

$$\psi^+(x, k, \theta) = e^{ik\theta\cdot x} + O(q) \approx e^{ik\theta\cdot x},$$

and consequently

$$A(k, \theta', \theta) \approx \int_{\mathbf{R}^3} e^{ik(\theta-\theta')\cdot x} q(x)\, dx = \mathcal{F}q(k(\theta-\theta')).$$

Hence a natural candidate to approximate a weak potential q is, with an appropriate choice of variables, the inverse Fourier transform of A, which is the contents of Definition 1.3.

The objective of this work is to establish certain connections between the approximate solutions given above and the true potential q. We start with a simple uniqueness result. While the result in itself is not very deep, as it turns out, the formulation of the proof contains a key observation of the later discussion.

PROPOSITION 1.4. *For potentials q satisfying* (1.2) *with $\delta > \frac{3}{2}$, the knowledge of $\tilde{q}_\theta$ with θ restricted to a one-dimensional semicircle, defines q uniquely.*

Proof. Writing the definition of $\tilde{q}_\theta$ together with formula (1.8) for the scattering amplitude A we get

$$\begin{aligned}\tilde{q}_\theta(x) &= \left(\frac{1}{2\pi}\right)^3 \int_{M_0} e^{-ik(\theta-\theta')\cdot x} A(k, \theta', \theta)\, d\mu_\theta(k, \theta') \\ &= \left(\frac{1}{2\pi}\right)^3 \int_{M_0} d\mu_\theta(k, \theta') \int_{\mathbf{R}^3} dy\, e^{-ik(\theta-\theta')\cdot(x-y)} q(y) v(y, k, \theta),\end{aligned}$$

where $v(y, k, \theta) = e^{-ik\theta\cdot y}\psi^+(y, k, \theta)$. After the change of variables (1.12), we have

$$\tilde{q}_\theta(x) = \left(\frac{1}{2\pi}\right)^3 \int_{\mathbf{R}^3} d\xi \int_{\mathbf{R}^3} dy\, e^{-i\xi\cdot(x-y)} q(y) v\left(y, \frac{|\xi|}{2\hat{\xi}\cdot\theta}, \theta\right). \tag{1.13}$$

Hence, the Fourier transform of $\tilde{q}_\theta$ is simply

$$\begin{aligned}\mathcal{F}\tilde{q}_\theta(\xi) &= \int_{\mathbf{R}^3} dy\, e^{i\xi\cdot y} q(y) v\left(y, \frac{|\xi|}{2\hat{\xi}\cdot\theta}, \theta\right) \\ &= \mathcal{F}q(\xi) + \int_{\mathbf{R}^3} dy\, e^{i\xi\cdot y} q(y)\left(v\left(y, \frac{|\xi|}{2\hat{\xi}\cdot\theta}, \theta\right) - 1\right).\end{aligned}$$

Note that by assumption (1.2), q is in L^1 and thus $\mathcal{F}q$ is continuous. By (1.7) and Proposition 1.2 we get for $\delta > \frac{1}{2}$

$$\|v - 1\|_{L^2_{-\delta}} = \|\mathcal{R}^+(k)(q\psi_0)\|_{L^2_{-\delta}} \leqq \frac{C}{|k|}\|q\|_{L^2_\delta}.$$

Thus, as $\hat{\xi}\cdot\theta$ approaches zero, we have

$$\begin{aligned}\left|\int_{\mathbf{R}^3} dy\, e^{i\xi\cdot y} q(y)\left(v\left(y, \frac{|\xi|}{2\hat{\xi}\cdot\theta}, \theta\right) - 1\right)\right| &\leqq \|q\|_{L^2_\delta}\left\|v\left(\cdot, \frac{|\xi|}{2\hat{\xi}\cdot\theta}, \theta\right) - 1\right\|_{L^2_{-\delta}} \\ &\leqq C\|q\|^2_{L^2_\delta}\frac{|2\hat{\xi}\cdot\theta|}{|\xi|} \to 0.\end{aligned}$$

But the data $\{\mathcal{F}q(\xi) \mid \hat{\xi} \cdot \theta = 0\}$, when θ runs through a semicircle, is enough for the complete recovery of q. □

It is not hard to see that the above theorem is simply a restatement of the well-known fact that the high-frequency limit of the scattering amplitude determines the potential uniquely (see, e.g., [13]). In fact, the band limited data $\{A(k, \theta', \theta) \mid \theta' \in S^2, |k| \leqq M\}$ determines $\mathcal{F}\tilde{q}_\theta$ only in the so-called Ewald spheres $\{\xi \mid |\xi \pm M\theta| \leqq M\}$ that touch the plane $\xi \cdot \theta = 0$ only at the origin, no matter how large M is. The important point, however, is (1.13). If we momentarily forget the q dependence of the function v, (1.13) says that $\tilde{q}_\theta$ is obtained from q by applying a pseudodifferential operator with the amplitude $v(y, |\xi|/2\hat{\xi} \cdot \theta, \theta)$ on q. Also, from the proof of Proposition 1.4, it is evident that the principal part of this symbol is 1. In the discussion that follows, we will use this observation.

From now on, we will state all the results for $\tilde{q}$ only. Most of the results remain valid for $\tilde{q}_\theta$, too. The approximation $\tilde{q}$ is chosen to make some formulas more symmetric.

In the sequel, the following notation will be used. Let $\mathcal{K}(k)$ denote the operator

$$\mathcal{K}(k)\varphi(x) = \int_{\mathbf{R}^3} |q(x)|^{1/2} G_k^+(x-y) q(y)^{1/2} \varphi(y)\, dy,$$

where G_k^+ is the kernel of the operator $\mathcal{R}_0^+(k)$, i.e.,

$$G_k^+(x) = -\frac{e^{ik|x|}}{4\pi|x|},$$

and in the commonly used way, $q(y)^{1/2} = \operatorname{sign} q(y)\, |q(y)|^{1/2}$. It is known by Proposition 1.2 that $\mathcal{K}(k)$ is a continuous mapping in L^2 with the norm estimate

$$\|\mathcal{K}(k)\| \leqq \frac{C}{|k|}$$

for large $|k|$. Moreover, $\mathcal{K}(k)$ is a Hilbert–Schmidt operator, since its kernel, also denoted by $\mathcal{K}$, satisfies

$$\|\mathcal{K}(k)\|_{\mathrm{HS}} = \left(\int_{\mathbf{R}^3} \int_{\mathbf{R}^3} |\mathcal{K}(x, y)|^2\, dx\, dy \right)^{1/2} = \frac{1}{4\pi} \left(\int_{\mathbf{R}^3} \int_{\mathbf{R}^3} \frac{|q(x)||q(y)|}{|x-y|^2}\, dx\, dy \right)^{1/2}$$

$$= \|q\|_{\mathrm{R}} < \infty,$$

$\|q\|_{\mathrm{R}}$ denoting the Rollnik norm of q. Further, let $\Phi_0(k)$ be the operator

$$\Phi_0(k) f(x) = |q(x)|^{1/2} \int_{S^2} e^{ik\theta \cdot x} f(\theta)\, d\theta,$$

and similarly, $\Phi(k)$ defined as

$$\Phi(k) f(x) = |q(x)|^{1/2} \int_{S^2} \psi^+(x, k, \theta) f(\theta)\, d\theta,$$

both defined for square integrable functions on S^2. The following theorem is similar to Theorem 1.2(ii) in [13].

THEOREM 1.5. *The operators $\Phi_0(k)$ and $\Phi(k)$ are continuous operators from $L^2(S^2)$ to $L^2(\mathbf{R}^3)$ with norm bounded by $C/|k|$ as $|k|$ is large.*

Theorem 1.5 is a straightforward consequence of the following lemma, which is a version of the well-known optical theorem. For completeness, we include a simple proof of it, following the lines of [12].

LEMMA 1.6. *Assume that q is a potential satisfying* (1.2) *and f belongs to $S(\mathbf{R}^3)$. Then the function $v = \mathcal{R}^+(k)f$ solves the nonhomogenous Schrödinger equation*

$$(-\Delta + q - k^2)v = f \tag{1.14}$$

with the outgoing radiation condition at infinity, i.e.,

$$v(x, k) = \frac{e^{ik|x|}}{4\pi|x|} A_f(k, \hat{x}) + h(x, k), \qquad h(x, k) = o\left(\frac{1}{|x|}\right).$$

Furthermore, for the far-field pattern $A_f(k, \hat{x})$ we have

$$\int_{S^2} |A_f(k, \hat{x})|^2 \, d\hat{x} = -\frac{4\pi}{k} \operatorname{Im} \int_{\mathbf{R}^3} v(x, k)\overline{f(x)} \, dx.$$

Proof. Equation (1.14) follows trivially from the definition of v. To verify the asymptotic behavior of this solution at infinity, we employ the resolvent equation

$$\mathcal{R}^+(k) = \mathcal{R}_0^+(k) - \mathcal{R}_0^+(k)q\mathcal{R}^+(k),$$

proved in [1]. Since $f \in S(\mathbf{R}^3)$, we then have, by the estimates given in [7],

$$\begin{aligned}\mathcal{R}^+(k)f(x) &= \mathcal{R}_0^+(k)(f - q\mathcal{R}^+(k)f)(x) \\ &= \frac{e^{ik|x|}}{4\pi|x|} \int_{\mathbf{R}^3} e^{-ik\hat{x}\cdot y}(f(y) - q(y)\mathcal{R}^+(k)f(y)) \, dy + o\left(\frac{1}{|x|}\right),\end{aligned}$$

since $\mathcal{R}^+(k)f$ is bounded. Hence, v has the claimed asymptotic form at infinity.

Let ρ be a smooth cutoff function on $[0, \infty)$, i.e., $0 \leqq \rho \leqq 1$, $\rho(r) = 1$ as $0 \leqq r < 1$ and $\rho(r) = 0$ as $r \geqq 2$. We set $\rho_n(r) = \rho(r/n)$, for integers n. By multiplying (1.14) with $\overline{v(x)}\rho_n(|x|)$, integrating over $\mathbf{R}^3$ and taking imaginary parts, we get

$$\operatorname{Im} \int_{\mathbf{R}^3} \rho_n(|x|)f(x)\overline{v(x, k)} \, dx = \operatorname{Im} \int_{\mathbf{R}^3} -\Delta v(x, k)\rho_n(|x|)\overline{v(x, k)} \, dx.$$

As n tends to infinity, the left-hand side converges to $\int f(x)\overline{v(x, k)} \, dx$. To get the desired limit for the right-hand side, we integrate by parts to get

$$\begin{aligned}&\operatorname{Im} \int_{\mathbf{R}^3} -\Delta v(x, k)\rho_n(|x|)\overline{v(x)} \, dx \\ &= \operatorname{Im} \int_{\mathbf{R}^3} \hat{x} \cdot \nabla v(x, k)\rho_n'(|x|)\overline{v(x, k)} \, dx \\ &= \operatorname{Im} \int_{\mathbf{R}^3} (\hat{x} \cdot \nabla v(x, k) - ikv(x, k))\rho_n'(|x|)\overline{v(x, k)} \, dx + k \int_{\mathbf{R}^3} \rho_n'(|x|)|v(x, k)|^2 \, dx.\end{aligned}$$

Since v satisfies the Sommerfeld radiation condition (see [7] and references therein), the first term tends to zero, whereas the second term is

$$\begin{aligned}k \int_{\mathbf{R}^3} \rho_n'(|x|)|v(x, k)|^2 \, dx &= \frac{k}{(4\pi)^2} \int_{S^2} |A_f(k, \hat{x})|^2 \, d\hat{x} \int_n^{2n} \rho_n'(r) \, dr + o(1) \\ &= k/(4\pi)^2 \int_{S^2} |A_f(k, \hat{x})|^2 \, dx(\rho(2) - \rho(1)) + o(1),\end{aligned}$$

yielding the claim. □

To obtain the result in Theorem 1.5, let $\mathcal{A}_q(k)$ denote the linear mapping that takes the inhomogeneity f to the corresponding scattering amplitude, i.e.,

$$\mathcal{A}_q(k): f(x) \mapsto A_f(k, \hat{x}).$$

By Lemma 1.6, we have

$$\|\mathcal{A}_q(k)f\|^2_{L^2(S^2)} \leqq \frac{C}{|k|}\|v\|_{L^2_{-\delta}}\|f\|_{L^2_\delta} \leqq \frac{C}{|k|^2}\|f\|^2_{L^2_\delta},$$

so the operator $\mathcal{A}_q(k)$ extends continuously to the space L^2_δ. From the proof of Lemma 1.6 we can also read off the integral kernel of the operator $\mathcal{A}_q(k)$. Indeed, for $k>0$,

$$\begin{aligned}\mathcal{A}_q(k)f(\hat{x}) &= \int_{\mathbf{R}^3} e^{-ik\hat{x}\cdot y}(f(y) - q(y)\mathcal{R}^+(k)f(y))\,dy \\ &= \langle \psi_0(\cdot, k, \hat{x}), f - q\mathcal{R}^+(k)f\rangle \\ &= \langle \psi^-(\cdot, k, \hat{x}), f\rangle \\ &= \int_{\mathbf{R}^3} \psi^+(y, k, -\hat{x})f(y)\,dy,\end{aligned}$$

the last equality following from (1.9). By a density argument, this holds for all $f \in L^2_\delta$. Thus, if we denote by $\mathcal{A}_q(k)^t: L^2(S^2) \to L^2_{-\delta}(\mathbf{R}^3)$ the transpose of $\mathcal{A}_q(k)$, and by $Q: L^2(S^2) \to L^2(S^2)$ the operator $Qf(\hat{x}) = f(-\hat{x})$, we obtain

$$\Phi(k): L^2(S^2) \xrightarrow{\mathcal{A}_q(k)^t Q} L^2_{-\delta}(\mathbf{R}^3) \xrightarrow{|q|^{1/2}} L^2(\mathbf{R}^3),$$

with the norm estimate as claimed. The negative values of k are treated in the same way. Similarly,

$$\Phi_0(k) = |q|^{1/2}\mathcal{A}_0(k)^t Q.$$

Later we will use the notation

$$\tilde{\Phi}_0(k) = Q\mathcal{A}_0(k)q^{1/2} = \Phi_0(k)^t \operatorname{sign} q: L^2(\mathbf{R}^3) \to L^2(S^2).$$

Consider next the Lippmann–Schwinger equation (1.6), which after n iterations gives

$$\psi^+ = \sum_{j=0}^{n} (\mathcal{R}_0^+(k)q)^j\psi_0 + (\mathcal{R}_0^+(k)q)^{n+1}\psi^+.$$

Substituting this expression into the integral (1.14) defining the scattering amplitude $A(k, \theta', \theta)$ we get

$$\begin{aligned}A(k, \theta', \theta) = \sum_{j=0}^{n} &\int_{\mathbf{R}^3} e^{-ik\theta'\cdot y} q(y)(\mathcal{R}_0^+(k)q)^j\psi_0(y, k, \theta)\,dy \\ &+ \int_{\mathbf{R}^3} e^{-ik\theta'\cdot y} q(y)(\mathcal{R}_0^+(k)q)^{n+1}\psi^+(y, k, \theta)\,dy,\end{aligned}$$

or, interpreting A as an operator from $L^2(S^2)$ to itself with each $k \neq 0$ fixed, we get

$$A = \sum_{j=0}^{n} \tilde{\Phi}_0\mathcal{K}^j\Phi_0 + \tilde{\Phi}_0\mathcal{K}^{n+1}\Phi.$$

For brevity, the k dependence of the operators is suppressed. Hence, with some abuse of notation, we can write

$$\tilde{q} = \sum_{j=0}^{n} \mathcal{F}_M^{-1}\tilde{\Phi}_0\mathcal{K}^j\Phi_0 + \mathcal{F}_M^{-1}\tilde{\Phi}_0\mathcal{K}^{n+1}\Phi. \tag{1.15}$$

Above, $\tilde{\Phi}_0\mathcal{K}^j\Phi_0$ and $\tilde{\Phi}_0\mathcal{K}^{n+1}\Phi$ denote the kernels of the corresponding operators, and $\mathcal{F}_M^{-1}$ acts on this function of k, θ, and θ'. The following lemma gives estimates of the smoothness of the various terms in (1.15).

LEMMA 1.7. *Assume that q belongs to Λ_δ^s with $\delta > \frac{3}{2}$ and $s > 0$. (In the case $s = 0$ we assume that q is in the class L_δ^∞.) Then, for $j \geqq 1$, $\mathcal{F}_M^{-1}\tilde{\Phi}_0\mathcal{K}^j\Phi_0$ and $\mathcal{F}_M^{-1}\tilde{\Phi}_0\mathcal{K}^j\Phi$ are in the Sobolev space H^t for all $t < s + j - \frac{1}{2}$.*

Proof. We start with the elementary case when $w_\delta q$ is bounded. By using the change of variables (1.14), we get with $0 \leqq t < j - \frac{1}{2}$.

$$
\begin{aligned}
&\|(1+|\xi|^2)^{t/2}\mathcal{F}(\mathcal{F}_M^{-1}\tilde{\Phi}_0\mathcal{K}^j\Phi_0)(\xi)\|_{L^2(\mathbf{R}^3)}^2 \\
&\quad = \frac{1}{4\pi}\int_{\mathbf{R}^3}(1+|\xi|^2)^t \left|\int_{S^2}\tilde{\Phi}_0\mathcal{K}^j\Phi_0\left(\frac{|\xi|}{2\hat{\xi}\cdot\theta}, \theta - 2(\hat{\xi}\cdot\theta)\hat{\xi}, \theta\right) d\theta\right|^2 d\xi \\
&\quad \leqq \int_{\mathbf{R}^3}(1+|\xi|^2)^t \int_{S^2}\left|\tilde{\Phi}_0\mathcal{K}^j\Phi_0\left(\frac{|\xi|}{2\hat{\xi}\cdot\theta}, \theta - 2(\hat{\xi}\cdot\theta)\hat{\xi}, \theta\right)\right|^2 d\theta\, d\xi \qquad (1.16) \\
&\quad = 4\pi\int_M d\mu(k,\theta',\theta)(1+|k(\theta-\theta')|^2)^t|\tilde{\Phi}_0\mathcal{K}^j\Phi_0(k,\theta',\theta)|^2 \\
&\quad \leqq C\int_{-\infty}^{\infty} dk\, k^2(1+|k|^2)^t \int_{S^2} d\theta' \int_{S^2} d\theta|\tilde{\Phi}_0\mathcal{K}^j\Phi_0(k,\theta',\theta)|^2.
\end{aligned}
$$

Hence, it suffices to find an appropriately strongly converging bound for the Hilbert-Schmidt norm of the operator $\tilde{\Phi}_0\mathcal{K}^j\Phi_0$, as $|k|$ grows. But the class of Hilbert-Schmidt operators is a norm ideal, so by Proposition 1.5 and remarks before it we have the estimate

$$
\|\tilde{\Phi}_0\mathcal{K}^j\Phi_0\|_{\mathrm{HS}} \leqq \|\tilde{\Phi}_0\|\,\|\mathcal{K}\|_{\mathrm{HS}}\|\mathcal{K}\|^{j-1}\|\Phi_0\| \leqq \frac{C}{|k|^{j+1}}
$$

for large $|k|$. Especially, the integral (1.16) converges.

Next, assume that $w_\delta q$ belongs to C^k with $k > 0$ an integer. Then, by the particular form of the weight function w_δ,

$$
w_\delta|D^\alpha q| \in L^\infty \quad \text{for } 0 \leqq |\alpha| \leqq k. \qquad (1.17)
$$

To show that $\mathcal{F}_M^{-1}\tilde{\Phi}_0\mathcal{K}^j\Phi_0$ is in H^t for $t < k + j - \frac{1}{2}$ it is obviously sufficient to prove that $D^\alpha\mathcal{F}_M^{-1}\tilde{\Phi}_0\mathcal{K}^j\Phi_0$ which obviously equals to $\mathcal{F}_M^{-1}((k(\theta-\theta'))^\alpha\tilde{\Phi}_0\mathcal{K}^j\Phi_0)$ belongs to H^t, when $t < k - \frac{1}{2}$ and $0 \leqq |\alpha| \leqq j$. By writing $\tilde{\Phi}_0\mathcal{K}^j\Phi_0(k,\theta',\theta)$ explicitly as

$$
\begin{aligned}
\tilde{\Phi}_0\mathcal{K}^j\Phi_0 = \int_{\mathbf{R}^3} dx_1 \int_{\mathbf{R}^3} dx_2 \cdots \int_{\mathbf{R}^3} dx_{j+1}\, e^{-ik\theta'\cdot x_1} q(x_1) \\
\cdot G_k^+(x_1 - x_2)q(x_2)\cdots G_k^+(x_j - x_{j+1})q(x_{j+1})\, e^{ik\theta\cdot x_{j+1}}
\end{aligned}
$$

it is not hard to see that by successive integrations by parts we have

$$
\begin{aligned}
&(k(\theta-\theta'))^\alpha\tilde{\Phi}_0\mathcal{K}^j\Phi_0(k,\theta',\theta) \\
&\quad = \sum_{\alpha_1+\cdots+\alpha_{j+1}=\alpha}(-i)^{|\alpha|}\int_{\mathbf{R}^3} dx_1\cdots\int_{\mathbf{R}^3} dx_{j+1}\, e^{-ik\theta'\cdot x_1} D^{\alpha_1}q(x_1) \\
&\qquad \cdot G_k^+(x_1 - x_2)\cdots G_k(x_j - x_{j+1})D^{\alpha_{j+1}}q(x_{j+1})\, e^{ik\theta\cdot x_{j+1}}.
\end{aligned}
$$

By defining the operators $\Phi_{o,\alpha}$ and $\mathscr{K}_{\alpha,\beta}$ by the formulas

$$\Phi_{o,\alpha} f(x) = |D^\alpha q(x)|^{1/2} \int_{S^2} e^{ik\theta\cdot x} f(\theta)\, d\theta,$$

$$\mathscr{K}_{\alpha,\beta}\varphi(x) = \int_{\mathbf{R}^3} |D^\alpha q(x)|^{1/2} G_k^+(x-y)(D^\beta q(y))^{1/2}\varphi(y)\, dy,$$

we can write the above formula as

$$(k(\theta-\theta'))^\alpha \tilde{\Phi}_0 \mathscr{K}^j \Phi_0(k,\theta',\theta) = \sum_{(-i)^{|\alpha|}\alpha_1+\cdots+\alpha_{j+1}=\alpha} \tilde{\Phi}_{0,\alpha_1}\mathscr{K}_{\alpha_1,\alpha_2}\cdots\mathscr{K}_{\alpha_j,\alpha_{j+1}}\Phi_{0,\alpha_{j+1}}.$$

The same reasoning as in the first step, together with (1.17), now gives an appropriate estimate for the Hilbert–Schmidt norm of this operator and hence for the Sobolev norm.

Consider now the general case $w_\delta q \in \Lambda^s$, $s>0$. Denote by F_{j+1} the multilinear mapping defined as

$$\begin{aligned} &F_{j+1}(w_\delta q_1, w_\delta q_2, \cdots, w_\delta q_{j+1})(x) \\ &\quad = \int_M d\mu(k,\theta',\theta)\, e^{-ik(\theta-\theta')\cdot x} \int_{\mathbf{R}^3} dx_1 \int_{\mathbf{R}^3} dx_2 \cdots \int_{\mathbf{R}^3} dx_{j+1} \\ &\qquad \cdot e^{-ik\theta'\cdot x_1} q_1(x_1) G_k^+(x_1-x_2) q_2(x_2)\cdots G_k^+(x_j-x_{j+1}) q_{j+1}(x_{j+1})\, e^{ik\theta\cdot x_{j+1}}. \end{aligned}$$

This mapping is, by the same argument as above, a continuous mapping

$$F_{j+1}: \prod_{i=1}^{j+1} C^k \to H^{t_k}$$

for all integers $k \geqq 0$ and reals $t_k < k+j-\frac{1}{2}$. Assume that $0<t=s+j-\frac{1}{2}-\varepsilon$ for some $\varepsilon>0$. We shall choose integers k_0 and k_1 so that $0 \leqq k_0 < s < k_1$, and a parameter δ satisfying $0<\delta<\min(\varepsilon, s-k_0)$. Further, let θ be an interpolation parameter, $0<\theta<1$, with the property that $(1-\theta)k_0+\theta k_1 = s-\delta$. Since $\delta<\varepsilon$, we then have $t_n = k_n+j-\frac{1}{2}-(\varepsilon-\delta) < k_n+j-\frac{1}{2}$, where $n=0,1$. By the choice of θ, $(1-\theta)t_0+\theta t_1 = t$. Hence, invoking the multilinear complex interpolation theorem (see, e.g., [3]), we get for $(w_\delta q_1, w_\delta q_2, \cdots, w_\delta q_{j+1}) \in \prod \Lambda^s$ the estimate

$$\begin{aligned} \|F_{j+1}(w_\delta q_1, w_\delta q_2, \cdots, w_\delta q_{j+1})\|_{H^t} &= \|F_{j+1}(w_\delta q_1, w_\delta q_2, \cdots, w_\delta q_{j+1})\|_{[H^{t_0}, H^{t_1}]_\theta} \\ &\leqq C \prod_{i=1}^{j+1} \|w_\delta q_i\|_{[C^{k_0}, C^{k_1}]_\theta} \leqq C \prod_{i=1}^{j+1} \|w_\delta q_i\|_{\Lambda^s}. \end{aligned}$$

The last inequality follows from Theorem 1.1. Choosing $q_1 = q_2 = \cdots = q_{j+1} = q$ we get the desired result.

Finally, we have to establish the smoothness estimate for the residual term $\tilde{\Phi}_0 \mathscr{K}^{n+1}\Phi$. Obviously, as $s=0$, the same argument as in the first step above applies to this term too, so we have $\tilde{\Phi}_0\mathscr{K}^{n+1}\Phi \in H^t$ for $t<n+\frac{1}{2}$. The case $s>0$ can now be treated by writing

$$\mathscr{F}_M^{-1}\tilde{\Phi}_0\mathscr{K}^{n+1}\Phi = \sum_{j=n+2}^{N} \mathscr{F}_M^{-1}\tilde{\Phi}_0\mathscr{K}^j\Phi_0 + \mathscr{F}_M^{-1}\tilde{\Phi}_0\mathscr{K}^{N+1}\Phi$$

and choosing N large enough, say $N \leqq n+[s]+1$. □

Now consider the first term in expansion (1.16). We have

$$\mathscr{F}_M^{-1}\tilde{\Phi}_0\Phi_0 = \mathscr{F}_M^{-1}\left(\int_{\mathbf{R}^3} e^{ik(\theta-\theta')\cdot y} q(y)\, dy\right)(x) = q(x).$$

Hence, we have proved the following asymptotic series expansion for $\tilde{q}$.

THEOREM 1.8. *Let q satisfy the assumption of the previous lemma. The linearized approximation $\tilde{q}$ of q admits the asymptotic expansion*

$$\tilde{q} \sim q + \sum_{j=1}^{\infty} \mathscr{F}_M \tilde{\Phi}_0 \mathscr{K}^j \Phi_0,$$

where the asymptotic equivalence $\sim$ is to be understood in the sense of

$$\tilde{q} - q - \sum_{j=1}^{n} \mathscr{F}_M^{-1} \tilde{\Phi}_0 \mathscr{K}^j \Phi_0 \in H^t, \qquad t < s + n + \frac{1}{2}.$$

The importance of this result is, of course, in the fact that we are able to analyse how the singularities of q affect the degree of singularity of $\tilde{q}$. The first nonlinear term $q_1(x) = \mathscr{F}_M \tilde{\Phi}_0 \mathscr{K} \Phi_0(x)$ in the above expansion will be studied separately in the next section. It will be proved, among other things, that q_1 belongs to the space Λ^1 if the true potential q is in L^∞_δ for $\delta > \frac{3}{2}$. The proof is based on the explicit form of the term q_1. However, it is possible, by using the mapping properties of the operators discussed in this section, to establish a counterpart of this result to the higher-order terms, too. In the following theorem we use the notation

$$q_j = \mathscr{F}_M^{-1} \tilde{\Phi}_0 \mathscr{K}^j \Phi_0.$$

THEOREM 1.9. *If $\delta > \frac{3}{2}$ and $q \in L^\infty_\delta$ then q_j belongs to the Lipschitz class* Lip_1 *for $j > 2$ and q_2 belongs to* $\mathrm{Lip}_s = \Lambda^s$ *for all s, $0 < s < 1$.*

Proof. For x_1, x_2 in $\mathbf{R}^3$ we have

$$\begin{aligned} q_j(x_1) - q_j(x_2) &= \frac{1}{16\pi^3} \int_{-\infty}^{\infty} k^2\, dk \int d\theta \int d\theta' (1 - \theta \cdot \theta') \tilde{\Phi}_0 \mathscr{K}^j \Phi_0(k, \theta', \theta) \\ &\qquad \cdot (e^{-ik(\theta - \theta') \cdot x_1} - e^{-ik(\theta - \theta') \cdot x_2}) \\ &= \frac{1}{16\pi^3} \int_{-\infty}^{\infty} k^2\, dk ((e_{x_1}, \tilde{\Phi}_0 \mathscr{K}^j \Phi_0\, e_{x_1})_{L^2(S^2)} - (e_{x_2}, \tilde{\Phi}_0 \mathscr{K}^j \Phi_0\, e_{x_2})_{L^2(S^2)} \\ &\qquad - (E_{x_1}, \tilde{\Phi}_0 \mathscr{K}^j \Phi_0 E_{x_1})_{L^2(S^2)} + (E_{x_2}, \tilde{\Phi}_0 \mathscr{K}^j \Phi_0 E_{x_2})_{L^2(S^2)}). \end{aligned}$$

Here, $e_{x_j} = e^{ik\theta \cdot x_j} \in L^2(S^2)$ and $E_{x_j} = \theta\, e_{x_j} \in (L^2(S^2))^3$, the space of vector-valued L^2 functions. The inner products in these spaces are both denoted by $(\cdot, \cdot)_{L^2(S^2)}$. Because $\|E_{x_1} - E_{x_2}\|_{L^2(S^2)} = \|e_{x_1} - e_{x_2}\|_{L^2(S^2)}$ and $\|e_{x_j}\|_{L^2(S^2)} = 4\pi$, we obtain the estimate

$$|q_j(x_1) - q_j(x_2)| \leqq \frac{1}{\pi^2} \int_{-\infty}^{\infty} k^2\, dk \|e_{x_1} - e_{x_2}\|_{L^2(S^2)} \|\tilde{\Phi}_0 \mathscr{K}^j \Phi_0\|.$$

But according to Theorem 1.5 and Proposition 1.2,

$$\|\tilde{\Phi}_0 \mathscr{K}^j \Phi_0\| \leqq \frac{C}{1 + |k|^{j+2}}$$

and on the other hand,

$$\begin{aligned} \|e_{x_1} - e_{x_2}\|^2_{L^2(S^2)} &= 2 \int_{S^2} (1 - e^{ik\theta \cdot (x_1 - x_2)})\, d\theta \\ &= 8\pi (1 - j_0(|k(x_1 - x_2)|). \end{aligned}$$

Thus, by denoting $r = |x_1 - x_2|$, we need an estimate for the integral

$$\int_0^\infty \frac{k^2}{1 + k^{j+2}} \left(1 - \frac{\sin kr}{kr}\right)^{1/2} dk.$$

We split this integral into two parts and estimate the first as

$$\int_{1/r}^{\infty} \frac{k^2}{1+k^{j+2}} \left(1-\frac{\sin kr}{kr}\right)^{1/2} dk \leqq \int_{1/r}^{\infty} \frac{k^2}{1+k^{j+2}}\, dk \leqq Cr^{j-1}.$$

For the remaining part of the integral we get for $j>2$

$$\int_{0}^{1/r} \frac{k^2}{1+k^{j+2}} \left(1-\frac{\sin kr}{kr}\right)^{1/2} dk \leqq r \int_{0}^{1/r} \frac{k^3}{1+k^{j+2}}\, dk \leqq Cr,$$

proving the first claim in the theorem. For $j=2$ we get

$$r \int_{0}^{1/r} \frac{k^3}{1+k^4}\, dk \leqq Cr \log\left(\frac{1}{r}\right) \leqq Cr^s$$

for $0<s<1$. Hence the proof is complete if we can show that $q_j \in L^\infty$. But this is obvious, since

$$|q_j(x)| \leqq C \int_0^{\infty} \frac{k^2}{1+k^{j+2}} \left(\|e_x\|_{L^2(S^2)} + \|E_x\|_{L^2(S^2)}\right) dk \leqq C < \infty. \qquad \square$$

Remark 1.10. It is not hard to see that the argument of Lemma 1.7 holds for potentials with the asymptotic behavior $q(x)=\mathcal{O}(|x|^{-2-\varepsilon})$ at infinity. Also, Theorem 1.5 remains valid (see [12]]). Only the asymptotic behaviour of the solution ψ^+ as x goes to infinity should be interpreted in the $L^2(S^2)$ sense, causing some notational changes in the above analysis.

2. Analysis of the first nonlinear term. In this section it is shown that the smoothness estimates obtained in the previous section are not the best possible. Since the terms in the asymptotic expansion of $\tilde{q}$ contain only the potential and simple, analytically known functions, we can try to write them in a more explicit form to refine the analysis of the degree of smoothness. We shall carry out this kind of analysis for the first nonlinear term here.

THEOREM 2.1. *The first nonlinear term admits the integral representation*

$$q_1(x) = \mathcal{F}_M^{-1} \tilde{\Phi}_0 \mathcal{K} \Phi_0(x) = -\frac{1}{2} \left| \frac{1}{4\pi} \int_{\mathbf{R}^3} \frac{x-y}{|x-y|^3}\, q(y)\, dy \right|^2.$$

The proof of this result is a straightforward calculation, and we postpone it to Appendix B.

To estimate the smoothness of q_1 we will prove the following result.

LEMMA 2.2. *For $\delta>1$ the operators*

$$T_j u(x) = \int \frac{(x-y)_j}{|x-y|^3}\, u(y)\, dy, \qquad j=1,2,3,$$

map $L^\infty_\delta(\mathbf{R}^3)$ to Λ^1.

Proof. We will employ the *Riesz transform* R_j, *Riesz potential* I_s, and *Bessel potential* J_s defined as

$$R_j f(x) = \mathcal{F}^{-1}\left(\frac{\xi_j}{|\xi|}\hat{f}\right)(x), \qquad i=1,2,3,$$

$$I_s f(x) = \mathcal{F}^{-1}(|\xi|^s \hat{f})(x),$$

$$J_s f(x) = \mathcal{F}^{-1}((1+|\xi|^2)^{s/2}\hat{f})(x).$$

Here $\hat{f}$ is the Fourier transform of the function f. It is well known that $J_\sigma : \Lambda^s \to \Lambda^{s-\sigma}$ is an isomorphism for all real values of s and σ (see [16] and [18]). On the other hand, the Riesz transforms are bounded operators in Λ^s. Observing that $\mathcal{F}(x_j/|x|^3) = c\xi_j/|\xi|^2$ for some constant c, we see that in fact

$$T_j f = R_j I_{-1}.$$

Thus we have to show only that I_{-1} maps L^∞_δ to Λ^1. To do so, let φ be a compactly supported C^∞ function in $\mathbf{R}^3$, $\varphi(\xi) = 1$ in some neighborhood of the origin, and let

$$m_1(\xi) = \frac{\varphi(\xi)}{|\xi|}, \qquad m_2(\xi) = \frac{1-\varphi(\xi)}{|\xi|}.$$

If $f \in L^\infty_\delta \subset L^2$, we have $\hat{f} \in L^2$ and, consequently, $m_1 \hat{f} \in L^1$ implying especially that m_1 defines a continuous operator

$$T_{m_1} : f \mapsto \mathcal{F}^{-1}(m_1 \hat{f}), \qquad T_{m_1} : L^\infty_\delta \to L^\infty.$$

On the other hand, since

$$\mathcal{F}(D_j T_{m_1} f)(\xi) = \frac{\xi_j}{|\xi|}\varphi(\xi)\hat{f}(\xi),$$

we see that $D_j T_{m_1} f \in L^\infty$ as well, so that

$$T_{m_1} : L^\infty_\delta \to \Lambda^1$$

continuously. It remains to prove that m_2 also defines a continuous mapping,

$$T_{m_2} : f \mapsto \mathcal{F}^{-1}(m_2 \hat{f}), \qquad T_{m_2} : L^\infty_\delta \to \Lambda^1.$$

In fact, a slightly stronger result can be obtained, namely

$$T_{m_2} : \Lambda^s \to \Lambda^{s+1}$$

for all real values of s. Since $L^\infty_\delta \subset \Lambda^0$ as is shown in Appendix A, this is clearly more than enough. But since J_1 maps Λ^{s+1} to Λ^s, it remains to be shown that

$$\left\| \mathcal{F}^{-1}\left(\frac{(1+|\xi|^2)^{1/2}}{|\xi|}(1-\varphi(\xi))\hat{f}(\xi) \right) \right\|_{\Lambda^s} \leqq C \|f\|_{\Lambda^s}.$$

But this follows from the Mikhlin type of multiplier theorem (Theorem A3 in Appendix A). □

By the pointwise multiplier theorem (Proposition A2) we obtain immediately an estimate of smoothness for q_1 that is clearly better than the one obtained from Lemma 1.7.

COROLLARY 2.3. *If the potential q satisfies $w_\delta q \in L^\infty$, then*

$$q_1 \in H^t \cap \Lambda^1$$

for all t, $0 < t < 1$.

Finally, we combine the results from Theorems 1.8 and 1.9 with Corollary 2.3.

THEOREM 2.4. *For potentials belonging to the space L^∞_δ,*

$$q - \tilde{q} \in \Lambda^s$$

for every s, $0 < s < 1$.

Proof. Since

$$q - \tilde{q} = \sum_{j=1}^{n} q_j + q_n^r,$$

where q_n^r is in the Sobolev space $H^{n+(1/2)-\varepsilon}$, the claim follows from the Sobolev embedding theorem, if we choose n large enough. □

Appendix A. We start this Appendix by defining some classical, "constructive" function spaces. Let k be an integer, $k \geqq 0$. By $C^k = C^k(\mathbf{R}^3)$ we denote the classical Hölder space of k times continuously differentiable fucntions, endowed with the norm

$$\|f\|_{C^k} = \sup_x \sum_{|\alpha| \leqq k} |D^\alpha f(x)|.$$

There are several candidates to extend the definition above to noninteger values by using various moduli of continuity. A useful scale of function spaces is obtained in the following way: Let $s = k + t$, where k is an integer, $0 \leqq k < s$ and $0 < t \leqq 1$. Then f belongs to the classical Zygmund space Λ^s, if

$$\|f\|_{\Lambda^s} := \|f\|_{L^\infty} + \sum_{|\alpha| \leqq k} \sup_{x,h} \frac{|D^\alpha f(x+h) - 2D^\alpha f(y) + D^\alpha f(x-h)|}{|h|^t} < \infty.$$

In general, the symmetric difference cannot be replaced by a simple difference in the above definition. However, when $0 < s < 1$, we have $\Lambda^s = \mathrm{Lip}_s$ with equivalent norms, where the Lipschitz class Lip_s, $0 < s \leqq 1$ is characterized by the condition

$$\|f\|_{\mathrm{Lip}_s} := \|f\|_{L^\infty} + \sup_{x,h} \frac{|f(x+h) - f(x)|}{|h|^s} < \infty.$$

It turns out that for $s = 1$, the space Λ^1 is strictly larger than Lip_1. Similarly, for integer values of s, the space Λ^s is strictly larger than the corresponding Hölder space C^s. Next, we want to extend the definition of Λ^s for values $s \leqq 0$. Similarly to the usual Sobolev space scale, this is accomplished through Fourier analysis. We define first a smooth dyadic partition of unity in the following way: Let φ and φ_0 be C^∞ functions in $\mathbf{R}^3$, $0 \leqq \varphi, \varphi_0 \leqq 1$ such that $\mathrm{supp}\, \varphi = \{\xi \mid \frac{1}{2} \leqq |\xi| \leqq 2\}$ and $\mathrm{supp}\, \varphi_0 = \{\xi \mid |\xi| \leqq 1\}$. Furthermore, assume that the functions are scaled so that

$$\sum_{k=0}^{\infty} \varphi_k(\xi) = 1 \quad \text{for all } \xi \in \mathbf{R}^3,$$

where $\varphi_k(\xi) = \varphi(2^{k-1}\xi)$, $k = 1, 2, 3, \cdots$. Now, by definition, a function f belongs to the Besov space $B^s_{p,q} = B^s_{p,q}(\mathbf{R}^3)$, $0 \leqq p, q \leqq \infty$ and $-\infty < s < \infty$, if

$$\left(\sum_{k=0}^{\infty} (2^{ks} \|\mathscr{F}^{-1}(\varphi_k \hat{f})\|_{L^p})^q \right)^{1/q} < \infty.$$

(If $q = \infty$, the sum is to be replaced by the supremum.) It can be shown that the definition of the Besov spaces is independent of the specific choice of the resolution of unity $\{\varphi_k\}_{k=0}^\infty$ (see [3], [17], [18].) The important tool for us is the following.

PROPOSITION A1. *For $s > 0$,*

$$B^s_{\infty,\infty} = \Lambda^s$$

with equivalent norms.

Therefore, we can define $\Lambda^s = B^s_{\infty,\infty}$ for all s, $-\infty < s < \infty$. It should be noted that the space $B^0_{\infty,\infty}$ is strictly larger than L^∞. The inclusion $L^\infty \subset B^0_{\infty,\infty}$ is simple to verify. If $f \in L^\infty$, then

$$\|\mathscr{F}^{-1}(\varphi_k \hat{f})\|_{L^\infty} = \|\mathscr{F}^{-1}\varphi_k * f\|_{L^\infty} \leqq \|\mathscr{F}^{-1}\varphi_k\|_{L^1} \|f\|_{L^\infty} = C\|f\|_{L^\infty},$$

the constant C being independent of k.

Especially, it follows that for $\delta > 0$,

$$L^\infty_\delta \subset L^\infty \subset \Lambda^0.$$

The Besov spaces are known to have several multiplication properties. The following ones are used in § 2.

PROPOSITION A2. (a) *Let $0 \leqq t < s$. Then, for every $f \in H^t$ and $h \in \Lambda^s$, $hf \in H^t$, and*

$$\|hf\|_{H^t} \leqq C\|h\|_{\Lambda^s}\|f\|_{H^t}.$$

(b) *For $s > 0$ the space Λ^s is a multiplication algebra, i.e., if $h, f \in \Lambda^s$, then $hf \in \Lambda^s$ with the estimate*

$$\|hf\|_{\Lambda^s} \leqq C\|h\|_{\Lambda^s}\|f\|_{\Lambda^s}.$$

Part (a) of the above proposition is a special case of Theorem 2.4.2 of [18]; part (b) follows from Theorem 2.8.3 of the same work. One important property of the Besov spaces is that the Mikhlin–Hörmander type of Fourier multiplier theorem is valid. The following proposition is, again, a special case of the more general result proved in [18, Thm. 2.3.7].

PROPOSITION A3. *Assume that m is a C^∞ function in $\mathbf{R}^3$ satisfying for some integer N large enough the estimate*

$$|D^\alpha m(\xi)| \leqq C_\alpha (1+|\xi|^2)^{-|\alpha|/2}, \qquad 0 \leqq |\alpha| \leqq N.$$

Then, for $f \in \Lambda^s$, $-\infty < s < \infty$, there is a constant $C > 0$ depending only on the constants C_α such that for all $f \in \Lambda^s$,

$$\|\mathscr{F}^{-1}(m\hat{f})\|_{\Lambda^s} \leqq C\|f\|_{\Lambda^s}.$$

Finally, we close this Appendix by proving Theorem 1.1.

Proof of Theorem 1.1. It is well known (see [17, p. 201]) that the Zygmund spaces are obtained from the Hölder spaces by the real interpolation method as

$$(C^{k_0}, C^{k_1})_{\theta,\infty} = \Lambda^s, \qquad s = (1-\theta)k_0 + \theta k_1.$$

On the other hand, by the extremal property of the real interpolation method,

$$(C^{k_0}, C^{k_1})_{\theta,1} \subset [C^{k_0}, C^{k_1}]_\theta \subset (C^{k_0}, C^{k_1})_{\theta,\infty}.$$

Therefore, to obtain the claim, it suffices to prove the following embedding result: If $A_1 \subset A_0$, then

$$(A_0, A_1)_{\eta,\infty} \subset (A_0, A_1)_{\theta,1},$$

where $0 < \theta < \eta < 1$. But this follows relatively easily by using the standard K-functional techniques. By definition,

$$K(t, a) = \inf_{a = a_0 + a_1} (\|a_0\|_{A_0} + t\|a_1\|_{A_1})$$

for $a = a_0 + a_1 \in A_0 + A_1$. But since $A_0 + A_1 \subset A_0$, we get $K(t, a) \leqq \|a\|_{A_0}$. Therefore, applying the $\Phi_{\theta,1}$-functional to $K(t, a)$,

$$\begin{aligned}\Phi_{\theta,1}(K(\cdot, a)) &= \int_0^\infty t^{-\theta} K(t, a)\frac{dt}{t} \leqq \sup_{t \geqq 0} t^{-\eta}K(t, a)\int_0^1 t^{\eta-\theta}\frac{dt}{t} + \|a\|_{A_0}\int_1^\infty t^{-\theta}\frac{dt}{t} \\ &= C_1\Phi_{\eta,\infty}(K(\cdot, a)) + C_2\|a\|_{A_0} \leqq C\|a\|_{(A_0,A_1)_{\eta,\infty}}.\end{aligned}$$

The last inequality follows from the fact that $(A_0, A_1)_{\eta,\infty}$ is an intermediate space, i.e., $(A_0, A_1)_{\eta,\infty} \subset A_0 + A_1 \subset A_0$. □

Appendix B. In this section we derive the expression for q_1 as given in the Theorem 2.1. We start with two rather technical but straightforward lemmas.

LEMMA B1. *For $x, y \in \mathbf{R}^3$, $k \in \mathbf{R}$, we have*

$$\int_{S^2} d\theta\, e^{-ik\theta\cdot x} \int_{S^2} d\theta'\, \theta\cdot(\theta-\theta')\, e^{ik\theta'\cdot y} = (4\pi)^2(j_0(k|x|)j_0(k|y|) - \hat{x}\cdot\hat{y} j_1(k|x|)j_1(k|y|)).$$

Here, j_0 and j_1 are the spherical Bessel functions of orders zero and 1, *and $\hat{x} = x/|x|$, $\hat{y} = y/|y|$.*

Proof. Start with the θ' integral,

$$I(\theta, k, y) = \int_{S^2} d\theta'\, \theta\cdot(\theta-\theta')\, e^{ik\theta'\cdot y}.$$

Using the expansion of the exponential function in spherical harmonics,

$$e^{ik\theta'\cdot y} = e^{i|k||y|\theta\cdot(\operatorname{sign} k\hat{y})} = 4\pi \sum_{l=0}^{\infty}\sum_{m=-l}^{l} i^l j_l(|ky|)\, Y_l^m(\operatorname{sign} k\,\hat{y})\, \bar{Y}_l^m(\theta'),$$

and choosing the coordinates with the $\hat{e}_3$ axis is parallel to θ so that

$$1-\theta\cdot\theta' = 1-\cos(\theta,\theta') = \sqrt{4\pi}\left(Y_0^0(\theta') - \frac{1}{\sqrt{3}} Y_1^0(\theta')\right),$$

we get by the orthogonality of the spherical harmonics

$$I(\theta, k, y) = (4\pi)^{3/2}\left(j_0(|ky|)\, Y_0^0(\operatorname{sign} k\,\hat{y}) - \frac{i}{\sqrt{3}} j_1(|ky|)\, Y_1^0(\operatorname{sign} k\,\hat{y})\right)$$

$$= 4\pi(j_0(|ky|) - i \operatorname{sign} k\,\hat{y}\cdot\theta j_1(|ky|)).$$

Invoking the spherical harmonics expansion once more, this time choosing the $\hat{e}_3$ axis parallel to sign $k\,\hat{y}$,

$$\int_{S^2} d\theta\, e^{-ik\theta\cdot x} I(k, \theta, y)$$

$$= (4\pi)^{5/2} \sum_{l=0}^{l=\infty}\sum_{m=-l}^{m=l} i^l j_l(|kx|)\, Y_l^m(-\operatorname{sign} k\,\hat{x})$$

$$\cdot \int_{S^2} \bar{Y}_l^m(\theta)\left(j_0(|ky|)\, Y_0^0(\theta) - i\frac{1}{\sqrt{3}} j_1(|ky|)\, Y_1^0(\theta)\right) d\theta$$

$$= (4\pi)^2(j_0(|kx|)j_0(|ky|) - \hat{x}\cdot\hat{y} j_1(|kx|)j_1(|ky|)),$$

which yields the claim, since the products of the Bessel functions are even functions. □

To deal with the k-integral we prove the following result.

LEMMA B2. *For α, β, and γ positive, we have the following formulas for the distributional Fourier transforms of the products of Bessel functions*:

$$\int_{-\infty}^{\infty} e^{ik\alpha} j_0(\beta k) j_0(\gamma k)\, dk = -\frac{\pi}{2\beta\gamma}(\delta(\alpha+\beta+\gamma) + \delta(\alpha-\beta-\gamma)$$

$$- \delta(\alpha+\beta-\gamma) - \delta(\alpha-\beta+\gamma)),$$

and similarly

$$\int_{-\infty}^{\infty} e^{ik\alpha} j_1(\beta k) j_1(\gamma k)\, dk$$

$$= \frac{\pi}{2\beta\gamma}\Big(\delta(\alpha+\beta-\gamma)+\delta(\alpha-\beta-\gamma)+\delta(\alpha+\beta-\gamma)+\delta(\alpha-\beta+\gamma)$$

$$-\frac{1}{\beta}(\Theta(\beta-\alpha-\gamma)+\Theta(\beta-|\alpha-\gamma|))-\frac{1}{\gamma}(\Theta(\gamma-\alpha-\beta)+\Theta(\gamma-|\alpha-\beta|))$$

$$+\frac{1}{\beta\gamma}(\min(\beta,\alpha+\gamma)+\operatorname{sign}(\gamma-\alpha)\min(\beta,|\alpha-\gamma|))\Big).$$

Here, δ is the delta distribution and Θ is the Heaviside function.

Proof. By the explicit formula

$$j_0(t)=\frac{\sin t}{t}=\frac{1}{2it}(e^{it}-e^{-it}),$$

the first identity is immediate. Similarly, by the formula

$$j_1(t)=\left(\frac{\sin t}{t^2}-\frac{\cos t}{t}\right)$$

we obtain for the second term

$$j_1(\beta k)j_1(\gamma k)=\frac{1}{\beta\gamma k^2}\Big(\cos\beta k\cos\gamma k-\frac{1}{\beta k}\sin\beta k\cos\gamma k-\frac{1}{\gamma k}\sin\gamma k\cos\beta k$$

$$+\frac{1}{\beta\gamma k^2}\sin\beta k\sin\gamma k\Big).$$

We will integrate this expression term by term. The first term gives the δ-terms appearing in the claim. To evaluate the second term we write

$$\int_{-\infty}^{\infty}\frac{1}{k}e^{ik\alpha}\sin\beta k\cos\gamma k\,dk=2\int_0^{\infty}\frac{1}{k}\cos\alpha k\sin\beta k\cos\gamma k\,dk$$

$$=\int_0^{\infty}\frac{1}{k}\sin\beta k\cos(\alpha+\gamma)k\,dk+\int_0^{\infty}\frac{1}{k}\sin\beta k\cos(\alpha-\gamma)k\,dk$$

$$=\frac{\pi}{2}\Theta(\beta-\alpha-\gamma)+\frac{\pi}{2}\Theta(\beta-|\alpha-\gamma|).$$

The last equality follows from [6, p. 414].

The integral of the third term equals that of the second with the roles of β and γ interchanged. Finally, to integrate the fourth term, we write

$$\int_{-\infty}^{\infty}\frac{1}{k^2}e^{ik\alpha}\sin\beta k\sin\gamma k\,dk=2\int_0^{\infty}\frac{1}{k^2}\cos\alpha k\sin\beta k\sin\gamma k\,dk$$

$$=\int_0^{\infty}\frac{1}{k^2}\sin\beta k\sin(\alpha+\gamma)k\,dk+\operatorname{sign}(\gamma-\alpha)\int_0^{\infty}\frac{1}{k^2}\sin\beta k\sin|\gamma-\alpha|k\,dk$$

$$=\frac{\pi}{2}\min(\beta,\alpha+\gamma)+\frac{\pi}{2}\operatorname{sign}(\gamma-\alpha)\min(\beta,|\gamma-\alpha|)$$

by [6, p. 414]. The proof of the lemma is thus complete. □

Now we are ready to prove Theorem 2.1. Let $\varphi \in S(\mathbf{R}^3)$. By the distributional definition of $\mathscr{F}_M^{-1}$, we have

$$\begin{aligned}\langle \mathscr{F}_M^{-1}\tilde{\Phi}_0 \mathscr{K} \Phi_0, \varphi\rangle &= \int_M d\mu(k,\theta',\theta)\tilde{\Phi}_0\mathscr{K}\Phi_0(k,\theta',\theta)\mathscr{F}^{-1}\varphi(k(\theta-\theta')) \\ &= \int_M d\mu(k,\theta',\theta)\mathscr{F}^{-1}\varphi(k(\theta-\theta'))\int_{\mathbf{R}^3} dx\, e^{-ik\theta'\cdot x} q(x) \\ &\quad \cdot \int_{\mathbf{R}^3} dy \frac{e^{ik|x-y|}}{4\pi|x-y|} q(y)\, e^{ik\theta\cdot y}.\end{aligned}$$

It is not hard to see, by using the mapping properties of $\mathscr{R}_0^+(k)$, that we can use the Fubini theorem in the above integral to get

$$\begin{aligned}\langle \mathscr{F}_M^{-1}\tilde{\Phi}_0 \mathscr{K} \Phi_0, \varphi\rangle = \frac{1}{4\pi}\int dx\, q(x)\int dy\, q(y)\frac{1}{|x-y|}\int_M d\mu(k,\theta',\theta)\mathscr{F}^{-1}\varphi(k(\theta-\theta')) \\ \cdot e^{-ik\theta'\cdot x}\, e^{ik|x-y|}\, e^{ik\theta\cdot y}.\end{aligned}$$

To obtain the claimed result, we have to prove that for $x \neq y$,

$$\begin{aligned}&\frac{1}{4\pi}\int_M d\mu(k,\theta',\theta)\mathscr{F}^{-1}\varphi(k(\theta-\theta'))\, e^{-ik\theta'\cdot x}\, e^{ik|x-y|}\, e^{ik\theta\cdot y} \\ &\quad = -\frac{1}{32\pi}\int_{\mathbf{R}^3} dz\, \varphi(z)|x-y|\frac{(x-z)\cdot(y-z)}{|x-z|^3|y-x|^3}.\end{aligned}$$

By Lemma B1, we get

$$\begin{aligned}&\frac{1}{4\pi}\int_M d\mu(k,\theta',\theta)\mathscr{F}^{-1}\varphi(k(\theta-\theta'))\, e^{-ik\theta'\cdot x}\, e^{ik|x-y|}\, e^{ik\theta\cdot y} \\ &\quad = \frac{1}{16\pi^3}\lim_{n\to\infty}\int_{-n}^{n} dk\, k^2\int_{\mathbf{R}^3} dz\, \varphi(z)\, e^{ik|x-y|}(j_0(k|z-y|)j_0(k|z-x|) \\ &\quad\quad -(\widehat{z-y})\cdot(\widehat{z-x})j_1(k|z-y|)j_1(k|z-x|)).\end{aligned}$$

Further, by Lemma A2 with $\alpha = |x-y|$, $\beta = |z-y|$ and $\gamma = |z-x|$, the above integral equals to $I_1 + I_2$, where

$$\begin{aligned}I_1 = -\frac{1}{32\pi^2}\int_{\mathbf{R}^3} dz \frac{\varphi(z)}{|z-y||z-x|} \\ \cdot\{(1+(\widehat{z-y})\cdot(\widehat{z-x}))(\delta(|x-y|+|z-y|+|z-x|) \\ +\delta(|x-y|-|z-y|-|z-x|))+(1-(\widehat{z-y})\cdot(\widehat{z-x})) \\ \cdot(\delta(|x-y|-|z-y|+|z-x|)+\delta(|x-y|+|z-y|-|z-x|))\},\end{aligned}$$

and the remaining part contains the less singular terms, i.e.,

$$\begin{aligned}I_2 = -\frac{1}{16\pi^2}\int_{\mathbf{R}^3} dz \frac{\varphi(z)}{|z-y||z-x|}(\widehat{z-y})\cdot(\widehat{z-x}) \\ \cdot\left\{\frac{1}{|z-y|}(\Theta(|z-y|-|x-y|-|z-x|)+\Theta(|z-y|-\|x-y|-|z-x\|)) \right. \\ \left. +\frac{1}{|z-x|}(\Theta(|z-x|-|x-y|-|z-y|)+\Theta(|z-x|-\|x-y|-|z-y\|))\right.\end{aligned}$$

$$-\frac{1}{|z-y||z-x|}(\min(|z-y|, |x-y|+|z-x|)$$
$$+\operatorname{sign}(|z-x|-|x-y|)\min(|z-y|, \|x-y|-|z-x\|))\Big\}.$$

First we show that the first integral I_1 vanishes. To get an interpretation for the one-dimensional delta distributions appearing in the first integral I_1, we use the prolate spheroidal coordinates with foci at x and y. If $|x-y|=2a>0$, we choose the new coordinates (ξ, ζ, ω) as

$$2a\xi = |x-z|+|y-z|, \qquad 1 \leqq \xi < \infty,$$
$$2a\zeta = |x-z|-|y-z|, \qquad -1 \leqq \zeta \leqq 1,$$
$$\omega = \text{angle around, the axis passing through } x \text{ and } y, \qquad 0 \leqq \omega \leqq 2\pi.$$

The Jacobian of the change of variables $z=(z_1, z_2, z_3) \to (\xi, \zeta, \omega)$ is

$$\left|\frac{\partial(z_1, z_2, z_3)}{\partial(\xi, \zeta, \omega)}\right| = a|x-z||y-z|$$

(see [2, p. 105]). Finally, denoting $\alpha_\pm = 1 \pm (\widehat{z-y}) \cdot (\widehat{z-x})$ and by u the coordinate transform $z=u(\xi, \zeta, \omega)$, we have

$$I_1 = -\frac{1}{64\pi^2}\int_1^\infty d\xi \int_{-1}^1 d\zeta \int_0^{2\pi} d\omega\, \varphi \circ u(\xi, \zeta, \omega)$$
$$\cdot (\alpha_+(\delta(1+\xi)+\delta(1-\xi))+\alpha_-(\delta(1-\zeta)+\delta(1+\zeta))),$$

where the identity $\delta(at)=\delta(t)/a$ is used. From this representation we deduce that, in fact, $I_1=0$, because by a simple geometrical argument $\alpha_+=0$ when $\xi=1$, and $\alpha_-=0$ when $\zeta=\pm 1$. To simplify the I_2 integral, note that $|z-y| \leqq |z-x|+|x-y|$ and $\|y-x|-|z-x\| \leqq |z-y|$ and similarly with x and y interchanged, so I_2 reduces simply to

$$I_2 = -\frac{1}{32\pi^2}\int_{\mathbf{R}^3} dz\, \varphi(z) \frac{(\widehat{z-y}) \cdot (\widehat{z-x})}{|z-y||z-x|}$$
$$\cdot \left(\frac{1}{|z-y|}+\frac{1}{|z-x|}-\frac{1}{|z-y||z-x|}(|z-y|+|z-x|-|x-y|)\right)$$
$$= -\frac{1}{32\pi^2}\int_{\mathbf{R}^3} dz\, \varphi(z) \frac{(\widehat{z-y}) \cdot (\widehat{z-x})}{|z-y|^2|z-x|^2}|x-y|.$$

The proof of Theorem 2.1 is thus complete. □

REFERENCES

[1] S. AGMON, *Spectral properties of Schrödinger operators and scattering theory*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 4 (2) (1975), pp. 151-218.

[2] G. ARFKEN, *Mathematical methods for physicists*, Second edition, Academic Press, New York, 1970.

[3] J. BERGH AND J. LÖFSTRÖM, *Interpolation Spaces. An Introduction*, Springer-Verlag, Berlin, New York, 1976.

[4] G. BEYLKIN, *Imaging of discontinuities in the inverse scattering problem by inversion of a causal generalized Radon transform*, J. Math. Phys., 26 (1985), pp. 99-108.

[5] R. BURRIDGE AND G. BEYLKIN, *On double integrals over spheres*, Inverse Problems, 4 (1988), pp. 1-10.

[6] I. S. GRADSTEYN AND I. M. RYZHIK, *Tables of Integrals, Series, and Products*, Fourth edition, Academic Press, New York, 1965.

[7] T. IKEBE, *Eigenfunction expansions associated with the Schrödinger operators and their application to scattering theory*, Arch. Mech. Anal. 5 (1960), pp. 1–34.

[8] R. NEWTON, *Noncentral potentials: the generalized Levinson theorem and the structure of the spectrum*, J. Math. Phys., 18 (1977), pp. 1348–1357.

[9] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics* III: *Scattering Theory*, Academic Press, New York, 1979.

[10] J. ROSE AND M. CHENEY, *Three-dimensional inverse scattering for the wave equation: weak scattering approximations with error estimates*, Inverse Problems, 4 (1988), pp. 435–447.

[11] Y. SAITO, *The principle of limiting absorption for the non-selfadjoint Schrödinger operator in* R^N ($N \neq 2$), Publ. Res. Inst. Math. Sci., 9 (1974), pp. 397–428.

[12] ———, *Spectral Representations for the Schrödinger Operators with Long-Range Potentials*, Lecture Notes in Math. 727, Springer-Verlag, Berlin, New York, 1979.

[13] ———, *Some properties of the scattering amplitude and inverse scattering problem*, Osaka J. Math., 19 (1982), pp. 527–547.

[14] E. SOMERSALO, *One-dimensional electromagnetic inverse reflection problem: formulation as a Riemann–Hilbert problem and imaging of discontinuities*, SIAM J. Appl. Math., 49 (1989), pp. 944–951.

[15] E. SOMERSALO, G. BEYLKIN, R. BURRIDGE, AND M. CHENEY, *Inverse scattering problem for the Schrödinger equation in three dimensions: Connections between exact and approximate methods*, IMA Preprint Ser. 449, University of Minnesota, Minneapolis, MN, 1988, pp. 1–10.

[16] E. STEIN, *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, Princeton, NJ, 1970.

[17] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.

[18] H. TRIEBEL, *Theory of Function Spaces*, Monographs in Mathematics 78, Birkhäuser, Basel, Boston, 1983.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 500-523, March 1991

013

TWO-DIMENSIONAL STATIONARY PHASE APPROXIMATION: STATIONARY POINT AT A CORNER*

J. P. McCLURE† AND R. WONG‡

Abstract. Asymptotic expansions are derived for the double integral

$$\int\int_D g(x, y) \exp(iNf(x, y))\, dx\, dy,$$

as $N \to +\infty$, where $f(x, y)$ has a stationary point at a corner of the boundary of D. Two different methods are given, one for the case of local extrema and one for saddlepoints. In the case of saddlepoints, our method allows the boundary of D to be tangent to a level curve of f.

Key words. stationary phase approximation, asymptotic expansion, double integral, stationary point at a corner

AMS(MOS) subject classification. primary 41A60

1. Introduction.

Consider the double integral

$$I(N) = \int\int_D g(x, y) \exp(iNf(x, y))\, dx\, dy, \tag{1.1}$$

where D is a bounded domain, f and g are real-valued C^∞-functions in D, and N is a large positive parameter. By a *stationary point* (x_0, y_0) of f, we mean a point at which the gradient of f vanishes, i.e., $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$. A stationary point is said to be *nondegenerate* if the Hessian matrix of f is nonsingular, i.e.,

$$\det f''(x_0, y_0) \equiv (f_{xx}f_{yy} - f_{xy}^2)(x_0, y_0) \neq 0. \tag{1.2}$$

If (x_0, y_0) is the only stationary point of f in D, if it is nondegenerate, and if it lies in the interior of D, then it is well known that

$$I(N) \sim g(x_0, y_0) \exp(iNf(x_0, y_0)) |\det f''(x_0, y_0)|^{-1/2} \left(\frac{2\pi\sigma}{N}\right) \tag{1.3}$$

as $N \to \infty$, where σ is equal to i, $-i$, or 1 depending on whether (x_0, y_0) is a local minimum, local maximum, or saddlepoint, respectively; see, for example, [2, eq. 18] or [11, Chap. VIII].

If the stationary point (x_0, y_0) is nondegenerate and lies on the boundary of D, and if the boundary curve is smooth near this point, then it is also known that $I(N)$ is asymptotic to half of the above approximation, i.e.,

$$I(N) \sim g(x_0, y_0) \exp(iNf(x_0, y_0)) |\det f''(x_0, y_0)|^{-1/2} \left(\frac{\pi\sigma}{N}\right) \tag{1.4}$$

as $N \to \infty$; see [6, p. 22] or [11, p. 442]. However, in the case where the stationary point (x_0, y_0) is a saddlepoint (i.e., $(f_{xy}^2 - f_{xx}f_{yy})(x_0, y_0) > 0$), the level set $f(x, y) = f(x_0, y_0)$, near (x_0, y_0), can be shown to consist of a pair of smooth curves intersecting at (x_0, y_0).

* Received by the editors October 2, 1989; accepted for publication (in revised form) March 27, 1990. This research was partially supported by Natural Sciences and Engineering Research Council of Canada grants A-8069 and A-7359.

† Department of Mathematics and Astronomy, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 2N2.

‡ Department of Applied Mathematics, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 2N2.

The validity of (1.4) requires the additional assumption that the boundary curve $h(x, y) = 0$ not be tangent to either of these curves at (x_0, y_0). This condition can be expressed analytically by

$$\{h_x f_{yy} + h_y(-f_{xy} \pm \sqrt{f_{xy}^2 - f_{xx} f_{yy}})\}(x_0, y_0) \neq 0. \tag{1.5}$$

(This reduces to the more familiar condition $(h_x^2 f_{yy} + h_y^2 f_{xx})(x_0, y_0) \neq 0$ given in [6, p. 23], when $f_{xy}(x_0, y_0) = 0$.)

If the nondegenerate stationary point (x_0, y_0) occurs at a corner, i.e., at a point on the boundary where the tangent line has a jump, then there is no result that can claim any generality. The problem of deriving an asymptotic formula for the integral (1.1) in this case has been known for some time, and is explicitly mentioned in [1, p. 88]. (For a discussion of corners that are not stationary points, see [2], [5], and [6].) The purpose of this paper is to provide a solution to this problem, and our discussion will be divided into two separate cases. In one case, the corner point (x_0, y_0) is a local extremum of the phase function $f(x, y)$; and in the other case, (x_0, y_0) is a saddlepoint. The methods that we use in these two cases are quite different. In the first case, our method consists of a reduction of the double integral (1.1) to a one-dimensional Fourier transform and an application of a Taylor-type expansion of the Dirac δ-function. This approach is motivated by a method of Jones and Kline [6] as elaborated in [12]. In the second case, we make repeated use of Green's theorem and follow with an application of a result due to Erdélyi [4] concerning the stationary phase method in the presence of a logarithmic singularity. From our discussion below, it will become apparent that our method works even when the boundary curve is smooth at the stationary point but condition (1.5) fails to hold.

2. Local extremum at a corner. For convenience, we shall assume that the corner point (x_0, y_0) is at the origin, i.e., $x_0 = y_0 = 0$, and that near this point the domain D is bounded by two smooth curves intersecting only at $(0, 0)$. We will represent these curves by $h(x, y) = 0$ and $k(x, y) = 0$, where h and k are C^∞-functions with nonvanishing gradients. Since these two curves intersect with a positive angle at $(0, 0)$, i.e., they are not tangent to each other there, the gradients (h_x, h_y) and (k_x, k_y) are linearly independent at $(0, 0)$. Without loss of generality, it can be assumed that these gradients point into D for (x, y) near $(0, 0)$. By using a partition of unity [10, pp. 147, 246] or a neutralizer [11, p. 427], we can always isolate the stationary point $(0, 0)$ from other critical points by multiplying g by a C^∞-function with compact support, which may be taken arbitrarily small. Thus we may assume that g vanishes C^∞-smoothly off an arbitrarily small neighborhood of $(0, 0)$. The situation is as depicted in Fig. 1.

In the integral (1.1), we now make the change of variables

$$s = h(x, y), \qquad t = k(x, y). \tag{2.1}$$

Since (h_x, h_y) and (k_x, k_y) are linearly independent at $(0, 0)$,

$$\frac{\partial(s, t)}{\partial(x, y)} = h_x k_y - h_y k_x \equiv \Delta(x, y) \tag{2.2}$$

does not vanish at $(x, y) = (0, 0)$. Consequently, the transformation (2.1) is invertible near $(0, 0)$. By shrinking the support of g if necessary, we may assume without loss of generality that this transformation is invertible, with nonzero Jacobian, on $\operatorname{supp}(g)$. Thus

$$I(N) = \int\int_{D_1} g_1(s, t) \exp(iNf_1(s, t))\, ds\, dt, \tag{2.3}$$

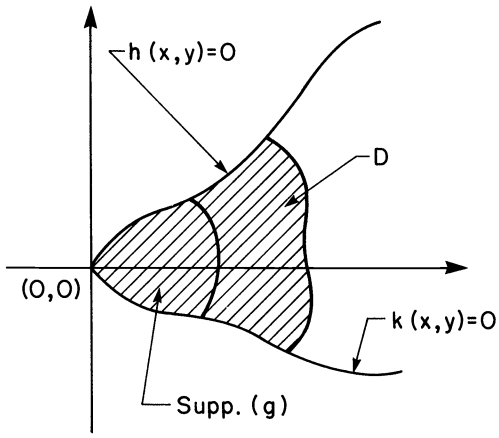


FIG. 1. *Domain D.*

where

$$f_1(s, t) = f(x, y), \qquad g_1(s, t) = \frac{g(x, y)}{|\Delta|}, \tag{2.4}$$

and D_1 is the image of D (see Fig. 2). Simple computation shows that $f_1(0, 0) = f(0, 0)$ and $(0, 0)$ is a stationary point of $f_1(s, t)$. Furthermore, near $(0, 0)$ we have

$$f_1(s, t) = f(0, 0) + \tfrac{1}{2}(\alpha s^2 + 2\beta st + \gamma t^2) + \cdots, \tag{2.5}$$

where

$$\alpha = \frac{\partial^2 f}{\partial x^2} k_y^2 - 2\frac{\partial^2 f}{\partial x\, \partial y} k_x k_y + \frac{\partial^2 f}{\partial y^2} k_x^2, \tag{2.6}$$

$$\beta = -\left(\frac{\partial^2 f}{\partial x^2} h_y k_y + \frac{\partial^2 f}{\partial y^2} h_x k_x\right) + \frac{\partial^2 f}{\partial x\, \partial y}(h_x k_y + h_y k_x), \tag{2.7}$$

$$\gamma = \frac{\partial^2 f}{\partial x^2} h_y^2 - 2\frac{\partial^2 f}{\partial x\, \partial y} h_x h_y + \frac{\partial^2 f}{\partial y^2} h_x^2, \tag{2.8}$$

all derivatives being evaluated at $(0, 0)$. Here we have also used the fact that the first-order partial derivatives of $f(x, y)$ vanish at $(0, 0)$.

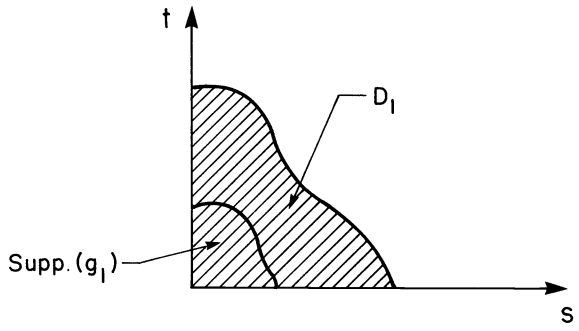


FIG. 2. *Domain* D_1.

Our next step is to eliminate the cross-product term st in the Maclaurin expansion (2.5). To do this, we make the second change of variables:

$$u = s + \frac{\beta}{\alpha} t \quad \text{and} \quad v = t. \tag{2.9}$$

It is easily verified that

$$\alpha s^2 + 2\beta st + \gamma t^2 = \lambda u^2 + \mu v^2,$$

where $\lambda = \alpha$ and $\mu = (\alpha\gamma - \beta^2)/\alpha$, and that the Jacobian of the transformation is 1. The integral $I(N)$ in (2.3) now becomes

$$I(N) = \int\int_{D_2} \psi(u, v) \exp(iN\phi(u, v))\, du\, dv, \tag{2.10}$$

where

$$\phi(u, v) = f(0, 0) + \frac{\lambda}{2} u^2 + \frac{\mu}{2} v^2 + \cdots, \tag{2.11}$$

$$\psi(u, v) = g_1(s, t). \tag{2.12}$$

The domain D_2 is as shown in Fig. 3.

We will apply the argument in [12] to the integral in (2.10). Before proceeding, we first observe that the nature of the stationary point $(0, 0)$ of $f(x, y)$ is preserved under the transformations (2.1) and (2.9). That is, if $(0, 0)$ is a local extremum of $f(x, y)$ then $(0, 0)$ is a local extremum of $\phi(u, v)$, and if $(0, 0)$ is a saddlepoint of $f(x, y)$ then $(0, 0)$ is a saddlepoint of $\phi(u, v)$. To demonstrate this, we let

$$a = \frac{\partial^2 f}{\partial x^2}(0, 0), \quad b = \frac{\partial^2 f}{\partial x\, \partial y}(0, 0), \quad c = \frac{\partial^2 f}{\partial y^2}(0, 0). \tag{2.13}$$

Simple computation yields

$$\alpha\gamma - \beta^2 = (ac - b^2)(k_x h_y - k_y h_x)^2, \tag{2.14}$$

$$\alpha = a\left[\left(k_y - \frac{b}{a} k_x\right)^2 + \frac{ac - b^2}{a^2} k_x^2\right], \tag{2.15}$$

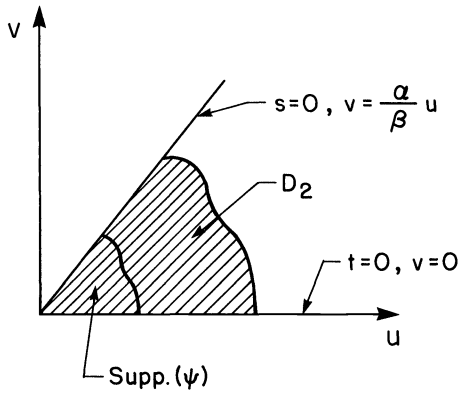


FIG. 3. *Domain* D_2.

where all derivatives are evaluated at $(0, 0)$. It is now evident that the signs of $\alpha\gamma - \beta^2$ and α, and hence those of λ and μ in (2.11), are determined by the signs of $ac - b^2$ and a. If, say, $(0, 0)$ is a local maximum of $f(x, y)$, then $ac - b^2 > 0$ and $a < 0$. Hence, $\alpha\gamma - \beta^2 > 0$ and $\alpha < 0$. Consequently, λ and μ are both negative, or equivalently, $(0, 0)$ is a local maximum of $\phi(u, v)$.

The argument in [12] first expresses the integral (2.10) in the form of a one-dimensional Fourier transform, using the method of resolution of double integrals [3, pp. 298–300]. Let m and M denote the infimum and supremum, respectively, of $\phi(u, v) - f(0, 0)$ in D_2. Then

$$I(N) = e^{iNf(0,0)} \int_m^M e^{iNt} h(t)\, dt, \tag{2.16}$$

where

$$h(t) = \int_\Gamma \frac{\psi(u, v)}{\sqrt{\phi_u^2 + \phi_v^2}}\, d\sigma \tag{2.17}$$

and Γ is the curve defined by

$$\Gamma = \{(u, v) \in D_2 : \phi(u, v) - f(0, 0) = t\}, \tag{2.18}$$

σ being the arclength of Γ.

Assume for the moment that $(0, 0)$ is a local minimum of $f(x, y)$. Then $(0, 0)$ is a local minimum of $\phi(u, v)$, in which case both λ and μ are positive. Put

$$u = \left(\frac{2\xi}{\lambda}\right)^{1/2} \cos\eta, \qquad v = \left(\frac{2\xi}{\mu}\right)^{1/2} \sin\eta. \tag{2.19}$$

Clearly

$$\xi = \frac{\lambda}{2} u^2 + \frac{\mu}{2} v^2, \tag{2.20}$$

$$\frac{\partial(u, v)}{\partial(\xi, \eta)} = \frac{1}{\sqrt{\lambda\mu}}. \tag{2.21}$$

In terms of the polar coordinates (ξ, η), the line integral becomes

$$h(t) = \int_{\Gamma'} \frac{\Phi(\xi, \eta)}{|\nabla(\xi + F)|}\, d\sigma', \tag{2.22}$$

where

$$\xi + F(\xi, \eta) = \phi(u, v) - f(0, 0), \qquad \Phi(\xi, \eta) = \frac{1}{\sqrt{\lambda\mu}} \psi(u, v), \tag{2.23}$$

$$\Gamma' = \{(\xi, \eta) \in D_3 : \xi + F(\xi, \eta) = t\}, \tag{2.24}$$

and $d\sigma'$ denotes the length of Γ'. The domain D_3 is shown in Fig. 4.

From (2.11), (2.19), and (2.23), it is easily seen that $F(\xi, \eta) = O(\xi^{3/2})$ and $F_\xi(\xi, \eta) = O(\xi^{1/2})$. Hence, by shrinking the support of Φ if necessary, we may assume without loss of generality that $1 + F_\xi > 0$. The implicit function theorem then ensures

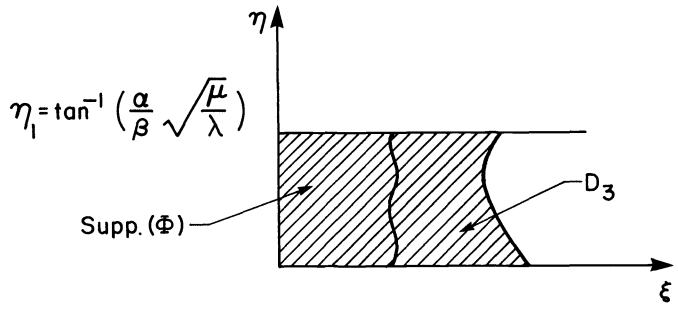


FIG. 4. *Domain* D_3.

that the equation $t = \xi + F(\xi, \eta)$ has a unique C^∞-solution $\xi_t(\eta)$. Consequently, we may write explicitly

$$d\sigma' = \frac{\sqrt{[1+F_\xi(\xi_t, \eta)]^2 + [F_\eta(\xi_t, \eta)]^2}}{1+F_\xi(\xi_t, \eta)}\, d\eta,$$

and the line integral in (2.22) becomes

$$h(t) = \int_0^{\eta_1} \frac{\Phi(\xi_t, \eta)}{1+F_\xi(\xi_t, \eta)}\, d\eta, \tag{2.25}$$

where $\eta_1 = \tan^{-1}[(\alpha/\beta)\sqrt{\mu/\lambda}]$. We will show next that $h(t)$ has an asymptotic expansion of the form

$$h(t) \sim a_0 + a_1 t^{1/2} + a_2 t + \cdots \quad \text{as } t \to 0^+. \tag{2.26}$$

To prove (2.26), we observe that in terms of generalized functions, the line integral (2.22) can be expressed as

$$h(t) = \langle \delta(t - \xi - F), \Phi \rangle,$$

where δ is the distribution concentrated on the curve Γ'; see [12, eqs. (3.5), (3.6)]. Thus the result of Lemma 2 in [12] applies immediately. Since Γ' lies inside domain D_3 (Fig. 4), the upper limit 2π in equation (4.6) of [12] is replaced by η_1. Consequently,

$$h(t) = \sum_{r=0}^{n} \frac{(-1)^r}{r!} \frac{\partial^r}{\partial t^r} \int_0^{\eta_1} \Phi(t, \eta) F^r(t, \eta)\, d\eta + R_{n+1}(t), \tag{2.27}$$

where the remainder satisfies

$$R_{n+1}(t) = O(t^{(n+1)/2}) \quad \text{as } t \to 0^+; \tag{2.28}$$

cf. [12, eq. (4.8)]. Since f and g are C^∞-functions, so are ϕ and ψ. Using (2.19), we can then write for any $p \geqq 1$

$$\Phi(\xi, \eta) F^r(\xi, \eta) = \sum_{\mu=0}^{p-1} c_{\mu r}(\eta) \xi^{\mu/2} + O(\xi^{p/2})$$

as $\xi \to 0^+$, where $c_{\mu r}(\eta)$ is a polynomial in $\cos\eta$ and $\sin\eta$, and the O-symbol is independent of η. From this it follows that each term in the series of (2.27) has an asymptotic expansion, as $t \to 0^+$, of the form $\sum_{\nu=0}^{\infty} a_{\nu r} t^{\nu/2}$. By rearranging the terms appropriately the desired result (2.26) is obtained.

In (2.10) it is tempting to make a nonlinear change of variables so that the phase function $\phi(u, v)$ becomes a sum of two squares. This would eliminate the use of the complicated expansion (2.27), and render the analysis much simpler; see [11, Chap. VIII, §§ 3, 6]. However, under this change of variable, the boundary curve $v=(\alpha/\beta)u$ of the domain D_2 would no longer be a straight line, and consequently D_3 would not be of the rectangular shape shown in Fig. 4.

The coefficients a_s in the expansion (2.26) can be calculated explicitly. The leading two are given by

$$a_0=\frac{1}{\sqrt{\lambda\mu}}\psi_{00}\eta_1, \tag{2.29}$$

$$\begin{aligned} a_1=\frac{1}{\sqrt{\lambda\mu}}\int_0^{\eta_1}\Bigg\{&\sqrt{\frac{2}{\lambda}}\psi_{10}\cos\eta+\sqrt{\frac{2}{\mu}}\psi_{01}\sin\eta \\ &-\frac{1}{4}\psi_{00}\Bigg[\left(\frac{2}{\lambda}\right)^{3/2}\phi_{30}\cos^3\eta+3\left(\frac{2}{\lambda}\right)\left(\frac{2}{\mu}\right)^{1/2}\phi_{21}\cos^2\eta\sin\eta \\ &+3\left(\frac{2}{\lambda}\right)^{1/2}\left(\frac{2}{\mu}\right)\phi_{12}\cos\eta\sin^2\eta+\left(\frac{2}{\mu}\right)^{3/2}\phi_{03}\sin^3\eta\Bigg]\Bigg\}\,d\eta, \end{aligned} \tag{2.30}$$

where $\psi_{ij}=\partial^{i+j}\psi/\partial u^i\,\partial v^j$ and a similar definition holds for ϕ_{ij}, all partial derivatives being evaluated at $(0, 0)$.

The asymptotic expansion of the integral in (2.16) will now follow from the theory for one-dimensional Fourier transforms. First, since in the present case $(0, 0)$ is a local minimum of $\phi(u, v)$, the lower limit m in (2.16) is zero. Second, since $\psi(u, v)$ vanishes outside the support of ψ and the support of ψ can be taken to be of any shape and arbitrarily small, the function $h(t)$ vanishes identically at the upper limit M. From (2.25), it is also evident that $h(t)$ is a C^∞-function for $0<t<M$. Therefore, upon termwise integration in the sense of Abel summability (see [9] or [11, p. 199]), we obtain

$$I(N)\sim e^{iNf(0,0)}\left[e^{i\pi/2}\frac{a_0}{N}+e^{i3\pi/4}\Gamma\left(\frac{3}{2}\right)\frac{a_1}{N^{3/2}}+e^{i\pi}\frac{a_2}{N^2}+\cdots\right] \tag{2.31}$$

as $N\to+\infty$.

The above analysis also works for the case when both λ and μ in (2.11) are negative, i.e., in the case when $(0, 0)$ is a local maximum for $f(x, y)$. The final result is

$$I(N)\sim e^{iNf(0,0)}\left[e^{-i\pi/2}\frac{a_0^*}{N}+e^{-i3\pi/4}\Gamma\left(\frac{3}{2}\right)\frac{a_1^*}{N^{3/2}}+e^{-i\pi}\frac{a_2^*}{N^2}+\cdots\right], \tag{2.32}$$

where

$$a_0^*=\frac{1}{\sqrt{\lambda\mu}}\psi_{00}\eta_1, \tag{2.33}$$

$$\begin{aligned} a_1^*=\frac{1}{\sqrt{\lambda\mu}}\int_0^{\eta_1}\Bigg\{&\psi_{10}\sqrt{\frac{2}{-\lambda}}\cos\eta+\psi_{01}\sqrt{\frac{2}{-\mu}}\sin\eta \\ &+\frac{1}{4}\psi_{00}\Bigg[\phi_{30}\left(\frac{2}{-\lambda}\right)^{3/2}\cos^3\eta+3\phi_{21}\left(\frac{2}{-\lambda}\right)\left(\frac{2}{-\mu}\right)^{1/2}\cos^2\eta\sin\eta \\ &+3\phi_{12}\left(\frac{2}{-\lambda}\right)^{1/2}\left(\frac{2}{-\mu}\right)\cos\eta\sin^2\eta \\ &+\phi_{03}\left(\frac{2}{-\mu}\right)^{3/2}\sin^3\eta\Bigg]\Bigg\}\,d\eta; \end{aligned} \tag{2.34}$$

cf. (2.29) and (2.30).

In terms of the original data f, g, h, and k, the leading term approximation in both (2.31) and (2.32) is

$$I(N) \sim e^{iNf(0,0)} \frac{g(0,0)}{\sqrt{ac-b^2}\,|k_x h_y - k_y h_x|^2} \tan^{-1}\left(\frac{1}{\beta}\sqrt{\alpha\gamma-\beta^2}\right)\frac{i}{N}, \tag{2.35}$$

where a, b, and c are given in (2.13) and α, β, and γ are given in (2.6)-(2.8). The partial derivatives in (2.35) are all evaluated at $(0, 0)$.

It is tempting to use the same method for the case of a saddlepoint at a corner, but it breaks down when one of the boundary curves $h(x, y) = 0$ or $k(x, y) = 0$ is tangent to the level curve $f(x, y) = 0$. To see this, we observe that the above procedure applies up to (2.16). Since the transformations (2.1) and (2.9) preserve the nature of the stationary point, if $(0, 0)$ is a saddlepoint of $f(x, y)$ then $(0, 0)$ is also a saddlepoint of $\phi(u, v)$, and consequently λ and μ have opposite signs. Suppose that $\lambda > 0$ and $\mu < 0$. Then the natural change of variable corresponding to (2.19) is

$$u = (2\xi)^{1/2}\frac{\cosh\eta}{\lambda^{1/2}}, \qquad v = (2\xi)^{1/2}\frac{\sinh\eta}{(-\mu)^{1/2}}.$$

The boundary curve $v = (\alpha/\beta)u$ of the domain D_2 is now mapped into the curve

$$\eta = \eta_1, \qquad \eta_1 = \tanh^{-1}\left(\frac{\alpha}{\beta}\sqrt{\frac{-\mu}{\lambda}}\right). \tag{2.36}$$

Suppose that the curve $h(x, y) = 0$ is tangent to the level curve $f(x, y) = 0$ at $(0, 0)$. Then (1.5) does not hold (with $x_0 = y_0 = 0$). In terms of the quantities α, β, and γ, this means that γ is zero, which in turn implies $\alpha/\beta = \sqrt{\lambda/-\mu}$, or equivalently, $\eta_1 = +\infty$ in (2.36). As a consequence, the arguments for the results corresponding to (2.26)-(2.28) are no longer valid. In particular, for the analogue of $F(\xi, \eta)$ as defined in (2.23), the estimate $F(\xi, \eta) = O(\xi^{3/2})$, as $\xi \to 0^+$, fails if $\eta_1 = \infty$.

Example. As a simple illustration, we consider the example in which

$$f(x, y) = (x^2 + xy + y^2)(1 - \tfrac{1}{2}x), \qquad g(x, y) = \cos(x + y), \tag{2.37}$$

and D is the domain bounded by the circles $(x-1)^2 + y^2 = 1$ and $x^2 + (y-1)^2 = 1$. It is easily verified that $f(x, y)$ has only two stationary points located at $(0, 0)$ and $(\frac{4}{3}, -\frac{2}{3})$. Hence, inside and on the boundary of D, $(0, 0)$ is the only stationary point of $f(x, y)$ and is a local minimum. Let $h(x, y) = x^2 - 2x + y^2$ and $k(x, y) = x^2 - 2y + y^2$. Simple calculations give

$$a = 2, \quad b = 1, \quad c = 2, \tag{2.38}$$

$$\alpha = 8, \quad \beta = 4, \quad \gamma = 8. \tag{2.39}$$

From (2.35) it follows that

$$\int\int_D g(x, y)\, e^{iNf(x,y)}\, dx\, dy \sim \frac{i}{48\sqrt{3}}\frac{\pi}{N} \tag{2.40}$$

as $N \to \infty$.

3. Saddlepoint at a corner: Transformation to canonical forms. As in § 2, we again assume that the corner point is at $(0, 0)$. By Theorem 4.1 in [7, p. 10], we may assume that the functions f, g, h, and k are extended to C^∞-functions in some neighborhood, say U of the origin. Although no actual use of the values of these extensions will be made outside the original domain D, the existence of these extensions will be a

convenience in our discussion. By shrinking the support of g, we may further assume that the support of g is contained in U. For definiteness, we also assume that for points in D we have $h(x, y) \geqq 0$ and $k(x, y) \geqq 0$, and that the angle of the corner of D at $(0, 0)$ is between 0 and π. The case of angles between π and 2π can be included by adding copies of the cases considered here. An illustration of the situation is depicted in Fig. 5.

Since $(0, 0)$ is a stationary point of $f(x, y)$, the initial terms in the Maclaurin expansion of $f(x, y)$ are

$$f(x, y) = f_{00} + f_{20}x^2 + f_{11}xy + f_{02}y^2 + \cdots,$$

where

$$f_{ij} = \frac{1}{i!\,j!} \frac{\partial^{i+j} f}{\partial x^i\, \partial y^j}(0, 0).$$

Without loss of generality, we may take $f_{00} = 0$, as it contributes only a factor $\exp\{iNf_{00}\}$ to the integral $I(N)$ in (1.1) (cf. (2.31)). Since it is always possible to eliminate the cross-product term xy by a linear change of variable, we may without loss of generality assume that

$$f(x, y) = f_{20}x^2 + f_{02}y^2 + O[(x^2 + y^2)^{3/2}] \tag{3.1}$$

as $(x, y) \to (0, 0)$.

The origin being a saddlepoint, the coefficients f_{20} and f_{02} have opposite signs. Hence, near $(0, 0)$, the level set, $f(x, y) = 0$ consists of two smooth curves intersecting at and only at $(0, 0)$; for a rigorous analysis, see the explanation following (3.9). We shall refer to these curves as the *critical curves* of $f(x, y)$. Also, we will assume that the following condition holds:

(C) Within the neighborhood U, each of the curves $h(x, y) = 0$ and $k(x, y) = 0$ meets the level set $f(x, y) = 0$ only at the point $(0, 0)$.

As indicated in § 1, previous treatments of a saddlepoint on a smooth boundary requires condition (1.5) to hold, i.e., that the critical curves of f not be tangent to the boundary of D at $(0, 0)$. Our condition (C) is much weaker; it allows the critical curves of f to meet the boundary of D tangentially, but rules out only situations in which a critical curve intersects the boundary infinitely often in the neighborhood U.

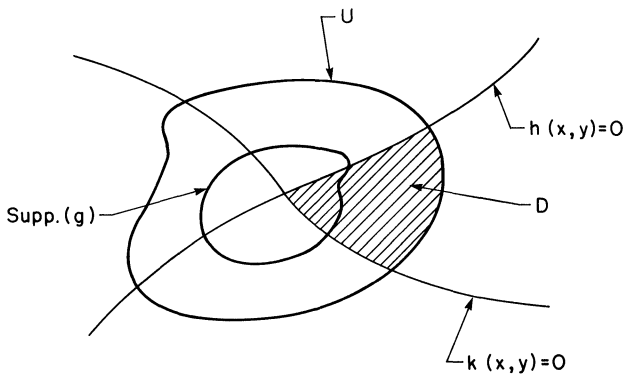


FIG. 5. *Neighborhood U.*

Since f_{20} and f_{02} have opposite signs, we may assume without loss of generality that $f_{20} > 0$ and $f_{02} < 0$. Define functions P and Q by

$$\begin{cases} P(x, y) = \dfrac{1}{x^2}[f(x, 0) + yf_y(x, 0) - f_{20}x^2] & \text{if } x \neq 0, \\ P(0, y) = \dfrac{1}{2} yf_{xxy}(0, 0), \end{cases} \tag{3.2}$$

$$\begin{cases} Q(x, y) = \dfrac{1}{y^2}[f(x, y) - f(x, 0) - yf_y(x, 0) - f_{02}y^2] & \text{if } y \neq 0, \\ Q(x, 0) = \dfrac{1}{2} f_{yy}(x, 0) - f_{02}. \end{cases} \tag{3.3}$$

By using l'Hôpital's rule, it is readily verified that each of P and Q is infinitely differentiable in U, and that $P(0, 0) = Q(0, 0) = 0$. It is also easily checked that

$$f(x, y) = f_{20}x^2 + x^2P(x, y) + f_{02}y^2 + y^2Q(x, y). \tag{3.4}$$

We now make the change of variables

$$s = x[f_{20} + P(x, y)]^{1/2}, \qquad t = y[-f_{02} - Q(x, y)]^{1/2}. \tag{3.5}$$

Clearly this transformation is infinitely differentiable in U, and from (3.4) we have

$$f(x, y) = s^2 - t^2. \tag{3.6}$$

Simple computation gives

$$\left.\frac{\partial(s, t)}{\partial(x, y)}\right|_{(0,0)} = \sqrt{-f_{20}f_{02}}. \tag{3.7}$$

By shrinking the neighborhood U in Fig. 5 if necessary, we may assume that the Jacobian $\partial(s, t)/\partial(x, y)$ is positive in U, and that the transformation (3.5) is one-to-one there. In terms of the variables s and t, the integral $I(N)$ in (1.1) becomes

$$I(N) = \int\int_{D_1} g_1(s, t) \exp(iN(s^2 - t^2))\, ds\, dt, \tag{3.8}$$

where D_1 is the image of D and

$$g_1(s, t) = g(x, y)\frac{\partial(x, y)}{\partial(s, t)}. \tag{3.9}$$

Let U_1 denote the image of U, and put $h_1(s, t) = h(x, y)$ and $k_1(s, t) = k(x, y)$. Parts of the curves $h_1(s, t) = 0$ and $k_1(s, t) = 0$ determine the boundary of D_1 near $(0, 0)$. For points (s, t) in D_1, we have $h_1(s, t) \geqq 0$ and $k_1(s, t) \geqq 0$. In view of (3.6), the set in U_1 corresponding to the level set $f(x, y) = 0$ is determined by $s^2 - t^2 = 0$. Consequently, as claimed earlier, $f(x, y) = 0$ determines, within U, two smooth curves intersecting non-tangentially at and only at $(0, 0)$. Since the transformation (3.5) is one-to-one in U, condition (C) implies that the curves $h_1(s, t) = 0$ and $k_1(s, t) = 0$ intersect each other, and the lines $s \pm t = 0$, at and only at $(0, 0)$ inside U_1.

To derive the asymptotic expansion of the integral (3.8), we make one further simple, linear change of variables:

$$u = s - t, \qquad v = s + t \tag{3.10}$$

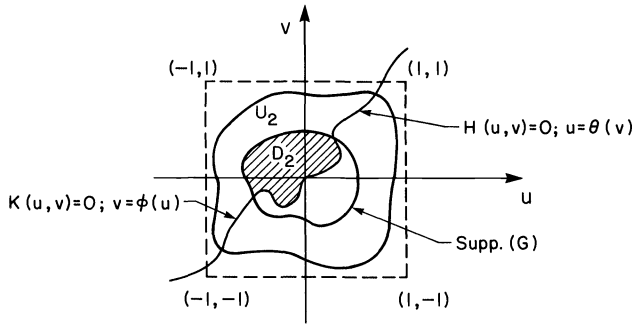


FIG. 6. *Neighborhood* U_2.

so that

$$I(N) = \int\int_{D_2} G(u, v)\, e^{iNuv}\, du\, dv, \tag{3.11}$$

where D_2 is the image of D_1 and

$$G(u, v) = g_1(s, t)\left|\frac{\partial(s, t)}{\partial(u, v)}\right| = \frac{1}{2}\, g_1(s, t). \tag{3.12}$$

Let U_2 denote the image of U_1, and put $H(u, v) = h_1(s, t)$ and $K(u, v) = k_1(s, t)$. For $(u, v) \in D_2$, we have $H(u, v) \geqq 0$ and $K(u, v) \geqq 0$.

By hypothesis, the gradients of $h(x, y)$ and $k(x, y)$ are both nonvanishing in U. Hence, in view of the nonsingularity of the transformations (3.5) and (3.10), each of the functions $H(u, v)$ and $K(u, v)$ has a nonvanishing gradient in U_2. As a consequence, locally within U_2, the equations $H(u, v) = 0$ and $K(u, v) = 0$ either determine v as a smooth function of u, say $v = \phi(u)$, or u as a function of v, say $u = \theta(v)$. We assume that U_2 is sufficiently small so that each of the curves $H(u, v) = 0$ and $K(u, v) = 0$ has such a representation globally within U_2. We also assume that U_2 is contained in the square with vertices at $(\pm 1, \pm 1)$; see Fig. 6. The integral (3.11) can then be written as a sum of not more than four terms, each of which is of one of the following forms:

$$\pm\int_0^{\pm 1}\int_0^{\phi(u)} G(u, v)\, e^{iNuv}\, dv\, du, \tag{3.13}$$

$$\pm\int_0^{\pm 1}\int_0^{\theta(v)} G(u, v)\, e^{iNuv}\, du\, dv, \tag{3.14}$$

$$\int_0^{\pm 1}\int_0^{\pm 1} G(u, v)\, e^{iNuv}\, du\, dv. \tag{3.15}$$

Note that the integral in (3.14) can be converted to that in (3.13). Thus our problem is solved if we can derive asymptotic expansions for the integrals in (3.13) and (3.15).

4. Saddlepoint at a corner: Derivation of expansions. We first consider the integral in (3.13), which we will rewrite as

$$I_1(N) = \int_0^1\int_0^{\phi(u)} G_0(u, v)\, e^{iNuv}\, dv\, du. \tag{4.1}$$

Recall that the function $\phi(u)$ corresponds to a part of the boundary of D_2 determined either by $H(u, v) = 0$ or $K(u, v) = 0$, and that it is a C^∞-function with $\phi(0) = 0$ and

$|\phi(u)| < 1$ for $0 \leqq u \leqq 1$. In view of (3.10), the curves corresponding to the critical curves of the original phase function $f(x, y)$ are determined by $u = 0$ and $v = 0$. By condition (C) in § 3, we have $\phi(u) \neq 0$ if $u \neq 0$. The function $G_0(u, v)$ is infinitely differentiable and has compact support inside the square with vertices at $(\pm 1, \pm 1)$.

Define

$$H_0^{(1)} = \frac{1}{v}[G_0(u, v) - G_0(u, 0)],$$
$$H_0^{(2)} = \frac{1}{u}[G_0(u, 0) - G_0(0, 0)], \tag{4.2}$$

and put $\mathbf{H}_0 = (H_0^{(1)}, H_0^{(2)})$. It is easily verified that

$$G_0(u, v) = G_0(0, 0) + (v, u) \cdot \mathbf{H}_0, \tag{4.3}$$

$$\nabla \cdot (\mathbf{H}_0 e^{iNuv}) = (\nabla \cdot \mathbf{H}_0) e^{iNuv} + iN[(v, u) \cdot \mathbf{H}_0] e^{iNuv}. \tag{4.4}$$

Inserting (4.3) and (4.4) into (4.1), we obtain

$$\begin{aligned} I_1(N) = G_0(0, 0) &\int_0^1 \int_0^{\phi(u)} e^{iNuv}\, dv\, du \\ &+ \frac{1}{iN} \int_0^1 \int_0^{\phi(u)} \nabla \cdot (\mathbf{H}_0 e^{iNuv})\, dv\, du \\ &+ \frac{i}{N} \int_0^1 \int_0^{\phi(u)} (\nabla \cdot \mathbf{H}_0) e^{iNuv}\, dv\, du. \end{aligned} \tag{4.5}$$

The first integral on the right can be written as

$$\frac{1}{iN} \int_0^1 \frac{1}{u}[e^{iNu\phi(u)} - 1]\, du. \tag{4.6}$$

In view of the fact that $G_0(u, v)$, and hence also $H_0^{(1)}(u, v)$, vanishes on the line $u = 1$, we readily verify by using Green's theorem that the second integral on the right-hand side of (4.5) is equal to

$$\begin{aligned} -\int_0^1 H_0^{(2)}(u, 0)\, du &+ \int_0^1 H_0^{(2)}(u, \phi(u))\, e^{iNu\phi(u)}\, du \\ &- \int_0^1 H_0^{(1)}(u, \phi(u))\, e^{iNu\phi(u)} \phi'(u)\, du, \end{aligned}$$

which in turn is equal to

$$\begin{aligned} -\int_0^1 \frac{1}{u}[G_0(u, 0) - G_0(0, 0)]\, du &\\ + \int_0^1 \frac{1}{u}[G_0(u, 0) - G_0(0, 0)]\, e^{iNu\phi(u)}\, du &\\ - \int_0^1 \frac{1}{\phi(u)}[G_0(u, \phi(u)) - G_0(u, 0)]\, e^{iNu\phi(u)} \phi'(u)\, du. & \end{aligned} \tag{4.7}$$

Combining the integral in (4.6) with the first two integrals in (4.7) gives

$$I_1(N)=\frac{1}{iN}\int_0^1 G_0(u,0)\frac{1}{u}[e^{iNu\phi(u)}-1]\,du$$

$$-\frac{1}{iN}\int_0^1 \frac{1}{\phi(u)}[G_0(u,\phi(u))-G_0(u,0)]\,e^{iNu\phi(u)}\phi'(u)\,du \tag{4.8}$$

$$+\frac{i}{N}\int_0^1\int_0^{\phi(u)} G_1(u,v)\,e^{iNuv}\,dv\,du,$$

where $G_1(u,v)=\nabla\cdot\mathbf{H}_0$. Note that the last integral on the right-hand side of (4.8) has exactly the same form as the integral $I_1(N)$ in (4.1), except that it is now multiplied by a factor $1/N$. Thus we may repeat the above procedure with $G_0(u,v)$ replaced by $G_1(u,v)$. Each time we perform this procedure, we produce a factor $1/N$ in the new "remainder term." Therefore, it suffices to consider just the two one-dimensional integrals in (4.8). To the first integral in (4.8), we apply an integration by parts. The result is

$$I_1(N)=-\int_0^1 (\log u)G_0(u,0)[\phi(u)+u\phi'(u)]\,e^{iNu\phi(u)}\,du$$

$$-\frac{1}{iN}\int_0^1 (\log u)\frac{\partial G_0}{\partial u}(u,0)\,e^{iNu\phi(u)}\,du+\frac{1}{iN}\int_0^1 (\log u)\frac{\partial G_0}{\partial u}(u,0)\,du$$

$$-\frac{1}{iN}\int_0^1 \frac{1}{\phi(u)}[G_0(u,\phi(u))-G_0(u,0)]\phi'(u)\,e^{iNu\phi(u)}\,du \tag{4.9}$$

$$+\frac{i}{N}\int_0^1\int_0^{\phi(u)} G_1(u,v)\,e^{iNuv}\,dv\,du.$$

Note that the third integral in (4.9) is independent of N; hence it is simply a coefficient of $1/N$ in the final expansion of $I_1(N)$. Here a comment is in order. Recall that the function $G_0(u,v)$ may involve an arbitrary C^∞-function used in isolating $(0,0)$ from other critical points. Thus it may seem strange that the coefficient

$$\int_0^1 (\log u)\frac{\partial G_0}{\partial u}(u,0)\,du$$

of $1/N$ in (4.9) depends on the values of $G_0(u,v)$ on the whole interval $[0,1]$. However, since $I_1(N)$ is only part of the original integral $I(N)$ in (1.1), there are contributions from other parts of $I(N)$ that in many cases will cancel this term in (4.9); see § 6 for a discussion of a complete case.

We now examine the three one-dimensional integrals in (4.9) which involve N. Observe that the first two are of the form

$$A(N)=\int_0^c (\log u)g(u)\,e^{iNh(u)}\,du, \tag{4.10}$$

to which we can apply the following result of Erdélyi [4], as corrected by McKenna [8].

Let $g(u)$ be an M times continuously differentiable function in $[0,c)$, which vanishes in some neighborhood of c; let $h(u)$ be differentiable and

$$h'(u)=u^{\rho-1}h_1(u), \tag{4.11}$$

where $\rho \geqq 1$, and $h_1(u)$ is positive and M times continuously differentiable for $0 \leqq u < c$. Introduce a variable w defined by

$$w^\rho = h(u) - h(0) \tag{4.12}$$

and put $w_1 = [h(c) - h(0)]^{1/\rho}$. Equation (4.12) can be inverted to give an expression of the form

$$u = w\left\{\left[\frac{\rho}{h_1(0)}\right]^{1/\rho} + R(w)\right\}, \tag{4.13}$$

where $R(0) = 0$, and $R(w)$ is M times continuously differentiable in $[0, w_1)$. Define functions $k_0(w)$ and $k_1(w)$ by

$$k_0(w) = g(u)\frac{du}{dw}, \tag{4.14}$$

$$k_1(w) = k_0(w)\log\left(\left[\frac{\rho}{h_1(0)}\right]^{1/\rho} + R(w)\right). \tag{4.15}$$

THEOREM (Erdélyi [4], McKenna [8]). *With the above conditions,*

$$A(N) = e^{iNh(0)}\sum_{n=0}^{M-1}\frac{1}{n!\rho}\Gamma\left(\frac{n+1}{\rho}\right) \cdot \exp\left(i(n+1)\pi/2\rho\right)A_n N^{-(n+1)/\rho} + o(N^{-M/\rho}) \tag{4.16}$$

as $N \to \infty$, *where*

$$A_n = k_0^{(n)}(0)\frac{1}{\rho}\left[\psi\left(\frac{n+1}{\rho}\right) - \log N + i\frac{\pi}{2}\right] + k_1^{(n)}(0), \tag{4.17}$$

ψ being the logarithmic derivative of the Gamma function.

In our application to the integrals in (4.9), we have $h(u) = u\phi(u)$ with $\phi(0) = 0$. We will assume

$$\phi(u) = u^{\rho-1}\phi_1(u), \tag{4.18}$$

where ρ is an integer, $\rho \geqq 2$, and $\phi_1(0) \neq 0$. This assumption is slightly stronger than condition (C); specifically, (C) would allow $\phi(u) = e^{-1/u}\phi_1(u)$, but (4.18) does not. We will first consider the case $\phi(u)$ positive in a neighborhood of the origin. By shrinking the support of $G_0(u, v)$, we may assume without loss of generality that

$$h'(u) = \phi(u) + u\phi'(u) = u^{\rho-1}[\rho\phi_1(u) + u\phi_1'(u)]$$

is positive in an interval $[0, c)$ for some $c \leqq 1$ so that the hypothesis of the above theorem for the phase function $h(u)$ is satisfied.

Since $h(u) = u\phi(u)$ in our case, coupling (4.12) and (4.18) gives

$$w = u[\phi_1(u)]^{1/\rho}, \tag{4.19}$$

from which we can compute dw/du and obtain

$$\frac{du}{dw} = \frac{\rho[\phi_1(u)]^{1-(1/\rho)}}{\rho\phi_1(u) + u\phi_1'(u)}. \tag{4.20}$$

A straightforward calculation then yields

$$\frac{du}{dw}(0) = [\phi_1(0)]^{-1/\rho}, \tag{4.21}$$

$$\frac{d^2u}{dw^2}(0) = -\frac{2\phi_1'(0)}{\rho[\phi_1(0)]^{1+(2/\rho)}}. \tag{4.22}$$

From (4.19), we also have

$$\log u = \log w - \frac{1}{\rho}\log \phi_1(u). \tag{4.23}$$

Thus the function $R(w)$ in (4.13) satisfies

$$\log\left(\left[\frac{\rho}{h_1(0)}\right]^{1/\rho} + R(w)\right) = -\frac{1}{\rho}\log \phi_1(u).$$

In the case of the first integral in (4.9), the amplitude function $g(u)$ in (4.10) is given by

$$g(u) = G_0(u, 0)[\phi(u) + u\phi'(u)].$$

Hence, in view of (4.18), the functions $k_0(w)$ and $k_1(w)$ in (4.14) and (4.15) are given by

$$k_0(w) = G_0(u, 0)u^{\rho-1}[\rho\phi_1(u) + u\phi_1'(u)]\frac{du}{dw},$$

$$k_1(w) = -\frac{1}{\rho}k_0(w)\log \phi_1(u).$$

Clearly, these functions vanish at least to order $\rho - 1$. Elementary computation yields

$$k_0^{(\rho-1)}(0) = \rho!\, G_0(0, 0),$$

$$k_0^{(\rho)}(0) = \frac{\rho!\rho}{[\phi_1(0)]^{1/\rho}}\frac{\partial G_0}{\partial u}(0, 0),$$

$$k_1^{(\rho-1)}(0) = -\frac{1}{\rho}k_0^{(\rho-1)}(0)\log \phi_1(0),$$

$$k_1^{(\rho)}(0) = -\frac{1}{\rho}k_0^{(\rho)}(0)\log \phi_1(0) - k_0^{(\rho-1)}(0)\frac{\phi_1'(0)}{[\phi_1(0)]^{(\rho+1)/\rho}}.$$

The above theorem of Erdélyi and McKenna then gives

$$\begin{aligned}\int_0^1 (\log u)G_0(u, 0)[\phi(u) + u\phi'(u)]\, e^{iNu\phi(u)}\, du \\ = (a_0 \log N + b_0)N^{-1} + (a_1 \log N + b_1)N^{-1-(1/\rho)} + o(N^{-1-(1/\rho)}),\end{aligned} \tag{4.24}$$

where

$$\begin{aligned} a_0 &= -\frac{i}{\rho}G_0(0, 0), \\ b_0 &= \frac{i}{\rho}G_0(0, 0)\left[\psi(1) + i\frac{\pi}{2} - \log \phi_1(0)\right], \\ a_1 &= -\frac{1}{\rho!\rho^2}\Gamma\left(\frac{\rho+1}{\rho}\right)\exp(i\pi(\rho+1)/2\rho)k_0^{(\rho)}(0), \\ b_1 &= \frac{1}{\rho!\rho^2}\Gamma\left(\frac{\rho+1}{\rho}\right)\exp(i\pi(\rho+1)/2\rho)\left\{k_0^{(\rho)}(0)\left[\psi\left(\frac{\rho+1}{2}\right) + i\frac{\pi}{2}\right] + \rho k_1^{(\rho)}(0)\right\}. \end{aligned} \tag{4.25}$$

In exactly the same manner, we have

$$\frac{1}{iN}\int_0^1 (\log u)\frac{\partial G_0}{\partial u}(u,0)\, e^{iNu\phi(u)}\, du \tag{4.26}$$
$$= (c_1 \log N + d_1)N^{-1-(1/\rho)} + o(N^{-1-(1/\rho)}),$$

where

$$c_1 = \frac{1}{\rho^2}\Gamma\left(\frac{1}{\rho}\right)\exp\left(i\pi(\rho+1)/2\rho\right)[\phi_1(0)]^{-1/\rho}\frac{\partial G_0}{\partial u}(0,0), \tag{4.27}$$
$$d_1 = c_1\left[\psi\left(\frac{1}{\rho}\right) + i\frac{\pi}{2} - \log\phi_1(0)\right].$$

For the penultimate integral on the right-hand side of (4.9), we need only the usual version of the stationary phase approximation [4, Thm. 3], and the result is

$$\frac{1}{iN}\int_0^1 \frac{1}{\phi(u)}[G_0(u,\phi(u)) - G_0(u,0)]\phi'(u)\, e^{iNu\phi(u)}\, du \tag{4.28}$$
$$= \frac{1}{\rho}\Gamma\left(\frac{1}{\rho}\right)\exp\left(-i\pi(\rho-1)/(2\rho)\right)\frac{\partial G_0}{\partial v}(0,0)\phi'(0)$$
$$\cdot[\phi_1(0)]^{-1/\rho}N^{-1-(1/\rho)} + o(N^{-1-(1/\rho)}).$$

(Note that $\rho \geqq 2$.) If $\rho > 2$, then $\phi'(0) = 0$ and hence the terms in (4.28) are of asymptotically smaller order than the terms in (4.26). A combination of (4.9), (4.24), (4.26), and (4.28) gives

$$I_1(N) \sim (\alpha_1 \log N + \beta_1)N^{-1} + (\gamma_1 \log N + \delta_1)N^{-1-(1/\rho)} + \cdots, \tag{4.29}$$

where

$$\alpha_1 = \frac{i}{\rho}G_0(0,0), \tag{4.30}$$

$$\beta_1 = -\frac{i}{\rho}G_0(0,0)\left[\psi(1) + i\frac{\pi}{2} - \log\phi_1(0)\right] - i\int_0^1 (\log u)\frac{\partial G_0}{\partial u}(u,0)\, du. \tag{4.31}$$

The coefficients γ_1 and δ_1 can be expressed in terms of the coefficients a_1, b_1, c_1, and d_1 given in (4.25) and (4.27). More precisely, we have $\gamma_1 = -a_1 - c_1$, $\delta_1 = -b_1 - c_1$, if $\rho > 2$, and

$$\delta_1 = -b_1 - c_1 - \frac{1}{\rho}\Gamma\left(\frac{1}{\rho}\right)\exp\left(-i\pi(\rho-1)/(2\rho)\right)\frac{\partial G_0}{\partial v}(0,0)\phi'(0)[\phi_1(0)]^{-1/\rho}$$

if $\rho = 2$.

5. Some variations. Recall that the integral $I_1(N)$ in (4.1) is only one of those in (4.13)-(4.15) that must be considered for the original problem. We now examine a number of simple variations of the computations leading to (4.29). First, note that by taking complex conjugates we have

$$I_1(-N) = \int_0^1\int_0^{\phi(u)} G_0(u,v)\, e^{-iNuv}\, dv\, du \tag{5.1}$$
$$\sim (\bar{\alpha}_1 \log N + \bar{\beta}_1)N^{-1} + (\bar{\gamma}_1 \log N + \bar{\delta}_1)N^{-1-(1/\rho)} + \cdots,$$

where a bar $^-$ denotes the complex conjugate.

Next, we consider the case in which $\phi(u)$ is negative for $0<u\leqq 1$. As in (4.18), we again write $\phi(u)=u^{\rho-1}\phi_1(u)$. Then $\phi_1(u)$ is negative and $|\phi_1(u)|=-\phi_1(u)$. The integral corresponding to (4.1) is now

$$\begin{aligned} I_2(N) &= \int_0^1 \int_{\phi(u)}^0 G_0(u, v)\, e^{iNuv}\, dv\, du \\ &= \int_0^1 \int_0^{|\phi(u)|} G_0(u, -v)\, e^{-iNuv}\, dv\, du. \end{aligned} \tag{5.2}$$

From (5.1) it follows that

$$I_2(N)\sim(\alpha_2\log N+\beta_2)N^{-1}+(\gamma_2\log N+\delta_2)N^{-1-(1/\rho)}+\cdots \tag{5.3}$$

as $N\to\infty$, where

$$\alpha_2=-\frac{i}{\rho}G_0(0,0), \tag{5.4}$$

$$\beta_2=\frac{i}{\rho}G_0(0,0)\left[\psi(1)-i\frac{\pi}{2}-\log|\phi_1(0)|\right]+i\int_0^1(\log u)\frac{\partial G_0}{\partial u}(u,0)\,du. \tag{5.5}$$

Explicit formulas can also be given for the coefficients γ_2 and δ_2.

If the domain of integration is bounded by the v-axis, the line $v=1$, and a smooth curve $u=\theta(v)$, where $\theta(v)$ is positive, then the integral corresponding to (4.1) is

$$I_3(N)=\int_0^1\int_0^{\theta(v)}G_0(u,v)\,e^{iNuv}\,du\,dv. \tag{5.6}$$

We assume, as in (4.18), that $\theta(v)=v^{\rho-1}\theta_1(v)$, where ρ is an integer greater than or equal to 2. Interchanging the variables u and v gives

$$I_3(N)=\int_0^1\int_0^{\theta(u)}G_0^*(u,v)\,e^{iNuv}\,dv\,du,$$

where $G_0^*(u,v)=G_0(v,u)$. By (4.29), we have

$$I_3(N)\sim(\alpha_3\log N+\beta_3)N^{-1}+(\gamma_3\log N+\delta_3)N^{-1-(1/\rho)}+\cdots, \tag{5.7}$$

where $\alpha_3=iG_0(0,0)/\rho$ and

$$\beta_3=-\frac{i}{\rho}G_0(0,0)\left[\psi(1)+i\frac{\pi}{2}-\log\theta_1(0)\right]-i\int_0^1(\log v)\frac{\partial G_0}{\partial v}(0,v)\,dv. \tag{5.8}$$

Similarly, if $\theta(v)<0$ for $0<v\leqq 1$, then by (5.3)

$$I_4(N)=\int_0^1\int_{\theta(v)}^0 G_0(u,v)\,e^{iNuv}\,du\,dv\sim(\alpha_4\log N+\beta_4)N^{-1}+\cdots, \tag{5.9}$$

where $\alpha_4=-iG_0(0,0)/\rho$ and

$$\beta_4=\frac{i}{\rho}G_0(0,0)\left[\psi(1)-i\frac{\pi}{2}-\log|\theta_1(0)|\right]+i\int_0^1(\log v)\frac{\partial G_0}{\partial v}(0,v)\,dv. \tag{5.10}$$

There are many other possibilities of this type. However, we will consider just one more; the reader can easily derive corresponding results for other types. For this case, we suppose we have the domain of integration in the second quadrant; specifically, we consider the integral

$$I_5(N)=\int_{-1}^0\int_0^{\phi(u)}G_0(u,v)\,e^{iNuv}\,dv\,du, \tag{5.11}$$

where the function $\tilde{\phi}(u)=\phi(-u)$ is positive for $0<u\leqq 1$ and of the form $\tilde{\phi}(u)=u^{\rho-1}\tilde{\phi}_1(u)$ with $\rho\geqq 2$; cf. (4.18). Changing variables by replacing u by $-u$, we get

$$I_5(N)=\int_0^1\int_0^{\tilde{\phi}(u)} G_0(-u, v)\, e^{-iNuv}\, dv\, du.$$

From (5.1) it follows that

$$I_5(N)\sim(\alpha_5\log N+\beta_5)N^{-1}+\cdots, \tag{5.12}$$

where $\alpha_5=-iG_0(0,0)/\rho$ and

$$\beta_5=\frac{i}{\rho}G_0(0,0)\left[\psi(1)-i\frac{\pi}{2}-\log\tilde{\phi}_1(0)\right]-i\int_{-1}^0(\log|u|)\frac{\partial G_0}{\partial u}(u,0)\,du. \tag{5.13}$$

Finally, we consider integrals of the form (3.15). Specifically, we consider the integral

$$I_6(N)=\int_0^1\int_0^1 G_0(u, v)\, e^{iNuv}\, du\, dv, \tag{5.14}$$

where G_0 has compact support inside the square with vertices at $(\pm1, \pm1)$. Clearly, we may write

$$I_6(N)=I_1(N)+I_3(N),$$

where $I_1(N)$ and $I_3(N)$ are the integrals given in (4.1) and (5.6), respectively, with $\phi(u)=u$ and $\theta(v)=v$. Thus, in both cases, we have $\rho=2$, and $\phi_1(u)=1$ and $\theta_1(v)=1$. By (4.29) and (5.7) we obtain

$$I_6(N)\sim(\alpha_6\log N+\beta_6)N^{-1}+\cdots, \tag{5.15}$$

where $\alpha_6=iG_0(0,0)$ and

$$\beta_6=-iG_0(0,0)\left[\psi(1)+i\frac{\pi}{2}\right]-i\int_0^1(\log w)\left[\frac{\partial G_0}{\partial u}(w,0)+\frac{\partial G_0}{\partial v}(0,w)\right]dw. \tag{5.16}$$

In exactly the same manner, we may write

$$I_7(N)\equiv\int_{-1}^0\int_0^1 G_0(u, v)\, e^{iNuv}\, dv\, du=I_5(N)+I_4(N), \tag{5.17}$$

where $I_5(N)$ and $I_4(N)$ are the integrals given in (5.11) and (5.9), respectively, with $\phi(u)=-u$ and $\theta(v)=-v$. Coupling (5.9) and (5.12) gives

$$I_7(N)\sim(\alpha_7\log N+\beta_7)N^{-1}+\cdots, \tag{5.18}$$

where $\alpha_7=-iG_0(0,0)$ and

$$\beta_7=iG_0(0,0)\left[\psi(1)-i\frac{\pi}{2}\right]+i\int_0^1(\log w)\left[\frac{\partial G_0}{\partial v}(0,w)-\frac{\partial G_0}{\partial u}(-w,0)\right]dw. \tag{5.19}$$

It is interesting to observe that by adding (5.15) and (5.18) we have

$$\int_{-1}^1\int_0^1 G_0(u, v)\, e^{iNuv}\, dv\, du\sim\left[\pi G_0(0,0)-i\int_{-1}^1(\log|u|)\frac{\partial G_0}{\partial u}(u,0)\,du\right]\frac{1}{N} \tag{5.20}$$

as $N\to\infty$. Carrying out calculations analogous to (5.15) and (5.18) for the other two quadrants, and adding, gives

$$\int_{-1}^1\int_{-1}^1 G_0(u, v)\, e^{iNuv}\, du\, dv\sim\frac{2\pi}{N}G_0(0,0) \tag{5.21}$$

as $N \to \infty$. Note that $(0, 0)$ is a saddlepoint of the phase function in (5.21) and lies in the interior of the support of G_0. Hence (5.21) follows directly from (1.3). However, (5.20) does not follow from (1.4), although $(0, 0)$ is a boundary saddlepoint. This is due to the fact that the boundary curve $v = 0$ in (5.20) is tangent to, and in fact coincides with, a level curve of the phase function at $(0, 0)$.

6. The final solution. From the above discussion, it is evident that the solution to our original problem depends on many factors. In this section we consider only a particular case, in which we derive the first term of the asymptotic expansion.

First we recall that the functions $\phi(u)$ and $\theta(v)$ in (4.1) and (5.6) are determined implicitly by the equations $H(u, v) = 0$ and $K(u, v) = 0$, which represent parts of the boundary of D_2 in a neighborhood of $(0, 0)$ (see Fig. 6). If these boundary curves are tangent to a coordinate axis, i.e., if the parameter ρ in (4.18) is greater than 2, then determining $\phi_1(0)$ or $\theta_1(0)$ will require higher-order implicit differentiation. Although this can be done for any specific example, it is difficult to give a formula for the general case. We will therefore make some simplifying assumptions. More precisely, we assume that after the transformations in (3.5) and (3.10), the domain D_2 is as depicted in Fig. 7, where $0 < \phi(u) < 1$ for $0 < u \leqq 1$, $\phi'(0) = \phi_1(0) > 0$, $-1 < \theta(v) < 0$ for $0 < v \leqq 1$, and $\theta'(0) = \theta_1(0) < 0$. In this case, the value of ρ for both ϕ and θ is 2. Since ϕ and θ are defined implicitly by $H(u, v) = 0$ and $K(u, v) = 0$, respectively, we have

$$\phi_1(0) = -\frac{H_u(0, 0)}{H_v(0, 0)}, \qquad \theta_1(0) = -\frac{K_v(0, 0)}{K_u(0, 0)}. \tag{6.1}$$

The integral (3.11) can be written as

$$\int\int_{D_2} G(u, v)\, e^{iNuv}\, du\, dv = I_6(N) - I_1(N) + I_4(N), \tag{6.2}$$

where $I_1(N)$ is given by (4.1), $I_4(N)$ by (5.9), and $I_6(N)$ by (5.14), with $G_0(u, v) = G(u, v)$ in all cases. A combination of (4.29), (5.9), and (5.15) gives

$$\int\int_{D_2} G(u, v)\, e^{iNuv}\, du\, dv \sim G(0, 0)[\pi - i \log \phi_1(0) - i \log |\theta_1(0)|]\, \frac{1}{2N} \tag{6.3}$$

as $N \to \infty$, where $\phi_1(0)$ and $\theta_1(0)$ are given in (6.1). Note that the terms involving integrals in (4.31), (5.10), and (5.16) all cancel out in the calculation of (6.3), confirming a remark made in § 4 following (4.9). Now, from § 3, we recall that

$$G(0, 0) = \tfrac{1}{2} g_1(0, 0) = \tfrac{1}{2} g(0, 0)/\sqrt{-f_{20} f_{02}}. \tag{6.4}$$

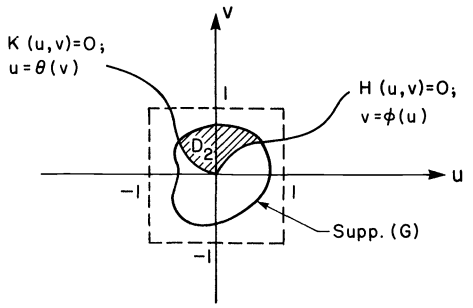


FIG. 7

Using (6.1), (3.10), and (3.5), we also find

$$\phi_1(0) = \frac{-\sqrt{-f_{02}}\, h_x(0,0) + \sqrt{f_{20}}\, h_y(0,0)}{\sqrt{-f_{02}}\, h_x(0,0) + \sqrt{f_{20}}\, h_y(0,0)}, \tag{6.5}$$

$$\theta_1(0) = \frac{\sqrt{-f_{02}}\, k_x(0,0) + \sqrt{f_{20}}\, k_y(0,0)}{-\sqrt{-f_{02}}\, k_x(0,0) + \sqrt{f_{20}}\, k_y(0,0)}. \tag{6.6}$$

Substituting (6.4)-(6.6) in (6.3), we obtain the leading term of an asymptotic expansion of the original integral $I(N)$ in terms of the original data f, g, h, and k.

The case where the boundary of D is smooth at the saddlepoint $(0, 0)$, but not tangent to a critical curve of the phase function $f(x, y)$, can be recovered from the result above. In the notation of § 5, a typical case of this would give

$$I(N) = I_6(N) + I_2(N) + I_4(N), \tag{6.7}$$

where $I_2(N)$, $I_4(N)$, and $I_6(N)$ are given, respectively, by (5.2), (5.9), and (5.14), with $G_0(u, v) = G(u, v)$ in all three cases. Since the boundary is not tangent to a critical curve of $f(x, y)$, the value of ρ for both ϕ and θ is again 2. The condition that the boundary of D be smooth at $(0, 0)$, i.e., have a tangent at that point, can be expressed in terms of ϕ_1 and θ_1 as

$$\phi_1(0) = \theta_1(0)^{-1}. \tag{6.8}$$

Noting that $\log|\phi_1(0)| = -\log|\theta_1(0)|$, it follows from (5.3), (5.9), and (5.15) that

$$I(N) \sim g(0,0)[-f_{20}f_{02}]^{-1/2}\left(\frac{\pi}{2N}\right) \tag{6.9}$$

as $N \to \infty$, in agreement with (1.4).

From the above discussion, it is now evident that our method is applicable when the boundary curve is smooth in a neighborhood of the saddlepoint and is tangent to a critical curve of f. For example, if in the uv-coordinates the boundary is smooth with a horizontal tangent at $(0, 0)$, then the integral can be expressed in the form $I_6(N) - I_1(N) + I_7(N) - I_5(N)$. The functions ϕ in $I_1(N)$ and $I_5(N)$ may be different, but both will have $\phi'(0) = 0$; or equivalently, $\rho > 2$. We emphasize that the result in such a case will be different from (6.9).

7. An example. As an illustration of the above calculation, we work out the first two terms in the asymptotic expansion of the integral

$$I(N) = \int_0^1 \int_{-x}^{x} \exp(iN \cosh x \cos y)\, dy\, dx, \tag{7.1}$$

which we will write in the form

$$I(N) = e^{iN} \int\int_D e^{iNf(x,y)}\, dx\, dy, \tag{7.2}$$

where D is the triangle with vertices at $(0, 0)$, $(1, -1)$, and $(1, 1)$, and

$$f(x, y) = \cosh x \cos y - 1 = \frac{x^2}{2} - \frac{y^2}{2} + O[(x^2+y^2)^2] \tag{7.3}$$

as $(x, y) \to (0, 0)$. Clearly, $f(x, y)$ has one and only one stationary point inside and on the boundary of D; it occurs at $(0, 0)$ and is a saddlepoint. Note that $(0, 0)$ is also a corner point of the domain D. By using a partition of unity (or neutralizers), we can

isolate it from the other two corner points $(1,-1)$ and $(1,1)$. Since the contribution from a corner point that is not a critical point of the phase function is $O(N^{-2})$ (see, e.g., [2]), we can rewrite (7.2) as

$$I(N)=e^{iN}\int\int_D g(x,y)\,e^{iNf(x,y)}\,dx\,dy+O(N^{-2}), \tag{7.4}$$

where $g(x,y)$ is a C^∞-function that is equal to 1 in a neighborhood of the origin and vanishes identically outside a disk with radius less than 1. From (7.3) it is clear that $f_{20}=\frac{1}{2}$ and $f_{02}=-\frac{1}{2}$. The functions P and Q in (3.2) and (3.3) are, in the present case, given by

$$P(x,y)=\frac{1}{x^2}\left[\cosh x-1-\frac{x^2}{2!}\right]=\frac{x^2}{4!}+O(x^4),$$

$$Q(x,y)=\frac{1}{y^2}\left[\cosh x(\cos y-1)+\frac{y^2}{2}\right]=-\frac{x^2}{(2!)^2}+\frac{y^2}{4!}+O[(x^2+y^2)^2].$$

Note that in this example $P(x,y)$ is independent of y. Define

$$\begin{aligned} p(x,y)&=\left[\frac{1}{2}+P(x,y)\right]^{1/2}=\left[\frac{1}{2}+\frac{x^2}{4!}+O(x^4)\right]^{1/2},\\ q(x,y)&=\left[\frac{1}{2}-Q(x,y)\right]^{1/2}=\left[\frac{1}{2}+\frac{x^2}{4}-\frac{y^2}{4!}+O\{(x^2+y^2)^2\}\right]^{1/2} \end{aligned} \tag{7.5}$$

so that the transformation (3.5) becomes

$$s=xp(x,y),\qquad t=yq(x,y). \tag{7.6}$$

Upon making the transformations (7.6) and (3.10), we have

$$I(N)=e^{iN}\int\int_{D_2} G(u,v)\,e^{iNuv}\,du\,dv, \tag{7.7}$$

where D_2 is the image of D and

$$G(u,v)=g(x,y)\frac{\partial(x,y)}{\partial(s,t)}\frac{\partial(s,t)}{\partial(u,v)}.$$

From (7.5), (7.6), and (3.12), it follows that

$$G(0,0)=g(0,0)=1. \tag{7.8}$$

Now we consider the boundary of D_2 near the corner $(0,0)$, which is determined by the two curves

$$H(u,v)=x+y=0,\qquad K(u,v)=x-y=0, \tag{7.9}$$

where $H(u,v)$ and $K(u,v)$ denote, respectively, the transforms of the functions $x+y$ and $x-y$ in terms of the variables u and v. From (7.9) and (3.10), we have

$$\begin{aligned} H_u(u,v)&=\tfrac{1}{2}(x_s+y_s)-\tfrac{1}{2}(x_t+y_t),\\ H_v(u,v)&=\tfrac{1}{2}(x_s+y_s)+\tfrac{1}{2}(x_t+y_t),\\ K_u(u,v)&=\tfrac{1}{2}(x_s-y_s)-\tfrac{1}{2}(x_t-y_t),\\ K_v(u,v)&=\tfrac{1}{2}(x_s-y_s)+\tfrac{1}{2}(x_t-y_t). \end{aligned} \tag{7.10}$$

Simple computation from (7.5) and (7.6) gives

$$x_s(0,0) = y_t(0,0) = \sqrt{2}, \qquad x_t(0,0) = y_s(0,0) = 0. \tag{7.11}$$

Hence

$$H_u(0,0) = 0, \qquad H_v(0,0) = \sqrt{2}, \tag{7.12}$$

$$K_u(0,0) = \sqrt{2}, \qquad K_v(0,0) = 0. \tag{7.13}$$

By the implicit function theorem, there exists $\phi(u)$ such that $v = \phi(u)$ satisfies $H(u, v) = 0$ near $(0, 0)$. Furthermore,

$$\phi'(u) = -\left.\frac{H_u(u,v)}{H_v(u,v)}\right|_{v=\phi(u)}, \tag{7.14}$$

and, in particular, $\phi'(0) = 0$. This means that the boundary curve $H(u, v) = 0$ is tangent to the u-axis at $(0, 0)$. Similarly, there exists $\theta(v)$ such that $u = \theta(v)$ satisfies $K(u, v) = 0$ near $(0, 0)$, with

$$\theta'(v) = -\left.\frac{K_v(u,v)}{K_u(u,v)}\right|_{u=\theta(v)} \tag{7.15}$$

and $\theta'(0) = 0$.

To apply our results, we need to determine a value of ρ for each of the boundary curves $v = \phi(u)$ and $u = \theta(v)$ so that $\phi(u) = u^{\rho-1}\phi_1(u)$ and $\theta(v) = v^{\rho-1}\theta_1(v)$, and the values of $\phi_1(0)$ and $\theta_1(0)$. (The value of ρ need *not* be the same for ϕ and θ.) Now, $\phi''(u)$ can be computed from (7.14). Using (7.12) and the fact that $\phi'(0) = 0$, we find

$$\phi''(0) = -\frac{1}{\sqrt{2}} H_{uu}(0,0). \tag{7.16}$$

To evaluate $H_{uu}(0, 0)$, we begin with the first equation in (7.10). From (3.10) we obtain

$$H_{uu}(u,v) = \tfrac{1}{4}(x_{ss} + y_{ss}) - \tfrac{1}{2}(x_{st} + y_{ts}) + \tfrac{1}{4}(x_{tt} + y_{tt}). \tag{7.17}$$

Differentiating the equations in (7.6) gives

$$\begin{aligned}
&1 = px_s + xp_s, \qquad 0 = px_t + xp_t,\\
&0 = px_{ss} + 2x_s p_s + xp_{ss},\\
&0 = px_{st} + p_t x_s + x_t p_s + xp_{st},\\
&0 = px_{tt} + 2x_t p_t + xp_{tt},\\
&0 = qy_s + yq_s, \qquad 1 = qy_t + yq_t,\\
&0 = qy_{ss} + 2q_s y_s + yq_{ss},\\
&0 = qy_{st} + q_t y_s + q_s y_t + yq_{st},\\
&0 = qy_{tt} + 2y_t q_t + yq_{tt}.
\end{aligned} \tag{7.18}$$

Since $(x, y) = (0, 0)$ corresponds to $(s, t) = (0, 0)$, and since $p(0, 0) = q(0, 0) = 1/\sqrt{2}$, we have from (7.11)

$$\begin{aligned}
&x_{ss} = -4p_s, \quad x_{st} = -2p_t, \quad x_{tt} = 0,\\
&y_{ss} = 0, \qquad y_{st} = -2q_s, \quad y_{tt} = -4q_t,
\end{aligned} \tag{7.19}$$

where it is understood that all derivatives are evaluated at $(0, 0)$. From (7.11), we also have $p_s = p_x x_s + p_y y_s = \sqrt{2}\, p_x$ at $(0, 0)$. Hence it follows from (7.5) that $p_s(0, 0) = \sqrt{2}\, p_x(0, 0) = 0$. Similarly, we find $p_t(0, 0) = 0$. (The argument for the latter case is actually simpler, since p is independent of y in this example.) By the same reasoning, we have $q_s = q_x x_s + q_y y_s = \sqrt{2}\, q_x$ and $q_s(0, 0) = \sqrt{2}\, q_x(0, 0) = 0$. Similarly, $q_t(0, 0) = 0$. A combination of (7.16), (7.17), and (7.19) gives

$$\phi''(0) = -\frac{1}{\sqrt{2}} H_{uu}(0, 0) = 0. \tag{7.20}$$

Our next step is to find $\phi^{(3)}(0)$, which can be computed from (7.14). The full expression is very complicated; however, since $\phi'(0) = \phi''(0) = 0$, we have from (7.12) and (7.20) the relatively simple equation

$$\phi^{(3)}(0) = -\frac{1}{\sqrt{2}} H_{uuu}(0, 0). \tag{7.21}$$

Differentiating (7.17), we obtain

$$H_{uuu}(u, v) = \tfrac{1}{8}(x_{sss} + y_{sss}) - \tfrac{3}{8}(x_{sst} + y_{sst}) + \tfrac{3}{8}(x_{stt} + y_{stt}) - \tfrac{1}{8}(x_{ttt} + y_{ttt}) \tag{7.22}$$

on account of (3.10). To evaluate the partial derivatives on the right-hand side, we turn to the equations in (7.18). First, from the third equation in (7.18), we have

$$p x_{sss} + 3 p_s x_{ss} + 3 p_{ss} x_s + x p_{sss} = 0.$$

Since $(x, y) = (0, 0)$ corresponds to $(s, t) = (0, 0)$, and since all the second-order partial derivatives in (7.19) vanish at $(0, 0)$, the last equation yields

$$x_{sss}(0, 0) = -6 p_{ss}(0, 0). \tag{7.23}$$

Now, $p_{ss} = p_{xx} x_s^2 + 2 p_{xy} x_s y_s + p_{yy} y_s^2 + p_x x_{ss} + p_y y_{ss}$. In view of the fact that p is independent of y, and again by the vanishing at $(0, 0)$ of all the second-order partial derivatives in (7.19), this gives

$$p_{ss}(0, 0) = 2 p_{xx}(0, 0) = \frac{1}{6\sqrt{2}}. \tag{7.24}$$

The last equality follows from (7.5). Coupling (7.23) and (7.24), we obtain

$$x_{sss}(0, 0) = -\frac{1}{\sqrt{2}}. \tag{7.25}$$

Similar computations, beginning with (7.18) and using analogues of (7.24), yield

$$\begin{aligned} &x_{sst} = 0, \quad x_{stt} = 0, \quad x_{ttt} = 0, \\ &y_{sss} = 0, \quad y_{sst} = -\sqrt{2}, \quad y_{stt} = 0, \quad y_{ttt} = \frac{1}{\sqrt{2}}, \end{aligned} \tag{7.26}$$

where all derivatives are again evaluated at $(0, 0)$. Returning to (7.21), we have from (7.22), (7.25), and (7.26)

$$\phi^{(3)}(0) = -\tfrac{1}{4}. \tag{7.27}$$

Hence it follows that for the boundary curve $H(u, v) = x + y = 0$ we have

$$\rho_H = 4, \qquad \phi_1(0) = -\tfrac{1}{24}. \tag{7.28}$$

In the same manner, beginning with (7.15), we find that for the boundary curve $K(u, v) = x - y = 0$, we have $\theta'(0) = \theta''(0) = 0$ and $\theta^{(3)}(0) = -\frac{1}{4}$. Consequently, we again have

$$\rho_K = 4, \qquad \theta_1(0) = -\tfrac{1}{24}. \tag{7.29}$$

Note that $\phi^{(3)}(0)$ and $\theta^{(3)}(0)$ negative implies that both $\phi(u)$ and $\theta(v)$ are negative for sufficiently small positive u and v. Consequently, (5.3), (5.9), and (5.15) apply, with the integral (7.1) taken in the form

$$I(N) = I_2(N) + I_6(N) + I_4(N).$$

Using these equations and (7.8), we obtain

$$I(N) \sim \frac{i}{2}\frac{\log N}{N} + \left\{\frac{3\pi}{4} + i\left[\frac{\gamma}{2} + \log(2\sqrt{6})\right]\right\}\frac{1}{N} \tag{7.30}$$

as $N \to \infty$, where $\gamma = -\psi(1)$ is the Euler constant.

The above example may seem to be too complicated for illustration purposes, and simpler examples could have been given instead. However, we find that this example is rather realistic, and demonstrates the fact that our method indeed works in practical situations. It also illustrates well the computations involved in our procedure, particularly in cases where parameter ρ is greater than 2.

REFERENCES

[1] N. BLEISTEIN, *Mathematical Methods for Wave Phenomena*, Academic Press, Orlando, FL, 1984.

[2] J. C. COOKE, *Stationary phase in two dimensions*, IMA J. Appl. Math., 29 (1982), pp. 25-37.

[3] R. COURANT, *Differential and Integral Calculus*, Vol. 2, Blackie & Son, London, 1970.

[4] A. ERDÉLYI, *Asymptotic expansions of Fourier integrals involving logarithmic singularities*, SIAM J. Appl. Math., 4 (1956), pp. 38-47.

[5] D. S. JONES, *The Theory of Generalized Functions*, Cambridge University Press, Cambridge, U.K., 1982.

[6] D. S. JONES AND M. KLINE, *Asymptotic expansions of multiple integrals and the method of stationary phase*, J. Math. Phys., 37 (1958), pp. 1-28.

[7] B. MALGRANGE, *Ideals of Differentiable Functions*, Oxford University Press, Bombay, 1966.

[8] J. MCKENNA, *Note on asymptotic expansions of Fourier integrals involving logarithmic singularities*, SIAM J. Appl. Math., 15 (1967), pp. 810-812.

[9] F. W. J. OLVER, *Error bounds for stationary phase approximations*, SIAM J. Math. Anal., 5 (1974), pp. 19-29.

[10] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.

[11] R. WONG, *Asymptotic Approximations of Integrals*, Academic Press, Boston, 1989.

[12] R. WONG AND J. P. MCCLURE, *On a method of asymptotic evaluation of multiple integrals*, Math. Comp., 37 (1981), pp. 509-521.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 524-542, March 1991

014

ON THE ASYMPTOTIC BEHAVIOR OF THE COEFFICIENTS OF ASYMPTOTIC POWER SERIES AND ITS RELEVANCE TO STOKES PHENOMENA*

G. K. IMMINK†

Abstract. This paper discusses the relevance of the asymptotic behavior of the coefficients of asymptotic power series for the study of Stokes phenomena. By way of illustration a connection problem is considered in the theory of linear difference equations.

Key words. asymptotic expansion, isomorphism of Malgrange, Cauchy-Heine transform, saddle-point method, Stokes phenomenon, linear analytic functional equation, difference equation

AMS(MOS) subject classifications. 30E15, 39

Introduction. In this paper we extend and apply ideas of Malgrange [10] and Ramis [12] concerning the connection between Stokes phenomena, in a wider sense, and formal power series. We start with an illustrative example.

Let y be an analytic function on the Riemann surface of $\log z$, with the following properties.

(i) y admits an asymptotic expansion of the form $\sum_{n=0}^{\infty} y_n z^{-n}$ as $z \to \infty$ in the sector S: $-\pi/2 < \arg z < 5\pi/2$.

(ii) $y(z) - y(z e^{2\pi i}) = c e^{-z}$, $c \in \mathbb{C}^*$.

The second property implies that the asymptotic behavior of y changes abruptly as $\arg z$ becomes larger than $5\pi/2$ or less than $-\pi/2$. Such a change in asymptotic behavior will be called a Stokes phenomenon.

Now consider the function h defined by

$$h(z) = \frac{1}{2\pi i} \int_0^{\infty} \frac{e^{-t}}{t - z}\, dt, \qquad 0 < \arg z < 2\pi.$$

h is a Cauchy-Heine transform of e^{-z} (cf. [12]). By deformation of the path of integration it may be continued analytically to the Riemann surface of $\log z$. With the aid of residue calculus we readily verify that

$$h(z) - h(z e^{2\pi i}) = e^{-z}. \tag{0.1}$$

Moreover, h admits the asymptotic expansion $\sum_{n=1}^{\infty} h_n z^{-n}$ as $z \to \infty$, $z \in S$, where

$$h_n = -\frac{1}{2\pi i} \int_0^{\infty} e^{-t} t^{n-1}\, dt, \qquad n \in \mathbb{N}. \tag{0.2}$$

From (ii) and (0.1) we conclude that

$$y(z e^{2\pi i}) - ch(z e^{2\pi i}) = y(z) - ch(z).$$

Thus it turns out that $y - ch$ is a single-valued analytic function, admitting an asymptotic expansion of the form $\sum_{n=0}^{\infty} a_n z^{-n}$ as $z \to \infty$, $z \in S$, where

$$a_n = y_n - ch_n. \tag{0.3}$$

* Received by the editors January 19, 1989; accepted for publication (in revised form) February 14, 1990.

† University of Groningen, Institute of Econometrics, P.O. Box 800, 9700AV, Groningen, the Netherlands.

This implies that $y - ch$ is holomorphic at ∞ and, consequently, $\sum_{n=0}^{\infty} a_n z^{-n}$ is a convergent power series. From (0.2) and (0.3) it now follows that

$$c = -2\pi i \lim_{n\to\infty} \frac{y_n}{(n-1)!}.$$

Apparently, the constant c, which plays a central role in the Stokes phenomenon occurring in this example, is intimately related to the asymptotic behavior of the coefficients y_n. It is this relationship that forms the subject of this paper.

We shall consider the following situation. Suppose we are given a number of sectors S_ν, $\nu \in \{1, \cdots, N\}$, which cover a neighborhood of ∞ and a corresponding number of functions y_ν with the following properties: y_ν is analytic in S_ν and represented asymptotically by a series of the form $\sum_{n=0}^{\infty} \hat{y}_n z^{-n}$ (independent of ν) as $z \to \infty$, $z \in S_\nu$, $\nu \in \{1, \cdots, N\}$. Moreover, assume that

$$(0.4) \qquad y_{\nu+1}(z) - y_\nu(z) = \sum_{j=1}^{m} c_j^\nu \varphi_j^\nu(z), \qquad z \in S_\nu \cap S_{\nu+1}, \quad \nu \in \{1, \cdots, N\},$$

where $S_{N+1} = e^{2\pi i} S_1$, $y_{N+1}(z) \equiv y_1(z\, e^{-2\pi i})$, $c_j^\nu \in \mathbb{C}$, and the φ_j^ν belong to a certain class of analytic functions. We shall establish a relation between the complex numbers c_j^ν and the asymptotic behavior of $\hat{y}_n$ for $n \to \infty$. In some applications this relation may be exploited to "compute" at least part of the numbers c_j^ν from the coefficients $\hat{y}_n$ (cf. [9] and Remark 2 herein).

If the y_ν represent (sectorial models of) a resurgent function, our results could be derived from the work of Ecalle (cf. [4]). For the present purpose, however, this assumption is not needed and we shall establish the relation mentioned above in a more direct manner.

The argument is essentially the same as the one we used in [9]. It is based on the Propositions 1.1–1.3 herein. Proposition 1.1 concerns the properties of Cauchy–Heine transforms of functions like the φ_j^ν in (0.4). Proposition 1.2 enables us to construct, from the Cauchy–Heine transforms of the φ_j^ν, analytic functions H_ν with the same properties as the y_ν and only differing from the y_ν by a convergent power series in $1/z$. The coefficients of the asymptotic expansion $\hat{H}$ of the H_ν are given by the expression

$$\hat{H}_n = -\frac{1}{2\pi i} \sum_{\nu=1}^{N} \sum_{j=1}^{m} c_j^\nu \int_{\gamma_\nu} \varphi_j^\nu(t) t^{n-1}\, dt, \qquad \gamma_\nu \subset S_\nu.$$

Under certain conditions, like those mentioned in Proposition 1.3, the saddle-point method may be applied to the integral

$$\int_{\gamma_\nu} \varphi_j^\nu(t) t^{n-1}\, dt$$

to obtain its asymptotic behavior for $n \to \infty$. The main result is stated in Theorem 1.4. In § 2 this result is applied to a connection problem in the theory of homogeneous linear difference equations.

1. The general argument. Let $\mathbb{C}_\infty$ denote the Riemann surface of $\log z$. Let $z_0 \in \mathbb{C}_\infty$, $\alpha, \beta \in \mathbb{R}$, $\alpha < \beta$. By $S(\alpha, \beta)$ we denote the sector

$$S(\alpha, \beta) = \{z \in \mathbb{C}_\infty : \alpha < \arg z < \beta\}$$

and by $S(z_0, \alpha, \beta)$ the set

$$(1.1) \qquad S(z_0, \alpha, \beta) = \{z \in \mathbb{C}_\infty : \alpha < \arg (z - z_0) < \beta, |z| > |z_0|\}.$$

This will also be called a sector.

If S is a sector of the form $S=S(z_0, \alpha, \beta)$, then $\underline{S}$ will denote the sector $S(z_0, \alpha, \beta+2\pi)$.

Let $S=S(z_0, \alpha, \beta)$, $S'=S(z_1, \alpha', \beta')$ with $\alpha<\alpha'<\beta'<\beta$. We shall write

$$S' \Subset S$$

whenever $z_1 \in S$ and $S' \subset S(z_0, \alpha', \beta')$.

Let $\hat{h}=\sum_{n=0}^{\infty} h_n z^{-n}$ be a formal power series in z^{-1}, S a sector of the type (1.1), and h a function on S. We say that h is represented asymptotically by $\hat{h}$ as $z \to \infty$ in S, and write

$$h(z) \sim \sum_{n=0}^{\infty} h_n z^{-n}, \qquad z \to \infty \text{ in } S$$

if, for every $S' \Subset S$ and every $N \in \mathbb{N}$,

$$R_N(h; z) \equiv h(z) - \sum_{n=0}^{N-1} h_n z^{-n} = O(z^{-N}), \qquad z \to \infty, \quad z \in S'.$$

Any function φ which is analytic in a sector S and represented asymptotically by zero (i.e., the series with coefficients equal to zero) as $z \to \infty$ in S, may be written as the difference of two determinations of its Cauchy-Heine transform. The following proposition, due to Ramis, is concerned with the asymptotic properties of this Cauchy-Heine transform.

PROPOSITION 1.1 (cf. [12, Prop. 4.2]). *Let α and β be real numbers such that $\alpha<\beta$, $z_0 \in S(\alpha, \beta)$, and let φ be an analytic function on $S=S(z_0, \alpha, \beta)$. Suppose there exist positive numbers M_n, $n \in \mathbb{N}$, such that*

$$\sup_{z \in S} |z^n \varphi(z)| < M_n, \qquad n \in \mathbb{N}. \tag{1.2}$$

Then the function h defined by

$$h(z)=\frac{z}{2\pi i} \int_{\gamma} \frac{\varphi(\zeta)}{\zeta(\zeta-z)}\, d\zeta, \qquad z \in S(z_0, \Theta, \Theta+2\pi),$$

where γ is a half line in S from z_0 to ∞ with direction Θ, has the following properties:

(i) *h can be continued analytically to $\underline{S}$,*

(ii) *$h(z)-h(z e^{2\pi i})=\varphi(z)$ for all $z \in S$,*

(iii) *h is represented asymptotically by*

$$\sum_{n=0}^{\infty} -\frac{1}{2\pi i}\left(\int_{\gamma} \varphi(\zeta)\zeta^{n-1}\, d\zeta\right) z^{-n},$$

as $z \to \infty$ in $\underline{S}$. Moreover, for every $S' \Subset \underline{S}$ there exists a positive constant $C_{S'}$ such that

$$\sup_{z \in S'} |z^n R_n(h; z)| \leqq C_{S'} M_{n+1}, \qquad n \in \mathbb{N}.$$

Proof. Let us suppose that S is a convex set, i.e., $\beta-\pi/2<\arg z_0<\alpha+\pi/2$. In that case every half line from z_0 to ∞ with direction $\Theta \in (\alpha, \beta)$ lies in S. If γ has direction Θ, h is obviously analytic in $S(z_0, \Theta, \Theta+2\pi)$. The analytic continuation to $\underline{S}$ is obtained by varying Θ. Part (ii) follows immediately from Cauchy's theorem.

Now let $S'=S(z_1, \alpha', \beta') \Subset \underline{S}$. Then there is a number $\varepsilon \in (0, \pi/2)$ such that $\alpha+\varepsilon<\arg(z-z_0)<\beta+2\pi-\varepsilon$ for all $z \in S'$. Let $z \in S'$ and choose $\Theta \in (\alpha, \beta)$ in such

a way that $\Theta+\varepsilon<\arg(z-z_0)<\Theta+2\pi-\varepsilon$. Let γ_Θ be the half line from z_0 to ∞ with direction Θ. For all $\zeta\in\gamma_\Theta$ the following inequality holds:

$$|\zeta-z|>|z-z_0|\sin\varepsilon>|z|\left(1-\left|\frac{z_0}{z_1}\right|\right)\sin\varepsilon. \tag{1.3}$$

It is easily seen that

$$z^nR_n(h;z)=\frac{z}{2\pi i}\int_{\gamma_\Theta}\frac{\varphi(\zeta)}{\zeta-z}\,\zeta^{n-1}\,d\zeta,\qquad n\in\mathbb{N}.$$

With (1.2) and (1.3) it follows that

$$|z^nR_n(h;z)|<\frac{1}{2\pi\sin\varepsilon}\left(1-\left|\frac{z_0}{z_1}\right|\right)^{-1}\int_{\gamma_\Theta}\left|\frac{d\zeta}{\zeta^2}\right|M_{n+1}.$$

Hence

$$\sup_{z\in S'}|z^nR_n(h;z)|<\frac{1}{2\pi\sin\varepsilon}\left(1-\left|\frac{z_0}{z_1}\right|\right)^{-1}\sup_{\Theta\in(\alpha,\beta)}\int_{\gamma_\Theta}\left|\frac{d\zeta}{\zeta^2}\right|M_{n+1}$$

and this proves (iii).

If S is not convex the above argument must be adapted in an obvious manner.

PROPOSITION 1.2 (cf. [10], [12]). *Let $N\in\mathbb{N}$. Let α_ν, β_ν, $\nu\in\{1,\cdots,N\}$, be real numbers such that $\alpha_\nu\leqq\alpha_{\nu+1}<\beta_\nu\leqq\beta_{\nu+1}$ if $\nu<N$ and $\alpha_N\leqq\alpha_{N+1}\equiv\alpha_1+2\pi<\beta_N\leqq\beta_{n+1}\equiv\beta_1+2\pi$. Let $z_\nu\in S(\alpha_{\nu+1},\beta_\nu)$ and $S^\nu=S(z_\nu,\alpha_{\nu+1},\beta_\nu)$, $\nu=1,\cdots,N$.*

Suppose that, for every $\nu\in\{1,\cdots,N\}$, we are given an analytic function φ_ν on S^ν with the property that $\varphi_\nu(z)\sim0$ as $z\to\infty$ in S^ν. Let

$$h_\nu(z)=\frac{z}{2\pi i}\int_{\gamma_\nu}\frac{\varphi_\nu(\zeta)}{\zeta(\zeta-z)}\,d\zeta,\qquad z\in S(z_\nu,\Theta_\nu,\Theta_\nu+2\pi),\quad \nu\in\{1,\cdots,N\},$$

where γ_ν is a half line in S^ν from z_ν to ∞ with direction Θ_ν and let

$$H_\nu(z)=\sum_{\mu=1}^{\nu-1}h_\mu(z)+\sum_{\mu=\nu}^{N}h_\mu(z\,e^{2\pi i})\quad \text{if } \nu\in\{2,\cdots,N\},$$

$$H_1(z)=\sum_{\mu=1}^{N}h_\mu(z\,e^{2\pi i})\qquad \text{and}\qquad H_{N+1}(z)=\sum_{\mu=1}^{N}h_\mu(z).$$

The functions H_ν have the following properties:

(i) *For every $\nu\in\{1,\cdots,N+1\}$ there exists a $\tilde z_\nu\in S(\alpha_\nu,\beta_\nu)$ such that $\tilde z_{N+1}=\tilde z_1\,e^{2\pi i}$ and H_ν is analytic on $S_\nu=S(\tilde z_\nu,\alpha_\nu,\beta_\nu)$.*

(ii) *$H_{\nu+1}(z)-H_\nu(z)=\varphi_\nu(z)$ for all $z\in S_\nu\cap S_{\nu+1}$, $\nu\in\{1,\cdots,N\}$, and $H_{N+1}(z)=H_1(z\,e^{-2\pi i})$ for all $z\in e^{2\pi i}S_1$.*

(iii) *H_ν admits an asymptotic power series expansion $\hat H$ independent of ν, as $z\to\infty$ in S_ν.*

Moreover, if $\tilde H_\nu$, $\nu=1,\cdots,N+1$, are functions with the same properties, then there exists a function h, holomorphic at ∞, such that

$$\tilde H_\nu-H_\nu=h\quad \text{for all } \nu\in\{1,\cdots,N\}.$$

Proof. From Proposition 1.1(i) we deduce that H_ν is analytic in

$$\bigcap_{\mu=1}^{\nu-1}S(z_\mu,\alpha_{\mu+1},\beta_\mu+2\pi)\ \bigcap_{\mu=\nu}^{N}S(z_\mu,\alpha_{\mu+1}-2\pi,\beta_\mu)\quad \text{if } \nu\in\{2,\cdots,N\},$$

in

$$\bigcap_{\mu=1}^{N} S(z_\mu, \alpha_{\mu+1}-2\pi, \beta_\mu) \quad \text{if } \nu=1,$$

and in

$$\bigcap_{\mu=1}^{N} S(z_\mu, \alpha_{\mu+1}, \beta_\mu+2\pi) \quad \text{if } \nu=N+1,$$

and this set contains a sector of the form $S(\tilde{z}_\nu, \alpha_\nu, \beta_\nu)$ for a suitable choice of $\tilde{z}_\nu$. Part (ii) follows immediately from Proposition 1.1(ii) by observing that

$$H_{\nu+1}(z)-H_\nu(z)=h_\nu(z)-h_\nu(z\,e^{2\pi i}) \quad \text{for all } \nu\in\{1,\cdots,N\}.$$

Furthermore, Proposition 1.1(iii) implies that $H_\nu(z)\sim\sum_{n=0}^{\infty} H_n z^{-n}$, as $z\to\infty$ in S_ν, where

$$H_n=-\frac{1}{2\pi i}\sum_{\nu=1}^{N}\int_{\gamma_\nu}\varphi_\nu(\zeta)\zeta^{n-1}\,d\zeta, \qquad n\in\mathbb{N}. \tag{1.4}$$

Now suppose that $\tilde{H}_\nu$, $\nu=1,\cdots,N+1$, are functions with the properties (i)-(iii) mentioned in Proposition 1.2. Then there exist $z'_\nu\in S(\alpha_\nu,\beta_\nu)$ such that both H_ν and $\tilde{H}_\nu$ are analytic on $\tilde{S}_\nu=S(z'_\nu,\alpha_\nu,\beta_\nu)$ and we have

$$\tilde{H}_{\nu+1}(z)-\tilde{H}_\nu(z)=H_{\nu+1}(z)-H_\nu(z), \quad z\in\tilde{S}_\nu\cap\tilde{S}_{\nu+1}, \quad \nu\in\{1,\cdots,N\}$$

and

$$\tilde{H}_{N+1}(z)-\tilde{H}_1(z\,e^{-2\pi i})=H_{N+1}(z)-H_1(z\,e^{-2\pi i}), \qquad z\in\tilde{S}_{N+1}.$$

It follows that

$$\tilde{H}_{\nu+1}-H_{\nu+1}=\tilde{H}_\nu-H_\nu \quad \text{for all } \nu\in\{1,\cdots,N\}$$

and, moreover,

$$\tilde{H}_{N+1}(z)-H_{N+1}(z)=\tilde{H}_1(z\,e^{-2\pi i})-H_1(z\,e^{-2\pi i}).$$

Hence the function $h=\tilde{H}_1-H_1$ can be continued analytically to a reduced neighborhood of ∞. Furthermore, property (iii) implies that h admits an asymptotic power series expansion in z^{-1} as $z\to\infty$ in a neighborhood of ∞ and, consequently, h is analytic in a full neighborhood of ∞.

The next proposition concerns the asymptotic behavior of integrals of the type

$$\int_\gamma \varphi(z)z^n\,dz,$$

where γ is a half line and φ is an analytic function with the property that $\varphi(z)\sim 0$ as $z\to\infty$ in some sector S containing γ. The conditions (iii)-(v) below are purely technical and have been chosen in such a way that the result follows by a straightforward application of the saddle-point method. They might be relaxed or replaced by other conditions. We have merely tried to define a class of functions for which this method works.

PROPOSITION 1.3 (cf. [2, Thm. 7, Remark 6]). *Let α and β be real numbers such that $\alpha<\beta$, $z_0\in S(\alpha,\beta)$, and let ψ be an analytic function on $S=S(z_0,\alpha,\beta)$ with the property that*

(i) $\exp\psi(z)\sim 0$ *as* $z\to\infty$ *in* S.

Let $g: S\times\mathbb{N}\to\mathbb{C}$ be defined by

$$g(z,n)=\psi(z)+n\log z.$$

Suppose there exists $n_0 \in \mathbb{N}$ *such that for all* $n \geqq n_0$ *the following conditions hold*:

(ii) *The equation* $\partial g/\partial z = 0$ *has a solution* $s_n \in S$ *such that the half line* γ_n *from* z_0 *to* ∞ *through* s_n *is contained in* S. *Moreover,* $s_n \to \infty$ *as* $n \to \infty$.

(iii) *There exists a number* $\Theta \in (0, \pi/2)$ *such that*

$$\left| \arg - s_n^2 \frac{\partial^2 g}{\partial z^2}(s_n, n) \right| \leqq \Theta$$

and, furthermore, $s_n^2(\partial^2 g/\partial z^2)(s_n, n) \to \infty$ *as* $n \to \infty$.

(iv) *There exist positive numbers* ε_0 *and* K *such that*

$$\left| z \frac{\partial^3}{\partial z^3} g(z, n) \left\{ \frac{\partial^2}{\partial z^2} g(z, n) \right\}^{-1} \right| \leqq K$$

if $|z - s_n| < \varepsilon_0 |s_n|$.

(v) *Let* $\alpha_n = \arg(s_n - z_0) - \arg s_n$. *There exists a positive number* ε_1, *a function* $n_1 : (0, \varepsilon_1) \to \mathbb{N}$, *a bounded function* $g_1 : (0, \varepsilon_1) \times (-1, 0) \to \mathbb{R}$, *and a function* $g_2 : (0, \varepsilon_1) \times (0, \infty) \to \mathbb{R}$ *such that, for all* $\varepsilon \in (0, \varepsilon_1)$, $\exp g_2(\varepsilon, \cdot) \in \mathscr{L}(0, \infty)$, *and, for all* $n \geqq n_1(\varepsilon)$,

$$\operatorname{Re}\{g(s_n(1 + \tau e^{i\alpha_n}), n) - g(s_n(1 - \varepsilon e^{i\alpha_n}), n)\} \leqq g_1(\varepsilon, \tau)$$

if $\tau \in (-|1 - z_0/s_n|, -\varepsilon)$, *whereas*

$$\operatorname{Re}\{g(s_n(1 + \tau e^{i\alpha_n}), n) - g(s_n(1 + \varepsilon e^{i\alpha_n}), n)\} \leqq g_2(\varepsilon, \tau)$$

if $\tau \in (\varepsilon, \infty)$.

Furthermore, let f *be a bounded analytic function on* S *and suppose there exists a positive number* ε *such that*

(vi) $\sup_{z \in I_n(\varepsilon)} |f(z) - 1| \to 0$ *if* $n \to \infty$, *where* $I_n(\varepsilon)$ *denotes the segment between* $s_n(1 - \varepsilon e^{i\alpha_n})$ *and* $s_n(1 + \varepsilon e^{i\alpha_n})$.

Let

$$\varphi(z) = f(z) \exp \psi(z),$$

and

$$J_n = \frac{1}{2\pi i} \int_\gamma \varphi(z) z^n \, dz$$

where γ *is a half line in* S *from* z_0 *to* ∞. *Then we have*

$$J_n = \left\{ 2\pi s_n^2 \frac{\partial^2 g}{\partial z^2}(s_n, n) \right\}^{-1/2} s_n \exp g(s_n, n)(1 + o(1)), \quad n \to \infty,$$

where $\arg\{s_n^2(\partial^2 g/\partial z^2)(s_n, n)\}^{-1/2} \in (-\pi, 0)$.

Proof. We shall closely follow the proof of Theorem 7 in [2]. Let $n \geqq n_0$. Due to (i), (ii) and the properties of f, we may replace γ by γ_n. Let $\varepsilon > 0$. We begin by considering the integrand on the segment $I_n(\varepsilon)$. We put

$$\left| \frac{\partial^2 g}{\partial z^2}(z, n) - \frac{\partial^2 g}{\partial z^2}(s_n, n) \right| = h(z).$$

From (iv) we deduce that

$$h(z) \leqq K \left| \frac{\partial^2 g}{\partial z^2}(s_n, n) \right| \int_{s_n}^z \left| \frac{d\zeta}{\zeta} \right| + K \int_{s_n}^z h(\zeta) \left| \frac{d\zeta}{\zeta} \right|$$

provided $|z-s_n|<\varepsilon_0|s_n|$. With the aid of Gronwall's generalized inequality (cf. [5, p. 36]) we find

$$h(z)\leqq \tilde{K}\left|\frac{z-s_n}{s_n}\right|\left|\frac{\partial^2 g}{\partial z^2}(s_n, n)\right|,$$

where $\tilde{K}$ is a positive constant, provided $|z-s_n|<\varepsilon_0|s_n|$. Hence it follows that, for all $\varepsilon\in(0, \varepsilon_0)$,

$$\frac{\partial g}{\partial z}(z, n)=(z-s_n)\frac{\partial^2 g}{\partial z^2}(s_n, n)(1+\varepsilon O(1)), \qquad n\to\infty \tag{1.5}$$

$$g(z, n)-g(s_n, n)=\frac{1}{2}(z-s_n)^2\frac{\partial^2 g}{\partial z^2}(s_n, n)(1+\varepsilon O(1)), \qquad n\to\infty, \tag{1.6}$$

uniformly on $I_n(\varepsilon)$. Here $O(1)$ is uniformly bounded in ε.

We introduce a new variable w by means of

$$\tfrac{1}{2}w^2=g(s_n, n)-g(s_n(1+\tau e^{i\alpha_n}), n), \qquad |\tau|\leqq\varepsilon. \tag{1.7}$$

Due to (1.6) we have

$$w^2=-\tau^2 e^{2i\alpha_n}s_n^2\frac{\partial^2 g}{\partial z^2}(s_n, n)(1+\varepsilon O(1))$$

and we remove the ambiguity in the definition of w by demanding that

$$w=\tau e^{i\alpha_n}\left(-s_n^2\frac{\partial^2 g}{\partial z^2}(s_n, n)\right)^{1/2}(1+\varepsilon O(1)), \tag{1.8}$$

where the square root has its principal value. Equations (1.7) and (1.5) imply that

$$w\frac{dw}{d\tau}=-s_n e^{i\alpha_n}\frac{\partial g}{\partial z}(s_n(1+\tau e^{i\alpha_n}), n)=-s_n^2\tau e^{2i\alpha_n}\frac{\partial^2 g}{\partial z^2}(s_n, n)(1+\varepsilon O(1)).$$

Consequently,

$$\frac{dw}{d\tau}=e^{i\alpha_n}\left(-s_n^2\frac{\partial^2 g}{\partial z^2}(s_n, n)\right)^{1/2}(1+\varepsilon O(1)). \tag{1.9}$$

Let $w_\pm$ correspond to $\tau=\pm\varepsilon$. From (1.8) it follows that

$$w_\pm=\pm\varepsilon e^{i\alpha_n}\left(-s_n^2\frac{\partial^2 g}{\partial z^2}(s_n, n)\right)^{1/2}(1+\varepsilon O(1)). \tag{1.10}$$

From (1.7), (1.9), (1.10), and condition (vi) of Proposition 1.3 we deduce that

$$\begin{aligned}&\int_{s_n(1-\varepsilon e^{i\alpha_n})}^{s_n(1+\varepsilon e^{i\alpha_n})}\varphi(z)z^n\,dz\\ &\quad=s_n e^{i\alpha_n}\exp g(s_n, n)\int_{w_-}^{w_+}e^{-w^2/2}\left(\frac{dw}{d\tau}\right)^{-1}dw(1+o(1))\\ &\quad=\left(-s_n^2\frac{\partial^2 g}{\partial z^2}(s_n, n)\right)^{-1/2}s_n\exp g(s_n, n)\int_{w_-}^{w_+}e^{-w^2/2}\,dw(1+o(1))(1+\varepsilon O(1)).\end{aligned}$$

Using (1.10) and condition (iii) and noting that $\lim_{n\to\infty}\alpha_n=0$, we conclude that, for every $\varepsilon>0$,

$$\lim_{n\to\infty}\int_{w_-}^{w_+}e^{-w^2/2}\,dw=\sqrt{2\pi}.$$

Hence

$$\frac{1}{2\pi i}\int_{s_n(1-\varepsilon e^{i\alpha_n})}^{s_n(1+\varepsilon e^{i\alpha_n})}\varphi(z)z^n\,dz=\left(2\pi s_n^2\frac{\partial^2 g}{\partial z^2}(s_n,n)\right)^{-1/2}s_n$$
$$\cdot(\exp g(s_n,n)(1+o(1))(1+\varepsilon O(1)),$$

where $\arg\{s_n^2(\partial^2 g/\partial z^2)(s_n,n)\}^{-1/2}\in(-\pi,0)$.

Next we consider the integral

$$J_n^+(\varepsilon)=\int_{s_n(1+\varepsilon e^{i\alpha_n})}^{\infty}\varphi(z)z^n\,dz.$$

From (1.7), condition (v), and the properties of f we deduce that, for $n\geqq n_1(\varepsilon)$,

$$|J_n^+(\varepsilon)|\leqq C\left|s_n\exp\left\{g(s_n,n)-\frac{1}{2}w_+^2\right\}\right|\int_{\varepsilon}^{\infty}\exp g_2(\varepsilon,\tau)\,d\tau$$
$$\leqq C_1\left|s_n\exp\left\{g(s_n,n)-\frac{1}{2}w_+^2\right\}\right|,$$

where C and C_1 are positive constants. In view of (1.10) and condition (iii) this implies that

$$J_n^+(\varepsilon)=\left(s_n^2\frac{\partial^2 g}{\partial z^2}(s_n,n)\right)^{-1/2}s_n\exp g(s_n,n)o(1),\qquad n\to\infty$$

and the same property holds for the integral over the segment between z_0 and $s_n(1-\varepsilon e^{i\alpha_n})$. Combining the above estimates we find

$$J_n=\left(2\pi s_n^2\frac{\partial^2 g}{\partial z^2}(s_n,n)\right)^{-1/2}s_n\exp g(s_n,n)(1+\varepsilon O(1))(1+o(1)),\qquad n\to\infty.$$

Since this is true for every sufficiently small ε the result follows.

THEOREM 1.4. *Let $N\in\mathbb{N}$. Let α_ν, β_ν, $\nu\in\{1,\cdots,N\}$, be real numbers such that $\alpha_\nu\leqq\alpha_{\nu+1}<\beta_\nu\leqq\beta_{\nu+1}$ if $\nu<N$ and $\alpha_N\leqq\alpha_{N+1}\equiv\alpha_1+2\pi<\beta_N\leqq\beta_{N+1}\equiv\beta_1+2\pi$. Let $\tilde{z}_\nu\in S(\alpha_\nu,\beta_\nu)$ and $S_\nu=S(\tilde{z}_\nu,\alpha_\nu,\beta_\nu)$, $\nu=1,\cdots,N$, $S_{N+1}=e^{2\pi i}S_1$. Suppose that, for each $\nu\in\{1,\cdots,N\}$, we are given an analytic function y_ν on S_ν, admitting an asymptotic expansion $\sum_{n=0}^{\infty}\hat{y}_n z^{-n}$ as $z\to\infty$ in S_ν, independent of ν. Let*

$$y_{N+1}(z)=y_1(z\,e^{-2\pi i}),\qquad z\in S_{N+1},$$

and

$$\varphi_\nu(z)=y_{\nu+1}(z)-y_\nu(z),\qquad z\in S_\nu\cap S_{\nu+1},\quad \nu\in\{1,\cdots,N\}.$$

Suppose that for every $\nu\in\{1,\cdots,N\}$ there exists a sector $\tilde{S}^\nu\subset S_\nu\cap S_{\nu+1}$, a positive integer $m(\nu)$, and, for every $j\in\{1,\cdots,m(\nu)\}$, analytic functions ψ_j^ν and f_j^ν on $\tilde{S}^\nu$, satisfying the conditions of Proposition 1.3, and a complex number c_j^ν such that

$$\varphi_\nu(z)=\sum_{j=1}^{m(\nu)}c_j^\nu f_j^\nu(z)\exp\psi_j^\nu(z),\qquad z\in\tilde{S}^\nu.$$

Let $g_j^\nu(z,n)=\psi_j^\nu(z)+n\log z$, let $s_n^{\nu,j}$ denote its saddle point, and let

$$M_j^\nu(n)=\left\{2\pi(s_n^{\nu,j})^2\frac{\partial^2 g_j^\nu}{\partial z^2}(s_n^{\nu,j},n)\right\}^{-1/2}s_n^{\nu,j}\exp g_\nu(s_n^{\nu,j},n),$$
$$j\in\{1,\cdots,m(\nu)\},\quad \nu\in\{1,\cdots,N\},$$

where $\arg\{(s_n^{\nu,j})^2(\partial^2 g_j^\nu/\partial z^2)(s_n^{\nu,j}, n)\}^{-1/2} \in (-\pi, 0)$. *Then there exists a convergent power series* $\sum_{n=0}^{\infty} h_n z^{-n}$ *such that*

$$\hat{y}_n = h_n - \sum_{\nu=1}^{N} \sum_{j=1}^{m(\nu)} c_j^\nu \{M_j^\nu(n-1)(1+o(1))\}, \qquad n \to \infty. \tag{1.11}$$

Proof. There exists $z_\nu \in S_\nu \cap S_{\nu+1}$ such that $S_\nu \cap S_{\nu+1}$ contains the sector $S^\nu = S(z_\nu, \alpha_{\nu+1}, \beta_\nu)$. As y_ν and $y_{\nu+1}$ admit the same asymptotic expansion, it follows that

$$\varphi_\nu(z) = y_{\nu+1}(z) - y_\nu(z) \sim 0 \quad \text{as } z \to \infty \text{ in } S^\nu, \quad \nu \in \{1, \cdots, N\}.$$

Obviously, the functions y_ν possess the properties (i)–(iii) mentioned in Proposition 1.2.

According to Proposition 1.2 there exists a function h, holomorphic at ∞, such that $y_\nu = h + H_\nu$ for all $\nu \in \{1, \cdots, N\}$. Let $\sum_{n=0}^{\infty} h_n z^{-n}$ be the power series expansion of h. With (1.4) we find

$$\begin{aligned} \hat{y}_n &= h_n - \sum_{\nu=1}^{N} \frac{1}{2\pi i} \int_{\gamma_\nu} \varphi_\nu(z) z^{n-1}\, dz \\ &= h_n - \sum_{\nu=1}^{N} \sum_{j=1}^{m(\nu)} \frac{c_j^\nu}{2\pi i} \int_{\gamma_\nu} f_j^\nu(z) \exp \psi_j^\nu(z) z^{n-1}\, dz, \qquad n \in \mathbb{N}, \end{aligned}$$

where γ_ν is a half line $\tilde{S}^\nu$, $\nu \in \{1, \cdots, N\}$.

The proof is completed by application of Proposition 1.3 to each term of the sum in the right-hand side of the above identity.

Remark 1. If the y_ν as well as the functions $f_j^\nu \exp \psi_j^\nu$ are solutions of some homogeneous linear functional equation, the numbers c_j^ν play a role similar to the Stokes multipliers in the theory of linear differential equations.

Remark 2. If one of the functions M_j^ν in (1.11) dominates the rest for $n \to \infty$, the corresponding coefficient c_j^ν may be determined from the asymptotic behavior of $\hat{y}_n$.

Remark 3. Propositions 1.1 and 1.2 may also be used to obtain estimates of the growth of the remainder terms $R_n(y_\nu; z)$ as $n \to \infty$. This will be illustrated by the application to linear difference equations in the next section.

Example. The nonlinear differential equation

$$\frac{dy}{dz} = \frac{a}{z^2} + y + \frac{b}{z^2} y^3, \qquad a, b \in \mathbb{C}^* \tag{1.12}$$

possesses three formal solutions of the form $\sum_{n=-1}^{\infty} \hat{y}_n z^{-n}$. The coefficients $\hat{y}_n$ can be determined from the recursive relations

$$\begin{aligned} &-2\hat{y}_{n+2} + (n+4)\hat{y}_{n+1} + b \sum_{\substack{m_i \geqq n \\ m_1+m_2+m_3=n}} \hat{y}_{m_1}\hat{y}_{m_2}\hat{y}_{m_3} = 0, \qquad n \geqq -1, \\ &(\hat{y}_{-1})^2 = -\frac{1}{b}, \qquad \hat{y}_0 = -\frac{1}{2}\hat{y}_{-1} \end{aligned} \tag{1.13}$$

and

$$\begin{aligned} &\hat{y}_{n+2} + (n+1)\hat{y}_{n+1} + b \sum_{m_1+m_2+m_3=n} \hat{y}_{m_1}\hat{y}_{m_2}\hat{y}_{m_3} = 0, \qquad n \geqq 1, \\ &\hat{y}_{-1} = \hat{y}_0 = \hat{y}_1 = 0, \qquad \hat{y}_2 = -a. \end{aligned} \tag{1.14}$$

Let $\hat{y}$ denote one of the formal solutions and let S be a sector of aperture less than π. It is a well-known fact that there exists a solution of (1.12), analytic in S and

represented asymptotically by $\hat{y}$ as $z \to \infty$ in S, uniformly on S (cf. [13]). Suppose that y_1 and y_2 are two solutions with these properties. Obviously,

$$\frac{d}{dz}(y_1 - y_2) = y_1 - y_2 + \frac{b}{z^2}(y_1^2 + y_1 y_2 + y_2^2)(y_1 - y_2). \tag{1.15}$$

Let $\hat{y} = \sum_{n=-1}^{\infty} \hat{y}_n^- z^{-n}$ and suppose the coefficients $\hat{y}_n^-$ satisfy (1.13). Then we have

$$y_1^2 + y_1 y_2 + y_2^2 = -\frac{3}{b}(z^2 - z) + h(z), \tag{1.16}$$

where h is a bounded analytic function on S, admitting an asymptotic expansion as $z \to \infty$ in S. Inserting (1.16) into (1.15) we obtain

$$\frac{d}{dz}(y_1 - y_2) = \left\{-2 + \frac{3}{z} + \frac{b}{z^2} h(z)\right\}(y_1 - y_2)$$

and this implies that

$$y_1 - y_2 = c\, e^{-2z} z^3 \left(1 + O\left(\frac{1}{z}\right)\right) \quad z \to \infty \text{ in } S,$$

where c is a complex number. Hence it follows that (1.12) has a unique solution y^-, analytic in a left half plane and represented asymptotically by the series $\sum_{n=-1}^{\infty} \hat{y}_n^- z^{-n}$ as $z \to \infty$ in this half plane. Moreover, it is easily seen that y^- may be continued analytically to a sector of the form $S(z_1, -3\pi/2, 3\pi/2)$, with $z_1 \in \mathbb{C}_\infty$, without a change in asymptotic behavior.

Furthermore, we have

$$y^-(z) - y^-(z\, e^{2\pi i}) = c^-\, e^{-2z} z^3 \left(1 + O\left(\frac{1}{z}\right)\right), \qquad c^- \in \mathbb{C},$$

as $z \to \infty$ in $S(-3\pi/2 + \varepsilon, -\pi/2 - \varepsilon)$ for any $\varepsilon \in (0, \pi/2)$. Applying Theorem 1.4 we find

$$c^- = -2\pi i \lim_{n \to \infty} \frac{2^{n+3} \hat{y}_n^-}{(n+2)!}.$$

In a similar manner it is shown that (1.12) possesses a unique solution y^+ analytic in $S(z_2, -\pi/2, 5\pi/2)$ for some $z_2 \in \mathbb{C}_\infty$ and represented asymptotically by the series $\sum_{n=-1}^{\infty} \hat{y}_n^+ z^{-n}$ determined by (1.14), as $z \to \infty$ in this sector. Moreover, it turns out that

$$y^+(z) - y^+(z\, e^{2\pi i}) = c^+\, e^{z} \left(1 + O\left(\frac{1}{z}\right)\right), \qquad c^+ \in \mathbb{C},$$

as $z \to \infty$ in $S(-\pi/2 + \varepsilon, \pi/2 - \varepsilon)$ for any $\varepsilon \in (0, \pi/2)$. Application of Theorem 1.4 now yields the relation

$$c^+ = 2\pi i \lim_{n \to \infty} \frac{(-1)^{n-1} \hat{y}_n^+}{(n-1)!}.$$

2. An application to linear difference equations. We consider the mth-order homogeneous linear difference equation

$$\sum_{j=0}^{m} a_j(z) y(z + j) = 0, \tag{2.1}$$

where $a_j \in \mathbb{C}\{z^{-1}\}$, $j = 1, \cdots, m$ (or, equivalently, a system of m first-order difference equations). The "generic" case is when the characteristic equation of (2.1) has m distinct roots. This case has been treated in [8]. Here we shall deal with a more singular class of equations.

Under certain conditions, (2.1) possesses m linearly independent formal solutions of the form

$$\hat{y}_j(z) = \hat{h}_j(z) z^{\rho_j} \exp(d_j z \log z + \mu_j z), \qquad j = 1, \cdots, m, \tag{2.2}$$

where $\hat{h}_j(z) = \sum_{n=0}^{\infty} \hat{h}_{jn} z^{-n}$ with $\hat{h}_{j0} = 1$, $\rho_j \in \mathbb{C}$, $d_j \in \mathbb{Q}$ and $\mu_j \in \mathbb{C}$ for all $j \in \{1, \cdots, m\}$ (cf. [3], [11]).

We put

$$\rho_i - \rho_j = \rho_{ij}, \quad d_i - d_j = d_{ij}, \quad \text{and} \quad \mu_i - \mu_j = \mu_{ij}, \qquad i, j \in \{1, \cdots, m\}$$

and we assume that, for all $i, j \in \{1, \cdots, m\}$ such that $i \neq j$ and $d_{ij} = 0$,

$$\operatorname{Re} \mu_{ij} \neq 0. \tag{2.3}$$

For merely technical reasons we further assume that

$$\operatorname{Im} \mu_{ij} \notin \{0, -d_{ij}\pi\} \bmod 2\pi \quad \text{if } i \neq j, \quad i, j \in \{1, \cdots, m\} \tag{2.4}$$

but this condition can easily be removed. For all $i, j \in \{1, \cdots, m\}$ such that $i \neq j$ we shall denote by n_{ij} the integer determined by

$$\begin{aligned} &0 < \operatorname{Im} \mu_{ij} + 2n_{ij}\pi < 2\pi && \text{if } d_{ij} \leqq 0, \\ &0 < \operatorname{Im} \mu_{ij} + (2n_{ij} + d_{ij})\pi < 2\pi && \text{if } d_{ij} > 0. \end{aligned} \tag{2.5}$$

Let $S_1, \cdots, S_7$ be sectors of the following form:

$$S_1 = S(R e^{-i(\pi/2)}, -\pi, 0), \quad S_2 = e^{i(\pi/2)} S_1, \quad S_3 = S_4 = e^{i\pi} S_1,$$
$$S_5 = e^{i(3\pi/2)} S_1 \quad \text{and} \quad S_6 = S_7 = e^{2\pi i} S_1,$$

where $R > 0$. If R is chosen sufficiently large, equation (2.1) possesses, for each $j \in \{1, \cdots, m\}$ and $\nu \in \{1, 3, 4, 6, 7\}$, a unique solution y_j^{ν}, represented asymptotically by $\hat{y}_j$ as $z \to \infty$, uniformly on

$$\begin{aligned} &\left(\frac{\nu - 1}{3} - 1\right)\pi + \delta < \arg(z - R e^{(\nu/3 - 5/6)\pi i}) \leqq \frac{\nu - 1}{3}\pi \quad \text{if } \nu \in \{1, 4, 7\}, \\ &\left(\frac{\nu}{3} - 1\right)\pi \leqq \arg(z - R e^{(\nu/3 - 1/2)\pi i}) < \frac{\nu}{3}\pi - \delta \quad \text{if } \nu \in \{3, 6\} \end{aligned} \tag{2.6}$$

for every $\delta \in (0, \pi/2)$ (cf. [6, Thm. 2.4.5]; note that this is a stronger statement than $y_j^{\nu} \sim \hat{y}_j$ as $z \to \infty$ in S_{ν}). Moreover, we have

$$y_j^4 - y_j^3 = p_{jj}^3 y_j^3, \qquad y_j^7 - y_j^6 = p_{jj}^6 y_j^6, \tag{2.7}$$

where p_{jj}^3 and p_{jj}^6 are periodic functions of period 1 with the property that

$$\lim_{\operatorname{Im} z \to \infty} p_{jj}^3(z) = \lim_{\operatorname{Im} z \to -\infty} p_{jj}^6(z) = 0, \qquad j \in \{1, \cdots, m\}. \tag{2.8}$$

Furthermore, for each $j \in \{1, \cdots, m\}$, equation (2.1) possesses a unique solution y_j^2, analytic in S_2 and represented asymptotically by $\hat{y}_j$ as $z \to \infty$ in S_2, such that

$$y_j^2 - y_j^1 = \sum_{i=1}^{m} p_{ij}^1 y_i^1, \qquad y_j^3 - y_j^2 = \sum_{i=1}^{m} p_{ij}^2 y_i^2,$$

where p_{ij}^1 and p_{ij}^2 are periodic functions of period 1 with the following properties:

$$\begin{aligned} &p_{ij}^1 = p_{ij}^2 \equiv 0 \quad \text{if } d_{ij} > 0 \text{ or } d_{ij} = 0 \text{ and } \operatorname{Re} \mu_{ij} \geqq 0, \\ &\lim_{\operatorname{Im} z \to -\infty} p_{ij}^1(z) \exp\{-2(n_{ij} - 1)\pi i z\} \text{ and } \lim_{\operatorname{Im} z \to \infty} p_{ij}^2(z) \exp\{-2n_{ij}\pi i z\} \\ &\qquad \text{exist for all } i, j \in \{1, \cdots, m\} \text{ such that } i \neq j. \end{aligned} \tag{2.9}$$

Similarly, for each $j \in \{1, \cdots, m\}$, there exists a unique solution y_j^5, analytic in S_5 and represented asymptotically by $\hat{y}_j$ as $z \to \infty$ in S_5, such that

$$y_j^5 - y_j^4 = \sum_{i=1}^{m} p_{ij}^4 y_i^4, \qquad y_j^6 - y_j^5 = \sum_{i=1}^{m} p_{ij}^5 y_i^5,$$

where p_{ij}^4 and p_{ij}^5 are periodic functions of period 1 with the following properties:

$$p_{ij}^4 = p_{ij}^5 \equiv 0 \quad \text{if } d_{ij} < 0 \text{ or } d_{ij} = 0 \text{ and } \operatorname{Re} \mu_{ij} \leqq 0,$$

(2.10) $\lim_{\operatorname{Im} z \to \infty} p_{ij}^4(z) \exp\{-2n_{ij}\pi i z\}$ and $\lim_{\operatorname{Im} z \to -\infty} p_{ij}^5(z) \exp\{-2(n_{ij}-1)\pi i z\}$ exist for all $i, j \in \{1, \cdots, m\}$ such that $i \neq j$.

Now let

$$h_j^\nu(z) = y_j^\nu(z) z^{-\rho_j} \exp(-d_j z \log z - \mu_j z), \qquad j \in \{1, \cdots, m\}, \quad \nu \in \{1, \cdots, 7\}.$$

Obviously, h_j^ν is represented asymptotically by $\hat{h}_j$ as $z \to \infty$ in S_ν for all $j \in \{1, \cdots, m\}$ and all $\nu \in \{1, \cdots, 7\}$. Moreover, if $\nu \in \{1, 3, 4, 6, 7\}$, the asymptotic expansion is uniformly valid on (2.6) for every $\delta \in (0, \pi/2)$. The uniqueness of h_j^ν implies that

$$h_j^7(z) = h_j^1(z e^{-2\pi i}) \quad \text{for all } j \in \{1, \cdots, m\}. \tag{2.11}$$

Furthermore, we have, for all $j \in \{1, \cdots, m\}$ and $\nu \in \{1, \cdots, 6\}$,

$$h_j^{\nu+1}(z) - h_j^\nu(z) = \sum_{i=1}^{m} p_{ij}^\nu(z) h_i^\nu(z) z^{\rho_{ij}} \exp(d_{ij} z \log z + \mu_{ij} z). \tag{2.12}$$

For all $i, j \in \{1, \cdots, m\}$ and all $\nu \in \{1, \cdots, 6\}$ we define an integer n_{ij}^ν and complex numbers c_{ij}^ν and μ_{ij}^ν as follows:

(2.13)

$$n_{ij}^\nu = \begin{cases} \max\left\{ n \in \mathbb{Z} : \lim\limits_{\operatorname{Im} z \to \infty} p_{ij}^\nu(z) \exp(-2n\pi i z) \text{ exists} \right\} & \text{if } \nu \in \{2, 3, 4\} \text{ and } p_{ij}^\nu \not\equiv 0, \\ \min\left\{ n \in \mathbb{Z} : \lim\limits_{\operatorname{Im} z \to -\infty} p_{ij}^\nu(z) \exp(-2n\pi i z) \text{ exists} \right\} & \text{if } \nu \in \{1, 5, 6\} \text{ and } p_{ij}^\nu \not\equiv 0, \\ 0 \quad \text{otherwise,} \end{cases}$$

$$c_{ij}^\nu = \begin{cases} 0 & \text{if } p_{ij}^\nu \equiv 0, \\ \lim_{\operatorname{Im} z \to \pm\infty} p_{ij}^\nu(z) \exp(-2n_{ij}^\nu \pi i z) & \text{otherwise,} \end{cases} \tag{2.14}$$

$$\mu_{ij}^\nu = \mu_{ij} + 2n_{ij}^\nu \pi i z. \tag{2.15}$$

Furthermore, we define analytic functions f_{ij}^ν and φ_{ij}^ν by

$$f_{ij}^\nu(z) = \begin{cases} 0 & \text{if } c_{ij}^\nu = 0, \\ (c_{ij}^\nu)^{-1} p_{ij}^\nu(z) \exp(-2n_{ij}^\nu \pi i z) h_i^\nu(z) & \text{otherwise,} \end{cases} \tag{2.16}$$

$$\varphi_{ij}^\nu(z) = p_{ij}^\nu(z) h_i^\nu(z) z^{\rho_{ij}} \exp(d_{ij} z \log z + \mu_{ij} z). \tag{2.17}$$

Obviously,

$$\varphi_{ij}^\nu(z) = c_{ij}^\nu f_{ij}^\nu(z) z^{\rho_{ij}} \exp(d_{ij} z \log z + \mu_{ij}^\nu z), \qquad i, j \in \{1, \cdots, m\}, \quad \nu \in \{1, \cdots, 6\}. \tag{2.18}$$

In order to check whether the conditions of Theorem 1.4 are satisfied, we will first study the properties of the function $g : S \times \mathbb{N} \to \mathbb{C}$ defined by

$$g(z, n) = dz \log z + \mu z = (n + \rho) \log z,$$

where $d \in \mathbb{Q}$, $\mu \in \mathbb{C}$, $\rho \in \mathbb{C}$, and S is one of the sectors $S_\nu \cap S_{\nu+1}$, $\nu \in \{1, \cdots, 6\}$. From (2.18), (2.3), (2.7)-(2.10), and the definitions (2.13)-(2.15) we conclude that the following cases need to be considered:

1. $d = 0$, $\rho = 0$, $\mu = 2m\pi i$, $m \in \mathbb{N}$, $S = S_3$,
2. $d = 0$, $\rho = 0$, $\mu = -2m\pi i$, $m \in \mathbb{N}$, $S = S_6$,
3. $d = 0$, $\operatorname{Re} \mu < 0$, $\operatorname{Im} \mu < 0$, $S = S_1 \cap S_2$,
4. $d = 0$, $\operatorname{Re} \mu < 0$, $\operatorname{Im} \mu > 0$, $S = S_2 \cap S_3$,
5. $d = 0$, $\operatorname{Re} \mu > 0$, $\operatorname{Im} \mu > 0$, $S = S_4 \cap S_5$,
6. $d = 0$, $\operatorname{Re} \mu > 0$, $\operatorname{Im} \mu < 0$, $S = S_5 \cap S_6$,
7. $d < 0$, $\operatorname{Im} \mu < 0$, $S = S_1 \cap S_2$,
8. $d < 0$, $\operatorname{Im} \mu > 0$, $S = S_2 \cap S_3$,
9. $d > 0$, $\operatorname{Im} \mu + d\pi > 0$, $S = S_4 \cap S_5$,
10. $d > 0$, $\operatorname{Im} \mu + d\pi < 0$, $S = S_5 \cap S_6$.

In the first six cases, $\partial g / \partial z = 0$ has a unique solution s_n given by

$$s_n = -\frac{n+\rho}{\mu}. \tag{2.19}$$

Hence

$$\arg s_n = \arg\left(-\frac{1}{\mu}\right)(1+o(1)), \qquad n \to \infty. \tag{2.20}$$

Furthermore, we have

$$\frac{\partial^2 g}{\partial z^2}(s_n, n) = -\frac{\mu^2}{n+\rho}, \qquad s_n^2 \frac{\partial^2 g}{\partial z^2}(s_n, n) = -n-\rho, \tag{2.21}$$

$$z\frac{\partial^3 g}{\partial z^3}(z, n)\left\{\frac{\partial^2 g}{\partial z^2}(z, n)\right\}^{-1} = -2. \tag{2.22}$$

Let $S' \Subset S$. In each of the cases 1-6 there exists a positive number δ such that

$$\cos(\arg z + \arg \mu) < -\frac{\delta}{|\mu|} \quad \text{for all } z \in S'.$$

This implies that, for all $z \in S'$,

$$\operatorname{Re} g(z, n) \leqq -\delta|z| + (n + \operatorname{Re} \rho)\log|z| - \operatorname{Im} \rho \arg z.$$

Hence we easily deduce the existence of positive constants $A_{S'}$ and $C_{S'}$ such that

$$\sup_{z \in S'} |\exp g(z, n)| < C_{S'} A_{S'}^n n^n. \tag{2.23}$$

Now consider the cases 7-10. There $d \neq 0$ and the saddle point s_n is a solution of the equation

$$s_n\left(\log s_n + \frac{\mu}{d} + 1\right) = -\frac{n+\rho}{d}. \tag{2.24}$$

Let h be the inverse of the function $z \to z \log z$ (cf. [9, Ex. III], [4, § 3.6]). It has the following asymptotic behavior:

$$h(z) = \frac{z}{\log z}(1+o(1)), \qquad z \to \infty. \tag{2.25}$$

From (2.24) we deduce

$$
\begin{aligned}
s_n &= \exp\left(-\frac{\mu}{d}-1\right) h\left(-\frac{n+\rho}{d}\exp\left(\frac{\mu}{d}+1\right)\right) \\
&= -\frac{n+\rho}{d}\left\{\log h\left(-\frac{n+\rho}{d}\exp\left(\frac{\mu}{d}+1\right)\right)\right\}^{-1}.
\end{aligned} \tag{2.26}
$$

With (2.25) it follows that

$$
s_n = -\frac{n}{d\log n}(1+o(1)), \qquad n\to\infty. \tag{2.27}
$$

Equating the imaginary parts on both sides of (2.24), we get

$$
\operatorname{Im} s_n\left(\log|s_n|+\frac{\operatorname{Re}\mu}{d}+1\right)+\operatorname{Re} s_n\left(\arg s_n+\frac{\operatorname{Im}\mu}{d}\right)=\frac{\operatorname{Im}\rho}{d}.
$$

With (2.27) we find

$$
\operatorname{Im} s_n = \frac{n}{d(\log n)^2}\left(\arg s_n+\frac{\operatorname{Im}\mu}{d}\right)(1+o(1)), \qquad n\to\infty.
$$

Hence

$$
\operatorname{Im} s_n = \begin{cases} \dfrac{n}{d^2(\log n)^2}\operatorname{Im}\mu(1+o(1)), & n\to\infty \quad \text{if } d<0, \\ \dfrac{n}{d^2(\log n)^2}(\operatorname{Im}\mu+d\pi)(1+o(1)), & n\to\infty \quad \text{if } d>0. \end{cases} \tag{2.28}
$$

Furthermore, we have

$$
\frac{\partial^2 g}{\partial z^2}(s_n,n)=\frac{d}{s_n}-\frac{n+\rho}{s_n^2}=-\frac{d^2}{n+\rho}\log h\left(-\frac{n+\rho}{d}\exp\left(\frac{\mu}{d}+1\right)\right)(1+o(1)), \quad n\to\infty \tag{2.29}
$$

and hence

$$
s_n^2\frac{\partial^2 g}{\partial z^2}(s_n,n)=-n(1+o(1)), \qquad n\to\infty. \tag{2.30}
$$

We easily verify that

$$
z\frac{\partial^3 g}{\partial z^3}(z,n)\left\{\frac{\partial^2 g}{\partial z^2}(z,n)\right\}^{-1}=-\frac{2(n+\rho)-dz}{n+\rho-dz} \tag{2.31}
$$

and the expression on the right-hand side is obviously uniformly bounded on the half plane $-d\ \operatorname{Re} z>0$ and thus on S, provided $n\geqq n_0$, where n_0 is some sufficiently large number.

Let $S'\Subset S$. In each of the cases considered this implies the existence of a positive number δ such that

$$
d\cos\arg z<-\delta \quad \text{for all } z\in S'.
$$

Let $0<\varepsilon<\delta$. Then there exists a positive constant C such that

$$
|\exp g(z,n)|<C\exp(-\varepsilon|z|\log|z|)|z|^n, \qquad z\in S'.
$$

The expression to the right of the inequality sign attains its maximum as $|z| = h(ne/\varepsilon)/e$ and the maximum value is equal to

$$\exp\left(-2n+\frac{\varepsilon}{e}h\left(\frac{ne}{\varepsilon}\right)\right)h\left(\frac{ne}{\varepsilon}\right)^n.$$

In view of (2.25) it follows that there exist positive constants $A_{S'}$ and $C_{S'}$ such that

$$\sup_{z\in S'}|\exp g(z,n)| < C_{S'}A_{S'}^n\left(\frac{n}{\log n}\right)^n. \tag{2.32}$$

With the aid of (2.19), (2.21), (2.26), and (2.29) we can derive an explicit expression for the function $M:\mathbb{N}\to\mathbb{C}$ given by

$$M(n) = \left(2\pi s_n^2\frac{\partial^2 g}{\partial z^2}(s_n,n)\right)^{-1/2} s_n \exp g(s_n,n), \tag{2.33}$$

where $\arg\,(s_n^2(\partial^2 g/\partial z^2)(s_n,n))^{-1/2}\in(-\pi,0)$, in each of the cases considered above. With (2.30) we find

$$M(n) = \begin{cases} \{-2\pi(n+\rho)\}^{-1/2}\exp(-n-\rho)\left(\dfrac{n+\rho}{\mu}\right)^{n+\rho+1} & \text{if } d=0,\\ (-2\pi n)^{-1/2}\exp\{(n+\rho)\chi(n)^{-1}-1\}\left(\dfrac{n+\rho}{d\chi(n)}\right)^{n+\rho+1}(1+o(1)), & n\to\infty, \text{if } d\neq 0, \end{cases}$$

where $\chi(n) = \log h((n+\rho)/d\,\exp(\mu/d+1))$. Let us define a function $M_{d,\mu}:\mathbb{C}\to\mathbb{C}$ by

(2.34)

$$-2\pi i M_{d,\mu}(s) = \begin{cases} \Gamma(s)(-\mu)^{-s} & \text{if } d=0,\\ \Gamma(s)\exp\left\{\dfrac{s}{\log h(-s/d\,\exp(\mu/d+1))}\right\}\left\{-d\log h\left(-\dfrac{s}{d}\exp\left(\dfrac{\mu}{d}+1\right)\right)\right\}^{-s}, & \text{if } d\neq 0. \end{cases}$$

Using Stirling's formula and the properties of the function h, we readily verify that

$$-M(n-1) = M_{d,\mu}(n+\rho)(1+o(1)), \qquad n\to\infty. \tag{2.35}$$

Now let $\nu\in\{1,\cdots,6\}$, $\tilde{z}_\nu\in S_\nu\cap S_{\nu+1}$, and let $\tilde{S}^\nu$ be a sector of the following form:

$$\tilde{S}^\nu = S\left(\tilde{z}_\nu,\left(\frac{\nu}{3}-\frac{5}{6}\right)\pi+\delta,\frac{\nu-1}{3}\pi\right) \quad \text{if } \nu\in\{1,4\},$$

$$\tilde{S}^\nu = S\left(\tilde{z}_\nu,\frac{\nu-2}{3}\pi,\left(\frac{\nu}{3}-\frac{1}{6}\right)\pi\right) \quad \text{if } \nu\in\{2,5\},$$

$$\tilde{S}^\nu = S\left(\tilde{z}_\nu,\left(\frac{\nu}{3}-1\right)\pi+\delta,\frac{\nu}{3}\pi-\delta\right) \quad \text{if } \nu\in\{3,6\},$$

where $\delta\in(0,\pi/2)$. Let $i,j\in\{1,\cdots,m\}$ such that $c_{ij}^\nu\neq 0$, and let

$$g_{ij}^\nu(z,n) = d_{ij}z\log z+\mu_{ij}^\nu z+(n+\rho_{ij})\log z, \qquad z\in\tilde{S}^\nu,\quad n\in\mathbb{N}.$$

From (2.19)-(2.22) and (2.27)-(2.31) we deduce that conditions (ii)-(iv) of Proposition 1.3 are satisfied, provided δ is chosen sufficiently small. We readily verify that condition (v) holds as well (with $z_0=\tilde{z}_\nu$).

Next, we consider the function f_{ij}^{ν} defined by (2.16). The asymptotic properties of h_i^{ν} imply that

$$\lim_{z\to\infty} h_i^{\nu}(z) = 1 \quad \text{uniformly on } \tilde{S}^{\nu}. \tag{2.36}$$

Furthermore, from (2.14) and the fact that p_{ij}^{ν} is analytic on either a lower or an upper half plane it follows that

$$|p_{ij}^{\nu}(z)\exp(-2n_{ij}^{\nu}\pi i z) - c_{ij}^{\nu}| \leqq K\exp(-2\pi|\operatorname{Im} z|), \qquad z\in\tilde{S}^{\nu}, \tag{2.37}$$

where K is a positive constant. From (2.36) and (2.37) it is obvious that f_{ij}^{ν} is bounded on $\tilde{S}^{\nu}$. Moreover, with the aid of (2.20) it is easily seen that, in the case that $d_{ij}=0$, f_{ij}^{ν} satisfies condition (vi) of Proposition 1.3. Now suppose that $\nu\in\{1,2,4,5\}$ and $d_{ij}\neq 0$. Formulas (2.4) and (2.28) imply that $|\operatorname{Im} s_n|\to\infty$ as $n\to\infty$, where s_n denotes the saddle point of $g_{ij}^{\nu}(z,n)$. With (2.36) and (2.37) it follows that, also in this case, condition (vi) of Proposition 1.3 is fulfilled.

Apparently, all conditions of Theorem 1.4 are satisfied. Applying this theorem and using (2.33) and (2.35), we obtain the following result.

THEOREM 2.1. *For each $j\in\{1,\cdots,m\}$ there exists a convergent power series $\sum_{n=0}^{\infty} h_{jn}z^{-n}$ such that*

$$\hat{h}_{jn} = h_{jn} + \sum_{i=1}^{m}\sum_{\nu=1}^{6} c_{ij}^{\nu}\{M_{d_{ij},\mu_{ij}^{\nu}}(n+\rho_{ij})(1+o(1))\}, \qquad n\to\infty,$$

where c_{ij}^{ν} and $M_{d_{ij},\mu_{ij}^{\nu}}$ are defined by (2.14) *and* (2.34), *respectively.*

With the aid of Propositions 1.1 and 1.2 we are able to estimate the growth of the remainder terms $R_n(h_j^{\nu};z)$ for $n\to\infty$, $j\in\{1,\cdots,m\}$. Let $\nu\in\{1,\cdots,6\}$. $S_{\nu}\cap S_{\nu+1}$ is a sector of the form $S(z_{\nu},\alpha_{\nu},\beta_{\nu})$. We begin by considering the functions h_{ij}^{ν} defined by

$$h_{ij}^{\nu}(z) = \frac{z}{2\pi i}\int_{\gamma_{\nu}} \frac{\varphi_{ij}^{\nu}(\zeta)}{\zeta(\zeta-z)}\,d\zeta, \qquad i,j\in\{1,\cdots,m\}, \quad \nu\in\{1,\cdots,6\}, \tag{2.38}$$

where γ is a half line in $S_{\nu}\cap S_{\nu+1}$ from z_{ν} to ∞ and φ_{ij}^{ν} is defined by (2.17).

PROPOSITION 2.2. *Let $i,j\in\{1,\cdots,m\}$, $\nu\in\{1,\cdots,6\}$. The function h_{ij}^{ν} defined by* (2.38) *is analytic in $S_{\nu}\cap S_{\nu+1}$ and represented asymptotically by*

$$\sum_{n=0}^{\infty} -\frac{1}{2\pi i}\left(\int_{\gamma_{\nu}} \varphi_{ij}^{\nu}(\zeta)\zeta^{n-1}\,d\zeta\right) z^{-n}$$

as $z\to\infty$ in $S_{\nu}\cap S_{\nu+1}$. Moreover, for every $S'\Subset S_{\nu}\cap S_{\nu+1}$ there exist positive constants $A_{S'}$ and $C_{S'}$ such that, for all $n\in\mathbb{N}$,

$$\sup_{z\in S'}|z^n R_n(h_{ij}^{\nu};z)| \leqq \begin{cases} C_{S'}A_{S'}^n n! & \text{if } d_{ij}=0,\\ C_{S'}A_{S'}^n(n/\log n)^n & \text{if } d_{ij}\neq 0.\end{cases} \tag{2.39}$$

Proof. The first two statements follow immediately from Proposition 1.1 and the properties of φ_{ij}^{ν}. Now let $S'\Subset S_{\nu}\cap S_{\nu+1}$. We can choose a sector $S''\Subset S_{\nu}\cap S_{\nu+1}$ of the form $S''=S(\tilde{z}_{\nu},\tilde{\alpha}_{\nu},\tilde{\beta}_{\nu})$ such that $S'\Subset S''$. Let $\tilde{\gamma}_{\nu}$ be a half line in S'' from $\tilde{z}_{\nu}$ to ∞ and

$$\tilde{h}_{ij}^{\nu}(z) = \frac{z}{2\pi i}\int_{\tilde{\gamma}_{\nu}} \frac{\varphi_{ij}^{\nu}(\zeta)}{\zeta(\zeta-z)}\,d\zeta.$$

As $h_{ij}^{\nu}-\tilde{h}_{ij}^{\nu}$ is holomorphic at ∞, it is obviously sufficient to prove (2.39) for $\tilde{h}_{ij}^{\nu}$ instead of h_{ij}^{ν}. Using (2.18), (2.23), and (2.32) and noting that, due to (2.16), (2.36), and (2.37),

f_{ij}^{ν} is bounded on S'', we conclude that there exist positive numbers $A_{S''}$ and $C_{S''}$ such that, for all $n \in \mathbb{N}$,

$$\sup_{z \in S''} |z^n \varphi_{ij}^{\nu}(z)| \leqq \begin{cases} C_{S''} A_{S''}^n n! & \text{if } d_{ij} = 0, \\ C_{S''} A_{S''}^n (n/(\log n))^n & \text{if } d_{ij} \neq 0. \end{cases}$$

The result now follows by application of Proposition 1.1.

THEOREM 2.3 (cf. also [7]). *Let $j \in \{1, \cdots, m\}$, $\nu \in \{1, \cdots, 6\}$. For every $S' \Subset S_{\nu}$ there exist positive constants $A_{S'}$ and $C_{S'}$ such that*

$$\sup_{z \in S'} |z^n R_n(h_j^{\nu}; z)| < C_{S'} A_{S'}^n n!, \qquad n \in \mathbb{N}$$

Moreover, if the numbers c_{ij}^{μ} defined by (2.14) vanish for all $i \in \{1, \cdots, m\}$ such that $d_{ij} = 0$ and all $\mu \in \{1, \cdots, 6\}$, then there exist positive constants $\tilde{C}_{S'}$ and $\tilde{A}_{S'}$ such that

$$\sup_{z \in S'} |z^n R_n(h_j^{\nu}; z)| < \tilde{C}_{S'} \tilde{A}_{S'}^n \left(\frac{n}{\log n}\right)^n.$$

Proof. Using (2.11), (2.12), and the definitions (2.17) and (2.38), and applying Proposition 1.2, we conclude that there exists a function h_j, holomorphic at ∞, such that

$$h_j^{\nu}(z) = h_j(z) + \sum_{i=1}^{m} \left\{ \sum_{\mu=1}^{\nu-1} h_{ij}^{\nu}(z) + \sum_{\mu=\nu}^{6} h_{ij}^{\mu}(z\, e^{2\pi i}) \right\}.$$

Thus the statements of the theorem are seen to be an immediate corollary of Proposition 2.2.

To conclude this section we shall apply the above results to the second-order difference equation

$$\{(z+2)^2 + \alpha(z+2) + \beta\} y(z+2) - \{(z+1)^2 + \gamma(z+1)^2 + \gamma(z+1) + \delta\} y(z+1) + \sigma y(z) = 0,$$

where $\alpha, \beta, \gamma, \delta \in \mathbb{C}$, $\sigma \in \mathbb{C}^*$ (this is a particular case of the class of equations considered in [1]). This equation possesses two formal solutions $\hat{y}_1$ and $\hat{y}_2$ of the form

$$\hat{y}_1(z) = \hat{h}_1(z) z^{\gamma - \alpha - 2},$$

$$\hat{y}_2(z) = \hat{h}_2(z) z^{-\gamma - 2} \exp\{-2z \log z + (2 + \log \sigma) z\},$$

where $\hat{h}_j(z) = \sum_{n=0}^{\infty} \hat{h}_{jn} z^{-n}$ with $\hat{h}_{j0} = 1$, $j = 1, 2$. Thus we have

$$\rho_{12} = 2\gamma - \alpha = -\rho_{21}, \quad d_{12} = 2 = -d_{21}, \quad \mu_{12} = -(2 + \log \sigma) = -\mu_{21}.$$

Assumption (2.4) is equivalent to

$$\arg \sigma \neq 0 \bmod 2\pi.$$

We shall choose $\arg \sigma \in (0, 2\pi)$. With (2.5) it follows that $n_{12} = n_{21} = 0$. Hence, by (2.9) the following limits exist:

$$\lim_{\operatorname{Im} z \to -\infty} p_{21}^1(z) \exp 2\pi i z \quad \text{and} \quad \lim_{\operatorname{Im} z \to \infty} p_{21}^2(z).$$

From these and other considerations, based on the particular form of the equation, it can be deduced that the periodic functions p_{21}^1, p_{21}^2, p_{11}^3 and p_{11}^6 must be of the following form:

$$p_{11}^3(z) = \frac{c_{11}^3 \exp 2\pi i z + (\exp 2\pi i y - \exp 2\pi i \alpha) \exp 4\pi i z}{(1 - \exp 2\pi i (z - a))(1 - \exp 2\pi i (z - b))}, \tag{2.40}$$

$$p_{11}^6(z) = (1 + p_{11}^3(z))^{-1} \exp 2\pi i (\gamma - \alpha) - 1, \tag{2.41}$$

(2.42)

$$p_{21}^1(z) = -p_{21}^2(z)$$
$$= \frac{-c_{21}^2 + c_{21}^1 \exp 2\pi i(z-\gamma)}{1+\{c_{11}^3 - \exp(-2\pi i a) - \exp(-2\pi i b)\}\exp 2\pi i z + \exp 2\pi i(\gamma+2z)},$$

where a and b denote the roots of the polynomial $z^2+\alpha z+\beta$, and c_{21}^1, c_{21}^2, and c_{11}^3 are defined by (2.14). From (2.7), (2.9), and (2.10) it is seen that $c_{11}^\nu = 0$ for $\nu \in \{1, 2, 4, 5\}$ and $c_{21}^\nu = 0$ for $\nu \in \{3, 4, 5, 6\}$. According to Theorem 2.1 there exists a convergent power series $\sum_{n=0}^{\infty} h_{1n} z^{-n}$ such that

$$\begin{aligned}\hat{h}_{1n} = h_{1n} &+ c_{11}^3 M_{0,\mu_{11}^3}(n)(1+o(1)) + c_{11}^6 M_{0,\mu_{11}^6}(n)(1+o(1)) \\ &+ c_{21}^1 M_{-2,\mu_{21}^1}(n+\alpha-2\gamma)(1+o(1) \\ &+ c_{21}^2 M_{-2,\mu_{21}^2}(n+\alpha-2\gamma)(1+o(1)), \qquad n\to\infty.\end{aligned} \tag{2.43}$$

From (2.40)-(2.42) we deduce, with (2.13), that $n_{11}^3 = -n_{11}^6 = 1$, $n_{21}^1 = -1$, $n_{21}^2 = 0$ and hence, with (2.15), that

$$\mu_{11}^3 = -\mu_{11}^6 = 2\pi i, \quad \mu_{21}^1 = 2 + \log\sigma - 2\pi i, \quad \mu_{21}^2 = 2 + \log\sigma.$$

Using (2.34), we find

$$M_{0,\mu_{11}^3}(n) = (-1)^n M_{0,\mu_{11}^6}(n) = \Gamma(n)(-2\pi i)^{-n-1}.$$

As the dominating terms in (2.43) are the ones with coefficients c_{11}^3 and c_{11}^6 we conclude that

$$c_{11}^3 + c_{11}^6 = -\lim_{n\to\infty} \frac{\hat{h}_{12n}(2\pi i)^{2n+1}}{(2n-1)!},$$
$$c_{11}^3 - c_{11}^6 = \lim_{n\to\infty} \frac{\hat{h}_{12n+1}(2\pi i)^{2n+2}}{(2n)!}.$$

If $c_{11}^3 = 0$, then, by (2.41), $p_{11}^6 \equiv \exp 2\pi i(\gamma-\alpha) - 1$ and, in view of (2.8), this implies $c_{11}^6 = 0$ and $\gamma - \alpha \in \mathbb{Z}$. In that case (2.42) becomes

$$p_{21}^1(z) = -p_{21}^2(z) = \frac{-c_{21}^2 + c_{21}^1 \exp 2\pi i(z-\alpha)}{(1-\exp 2\pi i(z-a))(1-\exp 2\pi i(z-b))}$$

(where we have used the identity $a+b=-\alpha$), and the coefficient c_{21}^ν, $\nu \in \{1, 2\}$, of the dominating term in (2.43) may be determined from the asymptotic behavior of $\hat{h}_{1n}$ for $n\to\infty$.

On the other hand, if $c_{11}^3 \neq 0$, then the coefficients c_{21}^1 and c_{21}^2 cannot be determined by this method.

REFERENCES

[1] B. L. J. Braaksma and W. A. Harris, jr., *On an open problem in the theory of linear difference equations*, Nieuw Arch. Wisk. (3), 23 (1975), pp. 228-240.

[2] B. L. J. Braaksma and G. K. Immink, *A Borel-Ritt theorem with prescribed error bounds*, in Equations différentielles dans le champ complexe, Vol. I: équations différentielles ordinaires, Publication I.R.M.A. Strasbourg, 1988, pp. 1-33.

[3] A. Duval, *Lemmes de Hensel et factorisation formelle pour les opérateurs aux différences*, Funkcial. Ekvac., 26 (1983), pp. 349-368.

[4] J. Ecalle, *Les fonctions resurgentes*, tome III, Publ. Math. d'Orsay, Université de Paris-Sud, Paris, 1985.

[5] J. K. Hale, *Ordinary Differential Equations*, Interscience Publishers, New York, 1969.

[6] G. K. IMMINK, *Reduction to canonical forms and the Stokes phenomenon in the theory of linear difference equations*, SIAM J. Math. Anal., 22 (1991), pp. 239-260.

[7] ———, *Asymptotic expansions with error bounds for solutions of difference equations of "level* 1^+*,"* in Equations différentielles dans le champ complexe, Vol. I: équations différentielles ordinaires, Publication I.R.M.A. Strasbourg, 1988, pp. 35-60.

[8] ———, *Resurgent functions and connection matrices for a linear homogeneous system of difference equations*, Funkcial. Ekvac., 31 (1988), pp. 197-219.

[9] ———, *Note on the relation between Stokes multipliers and formal solutions of analytic differential equations*, SIAM J. Math. Anal., 21 (1990), pp. 782-792.

[10] B. MALGRANGE, *Remarques sur les équations différentielles à points singuliers irréguliers*, Lecture Notes in Math. 712, Springer-Verlag, Berlin, 1979, pp. 77-86.

[11] C. PRAAGMAN, *The formal classification of linear difference operators*, Proc. Kon. Ned. Ac. Wet. Ser. A, 86 (1983), pp. 249-261.

[12] J.-P. RAMIS, *Les séries k-sommables et leurs applications*, Lecture Notes in Phys. 126, Springer-Verlag, Berlin, 1980, pp. 178-199.

[13] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, Interscience Publishers, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 543–553, March 1991

015

INVERTIBILITY OF SHIFTED BOX SPLINE INTERPOLATION OPERATORS *

C. K. CHUI†, J. STÖCKLER‡, AND J. D. WARD†

Abstract. Cardinal interpolation by integer-translates of shifted box splines $M_{n,\alpha} := M_{nnn}(\cdot + \alpha)$ on the three-direction mesh is studied. It was recently shown by Sivakumar that for even integers n the imaginary part of a certain rotation of the symbol of $M_{n,\alpha}$ does not vanish on the torus T^2 for all α in the shift region $\Lambda = (-\frac{1}{2}, \frac{1}{2})^2 \cap \{(s,t)\colon |s-t| < \frac{1}{2}\}$, and consequently, cardinal interpolation at $\mathbb{Z}^2$ by using $M_{n,\alpha}(\cdot - j), j \in \mathbb{Z}^2$, is poised for all even n and all $\alpha \in \Lambda$. For odd n, however, since both the real and imaginary parts of the rotated symbol have nonempty zero sets on T^2 for certain $\alpha \in \Lambda$ close to the corners of $\partial\Lambda$, the analysis in Sivakumar's work does not directly apply to the study of this situation. In this paper we prove that the above mentioned zero sets are disjoint for all odd integers n and all $\alpha \in \Lambda$, and hence, the symbol of $M_{n,\alpha}$ never vanishes on T^2. In other words, the cardinal interpolation operators corresponding to $M_{n,\alpha}, \alpha \in \Lambda$ and $n = 1, 2, \cdots$, are invertible.

Key words. cardinal interpolation, poisedness, shifted box splines, symbols

AMS(MOS) subject classifications. 41A05, 41A15, 41A63

1. Introduction and results.

Let ϕ be a piecewise continuous real-valued function with compact support in $\mathbb{R}^s, s \geq 1$, and let

$$S(\phi) := \left\{ \sum_{j \in \mathbb{Z}^s} a_j \phi(\cdot - j)\colon\ a_j \in \mathbb{R} \right\}.$$

The problem of *cardinal interpolation* from $S(\phi)$ is to study the existence and uniqueness of a bounded "coefficient" sequence $\{a_j\} \subset \mathbb{R}$ corresponding to any given bounded "data" sequence $\{f_j\} \subset \mathbb{R}$, such that the "spline" function $\Sigma a_j \phi(\cdot - j)$ from $S(\phi)$ agrees with $\{f_j\}$ on $\mathbb{Z}^s$; that is,

$$\sum_{j \in \mathbb{Z}^s} a_j \phi(i-j) = f_i, \qquad i \in \mathbb{Z}^s. \tag{1.1}$$

This problem is said to be *poised* (or *correct*) if corresponding to any bounded data sequence $\{f_j\}$ there exists a unique bounded coefficient sequence $\{a_j\}$ such that (1.1) is satisfied. The *discrete Fourier transform* $\tilde{\phi}$ of ϕ, defined by

$$\tilde{\phi}(x) = \sum_{j \in \mathbb{Z}^s} \phi(j) e^{-ij\cdot x}, \qquad x \in \mathbb{R}^s, \tag{1.2}$$

plays a central role in the study of the above problem. (Note that $\tilde{\phi}$ agrees with the restriction on the torus T^s of the *symbol* or *z-transform* $\Sigma \phi(j) z^{-j}$ of ϕ.) In fact, *the*

*Received by the editors July 31, 1989; accepted for publication (in revised form) March 27, 1990.

†Center for Approximation Theory, Texas A&M University, College Station, Texas 77843. The works of these authors was supported by National Science Foundation grants DMS-8602337 and DMS-8701190.

‡Department of Mathematics, University of Duisburg, D-4100 Duisburg, Federal Republic of Germany.

problem of cardinal interpolation from $S(\phi)$ is poised if and only if $\tilde{\phi}$ never vanishes (cf. [1], [6], and [3]). Hence, in this paper, we will concentrate on establishing the strict positivity of $|\phi(x)|, x \in \mathbb{R}^s$.

A somewhat stronger condition is the so-called "metric condition"

$$|1 - \tilde{\phi}(x)| < 1, \qquad x \in \mathbb{R}^s, \tag{1.3}$$

first considered in this context in [4]. If $\tilde{\phi}$ satisfies (1.3), then not only the cardinal interpolation problem has a unique solution to any given bounded data sequence, but this solution can also be constructed using the Neumann series (cf. [5] and [3]). Another important reason for considering a condition such as the metric condition in (1.3) is that since a multivariate analogue of the univariate "total positivity" does not exist, some condition which is not described in terms of sign changes, such as the metric condition (1.3) (or perhaps more suitably a somewhat weaker one), is needed to take place of total positivity in $\mathbb{R}^s, s \geq 2$. For instance, let $B = B_k := \chi_{(-\frac{1}{2},\frac{1}{2})} * \cdots * \chi_{(-\frac{1}{2},\frac{1}{2})}$ be the kth order centered B-spline, where a k-fold convolution of the characteristic function of $(-\frac{1}{2}, \frac{1}{2})$ is taken. Then by using total positivity, it was shown in Micchelli [7] and de Boor and Schoenberg [2] that the cardinal interpolation operator corresponding to the shifted B-spline $B_\alpha := B(\cdot + \alpha), |\alpha| \leq \frac{1}{2}$ is invertible if and only if $\alpha \neq \pm\frac{1}{2}$. In Sivakumar [8], this invertibility result was recovered by showing that $\widetilde{B}_\alpha$ never vanishes for $|\alpha| < \frac{1}{2}$. Recently, in Smith and Ward [10], it was further shown that the metric condition

$$|1 - \widetilde{B}_\alpha(x)| < 1, \qquad x \in \mathbb{R},$$

is satisfied for $|\alpha| \leq \frac{1}{2}$ if and only if $\alpha \neq \pm\frac{1}{2}$. Hence, it would seem to be hopeful that the metric condition can sometimes replace the total positivity property in certain multivariate derivations.

In this paper, we will restrict attention to bivariate box splines on a three-direction mesh. Let M_{nnn} denote the centered box spline with directions (1,0), (0,1), and (1,1), each repeated n times (cf. [3], Chap. 2). We will study the shifted box spline

$$M_{n,\alpha} := M_{nnn}(\cdot + \alpha), \tag{1.4}$$

where α is in the shift region

$$\Lambda := (-\tfrac{1}{2}, \tfrac{1}{2})^2 \cap \{(s,t)\colon\ |s - t| < \tfrac{1}{2}\}, \tag{1.5}$$

(cf. Fig. 1). Note that Λ is a largest connected region for the shift parameter α in the sense that for each $\alpha \in \partial\Lambda$, the zero set of $\widetilde{M}_{n,\alpha}$ is nonempty (cf. [9] and [12]). Recently, Sivakumar [9] showed that when n is even, $Im(e^{-i2\pi\alpha\cdot x}\widetilde{M}_{n,\alpha}(x)) \neq 0$ for all $x \in \mathbb{R}^2$. For odd $n > 1$, however, this conclusion no longer holds, and in fact, both of the zero sets of the real and imaginary parts of $e^{-i2\pi\alpha\cdot x}\widetilde{M}_{n,\alpha}(x)$ are nonempty. The objective of this paper is to show that these two sets are disjoint. More precisely, the following result will be established.

THEOREM 1. *For each $n = 1, 2, \cdots, \widetilde{M}_{n,\alpha}(x) \neq 0$ for all $x \in \mathbb{R}^2$ and $\alpha \in \Lambda$.*

Of course, as mentioned above, for even n this result was already established in [9]. It is interesting to note that due to the nature of the zero sets of the real and imaginary parts, the proof for odd integers n is much more involved, and in fact we used computer experiments to locate a separation of these zero sets.

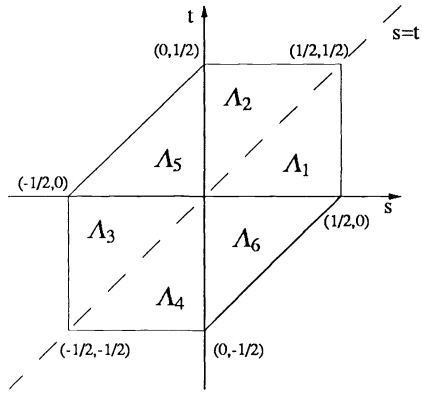


FIG. 1

A very brief outline of the proof of Theorem 1 will be given in §2. Necessary lemmas and Theorem 1 will be proved in §3. In the final section, numerical examples will be given to demonstrate that in contrast with the univariate B-spline situation, the metric condition (1.3) is not satisfied by every $\widetilde{M}_{n,\alpha}, \alpha \in \Lambda$, and this leaves room for further investigation.

2. Brief outline of proof. Let $\hat{\phi}$ denote, as usual, the Fourier transform of ϕ. Then by an application of the Poisson Summation formula (cf. [11, p. 49]), we have

$$\begin{aligned} \widetilde{M}_{n,\alpha}(x) &= \sum_{j\in\mathbb{Z}^s} \widehat{M}_{n,\alpha}(x+2\pi j) \\ &= \sum_{j\in\mathbb{Z}^s} \widehat{M}_{nnn}(x+2\pi j)e^{i\alpha\cdot(x+2\pi j)}, \end{aligned} \tag{2.1}$$

where $M_{n,\alpha}$ is the shift of M_{nnn} by α as defined in (1.4). Also, let $\mathcal{A}$ denote the multiplicative group of 2×2 matrices isomorphic to the permutation group S_3; that is, $\mathcal{A}$ has order 6 and consists of the matrices: $\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right), \left(\begin{smallmatrix}0&-1\\-1&0\end{smallmatrix}\right), \left(\begin{smallmatrix}1&1\\0&-1\end{smallmatrix}\right), \left(\begin{smallmatrix}-1&0\\1&1\end{smallmatrix}\right), \left(\begin{smallmatrix}-1&-1\\1&0\end{smallmatrix}\right), \left(\begin{smallmatrix}0&1\\-1&-1\end{smallmatrix}\right)$. As noted in [9], the following identities hold for all $A \in \mathcal{A}$:

$$\begin{cases} M_{n,\pm\alpha A}(\pm xA) = M_{n,\alpha}(x), \\ \widetilde{M}_{n,\pm\alpha A}(x) = \widetilde{M}_{n,\alpha}(\pm xA^t), \\ \widehat{M}_{n,\pm\alpha A}(x) = \widehat{M}_{n,\alpha}(\pm xA^t), \end{cases} \tag{2.2}$$

where, as usual, A^t denotes the transpose of A. The identities in (2.2) allow us to restrict our attention to a certain subregion of $\mathbb{R}^2$ in establishing $\widetilde{M}_{n,\alpha}(x) \neq 0, x \in \mathbb{R}^2$.

Let $x = (2\pi u, 2\pi v)$ and let Δ denote the closed triangular region with vertices at (0,0), $(\frac{1}{2},0)$, and $(\frac{1}{3},\frac{1}{3})$. Then the set

$$F := \bigcup_{\pm A\in\mathcal{A}} \{(u,v)A^t\colon\ (u,v)\in\Delta\},$$

shown in Fig. 2, is a *fundamental region* in the sense that with a portion of its boundary removed, the translates of F over $\mathbb{Z}^2$ form a disjoint partition of $\mathbb{R}^2$. It is not difficult to verify that F is invariant under the transformations $\pm A^t$ for all $A \in \mathcal{A}$, and that

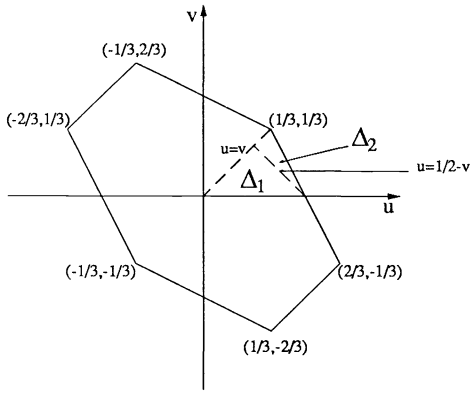


FIG. 2

the shift region Λ is invariant under the transformations $A \in \mathcal{A}$. Hence, in view of (2.2) and the definition of F, in order to prove that $\widetilde{M}_{n,\alpha}(x) \neq 0$ for all $x \in \mathbb{R}^2$ and $\alpha \in \Lambda$, it is sufficient to show that $\widetilde{M}_{n,\alpha}(x) \neq 0$ for all $x \in 2\pi\triangle$ and $\alpha \in \Lambda$. For this purpose, we divide $\triangle$ into two sub-triangles:

$$\begin{cases} \triangle_1 = \{(u,v) \in \triangle\colon\ u+v \leq \frac{1}{2}\}, \\ \triangle_2 = \left\{(u,v) \in \triangle\colon\ u+v > \frac{1}{2}\right\} \end{cases} \tag{2.3}$$

(cf. Fig. 2). Instead of showing that $\widetilde{M}_{n,\alpha}(x) \neq 0$, we follow [9] and show that a related series does not vanish on $\triangle$. First, consider the rotated symbol

$$\begin{aligned} P_{n,\alpha}(u,v) &:= e^{-i2\pi\alpha\cdot(u,v)}\widetilde{M}_{n,\alpha}(2\pi u, 2\pi v) \\ &= \sum_{j,k\in\mathbb{Z}} e^{i2\pi(js+kt)}\Big(\frac{\sin\pi u}{\pi(u+j)}\cdot\frac{\sin\pi v}{\pi(v+k)}\cdot\frac{\sin\pi(u+v)}{\pi(u+v+j+k)}\Big)^n. \end{aligned} \tag{2.4}$$

Here and throughout, we set $\alpha = (s,t)$. For $(u,v) \in \triangle$, the zero structure of $P_{n,\alpha}$ is the same as that of

$$\begin{aligned} Q_{n,\alpha}(u,v) &:= \Big(\frac{\pi u}{\sin\pi u}\cdot\frac{\pi v}{\sin\pi v}\cdot\frac{\pi(u+v)}{\sin\pi(u+v)}\Big)^n P_{n,\alpha}(u,v) \\ &= \sum_{j,k\in\mathbb{Z}} e^{i2\pi(js+kt)}\Big(\frac{u}{u+j}\Big)^n\Big(\frac{v}{v+k}\Big)^n\Big(\frac{u+v}{u+v+j+k}\Big)^n. \end{aligned} \tag{2.5}$$

We will first show that $Re\ Q_{n,\alpha}(u,v) > 0$ for $(u,v) \in \triangle_1$. To analyse $Q_{n,\alpha}$ on $\triangle_2$, we divide the shift region Λ into six parts $\Lambda_1, \cdots, \Lambda_6$ as shown in Fig. 1. Since we require $\Lambda = \Lambda_1 \cup \cdots \cup \Lambda_6$, the six edges separating $\Lambda_1, \cdots, \Lambda_6$ and joining the origin to the boundary of Λ must be included. For notational convenience, we simply include both such boundary edges of each Λ_i to Λ_i. For $(u,v) \in \triangle_2$ it will be shown that $Im\ Q_{n,\alpha}(u,v) < 0$ if $\alpha = (s,t) \in \Lambda_1 \cup \Lambda_2 \backslash \{0\}$ and $Im Q_{n,\alpha}(u,v) > 0$ if $\alpha = (s,t) \in \Lambda_3 \cup \Lambda_4 \backslash \{0\}$. On the other hand, for $\alpha = (s,t) \in \Lambda_5 \backslash \{0\}$ and $\alpha = (s,t) \in \Lambda_6 \backslash \{0\}$, we will see that $Im(e^{i2\pi t}Q_{n,\alpha}(u,v))$ is strictly positive and strictly negative, respectively, for all $(u,v) \in \triangle_2$. Of course the isolated situation $\alpha = 0$ not considered here corresponds to the centered box spline M_{nnn} result already established by de Boor, Höllig, and Riemenschneider in [1].

3. Proof of main result. Theorem 1 can be easily verified for $n = 1$, and, as mentioned above, has already been established by Sivakumar [9] for even n. In this section, we will present a proof for all integers $n \geq 3$, although some of the estimates still hold for $n = 2$. To facilitate our argument, several simple observations are included in the following lemma.

LEMMA 1. *Let* $k, \ell, m \in \mathbb{Z}\backslash\{0\}$ *and* $0 \leq v \leq u < 1$. *Then*

(a) $|\sin 2\pi yk| \leq |k||\sin 2\pi y|$ *for all* $y \in \mathbb{R}$;

(b) *The functions*

$$h_{\pm}(x,y,z) = (1-x)^{-|k|}(1-y)^{-|\ell|}(1-z)^{-|m|} \pm (1+x)^{-|k|}(1+y)^{-|\ell|}(1+z)^{-|m|}$$

are increasing in each of the variables x, y, z *on the interval* $[0,1)$; *moreover,* $|h_-(x,y,z)| \leq h_-(|x|,|y|,|z|)$ *for* $x, y, z \in (-1,1)$.

(c) $|\frac{1-u}{u+k}| \leq |\frac{1-v}{v+k}|$;

(d) $|\frac{v}{v+k}| \leq |\frac{u}{u+k}|$; *and*

(e) $|\frac{1+v}{v+k}| \leq |\frac{1+u}{u+k}|$.

Note that the inequalities (c), (d), and (e) were also used in [9]. We now proceed with our estimates.

LEMMA 2. *Let* $n \geq 2$ *be any integer. Then* $ReQ_{n,(s,t)}(u,v) > 0$ *for all* $(u,v) \in \triangle_1$ *and* $(s,t) \in \Lambda$.

Proof. From (2.5), we have

$$ReQ_{n,(s,t)}(u,v) = \sum_{j,k} \cos 2\pi(js+kt)\Big[\frac{u}{u+j}\cdot\frac{v}{v+k}\cdot\frac{u+v}{u+v+j+k}\Big]^n. \tag{3.1}$$

We split this series into two parts, namely:

$$\begin{aligned}\sum_{j,k} &= \sum_{j\in\mathbb{Z},k=0} + \sum_{j\in\mathbb{Z},k\neq 0} \\ &=: S + T,\end{aligned}$$

and will show that two terms in the series S dominate. To see this, the series S is again written as the sum of two subseries as follows:

$$\begin{aligned} S &= \sum_{j\in\mathbb{Z}} \cos 2\pi js\Big[\frac{u}{u+j}\cdot\frac{u+v}{u+v+j}\Big]^n \\ &= \sum_{j\in\mathbb{Z}}(-1)^j\Big[\frac{u}{u+j}\cdot\frac{u+v}{u+v+j}\Big]^n \\ &\quad + 2\sum_{j\in\mathbb{Z}}(-1)^{j+1}\sin^2\pi j\Big(\frac{1}{2}-s\Big)\Big[\frac{u}{u+j}\cdot\frac{u+v}{u+v+j}\Big]^n \\ &=: S_1 + S_2. \end{aligned} \tag{3.2}$$

For S_1, set $\varepsilon := \frac{1}{2} - u \geq 0$ and

$$\begin{aligned} a_j &:= (-1)^j\Big[\Big(\frac{u}{u+j}\cdot\frac{u+v}{u+v+j}\Big)^n - \Big(\frac{u}{u-j-1}\cdot\frac{u+v}{u+v-j-1}\Big)^n\Big] \\ &= (-1)^j\Big[\Big(\frac{1-2\varepsilon}{2j+1-2\varepsilon}\Big)^n\Big(\frac{1-2(\varepsilon-v)}{2j+1-2(\varepsilon-v)}\Big)^n \\ &\qquad - \Big(\frac{1-2\varepsilon}{2j+1+2\varepsilon}\Big)^n\Big(\frac{1-2(\varepsilon-v)}{2j+1+2(\varepsilon-v)}\Big)^n\Big]. \end{aligned}$$

Here, as a function of (u,v), a_0 is nonnegative and vanishes only when $(u,v)=(\frac{1}{2},0)$. Hence, it follows from Lemma 1(b) and $0 \le v \le \varepsilon \le \frac{1}{2}$ that

$$|a_j| \le \frac{1}{(2j+1)^{2n}} a_0$$

for all $j \ge 1$, and consequently, for all $n \ge 2$,

$$S_1 = \sum_{j=0}^{\infty} a_j \ge a_0 \left[1 - \sum_{j=1}^{\infty} \frac{1}{(2j+1)^{2n}}\right] \ge \frac{67}{68} a_0. \tag{3.3}$$

The subseries S_2 defined in (3.2) can be treated in a similar manner. Let

$$b_j := (-1)^{j+1} 2 \sin^2 \pi j \Big(\frac{1}{2} - s\Big) \Big[\Big(\frac{u}{j-u} \cdot \frac{u+v}{j-u-v}\Big)^n + \Big(\frac{u}{j+u} \cdot \frac{u+v}{j+u+v}\Big)^n\Big] \tag{3.4}$$

where $b_1 \ge 0$ and is strictly positive for $|s| < \frac{1}{2}$ and $(u,v) \ne (0,0)$. By factoring out j^{-2n} in (3.4) and applying Lemma 1 (a) and (b), we have

$$\begin{aligned} S_2 &\ge b_1 \left[1 - \sum_{j=2}^{\infty} \frac{1}{j^{2n}} \frac{\sin^2 \pi j(\frac{1}{2} - s)}{\sin^2 \pi(\frac{1}{2} - s)}\right] \\ &\ge b_1 \left[1 - \sum_{j=2}^{\infty} \frac{1}{j^{2n-2}}\right] \ge 0.35 b_1 \end{aligned} \tag{3.5}$$

for all $n \ge 2$. Hence, combining the information from (3.3) and (3.5), we have

$$S \ge \frac{67}{68} a_0 + 0.35 b_1, \tag{3.6}$$

which is strictly positive for all $(u,v) \in \triangle_1$ and $|s| < \frac{1}{2}$. To study the series T, note that for $(u,v) \in \triangle_1$, the inequalities

$$0 \le u, u+v \le \tfrac{1}{2} \quad \textit{and} \quad \tfrac{3}{4} \le 1 - v \le 1$$

hold. It follows that for $n \ge 2$,

$$\begin{aligned} |T| &\le \Big(\frac{v}{1-v}\Big)^n \sum_{j \in \mathbb{Z}, k \ne 0} \Big|\frac{1}{1+2j}\Big|^n \Big|\frac{1}{k}\Big|^n \Big|\frac{1}{1+2j+2k}\Big|^n \\ &\le \frac{4}{9} v \sum_{j \in \mathbb{Z}, k \ne 0} \Big|\frac{1}{1+2j}\Big|^2 \Big|\frac{1}{k}\Big|^2 \Big|\frac{1}{1+2j+2k}\Big|^2 \\ &\le \frac{5}{4} v. \end{aligned}$$

On the other hand, writing $\varepsilon = \frac{1}{2} - u$, we have

$$\begin{aligned} a_0 &\ge 1 - \Big(\frac{u}{1-u}\Big)^2 = \frac{(1+2\varepsilon)^2 - (1-2\varepsilon)^2}{(1+2\varepsilon)^2} \\ &\ge 2\varepsilon \ge 2v. \end{aligned}$$

This together with the lower bound estimate (3.6) gives

$$ReQ_{n,(s,t)}(u,v) \geq S - |T| \geq \Big(\frac{67}{68} - \frac{5}{8}\Big)a_0 + 0.35b_1 > 0$$

for all $|s| < \frac{1}{2}$, completing the proof of the lemma. □

For $(u,v) \in \triangle_2$, we first consider the shift region $\Lambda_1 \cup \Lambda_2 \cup \Lambda_3 \cup \Lambda_4$ as follows.

LEMMA 3. *Let $n \geq 3$ be an integer and $(u,v) \in \triangle_2$. Then*

$$ImQ_{n,\alpha}(u,v) < 0 \quad \text{for } \alpha \in \Lambda_1 \cup \Lambda_2 \backslash \{0\} \tag{3.7}$$

and

$$ImQ_{n,\alpha}(u,v) > 0 \quad \text{for } \alpha \in \Lambda_3 \cup \Lambda_4 \backslash \{0\}. \tag{3.8}$$

Proof. We first observe that since

$$Q_{n,\alpha}(u,v) = \overline{Q_{n,-\alpha}(u,v)},$$

the conclusions (3.7) and (3.8) are equivalent. To verify (3.7), write

$$ImQ_{n,(s,t)}(u,v) =: T_1 + T_2, \tag{3.9}$$

where

$$T_1 := \sum_{j,k\in\mathbb{Z}} \sin 2\pi sj \cos 2\pi tk \Big[\frac{u}{u+j} \cdot \frac{v}{v+k} \cdot \frac{u+v}{u+v+j+k}\Big]^n$$

and

$$T_2 := \sum_{j,k\in\mathbb{Z}} \cos 2\pi sj \sin 2\pi tk \Big[\frac{u}{u+j} \cdot \frac{v}{v+k} \cdot \frac{u+v}{u+v+j+k}\Big]^n.$$

Our plan is to show that $T_1, T_2 \leq 0$ on $\triangle_2$ and have disjoint zeros there. For the series T_1, we factor out the dominant term

$$c_{-1,0} := -\sin 2\pi s \Big[\frac{u}{1-u} \cdot \frac{u+v}{1-u-v}\Big]^n \leq 0, \tag{3.10}$$

and using the condition $u+v > \frac{1}{2}$ and $u > \frac{1}{4}$ for $(u,v) \in \triangle_2$ and applying Lemma 1 (c)–(e), we conclude that

$$\begin{aligned} T_1 &\leq c_{-1,0}\Big\{1 - \sum_{(j,k)\neq(-1,0)} \Big|\frac{\sin 2\pi sj}{\sin 2\pi s}\Big| |\cos 2\pi tk| \Big|\frac{1-u}{u+j}\Big|^n \Big|\frac{v}{v+k}\Big|^n \Big|\frac{1-u-v}{u+v+j+k}\Big|^n\Big\} \\ &\leq c_{-1,0}\Big\{1 - \sum_{(j,k)\neq(-1,0)} |j| \Big|\frac{3}{1+4j} \cdot \frac{1}{1+3k} \cdot \frac{1}{1+2j+2k}\Big|^n\Big\} \\ &\leq 0.96 c_{-1,0}, \end{aligned}$$

where the numerical constant is obtained by using $n = 3$. Hence, $T_1 \leq 0$ and has zeros in $\triangle_2$ (and in fact vanishes identically) if and only if $s = 0$. To study T_2, we first restrict our attention to $v \geq 0.05$. In this case, factoring out the dominant term

$$d_{0,-1} := -\sin 2\pi t \Big(\frac{v}{1-v} \cdot \frac{u+v}{1-u-v}\Big)^n \leq 0,$$

and applying Lemma 1 (c)–(e), we have, similarly,

$$T_2 \le d_{0,-1}\Big\{1 - \sum_{(j,k)\neq(0,-1)} |k|\Big|\frac{1}{1+2j}\cdot\frac{19}{1+20k}\cdot\frac{1}{1+2j+2k}\Big|^n\Big\}$$
$$\le 0.08 d_{0,-1},$$

where, in addition to the constraint $v \ge 0.05$, $n = 3$ has been used for computing the numerical constant. Note that $T_2 < 0$ if $t \neq 0$, and vanishes only if $t = 0$. Now suppose that $v \le 0.05$. We keep $(u, v) \in \triangle_2$ and t fixed and consider $g(s) := T_2(s, t, u, v)$ as a function of s. We will first show that $g(\frac{1}{2}) < 0$ for $t \neq 0$ and odd n, and $g(0) < 0$ for $t \neq 0$ and even n. Next we will verify that g is an increasing (respectively, decreasing) function on $[0, \frac{1}{2}]$ depending on n odd or even; hence, the proof of the lemma will be complete.

To establish the first claim for odd n, set $\varepsilon = \frac{1}{2} - u$ and consider the sum of the (j, k) and $(-j-1, -k)$ terms of $[uv(u+v)]^{-n} g(1/2)$, namely:

(3.11)
$$\begin{aligned} g_{jk} &:= (-1)^j 2^{2n} \sin 2\pi tk\{(2j+1-2\varepsilon)^{-n}(k+v)^{-n}(2j+2k+1+2(v-\varepsilon))^{-n} \\ &\qquad - (2j+1+2\varepsilon)^{-n}(k-v)^{-n}(2j+2k+1-2(v-\varepsilon))^{-n}\} \\ &= (-1)^j 2^{2n}(\sin 2\pi tk)(2j+1)^{-n}k^{-n}(2j+2k+1)^{-n} \\ &\quad \times \Big\{\Big(1-\frac{2\varepsilon}{2j+1}\Big)^{-n}\Big(1+\frac{v}{k}\Big)^{-n}\Big(1+\frac{2(v-\varepsilon)}{2j+2k+1}\Big)^{-n} \\ &\qquad - \Big(1+\frac{2\varepsilon}{2j+1}\Big)^{-n}\Big(1-\frac{v}{k}\Big)^{-n}\Big(1-\frac{2(v-\varepsilon)}{2j+2k+1}\Big)^{-n}\Big\}. \end{aligned}$$

Here, the assumption that n is odd has been used to determine the signs of both terms in the above formulation. By Lemma 1 (a),(b),

$$|g_{jk}| \le |2j+1|^{-n}|k|^{-n+1}|2j+2k+1|^{-n}|g_{-1,1}|,$$

where $g_{-1,1} \le 0$ and $= 0$ only for $t = 0$. Hence, it follows that

$$g\Big(\frac{1}{2}\Big) \le g_{-1,1}[uv(u+v)]^n\Big\{1 - \sum_{\substack{j\in\mathbb{Z}\\ k\ge 1}} |2j+1|^{-n}|k|^{-n+1}|2j+2k+1|^{-n}\Big\}$$
$$\le 0.89 g_{-1,1}[uv(u+v)]^n.$$

If n is even, then $g_{-1,1} \ge 0$ and the corresponding term in the series of $[uv(u+v)]^{-n}g(0)$ is $\tilde{g}_{jk} := (-1)^j g_{jk}$. Since $\tilde{g}_{-1,1} \le 0$ and $= 0$ only for $t = 0$, the above analysis directly applies to showing $g(0) \le 0$.

To establish the claim that g is a monotone function on $[0, \frac{1}{2}]$ for fixed $(u, v) \in \triangle_2$ with $v \le 0.05$ and $t \neq 0$, we note that

$$g'(s) = -2\pi \sum_{j,k\neq 0} j \sin 2\pi js \sin 2\pi kt\Big[\frac{u}{u+j}\cdot\frac{v}{v+k}\cdot\frac{u+v}{u+v+j+k}\Big]^n.$$

In particular, the series is dominated by the $(-1, 1)$ term

$$g'_{-1,1} := -2\pi \sin 2\pi s \sin 2\pi t\Big[\frac{u}{u-1}\cdot\frac{v}{1+v}\Big]^n,$$

which is positive for odd n and negative for even n. Assuming $0 \le v \le 0.05, 0.45 \le u \le 0.5$, and $0.5 \le u+v \le 0.525$, Lemma 1 (c)–(e) is used as follows:

$$\begin{aligned}
&\Big| \sum_{\substack{j,k\neq 0\\ (j,k)\neq(-1,1)}} j \sin 2\pi js \sin 2\pi kt \Big[\frac{u}{u+j}\cdot\frac{v}{v+k}\cdot\frac{u+v}{u+v+j+k}\Big]^n \Big| \\
&\qquad \le g'_{-1,1} \sum_{(j,k)\neq(-1,1)} j^2|k| \Big|\frac{1-u}{u+j}\cdot\frac{1+v}{v+k}\cdot\frac{u+v}{u+v+j+k}\Big|^n \\
&\qquad \le g'_{-1,1} \sum_{(j,k)\neq(-1,1)} j^2|k| \Big|\frac{1-u}{u+j}\cdot\frac{1+v}{v+k}\cdot\frac{u+v}{u+v+j+k}\Big|^3 \\
&\qquad \le 0.45 g'_{-1,1}.
\end{aligned}$$

Hence, combining the estimates for T_1 and T_2 into (3.9), we have established (3.7) and consequently completed the proof of the lemma. □

To complete the proof of Theorem 1, the only remaining task is to establish the following.

LEMMA 4. *Let $n \ge 3$. Then for all $(u,v) \in \triangle_2$,*

$$Im\ e^{i2\pi t} Q_{n,(s,t)}(u,v) > 0 \quad \text{if } (s,t) \in \Lambda_5 \backslash \{0\}, \tag{3.12}$$

$$Im\ e^{i2\pi t} Q_{n,(s,t)}(u,v) < 0 \quad \text{if } (s,t) \in \Lambda_6 \backslash \{0\}. \tag{3.13}$$

Proof. The proof adapts the method in [9] (cf. Case IV in the proof there) applied to a different $\triangle_2$ as follows. Again, it is sufficient to establish (3.12). Let $\alpha = (s,t) \in \Lambda_5 \backslash \{0\}$. Then

$$\begin{aligned}
&Im\ e^{i2\pi t} Q_{n,(s,t)}(u,v) \\
&\qquad = \sum_{j,k\in\mathbb{Z}} \sin 2\pi(js + (k+1)t) \Big\{\frac{u}{u+j}\cdot\frac{v}{v+k}\cdot\frac{u+v}{u+v+j+k}\Big\}^n \\
&\qquad =: T' + T'',
\end{aligned} \tag{3.14}$$

where T' and T'' are defined as in [9] (see also (3.15)–(3.16) to follow). For $(u,v) \in \triangle_2$, $0 \le u \le \frac{1}{2}, 0 \le v \le \frac{1}{3}$, and $0 \le u+v < \frac{2}{3}$, it follows from [9, Eq. (4.25)] that T' is dominated by $\sin 2\pi t$, and hence is nonnegative and vanishes only when $t = 0$. For T'', we must modify the argument in [9] and rely on the restrictions $u \ge \frac{1}{4}, v \le \frac{1}{3}$ and $u+v \ge \frac{1}{2}$ in applying Lemma 1 (c)–(d). With the dominant term

$$t''_{-1,0} := \Big(\frac{u}{1-u}\cdot\frac{u+v}{1-u-v}\Big)^n \sin 2\pi(t-s) \ge 0, \tag{3.15}$$

we have

$$\begin{aligned}
T'' &:= t''_{-1,0}\Big[1 - \sum_{\substack{(j,k)\neq\\ (0,0),(-1,0)}} \Big(\frac{1-u}{u+j}\cdot\frac{v}{v+k}\cdot\frac{1-u-v}{u+v+j+k}\Big)^n \\
&\qquad \times \frac{\sin 2\pi j(t-s)}{\sin 2\pi(t-s)} \cos 2\pi jt \cos 2\pi(k+1)t\Big] \qquad (3.16)\\
&\ge t''_{-1,0}\Big[1 - \sum_{\substack{(j,k)\neq\\ (0,0),(-1,0)}} \Big|\frac{3}{1+4j}\cdot\frac{1}{1+3k}\cdot\frac{1}{1+2j+2k}\Big|^n |j|\Big] \\
&\ge 0.96 t''_{-1,0},
\end{aligned}$$

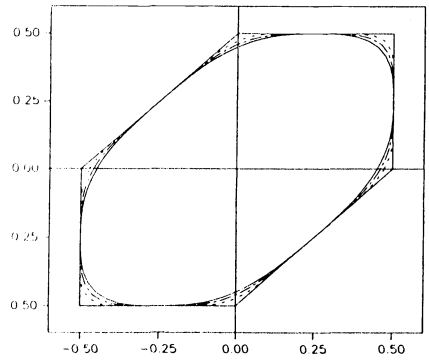


FIG. 3

where the numerical lower bound is obtained by using $n = 3$. In view of (3.15), $T'' \geq 0$ and vanishes only for $s = t$. Combining this with the estimate for T', we conclude that $Im\, e^{i2\pi t} Q_{n,(s,t)}(u,v)$ is strictly positive for all $(u,v) \in \triangle_2$ and all shift parameters $(s,t) \in \triangle_5 \backslash \{0\}$. This completes the proof of Lemma 4, and hence, of Theorem 1. □

4. The metric condition. The purpose of establishing $\tilde{\phi}(x) \neq 0$ for all $x \in \mathbb{R}^s$ is to guarantee that cardinal interpolation by using $\phi(\cdot - j), j \in \mathbb{Z}^s$ is poised. If $\tilde{\phi}$ satisfies the metric condition (1.3), then not only $\tilde{\phi}(x) \neq 0$ for all $x \in \mathbb{R}^s$, but the cardinal interpolants can also be constructed by using the Neumann series [5]. For B-splines B_k in one variable, it was shown in [10] that $|1 - \widetilde{B}_{k,\alpha}(x)| < 1$ for all x and $|\alpha| < \frac{1}{2}$, where $B_{k,\alpha} := B_k(\cdot + \alpha)$; hence, the important univariate notion of total positivity is not required in establishing invertibility of the cardinal interpolation operators corresponding to the shifted B-splines. Unfortunately, in the case of bivariate splines on the three-direction mesh considered in this paper, the metric condition

$$|1 - \widetilde{M}_{n,\alpha}(x)| < 1, \qquad x \in \mathbb{R}^2, \tag{4.1}$$

is no longer equivalent to the condition

$$\widetilde{M}_{n,\alpha}(x) \neq 0, \qquad x \in \mathbb{R}^2.$$

Let Λ^n denote the largest subregion of Λ on which the metric condition (4.1) holds. In Fig. 3, we have plotted subregions $\Lambda^1, \Lambda^2, \Lambda^3, \Lambda^4$ of the shift region Λ. Observe that $\Lambda = \Lambda^1 \supset \Lambda^2 \supset \Lambda^3 \supset \Lambda^4$, where all the containments are proper. However, at this writing, we do not know if $\{\Lambda^n\}$ is a monotone sequence. It would also be interesting to know its "limit" or "intersection."

Acknowledgment. The authors thank N. Sivakumar for several stimulating conversations.

REFERENCES

[1] C. DE BOOR, K. HÖLLIG, AND S. D. RIEMENSCHNEIDER, *Bivariate cardinal interpolation by splines on a three direction mesh*, Illinois J. Math., 29 (1985), pp. 533–566.

[2] C. DE BOOR AND I. J. SCHOENBERG, *Cardinal interpolation and spline functions* VIII: *The Budan-Fourier theorem for splines and applications,* Lecture Notes in Mathematics 501, Springer-Verlag, 1976.

[3] C. K. CHUI, *Multivariate Splines,* CBMS-NSF Regional Conference Series in Applied Math. 54, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1988.

[4] C. K. CHUI AND H. DIAMOND, *A natural formulation of quasi-interpolation by multivariate splines,* Proc. Amer. Math. Soc., 99 (1987), pp. 643–646.

[5] C. K. CHUI, H. DIAMOND, AND L. A. RAPHAEL, *Interpolation by multivariate splines,* Math. Comp., 51 (1988), pp. 203–218.

[6] C. K. CHUI, K. JETTER, AND J. D. WARD, *Cardinal interpolation by multivariate splines,* Math. Comp., 48 (1987), pp. 711–724.

[7] C. A. MICCHELLI, *Cardinal L-splines,* in Studies in Spline Functions and Approximation Theory, Karlin, Micchelli, Pinkus, and Schoenberg, eds., Academic Press, New York, 1976, pp. 163–202.

[8] N. SIVAKUMAR, On univariate cardinal interpolation by shifted splines, Rocky Mountain J. Math., 19 (1989), pp. 481–489.

[9] ———, *On bivariate cardinal interpolation by shifted splines on a three-direction mesh,* J. Approx. Theory, 61 (1990), pp. 178–193.

[10] P. W. SMITH AND J. D. WARD, *Quasi-interpolants from spline interpolation operators,* Constr. Appr., 6 (1990), pp. 97–110.

[11] E. M. STEIN AND G. WEISS, *Introduction to Fourier Analysis on Euclidean Spaces,* Princeton University Press, Princeton, NJ, 1971.

[12] J. STÖCKLER, *Cardinal interpolation with translates of shifted bivariate box-splines,* in Mathematical Methods in Computer Aided Geometric Design, T. Lyche and L. L. Schumaker, eds., Academic Press, New York, 1989, pp. 583–592.

SIAM J. MATH. ANAL.
Vol. 22, No. 2, pp. 554-572, March 1991

016

A SIMPLE WILSON ORTHONORMAL BASIS WITH EXPONENTIAL DECAY*

INGRID DAUBECHIES†, STÉPHANE JAFFARD‡, AND JEAN-LIN JOURNÉ§

Abstract. Following a basic idea of Wilson ["Generalized Wannier functions," preprint] orthonormal bases for $L^2(\mathbb{R})$ which are a variation on the Gabor scheme are constructed. More precisely, $\phi \in L^2(\mathbb{R})$ is constructed such that the ψ_{ln}, $l \in \mathbb{N}$, $n \in \mathbb{Z}$, defined by

$$\psi_{0n}(x) = \phi(x-n)$$

$$\psi_{ln}(x) = \sqrt{2}\,\phi\left(x-\frac{n}{2}\right)\cos(2\pi l x) \quad \text{if } l \neq 0,\ l+n \in 2\mathbb{Z}$$

$$= \sqrt{2}\,\phi\left(x-\frac{n}{2}\right)\sin(2\pi l x) \quad \text{if } l \neq 0,\ l+n \in 2\mathbb{Z}+1,$$

constitute an orthonormal basis. Explicit examples are given in which both ϕ and its Fourier transform $\hat{\phi}$ have exponential decay. In the examples ϕ is constructed as an infinite superposition of modulated Gaussians, with coefficients that decrease exponentially fast. It is believed that such orthonormal bases could be useful in many contexts where lattices of modulated Gaussian functions are now used.

Key words. orthonormal bases, phase space localization, time-frequency analysis

AMS(MOS) subject classifications. 46C10, 81C40, 94A11

1. Introduction.

In several applications in quantum mechanics and in signal analysis, sets of functions generated from one single function by phase space translations are encountered:

$$g_{mn}(x) = e^{2\pi i \alpha m x} g(x-\beta n), \qquad m, n \in \mathbb{Z}. \tag{1.1}$$

If the function g and its Fourier transform $\hat{g}$,

$$\hat{g}(\xi) = \int dx\, e^{2\pi i x \xi} g(x),$$

are both centered around zero, then the function g_{mn} is centered around the phase space point $(\alpha m, \beta n)$. We can then hope to use the functions g_{mn} for expansions of functions with good phase space localization. More concretely, we would like expansions of the type

$$f = \sum_{m,n} c_{mn}(f) g_{mn}, \tag{1.2}$$

with the property that the $c_{mn}(f)$ are nonnegligible only for those values of (m, n) associated to phase space points where f is nonnegligible. For example, if $\int_{|t|\geq T} dt\, |f(t)|^2 \leqq \varepsilon \|f\|^2$ and $\int_{|\xi|\geq \Omega} d\xi\, |\hat{f}(\xi)|^2 = \varepsilon \|f\|^2$, then we would prefer most of the "content" of f to be concentrated in the $c_{mn}(f)$ with $(m\alpha, n\beta)$ within or close to the rectangle $[-\Omega, \Omega] \times [-T, T]$. More concretely, this can be translated into the

* Received by the editors September 5, 1989; accepted for publication (in revised form) March 13, 1990.

† AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974. This author is "Bevoegdverklaand Navonsen" at the Belgian National Science Foundation (on leave); also on leave from the Department of Theoretical Physics, Vrije Universiteit, Brussels, Belgium.

‡ Centre d'Etude et de Recherche en Mathématique Appliquée, Ecole des Ponts et Chaussées, La Courtine, 93167 Noisy-le-Grand, France.

§ Princeton University, Princeton, New Jersey 08544. This author's work was supported by the National Science Foundation.

requirement

$$\sum_{\substack{|\alpha m|\geqq T+\Delta T \\ |\beta n|\geqq \Omega+\Delta\Omega}} |c_{mn}(f)|^2 \leqq C\varepsilon \|f\|^2, \tag{1.3}$$

where C should be independent of ε, T, and Ω, and where ΔT, $\Delta\Omega$ should only depend on the desired precision ε.

One example of a set of functions of type (1.1) are the phase space Wannier functions used in solid state physics. In the absence of a potential they are obtained by the choice

$$g(x) = \frac{\sin \pi x}{\pi x},$$

which corresponds to

$$\hat{g}(\xi) = \begin{cases} 1, & |\xi| < \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

(When the potential is nonzero, the Wannier functions are more complicated [1].) For this choice of g, and for the parameter choice $\alpha = \beta = 1$, the functions (1.1) constitute an orthonormal basis of $L^2(\mathbb{R})$. Expansions of type (1.2) are therefore simple to obtain: it suffices to take $c_{mn}(f) = \int dx\, \overline{g_{mn}(x)}\, f(x)$. Unfortunately, the localization of g is not very good. The function g has a rather long tail, so that

$$\int dx\, x^2 |g(x)|^2 = \infty.$$

As a consequence of this, expansions of functions with respect to the phase space Wannier functions do not have the good phase space localization features described above.

Another example of a set of functions of type (1.1) is given by the "Gabor expansions." These correspond to the choice

$$g(x) = 2^{1/4} \exp(-\pi x^2).$$

In the original proposal of Gabor [2], the parameter choice $\alpha = \beta = 1$ is made. Unfortunately, this parameter choice leads to numerically unstable expansions: for any $\varepsilon > 0$, there exists $f \in L^2(\mathbb{R})$ such that $\|f\| = 1$ but $\sum_{m,n} |c_{mn}(f)|^2 \leqq \varepsilon$. It can be shown that this phenomenon happens for any choice of α, β such that $\alpha\beta = 1$ [3a, b], [4a, b]. If $\alpha\beta > 1$, then the g_{mn} do not span all of $L^2(\mathbb{R})$ [5], [6]. If $\alpha\beta < 0.996$, then numerically stable expansions of type (1.2) do exist, with the "good" phase space localization described by (1.3) (see [7], [8]; it is conjectured that this situation persists for $\alpha\beta < 1$). There is, however, a price to pay: for $\alpha\beta < 1$, the g_{mn} are highly redundant, in the sense that any finite number of them lies in the closed linear span of all the others. While Gabor expansions with $\alpha\beta < 1$ are indeed used in practical computations in atomic and nuclear physics, this redundancy can be quite a nuisance.

These two examples illustrate how convenient it would be to have a nice orthonormal basis ($\rightarrow$ no redundancy) of type (1.1), based on a function g such that both g and $\hat{g}$ have good decay properties ($\rightarrow$ expansions with good phase space localization). Unfortunately, such an orthonormal basis does not exist. A theorem stated by Balian [9] and Low [10] asserts that a set of functions of type (1.1) can only constitute an orthonormal basis if either $\int dx\, x^2 |g(x)|^2 = \infty$ or $\int d\xi\, \xi^2 |\hat{g}(\xi)|^2 = \infty$. Balian's and Low's proofs contain a technical gap that was filled by Coifman and Semmes, as reported in [7]; a much simpler proof was subsequently found by Battle [11]. Even if the orthonormality, but not the "basis" requirement, is given up, the same conclusion still holds, as shown by the extension of Battle's argument in [12]. Both the original proof

and Battle's proof of the Balian-Low theorem rely heavily on the special structure of the g_{mn} as defined by (1.1). We might therefore wonder whether there exist more general bases, $\psi_{mn}(x)$, with phase space localizations distributed more or less regularly over phase space, and such that uniform bounds on the decay of all the ψ_{mn} and $(\psi_{mn})\hat{}$, away from their central value, would hold. It turns out that there is indeed improvement from giving up the simplicity of (1.1), but only very little. Bourgain [13] has constructed an orthonormal basis of ψ_{mn} such that

$$\begin{aligned} &\int dx\ (x-\bar{x}_{mn})^2|\psi_{mn}(x)|^2 \leqq C, \\ &\int d\xi\ (\xi-\bar{\xi}_{mn})^2|(\psi_{mn})\hat{}(\xi)|^2 \leqq C, \end{aligned} \tag{1.4}$$

uniformly in m, n, where $\bar{x}_{mn} = \int dx\, x|\psi_{mn}(x)|^2$, and $\bar{\xi}_{mn}$ is defined analogously. However, as soon as a slightly sharper localization is required, we hit another no-go-theorem, even for these more general constructions: Steger [14] proved that $L^2(\mathbb{R})$ does not admit an orthonormal basis ψ_{mn} satisfying

$$\begin{aligned} &\int dx\ (x-\bar{x}_{mn})^{2(1+\varepsilon)}|\psi_{mn}(x)|^2 \leqq C, \\ &\int d\xi\ (\xi-\bar{\xi}_{mn})^{2(1+\varepsilon)}|(\psi_{mn})\hat{}(\xi)|^2 \leqq C. \end{aligned} \tag{1.5}$$

Orthonormality, or, what is weaker, the existence of numerically stable expansions of type (1.2) with nonredundant functions ψ_{mn}, is therefore incompatible with good phase space localization.

In all the above, "good phase space localization" stands for strong decay properties of the ψ_{mn}, $(\psi_{mn})\hat{}$ away from the average values $\bar{x}_{mn}$, $\bar{\xi}_{mn}$. This corresponds to a picture in which both ψ_{mn} and $(\psi_{mn})\hat{}$ have essentially *one* peak. In [15] Wilson proposes instead to construct orthonormal bases ψ_{mn} of the type

$$\psi_{mn}(x) = f_m(x-n), \qquad m \in \mathbb{N}, \quad n \in \mathbb{Z}, \tag{1.6}$$

where $\hat{f}_m$ has *two* peaks, situated near $m/2$ and $-m/2$,

$$\hat{f}_m(\xi) = \phi_m^+\left(\xi - \frac{m}{2}\right) + \phi_m^-\left(\xi + \frac{m}{2}\right), \tag{1.7}$$

with ϕ_m^+, ϕ_m^- centered around zero. He proposes numerical evidence for the existence of such an orthonormal basis, with uniform exponential decay for f_m and ϕ_m^+, ϕ_m^-. In his numerical construction he further "optimizes" the localization by requiring

$$\int d\xi\, \xi^2 \overline{(\psi_{mn})\hat{}(\xi)}\, \psi_{m'n'}(\xi) = 0 \quad \begin{array}{l} \text{if } |m-m'|>1, \\ \text{or if } |m-m'|=1,\ |n-n'|>1. \end{array} \tag{1.8}$$

In [16] Sullivan et al. present arguments explaining both the existence of Wilson's basis and its exponential decay. In both [15] and [16] there are infinitely many functions $\phi_m^{\pm}$; as m tends to ∞, the $\phi_m^{\pm}$ tend to a limit function $\phi_\infty^{\pm}$.

The moral of Wilson's construction is that orthonormal bases with good phase space localization are possible after all if bimodal functions as in (1.7) are used. This is reminiscent of what happens for orthonormal wavelet bases, i.e., orthonormal bases of $L^2(\mathbb{R})$ of the type

$$h_{mn}(x) = 2^{-m/2} h(2^{-m}x - n), \qquad m, n \in \mathbb{Z}. \tag{1.9}$$

There exist functions h with excellent phase space localization properties such that the functions (1.9) constitute an orthonormal basis. In [17] Meyer constructs such a function h with compactly supported, C^∞ Fourier transform $\hat{h}$; [18]–[20] give examples of exponentially decaying $h \in C^k$; and [21] constructs compactly supported $h \in C^k$. In all these examples, $|\hat{h}|$ has two peaks, one for positive and one for negative frequencies. It has been shown [22] that these two peaks need not be symmetrical in order for the h_{mn} to constitute an orthonormal basis (the examples in [17]–[21] all have symmetric peaks for $|\hat{h}|$). However, there is no example, so far, of reasonably well-localized functions $h^\pm$ such that support $(h^\pm) \subset \mathbb{R}_\pm$ and such that the $h^\pm_{mn}$ constitute an orthonormal basis of $L^2(\mathbb{R})$, corresponding to wavelet bases with only one "peak" in frequency. (Equivalently, there is no example of a reasonably smooth function $\phi = \hat{h}^+$ such that the functions $2^{m/2} \exp(2\pi i 2^m n\xi)\phi(2^m\xi)$ are an orthonormal basis of $L^2(\mathbb{R}_+)$.) It is believed, without proof so far, that no such basis exists. This seems to be the analogue, for the wavelet situation, of the Balian–Low theorem.

In this paper we construct an explicit bimodal orthonormal basis of the type (1.6), (1.7). Our basis is especially simple because it is again generated by one single function, unlike the bases in [15], [16]. More explicitly, we construct a real function ϕ such that with the definitions

$$\hat{f}_1(\xi) = \phi(\xi),$$

$$\hat{f}_{2\ell+\kappa}(\xi) = \frac{1}{\sqrt{2}}[\phi(\xi-\ell) + (-1)^{\ell+\kappa}\phi(\xi+\ell)]\, e^{i\pi\kappa\xi}, \qquad \ell \in \mathbb{N}\setminus\{0\}, \quad \kappa = 0 \text{ or } 1, \tag{1.9a}$$

the family

$$\psi_{mn}(x) = f_m(x-n), \qquad m \in \mathbb{N}\setminus\{0\}, \quad n \in \mathbb{Z} \tag{1.9b}$$

constitutes an orthonormal basis. Both ϕ and its Fourier transform $\hat{\phi}$ have exponential decay. Moreover, ϕ can be explicitly constructed as a rapidly converging superposition of Gaussians. All these features should make the basis constructed here especially attractive for the computations in atomic and nuclear physics where the Gabor functions are now used. The price we pay for the simplicity of our Wilson basis is that the near-diagonalization (1.8) of ξ^2 no longer holds.

This paper is organized as follows. In § 2 we derive necessary and sufficient conditions on ϕ for the ψ_{mn}, defined by (1.9), to be an orthonormal basis of $L^2(\mathbb{R})$. In § 3 we rewrite these conditions in another form, via the Zak transform. In their new form, it is easy to see how to satisfy these conditions. We use this in § 4 to construct an explicit Wilson basis with all the properties mentioned above. It turns out that our construction is related to "tight frames" [23], [7]. We review this concept in § 5, and explain how it is linked to the present construction. This leads to an alternate construction method, given in § 6, which is easier to implement numerically. Finally § 7 gives some concluding remarks. In particular, we show how a relabelling of the ψ_{mn} in (1.9) reduces the construction to the formula given in the Abstract.

2. Necessary and sufficient conditions. It suffices to prove that

$$\|\psi_{mn}\| = 1, \qquad m \in \mathbb{N}\setminus\{0\}, \quad n \in \mathbb{Z}, \tag{2.1}$$

and

$$\sum_{m=1}^{\infty} \sum_{n=-\infty}^{\infty} \langle g, \psi_{mn}\rangle\langle\psi_{mn}, h\rangle = \langle g, h\rangle \tag{2.2}$$

for all $g, h \in L^2(\mathbb{R})$. Indeed, from (2.1) and (2.2) we obtain

$$1 = \|\psi_{m'n'}\|^2 = \sum_{m,n} |\langle \psi_{m'n'}, \psi_{mn}\rangle|^2$$

$$= 1 + \sum_{(m,n)\neq(m',n')} |\langle \psi_{m'n'}, \psi_{mn}\rangle|^2,$$

whence $\langle \psi_{mn}, \psi_{m'n'}\rangle = \delta_{mm'}\delta_{nn'}$. It follows that (2.1) and (2.2) imply that ψ_{mn} constitute a total orthonormal set.

We first concentrate on (2.2). Using Parseval's identity and the Poisson summation formula, we find

$$\sum_{m=1}^{\infty}\sum_{n=-\infty}^{\infty} \langle g, \psi_{mn}\rangle\langle \psi_{mn}, h\rangle = \sum_{m=1}^{\infty}\sum_{k=-\infty}^{\infty} \int d\xi\, \overline{\hat{g}(\xi)}\, \hat{h}(\xi+k)\hat{f}_m(\xi)\overline{\hat{f}_m(\xi+k)}.$$

(Note that we use the physicist's convention for the inner product in $L^2(\mathbb{R})$, which is linear in the *second* argument, $\langle g, h\rangle = \int dx\, \overline{g(x)}h(x)$.) In order to have (2.2), it is therefore necessary and sufficient that

$$\sum_{m=1}^{\infty} \hat{f}_m(\xi)\overline{\hat{f}_m(\xi+k)} = \delta_{k0}. \tag{2.3}$$

Let us write this out in terms of ϕ. For the time being, we disregard any convergence questions; for the function we will construct all series converge absolutely and uniformly. We also assume ϕ to be real.

$$\sum_{m=1}^{\infty} \hat{f}_m(\xi)\overline{\hat{f}_m(\xi+k)}$$

$$= \phi(\xi)\phi(\xi+k) + \frac{1}{2}\sum_{\ell=1}^{\infty}\sum_{\kappa=0}^{1} [\phi(\xi-\ell) + (-1)^{\ell+\kappa}\phi(\xi+\ell)]$$

$$\cdot [\phi(\xi\,\xi\varsigma\epsilon'' \eta\!\!\!/\, \ell + k) + (\overset{\ell+}{-}1) \quad \phi(\xi+\ell+k)]^{i\pi\kappa k} \tag{2.4}$$

$$= \phi(\xi)\phi(\xi+k) + \sum_{\ell\in\mathbb{Z},\ell\neq 0} \phi(\xi+\ell)\phi(\xi+\ell+k)\frac{1}{2}(1+(-1)^k)$$

$$+ \sum_{\ell\in\mathbb{Z},\ell\neq 0} (-1)^{\ell}\phi(\xi+\ell)\phi(\xi-\ell+k)\frac{1}{2}(1-(-1)^k).$$

If k is even, $k = 2j$, then

$$(2.4) = \sum_{\ell\in\mathbb{Z}} \phi(\xi+\ell)\phi(\xi+\ell+2j).$$

If k is odd, $k = 2j+1$, then

$$(2.4) = \sum_{\ell\in\mathbb{Z}} (-1)^{\ell}\phi(\xi+\ell)\phi(\xi-\ell+2j+1) = 0,$$

as is easily shown by the change of summation index $\ell' = -\ell + 2j + 1$. It follows that (2.3) is equivalent to

$$\sum_{\ell\in\mathbb{Z}} \phi(\xi+\ell)\phi(\xi+\ell+2j) = \delta_{j0}. \tag{2.5}$$

We now turn to (2.1). It clearly suffices to prove $\|\hat{f}_m\| = 1$, $m \in \mathbb{N}\setminus\{0\}$. For $m = 1$ this gives

$$\int d\xi\, |\phi(\xi)|^2 = 1.$$

For $m=2\ell+\sigma$, $\ell\geqq 1$, we find

$$\|\hat{f}_m\|^2=\frac{1}{2}\int d\xi\,|\phi(\xi-\ell)+(-1)^{\ell+\sigma}\phi(\xi+\ell)|^2$$

$$=1+(-1)^{\ell+\sigma}\int d\xi\,\phi(\xi-\ell)\phi(\xi+\ell).$$

It follows that (2.1) is equivalent to

$$\int d\xi\,\phi(\xi)\phi(\xi+2\ell)=\delta_{\ell 0}. \tag{2.6}$$

This condition is automatically satisfied if (2.5) holds:

$$\int_{-\infty}^{\infty} d\xi\,\phi(\xi)\phi(\xi+2\ell)=\sum_{k\in\mathbb{Z}}\int_0^1 d\xi\,\phi(\xi+k)\phi(\xi+k+2\ell)$$

$$=\int_0^1 d\xi\,\delta_{\ell 0}=\delta_{\ell 0}.$$

We have again assumed that ϕ is sufficiently well behaved so that the summation and integration may be commuted in this computation. It is easily checked that it is sufficient that ϕ decays faster than $|\xi|^{-1}$ for $|\xi|\to\infty$. For the examples we will construct, this is no problem. The following proposition summarizes our findings.

PROPOSITION 2.1. *Suppose that ϕ is a real function on $\mathbb{R}$ satisfying*

$$|\phi(\xi)|\leqq C(1+|\xi|)^{-1-\varepsilon}$$

for some C, $\varepsilon>0$. Then the functions ψ_{mn} defined by (1.9) *constitute an orthonormal basis for $L^2(\mathbb{R})$ if and only if*

$$\sum_{\ell\in\mathbb{Z}}\phi(\xi+\ell)\phi(\xi+\ell+2j)=\delta_{j0}.$$

We therefore have only one set of conditions, namely (2.5). This condition can be almost trivially satisfied if we choose ϕ to be supported in $[-1,1]$. In this case $\phi(\xi)\phi(\xi+2\ell)=0$ if $\ell\neq 0$, for any $\xi\in\mathbb{R}$. It follows that (2.5) is satisfied if $\sum_\ell \phi(\xi+\ell)^2=1$. Since this sum is periodic in ξ with period 1, we only need to check what happens for $0\leqq\xi\leqq 1$. For ϕ supported in $[-1,1]$, this means we only need to ascertain that $\phi(\xi)^2+\phi(\xi-1)^2=1$ for $0\leqq\xi\leqq 1$. Such ϕ are easy to construct: for any function F such that

$$F:\mathbb{R}\to\mathbb{R}$$

$$F(x)=\begin{cases}0, & x\leqq 0,\\ 1, & x\geqq 1,\end{cases}$$

$$0\leqq F(x)\leqq 1 \quad \text{for all } x,$$

the function ϕ defined by

$$\phi(\xi)=\begin{cases}\sin\left[\frac{\pi}{2}F(\xi+1)\right], & \xi\leqq 0,\\ \cos\left[\frac{\pi}{2}F(\xi)\right], & \xi\geqq 0\end{cases}$$

is a function supported in $[-1, 1]$ which satisfies (2.5), hence (2.1) and (2.2). If F is C^k (where k may be ∞), then ϕ is C^k. The corresponding ψ_{mn} are C^∞-functions; their decay at ∞ is regulated by the regularity of F. If F is C^∞, then the ψ_{mn} have "fast decay," i.e., for all $N \in \mathbb{N}$, there exists C_N such that

$$|\psi_{mn}(x)| \leqq C_N(1+|x-n|^2)^{-N}.$$

In practice, however, the constants C_N turn out to be rather large, so that the *numerical* localization of the ψ_{mn} is not very good. The examples we construct in § 4, corresponding to noncompactly supported ϕ, have better effective localization.

3. The Zak transform—rewriting the conditions. Using a unitary transformation, we will rewrite the infinitely many conditions (2.5) (one for every j) into a different form, reducing them to one single condition which is then easy to satisfy. The unitary map we shall use is the Zak transform. For the purposes of this paper, we define the Zak transform by

$$(U_Z g)(t, s) = \sqrt{2} \sum_{k \in \mathbb{Z}} e^{2\pi i t k} g(2(s-k)). \tag{3.1}$$

This is well defined for functions g with sufficient decay, $|g(x)| \leqq C(1+|x|^2)^{-1/2-\varepsilon}$. The two-variable function $G = U_Z g$ is periodic in the first and "semi-periodic" in the second variable,

$$\begin{aligned} G(t+1, s) &= G(t, s), \\ G(t, s+1) &= e^{2\pi i t} G(t, s). \end{aligned} \tag{3.2}$$

The set of all functions G of two variables satisfying the periodicity conditions (3.2) can be equipped with the norm

$$\|G\|^2 = \int_0^1 dt \int_0^1 ds\, |G(t, s)|^2. \tag{3.3}$$

We will denote the closure of this set, under the norm (3.3), by $\mathcal{Z}$. A function G is in $\mathcal{Z}$ if and only if its restriction to $[0, 1[\times [0, 1[$ is square integrable, and it satisfies the periodicity conditions (3.2) almost everywhere. It follows that $\mathcal{Z}$ is isomorphic with $L^2([0, 1]^2)$. The functions $E_{mn}(t, s)$, defined by

$$E_{mn}(t, s) = e^{2\pi i n t}\, e^{2\pi i m s} \quad \text{for } t, s \in [0, 1[,$$

extended by (3.2) to all of $\mathbb{R}^2$, constitute an orthonormal basis for $\mathcal{Z}$.

The map U_Z defined by (3.1) can be extended to a unitary map from $L^2(\mathbb{R})$ to $\mathcal{Z}$. This follows from the fact that U_Z maps the orthonormal basis $e_{mn}(x) = e^{\pi i m x} \chi(x-2n)$, where $\chi(x) = 2^{-1/2}$ if $0 \leqq x < 2$, $\chi(x) = 0$ otherwise, to the orthonormal basis E_{mn} of $\mathcal{Z}$, $U_Z e_{mn} = E_{mn}$.

The Zak transform has many interesting properties; it derives its name from its systematic study by J. Zak, who introduced it as a tool in solid state physics [24a-c]. It had already been studied sporadically before Zak's work, and it is claimed that even Gauss was already aware of some of its properties. An excellent review of the mathematical properties of U_Z and its applications to signal analysis is Janssen's paper [25], which also contains an extensive reference list.

The inverse transform of (3.1) is given by

$$(U_Z^{-1} G)(x) = \frac{1}{\sqrt{2}} \int_0^1 dt\, G\left(t, \frac{x}{2}\right). \tag{3.4}$$

Again this is only well defined for some G in $\mathscr{Z}$ (including all bounded G), but it can be extended to all of $\mathscr{Z}$.

There exists a relationship between $U_Z g$ and $U_Z \hat{g}$. Using the Poisson summation formula, we find

$$
\begin{aligned}
(U_Z\hat{g})(t,s) &= \sqrt{2}\sum_{\ell\in\mathbb{Z}} e^{2\pi i t\ell}\int_{-\infty}^{\infty} dx\, e^{4\pi i(s-\ell)x} g(x) \\
&= \frac{1}{\sqrt{2}} e^{2\pi i s t}\sum_{k\in\mathbb{Z}} e^{-2\pi i s k} g\left(\frac{t-k}{2}\right) \qquad \text{(3.5a)} \\
&= \frac{1}{2}\sum_{j=0}^{3} e^{2\pi i s(t+j)}(U_Z g)\left(-4s, \frac{t+j}{4}\right).
\end{aligned}
$$

Similarly,

$$
(U_Z g)(t,s) = \frac{1}{2}\sum_{j=0}^{3} e^{2\pi i s(t+j)}(U_Z \hat{g})\left(4s, -\frac{t+j}{4}\right). \qquad \text{(3.5b)}
$$

Let us now apply all this to the problem at hand. We define $\Phi = U_Z\phi$, and we rewrite (2.5) in terms of Φ. We have

$$
\begin{aligned}
\sum_{\ell\in\mathbb{Z}} \phi(\xi+\ell)\phi(\xi+\ell+2j) &= \frac{1}{2}\sum_{\ell\in\mathbb{Z}}\int_0^1 dt\int_0^1 dt'\,\Phi\left(t,\frac{\xi+\ell}{2}\right)\Phi\left(t',\frac{\xi+\ell}{2}+j\right) \\
&= \frac{1}{2}\sum_{k\in\mathbb{Z}}\int_0^1 dt\int_0^1 dt'\left[\Phi\left(t,\frac{\xi}{2}+k\right)\Phi\left(t',\frac{\xi}{2}+k+j\right)\right. \\
&\qquad \left. +\Phi\left(t,\frac{\xi+1}{2}+k\right)\Phi\left(t',\frac{\xi+1}{2}+k+j\right)\right] \\
&= \frac{1}{2}\sum_{k\in\mathbb{Z}}\int_0^1 dt\int_0^1 dt'\, e^{2\pi i k(t+t')}\, d^{2\pi i j t'} \\
&\qquad \cdot\left[\Phi\left(t,\frac{\xi}{2}\right)\Phi\left(t',\frac{\xi}{2}\right)+\Phi\left(t,\frac{\xi+1}{2}\right)\Phi\left(t',\frac{\xi+1}{2}\right)\right] \\
&\qquad\qquad \text{(use (3.2))} \\
&= \frac{1}{2}\int_0^1 dt\, e^{-2\pi i j t}\left[\left|\Phi\left(t,\frac{\xi}{2}\right)\right|^2+\left|\Phi\left(t,\frac{\xi+1}{2}\right)\right|^2\right].
\end{aligned}
$$

In the last step we have assumed that $\Phi(\cdot, s)$ is square integrable for all s; by the definition (3.1) of the Zak transform we easily check that this is equivalent to the requirement that $\sum_k |\phi(2s-2k)|^2$ be bounded for all s, which is certainly true if, as in Proposition 2.1, ϕ decays faster than $|\xi|^{-1}$. Note that we have used $\Phi(-t, s) = \overline{\Phi(t, s)}$, which is true for real functions ϕ. All this proves the following proposition.

PROPOSITION 3.1. *Let ϕ be as in Proposition* 2.1. *Then* (2.5) *is satisfied, i.e.,*

$$
\sum_{\ell\in\mathbb{Z}} \phi(\xi+\ell)\phi(\xi+\ell+2j) = \delta_{j0}
$$

if and only if the Zak transform $\Phi = U_Z\phi$ of ϕ, as defined by (3.1) *satisfies*

$$
|\Phi(t,s)|^2 + |\Phi(t, s+\tfrac{1}{2})|^2 = 2 \qquad \text{(3.6)}
$$

for almost all $t, s \in [0,1]^2$.

4. Constructing solutions. Now that we have reduced the infinitely many conditions (2.5) to the single condition (3.6), we can get down to the business of constructing explicit "nice" ϕ satisfying (2.5). Typically, we start with a real function g with exponential decay,

$$|g(x)| \leqq C e^{-\lambda|x|}, \tag{4.1}$$

such that its Fourier transform has exponential decay as well,

$$|\hat{g}(y)| \leqq C e^{-\mu|y|}. \tag{4.2}$$

Define $G = U_Z g$; G is well defined and continuous. Since g is real, we have, for all $t, s \in \mathbb{R}$,

$$G(-t, s) = \overline{G(t, s)}. \tag{4.3}$$

Assume that

$$\inf_{t,s\in[0,1]} [|G(t, s)|^2 + |G(t, s+\tfrac{1}{2})|^2] > 0. \tag{4.4}$$

We then define

$$\phi = U_Z^{-1}\Phi, \tag{4.5}$$

where

$$\Phi(t, s) = \sqrt{2}\,\frac{G(t, s)}{[|G(t, s)|^2 + |G(t, s+\tfrac{1}{2})|^2]^{1/2}}. \tag{4.6}$$

Then the following theorem holds.

THEOREM 4.1. *The function ϕ, defined by* (4.5), *is a real function, and satisfies* (2.5). *Furthermore, both ϕ and $\hat{\phi}$ have exponential decay.*

Proof. 1. It follows from (4.3) and (4.6) that

$$\Phi(-t, s) = \overline{\Phi(t, s)},$$

so that, using (3.4) and (3.2),

$$\begin{aligned}\overline{\phi(x)} &= \frac{1}{\sqrt{2}}\int_0^1 dt\, \overline{\Phi\left(t, \frac{x}{2}\right)} = \frac{1}{\sqrt{2}}\int_0^1 dt\, \Phi\left(-t, \frac{x}{2}\right) \\ &= \frac{1}{\sqrt{2}}\int_{-1}^0 dt\, \Phi\left(t, \frac{x}{2}\right) = \frac{1}{\sqrt{2}}\int_0^1 dt\, \Phi\left(t, \frac{x}{2}\right) = \phi(x).\end{aligned}$$

2. To prove that ϕ has exponential decay, we first extend the definition domain of G from $\mathbb{R}^2$ to $(\mathbb{R} + i(-\lambda/\pi, \infty)) \times \mathbb{R}$. From (4.1) we see that the series

$$G(t + i\tau, s) = \sqrt{2} \sum_{\ell\in\mathbb{Z}} e^{2\pi i(t+i\tau)\ell} g(2(s-l)) \tag{4.7}$$

converges absolutely for $\tau > -\lambda/\pi$. The function $G(z, s)$ is continuous on $(\mathbb{R} + i(-\lambda/\pi, \infty)) \times \mathbb{R}$, and $G(\cdot, s)$ is analytic on $\mathbb{R} + i(-\lambda/\pi, \infty)$ for every $s \in \mathbb{R}$. Moreover,

$$\begin{aligned} G(z, s+1) &= e^{2\pi i z} G(z, s), \\ G(z+1, s) &= G(z, s). \end{aligned} \tag{4.8}$$

We also define, for $z \in \mathbb{R} + i(-\lambda/\pi, \infty)$, $s \in \mathbb{R}$

$$\mathcal{G}(z, s) = G(z, s)G(-z, s) + G(z, s+\tfrac{1}{2})G(-z, s+\tfrac{1}{2}). \tag{4.9}$$

Then $\mathcal{G}(\cdot, s)$ is analytic on $\mathbb{R} + i(-\lambda/\pi, \infty)$ for every $s \in \mathbb{R}$, and

$$\mathcal{G}(z+1, s) = \mathcal{G}(z, s) = \mathcal{G}(z, s+\tfrac{1}{2}) \tag{4.10}$$

for all $z \in \mathbb{R} + i(-\lambda/\pi, \infty)$, $s \in \mathbb{R}$. Using (4.2), we can show that G is uniformly continuous on $(\mathbb{R} + i[-\lambda/\pi, \infty)) \times [0, 1]$; together with (4.10) this implies that $\mathcal{G}$ is uniformly continuous on $(\mathbb{R} + i[-\lambda/\pi, \infty)) \times \mathbb{R}$. On the other hand, the restriction of $\mathcal{G}$ to $\mathbb{R} \times \mathbb{R}$ is real, and bounded below away from zero by (4.4). It follows that there exists $\tilde{\lambda} > 0$ so that $|\mathcal{G}|$ is bounded below away from zero on $(\mathbb{R} + i[-\tilde{\lambda}, \tilde{\lambda}]) \times \mathbb{R}$. We can therefore define $\mathcal{G}^{-1/2}$ as a uniformly continuous function on $(\mathbb{R} + i[-\tilde{\lambda}, \tilde{\lambda}]) \times \mathbb{R}$; $\mathcal{G}(z, s)^{-1/2}$ is analytic in $z \in \mathbb{R} + i(-\tilde{\lambda}, \tilde{\lambda})$ for all $s \in \mathbb{R}$. We can therefore extend (4.6), and define for $z \in \mathbb{R} + i(-\tilde{\lambda}, \tilde{\lambda})$, $s \in \mathbb{R}$,

$$\Phi(z, s) = \sqrt{2}\, \mathcal{G}(z, s)^{-1/2} G(z, s).$$

By (4.8) and (4.10) this extension satisfies

$$\begin{aligned} &\Phi(z+1, s) = \Phi(z, s), \\ &\Phi(z, s+1) = e^{2\pi i z} \Phi(z, s). \end{aligned} \tag{4.11}$$

We can now use this extension to prove exponential decay of ϕ. By (4.5) and (3.4)

$$\phi(x) = \frac{1}{\sqrt{2}} \int_0^1 dt\, \Phi\left(t, \frac{x}{2}\right).$$

Assume that $x \geqq 0$. (We will treat $x \leqq 0$ afterwards.) Using the analyticity of Φ in $t + i\tau$, we can deform the integration path,

$$\phi(x) = \frac{1}{\sqrt{2}} \left[\int_0^\Lambda d\tau\, \Phi\left(i\tau, \frac{x}{2}\right) + \int_0^1 dt\, \Phi\left(t + i\Lambda, \frac{x}{2}\right) + \int_\Lambda^0 d\tau\, \Phi\left(1 + i\tau, \frac{x}{2}\right) \right],$$

where we assume $0 < \Lambda < \tilde{\lambda}$. Since $\Phi(1 + i\tau, x/2) = \Phi(i\tau, x/2)$, the first and third integral cancel out. If $x = 2n + 2x_1$, with $x_1 \in [0, 1]$, then, by (4.11),

$$\begin{aligned} |\phi(x)| &= \frac{1}{\sqrt{2}} \left| \int_0^1 dt\, e^{2\pi i n (t + i\Lambda)} \Phi(t + i\Lambda, x_1) \right| \\ &\leqq \frac{1}{\sqrt{2}}\, e^{-2\pi\Lambda n} \sup_{\substack{z \in [0,1] + i[-\tilde{\lambda}, \tilde{\lambda}] \\ s \in [0,1]}} |\Phi(z, s)| \\ &\leqq C'\, e^{-\pi\Lambda x}. \end{aligned}$$

For $x \leqq 0$ we use the same argument, but we deform the integration path by going into the Im $z < 0$ half plane. It follows that for all Λ such that

$$\Lambda < \min\left(\frac{\lambda}{\pi}, \inf\{|\tau|;\ \mathcal{G}(t + i\tau, s) = 0 \text{ for some } t, s \in [0, 1]\} \right),$$

there exists a constant C_Λ such that

$$|\phi(x)| \leqq C_\Lambda\, e^{-\pi\Lambda|x|}. \tag{4.12}$$

3. To prove the exponential decay of $\hat{\phi}$, we use the connection (3.5) between the Zak transforms of a function and of its Fourier transform. Because of (4.2) and (3.5b), arguments similar to those in step 2 above show that G can be extended to a uniformly continuous function on $\mathbb{R} \times (\mathbb{R} + i(\mu/4\pi, \infty))$, and that, for every $t \in \mathbb{R}$, $G(t, s + i\sigma)$ is analytic in $s + i\sigma \in \mathbb{R} + i(\mu/4\pi, \infty)$. We can now define, for $t \in \mathbb{R}$, $w = s + i\sigma \in \mathbb{R} + i(\mu/4\pi, \infty)$,

$$\Gamma(t, w) = G(t, w) G(-t, w) + G(t, w + \tfrac{1}{2}) G(-t, w + \tfrac{1}{2}).$$

Again $\Gamma(t, w)$ is analytic, and there exists $\tilde{\mu} > 0$ so that $|\Gamma|$ is bounded below away from zero on $\mathbb{R} \times (\mathbb{R} + i[-\tilde{\mu}, \tilde{\mu}])$. It follows that Φ has an extension to $\mathbb{R} \times (\mathbb{R} + i[-\tilde{\mu}, \tilde{\mu}])$,

$$\Phi(t, s+i\sigma) = \sqrt{2}\, G(t, s+i\sigma)\Gamma(t, s+i\sigma)^{-1/2},$$

which is analytic in $s+i\sigma$ for every fixed t, and which satisfies

$$\Phi(t, w+1) = e^{2\pi i t}\Phi(t, w),$$

$$\Phi(t+1, w) = \Phi(t, w).$$

By (3.4) and (3.5a) we have

$$\hat{\phi}(y) = \frac{1}{\sqrt{2}} \int_0^1 ds\, (U_Z \hat{\phi})\left(s, \frac{y}{2}\right)$$

$$= \frac{1}{2\sqrt{2}} \sum_{j=0}^{3} \int_0^1 ds\, e^{\pi i y(s+j)} \Phi\left(-2y, \frac{s+j}{4}\right).$$

We can now play the same game as before (deform the integral over s into the complex plane, $\cdots$). The result is that for all Δ such that

$$\Delta < \min\left(\frac{\mu}{\pi}, 4 \inf\{|\sigma|;\, \Gamma(t, s+i\sigma) = 0 \text{ for some } t, s \in [0, 1]\}\right),$$

there exists a constant $\hat{C}_\Delta$ such that

$$|\hat{\phi}(y)| \leqq \hat{C}_\Delta\, e^{-\pi\Delta|y|}. \tag{4.13}$$

4. It remains to show that ϕ satisfies (2.5). It is obvious from $|G(t, s+1)| = |G(t, s)|$ and from (4.6) that

$$|\Phi(t, s)|^2 + |\Phi(t, s+\tfrac{1}{2})|^2 = 2 \tag{4.14}$$

for all $t, s \in \mathbb{R}$. Because of the exponential decay of ϕ and $\hat{\phi}$, all the manipulations of § 3 are indeed allowed, and (4.14) implies (2.5). □

Any function g satisfying (4.1), (4.2), and (4.4) can therefore be used to construct an orthonormal Wilson basis of type (1.9). An explicit example is given by the Gaussian

$$g(x) = (2\nu)^{1/4}\, e^{-\nu\pi x^2}. \tag{4.15}$$

The Zak transform of g is related to one of Jacobi's theta functions,

$$\begin{aligned} G(t, s) &= \sqrt{2}\, (2\nu)^{1/4}\, e^{-4\nu\pi s^2} \sum_{\ell} e^{-4\nu\pi\ell^2}\, e^{2\pi\ell(4\nu s + it)} \\ &= \sqrt{2}\, (2\nu)^{1/4}\, e^{-4\nu\pi s^2} \theta_3(t - 4i\nu s \,|\, 4i\nu), \end{aligned} \tag{4.16}$$

with Bateman's notation [26]

$$\theta_3(z \,|\, \tau) = 1 + 2 \sum_{\ell=1}^{\infty} \cos(2\pi\ell z)\, e^{i\pi\tau\ell^2}.$$

As defined by (4.16), the function G has only one zero in $[0, 1]^2$, namely in $t = s = \frac{1}{2}$ [26]. Consequently, (4.4) is satisfied. Since g and $\hat{g}(y) = (2/\nu)^{1/4}\, e^{-\pi y^2/\nu}$ obviously have exponential decay, the construction (4.5)–(4.6) does lead to a Wilson basis with exponential phase space localization. For $\nu = 0.5$ we find

$$\inf\{|\tau|;\, \mathcal{G}(t + i\tau, s) = 0 \text{ for some } t, s \in [0, 1]\} = 0.5,$$

$$\inf\{|\sigma|;\, \Gamma(t, s+i\sigma) = 0 \text{ for some } t, s \in [0, 1]\} = 0.25.$$

Consequently, for every $\varepsilon > 0$ there exists C_ε such that

$$|\phi(x)| \leqq C_\varepsilon e^{-(\pi-\varepsilon)|x|/2}, \qquad |\hat{\phi}(y)| \leqq C_\varepsilon e^{-(\pi-\varepsilon)|x|}. \tag{4.17}$$

Remarks. 1. The decay rates in (4.17) can be adjusted by starting with a Gaussian different from (4.15). For $\nu = 2^{-1/2}$ e.g., we find that the corresponding ϕ and $\hat{\phi}$ are bounded by

$$|\phi(x)| \leqq C_\varepsilon \exp(-(\pi-\varepsilon)|x|/\sqrt{2}), \qquad |\hat{\phi}(y)| \leqq C_\varepsilon \exp(-(\pi-\varepsilon)|y|/\sqrt{2}). \tag{4.18}$$

2. It is easy to show that if g is an even function, then ϕ is even as well.

3. In [16] the explanation for the existence and exponential decay of the basis constructed by Wilson in [15] starts from an ansatz different from (1.9): the bimodal functions used as a starting point are of the form

$$g\left(\xi - \frac{2m+1}{4}\right) + (-1)^m g\left(\xi + \frac{2m+1}{4}\right). \tag{4.19}$$

For this ansatz the normalization (2.1) and the "completeness requirement" (2.2) do not reduce to the same condition. The orthonormalization of the functions in (4.19), starting from a "nice" g, results therefore in

$$\hat{f}_m(\xi) = \phi_m^1\left(\xi - \frac{2m+1}{4}\right) + \phi_m^2\left(\xi + \frac{2m+1}{4}\right),$$

where the $\phi_m^{1,2}$ depend on m. In the orthonormalization procedure in [16] the "overlap matrix" of the functions (4.19) is used. This overlap matrix also contains the quantity $|G(t, s)|^2 + |G(t, s+\frac{1}{2})|^2$, where G is the Zak transform of g (see Appendix B in [16]; the notation is very different, however). The merit of the present construction, starting from (1.9), is that the orthonormality (2.1) automatically follows once (2.2) is established; moreover, (2.1) + (2.2) are equivalent to the single condition (3.6), which enables us to construct, via (4.5)-(4.6) a *single* function ϕ generating the whole Wilson basis.

5. The link with tight frames. We start by briefly reviewing some material concerning "frames." Frames were introduced by Duffin and Schaeffer [27] in the context of nonharmonic Fourier series; in [23] and [7] special frames, constituted by families of functions of type (1.1), were studied in connection with the windowed Fourier transform. We review here some results from [7].

A family of g_{mn}, as defined in (1.1), constitutes a frame if there exist $A > 0$, $B < \infty$ such that, for all f in $L^2(\mathbb{R})$,

$$A\|f\|^2 \leqq \sum_{m,n} |\langle g_{mn}, f\rangle|^2 \leqq B\|f\|^2. \tag{5.1}$$

This condition can also be rewritten as

$$A\,\mathrm{Id} \leqq \mathbf{P} \leqq B\,\mathrm{Id}, \tag{5.2}$$

where $\mathbf{P}$ is the positive operator

$$\mathbf{P} = \sum_{m,n} P_{mn}, \qquad P_{mn} f = \langle g_{mn}, f\rangle g_{mn}. \tag{5.3}$$

If the g_{mn} constitute a frame, then functions $f \in L^2(\mathbb{R})$ can be completely characterized by the family of inner products $(\langle g_{mn}, f\rangle)_{m,n\in\mathbb{Z}}$, and there exists a numerically stable inversion procedure to reconstruct f from these inner products,

$$f = \sum_{m,n} \langle g_{mn}, f\rangle \tilde{g}_{mn},$$

with

$$\tilde{g}_{mn}(x) = e^{2\pi i\alpha mn}\tilde{g}(x-\beta n),$$
$$\tilde{g} = \mathbf{P}^{-1}g,$$

with $\mathbf{P}$ as defined by (5.3). Because of (5.2), $\mathbf{P}$ has a bounded inverse, so that $\tilde{g}$ is well defined. A special case arises when the frame is *tight*, i.e., when the frame bounds A and B are equal,

$$\sum_{m,n\in\mathbb{Z}} |\langle g_{mn}, f\rangle|^2 = A\|f\|^2.$$

It then follows that

$$\mathbf{P} = A\,\mathrm{Id},$$
$$\tilde{g} = A^{-1}g,$$
$$f = A^{-1}\sum_{m,n\in\mathbb{Z}} \langle g_{mn}, f\rangle g_{mn}.$$

In general, frames are redundant (they contain "too many" vectors, or more precisely, any frame vector lies in the closed linear span of all the others). If the frame is tight, then A indicates how redundant the frame is; for tight frames of type (1.1) we find [7]

$$A = (\alpha\beta)^{-1}\|g\|^2. \tag{5.4}$$

A frame of type (1.1) can only be an orthonormal basis if $\alpha\beta = 1$ (and if, moreover, g is chosen appropriately), corresponding to $A = 1$, or no redundancy. Tight frames with "nice" g exist if and only if $\alpha\beta < 1$; see [23] for a construction with compactly supported g.

Let us now specialize to the case $\alpha = .5$, $\beta = 1$,

$$g_{mn}(x) = e^{im\pi x}g(x-n).$$

The density of the phase space lattice corresponding to the g_{mn} is then twice as high as for an orthonormal basis. Suppose g is "nice," i.e., both g and $\hat{g}$ have fast decay at ∞. Let us investigate under which conditions on g the g_{mn} constitute a frame (respectively, tight frame). Because $(\alpha\beta)^{-1} = 2$ is an integer, the Zak transform is a natural tool to study these questions, as observed in [8]. Using (3.1), we find

$$(U_Z g_{m2n})(t, s) = e^{-2\pi itn}\, e^{2\pi ims} G(t, s),$$
$$(U_Z g_{m2n-1})(t, s) = e^{-2\pi itn}\, e^{2\pi ims} G(t, s+\tfrac{1}{2}),$$

where $G = U_Z g$. It follows that, for all $h_1, h_2 \in L^2(\mathbb{R})$,

$$\begin{aligned}\sum_{m,n\in\mathbb{Z}} \langle h_1, P_{m2n}h_2\rangle &= \sum_{m,n\in\mathbb{Z}} \langle h_1, g_{m2n}\rangle\langle g_{m2n}, h_2\rangle \\ &= \sum_{m,n\in\mathbb{Z}} \left[\int_0^1 dt \int_0^1 ds\, U_Z h_1(t, s)\overline{G(t, s)}\, e^{2\pi itn}\, e^{-2\pi ims}\right]^* \\ &\quad \cdot \left[\int_0^1 dt \int_0^1 ds\, U_Z h_2(t, s)\overline{G(t, s)}\, e^{2\pi itn}\, e^{-2\pi ims}\right] \\ &= \int_0^1 dt \int_0^1 ds\, \overline{U_Z h_1(t, s)}\, U_Z h_2(t, s)|G(t, s)|^2.\end{aligned}$$

Consequently, $U_Z[\sum_{m,n\in\mathbb{Z}} P_{m2n}]U_Z^{-1}$ is multiplication by $|G(t,s)|^2$ in $\mathscr{Z}$. Similarly, $U_Z[\sum_{m,n\in\mathbb{Z}} P_{m2n-1}]U_Z^{-1}$ is multiplication by $|G(t,s+\frac{1}{2})|^2$. Consequently, $\mathbf{P}=\sum_{m,n} P_{mn}$ is unitarily equivalent to multiplication by $|G(t,s)|^2+|G(t,s+\frac{1}{2})|^2$ on $\mathscr{Z}$. It follows that the g_{mn} constitute a frame, or equivalently that $\mathbf{P}$ satisfies (5.2), if and only if

$$0<A\leqq|G(t,s)|^2+|G(t,s+\tfrac{1}{2})|^2\leqq B<\infty$$

for all $t, s\in[0,1]$. All this is summarized in the following proposition.

PROPOSITION 5.1. *The functions $g_{mn}(x)=e^{im\pi x}g(x-n)$ constitute a frame if and only if the Zak transform $G=U_Zg$ of g, as defined by* (3.1), *satisfies*

$$A=\inf_{t,s\in[0,1]}[|G(t,s)|^2+|G(t,s+\tfrac{1}{2})|^2]>0$$

and

$$B=\sup_{t,s\in[0,1]}[|G(t,s)|^2+|G(t,s+\tfrac{1}{2})|^2]<\infty.$$

Note that if $|g(x)|\leqq C(1+|x|)^{-1-\varepsilon}$, then G is bounded, and B is automatically finite. There are other procedures than the Zak transform to check whether the g_{mn} constitute a frame [7]. The point of Proposition 5.1 is that any reasonably well-localized g such that the g_{mn} constitute a frame can be used as a starting point in the construction of ϕ in § 4. Note that the computations above also prove the following proposition.

PROPOSITION 5.2. *Let ϕ be a real function such that $|\phi(x)|\leqq C(1+|x|)^{-1-\varepsilon}$ and $\int dx\,|\phi(x)|^2=1$. Then the following are equivalent*:

(1) *The ψ_{mn}, as defined by* (1.9), *constitute an orthonormal basis,*

(2) *The Zak transform $\Phi=U_Z\phi$ of ϕ satisfies*

$$|\Phi(t,s)|^2+|\Phi(t,s+\tfrac{1}{2})|^2=2,$$

(3) *The functions $\phi_{mn}(x)=e^{im\pi x}\phi(x-n)$, $m, n\in\mathbb{Z}$, constitute a tight frame.*

Proof.

(1) $\Leftrightarrow$ (2) is proved in Proposition 5.2.

Define now $\mathbf{P}(\phi)$ by

$$\mathbf{P}(\phi)f=\sum_{m,n}\langle\phi_{mn},f\rangle\phi_{mn}.$$

Then, by the computation above,

$$\mathbf{P}(\phi)=U_Z^{-1}\{\text{multiplication with } [|\Phi(t,s)^2|+|\Phi(t,s+\tfrac{1}{2})|^2]\}U_Z. \tag{5.5}$$

If (2) holds, then it follows that $\mathbf{P}(\phi)=2\,\mathrm{Id}$, i.e.,

$$\sum_{m,n}|\langle\phi_{mn},f\rangle|^2=2\|f\|^2,$$

so the ϕ_{mn} constitute a tight frame.

On the other hand, if the ϕ_{mn} constitute a tight frame, i.e.,

$$\sum_{m,n}|\langle\phi_{mn},f\rangle|^2=A\|f\|^2,$$

then $A=2$ by (5.4) ($\alpha=.5$, $\beta=1$, and $\|g\|=1$). It follows that $\mathbf{P}(\phi)=2\,\mathrm{Id}$, which by (5.5) implies (2). □

Remark. From this analysis it follows that the construction in § 4 and § 6 below can also be used to generate tight frames with exponential localization in both time and frequency. The construction in § 6 can easily be extended to tight frames with arbitrary redundancy. These tight frames contrast with those constructed in [23], where either ϕ or $\hat{\phi}$ had compact support.

Proposition 5.2 leads to the following interpretation of our Wilson bases. Suppose that the ϕ_{mn} constitute a tight frame. Since $\alpha\beta = .5$, this tight frame has redundancy 2, i.e., it has "two times as many vectors" as an orthonormal basis. The Wilson basis vectors generated by ϕ via (1.9) are given by

$$(\psi_{mn})\hat{}(\xi) = e^{2\pi i n\xi}\hat{f}_m(\xi),$$

or

$$(\psi_{1n})\hat{}(\xi) = e^{2\pi i n\xi}\phi(\xi) = \phi_{02n}(\xi), \tag{5.6a}$$

$$(\psi_{2\ell+\kappa n})\hat{}(\xi) = \frac{1}{\sqrt{2}}\, e^{2\pi i n\xi}\, e^{i\pi\kappa\xi}[\phi(\xi-\ell)+(-1)^{\ell+\kappa}\phi(\xi+\ell)]$$

$$= \frac{1}{\sqrt{2}}(\phi_{\ell 2n+\kappa}+(-1)^{\ell+\kappa}\phi_{-\ell 2n+\kappa})(\xi), \tag{5.6b}$$

$$\ell \in \mathbb{N}\backslash\{0\}, \qquad \kappa = 0 \text{ or } 1.$$

Formula (1.9) can therefore be viewed as a procedure eliminating the redundancy factor 2 from the tight frame ϕ_{mn} by choosing only the ϕ_{0n} with even n, and replacing every pair $\phi_{\ell n}$, $\phi_{-\ell n}$ ($\ell \neq 0$) by one judiciously chosen linear combination of these two vectors. It seems a small miracle that the result is an orthonormal basis!

Note that (5.6) can be made even simpler by a relabelling of the ψ_{mn}. Denote

$$\Psi_{\ell 2n+\kappa} = \psi_{2\ell+\kappa n}, \qquad \ell \neq 0, \quad \kappa = 0 \text{ or } 1,$$

$$\Psi_{0n} = \psi_{1n}.$$

Then (5.6) becomes

$$(\Psi_{0n})\hat{} = \phi_{02n},$$

$$(\Psi_{\ell n})\hat{} = \frac{1}{\sqrt{2}}(\phi_{\ell n}+(-1)^{\ell+n}\phi_{-\ell n}), \qquad \ell \neq 0,$$

making the reduction from the right frame with redundancy 2 to the orthonormal basis even more elegant.

6. The construction revisited. The equivalence between (4.4) and the frame condition (5.1) leads to an alternate construction for the function ϕ which is very easy to implement numerically.

Choose g such that (4.1), (4.2), and (4.4) are satisfied. Then, by the argument in § 5,

$$U_Z \mathbf{P} U_Z^{-1} = \text{multiplication by } |G(t,s)|^2+|G(t,s+\tfrac{1}{2})|^2;$$

consequently,

$$\phi = U_Z^{-1}\frac{\sqrt{2}\,G}{\sqrt{|G(t,s)|^2+|G(t,s+\frac{1}{2})|^2}} = U_Z^{-1}U_Z\mathbf{P}^{-1/2}U_Z^{-1}\sqrt{2}\,G$$

$$= \sqrt{2}\,\mathbf{P}^{-1/2}U_Z^{-1}G = \sqrt{2}\,\mathbf{P}^{-1/2}g.$$

The operator $\mathbf{P}^{-1/2}$ can be written as a convergent series. Since

$$A\,\mathrm{Id} \leqq \mathbf{P} \leqq B\,\mathrm{Id}$$

for all $A>0$, $B<\infty$ satisfying

$$A \leqq \inf_{t,s\in[0,1]} [|G(t,s)|^2+|G(t,s+\tfrac{1}{2})|^2],$$

$$B \geqq \sup_{t,s\in[0,1]} [|G(t,s)|^2+|G(t,s+\tfrac{1}{2})|^2],$$

we have

$$\begin{aligned}\mathbf{P}^{-1/2} &= \left(\frac{2}{A+B}\right)^{1/2}\left[\mathrm{Id}-\left(\mathrm{Id}-\frac{2\mathbf{P}}{A+B}\right)\right]^{-1/2} \\ &= \left(\frac{2}{A+B}\right)^{1/2}\sum_{k=0}^{\infty}\frac{(2k)!}{2^{2k}(k!)^2}\left(\mathrm{Id}-\frac{2\mathbf{P}}{A+B}\right)^k,\end{aligned} \tag{6.1}$$

where the series converges because

$$\left\|\mathrm{Id}-\frac{2\mathbf{P}}{A+B}\right\| \leqq \frac{B-A}{B+A}<1.$$

This can be used to write ϕ as a combination of g_{mn}, with coefficients computed recursively. For instance, if g is Gaussian, $g^{\nu}(x)=(2\nu)^{1/4}\,e^{-\nu\pi x^2}$, then we find

$$\phi(x)=\frac{2}{\sqrt{A_\nu+B_\nu}}\sum_{m,n\in\mathbb{Z}} a_{mn}g^{\nu}_{mn}(x) \tag{6.2}$$

with

$$\begin{aligned}a_{mn} &= \sum_{k=0}^{\infty}\frac{(2k)!}{2^{2k}(k!)^2}\,b_{mn}^k, \\ b_{mn}^k &= \left(1-\frac{2}{A_\nu+B_\nu}\right)b_{mn}^{k-1}-\frac{2}{A_\nu+B_\nu}\sum_{(m',n')\neq(m,n)}\omega_{mn,m'n'}b_{m'n'}^{k-1},\end{aligned} \tag{6.3}$$

where

$$\omega_{mn,m'n'}=\exp\left[i(m'-m)(n+n')\frac{\pi}{2}-\frac{\nu\pi}{2}(n-n')^2-\frac{\pi}{8\nu}(m-m')^2\right],$$

$$b_{mn}^0=\delta_{m0}\delta_{n0}.$$

While this seems lengthy, it is very easy to program on a computer. The procedure converges at least as fast as a geometric series in $(B_\nu-A_\nu)/(B_\nu+A_\nu)$. For $\nu=.5$ we find $A_\nu=1.670$, $B_\nu=2.361$, $(B_\nu-A_\nu)/(B_\nu+A_\nu)=.1712$; for $\nu=2^{-1/2}$, we have $A_\nu=1.533$, $B_\nu=2.492$, $(B_\nu-A_\nu)/(B_\nu+A_\nu)=.2381$. Figures 1 and 2 give graphs of ϕ and $\hat{\phi}$, for $\nu=.5$ and $\nu=2^{-1/2}$, respectively.

Remarks. 1. A and B can be computed via the Zak transform:

$$A=\inf_{t,s\in[0,1]} [|(U_Zg)(t,s)|^2+|(U_Zg)(t,s+\tfrac{1}{2})|^2],$$

$$B=\sup_{t,s\in[0,1]} [|(U_Zg)(t,s)|^2+|(U_Zg)(t,s+\tfrac{1}{2})|^2].$$

In [7] an alternative way of estimating A and B is given, leading to a lower bound for A and an upper bound for B, without recourse to the Zak transform. Using the Poisson summation formula, we find

$$\mathbf{m}(g)-\mathbf{r}(g)\leqq A\leqq B\leqq \mathbf{M}(g)+\mathbf{r}(g), \tag{6.4}$$

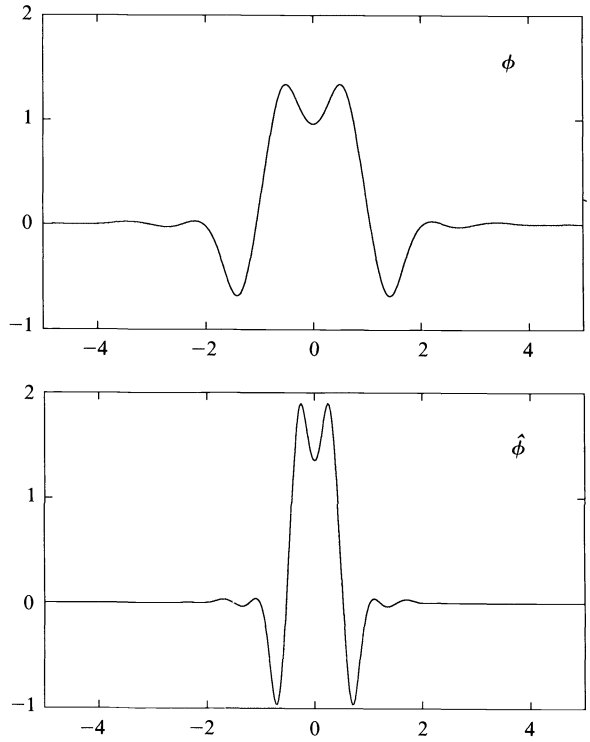


FIG. 1. *Plots of ϕ and $\hat{\phi}$, constructed from $g(x) = e^{-\pi x^2/2}$. For this choice of g, we have $\hat{\phi}(y) = \sqrt{2}\,\phi(2y)$ (see (6.6)). To draw these graphs, the recursive computation* (6.2), (6.3) *was used.*

where

$$\mathbf{m}(g) = \inf_{x\in[0,1]} \sum_n |g(x-n)|^2,$$

$$\mathbf{M}(g) = \sup_{x\in[0,1]} \sum_n |g(x-n)|^2,$$

$$\mathbf{r}(g) = 2\sum_{k=1}^{\infty} [\beta(2k)\beta(-2k)]^{1/2},$$

$$\beta(s) = \sup_{x\in[0,1]} \sum_n |g(x-n)g(x-n+s)|.$$

Note that the lower bound in (6.4) also gives a sufficient condition ensuring that (4.4) holds, without having to check the Zak transform. In some cases, more efficient bounds can be computed from the Fourier transform $\hat{g}$ of g. We obtain [7]

$$\tilde{\mathbf{m}}(g) - \tilde{\mathbf{r}}(g) \leqq A \leqq B \leqq \tilde{\mathbf{M}}(g) - \tilde{\mathbf{r}}(g)$$

with

$$\tilde{\mathbf{m}}(g) = \inf_{\xi\in[0,1/2]} \sum_n |\hat{g}(\xi - n/2)|^2,$$

$$\tilde{\mathbf{M}}(g) = \sup_{\xi\in[0,1/2]} \sum_n |\hat{g}(\xi - n/2)|^2,$$

$$\tilde{\mathbf{r}}(g) = 2\sum_{k=1}^{\infty} [\tilde{\beta}(k)\tilde{\beta}(-k)]^{1/2},$$

$$\tilde{\beta}(s) = \sup_{\xi\in[0,1/2]} \sum_n |\hat{g}(\xi - n/2)\hat{g}(\xi - n/2 + s)|.$$

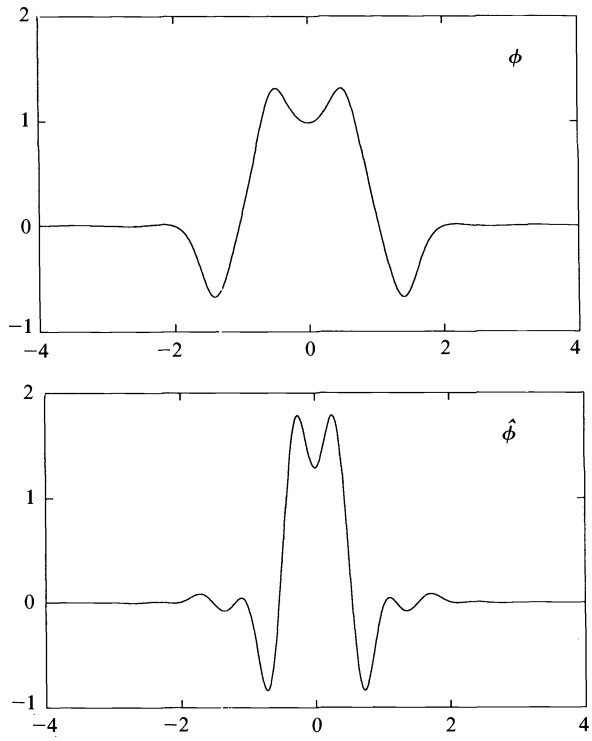


FIG. 2. *Plots of ϕ and $\hat{\phi}$ constructed from $g(x) = 2^{1/8} e^{-\pi x^2/\sqrt{2}}$. For this choice of g, the decay rates of ϕ and $\hat{\phi}$ are identical (see (4.18)). We have again used (6.2), (6.3) to compute ϕ; $\hat{\phi}$ is simply obtained by replacing g^{ν}_{mn} in (6.2) by $(g^{\nu}_{mn})\hat{}(y) = (-1)^{mn} e^{2\pi i y n} g^{1/\nu}(y + (m/2))$.*

2. Let us introduce the notation F for the Fourier transform and D_a for the dilations $(D_a f)(x) = |a|^{1/2} f(ax)$, and let us write $\mathbf{P}(g)$, $\phi(g)$ to make the dependence of the operator $\mathbf{P}$ and the function ϕ on the function g more explicit. Then we easily check that

$$D_{1/2} F \mathbf{P}(g) = \mathbf{P}(D_{1/2} F g) D_{1/2} F,$$

implying

$$D_{1/2} F \phi(g) = \phi(D_{1/2} F g). \tag{6.5}$$

Denoting $\phi(g^{\nu})$ by ϕ_{ν}, where $g^{\nu}(x) = (2\nu)^{1/4} \exp(-\pi \nu x^2)$, we find therefore

$$(\phi_{\nu})\hat{}(y) = \sqrt{2}\, \phi_{(4\nu)^{-1}}(2y). \tag{6.6}$$

In particular, for $\nu = .5$, $(\phi_{1/2})\hat{}(y) = \sqrt{2}\, \phi_{1/2}(2y)$.

7. Conclusion. We have shown how to construct very simple Wilson bases ψ_{mn}, generated by a single function ϕ, via

$$\begin{aligned} &\psi_{mn}(x) = f_m(x - n), \qquad n \in \mathbb{Z}, \quad m \in \mathbb{N}\backslash\{0\}, \\ &\hat{f}_1(\xi) = \phi(\xi), \\ &\hat{f}_{2\ell+\kappa}(\xi) = \frac{1}{\sqrt{2}} [\phi(\xi - \ell) + (-1)^{\ell+\kappa} \phi(\xi + \ell)]\, e^{i\pi\kappa\xi}, \qquad \ell \in \mathbb{N}\backslash\{0\}, \quad \kappa = 0 \text{ or } 1. \end{aligned} \tag{7.1}$$

We have explicitly constructed such bases; in order to obtain exponential decay for both the ψ_{mn} and their Fourier transforms $\hat{\psi}_{mn}$, it suffices to choose a function g such that g and $\hat{g}$ have exponential decay, and such that condition (4.4) is satisfied, or equivalently, such that the $g_{mn}(x) = e^{\pi i m x} g(x-n)$ constitute a frame. (For this it is sufficient that $\mathbf{m}(g) - \mathbf{r}(g) > 0$ or $\tilde{\mathbf{m}}(g) - \tilde{\mathbf{r}}(g) > 0$—see § 6.) The function ϕ can then be constructed from g either via the Zak transform (see § 4) or via a recursive algorithm (see § 6).

The functions f_m in (7.1) are given by the inverse Fourier transform of ϕ. If g is real and even, then so is ϕ, so that its Fourier transform and inverse Fourier transform coincide. We then have

$$f_1(x) = \hat{\phi}(x),$$

$$f_{2\ell+\kappa}(x) = \frac{1}{\sqrt{2}} \hat{\phi}\left(x + \frac{\kappa}{2}\right) e^{\pi i \ell \kappa} \left[e^{2\pi i \ell x} + (-1)^{\ell+\kappa} e^{-2\pi i \ell x} \right].$$

Using the relabelling

$$\Psi_{0n} = \psi_{1n}, \qquad \Psi_{\ell 2n+\kappa} = \psi_{2\ell+\kappa n}$$

we find

$$\Psi_{0n}(x) = \hat{\phi}(x-n),$$

$$\Psi_{\ell n}(x) = \sqrt{2}\, \hat{\phi}\left(x - \frac{n}{2}\right) \begin{cases} \cos(2\pi \ell x) & \text{if } \ell + n \text{ is even,} \\ \sin(2\pi \ell x) & \text{if } \ell + n \text{ is odd.} \end{cases}$$

It follows that the Wilson bases constructed here are very similar to the functions (1.1): the only difference is the alternate use of sines and cosines instead of complex exponentials. This trick is sufficient to beat the no-go Balian-Low theorem.

REFERENCES

[1] W. KOHN, *Analytic properties of Bloch waves and Wannier functions*, Phys. Rev., 115 (1959), pp. 809-821.

[2] D. GABOR, *Theory of communication*, J. Inst. Electr. Engrg. London (III), 93 (1946), pp. 429-457.

[3a] M. J. BASTIAANS, *Gabor's signal expansion and degrees of freedom of a signal*, Proc. IEEE, 68 (1980), pp. 538-539.

[3b] ———, *A sampling theorem for the complex spectrogram and Gabor's expansion of a signal in Gaussian elementary signals*, Optical Engrg., 20 (1981), pp. 594-598.

[4a] A. J. E. M. JANSSEN, *Gabor representation of generalized functions*, J. Math. Appl., 80 (1981), pp. 377-394.

[4b] ———, *Gabor representation and Wigner distribution of signals*, Proc. IEEE (1984), pp. 41.B.2.1-41.B.2.4.

[5] V. BARGMANN, P. BUTERA, L. GIRARDELLO, AND J. R. KLAUDER, *On the completeness of coherent states*, Rep. Math. Phys., 2 (1971), pp. 221-228.

[6] A. M. PERELOMOV, *Note on the completeness of systems of coherent states*, Teor. i Matem. Fis., 6 (1971), pp. 213-224.

[7] I. DAUBECHIES, *The wavelet transform, time-frequency localization and signal analysis*, IEEE Trans. Inf. Theory, 36 (1990) pp. 961-1005.

[8] I. DAUBECHIES AND A. GROSSMANN, *Frames of entire functions in the Bargmann space*, Comm. Pure Appl. Math., 41 (1988), pp. 151-164.

[9] R. BALIAN, *Un principe d'incertitude fort en théorie du signal ou en mécanique quantique*, C.R. Acad. Sci. Paris, Sér. 2, 292 (1981), pp. 1357-1361.

[10] F. LOW, *Complete Sets of Wave-Packets*, in A Passion for Physics—Essays in Honor of Geoffrey Chew, World Scientific, Singapore, 1985, pp. 17-22.

[11] G. BATTLE, *Heisenberg proof of the Balian-Low theorem*, Lett. Math. Phys., 15 (1988), pp. 175-177.

[12] I. DAUBECHIES AND A. J. E. M. JANSSEN, *Two theorems on lattice expansions*, preprint.

[13] J. BOURGAIN, *A remark on the uncertainty principle for Hilbertian basis*, J. Funct. Anal., 79 (1988), pp. 136-143.
[14] T. STEGER, private communication, 1986.
[15] K. G. WILSON, *Generalized Wannier Functions*, Cornell University preprint, 1987.
[16] D. J. SULLIVAN, J. J. REHR, J. W. WILKINS, AND K. G. WILSON, *Phase Space Wannier Functions in Electronic Structure Calculations*, Cornell University, preprint, 1987.
[17] Y. MEYER, *Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs*, Séminaire Bourbaki, Paris, France, No. 662, 1985-1986.
[18] J. O. STROMBERG, *A modified Franklin system and higher order spline systems on $\mathbb{R}^n$ as unconditional bases for Hardy spaces*, Conf. in Honor of A. Zygmund, Vol. II, W. Beckner, A. P. Calderon, R. Fefferman, and P. Jones, eds., Wadsworth Math. Series, Belmont, CA, 1983, pp. 475-493.
[19] G. BATTLE, *A block spin construction of ondelettes. Part* I. *Lemarié functions*, Comm. Math. Phys., 110 (1987), pp. 601-615.
[20] P. G. LEMARIÉ, *Une nouvelle base d'ondelettes de $L^2(\mathbb{R}^n)$*, J. Math. Pures Appl., 67 (1988), pp. 227-236.
[21] I. DAUBECHIES, *Orthonormal basis of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909-996.
[22] A. COHEN, private communication, 1988.
[23] I. DAUBECHIES, A. GROSSMANN, AND Y. MEYER, *Painless non-orthogonal expansions*, J. Math. Phys., 27 (1986), pp. 1271-1283.
[24a] J. ZAK, *Finite translations in solid state physics*, Phys. Rev. Lett., 19 (1967), pp. 1385-1397.
[24b] ———, *Dynamics of electrons in solids in external fields*, Phys. Rev., 168 (1968), pp. 686-695.
[24c] ———, *The kq representation in the dynamics of electrons in solids*, Solid State Physics, 27 (1972), pp. 1-62.
[25] A. J. E. M. JANSSEN, *The Zak transform: a signal transform for sampled time-continuous signals*, Philips J. Res., 43 (1988), pp. 23-69.
[26] BATEMAN PROJECT, *Higher Transcendental Functions, Vol.* 2, McGraw-Hill, New York, 1955.
[27] R. J. DUFFIN AND A. C. SCHAEFFER, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc., 72 (1952), pp. 341-366.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 575-586, May 1991

A HYPERBOLIC THEORY FOR THE EVOLUTION OF PLANE CURVES*

MORTON E. GURTIN† AND PAOLO PODIO-GUIDUGLI‡

Abstract. A theory is developed for the evolution of plane curves. This theory is based on balance laws for mass and momentum in conjunction with constitutive equations appropriate to a phase interface such as that between a crystal and its melt. The resulting evolution equation is hyperbolic and has solutions with aspects qualitatively reminiscent of the melting-freezing waves observed at the surface of He^4 crystals.

Key words. evolving curves, phase interfaces, melting-freezing waves

AMS(MOS) subject classifications. 35L70, 70K99

1. Introduction. It is the purpose of this paper to develop a theory for the evolution of plane curves which is based on balance laws for mass and momentum in conjunction with constitutive equations appropriate to a phase interface, and which leads to hyperbolic evolution equations. We have three reasons for presenting such a theory:

(1) The form of the balance laws is not at all obvious, and, in fact, represents an intriguing problem in continuum mechanics whose solution requires a nonstandard conceptual framework.

(2) The *parabolic* theory for the evolution of plane curves, which in its simplest form is based on the *curve-shortening equation*[1]

$$v = \kappa \tag{1.1}$$

(relating the normal velocity v and curvature κ) has been extremely successful, providing geometers with great insight; to our knowledge there is no *hyperbolic* version of (1.1).

(3) Crystals of helium in their melt exhibit a phenomenon generally not found in other materials: oscillations[2] of the solid-liquid interface in which atoms of the solid move only when they melt and enter the liquid. Motivated by Andreev and Parshin's [AP] classical discussion of such *melting-freezing waves*, a continuum model was developed in [G4][3] for a rigid crystal in an incompressible,[4] inviscid melt. This model, which we shall refer to as the **CM model**, leads to a free-boundary problem for the evolution of the interface; coupling between the interface and the melt renders this problem difficult, and it would seem useful to have a simple model in which the motion of the interface is governed by a hyperbolic analog of (1.1).

* Received by the editors December 4, 1989; accepted for publication May 3, 1990.

† Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890. The research of this author was supported by the U.S. Army Research Office and the National Science Foundation.

‡ Dipartimento di Ing. Civele, Secunda Universitá di Roma, Roma, Italy 00173. The research of this author was supported by the Ministero della Pubblica Istruzione.

[1] See, e.g., Brakke [B], Gage and Hamilton [GH], and Grayson [Gr], as well as the references therein and those cited in § 5 of [AG].

[2] Such waves were predicted by Andreev and Parshin [AP] in 1978 and exhibited experimentally by Keshishev, Parshin, and Babkin [KPB] in 1979.

[3] Using, as a basis, a framework developed in [G1]-[G3], [AG], [GS].

[4] Andreev and Parshin [AP] note that the phase velocity of melting-freezing waves is generally well below the sound velocity.

Here we develop a theory in which *only the interface is endowed with mathematical structure*:[5] we model the inertia of the melt through an "*effective*" inertia for the interface, with the melt considered only as a source of atoms for the crystallization process; and we characterize this inertia constitutively through the corresponding interfacial mass density. As in the CM model, we restrict our attention to a purely mechanical[6] theory, and, to avoid the geometric complications that accompany evolving surfaces, to a two-dimensional theory in which the interface evolves as a plane curve.

Because of the presence of an "effective inertia," the balance laws for mass and momentum are not obvious. We derive these laws as a consequence of the requirement that the mechanical energy production—the rate at which the kinetic energy is changing minus the power expended by capillary forces—be invariant under Galilean changes in observer.

Constitutive equations, of the form derived in [G1] as a consequence of thermodynamical arguments, are assumed for the relevant interfacial fields. These equations and the underlying balance laws yield a single equation for the evolution of the interface:

$$\rho(\theta)v^\circ+\beta(\theta)v=[\psi(\theta)+\psi''(\theta)]\kappa-F. \tag{1.2}$$

Here $\psi(\theta)$, $\rho(\theta)$, and $\beta(\theta)$ are the energy, effective density, and kinetic coefficients for the interface; F is a constant which represents the driving force for crystallization; v° is the time derivative of v following the interface; and θ is the angle to the interface-normal $\mathbf{m}$. We assume that

$$\psi(\theta)+\psi''(\theta)>0, \qquad \rho(\theta)>0 \tag{1.3}$$

for all values of θ; this ensures that when (1.2) is combined with standard kinematical conditions for the evolution of a plane curve, the resulting partial differential equations are *hyperbolic*.

Equation (1.2) with $\beta(\theta)=0$, *linearized about a flat interface at equilibrium*, reduces to the classical wave equation

$$\rho_0 h_{tt}=(\psi+\psi'')_0 h_{xx} \tag{1.4}$$

for the interface expressed as a graph $y=h(x,t)$. This equation exhibits *melting-freezing waves*; that is, oscillatory solutions of the form

$$h(x,t)=C\,e^{i\lambda x}\,e^{-i\omega t} \tag{1.5}$$

with

$$\omega^2=\frac{(\psi+\psi'')_0\lambda^2}{\rho_0}. \tag{1.6}$$

As noted by Andreev and Parshin [AP] (cf. [G4]), the CM model exhibits melting-freezing waves of the form (1.5), but there ω^2 is proportional[7] to λ^3, rather than λ^2 as in (1.6). Thus the agreement between the simple model developed here and the more detailed CM model *is at most qualitative*; because of the simplified modeling of inertia, this is not unexpected.

[5] This is the point of view taken by Brower et al. [BK] and Ben-Jacob et al. [BG], who use equations involving only the interface to model interfacial evolution governed by bulk diffusion.

[6] As noted by Maris and Andreev [MA], for superconductors such as solid helium and its melt, solidification is "essentially a mechanical process, rather than a thermal process as it is for ordinary materials."

[7] This proporationality of ω^2 to λ^3 is confirmed by the experiments of Keshishev, Parshin, and Babkin [KPB].

For completeness we discuss the form the basic equations take when the interface is an evolving *surface* in $\mathbb{R}^3$. There (1.2) is replaced by

$$\rho(\mathbf{m})v^\circ+\beta(\mathbf{m})v=\psi(\mathbf{m})\kappa+\psi_{mm}(\mathbf{m})\cdot\mathbf{L}-F, \tag{1.7}$$

where $\mathbf{L}$ is the curvature tensor, $\kappa=\operatorname{tr}\mathbf{L}$ is twice the mean curvature, and $\psi_{mm}(\mathbf{m})$ is the second gradient of $\psi(\mathbf{m})$ on the surface of the unit ball.

We solve the problem (in $\mathbb{R}^2$ and $\mathbb{R}^3$) of radially symmetric crystallization of an isotropic crystal in an infinite melt. If the phase interface is initially at rest, then:

(i) For $F\geqq 0$ the crystal melts completely in finite time.

(ii) For $F<0$ there is a *critical radius* $R_{\text{crit}}:=(n-1)\psi/|F|$ (in $\mathbb{R}^n$) such that a crystal of radius $R(0)<R_{\text{crit}}$ melts in finite time, a crystal of radius $R(0)>R_{\text{crit}}$ grows unboundedly as $t\to\infty$.

An analogous problem is discussed in [G4] for the CM model in $\mathbb{R}^3$. The results are qualitatively the same as those described in (i) and (ii). In fact, if we identify F with the constant $\Psi_c+P-\zeta(\Psi+P)$ of the CM model, where Ψ_c and Ψ are the crystal and melt energies, ζ is the ratio of crystal density to melt density, and P is the far-field pressure in the melt, then the critical radii of the two theories coincide. As would be expected, the two theories exhibit quantitative differences: for example, during unbounded growth the radius grows asymptotically as t^2 within the present theory and as t within the CM model.

Although the CM model does exhibit oscillatory behavior, it is not clear whether or not shocks and other propagating discontinuities are possible.[8] To the contrary, such phenomena are generated within the present theory. We study the propagation of fronts across which the curvature is discontinuous. We show that, in the presence of anisotropy, fronts whose amplitude is sufficiently large and of the right sign grow to infinity in finite time, strongly suggesting that the interface develops a corner. Guided by other theories[9] of hyperbolic behavior, this result seems to indicate that there is global existence of classical solutions of (1.2) for initial data that are both sufficiently small and sufficiently smooth, but that smooth solutions corresponding to large data develop singularities in finite time.

2. Crystals. We consider an *infinite* **crystal lattice** modeled as a two-dimensional continuum, in fact as $\mathbb{R}^2$. A **crystal** $\mathscr{C}$ is then a compact subset of the lattice with boundary $\partial\mathscr{C}$, a smooth, simple closed curve. $\partial\mathscr{C}$ represents the **interface** between the crystal and its **melt**; we write $\mathbf{m}(\mathbf{x})$ for the outward unit normal to $\partial\mathscr{C}$ and define a unit tangent $\boldsymbol{\ell}(\mathbf{x})$ so that $\{\boldsymbol{\ell}(\mathbf{x}),\mathbf{m}(\mathbf{x})\}$ is a positively oriented basis of $\mathbb{R}^2$ (cf. Fig. 1). We let ds denote the element of arclength on $\partial\mathscr{C}$ and write f_s for the derivative, sometimes partial, of f with respect to arclength on $\partial\mathscr{C}$. (Our convention is that arclength increase in the direction of $\boldsymbol{\ell}$.) We then have the Frenet formulas

$$\mathbf{m}_s=-\kappa\boldsymbol{\ell},\qquad \boldsymbol{\ell}_s=\kappa\mathbf{m} \tag{2.1}$$

with $\kappa(\mathbf{x})$ the **curvature** of $\partial\mathscr{C}$. We define the **angle** $\theta(\mathbf{x})$, as a smooth function of $\mathbf{x}$, through

$$\mathbf{m}=(\cos\theta,\sin\theta),\qquad \boldsymbol{\ell}=(\sin\theta,-\cos\theta). \tag{2.2}$$

Our goal is to model situations in which crystals grow or shrink by processes such as solidification and melting. We therefore consider crystals $\mathscr{C}(t)$ that evolve with time

[8] Rogers [R] shows that such phenomena are not possible within the *linearized* CM theory.

[9] See, e.g., Renardy, Hrusa, and Nohel [RHN].

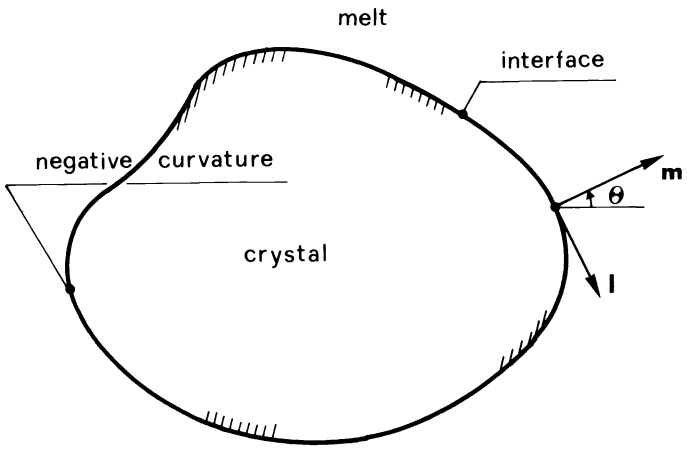


FIG. 1. *Sign conventions for interfacial motions.*

t, under the assumption that $\partial\mathscr{C}(t)$ is a smooth evolving curve (in the sense of [AG]). We write $v(\mathbf{x}, t)$ for the **normal velocity** of $\partial\mathscr{C}(t)$ in the direction $\mathbf{m}(\mathbf{x}, t)$, so that

$$\mathbf{v}(\mathbf{x}, t) = v(\mathbf{x}, t)\mathbf{m}(\mathbf{x}, t) \tag{2.3}$$

represents the **velocity** of $\partial\mathscr{C}(t)$. Fix t and $\mathbf{x} \in \partial\mathscr{C}(t)$ and (for τ sufficiently close to t) let $\mathbf{y}(\tau)$ denote the curve that passes through $\mathbf{x}$ at time t and has

$$\mathbf{y}^{\cdot}(\tau) = \mathbf{v}(\mathbf{y}(\tau), \tau)$$

($\mathbf{y}^{\cdot}(\tau) = d\mathbf{y}(\tau)/d\tau$; we use this notation for functions of time alone). Then the **normal time-derivative** $\Phi^{\circ}(\mathbf{x}, t)$ (following $\partial\mathscr{C}(t)$) of a scalar or vector function $\Phi(\mathbf{x}, t)$ is defined by

$$\Phi^{\circ}(\mathbf{x}, t) = \frac{d}{d\tau}\Phi(\mathbf{y}(\tau), \tau)|_{\tau=t}. \tag{2.4}$$

The identities

$$\begin{aligned} &\theta^{\circ} = \boldsymbol{\ell}^{\circ} \cdot \mathbf{m} = -\mathbf{m}^{\circ} \cdot \boldsymbol{\ell} = v_s, \\ &\boldsymbol{\ell}^{\circ} = \mathbf{m}\theta^{\circ} = \mathbf{v}_s + \kappa v \boldsymbol{\ell} \end{aligned} \tag{2.5}$$

are standard.[10]

By an **interfacial chunk** we mean a smoothly evolving curve $\sigma(t)$ with $\sigma(t) \subset \partial\mathscr{C}(t)$ at each time t; we say that $\sigma(t)$ *evolves normally* if its endpoints $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ evolve with the normal velocity of the interface:

$$\mathbf{x}_1^{\cdot}(t) = \mathbf{v}(\mathbf{x}_1(t), t), \qquad \mathbf{x}_2^{\cdot}(t) = \mathbf{v}(\mathbf{x}_2(t), t). \tag{2.6}$$

For any function $\Phi(\mathbf{x}, t)$ we write

$$\int_{\partial\sigma(t)} \Phi = \Phi(\mathbf{x}_2(t), t) - \Phi(\mathbf{x}_1(t), t). \tag{2.7}$$

[10] See, e.g., [AG, eqs. (2.4), (2.6), (2.18)].

We then have the identities[11]

$$\int_{\partial\sigma(t)} \Phi = \int_{\sigma(t)} \Phi_s \, ds,$$

$$\frac{d}{dt}\int_{\sigma(t)} \Phi \, ds = \int_{\sigma(t)} (\Phi^\circ - \Phi\kappa v) \, ds. \tag{2.8}$$

For convenience, we will generally omit the argument t when writing such integrals.

3. Interfacial forces. Inertia. Balance laws.

3.1. Basic concepts. We describe the **micromechanics** of an evolving crystal $\mathscr{C}(t)$ by four functions of $\mathbf{x} \in \partial\mathscr{C}(t)$ and t:

$\mathbf{C}(\mathbf{x}, t)$ **interfacial stress,**
$\mathbf{b}(\mathbf{x}, t)$ **interaction force,**
$\rho(\mathbf{x}, t)$ (effective) **inertial density,**
$r(\mathbf{x}, t)$ **mass supply.**

$\mathbf{C}(\mathbf{x}, t)$ represents the force *within the interface* exerted across $\mathbf{x}$ at time t; if we let σ^+ and σ^-, respectively, denote right and left neighborhoods of $\mathbf{x}$ in $\partial\mathscr{C}(t)$, then $\mathbf{C}(\mathbf{x}, t)$ is the force exerted *on* σ^- *by* σ^+. Concerning the remaining functions, $\mathbf{b}(\mathbf{x}, t)$ represents the net force exerted on the interface by the bulk material of the crystal and the melt; $\rho(\mathbf{x}, t)$ gives the inertial density of the interface; $r(\mathbf{x}, t)$ represents the *rate* at which mass is supplied to the interface.

We characterize forces by the manner in which they expend power, and inertia by the manner in which it affects kinetic energy. In particular, we assume that $\mathbf{C}(\mathbf{x}, t)$ and $\mathbf{b}(\mathbf{x}, t)$ expend power through the velocity[12] $\mathbf{v}(\mathbf{x}, t)$, and that it is this velocity that is appropriate to the kinetic energy of the interface. Precisely, given any normally evolving interfacial chunk $\sigma(t)$,

$$\frac{1}{2}\int_\sigma \rho|\mathbf{v}|^2 \, ds \tag{3.1}$$

is the **kinetic energy** of $\sigma(t)$,

$$\frac{1}{2}\int_\sigma r|\mathbf{v}|^2 \, ds \tag{3.2}$$

is the **rate at which kinetic energy is supplied** to $\sigma(t)$, and

$$\int_{\partial\sigma} \mathbf{C}\cdot\mathbf{v} + \int_\sigma \mathbf{b}\cdot\mathbf{v} \, ds \tag{3.3}$$

is the **power expended** on $\sigma(t)$. We will refer to

$$\mathscr{E}(\sigma)(t) = \frac{d}{dt}\left\{\frac{1}{2}\int_\sigma \rho|\mathbf{v}|^2 \, ds\right\} - \frac{1}{2}\int_\sigma r|\mathbf{v}|^2 \, ds - \int_{\partial\sigma} \mathbf{C}\cdot\mathbf{v} - \int_\sigma \mathbf{b}\cdot\mathbf{v} \, ds \tag{3.4}$$

as the **mechanical energy production** at time t. The first law of thermodynamics requires that this quantity be balanced by the addition of heat and by changes in the internal energy.

[11] Identities $(2.8)_2$ are generally not true if the chunk does *not* evolve normally: an additional term is needed to account for the flux of Φ across $\partial\sigma(t)$ (cf., e.g., [AG, eq. (2.34)]).

[12] See [G1, eq. (1.3) and § 3].

3.2. Invariance under observer changes. Balance laws for interfacial mass and momentum. A basic assumption of our theory is that

> The mechanical energy production be invariant
> under Galilean changes in observer;

precisely, we assume that, given any normally evolving interfacial chunk $\mathcal{S}(t)$ and any time t,

$$\begin{aligned}\frac{d}{dt}\left\{\frac{1}{2}\int_{\mathcal{S}}\rho|\mathbf{v}+\mathbf{c}|^2\,ds\right\}-\frac{1}{2}\int_{\mathcal{S}}r|\mathbf{v}+\mathbf{c}|^2\,ds-\int_{\partial\mathcal{S}}\mathbf{C}\cdot(\mathbf{v}+\mathbf{c})-\int_{\mathcal{S}}\mathbf{b}\cdot(\mathbf{v}+\mathbf{c})\,ds\\ =\frac{d}{dt}\left\{\frac{1}{2}\int_{\mathcal{S}}\rho|\mathbf{v}|^2\,ds\right\}-\frac{1}{2}\int_{\mathcal{S}}r|\mathbf{v}|^2\,ds-\int_{\partial\mathcal{S}}\mathbf{C}\cdot\mathbf{v}-\int_{\mathcal{S}}\mathbf{b}\cdot\mathbf{v}\,ds\end{aligned}\tag{3.5}$$

for any constant vector $\mathbf{c}$. Here $\mathbf{c}$ is the velocity defining an arbitrary observer change, and underlying (3.5) is the tacit assumption that ρ, $\mathbf{C}$, r, and $\mathbf{b}$ be invariant, while $\mathbf{v}$ transform to $\mathbf{v}+\mathbf{c}$.

If we expand the left side of (3.5) in terms of $\mathbf{c}$, we find that

$$\mathbf{c}\cdot\left\{\frac{d}{dt}\int_{\mathcal{S}}\rho\mathbf{v}\,ds-\int_{\mathcal{S}}r\mathbf{v}\,ds-\int_{\partial\mathcal{S}}\mathbf{C}-\int_{\mathcal{S}}\mathbf{b}\,ds\right\}+\frac{1}{2}|\mathbf{c}|^2\left\{\frac{d}{dt}\int_{\mathcal{S}}\rho\,ds-\int_{\mathcal{S}}r\,ds\right\}=0;\tag{3.6}$$

since $\mathbf{c}$ is *arbitrary*, this yields **balance of mass**

$$\frac{d}{dt}\int_{\mathcal{S}}\rho\,ds=\int_{\mathcal{S}}r\,ds\tag{3.7}$$

and **balance of momentum**

$$\frac{d}{dt}\left\{\int_{\mathcal{S}}\rho\mathbf{v}\,ds\right\}-\int_{\mathcal{S}}r\mathbf{v}\,ds=\int_{\partial\mathcal{S}}\mathbf{C}+\int_{\mathcal{S}}\mathbf{b}\,ds.\tag{3.8}$$

The relations (3.7) and (3.8) are required to hold for every normally evolving interfacial chunk $\mathcal{S}(t)$; using (2.8), we have the **local balance laws:**

$$\begin{aligned}\rho^{\circ}-\rho\kappa v&=r,\\ \rho\mathbf{v}^{\circ}&=\mathbf{C}_s+\mathbf{b}.\end{aligned}\tag{3.9}$$

We have taken the normal velocity as the kinematic variable that characterizes the manner in which power is expended: tangential motion does not expend power. As is consistent with a "constraint" of this type, *we leave as indeterminate the tangential component of the interaction* $\mathbf{b}$, *and therefore concern ourselves only with the normal component of the momentum balance law* $(3.9)_2$.

Using the local balance laws (3.9) in conjunction with (2.8), we can write the mechanical energy production (3.4) in the simple form

$$\mathscr{E}(\mathcal{S})(t)=-\int_{\mathcal{S}}\mathbf{C}\cdot\mathbf{v}_s\,ds.\tag{3.10}$$

For convenience, we decompose the interfacial stress into normal and tangential components:

$$\mathbf{C}=\sigma\boldsymbol{\ell}+\xi\mathbf{m}\tag{3.11}$$

with $\sigma(\mathbf{x}, t)$ the **surface tension** (line tension) and $\xi(\mathbf{x}, t)$ the **surface shear**. Then, writing

$$b=\mathbf{b}\cdot\mathbf{m}\tag{3.12}$$

for the **normal interaction**, the normal component of $(3.9)_1$ becomes

$$\rho v^\circ = \xi_s + \sigma\kappa + b. \tag{3.13}$$

4. Constitutive equations. To state the constitutive equations that form the basis of our theory, we associate, with each evolving crystal, an **interfacial energy** ψ. As **constitutive equations** we allow the interfacial energy, the interfacial stress, the inertial density, and the normal interaction to depend on the orientation of the interface through a dependence on the angle θ, and we allow the kinetics of the interface to affect the normal interaction through a dependence on the normal velocity[13] v:

$$\psi = \psi(\theta), \quad \mathbf{C} = \mathbf{C}(\theta), \quad \rho = \rho(\theta), \quad b = b(\theta, v). \tag{4.1}$$

We assume that:[14]

(i) $\psi(\theta)$ generates the interfacial stress through the thermodynamic relation[15]

$$\mathbf{C}(\theta) = \psi(\theta)\boldsymbol{\ell}(\theta) + \psi'(\theta)\mathbf{m}(\theta); \tag{4.2}$$

(ii) The normal interaction is given by a relation of the form

$$b(\theta, v) = -F - \beta(\theta)v, \tag{4.3}$$

with F a constant;

(iii) The following inequalities hold:

$$\rho(\theta), \psi(\theta), \beta(\theta) > 0. \tag{4.4}$$

Trivially, (4.2) implies that $\sigma = \sigma(\theta)$ and $\xi = \xi(\theta)$ with

$$\sigma(\theta) = \psi(\theta), \qquad \xi(\theta) = \psi'(\theta). \tag{4.5}$$

Concerning (4.3), the constant F represents a **driving force** for the crystallization process, while $-\beta(\theta)v$ represents a "drag force" which, by $(4.4)_3$, opposes interfacial motion.[16]

A consequence of (3.9) and the constitutive relations (4.2) and (4.3) is the *balance law for energy*:[17]

$$\frac{d}{dt}\left\{\int_{\partial\mathscr{C}}\left(\frac{1}{2}\rho|\mathbf{v}|^2 + \psi\right) ds + F \text{ area } (\mathscr{C})\right\} = \frac{1}{2}\int_{\partial\mathscr{C}} r|\mathbf{v}|^2\, ds - \int_{\partial\mathscr{C}} \beta v^2\, ds, \tag{4.6}$$

with $\partial\mathscr{C} = \partial\mathscr{C}(t)$. This result allows us to identify the last term as energy dissipated during crystal growth. The derivation of (4.6) is not difficult. First of all we have the standard relation

$$\frac{d}{dt}\text{area } (\mathscr{C}) = \int_{\partial\mathscr{C}} v\, ds, \tag{4.7}$$

as well as the following consequences of (2.5), (4.2), and (4.3):

$$\begin{aligned} &\mathbf{C}\cdot\mathbf{v}_s = \psi'(\theta)\theta^\circ - \psi(\theta)v\kappa = \psi(\theta)^\circ - \psi(\theta)v\kappa, \\ &bv = -Fv + \beta(\theta)v^2. \end{aligned} \tag{4.8}$$

---

[13] Here v represents the normal velocity of the interface relative to the crystal, so that v is trivially invariant under observer changes.

[14] See [G1] and [AG], where (4.2) and (4.3) (with $\beta = \beta(\theta, v) \geqq 0$) are derived using a thermodynamic argument.

[15] $\psi'(\theta) = d\psi(\theta)/d\theta$.

[16] See [G1, Remark 4.1].

[17] Generalizing (7.6) of [AG]. The integral involving ψ over $\partial\mathscr{C}$ plus F area $(\mathscr{C})$ represents a basic Gibbsian functional for the statical theory of crystals (cf. [G2, § 3.2]).

The relation (4.6) follows from (3.4) and (3.10) with $\mathscr{s} = \partial\mathscr{C}$ (so that $\partial\mathscr{s} = \varnothing$) in conjunction with (2.3), (3.12), (4.7), and (4.8).

5. Partial differential equations. The equations of our theory are balance of mass $(3.9)_1$ and the normal component (3.13) of balance of momentum in conjunction with the constitutive relations (4.2) and (4.3). Balance of mass $(3.9)_1$ coupled with $(2.5)_1$ yield

$$\rho'(\theta)v_s - \rho(\theta)\kappa v = r. \tag{5.1}$$

The melt serves only as a source of atoms for the evolving crystal,[18] so we may regard (5.1) as defining r. On the other hand, (3.13), by virtue of $(2.5)_3$, (4.3), and (4.4), yields the **evolution equation**

$$\rho(\theta)v^\circ + \beta(\theta)v = [\psi(\theta) + \psi''(\theta)]\kappa - F, \tag{5.2}$$

which forms the basis of our theory. Note that, for an *isotropic interface*, ψ, β, and ρ are constants and (5.2) reduces to

$$\rho^\circ + \beta v = \psi\kappa - F. \tag{5.3}$$

When the crystallization process takes place in $\mathbb{R}^3$, the interface evolves as a *surface*, rather than as a curve, but apart from notation the theory is identical. Following the notation and terminology of [G1], we write ∇_{surf} for the *surface gradient*, $\mathbf{L} = -\nabla_{\text{surf}}\mathbf{m}$ for the *curvature tensor*, and $\kappa = \operatorname{tr}\mathbf{L}$ for twice the *mean curvature*. Then the only essential changes regarding the theory presented thus far are the replacement of $\psi(\theta)$ by $\psi(\mathbf{m})$, $\beta(\theta)$ by $\beta(\mathbf{m})$, $\rho(\theta)$ by $\rho(\mathbf{m})$, and (5.2) by

$$\rho(\mathbf{m})v^\circ + \beta(\mathbf{m})v = \psi(\mathbf{m})\kappa + \psi_{mm}(\mathbf{m}) \cdot \mathbf{L} - F, \tag{5.4}$$

where $\psi_{mm}(\mathbf{m})$ is the second gradient of $\psi(\mathbf{m})$ on the surface of the unit ball.

Consider now the general theory in $\mathbb{R}^2$ as defined by (5.2). *Locally*, the interface may be represented as the graph of a function $y = h(x, t)$, provided the x and y axes are chosen appropriately. Consider the choice indicated in Fig. 2 (with orientation such that arclength increase with increasing x) and let

$$p = h_x, \tag{5.5}$$

where a subscript denotes partial differentiation with respect to the corresponding variable. Then

$$p \tan\theta = -1 \tag{5.6}$$

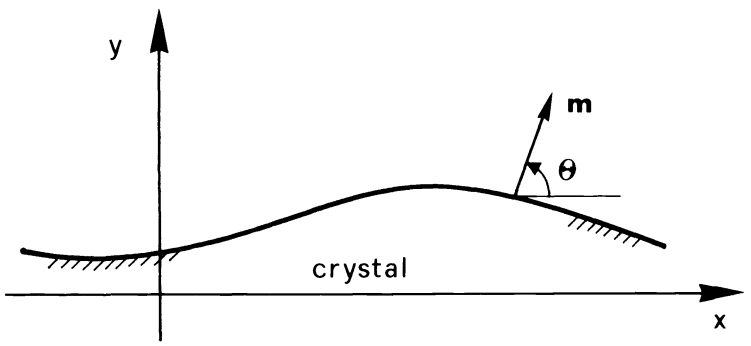


FIG. 2. *Sign conventions when the interface is a graph* $y = h(x, t)$.

[18] Reference [G4] considers a more detailed structure for the melt.

and, considering v as a function $v(x, t)$,

$$v = h_t \sin\theta, \qquad \kappa = h_{xx} \sin^3\theta, \tag{5.7}$$

$$v^\circ = v_t + v_x v \cos\theta = \sin\theta[h_{tt} + 2\sin\theta\cos\theta h_t h_{tx} + (h_t \sin\theta\cos\theta)^2 h_{xx}].$$

Thus, defining

$$B(\theta) = \beta(\theta)/\rho(\theta), \qquad G(\theta) = [\psi(\theta) + \psi''(\theta)]/\rho(\theta), \tag{5.8}$$
$$D(\theta) = F/[\rho(\theta)\sin\theta],$$

the evolution equation (5.2) takes the form

$$h_{tt} + B(\theta)h_t + 2\sin\theta\cos\theta h_t h_{tx} = \sin^2\theta[G(\theta) - (h_t\cos\theta)^2]h_{xx} - D(\theta), \tag{5.9}$$

with θ a function of h_x through (5.5) and (5.6).

Equation (5.9) is hyperbolic if and only if the discriminant Δ of the coefficients of its principal part is strictly positive. Since $\Delta = 4\sin^2\theta G(\theta)$, it is clear from (4.4) that (5.9) *is hyperbolic if and only if*

$$\psi(\theta) + \psi''(\theta) > 0. \tag{5.10}$$

We will assume that (5.10) *is satisfied for all angles of interest.*

Remark. The inequality (5.10) is essentially a condition of static stability for the interface: it follows from the requirement that straight line-segments locally minimize interfacial energy.[19] When the inequality (5.10) is reversed the partial differential equation (5.9) is elliptic and yields unstable behavior for standard initial value problems. There is no compelling physical reason to suppose that (5.10) is satisfied; in fact, material scientists often consider energies which violate (5.10) for particular ranges of θ (cf. Gjostein [G], Cahn and Hoffman [CH]). Since $\psi(\theta) > 0$ and is periodic, at worst we can have an equation which is elliptic for some but not all values of θ. Such elliptic intervals can be treated by introducing corners in the evolving crystal (cf. [AG]).

6. Some simple solutions.

6.1. Radial solutions for an isotropic crystal. In view of (5.3) and (5.4), isotropic, radially symmetric crystals in $\mathbb{R}^n$ $(n = 2, 3)$ evolve according to

$$\rho R^{\cdot\cdot}(t) + \beta R^{\cdot}(t) + (n-1)\psi R(t)^{-1} = -F, \tag{6.1}$$

with $R(t)$, the radius of the interface. Assuming that $R^{\cdot}(0) = 0$ and appealing to the phase portrait for (6.1), it is not difficult to verify that:

(i) For $F \geqq 0$, crystals melt completely in finite time;

(ii) For $F < 0$, crystals of radius $R(0) < R_{\text{crit}} := (n-1)\psi/|F|$ melt in finite time, while crystals of radius $R(0) > R_{\text{crit}}$ grow unboundedly as $t \to \infty$.

These conclusions are true whether or not $\rho = 0$ (provided we drop the initial condition $R^{\cdot}(0) = 0$ for $\rho = 0$). Furthermore, for $F < 0$ and $R(0) > R_{\text{crit}}$, $R(t)$ grows (for large t) as t when $\rho = 0$ and as t^2 when $\rho \neq 0$.

6.2. Small oscillations about a flat interface. Assume that

$$F = 0. \tag{6.2}$$

Then flat interfaces ($\kappa = 0$) describe equilibrium solutions of the general anisotropic equation (5.2). We now consider *interfacial motions which are close to equilibria of this*

[19] See Herring [H], Frank [Fr], Gjostein [G], Gruber (as referred to in [G]), Taylor [T], Fonseca [F], and Angenent and Gurtin [AG].

form. Precisely, we assume, without loss of generality, that the interface has angle $\theta = \pi/2$ at equilibrium and consider interfacial motions represented as a graph $y = h(x, t)$ with h and its derivatives "small." In view of (5.6), the angle $\theta(x, t)$, to first order in h, is given by

$$\theta = (\pi/2) + h_x. \tag{6.3}$$

Therefore (5.9) *linearized about this equilibrium* has the form

$$h_{tt} + B_0 h_t = G_0 h_{xx}, \tag{6.4}$$

where the subscript zero signifies that the corresponding quantity is to be evaluated at $\theta = \pi/2$. Equation (6.4) has the solution

$$\begin{aligned} &h(x, t) = C\, e^{i\lambda x}\, e^{-(i\omega+\zeta)t}, \\ &\omega^2 = G_0\lambda^2 - \zeta^2, \qquad \zeta = B_0/2, \end{aligned} \tag{6.5}$$

which represents *damped oscillations* of the interface.

6.3. Curvature waves advancing on a flat interface. In continuum mechanics the breakdown of solutions due to the formation of shocks (jumps in velocity) can be discussed qualitatively in terms of the blowup of waves involving jumps in acceleration.[20] An analogous discussion applies here with corners (jumps in angle) playing the role of shocks and curvature the role of acceleration.

We continue to assume that $F = 0$. We consider a front of discontinuous curvature advancing into a flat interface (with $\theta = \pi/2$). Precisely, we consider an interface described by a graph $y = h(x, t)$ and assume that there is a curve $\mathscr{K}$ in the (x, t)-plane which has the form $x = \xi(t)$ and is such that:

(W1) $h(x, t) = 0$ for $x \geqq \xi(t)$;

(W2) h and its first derivatives are continuous across $\mathscr{K}$, but second and higher derivatives of h suffer possible jump discontinuities across $\mathscr{K}$;

(W3) $[\kappa] \neq 0$.

Here and in what follows, $[g](t)$ denotes the **jump** in a function $g(x, t)$ across $\mathscr{K}$:

$$[g](t) = g(\xi(t)+0, t) - g(\xi(t)-0, t). \tag{6.6}$$

Because of (W3), we will refer to $\mathscr{K}$ as a **curvature wave**.

Standard kinematical conditions[21] give

$$[h_{tt}] = -c[h_{xt}] = c^2[h_{xx}], \tag{6.7}$$

where

$$c = \frac{d\xi}{dt} \tag{6.8}$$

is the **velocity of propagation**. Furthermore, by (W1) and (W2),

$$\theta = \pi/2 \quad \text{and} \quad h_t = h_x = 0 \quad \text{on } \mathscr{K}; \tag{6.9}$$

hence (5.6) and (5.7) yield

$$[\theta_t] = [h_{xt}], \qquad [\kappa] = [h_{xx}]. \tag{6.10}$$

[20] See, e.g., § 2 of [CG] for a discussion of such waves.

[21] See, e.g., [CG, eq. (2.5)].

As before, we write ϕ_0 for a function $\phi(\theta)$ evaluated at $\theta = \pi/2$. The jump in (5.9) across $\mathcal{K}$ then yields

$$c^2 = G_0 = (\psi + \psi'')_0/\rho_0, \tag{6.11}$$

and the *velocity of propagation is constant.*

We define the **amplitude** of the wave by

$$a = [\kappa] = [h_{xx}]; \tag{6.12}$$

a standard identity[22] then yields

$$2c^2 \frac{da}{dt} = [h_{ttt}] - c[h_{xxt}]. \tag{6.13}$$

Next, we differentiate (5.9) with respect to t and take the jump in the resulting equation; because of (6.9) and (6.10), this yields

$$[h_{ttt}] + B_0[h_{tt}] = G_0[h_{xxt}] + (G')_0[h_{xx}][h_{xt}]. \tag{6.14}$$

In deriving (6.14), we used the fact that, by (W1), $[h_{xx}\theta_t] = [h_{xx}][\theta_t]$. The relations (6.7) and (6.11)-(6.14) yield a nonlinear differential equation for the amplitude

$$\frac{2da}{dt} + B_0 a + \left[\frac{(G')_0}{c}\right] a^2 = 0. \tag{6.15}$$

For an isotropic crystal, $G(\theta)$ is independent of θ and $(G')_0 = 0$. In this case,

$$a(t) = a(0)\, e^{-B_0 t/2} \tag{6.16}$$

and curvature waves *decay.*

The results are far more interesting when the crystal is anisotropic. Assume that $(G')_0 \neq 0$. Then (6.15) has the explicit solution

$$a(t) = \frac{a(0)[1 - A]}{1 - A\, e^{B_0 t/2}}, \qquad A = 1 - \frac{\lambda}{a(0)}, \qquad \lambda = -\frac{B_0 c}{(G')_0}. \tag{6.17}$$

An elementary analysis of the solution (6.17) leads to the following conclusions:

(i) $a(0)/\lambda = 1$ implies $a(t) \equiv a(0)$;

(ii) $a(0)/\lambda < 1$ implies $a(t) \to 0$ monotonically as $t \to \infty$,

(iii) $a(0)/\lambda > 1$ implies $a(t) \to \infty$ in the finite time

$$t_\infty = -2(\ln A)/B_0. \tag{6.18}$$

The result (iii) asserts that if initially the jump in curvature is sufficiently large and of the right sign, then this jump becomes infinite in finite time, strongly suggesting that *the interface develops a corner*: $[\theta] \neq 0$. The number λ represents a *critical amplitude*: $a(t)$ grows without bound or decays to zero according as $a(0)/\lambda > 1$ or $a(0)/\lambda < 1$. This critical amplitude captures the competing effects of anisotropy and dissipation: anisotropy, measured by $(G')_0$, promotes blowup; dissipation, measured by B_0, promotes decay.

[22] See, e.g., [CG, eq. (2.10)].

REFERENCES

[AG] S. ANGENENT AND M. E. GURTIN, *Multiphase thermomechanics with interfacial structure. 2. Evolution of an isothermal interface*, Arch. Rational Mech. Anal., to appear.

[AP] A. F. ANDREEV AND A. Y. PARSHIN, *Equilibrium shape and oscillations of the surface of quantum crystals*, Soviet Phys. JETP, 48 (1978), pp. 763–766.

[B] K. A. BRAKKE, *The Motion of a Surface by Its Mean Curvature*, Princeton University Press, Princeton, NJ, 1978.

[BG] E. BEN-JACOB, N. GOLDENFELD, J. S. LANGER, AND G. SCHON, *String models of interfacial pattern formation*, Phys. D, 12 (1984), pp. 245–252.

[BK] R. BROWER, D. KESSLER, J. KOPLIK, AND H. LEVINE, *Simple models of interface growth*, Phys. D, 12 (1984), pp. 241–244.

[CG] B. D. COLEMAN AND M. E. GURTIN, *Waves in materials with memory. 2. On the growth and decay of one-dimensional acceleration waves*, Arch. Rational Mech. Anal., 19 (1965), pp. 239–265.

[CH] J. W. CAHN AND D. W. HOFFMAN, *A vector thermodynamics for anisotropic surfaces. 2. Curved and faceted surfaces*, Acta Metall. 22 (1974), pp. 1205–1214.

[F] I. FONSECA, *Interfacial energy and the Maxwell rule*, Report 88-18, Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA, 1988.

[Fr] F. C. FRANK, *The geometrical thermodynamics of surfaces*, in Metal Surfaces: Structure, Energetics, and Kinetics, Amer. Soc. Metals, Metals Park, OH, 1963.

[G] N. A. GJOSTEIN, *Adsorption and surface energy. 2. Thermal faceting from minimization of surface energy*, Acta Metall., 11 (1963), pp. 969–978.

[Gr] M. A. GRAYSON, *The heat equation shrinks embedded plane curves to points*, J. Differential Geom., 26 (1987), pp. 285–314.

[GH] M. GAGE AND R. S. HAMILTON, *The heat equation shrinking convex plane curves*, J. Differential Geom., 23 (1986), pp. 69–96.

[G1] M. E. GURTIN, *Multiphase thermomechanics with interfacial structure. 1. Heat conduction and the capillary balance law*, Arch. Rational Mech. Anal., 104 (1988), pp. 195–221.

[G2] ———, *On the two-phase Stefan problem with interfacial energy and entropy*, Arch. Rational Mech. Anal., 96 (1986), pp. 199–241.

[G3] ———, *Toward a nonequilibrium thermodynamics of two phase materials*, Arch. Rational Mech. Anal., 100 (1988), pp. 1275–1312.

[G4] ———, *A mechanical theory for crystallization of a rigid solid in a liquid melt: melting-freezing waves*, Arch. Rational Mech. Anal., 110 (1990), pp. 287–312.

[GS] M. E. GURTIN AND A. STRUTHERS, *Multiphase thermomechanics with interfacial structure. 3*, Arch. Rational Mech. Anal., to appear.

[H] C. HERRING, *Some theorems on the free energies of crystal surfaces*, Phys. Rev., 82 (1951), pp. 87–93.

[KPB] K. O. KESHISHEV, A. Y. PARSHIN, AND A. V. BABKIN, *Experimental detection of crystallization waves in* He^4, JETP Lett., 30 (1979), pp. 56–59.

[MA] H. J. MARIS AND A. F. ANDREEV, *The surface of crystalline helium* 4, Phys. Today (Feb. 1987), pp. 25–30.

[RHN] M. RENARDY, W. J. HRUSA, AND J. A. NOHEL, *Mathematical Problems in Viscoelasticity*, John Wiley, New York, 1987.

[R] R. C. ROGERS, *Notes on a linear model of mechanical phase transitions*, preprint.

[T] J. E. TAYLOR, *Crystalline variational problems*, Bull. Amer. Math. Soc., 84 (1978), pp. 568–588.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 587–601, May 1991

002

BLOWUP ASYMPTOTICS FOR THE REACTIVE-EULER GAS MODEL*

ALBERTO BRESSAN†

Abstract. This paper is concerned with a semilinear hyperbolic system modeling the behavior of a reactive gas. Because of the presence of a nonlinear reaction term, solutions become unbounded within finite time; generically, the blowup occurs at a single point. As the explosion time is approached, a precise description of the asymptotic profile of a solution is obtained by the use of a rescaled set of coordinates. Relying on a topological argument, we prove that the rescaled solution has a nontrivial, nonsingular limit, whose analytical expression depends on a finite set of parameters, determined by the initial conditions.

Key words. reactive-Euler equations, blowup, asymptotic profile

AMS(MOS) subject classification. 35L40

1. Introduction.

A semilinear hyperbolic system of the form

$$\begin{cases} \vartheta_t - aP_t = be^{\vartheta} \\ v_t + cP_x = 0 \\ v_x + dP_t = be^{\vartheta} \end{cases} \tag{1.1}$$

was first derived in [4] and, independently, in [9] as a model for the behavior of a reactive gas where viscous and diffusive effects are sufficiently weak. See also [3], [11] for a unified treatment. Here the variables P, v, and ϑ denote the perturbations from a spatially uniform equilibrium state of pressure, velocity, and temperature, respectively, while a, b, c, d are positive constants. We first study the system (1.1) in the bounded domain $I = [-1, 1]$, with initial and boundary conditions

$$P(0,x) = \bar{P}(x), \quad v(0,x) = \bar{v}(x), \quad \vartheta(0,x) = \bar{\vartheta}(x), \tag{1.2}$$

$$v(t,1) = v(t,-1) = 0. \tag{1.3}$$

Here $\bar{P}, \bar{v}, \bar{\vartheta}$ are continuous functions on I, with $\bar{v}(1) = \bar{v}(-1) = 0$. The presence in (1.1) of a reaction term which grows exponentially with the temperature allows solutions to become unbounded within finite time. Our main concern is the location of the blowup and the description of the profile of a solution as the explosion time is approached. In §3 we prove that a generic solution of (1.1)–(1.3) blows up at a single point. In §4, assuming that the initial conditions are smooth and that the initial temperature has a sufficiently large and "well-focused" maximum (in a sense to be specified later), we determine the asymptotic profile of a solution close to the blowup point. A topological argument [8, p. 278] combined with standard comparison techniques shows that, in a suitable set of rescaled coordinates, the solution has a nontrivial nonsingular limit. This asymptotic limit depends on finitely many parameters, determined by the initial conditions. For a single parabolic equation, a similar rescaling of coordinates was studied in [7] and later in [2] for an exponential nonlinearity. The present paper supplies a rigorous proof to the formal computations

* Received by the editors May 26, 1989; accepted for publication (in revised form) July 11, 1989. This research supported by National Science Foundation grant DMS 8802201.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80309.

on the reactive-Euler model which appeared in [6], [10], relative to the case where the blowup occurs in the interior of the domain I. If the blowup point lies on the boundary of I, the variable rescaling and the consequent mathematical analysis will be different [10].

2. Preliminaries. The change of variables

$$\begin{cases} z_1 = \vartheta - aP \\ z_2 = \dfrac{a}{2}\,[P + (cd)^{-1/2} v] \\ z_3 = \dfrac{a}{2}\,[P - (cd)^{-1/2} v] \end{cases} \tag{2.1}$$

transforms (1.1), (1.3) into the system

$$\begin{cases} z_{1,t} = A\, e^{z_1+z_2+z_3} \\ z_{2,t} + \lambda z_{2,x} = B\, e^{z_1+z_2+z_3} \\ z_{3,t} - \lambda z_{3,x} = B\, e^{z_1+z_2+z_3} \end{cases} \tag{2.2}$$

with boundary conditions

$$z_2(t,1) = z_3(t,1), \qquad z_2(t,-1) = z_3(t,-1). \tag{2.3}$$

Here $A = b$, $B = ab/2d$, $\lambda = (c/d)^{1/2}$. Call $E \subset \mathcal{C}(I, \mathbb{R}^3)$ the Banach space of all continuous functions $z = (z_1, z_2, z_3)$ from $[-1,1]$ into $\mathbb{R}^3$ which satisfy

$$z_2(1) = z_3(1), \qquad z_2(-1) = z_3(-1),$$

with the usual norm

$$\| z \| = \max \{|z_i(x)|;\ |x| \leq 1,\ i = 1,2,3\}.$$

For a given set of initial conditions

$$z_i(0,x) = \bar{z}_i(x), \tag{2.4}$$

with $\bar{z} \in E$, it is well known [5], [12] that the system (2.2), (2.3) has a unique forward solution z defined on some maximal interval $[0,T)$. Using the semigroup notation, we write $S^t(\bar{z}) = z(t,\cdot) \in E$ for the value of this solution at time t. Observe that either $T = +\infty$ or $\| S^t(\bar{z}) \| \to \infty$ as $t \to T^-$. In this second case, we call $T = T(\bar{z})$ the blowup time for the initial conditions (2.4). A point $\bar{x} \in [-1,1]$ is a blowup point if there exist sequences $x_n \to \bar{x}$, $t_n \to T-$ such that $|z(t_n, x_n)| \to +\infty$. The set of blowup points, given the initial data $z(0,\cdot) = \bar{z}(x)$, is denoted by $\mathcal{B}\ell(\bar{z})$. On the space E, define the ordering

$$u \preceq v \quad \text{iff} \quad u_i(x) \leq v_i(x) \qquad \forall\, x \in I, \quad i = 1,2,3.$$

By a standard comparison theorem, if $u \preceq v$, then $S^t(u) \preceq S^t(v)$ for all $t \geq 0$ at which both are defined. In particular, $u \preceq v$ implies $T(u) \geq T(v)$.

In a metric space Y, $B(y,\delta)$ denotes the open ball centered at y with radius δ. We recall that a set-valued map F is upper semicontinuous [1, p. 66] if, for all $\bar{z}$ and $\varepsilon > 0$, there exists $\delta > 0$ such that

$$z \in B(\bar{z},\delta) \Longrightarrow F(z) \subseteq B(F(\bar{z}),\varepsilon).$$

Here $B(F(\bar{z}), \varepsilon)$ denotes the neighborhood of radius ε around the set $F(\bar{z})$. If Y is a complete metric space, a subset $R \subseteq Y$ is residual in Y (i.e., of second category) if and only if its complement $Y \setminus R$ lies in the union of countably many closed nowhere dense sets. A property $\mathcal{P}$ for the elements of Y is *generic* if it holds true for all y in a residual subset of Y. In the next section we prove that the property "$\mathcal{B}\ell(\bar{z})$ is a singleton" is generic in the space E. This is a mathematically precise way of expressing the fact that for "most" initial conditions $\bar{z} \in E$ the corresponding solution of (2.2) – (2.4) blows up at a single point. All of our results will be stated for the system (2.2), which is more tractable than (1.1). Performing the change of variables inverse of (2.1), analogous results for the original system (1.1) can be easily obtained.

3. Properties of the blowup set. As a preliminary, observe that for every initial condition $\bar{z} \in E$ the solution z of (2.2), (2.3) blows up in finite time. Indeed, following [3] define

$$c^+ = \max\{A, B\}, \qquad c^- = \min\{A, B\}, \tag{3.1}$$

$$\begin{aligned} m^+ &= \max\{\bar{z}_i(x); \quad x \in I, \quad i = 1,2,3\}, \\ m^- &= \min\{\bar{z}_i(x); \quad x \in I, \quad i = 1,2,3\}. \end{aligned} \tag{3.2}$$

Comparing the components of z with the solutions of the scalar Cauchy problems

$$y' = c^{\pm} e^{3y}, \qquad y(0) = m^{\pm} \tag{3.3}$$

we deduce

$$-\tfrac{1}{3}\ln[e^{-3m^-} - 3c^- t] \le z_i(t,x) \le -\tfrac{1}{3}\ln[e^{-3m^+} - 3c^+ t] \tag{3.4}$$

for $t > 0$, $x \in I$, $i = 1,2,3$. Hence the solution z blows up at some time $T = T(\bar{z})$ with

$$(3c^+ e^{3m^+})^{-1} \le T(\bar{z}) \le (3c^- e^{3m^-})^{-1}. \tag{3.5}$$

Observe, however, that all solutions can be extended backwards for all times $t \in (-\infty, 0]$. Indeed, another simple comparison argument yields

$$m^- + c^+ e^{3m^+} t \le z_i(t,x) \le m^+ \qquad \forall\, t \le 0.$$

In this section we study the dependence of the blowup time and of the blowup set on the initial data and prove that, generically, solutions blow up at a single point.

PROPOSITION 3.1. *The map $\bar{z} \to T(\bar{z})$ is continuous on E.*

Proof. Fix $\bar{z} \in E$. If $\tau < T(\bar{z})$, then $\| S^\tau(\bar{z}) \| < \infty$ and $S^{-\tau}$ provides a homeomorphism of a bounded neighborhood V of $S^\tau(\bar{z})$ onto a neighborhood U of $\bar{z}$. This implies $T(\bar{u}) > \tau$ for all $\bar{u} \in U$, hence

$$\liminf_{\bar{u} \to \bar{z}} T(\bar{u}) \ge T(\bar{z}). \tag{3.6}$$

To prove the converse inequality, fix $\varepsilon > 0$ and call $z = z(t,x), v = v(t,x)$ the solutions of (2.2), (2.3) with initial conditions $\bar{z}, \bar{v}$, respectively. For all t, x we have

$$v_i(t,x) \ge \min\{\bar{v}_i(x); \quad |x| \le 1, \ i = 1,2,3\}. \tag{3.7}$$

Define m^- as in (3.2) and choose $\tau < T(\bar z)$ and K large enough so that

$$e^{1-K}\, e^{2-2m^-} < \varepsilon c^-, \tag{3.8}$$

$$\max\{z_i(\tau, x); \quad |x| \leq 1,\ i = 1,2,3\} > K. \tag{3.9}$$

As before, $S^{-\tau}$ provides a homeomorphism between a neighborhood V of $S^\tau(\bar z)$ and a neighborhood U of $\bar z$. By possibly shrinking U and V, it is not restrictive to assume that

$$\| \bar v - \bar z \| < 1, \qquad \| S^\tau(\bar v) - S^\tau(\bar z) \| < 1 \quad \forall\, \bar v \in U. \tag{3.10}$$

From (3.9), (3.10) we deduce

$$\max\{v_i(\tau, x); \quad |x| \leq 1,\ i = 1,2,3\} > K - 1 \qquad \forall\, \bar v \in U. \tag{3.11}$$

Assume that this maximum is $v_j(\tau, \bar x)$. Integrating the jth equation in (2.2) along the jth characteristic line through $(\tau, \bar x)$ and estimating the other components v_i, $i \neq j$ by $v_i(t,x) \geq m^- - 1$ we obtain

$$\frac{d}{dt}\, v_j(\tau + t, \bar x + \lambda_j t) \geq c^-\, e^{2m^- - 2 + v_j}. \tag{3.12}$$

Here $\lambda_1 = 0$, $\lambda_2 = \lambda$, $\lambda_3 = -\lambda$. If $j = 2,3$, the corresponding characteristic may hit the boundary of $[-1,1]$. In such cases, the estimate can be continued along the reflected characteristic, through the same boundary point. A comparison with the solution of the Cauchy problem

$$y' = c^-\, e^{2m^- - 2 + y}, \qquad y(0) = K - 1$$

yields the estimate

$$v_j(\tau + t,\ \bar x + \lambda_j t) \geq -\ln[e^{1-K} - c^-\, e^{2m^- - 2} t]. \tag{3.13}$$

By (3.13), (3.8), every solution v with initial data $\bar v \in U$ blows up within time $\tau + \varepsilon$. This implies

$$\limsup_{\bar v \to \bar z} T(\bar v) \leq T(\bar z),$$

completing the proof.

PROPOSITION 3.2. *The map $\bar z \to \mathcal{B}\ell(\bar z)$ is upper semicontinuous, compact valued, from E into $[-1,1]$.*

Proof. It is clear that each set $\mathcal{B}\ell(\bar z)$ is closed, hence compact. The upper semicontinuity of the map $\mathcal{B}\ell$ will be established by proving that for each $\bar z \in E$ and each $\bar x \notin \mathcal{B}\ell(\bar z)$ there exists $\varepsilon > 0$ and a neighborhood U of $\bar z$ such that

$$[\bar x - \varepsilon, \bar x + \varepsilon] \cap \mathcal{B}\ell(\bar v) = \phi \qquad \forall\, \bar v \in U. \tag{3.14}$$

Let $\bar z, \bar x$ be given. Since $\bar x \notin \mathcal{B}\ell(\bar z)$, there exists $\delta > 0$ and L such that the solution $z = z(t,x)$ of (2.2) – (2.4) satisfies

$$z_i(t,x) < L - 1 \qquad \forall\, i = 1,2,3, \quad |x - \bar x| \leq \delta, \quad 0 \leq t < T(\bar z). \tag{3.15}$$

Set $\varepsilon = \min\{e^{-3L}/7c^+, \delta/3\lambda\}$, with c^+ as in (3.1), and define $\tau = T(\bar z) - \varepsilon$. Then $S^{-\tau}$ provides a homeomorphism from a neighborhood V of $S^\tau(\bar z)$ onto a neighborhood U of $\bar z$. By possibly shrinking U and V, in view of (3.15) and of Proposition 3.1, we can assume that

$$v_i(\tau, x) < L \qquad \forall\, i = 1,2,3, \quad |x - \bar x| \le \delta, \tag{3.16}$$

for all solutions v with initial conditions $\bar v \in U$, and that

$$T(\bar v) < T(\bar z) + \varepsilon \qquad \forall\, \bar v \in U. \tag{3.17}$$

Comparing the values of v_i with the solution of the Cauchy problem

$$y' = c^+ e^{3y}, \qquad y(\tau) = L \tag{3.18}$$

from (3.16) it follows that

$$v_i(t,x) \le -\tfrac{1}{3} \ln(3c^+(\tau - t) + e^{-3L}) \tag{3.19}$$

for $t \in [\tau, T(\bar v)]$, $|x - \bar x| \le \delta - \lambda(t - \tau)$. Since $T(\bar v) < \tau + 2\varepsilon$, (3.19) and the choice of ε imply that, for every initial condition $\bar v \in U$, the corresponding blowup set does not contain the interval $[\bar x - \varepsilon, \bar x + \varepsilon]$.

THEOREM 3.3. *The set E^* of all $\bar z \in E$ such that $\mathcal{B}\ell(\bar z)$ consists of a single point is residual in E.*

Proof. For every integer $m \ge 1$ define the set

$$A_m = \{\bar z \in E, \quad \operatorname{diam} \mathcal{B}\ell(\bar z) \ge m^{-1}\},$$

where $\operatorname{diam} V = \sup\{|x - y|;\ x, y \in V\}$ denotes the diameter of the set V. Clearly $E^* = E \setminus \bigcup_{m \ge 1} A_m$. It thus suffices to show that each A_m is closed and nowhere dense.

To see that A_m is closed, assume $v_n \in A_m$ for all $n \ge 1$, $v_n \to u$. Let $x_n, y_n \in \mathcal{B}\ell(v_n)$ with $|x_n - y_n| \ge m^{-1}$. Choosing a suitable subsequence, we can assume that $x_{n'} \to \bar x$, $y_{n'} \to \bar y$. By Proposition 3.2, $\bar x, \bar y \in \mathcal{B}\ell(u)$, hence $\operatorname{diam} \mathcal{B}\ell(u) \ge |\bar x - \bar y| \ge m^{-1}$ and $u \in A_m$.

It remains to prove that the complement A_m^c of A_m is everywhere dense. Fix $u \in E$ and let $\bar x \in \mathcal{B}\ell(u)$. Define $\varepsilon = 1/9m$, $\delta = \varepsilon/\lambda$. Let $\varphi : [-1,1] \to [0,1]$ be a continuous function such that $\varphi(x) = 1$ if $|x - \bar x| \le 2\varepsilon$, $\varphi(x) = 0$ if $|x - \bar x| \ge 3\varepsilon$. For each $n \ge 1$ define the functions $v_n = (v_{n1}, v_{n2}, v_{n3})$ and $\bar w_n$ by setting

$$v_{ni}(x) = u_i(T(u) - \delta, x) + n^{-1}\varphi(x), \qquad \bar w_n = S^{-T(u)+\delta}(v_n).$$

Clearly $v_n \to S^{T(u)-\delta}(u)$ and $\bar w_n \to u$ as $n \to \infty$. We claim that $\bar w_n \notin A_m$ for every $n \ge 1$. Call $w_n(t,x)$ the solution of (2.2), (2.3) with initial conditions $\bar w_n(x)$. Then

$$w_n(T(u) - \delta, x) = v_n(x) > u(T(u) - \delta, x)$$

for each $n \ge 1$, $|x - \bar x| \le 3\varepsilon$. By the continuity of u with respect to time, there exists $\delta' \in (0, \delta)$ such that

$$w_{ni}(T(u) - \delta, x) \ge u_i(T(u) - \delta + \delta', x) \tag{3.20}$$

for $|x - \bar{x}| \leq 2\varepsilon$. From (3.20) we deduce

$$w_{ni}(t,x) \geq u_i(t+\delta', x) \tag{3.21}$$

whenever $t \geq T(u) - \delta$, $|x - \bar{x}| \leq 2\varepsilon - \lambda(t - T(u) + \delta)$.

Since $\bar{x}$ is a blowup point for u, (3.21) implies $T(\bar{w}_n) \leq T(u) - \delta'$. All blowup points of $\bar{w}_n$, however, are contained inside the set $\{x;\ |x - \bar{x}| \leq 4\varepsilon\}$, because $w_n(t,x) = u(t,x)$ whenever $|x - \bar{x}| > 3\varepsilon + \lambda(t - T(u) + \delta)$, $T(u) - \delta \leq t < T(\bar{w}_n)$. Hence $\operatorname{diam} \mathcal{B}\ell(\bar{w}_n) > m^{-1}$ and $\bar{w}_n \notin A_m$, for all $n \geq 1$.

4. Asymptotic estimates. The aim of this section is to describe the asymptotic profile of a solution of (2.2) as the blowup time is approached. Since what matters here is just the local behavior, for simplicity we neglect the boundary conditions (2.3) and work with the Banach space $\mathcal{C}^3(\mathbb{R}, \mathbb{R}^3)$ of functions $z = (z_1, z_2, z_3)$ which are three times continuously differentiable on $\mathbb{R}$. Furthermore, we assume $A + 2B = 1$. This condition is clearly not restrictive, because it can always be achieved by the time rescaling $t' = (A + 2B)t$.

THEOREM 4.1. *There exists a nonempty open set $U \subset \mathcal{C}^3(\mathbb{R}, \mathbb{R}^3)$ such that, if $\bar{z} \in U$, the corresponding solution z of (2.2) – (2.4) blows up at an isolated point x_0. Moreover, if $T = T(\bar{z})$ is the blowup time, there exist constants $\Omega < 0$ and z_i^∞ such that $z_1^\infty + z_2^\infty + z_3^\infty = 0$, with the following property. For every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\begin{aligned} &\left|z_1(t,x) - z_1^\infty + A \ln\left[(T-t) - \tfrac{\Omega}{2}(x - x_0)^2\right]\right| < \varepsilon, \\ &\left|z_i(t,x) - z_i^\infty + B \ln\left[(T-t) - \tfrac{\Omega}{2}(x - x_0)^2\right]\right| < \varepsilon \end{aligned} \tag{4.1}$$

$(i = 2, 3)$, whenever $T - \delta \leq t < T$, $|x - x_0| < \delta$.

Further estimates can be easily deduced from (4.1). For example, the temperature $\vartheta = z_1 + z_2 + z_3$ satisfies

$$\lim_{t \to T^-} \left[\vartheta(t,x) + \ln\left((T-t) - \frac{\Omega}{2}(x - x_0)^2\right)\right] = 0$$

uniformly on every domain of the form $|x| \leq \psi(T - t)$, ψ being any continuous function with $\psi(0) = 0$. Moreover, for $x \neq x_0$ the limit values of $z_i(t,x)$ as $t \to T^-$ are well defined. In particular, the final temperature profile $\vartheta(T, \cdot)$ satisfies

$$\lim_{x \to x_0} \left[\vartheta(T,x) + 2\ln|x - x_0| + \ln(-\Omega/2)\right] = 0. \tag{4.2}$$

A more precise version of the above theorem will be actually proved, providing computable conditions on the initial data $\bar{z}$ which imply (4.1). As a preliminary, observe that, given any $\alpha, \beta, \gamma > 0$, we can always find $\tilde{\tau}$ large enough so that the following conditions hold.

(C1) $e^{\tilde{\tau}/6} > \max\{2\lambda,\ 12\beta\lambda/\alpha,\ 4\gamma\lambda,\ e^{1/2}\}$,

(C2) The solution (μ, m, M, D) of the Cauchy problem

$$\begin{cases} \dot{\mu} = -4me^{-\tau/6} - \mu/2 & \mu(\tilde{\tau}) = \gamma/2, \\ \dot{m} = 2me^{-\tau/3} & m(\tilde{\tau}) = -\beta/2, \\ \dot{M} = 8m^2 e^{-\tau/3} & M(\tilde{\tau}) = -2\alpha, \\ \dot{D} = 2\beta + 8\beta^2 e^{-\tau/3} - D & D(\tilde{\tau}) = 2\beta, \end{cases} \tag{4.3}$$

satisfies

$$\mu(\tau) < \gamma, \quad m(\tau) > -\beta, \quad M(\tau) < -\alpha, \quad D(\tau) < 3\beta \qquad \forall \tau \geq \tilde{\tau}. \tag{4.4}$$

THEOREM 4.2. *Let $\alpha, \beta, \gamma, \tilde{\tau}$ be constants for which the conditions* (C1), (C2) *hold. Let $\bar{z} = (\bar{z}_1, \bar{z}_3, \bar{z}_3)$ be a $\mathcal{C}^3$ function such that $\bar{\vartheta} = \bar{z}_1 + \bar{z}_2 + \bar{z}_3$ attains a local maximum $\hat{\vartheta} > \tilde{\tau} + 1$ at a point $\hat{x}$. In addition, assume that*

(i) $\max |(\bar{z}_i)_x| < (\gamma/2) e^{(\hat{\vartheta}-1)/2}$,

(ii) $\max |(\bar{z}_i)_{xx}| < 2\beta\, e^{\hat{\vartheta}-1}$,

(iii) $-e^{\hat{\vartheta}-1}\beta/2 < \sum_i \min (z_i)_{xx} \leq \sum_i \max (z_i)_{xx} < -2\alpha\, e^{\hat{\vartheta}+1}$,

the max *and* min *being all taken over the set $|x - \hat{x}| \leq 2e^{2(1-\hat{\vartheta})/3}$. Then the solution z of* (2.2), (2.4) *blows up at an isolated point x_0, at some time T, with*

$$|\hat{x} - x_0| \leq e^{2(1-\hat{\vartheta})/3}, \qquad e^{-1-\hat{\vartheta}} \leq T \leq e^{1-\hat{\vartheta}}.$$

Moreover, the estimates (4.1) *hold with $-\beta \leq \Omega \leq -\alpha$.*

We outline here the main arguments in the proof, while the details will be worked out in §§5 and 6. Since all solutions of (2.2) have components which remain uniformly bounded from below, the blowup occurs precisely when their sum $\vartheta = z_1 + z_2 + z_3$ becomes unbounded. If ϑ blows up at $t = T$, $x = x_0$, we expect that a nontrivial, nonsingular asymptotic limit can be obtained using the rescaled variables

$$\tau = -\ln(T - t), \qquad \eta = \frac{x - x_0}{\sqrt{T - t}}, \tag{4.5}$$

$$\begin{cases} u = z_1 + A \cdot \ln(T - t), \\ v = z_2 + B \cdot \ln(T - t), \\ w = z_3 + B \cdot \ln(T - t), \end{cases} \tag{4.6}$$

$$S = u + v + w = \vartheta + \ln(T - t). \tag{4.7}$$

In these new variables, (2.2) takes the form

$$\begin{cases} u_\tau + \dfrac{\eta}{2} u_\eta = A(e^S - 1) \\ v_\tau + \left(\dfrac{\eta}{2} + \lambda e^{-\tau/2}\right) v_\eta = B(e^S - 1) \\ w_\tau + \left(\dfrac{\eta}{2} - \lambda e^{-\tau/2}\right) w_\eta = B(e^S - 1), \end{cases} \tag{4.8}$$

$$S_\tau + \frac{\eta}{2} S_\eta + \lambda e^{-\tau/2}(v_\eta - w_\eta) = e^S - 1. \tag{4.9}$$

The new initial conditions, at $\tau_0 = -\ln T$ are

$$u(\tau_0, \eta) = \bar{u}(\eta) = \bar{z}_1(x_0 + \eta T^{1/2}), \quad v(\tau_0, \eta) = \bar{v}(\eta) = \bar{z}_2(x_0 + \eta T^{1/2}),$$

$$w(\tau_0, \eta) = \bar{w}(\eta) = \bar{z}_3(x_0 + \eta T^{1/2}). \tag{4.10}$$

We remark, however, that the exact values of T and x_0 are now known a priori. Instead, we know the point $\hat{x}$ where the maximum value $\hat{\vartheta}$ of $\vartheta = z_1 + z_2 + z_3$ is initially attained. Of course, we expect $x_0 \simeq \hat{x}$ and $T \simeq e^{-\hat{\vartheta}}$, but equality need not hold, in general. For this reason, we have to consider a two-parameter family of transformations depending on T, x_0. Define the set

$$\mathcal{G} = \{(T, x_0);\ e^{-1-\hat{\vartheta}} \leq T \leq e^{1-\hat{\vartheta}},\ |\hat{x} - x_0| \leq e^{2(1-\hat{\vartheta})/3}\}. \tag{4.11}$$

Elements of $\mathcal{G}$ will be our "guesses" for the exact time and location of the blowup. For each fixed $(T, x_0) \in \mathcal{G}$, call $\xi(\tau)$ the point at which the rescaled variable $S(\tau, \cdot)$ attains its maximum. Assuming that $S_{\eta\eta}$ is negative at ξ, from (4.9) and the relations

$$S_\eta(\tau, \xi(\tau)) \equiv 0, \quad S_{\eta\tau}(\tau, \xi(\tau)) + S_{\eta\eta}(\tau, \xi(\tau))\dot{\xi}(\tau) \equiv 0, \tag{4.12}$$

$$S_{\eta\tau} + \frac{\eta}{2} S_{\eta\eta} + \frac{1}{2} S_\eta + \lambda e^{-\tau/2}(v_{\eta\eta} - w_{\eta\eta}) = e^S S_\eta, \tag{4.13}$$

we obtain

$$\frac{d\xi}{d\tau} = -\frac{S_{\eta\tau}(\tau, \xi)}{S_{\eta\eta}(\tau, \xi)} = \frac{\xi(\tau)}{2} + \frac{\lambda}{S_{\eta\eta}} e^{-\tau/2}(v_{\eta\eta} - w_{\eta\eta}), \tag{4.14}$$

$$\frac{d}{d\tau} S(\tau, \xi(\tau)) = S_\tau + S_\eta \dot{\xi} = e^{S(\tau,\xi)} - 1 - \lambda e^{-\tau/2}(v_\eta(\tau, \xi) - w_n(\tau, \xi)). \tag{4.15}$$

The initial conditions for $\xi, S(\xi)$ are

$$\xi(\tau_0) = (\hat{x} - x_0)T^{-1/2}, \qquad S(\xi(\tau_0)) = \hat{\vartheta} + \ln T. \tag{4.16}$$

In $\mathbb{R}^3$, define the tube with shrinking square section

$$\mathcal{T} = \{(\tau, \xi, S); |\xi|, |S| \leq e^{-\tau/3}\}. \tag{4.17}$$

As (T, x_0) range in $\mathcal{G}$, (4.14) – (4.16) determine a two-parameter family of trajectories $\tau \to (\xi(\tau), S(\tau, \xi(\tau)))$ depending continuously on T, x_0. We will show that these trajectories are well defined (in particular $S_{\eta\eta}(\tau, \xi(\tau)) < 0$) as long as they remain inside $\mathcal{T}$. Moreover, all boundary points of $\mathcal{T}$ are strict exit points. A topological argument thus implies the existence of some $T, x_0 \in \mathcal{G}$ whose corresponding solution of (4.14) – (4.16) remains forever inside $\mathcal{T}$. Such T and x_0 provide the exact timing and location of the blowup. In the second part of the proof we refine the estimates on u, v, w, and on their derivatives, and establish the limits (4.1) by means of a comparison theorem.

5. Proof of Theorem 4.2. Step 1. For $(T, x_0) \in \mathcal{G}$, let u, v, w, S be the corresponding solution of (4.8) – (4.10), defined for $\tau \geq \tau_0 = -\ln T$. Introduce the scalar quantities

$$\mu(\tau) = \max\{|u_\eta(\tau, \eta)|,\ |v_\eta(\tau, \eta)|,\ |w_\eta(\tau, \eta)|;\quad |\eta| \leq e^{-\tau/6}\} \tag{5.1}$$

$$\begin{cases} m^u(\tau) = \min u_{\eta\eta}(\tau, \eta) \\ m^v(\tau) = \min v_{\eta\eta}(\tau, \eta) \\ m^w(\tau) = \min w_{\eta\eta}(\tau, \eta) \end{cases} \qquad \begin{cases} M^u(\tau) = \max u_{\eta\eta}(\tau, \eta) \\ M^v(\tau) = \max v_{\eta\eta}(\tau, \eta) \\ M^w(\tau) = \max w_{\eta\eta}(\tau, \eta), \end{cases}$$

$$D(\tau) = \max\{|u_{\eta\eta}(\tau,\eta)|,\ |v_{\eta\eta}(\tau,\eta)|,\ |w_{\eta\eta}(\tau,\eta)|\}, \tag{5.2}$$

where the min and max are all taken over the set $|\eta| \le e^{-\tau/6}$. Defining

$$m = m^u + m^v + m^w, \qquad M = M^u + M^v + M^w, \tag{5.3}$$

we clearly have

$$m(\tau) \le S_{\eta\eta}(\tau,\eta) \le M(\tau) \tag{5.4}$$

whenever $|\eta| \le e^{-\tau/6}$. Estimates on the initial values of μ, m, M, and D can be derived from assumptions (i) – (iii) in Theorem 4.2. Recalling that

$$\hat\vartheta + 1 \ge \tau_0 = -\ln T \ge \hat\vartheta - 1 > \tilde\tau > 3, \tag{5.5}$$

if $|\eta| \le e^{-\tau_0/6}$, then

$$|x - x_0| = |\eta|\, e^{-\tau_0/2} \le e^{-2\tau_0/3} \le e^{2(1-\hat\vartheta)/3},$$

$$|x - \hat x| \le |x - x_0| + |x_0 - \hat x| \le 2\, e^{2(1-\hat\vartheta)/3}.$$

Hypotheses (i) – (iii) therefore imply

$$\begin{cases} \mu(\tau_0) < (\gamma/2)^{(\hat\vartheta-1)/2} \cdot e^{-\tau_0/2} \le \gamma/2 \\ m(\tau_0) > [-e^{\hat\vartheta-1} \cdot \beta/2] \cdot e^{-\tau_0} \ge -\beta/2 \\ M(\tau_0) < [-2\alpha\, e^{\hat\vartheta+1}] e^{-\tau_0} \le -2\alpha \\ D(\tau_0) < [e^{\hat\vartheta-1}\beta/2] e^{-\tau_0} \le \beta/2. \end{cases} \tag{5.6}$$

To estimate μ, m, M, D when $\tau > \tau_0$, we differentiate (4.8) and obtain

$$\begin{cases} u_{\eta\tau} + \dfrac{\eta}{2} u_{\eta\eta} + \dfrac{u_\eta}{2} = A\, e^S S_\eta \\ v_{\eta\tau} + \dfrac{\eta}{2} v_{\eta\eta} + \dfrac{v_\eta}{2} = B\, e^S S_\eta - \lambda\, e^{-\tau/2} v_{\eta\eta} \\ w_{\eta\tau} + \dfrac{\eta}{2} w_{\eta\eta} + \dfrac{w_\eta}{2} = B\, e^S S_\eta + \lambda\, e^{-\tau/2} w_{\eta\eta}, \end{cases} \tag{5.7}$$

$$\begin{cases} u_{\eta\eta\tau} + u_{\eta\eta} + \dfrac{\eta}{2} u_{\eta\eta\eta} = A\, e^S (S_\eta^2 + S_{\eta\eta}) \\ v_{\eta\eta\tau} + v_{\eta\eta} + \dfrac{\eta}{2} v_{\eta\eta\eta} = B\, e^S (S_\eta^2 + S_{\eta\eta}) - \lambda\, e^{-\tau/2} v_{\eta\eta\eta} \\ w_{\eta\eta\tau} + w_{\eta\eta} + \dfrac{\eta}{2} w_{\eta\eta\eta} = B\, e^S (S_\eta^2 + S_{\eta\eta}) + \lambda\, e^{-\tau/2} w_{\eta\eta\eta}. \end{cases} \tag{5.8}$$

Observe that the three families of characteristics for the system (4.8) are determined by the equations

$$\dot\eta = \frac{\eta}{2}, \quad \dot\eta = \frac{\eta}{2} + \lambda\, e^{-\tau/2}, \quad \dot\eta = \frac{\eta}{2} - \lambda\, e^{-\tau/2}.$$

By (5.5) and assumption (C1) on $\tilde{\tau}$, when $\tau \geq \tilde{\tau}$ all characteristics are flowing out from the domain $\{(\tau, \eta); |\eta| \leq e^{-\tau/6}\}$. Assume that $(\tau, \xi(\tau), S(\xi(\tau))) \in \mathcal{T}$ and $m(\tau) \leq M(\tau) < 0$ for τ in some initial interval $[\tau_0, \tau')$. Then the following estimates hold.

$$S(\tau, \eta) \leq e^{-\tau/3}, \qquad e^{S(\tau,\eta)} \leq 1 + 2\,e^{-\tau/3} \leq 2, \tag{5.9}$$

$$|S_\eta(\tau, \eta)| \leq |\eta - \xi(\tau)||m(\tau)| \leq 2|m(\tau)|e^{-\tau/6}. \tag{5.10}$$

Using (5.9), (5.10), from (5.7), (5.8) we deduce

$$\begin{aligned} \dot{\mu}(\tau) &\leq -\mu(\tau)/2 + \max\{|S_\eta(\tau, \eta)| \cdot e^{S(\tau,\eta)};\ |\eta| \leq e^{-\tau/6}\} \\ &\leq -\mu(\tau)/2 + 4|m(\tau)|e^{-\tau/6}, \end{aligned} \tag{5.11}$$

$$\begin{cases} \dot{m}^u(\tau) \geq -m^u(\tau) + A \cdot m(\tau)(1 + 2e^{-\tau/3}) \\ \dot{m}^v(\tau) \geq -m^v(\tau) + B \cdot m(\tau)(1 + 2e^{-\tau/3}) \\ \dot{m}^w(\tau) \geq -m^w(\tau) + B \cdot m(\tau)(1 + 2e^{-\tau/3}), \end{cases}$$

$$\dot{m}(\tau) \geq 2e^{-\tau/3} m(\tau), \tag{5.12}$$

$$\begin{cases} \dot{M}^u(\tau) \leq -M^u(\tau) + 2A(2|m(\tau)|e^{-\tau/6})^2 + A(1 + 2e^{-\tau/3})M(\tau) \\ \dot{M}^v(\tau) \leq -M^v(\tau) + 2B(2|m(\tau)|e^{-\tau/6})^2 + B(1 + 2e^{-\tau/3})M(\tau) \\ \dot{M}^w(\tau) \leq -M^w(\tau) + 2B(2|m(\tau)|e^{-\tau/6})^2 + B(1 + 2e^{-\tau/3})M(\tau), \end{cases}$$

$$\dot{M}(\tau) \leq 8m^2(\tau)e^{-\tau/3}. \tag{5.13}$$

Assumptions (C1), (C2) now imply

$$\mu(\tau) < \gamma, \quad m(\tau) > -\beta, \quad M(\tau) < -\alpha \tag{5.14}$$

for all $\tau \in [\tau_0, \tau')$. Since $S_{\eta\eta} < M(\tau) < 0$, (5.14) shows that solutions (ξ, S) of (4.14), (4.15) are well defined and depend continuously on the parameters T, x_0, as long as (ξ, S) remain inside the tube $\mathcal{T}$. Moreover, from (5.8) it follows

$$\begin{aligned} \dot{D}(\tau) &\leq -D(\tau) + 2\max\{(S_\eta^2 + |S_{\eta\eta}|);\ |\eta| \leq e^{-\tau/6}\} \\ &\leq -D(\tau) + 2[(2|m(\tau)|e^{-\tau/6})^2 + |m(\tau)|] \\ &\leq -D(\tau) + 8\beta^2 e^{-\tau/3} + 2\beta, \end{aligned} \tag{5.15}$$

hence, by condition (C2),

$$D(\tau) < 3\beta \qquad \forall\, \tau \in [\tau_0, \tau'). \tag{5.16}$$

We now define a continuous map φ from the unit square $Q = \{(y_1, y_2);\ |y_1|, |y_2| \leq 1\}$ onto its boundary as follows. Given $(y_1, y_2) \in Q$, there exists a unique $(T, x_0) \in \mathcal{G}$ such that, setting $\tau_0 = -\ln T$, we have

$$(y_1, y_2) = e^{\tau_0/3}(\xi(\tau_0), \qquad S(\xi(\tau_0))) = T^{-1/3}((\hat{x} - x_0)T^{-1/2},\ \hat{\vartheta} + \ln T). \tag{5.17}$$

If the corresponding solution of (4.14) – (4.16) escapes through the boundary of $\mathcal{T}$ at time τ, define

$$\varphi(y_1, y_2) = e^{\tau/3}(\xi(\tau), \qquad S(\tau, \xi(\tau))) \in \partial Q. \tag{5.18}$$

As customary, the continuity of φ can be proved by showing that every boundary point of $\mathcal{T}$ with $\tau \geq \tilde{\tau}$ is a strict exit point for solutions of (4.14) – (4.16). Indeed, if $|\xi(\tau)| = e^{-\tau/3}$, using (5.16) in (4.14) we obtain

$$\begin{aligned} \frac{d}{d\tau}|\xi| &\geq \frac{1}{2}e^{-\tau/3} - \frac{\lambda}{|M(\tau)|}e^{-\tau/3} \cdot 2D(\tau) \\ &\geq \frac{1}{2}e^{-\tau/3} - \frac{6\beta\lambda}{\alpha}e^{-\tau/2} > 0 \end{aligned}$$

because of (C1). On the other hand, if $|S(\tau, \xi(\tau))| = e^{-\tau/3}$, using (5.14) in (4.15) we now have

$$\frac{d}{d\tau}|S(\tau, \xi(\tau))| \geq \frac{1}{2}e^{-\tau/3} - \lambda e^{-\tau/2} \cdot 2\mu(\tau) > \frac{1}{2}e^{-\tau/3} - 2\lambda\gamma e^{-\tau/2} > 0.$$

If every choice $(T, x_0) \in \mathcal{G}$ yields a trajectory which escapes from $\mathcal{T}$ at some finite time τ, the map φ would then be a continuous surjection from Q onto ∂Q which leaves the points of ∂Q fixed. Since no such map exists, we have proved the existence of some $(T, x_0) \in \mathcal{G}$ whose corresponding trajectory $\tau \to (\xi(\tau), S(\tau, \xi(\tau)))$ remains inside $\mathcal{T}$ for all times $\tau \geq \tau_0$.

6. Proof of Theorem 4.2. Step 2. From now on, everything is referred to a unique coordinate transformation, i.e., the one determined by the element $(T, x_0) \in \mathcal{G}$ singled out at the end of §5. Differentiating (5.8) once again we find

$$\begin{cases} u_{\eta\eta\eta\tau} + \frac{3}{2}u_{\eta\eta\eta} + \frac{\eta}{2}u_{\eta\eta\eta\eta} = Ae^S S_{\eta\eta\eta} + Ae^S S_\eta(S_\eta^2 + 3S_{\eta\eta}) \\ v_{\eta\eta\eta\tau} + \frac{3}{2}v_{\eta\eta\eta} + \left(\frac{\eta}{2} + \lambda e^{-\tau/2}\right)v_{\eta\eta\eta\eta} = Be^S S_{\eta\eta\eta} + Be^S S_\eta(S_\eta^2 + 3S_{\eta\eta}) \\ w_{\eta\eta\eta\tau} + \frac{3}{2}w_{\eta\eta\eta} + \left(\frac{\eta}{2} - \lambda e^{-\tau/2}\right)w_{\eta\eta\eta\eta} = Be^S S_{\eta\eta\eta} + Be^S S_\eta(S_\eta^2 + 3S_{\eta\eta}). \end{cases} \tag{6.1}$$

Call $E^u(\tau)$ the maximum of $|u_{\eta\eta\eta}(\tau, \cdot)|$ over the set $|\eta| \leq e^{-\tau/6}$, and define $E^v(\tau)$, $E^w(\tau)$ similarly. Set $E = E^u + E^v + E^w$. Since $|S_{\eta\eta\eta}(\tau, \eta)| \leq E(\tau)$, by (6.1) E satisfies the differential inequality

$$\dot{E}(\tau) \leq \left(-\frac{3}{2} + \max_\eta e^{S(\tau,\eta)}\right)E(\tau) + \max_\eta |S_\eta| \cdot |e^S S_\eta^2 + 3e^S S_{\eta\eta}|. \tag{6.2}$$

The estimates (5.10) and (5.14) together imply

$$\lim_{\tau\to\infty} \max\{|S_\eta(\tau, \eta)|; |\eta| \leq e^{-\tau/6}\} = 0. \tag{6.3}$$

Using (6.3) and the uniform bounds on $S, |S_{\eta\eta}|$ in (6.2) we obtain

$$\lim_{\tau\to\infty} E(\tau) = 0,$$

therefore, recalling the definitions (5.3),

$$\lim_{\tau\to\infty} |M(\tau) - m(\tau)| = 0. \tag{6.4}$$

By (5.4), (5.14), from (6.4) it follows that

$$\Omega = \lim_{\substack{\tau\to\infty \\ |\eta|\leq e^{-\tau/6}}} S_{\eta\eta}(\tau,\eta) \tag{6.5}$$

exists and satisfies

$$-\beta \leq \Omega \leq -\alpha. \tag{6.6}$$

Using (6.5) in (5.8) we now obtain

$$\begin{cases} \lim_{\tau\to\infty} u_{\eta\eta} = A\Omega, \\ \lim_{\tau\to\infty} v_{\eta\eta} = \lim_{\tau\to\infty} w_{\eta\eta} = B\Omega, \end{cases} \tag{6.7}$$

all limits being taken inside the region $|\eta| \leq e^{-\tau/6}$. The equations (4.14), (4.15) in view of (6.5), (6.7) yield

$$|\xi(\tau)| \leq C\, e^{-\tau/2}, \qquad |S(\tau,\xi(\tau))| \leq C_1\, e^{-\tau/2} \tag{6.8}$$

for some constant C_1 and all τ large enough. Since $S_\eta(\tau,\xi(\tau)) \equiv 0$ and $|S_{\eta\eta}| < \beta$, (6.8) implies an estimate of the form

$$|S_\eta(\tau,\eta)| \leq C_2 e^{-\tau/2} \tag{6.9}$$

whenever $|\eta| \leq (C+2\lambda)e^{-\tau/2}$. In particular, (6.8) and (6.9) yield

$$|S(\tau,0)| \leq C_3 e^{-\tau/2}, \tag{6.10}$$

valid for some constant C_3 and all τ large enough. Define

$$\mu'(\tau) = \max\{|u_\eta(\tau,\eta)|,\ |v_\eta(\tau,\eta)|,\ |w_\eta(\tau,\eta)|;\ |\eta| \leq 2\lambda\, e^{-\tau/2}\}.$$

An estimate entirely similar to (5.11) now yields

$$\begin{aligned} \dot\mu' &\leq -\mu'(\tau)/2 + \max\{|S_\eta(\tau,\eta)|e^{S(\tau,\eta)};\ |\eta| \leq 2\lambda\, e^{-\tau/2}\} \\ &\leq -\mu'(\tau)/2 + 2C_2\, e^{-\tau/2}, \end{aligned}$$

which implies, in particular,

$$|u_\eta(\tau,0)|, \quad |v_\eta(\tau,0)|, \quad |w_\eta(\tau,0)| \leq C_4\tau\, e^{-\tau/2}. \tag{6.11}$$

To extend our estimates beyond the narrow strip $|\eta| \leq 2\lambda\, e^{-\tau/2}$ we rely on a comparison technique.

LEMMA. *Let (u,v,w) be a solution of* (5.2) *and let $Z = Z(\tau,\eta)$ be a scalar function such that*

$$Z_\tau + \frac{\eta}{2} Z_\eta = e^Z - 1. \tag{6.12}$$

Call

$$\begin{aligned} u^-(\tau,\eta) &= \min\{u(\tau,\eta');|\eta'-\eta| \leq 2\lambda\, e^{-\tau/2}\}, \\ u^+(\tau,\eta) &= \max\{u(\tau,\eta');|\eta'-\eta| \leq 2\lambda\, e^{-\tau/2}\} \end{aligned}$$

and similarly for v^-, v^+, w^-, w^+. *If at some point* (τ, η) *we have* $(u^-+v^-+w^-)(\tau,\eta) \geq Z(\tau, \eta)$, *then*

$$(u^- + v^- + w^-)(\tau + t,\ \eta e^{t/2}) \geq Z(\tau + t,\ \eta e^{t/2}) \qquad \forall t > 0. \tag{6.13}$$

On the other hand, if $(u^+ + v^+ + w^+)(\tau,\eta) \leq Z(\tau,\eta)$, *then*

$$(u^+ + v^+ + w^+)(\tau + t,\ \eta e^{t/2}) \leq Z(\tau + t,\ \eta e^{t/2}) \qquad \forall t > 0. \tag{6.14}$$

Indeed, (6.13) and (6.14) can be proved by defining

$$Y^{\pm}(t) = (u^{\pm} + v^{\pm} + w^{\pm})(\tau + t,\ \eta e^{t/2})$$

and checking that

$$\frac{dY^-(t)}{dt} \geq \frac{d}{dt} Z(\tau + t,\ \eta e^{t/2}),$$

$$\frac{dY^+(t)}{dt} \leq \frac{d}{dt} Z(\tau + t,\ \eta e^{t/2})$$

whenever $Y^- \geq Z$ or $Y^+ \leq Z$, respectively.

Using the lemma, we now prove that as $\tau \to \infty$ the functions u, v, w, S converge to some limit, uniformly on bounded sets. Fix $\varepsilon > 0$ and define

$$Z^+(\tau,\eta) = -\ln\left(1 - \frac{\Omega + \varepsilon}{2}\eta^2\right),$$

$$Z^-(\tau,\eta) = -\ln\left(1 - \frac{\Omega - \varepsilon}{2}\eta^2\right).$$

Observe that Z^+, Z^- are time-invariant solutions of (6.12), with $Z^+_{\eta\eta}(0) = \Omega + \varepsilon$, $Z^-_{\eta\eta}(0) = \Omega - \varepsilon$. For any $\varepsilon' > 0$, by (6.7) there exists τ_1 so large that

$$\begin{cases} u(\tau,\eta) \leq u(\tau,0) + u_\eta(\tau,0)\cdot\eta + (A\Omega + \varepsilon')\eta^2/2 \\ v(\tau,\eta) \leq v(\tau,0) + v_\eta(\tau,0)\cdot\eta + (B\Omega + \varepsilon')\eta^2/2 \\ w(\tau,\eta) \leq w(\tau,0) + w_\eta(\tau,0)\cdot\eta + (B\Omega + \varepsilon')\eta^2/2 \end{cases} \tag{6.15}$$

whenever $|\eta| \leq e^{-\tau/6}$, $\tau \geq \tau_1$.

Call $u_\infty, v_\infty, w_\infty$, respectively, the limits of $u(\tau,0), v(\tau,0), w(\tau,0)$ as $\tau \to \infty$. Using (6.10), (6.11) in (4.8), it is clear that these limits exist. In fact, we have the estimates

$$|u(\tau,0) - u_\infty|,\ |v(\tau,0) - v_\infty|, |w(\tau,0) - w_\infty| \leq C_5 \tau e^{-\tau/2} \tag{6.16}$$

for some constant C_5 and all τ large enough. Moreover,

$$u_\infty + v_\infty + w_\infty = \lim_{\tau\to\infty} S(\tau,0) = 0.$$

By (6.15), (6.16), there exists $\bar\tau$ so large that

$$u^+ - u_\infty \leq AZ^+, \quad v^+ - v_\infty \leq BZ^+, \quad w^+ - w_\infty \leq BZ^+, \tag{6.17}$$

at every point of the set $\{(\tau,\eta);\ \tau\geq\bar\tau,\ |\eta|=\frac12\,e^{-\tau/6}\}$. In particular,

$$(u^+ + v^+ + w^+)(\tau,\pm e^{-\tau/6}) \leq Z^+\left(\tau,\pm\tfrac12\,e^{-\tau/6}\right) \tag{6.18}$$

for all $\tau\geq\bar\tau$. The previous lemma now yields

$$S(\tau,\eta)\leq (u^+ + v^+ + w^+)(\tau,\eta)\leq Z^+(\tau,\eta) \tag{6.19}$$

on the region

$$\Sigma_{\bar\tau}=\left\{(\tau,\eta);\ \tau\geq\bar\tau,\ \tfrac12\,e^{-\tau/6}\leq|\eta|\leq\tfrac12\,e^{(\tau-\bar\tau)/2}\cdot e^{-\bar\tau/6}\right\}.$$

For $\bar\tau$ suitably large, an entirely similar argument yields

$$S(\tau,\eta)\geq (u^- + v^- + x^-)(\tau,\eta)\geq Z^-(\tau,\eta) \tag{6.20}$$

for all $(\tau,\eta)\in\Sigma_{\bar\tau}$. Since $S(\tau,\eta)\to 0$ as $\tau\to\infty$ on the strip $|\eta|\leq\frac12\,e^{-\tau/2}$, (6.19) and (6.20) prove the following. For every $\varepsilon>0$ there exists $\bar\tau$ large enough so that

$$\left|S(\tau,\eta)+\ln\left(1-\frac{\Omega}{2}\,\eta^2\right)\right|<\varepsilon \tag{6.21}$$

for every (τ,η) in the region

$$R_{\bar\tau}=\left\{(\tau,\eta);\ \tau\geq\bar\tau,\ |\eta|\leq\tfrac12\,e^{\tau/2}\,e^{-2\bar\tau/3}\right\}.$$

Using (6.19) and (6.20) in (4.8), another comparison argument shows that on $\Sigma_{\bar\tau}$ the functions $u^\pm$ satisfy

$$A\cdot Z^-(\eta)\leq u^-(\tau,\eta)-u_\infty\leq u^+(\tau,\eta)-u_\infty\leq A\cdot Z^+(\eta).$$

Similarly, the functions $v^\pm-v_\infty$ and $w^\pm-w_\infty$ are bounded by $B\cdot Z^\pm$. On the strip $|\eta|\leq\frac12\,e^{-\tau/6}$, estimates on u,v,w are already known. Therefore we conclude that for every $\varepsilon>0$ there exists $\bar\tau$ such that

$$\begin{aligned}
&\left|u(\tau,\eta)-u_\infty+A\cdot\ln\left(1-\frac{\Omega}{2}\,\eta^2\right)\right|<\varepsilon,\\
&\left|v(\tau,\eta)-v_\infty+B\cdot\ln\left(1-\frac{\Omega}{2}\,\eta^2\right)\right|<\varepsilon,\\
&\left|w(\tau,\eta)-w_\infty+B\cdot\ln\left(1-\frac{\Omega}{2}\,\eta^2\right)\right|<\varepsilon,
\end{aligned} \tag{6.22}$$

for every $(\tau,\eta)\in R_{\bar\tau}$. By setting $z_1^\infty=u_\infty$, $z_2^\infty=v_\infty$, $z_3^\infty=w_\infty$, the reinterpretation of (6.22) in the original variables t,x,z is the following. For every $\varepsilon>0$ there exists $\bar t<T$ such that the estimates (4.1) hold whenever $\bar t\leq t<T$, $|x-x_0|\leq\frac12\,(T-t)^{2/3}$. This proves Theorem 4.2.

The statements in Theorem 4.1 now follow as corollaries. Indeed, for any given $\alpha,\beta,\gamma,\tilde\tau$, the hypotheses in Theorem 4.2 are satisfied by a nonempty, open set of functions. The previous proof also indicates that the parameters T,x_0,z_i^∞, and Ω, which characterize the self-similar blowup, depend continuously on the initial data $\bar z$ in the $\mathcal{C}^2$ topology.

REFERENCES

[1] J. P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[2] J. BEBERNES, A. BRESSAN, AND D. EBERLY, *A description of blowup for the solid fuel ignition model*, Indiana Univ. Math. J., 36 (1987), pp. 295–305.

[3] J. BEBERNES AND D. KASSOY, *Reactive-Euler induction models*, Proc. of the Sixth Army Conference on Applied Math. & Computing, ARO Report 89-1, pp. 473-482.

[4] J. F. CLARKE AND R. S. CANT, *Nonsteady gasdynamic effects in the induction domain behind a strong shock wave*, Dynamics of Flames and Reactive Systems, Progress in Aeronautics and Astronautics, 95 (1985), pp. 142–163.

[5] R. COURANT AND D. HILBERT, *Methods of mathematical physics*, Vol. II, Wiley Interscience, New York, 1962.

[6] J. W. DOLD, *Dynamic transition of a self-igniting region*, In Mathematical Modeling in Combustion and Related Topics, C.-M. Brauner and C. Schmidt-Lainé, eds., NATO ASI series, M. Nijhoff, 1988, pp. 461–470.

[7] Y. GIGA AND R. KOHN, *Asymptotically self-similar blow-up of semilinear heat equations*, Comm. Pure Appl. Math., 29 (1980), pp. 21–36.

[8] P. HARTMAN, *Ordinary Differential Equations*, Birkhäuser, Boston, 1982.

[9] T. L. JACKSON AND A. K. KAPILA, *Shock-induced thermal runaway*, SIAM J. Appl. Math., 45 (1985), pp. 130–137.

[10] T. L. JACKSON, A. K. KAPILA, AND D. S. STEWART, *Evolution of a reaction center in an explosive material*, Univ. of Illinois T. & A.M. Report, No. 484, 1987.

[11] D. R. KASSOY, A. K. KAPILA, AND D. S. STEWART, *A unified formulation for diffusive and nondiffusive thermal explosion theory*, Combust. Sci. Tech., to appear.

[12] R. H. MARTIN, JR., *Nonlinear Operators and Differential Equations in Banach Spaces*, John Wiley, New York, 1976.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 602–622, May 1991

003

AN INTERIOR DISCONTINUITY OF A NONLINEAR ELLIPTIC-HYPERBOLIC SYSTEM*

SENHUEI E. CHEN† AND R. BRUCE KELLOGG‡

Abstract. A nonlinear elliptic-hyperbolic system of partial differential equations which is a simplified form of the equations of viscous, compressible, barotropic flow at steady-state is studied. A boundary value problem for the system on a strip $D = (0, a) \times (-\infty, \infty)$ is considered. Zero boundary conditions for the velocities u and v on the sides of the strip D are imposed, and for pressure $p(0, y) = p_0(y)$ is imposed, where $p_0(y)$ has a jump at $y = 0$. Jump conditions for the system show that u and v are continuous. However, their derivatives and pressure have a curve of discontinuity. With sufficient small width of the strip D the Schauder fixed-point theorem gives a solution with a curve of discontinuity. The results suggest that there are two-dimensional, viscous, barotropic, steady-state compressible flows with discontinuity.

Key words. viscous compressible flow, discontinuous solutions

AMS(MOS) subject classifications. 76N10, 35Q10

1. Introduction.

The mathematical theory of the compressible Navier–Stokes equations is far from complete, and there are many unsolved questions. Among these questions is the interior regularity of the solutions. There are some situations where it is quite clear what to expect. For example, in the case of compressible inviscid flow, the governing system of equations is hyperbolic and solutions will, in general, contain shock discontinuities. In the case of incompressible viscous flow, the governing system of equations may be thought of as elliptic, and it is reasonable to expect that the solutions will be regular interior to the flow region. In the case of viscous compressible flow, the governing system of equations is neither elliptic nor hyperbolic, and the issue of interior regularity is open. In this paper, we make a contribution to this problem.

There has been some work related to the problem of interior regularity of the solutions to the compressible flow equations. D. Hoff [1], [2] has considered the time-dependent, viscous, compressible flow equations in one space dimension. He shows that there are indeed solutions for which the pressure has a jump discontinuity; however, he also shows that this discontinuity decays as time goes to infinity. In [4] Valli considers the three-dimensional, steady-state, compressible flow equations. He shows that if the boundary data is small and smooth, then there is a smooth solution to the system. Kellogg [6] studies the two-dimensional, steady-state, viscous, compressible flow equations. The equations are linearized around an ambient nonzero solution to the nonlinear system. He shows that there is a solution to this linearized system that contains an interior discontinuity. The pressure takes a jump across this curve. The curve is a streamline of the ambient flow field that emanates from a jump in the specified boundary values of the pressure.

From the above discussion it seems possible that there are solutions to the steady-state, viscous, compressible flow equations which have an interior curve of discontinuity. We are unable to prove this result. Instead, we construct a simplified system of equations that still retains the "elliptic-hyperbolic" character of the original flow equations. Our simplification enables us to carry through an existence theorem for a

* Received by the editors June 12, 1989; accepted for publication (in revised form) May 14, 1990.
† Department of Mathematics, Howard University, Washington, D.C. 20059
‡ Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. Supported in part by the Army Research Office.

discontinuous solution, which is the purpose of this paper.

Let the primary unknowns be the velocity $u(x,y)$, $v(x,y)$ in the x, y directions, respectively, and the pressure $p(x,y)$, and let $\rho(p)$ be the density, which is related to the pressure through the equation of state. We consider the case of barotropic flow, in which there is no dependence on temperature. Let μ and λ denote the two coefficients of viscosity. The two-dimensional, barotropic, steady-state, compressible Navier–Stokes equations are:

$$-(2\mu+\lambda)u_{xx}-(\mu+\lambda)v_{xy}-\mu u_{yy}+(\rho u^2)_x+(\rho uv)_y+p_x=0, \tag{1.1a}$$

$$-\mu v_{xx}-(\mu+\lambda)u_{xy}-(2\mu+\lambda)v_{yy}+(\rho uv)_x+(\rho v^2)_y+p_y=0, \tag{1.1b}$$

$$(\rho u)_x+(\rho v)_y=0. \tag{1.1c}$$

To obtain the simplified system we set $\mu=-\lambda=1$ and we drop the convective terms in (1.1a,b). More crucially, we replace the continuity equation by a modified "continuity" equation, $u\rho_x+v\rho_y=0$. Thus, we have dropped the term $\rho u_x+\rho v_y$ from the continuity equation. This modified "continuity" equation has no physical meaning. Our only justification in considering it is that the elliptic-hyperbolic character of the system is unchanged, and that we are able to demonstrate the existence of a solution of the modified system with an interior discontinuity. The difficulties encountered in using (1.1c) are explained in §2. With the assumption that the system is barotropic, the density $\rho=\rho(p)$ is a function of pressure only and the system can be simplified further. Writing $\rho_x=(d\rho/dp)p_x$, $\rho_y=(d\rho/dp)p_y$, we obtain from (1.1c) the equation $up_x+vp_y=0$. We make a further notational change. The solution that we obtain will be such that the flow is close to uniform flow in the x direction. We therefore replace the unknown u by $1+u$. This leaves the modified momentum equations unchanged. The modified compressible flow equations studied in this paper are therefore

$$-u_{xx}-u_{yy}+p_x=0, \tag{1.2a}$$

$$-v_{xx}-v_{yy}+p_y=0, \tag{1.2b}$$

$$(1+u)p_x+vp_y=0. \tag{1.2c}$$

The system (1.2) will be considered in the strip $D=(0,a)\times(-\infty,\infty)$. The width a will be chosen later. We impose zero boundary conditions for u and v on the sides of the strip D, and we impose the boundary condition for the pressure on the side of the strip D through which the flow enters. Thus, we are led to the boundary conditions:

$$u(0,y)=u(a,y)=0, \tag{1.2d}$$

$$v(0,y)=v(a,y)=0, \tag{1.2e}$$

$$p(0,y)=p_0(y). \tag{1.2f}$$

The function $p_0(y)$ will be chosen to have a simple jump discontinuity at $y = 0$, to vanish outside $[0, a]$, and to be smooth for $y \geq 0$. We let $\delta p_0 = p_0(+0)$ denote the jump in $p_0(y)$ at $y = 0$. We will show that the system (1.2) has a solution with the property that $p(x, y)$ is discontinuous across a curve $\mathcal{C}$. The curve $\mathcal{C}$ is the streamline of the flow emanating from the point (0,0), which is the point of discontinuity of p_0. The first derivatives of u and v undergo jump discontinuities across $\mathcal{C}$, and satisfy certain jump conditions analogous to the Rankine–Hugoniot conditions for inviscid flow. The proof reformulates the problem as a fixed-point mapping T in a Banach space and uses the Schauder fixed-point theorem. In the course of the proof, we must choose a to be suitably small. The specific restriction on a is given below, and is used to guarantee that T is a self-mapping of a ball in the Banach space. Thus, both the width of the strip D and the support of p_0 are made small. On the other hand, there is no restriction on the size of the jump δp_0.

In §2 we define precisely what is meant by a weak solution of (1.2), we reformulate (1.2) as finding the fixed point of a map in a certain Banach space, and we state our existence theorem (Theorem 2.1). In addition, we derive in a heuristic way the jump conditions that the solution satisfies across the curve $\mathcal{C}$.

Section 3 presents a formula for the solution of the momentum equations (1.2a,b) in terms of certain weakly singular integrals. From this formula we obtain a modulus of continuity of u and v in the y variable of the form $r(|\ln|r||+1)$, which is just enough to ensure the unique solvability of the characteristic equations associated with (1.2c). Also in this section, we verify the jump conditions for a weak solution of (1.2), and we derive a formula for the solution of (1.2c).

In §4 an analysis of the interaction between the weakly singular integrals occurring in the solution of (1.2a,b) and the characteristic equation occurring in the solution of (1.2c) proves the continuity of the fixed-point mapping T. Following that, some a priori inequalities are derived.

In §5, a value of the width of the strip D is determined, and the Schauder fixed-point theorem (see, for example, [3]), is applied to give the existence of a solution with a curve $\mathcal{C}$ of discontinuity.

2. Existence theorem and jump conditions. In this section we define precisely what we mean by a weak solution of the system (1.2). We then state our main existence theorem as Theorem 2.1, whose proof is given in §5. The existence theorem establishes the existence of a weak solution (u, v, p) with the property that p has a curve $\mathcal{C}$ of discontinuity. The weak solution must satisfy certain jump conditions across the curve $\mathcal{C}$ of discontinuity, and we end this section with a discussion and heuristic derivation of these jump conditions.

Let $C^0(\bar{D})$ be the space of continuous functions in $\bar{D}$. Let

$$\mathcal{A} = \{(u, v) : u \in C^0(\bar{D}), v \in C^0(\bar{D}), ||u|| + ||v|| < \infty\},$$

where

$$||u|| = \sup_{x,y \in D} |u(x, y)| + \sup_{x,y,\bar{y} \in D, y \neq \bar{y}} \frac{|u(x, y) - u(x, \bar{y})|}{|y - \bar{y}|(|\ln|y - \bar{y}|| + 1)},$$

and similarly for $||v||$. It is easy to see that $\mathcal{A}$ is a Banach space with the norm $||(u, v)|| = ||u|| + ||v||$. The particular form of the norm $||u||$ reflects the elliptic-hyperbolic character of our system. The modulus of continuity of u in the y variable is $r[|\ln|r|| + 1]$. This modulus of continuity is just enough to guarantee the solvability

of the hyperbolic equation (1.2c), see the beginning of §3, and to provide the a priori estimates that arise from this solution (note Lemmas 4.1 and 4.2). On the other hand, this modulus of continuity is provided by estimates for the weakly singular integrals that occur in the solution of the elliptic equations (1.2a,b), and it seems that no better modulus of continuity arises from these weakly singular integrals.

For system (1.2), we say that (u, v, p) is a weak solution, if u and v belong to $\mathcal{A}$, p is a locally integrable function in D, and (u, v, p) satisfy the following conditions: for any $\phi(x, y) \in C_0^\infty(D)$, the equations

$$\int\int_D [-u\phi_{xx} - u\phi_{yy} - p\phi_x]dxdy = 0, \tag{2.1a}$$

$$\int\int_D [-v\phi_{xx} - v\phi_{yy} - p\phi_y]dxdy = 0 \tag{2.1b}$$

hold; on each curve $y = h(x, y_0)$ defined by the equation $h'(x, y_0) = v(x, h(x, y_0))/(1+u(x, h(x, y_0)))$, $h(0, y_0) = y_0$, p is constant and hence satisfies (1.2c). As mentioned above, since $(u, v) \in \mathcal{A}$, the function $h(x, y_0)$ is well defined. We now state the main existence theorem.

THEOREM 2.1. *Let p_0 be a function with $p_0(y) = 0$ for $y < 0$ and $y \geq a$, $p_0(y)$ infinitely differentiable for $y \geq 0$, and $\lim_{y \to +0} p_0(y) = \delta p_0 \neq 0$. Then there is a number $a^* > 0$, satisfying* (R1), (R2), *and* (R3) *defined in* §§4 *and* 5 *such that, if $0 < a < a^*$, the system* (1.2) *has a weak solution (u, v, p) with $(u, v) \in \mathcal{A}$. The solution has a curve C of discontinuity on which the jump conditions* (J1), (J2), *and* (J3) *defined below are satisfied.*

Let $\mathcal{B}$ be the closed ball in $\mathcal{A}$ with center 0 and radius 1/4. We define a mapping T on $\mathcal{B}$ as follows. Let $T_1 : (u, v) \to p$, where p is obtained by solving

$$(1 + u)p_x + vp_y = 0, \qquad p(0, y) = p_0(y),$$

and let $T_2 : p \to (\bar{u}, \bar{v})$, where $\bar{u}$, $\bar{v}$ satisfy

$$-\bar{u}_{xx} - \bar{u}_{yy} + p_x = 0, \qquad \bar{u}(0, y) = \bar{u}(a, y) = 0,$$

$$-\bar{v}_{xx} - \bar{v}_{yy} + p_y = 0, \qquad \bar{v}(0, y) = \bar{v}(a, y) = 0.$$

Let $T = T_2 \circ T_1$ be the composition of T_2 and T_1. Theorem 2.1 will be proved in §5 by showing that the mapping T has a fixed point in $\mathcal{B}$.

We may now explain the difficulties encountered in using the true continuity equation (1.1c), instead of (1.2c). Presumably, T_1 should be the solution operator of (1.1c). When written out, (1.1c) becomes $u\rho_x + v\rho_y + (u_x + v_y)\rho = 0$; if this equation were used, the space $\mathcal{A}$ would need to be replaced by a space of functions for which $u_x + v_y$ has meaning. Alternately, we could define a solution operator T_1 of (1.1c) by means of a weak solution. If this is the case, it is hard to get enough information concerning ρ to establish continuity results of the form given in Propositions 4.3 and 4.7.

Let $\Gamma = (x(s), y(s))$ be a smooth curve, parametrized by s, which divides D into two parts, D_1 and D_2. The unit normal vector to the curve is $n(s) = (-\dot{y}(s), \dot{x}(s))$. Suppose $n(s)$ points into D_2, so $n(s)$ is the outward pointing normal to D_1. In each

subregion we suppose that u, v, and p are smooth, but they or their derivatives may have a simple jump across Γ. Let u_t denote the tangential derivative along the curve Γ. Let $[u]_1$, $[u]_2$, $[u_x]_1$, $[u_x]_2$, etc., denote the one-sided limits of u, u_x, etc., on Γ. We use the notation $\delta u = [u]_2 - [u]_1$, $\delta u_x = [u_x]_2 - [u_x]_1$ to measure the jump. In the following theorem, we present jump conditions for solutions with a curve Γ of discontinuity. The proof uses the standard technique of integration by parts, but assumes that the solution is smooth on either side of Γ. In Theorem 3.1 these jump conditions are established for a weak solution.

THEOREM 2.2. *Let the domain D be divided into two parts D_1, D_2, by a smooth curve Γ. Let (u, v, p) be a weak solution of* (1.2). *Suppose u, v, p are smooth in each subdomain with one-sided limits on Γ, and suppose p has a jump on Γ. Then Γ is a streamline of the flow, the first derivatives of u and v have jumps on Γ, and u, v, p satisfy the jump conditions*

$$\delta u = \delta v = \delta u_t = \delta v_t = 0, \tag{J1}$$

$$\dot{y}\delta u_x - \dot{x}\delta u_y = \dot{y}\delta p, \tag{J2}$$

$$\dot{y}\delta v_x - \dot{x}\delta v_y = -\dot{x}\delta p. \tag{J3}$$

Proof. Let γ denote a curve that crosses Γ at a point q_0. Let $q \in \gamma$, and let Γ_q be the streamline through q. Then $p = p(q)$ on Γ_q. As q varies over $\gamma \cap D_1$, $p(q)$ is continuous, and as p varies over $\gamma \cap D_2$, $p(q)$ is continuous. Hence, for p to be discontinuous across Γ, we must have $\gamma = \Gamma_{q_0}$; that is, Γ is a streamline.

Since $(u, v) \in \mathcal{A}$, u and v are continuous. Hence $\delta u = \delta v = 0$, and since u and v are assumed to be smooth on either side of Γ, $\delta u_t = \delta v_t = 0$, and we have (J1).

Let ϕ be a smooth function of compact support contained in D_1. From a well-known property of the Laplacian operator and the fact that p is smooth in D_1, (2.1a) implies that u satisfies (1.2a) in D_1. Similarly, u satisfies (1.2a) in D_2. Next, let ϕ have support which intersects Γ. Applying (2.1a), we write the integral as a sum of integrals over D_1 and D_2. Integrating by parts and using (1.2a) in D_1 and D_2, we obtain

$$\int_\Gamma [\dot{y}\delta u_x - \dot{x}\delta u_y - \dot{y}\delta p + \delta u(\dot{y}\delta p_x - \dot{x}\delta p_y)]\phi ds = 0.$$

Since this is true for all smooth ϕ with compact support, and since $\delta u = 0$, we obtain (J2). The proof of (J3) is similar.

3. Properties of solutions. The aim of this section is to derive formulas for the mappings T_1 andT_2. These formula are then used to verify jump conditions and derive a formula for the modulus of continuity for a pair $(\bar{u}, \bar{v})$ in the range of T_2. First, we derive a formula for T_1 by solving (1.2c) and (1.2f) for p with (u, v) a given pair in $\mathcal{A}$. The characteristic curves $y = h(x, y_0)$ corresponding to the hyperbolic equation (1.2c) are those defined by the solutions of the nonlinear ordinary differential equation:

$$\frac{dh}{dx} = \frac{v(x, h(x, y_0))}{1 + u(x, h(x, y_0))}, \qquad h(0, y_0) = y_0. \tag{3.1}$$

In particular, when $y_0 = 0$, we obtain the characteristic curve $y = H(x)$ emanating from $(0,0)$:

$$\frac{dH}{dx} = \frac{v(x,H(x))}{1+u(x,H(x))}, \qquad H(0) = 0. \tag{3.2}$$

For the existence and uniqueness of solution of (3.1) and (3.2) we refer to Hartman [5]. Since u and v are continuous, the Peano existence theorem [5, Chap. II, Thm. 2.1] shows that solutions $h(x,y_0)$ and $H(x)$ exist. Since u and v are in $\mathcal{A}$, the Osgood criterion [5, Chap. III, Cor. 6.2] is satisfied and the solutions h and H are unique. As a consequence of this uniqueness, $h(x,y_0)$ is a strictly increasing function in y_0, so the inverse function exists. For each x, let the inverse function of $h(x,y_0)$ be $g(x,y)$, i.e.,

$$g(x,h(x,y_0)) = y_0, \qquad g(x,h(x,0)) = g(x,H(x)) = 0. \tag{3.3}$$

Through $g(x,y)$, the solution $p = T_1(u,v)$ can be expressed as

$$p(x,y) = p_0(g(x,y)).$$

In particular,

$$p(x,H(x)) = p_0(g(x,H(x))) = p_0(+0) = \delta p_0. \tag{3.4}$$

Thus, p is constant along each curve $y = h(x,y_0)$. Due to the jump discontinuity of $p_0(y)$ at $y = 0$, $p(x,y)$ has a jump discontinuity across the characteristic $y = H(x)$, that is, the one emanating from $(0,0)$ and denoted by $\mathcal{C}$. Moreover $p(x,y) = 0$, if $y < H(x)$, $p(x,y)$ is continuous and compactly supported in $y \geq H(x)$.

Now we are in position to solve (1.2a,b) subject to (1.2d,e) for $\bar{u}(x,y)$ and $\bar{v}(x,y)$ with given $p(x,y)$ in the range of T_1. We will take the Fourier transform of (1.2a,b) with respect to y. Since $p(x,y)$, and the first derivatives of $\bar{u}(x,y)$ and $\bar{v}(x,y)$ possibly are discontinuous on the curve $y = H(x)$, the usual formulas for the Fourier transform of the derivatives must be modified and the transformed equations include terms containing these jumps. When we apply the jump conditions (J1) and (J2), these extra terms cancel out and we obtain two linear ordinary differential equations in x with coefficients in t:

$$-\hat{u}_{xx} + t^2\hat{u} + \hat{p}_x = 0,$$

$$-\hat{v}_{xx} + t^2\hat{v} + it\hat{p} = 0,$$

where $\hat{u}$, $\hat{v}$, and $\hat{p}$ denote the Fourier transform of $\bar{u}$, $\bar{v}$, $\bar{p}$, and t denotes the transformed variable. These equations may be solved subject to boundary conditions.

For this, let

$$k_1(x-s,y-t) = \frac{-\sin \pi(x-s)/a}{8a(\sinh^2 \pi(y-t)/2a + \sin^2 \pi(x-s)/2a)},$$

$$k_2(x+s,y-t) = \frac{-\sin \pi(x+s)/a}{8a(\sinh^2 \pi(y-t)/2a + \sin^2 \pi(x+s)/2a)},$$

$$k_3(x-s,y-t)=\frac{-\sinh\pi(y-t)/a}{8a(\sinh^2\pi(y-t)/2a+\sin^2\pi(x-s)/2a)},$$

$$k_4(x+s,y-t)=\frac{-\sinh\pi(y-t)/a}{8a(\sinh^2\pi(y-t)/2a+\sin^2\pi(x+s)/2a)}.$$

Then, taking the inverse Fourier transform of the solutions and using some formulas in [7], we obtain the following formula for $(\bar{u},\bar{v})=T_2(p)$:

$$\begin{aligned}\bar{u}(x,y)&=\int\int_D p(s,t)(k_1(x-s,y-t)+k_2(x+s,y-t))dtds\\&\equiv\bar{u}_1(x,y)+\bar{u}_2(x,y),\end{aligned}\tag{3.5}$$

$$\begin{aligned}\bar{v}(x,y)&=\int\int_D p(s,t)(k_3(x-s,y-t)+k_4(x+s,y-t))dtds\\&\equiv\bar{v}_1(x,y)+\bar{v}_2(x,y).\end{aligned}\tag{3.6}$$

For some analysis in §4, it is convenient to introduce

$$k_7(x-s,y-t)=\frac{s-x}{2\pi((y-t)^2+(x-s)^2)}.$$

We may write $\bar{u}(x,y)$ as

$$\bar{u}(x,y)=[\bar{u}_1(x,y)-\bar{u}_3(x,y)]+\bar{u}_2(x,y)+\bar{u}_3(x,y),\tag{3.7}$$

where

$$\bar{u}_3(x,y)\equiv\int\int_D p(s,t)k_7(x-s,y-t)dtds.\tag{3.8}$$

Now we verify the jump conditions (J1), (J2), and (J3) for a pair $(\bar{u},\bar{v})$ in the range of T_2.

THEOREM 3.1. *Let u, v be in $\mathcal{A}$, and p the image of u, v under the mapping T_1, then $(\bar{u},\bar{v})=T_2(p)$ satisfies the jump conditions* (J1), (J2), *and* (J3).

Proof. Since $k_2(x+s,y-t)$ is C^1 as a function of x and y and bounded for all (s,t) in D, (x,y) in D, $\bar{u}_2(x,y)$ is C^1 in x and y. Hence $\bar{u}_2(x,y)$ takes no jump across $y=H(x)$. Using the fact that $p(s,t)$ is supported in $t\geq H(s)$ and making a change of variables $y-t=\tau$, $x-s=\sigma$ yields

$$\bar{u}_1(x,y)=\int_{x-a}^{x}\int_{-\infty}^{y-H(x-\sigma)}p(x-\sigma,y-\tau)k_1(\sigma,\tau)d\tau d\sigma.$$

We use this expression to evaluate the partial derivatives of $\bar{u}_1(x,y)$:

$$\begin{aligned}\frac{\partial\bar{u}_1}{\partial x}&=\int_{-\infty}^{y}p(0,y-\tau)k_1(x,\tau)d\tau-\int_{-\infty}^{y-H(a)}p(a,y-\tau)k_1(x-a,\tau)d\tau\\&\quad-\int_{x-a}^{x}H'(x-\sigma)p(x-\sigma,H(x-\sigma))k_1(\sigma,y-H(x-\sigma))d\sigma\\&\quad+\int_{x-a}^{x}\int_{-\infty}^{y-H(x-\sigma)}p_{,1}(x-\sigma,y-\tau)k_1(\sigma,\tau)d\tau d\sigma\end{aligned}$$

$$\equiv A_1 + A_2 + A_3 + A_4,$$

$$\frac{\partial \bar{u}_1}{\partial y} = \int_{x-a}^{x} \int_{-\infty}^{y-H(x-\sigma)} p_{,2}(x-\sigma, y-\tau) k_1(\sigma,\tau) d\tau d\sigma + \int_{x-a}^{x} p(x-\sigma, H(x-\sigma)) k_1(\sigma, y-H(x-\sigma)) d\sigma \equiv A_5 + A_6.$$

Here $p_{,i}$ means the partial derivative of p with respect to the ith variable. Terms A_1, A_2, A_4, A_5 are continuous across $y = H(x)$. However, A_3 and A_6 become singular as x, y approaches the curve $\mathcal{C}$, and these terms give rise to the jump discontinuity in $\partial \bar{u}_1/\partial x$ and $\partial \bar{u}_1/\partial y$. Using (3.4) we have

$$\begin{aligned}
A_3(x,y) &= -\int_0^a H'(s) p(s, H(s)) k_1(x-s, y-H(s)) ds \\
&= -\delta p_0 \int_0^a H'(s) k_1(x-s, y-H(s)) ds \\
&= -\delta p_0 \int_0^a (H'(s) - H'(x)) k_1(x-s, y-H(s)) ds \\
&\quad -\delta p_0 H'(x) \int_0^a \left(k_1(x-s, y-H(s)) - \frac{1}{2\pi} k_5(x-s, y-H(s)) \right) ds \\
&\quad \frac{-\delta p_0}{2\pi} H'(x) \int_0^a \left(k_5(x-s, y-H(s)) - k_6(x-s, y-H(s)) \right) ds \\
&\quad \frac{-\delta p_0}{2\pi} H'(x) \int_0^a k_6(x-s, y-H(s)) ds,
\end{aligned}$$

where

$$k_5(x-s, y-H(s)) = \frac{(s-x)}{((y-H(s))^2 + (x-s)^2)},$$

$$k_6(x-s, y-H(s)) = \frac{(s-x)}{(y-H(x)-H'(x)(s-x))^2 + (x-s)^2}.$$

Since $(H'(s) - H'(x))k_1$, $k_1 - k_5$, and $k_5 - k_6$ are continuous in x, y and bounded in D as a function of s, t for all (x, y) in D, $\int_0^a (H'(s) - H'(x)) k_1 ds$, $\int_0^a (k_1 - k_5) ds$ and $\int_0^a (k_5 - k_6) ds$ do not take a jump across $\mathcal{C}$. Further analysis of

$$-\frac{\delta p_0}{2\pi} H'(x) \int_0^a k_6(x-s, y-H(s)) ds$$

is needed to bring out the jump of $\partial u_1/\partial x$ across $\mathcal{C}$. By setting $x - s = \bar{s}$, $y - H(x) = \alpha$, and $\bar{s}/\alpha = \sigma$,

$$\begin{aligned}
\frac{-\delta p_0}{2\pi} H'(x) \int_0^a k_6(x-s, y-H(s)) ds &= \frac{\delta p_0}{2\pi} H'(x) \int_{x-a}^{x} \frac{\bar{s}}{(\alpha - H'(x)\bar{s})^2 + \bar{s}^2} d\bar{s} \\
&= \frac{\delta p_0}{2\pi} H'(x) \int_{(x-a)/\alpha}^{x/\alpha} \frac{\sigma}{(1 - H'(x)\sigma)^2 + \sigma^2} d\sigma \\
&\equiv \frac{\delta p_0}{2\pi} H'(x) \mathcal{F}(\alpha).
\end{aligned}$$

When $y - H(x) = \alpha > 0$,

$$\begin{aligned}\mathcal{F}^+(\alpha) = &\frac{1}{2(1+H'(x)^2)} \ln\left(\left(1 - H'(x)\frac{x}{\alpha}\right)^2 + H'(x)^2\left(\frac{x}{\alpha}\right)^2\right)\\ &+\frac{H'(x)}{1+H'(x)^2}\tan^{-1}\left((1+H'(x)^2)\frac{x}{\alpha} - H'(x)\right)\\ &-\frac{1}{2(1+H'(x)^2)} \ln\left(\left(1 - H'(x)\frac{x-a}{\alpha}\right)^2 + H'(x)^2\left(\frac{x-a}{\alpha}\right)^2\right)\\ &-\frac{H'(x)}{1+H'(x)^2}\tan^{-1}\left((1+H'(x)^2)\frac{x-a}{\alpha} - H'(x)\right).\end{aligned}$$

When $y - H(x) = -\alpha < 0$,

$$\begin{aligned}\mathcal{F}^-(\alpha) = &-\frac{1}{2(1+H'(x)^2)} \ln\left(\left(1 - H'(x)\frac{a-x}{\alpha}\right)^2 + H'(x)^2\left(\frac{a-x}{\alpha}\right)^2\right)\\ &-\frac{H'(x)}{1+H'(x)^2}\tan^{-1}\left((1+H'(x)^2)\frac{a-x}{\alpha} - H'(x)\right)\\ &+\frac{1}{2(1+H'(x)^2)} \ln\left(\left(1 - H'(x)\frac{-x}{\alpha}\right)^2 + H'(x)^2\left(\frac{-x}{\alpha}\right)^2\right)\\ &+\frac{H'(x)}{1+H'(x)^2}\tan^{-1}\left((1+H'(x)^2)\frac{-x}{\alpha} - H'(x)\right).\end{aligned}$$

Letting y go to $H(x)$ from above and below, we find that $\lim_{\alpha\to 0}\mathcal{F}^{\pm}(\alpha)$ exists and we obtain

$$\lim_{\alpha\to 0}\frac{\delta p_0}{2\pi}H'(x)(\mathcal{F}^+(\alpha) - \mathcal{F}^-(\alpha)) = \frac{\delta p_0 H'(x)H'(x)}{(1+H'(x)^2)} \tag{3.9}$$

Next, in the expression of $\partial u_1/\partial y$, $A_6(x,y)$ takes the same amount of jump across $\mathcal{C}$ as

$$-\frac{\delta p_0}{2\pi}\int_0^a k_6(x-s, y-H(s))ds.$$

By setting $x - s = \bar{s}$, $y - H(x) = \alpha$, $\bar{s}/\alpha = \sigma$,

$$\begin{aligned}\frac{-\delta p_0}{2\pi}\int_0^a k_6(x-s,y-H(s))ds &= \frac{\delta p_0}{2\pi}\int_0^a \frac{x-s}{(\alpha - H'(x)(x-s))^2 + (x-s)^2}ds\\ &= \frac{\delta p_0}{2\pi}\int_{x-a}^{x}\frac{\bar{s}}{(\alpha - H'(x)\bar{s})^2 + \bar{s}^2}d\bar{s}\\ &= \frac{\delta p_0}{2\pi}\int_{(x-a)/\alpha}^{x/\alpha}\frac{\sigma}{(1-H'(x)\sigma)^2+\sigma^2}d\sigma\\ &\equiv \frac{\delta p_0}{2\pi}\mathcal{G}(\alpha).\end{aligned}$$

When $y - H(x) = \alpha > 0$,

$$\begin{aligned}\mathcal{G}^+(\alpha) = &-\frac{1}{2(1+H'(x)^2)} \ln\left(\left(1 - H'(x)\frac{x}{\alpha}\right)^2 + H'(x)^2\left(\frac{x}{\alpha}\right)^2\right)\\ &-\frac{H'(x)}{1+H'(x)^2}\tan^{-1}\left(1+H'(x)^2)\frac{x}{\alpha} - H'(x)\right)\\ &+\frac{1}{2(1+H'(x)^2)} \ln\left(\left(1 - H'(x)\frac{x-a}{\alpha}\right)^2 + H'(x)^2\left(\frac{x-a}{\alpha}\right)^2\right)\\ &+\frac{H'(x)}{1+H'(x)^2}\tan^{-1}\left((1+H'(x)^2)\frac{x-a}{\alpha} - H'(x)\right).\end{aligned}$$

When $y - H(x) = -\alpha < 0$,

$$\begin{aligned}\mathcal{G}^-(\alpha) = &\frac{1}{2(1+H'(x)^2)} \ln\left(\left(1 - H'(x)\frac{a-x}{\alpha}\right)^2 + H'(x)^2\left(\frac{a-x}{\alpha}\right)^2\right)\\ &+\frac{H'(x)}{1+H'(x)^2}\tan^{-1}\left((1+H'(x)^2)\frac{a-x}{\alpha} - H'(x)\right)\\ &-\frac{1}{2(1+H'(x)^2)} \ln\left(\left(1 - H'(x)\frac{-x}{\alpha}\right)^2 + H'(x)^2\left(\frac{-x}{\alpha}\right)^2\right)\\ &-\frac{H'(x)}{1+H'(x)^2}\tan^{-1}\left((1+H'(x)^2)\frac{-x}{\alpha} - H'(x)\right).\end{aligned}$$

Letting y go to $H(x)$ from above and below, we find that $\lim_{\alpha\to 0}\mathcal{G}^{\pm}(\alpha)$ exists and we obtain

$$\lim_{\alpha\to 0}\frac{\delta p_0}{2\pi}(\mathcal{G}^+(\alpha) - \mathcal{G}^-(\alpha)) = -\frac{\delta p_0 H'(x)}{(1+H'(x)^2)} \quad . \tag{3.10}$$

Since $\dot{x} = 1$, $\dot{y} = H'(x)$, jump condition (J2) follows from equations (3.9) and (3.10), and (J3) is proved in a similar way. To verify (J1), note that from these formulas, $\bar{u}$ and $\bar{v}$ have one-sided derivatives on either side of $\mathcal{C}$, so (J1) follows from the continuity of $\bar{u}$ and $\bar{v}$ across $\mathcal{C}$. This completes the proof.

In the expression (3.7) for $\bar{u}(x,y)$, $\bar{u}_1(x,y) - \bar{u}_3(x,y)$ and $\bar{u}_2(x,y)$ are smoother than $\bar{u}_3(x,y)$. The crucial part concerning the modulus of the continuity of $\bar{u}(x,y)$ is $\bar{u}_3(x,y)$. We state the modulus of continuity of $\bar{u}_3(x,y)$ and give the proof as follows.

PROPOSITION 3.2. *Let $0 < a < 1$, and let p be a bounded function supported in $[0,a]\times[-2a,2a]$, then*

$$\bar{u}_3(x,y) = \int_0^a\int_{-2a}^{2a} p(s,t)k_7(x-s,y-t)dtds$$

satisfies

$$|\bar{u}_3(x,y) - \bar{u}_3(x,\bar{y})| \le aC_1||p|||y-\bar{y}|(|\ln|y-\bar{y}|| + 1), \tag{3.11}$$

where $||p|| = \sup_{(s,t)\in D}|p(s,t)|$, C_1 is a constant.

Proof. Define

$$f(x,y,\bar{y}) = \int_0^a \int_{-2a}^{2a} \frac{|(s-x)(y+\bar{y}-2t)|dtds}{((x-s)^2+(y-t)^2)((x-s)^2+(\bar{y}-t)^2)}.$$

Then

$$|\bar{u}_3(x,y) - \bar{u}_3(x,\bar{y})| \leq \frac{||p||}{2\pi}|y-\bar{y}|f(x,y,\bar{y}).$$

Consider, without loss of generality, the case $x = \bar{y} = 0$. Then,

$$f(0,y,0) \leq \int_0^a \int_{-2a}^{2a} \frac{|s||t|+|s||y-t|}{(s^2+(y-t)^2)(s^2+t^2)}dtds \equiv f_1(0,y,0) + f_2(0,y,0).$$

We rescale $f_1(0,y,0)$ by setting $s = y\sigma$, $t = ay\tau$. Then

$$f_1(0,y,0) \leq a\int_0^{a/y} \int_{-2/y}^{2/y} \frac{a|\sigma||\tau|d\tau d\sigma}{(\sigma^2+(1-a\tau)^2)(\sigma^2+a^2\tau^2)}.$$

Let

$$B_1 = \left\{(\sigma,\tau) : 0 \leq \sigma \leq \tfrac{a}{y}, \tfrac{-2}{y} \leq \tau \leq \tfrac{2}{y}\right\},$$

$$B_2 = \left\{(\sigma,\tau) : 0 \leq \sigma, 0 \leq \sigma^2+\tau^2 \leq \left(\tfrac{4}{y}\right)^2\right\},$$

$$B_3 = \left\{(\sigma,\tau) : 0 \leq \sigma, 0 \leq \sigma^2+\tau^2 \leq \tfrac{1}{4}\right\},$$

$$B_4 = \left\{(\sigma,\tau) : 0 \leq \sigma, \tfrac{1}{4} \leq \sigma^2+\tau^2 \leq \left(\tfrac{4}{y}\right)^2\right\},$$

$$B_5 = \left\{(\sigma,\tau) : 0 \leq \sigma, \tfrac{1}{4} \leq \sigma^2+(1-\tau)^2 \leq \left(\tfrac{4}{y}+1\right)^2\right\}.$$

Since $0 < a < 1$, $B_1 \subset B_2$, $a|\sigma||\tau|/(\sigma^2+a^2\tau^2) < \frac{1}{2}$, then

$$f_1(0,y,0) \leq \frac{a}{2}\int\int_{B_2} \frac{d\tau d\sigma}{\sigma^2+(1-a\tau)^2}.$$

With $B_2 = B_3 \cup B_4$,

$$\begin{aligned} f_1(0,y,0) &\leq \frac{a}{2}\int\int_{B_3} \frac{d\tau d\sigma}{(\sigma^2+(1-a\tau)^2)} + \frac{a}{2}\int\int_{B_4} \frac{d\tau d\sigma}{(\sigma^2+(1-a\tau)^2)} \\ &\equiv \frac{a}{2}((\mathrm{i})+(\mathrm{ii})). \end{aligned}$$

Since $B_4 \subset B_5$, $a < 1$, and

$$\frac{1}{(\sigma^2+(1-a\tau)^2)} \leq \frac{1}{(\sigma^2+(1-\tau)^2)},$$

we obtain

$$\begin{aligned}
\text{(i)} &\leq \int\int_{B_3} \frac{1}{(\sigma^2+(1-\tau)^2)} d\tau d\sigma \leq C_2, \\
\text{(ii)} &\leq \int\int_{B_4} \frac{d\sigma d\tau}{(\sigma^2+(1-\tau)^2)} \\
&\leq \int\int_{B_5} \frac{d\sigma d\tau}{(\sigma^2+(1-\tau)^2)} \\
&\leq \int_{-\pi}^{\pi} \int_{1/2}^{(4/y)+1} \frac{1}{r} dr d\theta \leq C_3(|\ln|y|| + 1),
\end{aligned}$$

for some constants C_2, C_3. Choosing $C_4 = \max(C_2, C_3)$,

$$f_1(0, y, 0) \leq \frac{a}{2} C_4(|\ln|y|| + 1).$$

Similarly,

$$f_2(0, y, 0) \leq \frac{a}{2} C_4(|\ln|y|| + 1).$$

We have shown that $f(x, y, \bar{y}) \leq aC_4(|\ln|y|| + 1)$. This completes the proof.

The functions $k_2(x-s, y-t)$ and $k_1(x-s, y-t) - k_7(x-s, y-t)$ are in C^1 as functions of x and y, and are bounded for all $(s,t) \in D$, $(x,y) \in D$. Hence $\bar{u}_2$ and $\bar{u}_1 - \bar{u}_3$ are continuously differentiable in D, so there is a constant C_5 such that

$$|\bar{u}_2(x,y) - \bar{u}_2(x,\bar{y})| \leq aC_5||p|||y-\bar{y}|, \tag{3.12}$$

$$|\bar{u}_1(x,y) - \bar{u}_3(x,y) - (\bar{u}_1(x,\bar{y}) - \bar{u}_3(x,\bar{y}))| \leq aC_5||p|||y-\bar{y}|. \tag{3.13}$$

In view of (3.11), (3.12), and (3.13) we have the following concerning the modulus of continuity of $\bar{u}(x,y)$. A similar analysis holds for $\bar{v}(x,y)$ andwe obtain the following proposition.

PROPOSITION 3.3. *There is a constant C_6 such that*

$$|\bar{u}(x,y) - \bar{u}(x,\bar{y})| \leq aC_6||p|||y-\bar{y}|(|\ln|y-\bar{y}|| + 1), \tag{3.14}$$

$$|\bar{v}(x,y) - \bar{v}(x,\bar{y})| \leq aC_6||p|||y-\bar{y}|(|\ln|y-\bar{y}|| + 1). \tag{3.15}$$

4. A priori estimates. By definition, $h(x, y_0)$ is the solution to the ordinary differential equation (3.1). Since $||u|| + ||v|| < \frac{1}{4}$, $|dh/dx| \leq (\frac{1}{4})/(1-\frac{1}{4}) < 1$. Thus, if $y > 2a$, $g(x,y) > 1$, then $p_0(g(x,y)) = 0$; if $y < -2a$, $g(x,y) < 0$, $p_0(g(x,y)) = 0$. Hence the support of p is contained in $[0,a] \times [-2a, 2a]$. Furthermore, the set of y_0 such that the curve $y = h(x, y_0)$ meets the rectangle $[0,a] \times [-2a, 2a]$ is contained in the interval $[-3a, 3a]$. For later purposes, we require this interval to have length less than 1; thus, we require

$$a < \tfrac{1}{6}. \tag{R1}$$

Let $(u^j, v^j) \in \mathcal{B}$, $j = 1, 2$, and let $h_j(x, y_0)$ be the corresponding characteristic curves satisfying (3.1). In particular, since $H_j(0) = 0$, we may write $H_j(x)$ as

$$H_j(x) = \int_0^x \frac{v^j(s, H_j(s))}{1 + u^j(s, H_j(s))} ds, \qquad j = 1, 2.$$

Let

$$||H(x)|| = \sup_{x \in (0,a)} |H(x)|, \qquad ||h(x, y_0)|| = \sup_{(x,y_0) \in D} |h(x, y_0)|.$$

With this notation we shall derive inequalities for perturbations $H_1 - H_2$ and $h_1 - h_2$ in terms of $u^1 - u^2$ and $v^1 - v^2$. For this, we first state and prove in Lemma 4.2 a variation of Gronwall's inequality.

LEMMA 4.1. *Let $K_4 > 0$, $K_5 > 0$, $1 > \varepsilon > 0$ be constants and let $L(x)$ solve*

$$(*) \qquad L'(x) = -K_4(L(x) \ln L(X) - L(x)), \quad L(0) = \varepsilon K_5 < 1,$$

then $L(x) \leq 3(\varepsilon K_5)^{e^{-aK_4}}$.

Proof. Equation $(*)$ has solution $L(x) = \exp[1 - e^{-xK_4} + e^{-xK_4} \ln(\varepsilon K_5)]$, so

$$L(x) \leq 3 \exp[e^{-xK_4} \ln(\varepsilon K_5)] \leq 3 \exp[e^{-aK_4} \ln(\varepsilon K_5)] = 3(\varepsilon K_5)^{e^{-aK_4}}.$$

LEMMA 4.2. *Let $f(x)$ be a nonnegative function satisfying*

$$f(x) = \int_0^x -K_4(f(s) \ln f(s) - f(s)) ds + \varepsilon K_5.$$

Suppose $\varepsilon K_5 < 1$ and $3(\varepsilon K_5)^{e^{-aK_4}} < 1$, then

$$f(x) \leq 3(\varepsilon K_5)^{e^{-aK_4}}.$$

Proof. It is sufficient to show that $L(x)$ dominates $f(x)$. We use the fact that $-x \ln x + x$ is an increasing function in $(0, 1)$. So we require $0 \leq L(x) \leq 1$. This is fulfilled by our assumption $3(\varepsilon K_5)^{e^{-aK_4}} < 1$. Define for $\delta > 0$,

$$L_\delta(x) = \int_0^x -K_4(L_\delta(s) \ln L_\delta(s) - L_\delta(s)) ds + K_5 \varepsilon + \delta, \qquad L_\delta(0) = \varepsilon K_5 + \delta > f(0).$$

Both $f(x)$ and $L_\delta(x)$ are continuous functions. Suppose $L_\delta(x) < f(x)$ for some x. Let $x^* > 0$ be the smallest number such that $L_\delta(x^*) = f(x^*)$. Of course, $x^* > 0$, since $L_\delta(0) > f(0)$. Hence $f(x) < L_\delta(x)$, $x \in (0, x^*)$. We have

$$\begin{aligned} 0 &= L_\delta(x^*) - f(x^*) \\ &= \int_0^{x^*} -K_4(L_\delta(s) \ln L_\delta(s) - L_\delta(s)) ds - \int_0^{x^*} -K_4(f(s) \ln f(s) - f(s)) ds + \delta > 0, \end{aligned}$$

a contradiction. Consequently, $f(x) \leq L_\delta(x)$. This is true for any $\delta > 0$. Therefore

$$f(x) \leq \lim_{\delta \to 0} L_\delta(x) = 3(\varepsilon K_5)^{e^{-aK_4}}.$$

PROPOSITION 4.3. *Let* (u^i, v^i) *be in* $\mathcal{B}$, *and let* H_i, h_i *be the corresponding solution of* (3.1), (3.2), *respectively, for* $i = 1, 2$. *Then*

$$||H_1 - H_2|| \leq K_1(||u^1 - u^2|| + ||v^1 - v^2||)^{K_2}, \tag{4.1}$$

$$||h_1 - h_2|| \leq K_1(||u^1 - u^2|| + ||v^1 - v^2||)^{K_2} \tag{4.2}$$

where $K_1 = 3(8a)^{K_2}$, $K_2 = e^{-4a}$.

Proof. Inequality (4.1) will be proved in detail. Similar arguments will derive (4.2), therefore we skip the proof of (4.2). We have

$$\begin{aligned}
&|H_1(x) - H_2(x)| \\
&\quad \leq \int_0^x |\frac{v^1(s, H_1(s))}{1 + u^1(s, H_1(s))} - \frac{v^2(s, H_2(s))}{1 + u^2(s, H_2(s))}| ds \\
&\quad \leq 4 \int_0^x [|v^1(s, H_1(s)) - v^1(s, H_2(s))| + |v^1(s, H_2(s)) - v^2(s, H_2(s))|] ds \\
&\qquad + 4 \int_0^x |v^1| \{|u^1(s, H_1(s)) - u^1(s, H_2(s))| + |u^1(s, H_2(s)) - u^2(s, H_2(s))|\} ds \\
&\qquad + 4 \int_0^x |u^2| \{|v^1(s, H_1(s)) - v^1(s, H_2(s))| + |v^1(s, H_2(s)) - v^2(s, H_2(s))|\} ds \\
&\quad \leq 8 \int_0^x \{|v^1(s, H_1(s)) - v^1(s, H_2(s))| + |u^1(s, H_1(s)) - u^1(s, H_2(s))|\} ds \\
&\qquad + 8 \int_0^x \{|v^1(s, H_2(s)) - v^2(s, H_2(s))| + |u^1(s, H_2(s)) - u^2(s, H_2(s))|\} ds.
\end{aligned}$$

Thus,

$$\begin{aligned}
|H_1(x) - H_2(x)| \leq 8 \int_0^x \{|v^1(s, H_1) - v^1(s, H_2)| + |u^1(s, H_1) - u^1(s, H_2)|\} ds \\
+ 8a(||v^1 - v^2|| + ||u^1 - u^2||).
\end{aligned} \tag{4.3}$$

Using (3.14) and (3.15) in the right-hand side of (4.3) yields

$$\begin{aligned}
|H_1(x) - H_2(x)| \leq \int_0^x -4(|H_1 - H_2| \ln |H_1 - H_2| - |H_1 - H_2|) ds \\
+ 8a(||v^1 - v^2|| + ||u^1 - u^2||).
\end{aligned} \tag{4.4}$$

It is sufficient to require

$$3(8a)^{K_2} < 1, \tag{R2}$$

to apply Lemma 4.2 to (4.4) with $\varepsilon = ||v^1 - v^2|| + ||u^1 - u^2||$, $f(x) = |H_1(x) - H_2(x)|$, $K_4 = 4$, $K_5 = 8a$. The proof is complete.

Next, two inequalities for $||\bar{u}^1 - \bar{u}^2||$ and $||\bar{v}^1 - \bar{v}^2||$ bounded in terms of $||H_1 - H_2||$, $||h_1 - h_2||$ are needed. These inequalities are given in Proposition 4.7. We first require some preliminary results.

LEMMA 4.4. *Suppose $f(x,y)$ is a function in D satisfying*

$$|f(x,y) - f(x,\bar{y})| \leq C_0|y-\bar{y}|(|\ln|y-\bar{y}|| + 1),$$

for all $x \in [0,a]$ and C_0 is a constant satisfying $0 < 2aC_0 < 1$. Let $z(x,z_0)$ and $\bar{z}(x,\bar{z}_0)$ be the solutions of

$$\frac{dz}{dx} = f(x,z), \quad z(0,z_0) = z_0, \quad x \in [0,a],$$

$$\frac{d\bar{z}}{dx} = f(x,\bar{z}), \quad \bar{z}(0,\bar{z}_0) = \bar{z}_0, \quad x \in [0,a],$$

respectively. Then, there are constants $\beta \in (0,1)$ and $K_6 > 0$, depending on a and C_0 only, such that

$$\bar{z}(x) - z(x) \geq K_6(\bar{z}_0 - z_0)^{1/\beta}, \quad \text{if } 0 \leq \bar{z}_0 - z_0 \leq 1.$$

Proof.

$$\bar{z}(x) - z(x) = \bar{z}_0 - z_0 + \int_0^x \frac{f(s,\bar{z}) - f(s,z)}{\bar{z} - z}(\bar{z} - z)ds.$$

Introducing $w(x) = \bar{z}(x) - z(x)$ and $(f(s,\bar{z}) - f(s,z))/(\bar{z} - z) = \zeta(s)$, we have

$$w(x) = w(0) + \int_0^x \zeta(s)w(s)ds,$$

which has the solution

$$w(x) = w(0)e^{\int_0^x \zeta(s)ds}.$$

Letting the max of $|\zeta|$ be assumed at x^*, we have

$$w(0)e^{-a|\zeta(x^*)|} \leq w(x) \leq w(0)e^{a|\zeta(x^*)|}.$$

Using the assumption in the case $x = x^*$, $|\zeta(x^*)| \leq C_0(1 + |\ln w(x^*)|)$, we obtain, for all x in $[0,a]$,

$$w(0)e^{-aC_0(1+|\ln w(x^*)|)} \leq w(x) \leq w(0)e^{aC_0(1+|\ln w(x^*)|)}. \tag{4.5}$$

If $0 \leq w(x^*) \leq 1$, the left side inequality in (4.5) becomes

$$w(x) \geq w(0)e^{-aC_0}w(x^*)^{aC_0}. \tag{4.6a}$$

In particular,

$$w(x^*) \geq w(0)e^{-aC_0}w(x^*)^{aC_0},$$

which implies

$$w(x^*) \geq w(0)^{1/(1-aC_0)}e^{-aC_0/(1-aC_0)}.$$

Then

$$w(x^*)^{aC_0} \geq w(0)^{aC_0/(1-aC_0)} e^{-a^2C_0^2/(1-aC_0)}.$$

Inserting this back into (4.6a), we obtain

$$w(x) \geq w(0)^{1/(1-aC_0)} e^{-aC_0/(1-aC_0)}. \tag{4.6b}$$

If $1 \leq w(x^*)$, the right-hand side inequality in (4.5) becomes

$$w(x) \leq w(0)e^{aC_0} w(x^*)^{aC_0}.$$

In particular,

$$w(x^*) \leq w(0)e^{aC_0} w(x^*)^{aC_0},$$

which implies

$$w(x^*) \leq w(0)^{1/(1-aC_0)} e^{aC_0/(1-aC_0)}.$$

Therefore,

$$w(x^*)^{-aC_0} \geq w(0)^{-aC_0/(1-aC_0)} e^{-a^2C_0^2/(1-aC_0)}. \tag{4.7a}$$

Meanwhile, the left-hand side inequality in (4.5) becomes

$$w(x) \geq w(0)e^{-aC_0} w(x^*)^{-aC_0}. \tag{4.7b}$$

Inserting (4.7a) into (4.7b), we obtain

$$w(x) \geq w(0)^{(1-2aC_0)/(1-aC_0)} e^{-aC_0/(1-aC_0)}. \tag{4.7c}$$

Since $0 \leq w(0) \leq 1$, (4.6b) and (4.7c) imply, for all x in $[0, a]$,

$$w(x) \geq w(0)^{1/(1-aC_0)} e^{-aC_0/(1-aC_0)}.$$

Let $\beta = 1 - aC_0$ and $K_6 = e^{-aC_0/(1-aC_0)}$, the proof is complete.

LEMMA 4.5. *Let (u^1, v^1) be in $\mathcal{B}$ and $g_1(s, t_j)$ be theinverse function of $h_1(x, y_j)$, which satisfies the differential equation* (3.1) *subject to initial value y_j, for $j = 1, 2$, and $0 \leq |y_1 - y_2| \leq 1$. Then*

$$|g_1(s, t_1) - g_1(s, t_2)| \leq K_7 |t_1 - t_2|^\beta, \tag{4.8}$$

for some constant K_7 and β as in the Lemma 4.4 *for all (s, t_1) and (s, t_2) in D.*

Proof. Since (u^1, v^1) is in $\mathcal{B}$, applying Lemma 4.4 to the differential equation (3.1) with $C_0 = \frac{1}{4}$ and using (R1), we find that $|y_1 - y_2|^{1/\beta} \leq K_6{}^{-1} |h_1(s, y_1) - h_1(s, y_2)|$. Through the definition of g_1 and h_1, this inequality is equivalent to (4.8) if $K_7 = K_6{}^{-\beta}$.

LEMMA 4.6. *Fix t_1, let (u^j, v^j) be in $\mathcal{B}$. Let $g_1(s, t_1)$ and $g_2(s, t_1)$ satisfy $0 \leq |g_1(s, t_1) - g_2(s, t_1)| \leq 1$. Let $h_j(s, y_j)$ be the inverse function of $g_j(s, t_1)$ for $j = 1, 2$, where $h_j(s, y_j)$ satisfies equation* (3.1). *Then*

$$|g_1(s, t_1) - g_2(s, t_1)| \leq K_7 |h_1(s, y_2) - h_2(s, y_2)|^\beta \tag{4.9}$$

for all (s, t_1) and (s, y_2) in D and constant K_7 as in Lemma 4.4.

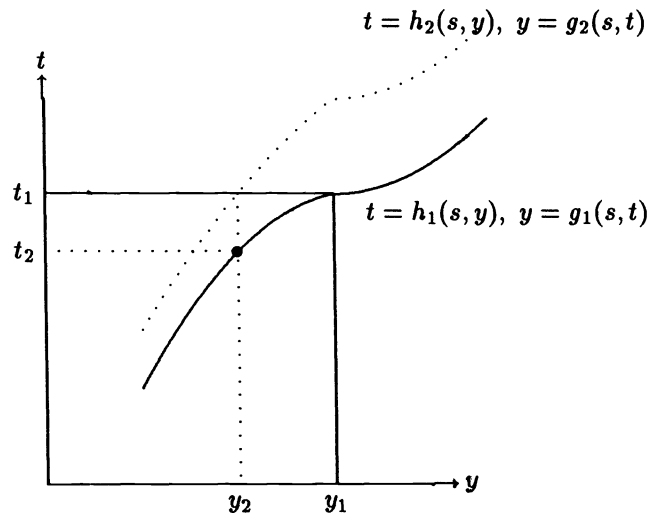


FIG. 4.1. *These curves show the relation between t and y for fixed s.*

Proof. Referring to Fig. 4.1 and applying Lemma 4.5, we have

$$|g_1(s,t_1) - g_2(s,t_1)| = |y_1 - y_2| = |g_1(s,t_1) - g_1(s,t_2)| \le K_7|t_1 - t_2|^\beta.$$

That is equivalent to (4.9).

As in (3.7), for $i = 1,2$, we write $\bar{u}^i = [\bar{u}_1^i - \bar{u}_3^i] + \bar{u}_2^i + \bar{u}_3^i$.

PROPOSITION 4.7.

$$\begin{aligned}(4.10)\qquad ||\bar{u}^1 - \bar{u}^2|| \le{}& K_{15}||p_0||||H_1 - H_2||^\alpha + K_{12}||p_0||||H_1 - H_2||^\alpha \\ &+K_{13}||p_0'||||h_1 - h_2||^\beta + K_7K_{10}||p_0'||||h_1 - h_2||^\beta\end{aligned}$$

$$\begin{aligned}(4.11)\qquad ||\bar{v}^1 - \bar{v}^2|| \le{}& K_{15}||p_0||||H_1 - H_2||^\alpha + K_{12}||p_0||||H_1 - H_2||^\alpha \\ &+K_{13}||p_0'||||h_1 - h_2||^\beta + K_7K_{10}||p_0'||||h_1 - h_2||^\beta,\end{aligned}$$

where $||p_0'|| = \sup_{y>0}|p_0'(y)|$, K_7, K_{10}, K_{12}, K_{13}, *and* K_{15} *are positive constants, with* $\alpha \in (0,1)$ *and* β *as in Lemma* 4.4.

Proof. Given $H_1(x)$ and $H_2(x)$, define $\bar{H}_1(x) = H_1(x) - 2||H_1 - H_2||$, $\bar{H}_2(x) = H_1(x) + 2||H_1 - H_2||$. Let us decompose D in the following way:

$$\begin{aligned}D_M &= \{(x,y) \in D, |y - H_1(x)| \le 2||H_1 - H_2||\},\\ D_U &= \{(x,y) \in D, y - \bar{H}_2(x) > 0\},\\ D_L &= \{(x,y) \in D, y - \bar{H}_1(x) < 0\}.\end{aligned}$$

$$\begin{aligned}\bar{u}_3^1(x,y)-\bar{u}_3^2(x,y) &= \int\int_{D_L}\frac{(x-s)(p_0(g_1(s,t))-p_0(g_2(s,t)))}{(x-s)^2+(y-t)^2}dtds\\ &\quad+\int\int_{D_M}\frac{(x-s)(p_0(g_1(s,t))-p_0(g_2(s,t)))}{(x-s)^2+(y-t)^2}dtds\\ &\quad+\int\int_{D_U}\frac{(x-s)(p_0(g_1(s,t))-p_0(g_2(s,t)))}{(x-s)^2+(y-t)^2}dtds\\ &\equiv \text{I+II+III}.\end{aligned}$$

For (s,t) in D_L, $p_0(g_1(s,t))=p_0(g_2(s,t))=0$, hence

$$|\text{I}|=0. \tag{4.12}$$

To estimate II we write

$$\begin{aligned}|\text{II}| &\le 2||p_0||\int_0^a\int_{\bar{H}_1(s)}^{\bar{H}_2(s)}\frac{dtds}{\sqrt{(x-s)^2+(y-t)^2}}\\ &\equiv 2||p_0||Q.\end{aligned}$$

Let $\bar{t}=t-\bar{H}_1(s)$, then

$$Q\le\int_0^a\int_0^{\bar{H}_2(s)-\bar{H}_1(s)}\frac{d\bar{t}ds}{\sqrt{(x-s)^2+(y-\bar{t}-\bar{H}_1(s))^2}}.$$

Making the further change of variables $\tau=y-\bar{t}-\bar{H}_1(s)$, $\sigma=x-s$, and setting $d=4||H_1-H_2||$,

$$\begin{aligned}Q &\le \int_{x-a}^{x}\int_{y-\bar{H}_1(x-\sigma)}^{y-\bar{H}_2(x-\sigma)}\frac{d\tau d\sigma}{\sqrt{\sigma^2+\tau^2}}\\ &\le \int_{-a}^{a}\int_{-d/2}^{d/2}\frac{dtds}{\sqrt{s^2+t^2}}\\ &= 4\int_0^a\int_0^d\frac{dtds}{\sqrt{s^2+t^2}}\\ &= 4\int_0^{\tan^{-1}d/a}\int_0^a\sec\theta dr d\theta+4\int_{\tan^{-1}d/a}^{\pi/2}\int_0^{d/2\sin\theta}dr d\theta\\ &\equiv \text{II}_1+\text{II}_2.\end{aligned}$$

Since $0\le\sec\theta\le 2$ in $0\le\theta\le\tan^{-1}d/a$, if $d/a<1$, and $\tan^{-1}d/a<C_7(d/a)^\alpha$ for $0<\alpha<1$, for some constant C_7 depending on α, we have

$$\text{II}_1\le C_8a^{1-\alpha}d^\alpha\equiv C_9d^\alpha.$$

$$\text{II}_2\le\pi d\int_{\tan^{-1}d/2a}^{\frac{\pi}{2}}\frac{1}{\theta}d\theta\le\pi d\left(\ln\frac{\pi}{2}-\ln\frac{d}{2a}\right)\le C_{10}d^\alpha,$$

for some constant C_{10}. Hence

$$|\text{II}|\le 2||p_0||(C_9+C_{10})d^\alpha\equiv K_{15}||p_0||||H_1-H_2||^\alpha. \tag{4.13}$$

It remains to estimate III. In D_U, both $g_1(s,t)$ and $g_2(s,t)$ are positive. They fall in the range in which $p_0(\cdot)$ is differentiable. Using $(R1)$, we see that $|g_1(s,t) - g_2(s,t)| < 1$. Hence we may use Lemma 4.6 to obtain $|p_0(g_1(s,t)) - p(g_2(s,t))| \leq ||p_0'||K_7|h_1 - h_2|^\beta$. Consequently, there is a constant K_{10} such that

$$|\mathrm{III}| \leq K_7 K_{10} ||p_0'|| |h_1 - h_2|^\beta. \tag{4.14}$$

Define

$$\begin{aligned}
&\bar{u}_3^1(x,y) - \bar{u}_3^2(x,y) - (\bar{u}_3^1(x,\bar{y}) - \bar{u}_3^2(x,\bar{y})) \\
&= \int\int_{D_U} (p_0(g_1) - p_0(g_2)) \frac{(y-\bar{y})(s-x)(y+\bar{y}-2t)dsdt}{s + ((x-s)^2 + (y-t)^2)((x-s)^2 + (\bar{y}-t)^2)} \\
&\quad + \int\int_{D_M} (p_0(g_1) - p_0(g_2)) \frac{(y-\bar{y})(s-x)(y+\bar{y}-2t)dsdt}{((x-s)^2 + (y-t)^2)((x-s)^2 + (\bar{y}-t)^2)} \\
&\equiv \mathrm{IV} + \mathrm{V}.
\end{aligned}$$

By Proposition 3.3 there is a constant K_{11} such that

$$|\mathrm{IV}| \leq ||p_0'|| ||g_1 - g_2|| K_{11} |y - \bar{y}| (|\ln |y - \bar{y}|| + 1).$$

By the proof of (4.13), there is a constant K_{12} such that

$$|\mathrm{V}| \leq ||p_0|| ||H_1 - H_2||^\alpha K_{12} |y - \bar{y}| (|\ln |y - \bar{y}|| + 1). \tag{4.15}$$

Finally, using (4.9),

$$|\mathrm{IV}| \leq ||p_0'|| ||h_1 - h_2||^\beta K_{13} |y - \bar{y}| (|\ln |y - \bar{y}|| + 1) \tag{4.16}$$

for some constant K_{13}. With these estimates of (4.12) through (4.16), the quantity $||\bar{u}_3^1 - \bar{u}_3^2||$ has been bounded by the right-hand side of (4.10). The other terms in the decomposition of $\bar{u}^1 - \bar{u}^2$ and $\bar{v}^1 - \bar{v}^2$ could be bounded in the same way, since they are more regular than $\bar{u}_3^1 - \bar{u}_3^2$. This completes the proof.

As a consequence of Propositions 4.3 and 4.7 we have Corollary 4.8.

COROLLARY 4.8. *The mapping T is continuous in $\mathcal{A}$.*

Proof. Inserting estimates (4.1), (4.2) into (4.10) and (4.11), we see that the mapping T is continuous in $\mathcal{A}$.

PROPOSITION 4.9. *There are constants C_6, K_{14} such that*

$$||\bar{u}|| + ||\bar{v}|| \leq 2aC_6 ||p_0|| + 8a||p_0|| + 2aK_{14} ||p_0||.$$

Proof. By the definition of $\bar{u}_3$, since $p(s,t)$ is supported in $[0,a] \times [-2a, 2a]$,

$$\bar{u}_3(x,y) = \int_0^a \int_{-2a}^{2a} p(s,t) k_7(x-s, y-t) dtds.$$

For those (x,y) such that $(y-t)^2 + (x-s)^2 > a^2$ for all (s,t) in $[0,a] \times [-2a, 2a]$ we have

$$\begin{aligned}
|\bar{u}_3(x,y)| &\leq ||p_0|| \int_0^a \int_{-2a}^{2a} \frac{|s-x|}{2\pi((y-t)^2 + (x-s)^2)} dsdt \\
&\leq ||p_0|| \int_0^a \int_{-2a}^{2a} \frac{1}{a\pi} dsdt \leq 4a||p_0||.
\end{aligned}$$

Otherwise, using the polar coordinate centered at (x, y),

$$|\bar{u}_3(x,y)| \le ||p_0|| \int_0^{4a} \int_{-\pi}^{\pi} \frac{1}{2\pi} d\theta dr = 4a||p_0||. \tag{4.17}$$

In either case, we conclude that

$$|\bar{u}_3(x,y)| \le 4a||p_0||. \tag{4.18}$$

Similarly, there is a constant K_{14} such that

$$|\bar{u}_1(x,y) - \bar{u}_3(x,y)| \le aK_{14}||p_0||, \tag{4.19}$$

$$|\bar{u}_2(x,y)| \le aK_{14}||p_0||. \tag{4.20}$$

Applying Proposition 3.3, we achieve

$$|\bar{u}(x,y) - \bar{u}(x,\bar{y})| + |\bar{v}(x,y) - \bar{v}(x,\bar{y})| \le 2aC_6||p_0|||y - \bar{y}|(|\ln|y - \bar{y}|| + 1). \tag{4.21}$$

By inequalities (4.18), (4.19), (4.20), and (4.21) the proof is complete.

In the next section, Proposition 4.9 will be used to show that T is self-mapping from $\mathcal{B}$ into $\mathcal{B}$. With Corollary 4.8 concerning the continuity of the mapping T, the Schauder fixed-point theorem will be applied to obtain the fixed point of T.

5. Proof of Theorem 2.1. First we show that T is a self-mapping on $\mathcal{B}$. Letting $(u, v) \in \mathcal{B}$, we construct $h(x, y_0)$, $H(x) = h(x, 0)$, $g(x, y)$ to obtain $p(x, y) = p_0(g(x, y)) = T_1(u, v)$. We must show that the pair $(\bar{u}, \bar{v}) = T_2(p)$ is in $\mathcal{B}$.

By Proposition 4.9 and choosing a small enough so that

$$2aC_6||p_0|| + 8a||p_0|| + 2aK_{14}||p_0|| \le \tfrac{1}{4}, \tag{R3}$$

for a given $||p_0||$, it follows that $T(\mathcal{B}) \subset \mathcal{B}$. The continuity of the mapping T is verified by the Corollary 4.8. To apply the Schauder fixed-point theorem, T must be a compact mapping from $\mathcal{B}$ into $\mathcal{B}$. To this end, we must show that any sequence in the image of T has a convergent subsequence. Given a sequence (u^j, v^j) in $\mathcal{B}$, through the mapping T_1, we have p_j, $h_j(x, y_0)$,$H_j(x)$, and $g_j(x, y)$; then through the mapping T_2, we end up with the sequence $(\bar{u}^j, \bar{v}^j)$. By the definitions of $h_j(x, y_0)$ and $H_j(x)$, it is easy to obtain

$$|H_j(x) - H_j(\bar{x})| \le C_{11}|x - \bar{x}|,$$

$$|h_j(x, y_0) - h_j(\bar{x}, \bar{y}_0)| \le C_{12}(|x - \bar{x}| + |y - \bar{y}_0|)$$

for some constants C_{11}, C_{12} independent of j. Thus $\{H_j(x)\}$ and $\{h_j(x, y_0)\}$ are equicontinuous families. Since $[0, a]$ is compact, by the Arzela–Ascoli theorem, there is a convergent subsequence $H_k(x)$ in $C[0, a]$. For the sequence $h_k(x, y_0)$, by the same argument, there is a convergent subsequence $h_l(x, y_0)$ in $C([0, a] \times [-2a, 2a])$. Therefore $\{H_l(x)\}$ and $\{h_l(x, y_0)\}$ are Cauchy sequences in the spaces $C[0, a]$, and $C([0, a] \times [-2a, 2a])$, respectively. Now we show that the sequence $(\bar{u}^l, \bar{v}^l)$ is Cauchy in $\mathcal{B}$. Let $R = [0, a] \times [-2a, 2a]$. Applying Proposition 4.7, we have

$$\begin{aligned} ||\bar{u}^{l+m} - \bar{u}^l|| \le{}& K_{15}||p_0||||H_{l+m} - H_l||^{\alpha} + K_7K_{10}||p_0'|| \sup_R |h_{l+m} - h_l|^{\beta} \\ &+ K_{12}||p_0||||H_{l+m} - H_l||^{\alpha} + K_{13}||p_0'|| \sup_R |h_{l+m} - h_l|^{\beta}, \end{aligned}$$

$$||\bar{v}^{l+m} - \bar{v}^{l}|| \leq K_{15}||p_0||||H_{l+m} - H_l||^{\alpha} + K_7 K_{10}||p_0'|| \sup_R |h_{l+m} - h_l|^{\beta}$$
$$+K_{12}||p_0||||H_{l+m}H_l||^{\alpha} + K_{13}||p_0'|| \sup_R |h_{l+m} - h_l|^{\beta}.$$

Hence $(\bar{u}^l, \bar{v}^l)$ is Cauchy in $\mathcal{B}$. Thus T is a continuous compact mapping from $\mathcal{B}$ into $\mathcal{B}$, if a is sufficiently small so that (R1), (R2), and (R3) are fulfilled. The Schauder fixed-point theorem completes the proof.

REFERENCES

[1] D. HOFF, *Construction of solutions for compressible, isentropic Navier–Stokes equations in one space dimension with on smooth initial data*, Proc. Royal Soc. Edinburgh, 103A, (1986), pp. 301–315.

[2] ———, *Global existence for one-dimensional, compressible, isentropic Navier–Stokes equations with large initial data*, Trans. Amer. Math. Soc., 303 (1987), pp. 169–181

[3] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Second edition, Springer-Verlag, Berlin, Heidelberg, 1983.

[4] A. VALLI, *On the existence of stationary solutions to compressible Navier–Stokes equations*, Anal. Nonlinéaire, 4 (1987), pp. 89–113.

[5] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[6] R. B. KELLOGG, *Discontinuous solutions of the linearized steady-state, compressible, viscous Navier–Stokes equations*, SIAM J. Math. Anal., 19 (1988), pp. 567–579.

[7] I. S. GRADSHTEYN AND I. M. RYZHIK, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980, p. 504.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 623-638, May 1991

004

ON THE OBSTACLE PROBLEM FOR QUASILINEAR ELLIPTIC EQUATIONS OF p LAPLACIAN TYPE*

HI JUN CHOE† AND JOHN L. LEWIS†

Abstract. This article considers an obstacle problem for quasilinear elliptic equations of p Laplacian type. It is shown, under certain smoothness assumptions on the obstacle, that the solution to the corresponding obstacle problem has interior Hölder continuous derivatives.

Key words. p Laplacian, obstacle problem, Hölder continuity of the gradient, interior regularity

AMS(MOS) subject classifications. 35J70, 35J60

1. Introduction. In this paper we consider solutions to an obstacle problem for the following partial differential equation in a bounded domain Ω of Euclidean n space ($\mathbb{R}^n$):

$$\operatorname{div}(|\nabla w|^{p-2}\nabla w) = \hat{b}(x, w, \nabla w), \tag{1.1}$$

when $1 < p < \infty$. Here, ∇w denotes the gradient of w. The left-hand side of this equation is the so-called p Laplacian, and is interesting from a partial differential equation (P.D.E.) standpoint, essentially because of its degeneracy. Thus to obtain the results in this paper we cannot directly apply standard obstacle theorems for uniformly elliptic equations of divergence type.

As for our results, under certain regularity assumptions on $\hat{b}$ and the obstacle $\hat{\psi}$, we shall show that a bounded solution to the obstacle problem for (1.1) has Hölder continuous derivatives on compact subsets of Ω. These results generalize work of Fuchs [2], Lindqvist [6], and Norando [9], who obtained similar regularity results for an obstacle problem corresponding to (1.1) under more restrictive assumptions.

Indeed, Lindqvist assumes $\Omega \subset \mathbb{R}^2$ and uses quasi-conformal mapping techniques. Fuchs and Norando consider only the case when $p > 2$, $\hat{b} \equiv 0$, and $\hat{\psi}$ has essentially bounded distributional second partials in $\mathcal{R}^n[\hat{\psi} \in W_{2,\infty}(\mathbb{R}^n)]$. Moreover, their methods do not appear extendable to the case when $1 < p < 2$. In contrast we assume

(a) $1 < p < \infty$,

(b) $\hat{\psi}$ has distributional second partials on $\mathbb{R}^n$ which are qth power integrable for some $q > n$, i.e.,

$$\hat{\psi} \in W_{2,q}(\mathbb{R}^n), \tag{1.2}$$

(c) For some $c_1 > 0$, $\hat{b}$ satisfies the structure conditions:

$$|\hat{b}(x, z, h)| \leqq c_1(f(x) + |h|^p), \tag{1.3}$$

when $x \in \Omega$ and $(z, h) \in \mathbb{R} \times \mathbb{R}^n$. Here, $f \geqq 0$ is measurable and qth power integrable on $\mathbb{R}^n$ for some $q > n$, i.e.,

$$f \in L_q(\mathbb{R}^n). \tag{1.4}$$

(a)-(c) are well known to be optimal in the classical case, $p = 2$.

* Received by the editors July 20, 1989; accepted for publication (in revised form) April 11, 1990.

† Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506-0027. The work of the second author was supported in part by the National Science Foundation and the Commonwealth of Kentucky through the Kentucky EPSCoR program.

Our results are new even for weak subsolutions to (1.1) (so no obstacle is present). Previously DiBenedetto [1] has shown regularity of weak subsolutions to (1.1) when either $f \equiv 0$ in (1.3), or when $|h|^p$ in (1.3) is deleted and $f \in L_{p'n}(\Omega)$, $p' = p/(p-1)$. Also, Tolksdorf [11] obtained regularity as above when $f \equiv 1$ in (1.3).

To get our results we use, among other things, an improved version of a method originally used by Lewis in [5]. We essentially show that a certain function of the gradient of our solution (say $\nabla \hat{u}$) is a subsolution to a uniformly elliptic equation in divergence form, whose right-hand side involves terms in $\hat{\psi}$, f, and a function which satisfies an inequality of reverse Hölder type on balls contained in Ω. Using elliptic theory, we then get an integral inequality for $|\nabla \hat{u}|$ which can be iterated to get Hölder continuity of $\nabla \hat{u}$.

We remark that, as this paper was in preparation, Bill Ziemer informed the authors that Jun Mu has an independent and quite different proof of regularity for the obstacle problem corresponding to (1.1) when $1 < p < \infty$, $\hat{\psi} \in W_{2,\infty}(\mathbb{R}^n)$, and $\hat{b} \equiv 0$. Next, we remark that our arguments apply to more general partial differential equations than (1.1). For example, if $|\nabla w|^{p-2}\nabla w$ is replaced in (1.1) by $A(x, w, \nabla w)$, $A = (A_1, \cdots, A_n)$ and A satisfies the conditions in either [1] or [11], then solutions to the corresponding obstacle problem have the same regularity as for the p Laplacian. Moreover, the proof we give can be adapted to equations with the same principal part as in [3] and [7]. Finally, we emphasize that the ultimate generality of the structure assumptions on $\hat{b}$ in (1.3) forces us to assume our solutions are bounded and makes it considerably difficult to follow more or less standard P.D.E. procedure in approximating by smooth solutions and establishing L_∞ bounds on the gradient.

To be more specific, recall that $\hat{u}$ is a solution to the obstacle problem for (1.1) with obstacle $\hat{\psi}$, provided $\hat{u} \in W_{1,p}(\Omega)$,

$$\hat{u}(x) \geqq \hat{\psi}(x), \qquad x \in \Omega \text{ a.e.}, \tag{1.5}$$

and

$$\int_\Omega [(|\nabla \hat{u}|^{p-2}\nabla \hat{u}) \cdot \nabla \eta + \hat{b}(x, \hat{u}, \nabla \hat{u})\eta]\, dx \geqq 0, \tag{1.6}$$

whenever $\eta \in C_0^\infty(\Omega)$ with $\hat{u} + \eta \geqq \hat{\psi}$ (almost everywhere in Ω), and so in particular when $\eta \geqq 0$. We prove the following theorem.

THEOREM 1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and E a given compact subset of Ω. For fixed p, $1 < p < \infty$, suppose $\hat{u}$, $\hat{\psi}$, $\hat{b}$, f, satisfy* (1.2)-(1.6), *and*

$$|\hat{u}(x)| \leqq c_2 < \infty, \qquad x \in \Omega. \tag{1.7}$$

Then $\hat{u}$ has a representative in $W_{1,p}(\Omega)$ with continuous derivatives and for this representative: Given E, a compact subset of Ω, there exists $\alpha \in (0, 1)$, and $c < \infty$, depending only on p, n, q, c_1, c_2, $\|\hat{\psi}\|_{2,q}$, $\|f\|_q$, and the distance from E to the boundary of Ω such that

$$\max_{x \in E} |\nabla \hat{u}(x)| \leqq c,$$

$$|\nabla \hat{u}(x) - \nabla \hat{u}(y)| \leqq c|x - y|^\alpha, \qquad x, y \in E.$$

In Theorem 1, $\|\cdot\|_{2,q}$ and $\|\cdot\|_q$, denote the norm in $W_{2,q}(\mathbb{R}^n)$ and $L_q(\mathbb{R}^n)$, respectively. In the sequel, we shall let c denote a positive constant, with the same dependence as in Theorem 1, not necessarily the same at each occurrence. We say that c depends on the data. Also a denotes a positive constant depending only on p and n, not necessarily the same at each occurrence. As for our proof, let $\delta(F_1, F_2)$ denote the

distance between the sets F_1, F_2, and $B(x, R)=\{y: |y-x|<R\}$. If $0<\varepsilon\leqq 1/10 \min[\delta(E, \partial\Omega), 1]$ then we first consider functions $u=u(\cdot, \varepsilon)$, $\psi=\psi(\cdot, \varepsilon)$, such that $\psi\in C^\infty(\mathbb{R}^n)$, where (1.2), (1.5) hold with $\hat{u}$, $\hat{\psi}$ replaced by u, ψ, and Ω by $B=B(x_0, R_0)$. Also, $x_0\in E$ and $0<R_0\leqq 1/10 \min[\delta(E, \partial\Omega), 1]$. Moreover,

$$\|\psi\|_{2,q}\leqq c\|\hat{\psi}\|_{2,q} \tag{1.8}$$

and

$$\int_B [(\varepsilon+|\nabla u|^2)^{(p/2-1)}\nabla u\cdot\nabla\eta+b_\varepsilon(x, u, \nabla u)\eta]\,dx\geqq 0 \tag{1.9}$$

whenever $\eta\in \dot{W}_{1,p}(\Omega)$ with $u+\eta\geqq\psi$ almost everywhere in B. Here $\dot{W}_{1,p}(\Omega)$ denotes the closure of $C_0^\infty(\Omega)$ in the norm of $W_{1,p}(\Omega)$, and

$$|b_\varepsilon(x, z, h)|\leqq 2c_1 \min\left[\frac{1}{\varepsilon}, f(x)+|h|^p\right], \tag{1.10}$$

when $(x, z, h)\in\Omega\times\mathbb{R}\times\mathbb{R}^n$ (c_1 as in (1.3)). Next we suppose for fixed ε as above,

$$|u(x)|\leqq c_3, \qquad x\in B, \tag{1.11}$$

where c_3 depends only on the data.

We note that (1.9) is easier to work with than (1.6), since $u=u(\cdot, \varepsilon)$ is a supersolution to a nondegenerate elliptic equation. In § 2 we make some preliminary remarks concerning Hölder continuity of u, $\hat{u}$, and point out that a difference quotient argument of Tolksdorf can be used to show for fixed σ, $0<\sigma<1$, ε, and R_0 as above that $u\in W_{2,2}[B(x_0, \sigma R_0)]$ and $|\nabla u|\in L_\infty[B(x_0, \sigma R_0)]$. In § 3, we obtain L_∞ estimates on $|\nabla u|$ in $B(x_0, \sigma R_0)$ which depend only on σ, R_0, and the data (i.e., the same quantities as c in Theorem 1). Thus our estimates are independent of ε, $0<\varepsilon<1/10 \min[1, \delta(E, \partial\Omega)]$. In §§ 4 and 5 we use the argument of Lewis, mentioned earlier, to obtain Theorem 1 with $\hat{u}$ replaced by $u(\cdot, \varepsilon)$, $\Omega=B$, $E=B(x_0, \sigma R_0)$, and constants which depend only on σ, R_0 and the data. In § 6 we show the existence of $u=u(\cdot, \varepsilon)$ for $R_0>0$ sufficiently small (depending only on the data). We then let $\varepsilon\to 0$ and get Theorem 1 from local uniqueness type arguments and our previous work.

2. Preliminary reductions. Let $u=u(\cdot, \varepsilon)$, $b=b_\varepsilon$, and $\psi=\psi(\cdot, \varepsilon)$ be as in § 1 for fixed ε, R_0, with $0<\varepsilon$, $R_0<1/10 \min[1, \delta(E, \partial\Omega)]$ and p, $1<p<\infty$. Then $u\in W_{1,p}(B)$ and u, ψ, satisfy (1.2), (1.5), with $\hat{u}$, $\hat{\psi}$, replaced by u, ψ, and also (1.8)-(1.11). For $\hat{u}$ in Theorem 1 we claim there exists c, $\bar{\lambda}>0$, depending only on the data, such that $\hat{u}$ has a representative in $W_{1,p}(\Omega)$ with

$$|\hat{u}(x)-\hat{u}(y)|\leqq c|x-y|^{\bar{\lambda}}, \tag{2.1a}$$

whenever

$$\max[\delta(\{x\}, E), \delta(\{y\}, E)]\leqq\tfrac{1}{2}\delta(E, \partial\Omega).$$

We also claim for $\varepsilon>0$ as above and fixed σ, $0<\sigma<1$, that there exists d, $\lambda>0$ depending only on σ, R_0, and the data such that

$$|u(x)-u(y)|\leqq d|x-y|^{\lambda}, \tag{2.1b}$$

whenever $x, y\in B_1=B(x_0, R_1)$. Here, $R_1=(1/100)(99+\sigma)R_0$. To prove (2.1) observe from Sobolev's theorem that

$$|\nabla\psi(x)-\nabla\psi(y)|\leqq c\|\psi\|_{2,q}|x-y|^{(1-n/q)}, \tag{2.2}$$

$x, y \in \mathbb{R}^n$. Inequality (2.1) now follows from (2.2) and the argument in [8] with slight modification. Essentially the modification consists in using the estimates for bounded sub- and supersolutions in [12] rather than those in [10]. We omit the details. We assume, as we may, that $0<\lambda$, $\bar{\lambda}<\min[\frac{1}{4}, 1-(n/q)]$. In the rest of this section and §§ 3–5, σ and R_0 will be fixed. We let d denote a positive constant which may depend only on σ, R_0, and the same quantities as c in Theorem 1 (the data), not necessarily the same at each occurrence.

With ε still fixed as above, let $K=\{x: u(x)=\psi(x)\}$, and observe from (2.1b) that K is relatively closed in B. For fixed e in $\mathbb{R}^n$ with $|e|=1$, $h>0$, and g a function on $\mathbb{R}^n$, put

$$\tau_h g(x)=[g(x+he)-g(x)]/h, \qquad x\in\mathbb{R}^n,$$

$$\tau_{\bar{h}} g(x)=[g(x-he)-g(x)]/h, \qquad x\in\mathbb{R}^n.$$

Let $\xi(\gamma)=(\varepsilon+|\gamma|^2)^{(p/2-1)}\gamma$, $\gamma\in\mathbb{R}^n$, and note that

$$(\xi(\gamma)-\xi(\gamma'))\cdot(\gamma-\gamma')\geqq a(\varepsilon+|\gamma|+|\gamma'|)^{(p-2)}|\gamma-\gamma'|^2, \tag{2.3}$$

when $\gamma, \gamma'\in\mathbb{R}^n$, where $a=a(p,n)$ depends only on p and n. Next, let $\phi\in C_0^\infty[B]$ and put $A^+=\max(A,0)$. Since, $(u-\psi-\beta)^+\phi$ vanishes in a neighborhood of $K\cap\bar{B}$ for each $\beta>0$, we see upon letting $\beta\to 0$ that equality holds in (1.9) for $\eta=(u-\psi)\phi$. If $0<h<((1-\sigma)/400)R_0$, it follows from this observation and (1.9) that

$$\int_B \{\tau_h[\xi(\nabla u)]\cdot\nabla\tau_h(\eta)+\tau_{\bar{h}}[\tau_h(\eta)]b\}\,dx\leqq 0. \tag{2.4}$$

Using this inequality, (2.3), and arguing as in Tolksdorf [11, § 2.2] we deduce first that $u\in W_{2,2}[B_1]$ for $p\geqq 2$ and $u\in W_{2,p}[B_1]$ for $1<p\leqq 2$. Second, it follows as in [11, § 2.3] that $\nabla u\in L_\infty[B_1]$ for $1<p<\infty$ and thereupon that $u\in W_{2,2}(B_1)$ for $1<p<\infty$. Here the norm of u, ∇u, in each space may depend in addition to R_0, σ, and the data, on ε, the smoothness of ψ, and $\|b\|_\infty$. Additional terms in the iteration due to the obstacle are easily handled using the smoothness of ψ. Again we omit the details.

From the above facts and (1.9) we see that

$$\nabla\cdot[\xi(\nabla u(x))]-b(x,u,\nabla u)=0 \quad \text{for a.e. } x\in B_1-K. \tag{2.5}$$

Let $g_e=\nabla g\cdot e$, $e\in\mathbb{R}^n$, $|e|=1$. Then from (2.5) and integration by parts we get

$$\int[\xi(\nabla u)]_e\cdot\nabla\eta\,dx=\int(\nabla\cdot\bar{\xi})\eta_e\,dx=\int_{B_1-K} b\eta_e\,dx+\int_K(\nabla\cdot\bar{\xi})\eta_e\,dx, \tag{2.6}$$

where $\bar{\xi}(x)=\xi(\nabla u(x))$ in the last two integrals and $\eta\in\dot{W}_{1,2}[B_1]$. For almost every $x\in B_1$ we note that the ith component of $\bar{\xi}_e$ at x is

$$(\bar{\xi}_e)_i=a_{ij}(\nabla u)(u_e)_{x_j}, \qquad 1\leqq i\leqq n, \tag{2.7}$$

where repeated indexes denote summation from 1 to n and

$$a_{ij}(\gamma)=(\varepsilon+|\gamma|^2)^{(p/2-2)}[(p-2)\gamma_i\gamma_j+\delta_{ij}(\varepsilon+|\gamma|^2)] \tag{2.8}$$

for $1\leqq i, j\leqq n$, $\gamma\in\mathbb{R}^n$. In (2.8), δ_{ij} is the Kronnecker δ. Also for $\mu=(\mu_1,\cdots,\mu_n)\in\mathbb{R}^n$,

$$a_1(\varepsilon+|\gamma|^2)^{(p/2-1)}|\mu|^2\leqq a_{ij}(\gamma)\mu_i\mu_j\leqq a_2(\varepsilon+|\gamma|^2)^{(p/2-1)}|\mu|^2 \tag{2.9}$$

for some $a_1=a_1(p,n)$, $a_2=a_2(p,n)>0$.

3. L_∞ estimates on $|\nabla u|$. Let $R_2=(1/25)(24+\sigma)R_0$ and put $B_2=B(x_0,R_2)$. In this section we prove the following lemma.

LEMMA 1. *For R_0, ε, σ, p fixed as in* § 2,

$$\operatorname{ess.max}_{x\in B_2} |\nabla u(x)| \leqq d. \tag{3.1}$$

To prove (3.1) we use Moser iteration. We will need some notation. Let $\bar{E}$, ∂E, $|E|$, denote, respectively, the closure, boundary, and outer Lebesgue measure of E. If g is a function on $B(x, r)$, $r>0$, put

$$\fint_{B(x,r)} g\,dx = |B(x,r)|^{-1}\int_{B(x,r)} g\,dx,$$

$$M(r, g) = M(r, g, x) = \operatorname{ess.max}_{y\in B(x,r)} g(y).$$

The proof consists of two parts. In the first part we essentially estimate $M(R_2, |\nabla u|)$ in terms of the L_μ norm of ∇u over $B_3 = B(x_0, R_3)$, where $R_3 = (1/50)(49+\sigma)R_0$. Here, $\mu = \mu(p, n)$. The second part of the proof consists of estimating this norm in terms of the L_p norm of $|\nabla u|$ and ultimately in terms of the Hölder norm of u in B_1. Since the proof is more or less standard (see [1, § 3] or [11, § 3]), we will omit many details. Let $0<R<((1-\sigma)R_0)/1{,}000)$, $z\in B(x_0, R_2)$, $16R<\rho<t<32R$, $\phi\in C_0^\infty[B(z,t)]$, and $\phi\equiv 1$ on $B(z,\rho)$ with $|\nabla\phi|\leqq 1{,}000/(t-\rho)$. Let $w(x)=(\varepsilon+|\nabla u|^2)(x)$, $x\in B$, and let $\theta\geqq 0$ be a nondecreasing Lipschitz function on $\mathbb{R}$. We put $\eta(x)=u_e(x)\theta[w(x)]\phi^2(x)$ in (2.6) and use (2.7)-(2.9) with $e=e(k)$, $1\leqq k\leqq n$, where $\{e(k), 1\leqq k\leqq n\}$ is an orthonormal set of coordinate vectors. Summing, we get

$$\begin{aligned} I_1 &= \int w^{(p/2-1)}\left[\sum_i |\nabla u_{x_i}|^2\theta(w)+\theta'(w)|\nabla w|^2\right]\phi^2\,dx \\ &\leqq a\int w^{(p/2-1)}\theta(w)|u_{x_i}|\left(\sum_j |u_{x_ix_j}|\right)|\nabla\phi|\phi\,dx \\ &\quad + ac_1\int w^{p/2}|\nabla\eta|\,dx + ac_1\int f|\nabla\eta|\,dx + a\int_K |\nabla\cdot\xi||\nabla\eta|\,dx = I_2, \end{aligned} \tag{3.2}$$

where $a=a(p,n)$ and we have used (1.10). We first let $\theta(w)=w^s$, $s\geqq\frac{1}{2}(p+1)$ in (3.2). From

$$\begin{aligned} w^{p/2}|\nabla\eta| &\leqq w^{p/2}\{[|u_{x_ix_i}|w^s + sw^{(s-1)}|\nabla u||\nabla w|]\phi^2 + 2|\nabla u|w^s\phi|\nabla\phi|\} \\ &\leqq (4ac_1)^{-1}w^{(p/2-1)}\left[sw^{(s-1)}|\nabla w|^2 + \left(\sum_i |\nabla u_{x_i}|^2\right)w^s\right]\phi^2 \\ &\quad + a(c_1+1)[(s+1)w^{(p/2+s+1)}\phi^2 + w^{(p/2+s)}|\nabla\phi|^2], \end{aligned} \tag{3.3}$$

and Hölder's inequality, it follows that for a large enough

$$\begin{aligned} I_2 - \frac{1}{2}I_1 &\leqq c\left[\int (|\nabla\phi|^2+(s+1)w\phi^2)w^{(p/2+s)}\,dx\right] + c(s+1)\left(\int f^2 w^{(s-p/2+1)}\phi^2\,dx\right) \\ &\quad + csM(R_0, 1+|\nabla\psi|^2)^{s+p/2}\left\|\sum_{i,j}\psi_{x_ix_j}\phi\right\|_2^2 = I_3. \end{aligned} \tag{3.4}$$

In the last term of (3.4), the maximum of $1+|\nabla\psi|^2$ is relative to $B(x_0, R_0)$. Also,

$$I_1 \geqq as\int w^{(s+p/2-2)}|\nabla w|^2\phi^2\,dx \geqq as^{-1}\left(\int |\nabla w^{(s/2+p/4)}|^2\phi^2\,dx\right). \tag{3.5}$$

Combining (3.2), (3.4), and (3.5), we get

$$\|\nabla[w^{(s/2+p/4)}\phi]\|_2^2 \leqq c(s+1)I_3. \tag{3.6}$$

Let $2^* = 2n/(n-2)$, when $n > 2$. Then it follows from (3.6) and Sobolev's theorem when $n > 2$ that

$$\|w^{(s/2+p/4)}\phi\|_{2^*} \leqq c(s+1)^{1/2} I_3^{1/2}. \tag{3.7}$$

Let $\tilde{q} = \min(q, n/(1-\lambda))$ and $\bar{q} = 2\tilde{q}/(\tilde{q}-2) < 2^*$. Now using Hölder's inequality again we see from (3.7) that for $n > 2$ and $s \geqq (p+1)/2$,

$$\begin{aligned} &c(s+1)^{-1/2} R^{-n/2^*} \|w^{(s/2+p/4)}\phi\|_{2^*} \\ &\leqq J(\rho, t, R)\left[\fint_{B(z,t)} w^{(p/4+s/2)\bar{q}}\, dx\right]^{1/\bar{q}} + c^{(p/4+s/2)} k(R), \end{aligned} \tag{3.8}$$

where $\bar{f} = f + 1$,

$$k(R) = c\left[R^{(\tilde{q}-n)} \int_{B(z,32R)} \left\{\bar{f}^{\tilde{q}} + \sum_{i,j} |\psi_{x_i x_j}|^{\tilde{q}}\right\} dx\right]^{1/\tilde{q}},$$

and

$$J(\rho, t, R) = cR\left\{(t-\rho)^{-1} + \left[\int_{B(z,32R)} w^{\tilde{q}/2}\, dx\right]^{1/\tilde{q}}\right\} + k(R).$$

If $n = 2$ put $2^* = 2\bar{q}$. Then $w^{(p/4+s/2)}\phi$ is of bounded mean oscillation in $\mathbb{R}^2$(BMO) with BMO norm less than or equal to $a\|\nabla[w^{(p/4+s/2)}\phi]\|_2$. Using this fact and arguments similar to the above, we find that (3.8) is still true. We note from our assumptions on ρ, t that

$$k(R) \leqq cR^{(1-n/\tilde{q})} \leqq cR_0^{(1-n/\tilde{q})} = d. \tag{3.9}$$

From (3.9) it follows for properly chosen ρ, t, that (3.8) can be iterated in the usual way starting with $s = (p+1)/2$. Doing this we get

$$M(16R, w) \leqq \left[1 + \left(\int_{B(z,32R)} w^{\tilde{q}/2}\, dx\right)^{1/\tilde{q}}\right]^c \left[\fint_{B(z,20R)} (w+d)^\tau\, dx\right]^{1/\tau}, \tag{3.10}$$

where $\tau = (p + \frac{1}{2})\bar{q}/2$. From (3.10) we see that in order to prove (3.1) it suffices to show there exists $R = R(d) > 0$ such that

$$\left(\fint_{B(z,20R)} w^\alpha\, dx\right)^{1/\alpha} \leqq d(R)^{(2\lambda-2)}, \tag{3.11}$$

where $\alpha = \max[\tilde{q}/2, \tau]$. Indeed, from our choice of $\tilde{q}$, we then get

$$\int_{B(z,20R)} w^{\tilde{q}/2}\, dx \leqq dR^{[(\lambda-1)\tilde{q}+n]} \leqq d.$$

To prove (3.11) we need to estimate the term $\int w^{p/2+s+1}\phi^2\, dx$ in (3.4) in a different way. To do this we consider two cases depending on whether (A) $B(z, 64R) \cap K \neq \{\phi\}$ or (B) $B(z, 64R) \cap K = \{\phi\}$. If (A) holds we put $\eta = (u-\psi)\sigma(w)\phi^2$ in (1.9) where $\sigma(w) = (w^{s+1}-1)^+$, and use Hölder continuity of u, ψ, as well as Hölder's inequality, to get for $\tilde{w} = \max(w, 1)$, $0 < s \leqq \alpha$, and $R = R(d) > 0$ small enough that

$$\begin{aligned} L_1 = \int \tilde{w}^{(p/2+s+1)}\phi^2\, dx &\leqq dR^{2\lambda}\tilde{I}_1 + dR^{2\lambda} \int_{\{w \geqq 1\}} f^2 w^{(s+1-p/2)}\phi^2\, dx \\ &\quad + [dR^{2\lambda}(t-\rho)^{-2} + d]\left(\int_{B(z,t)} \tilde{w}^{(p/2+s)}\, dx\right) + dR^n, \end{aligned} \tag{3.12}$$

where $\tilde{I}_1 = \int_{\{w \geqq 1\}} w^{(p/2+s-2)} |\nabla w|^2 \phi^2 \, dx$. Next put $\theta(w) = [w^s - 1]^+$ in (3.2), $(p/2)((2^*/2)-1) \leqq s \leqq \alpha$. Arguing as in (3.2), (3.4), (3.5), and using (3.12), it follows for $R = R(d) > 0$ small enough that

$$\begin{aligned}\int |\nabla[\tilde{w}^{(s/2+p/4)}\phi]|^2 \, dx &\leqq d(t-\rho)^{-2}\left(\int_{B(z,t)} \tilde{w}^{(p/2+s)} \, dx\right) \\ &\quad + d\left[\int (f\phi)^2 \tilde{w}^{(p/2+s)} \, dx\right] \\ &\quad + d\left[\int_{B(z,t)} \sum_{i,j} (\psi_{x_i x_j})^2 \, dx\right] + dR^n.\end{aligned}$$

This inequality and Sobolev's and Hölder's inequalities imply as in (3.8) that for $(p/2)((2^*/2)-1) \leqq s \leqq \alpha$,

$$\begin{aligned} &R^{-n/2^*} \| \tilde{w}^{(s/2+p/4)}\phi \|_{2^*} \\ &\quad \leqq k(R)\left[\fint \phi^{\tilde{q}} \tilde{w}^{(p/4+s/2)\tilde{q}} \, dx\right]^{1/\tilde{q}} \\ &\quad\quad + dR(t-\rho)^{-1}\left[\fint_{B(z,t)} \tilde{w}^{(s+p/2)} \, dx\right]^{1/2} + k(R) \\ &\quad \leqq dR(t-\rho)^{-1}\left[\fint_{B(z,t)} \tilde{w}^{(s+p/2)} \, dx\right]^{1/2} + k(R)\end{aligned} \tag{3.14}$$

for R small, since (3.9) holds and $\tilde{q} < 2^*$. Now iterate (3.14), starting with $s = ((2^*/2)-1)(p/2)$. After at most $N = N(c)$ times, we obtain

$$\left(\fint_{B(z,20R)} \tilde{w}^{\alpha} \, dx\right)^{1/\alpha} \leqq d\left[\fint_{B(z,22R)} \tilde{w}^{(2^*p/4)} \, dx\right]^{4/(2^*p)} + dk(R). \tag{3.15}$$

Finally to estimate the right-hand side of (3.15) let

$$\begin{aligned} \bar{u}_{x_k} &= [u_{x_k} - (4n)^{-1}]^+, \qquad 1 \leqq k \leqq n, \\ u^*_{x_k} &= [-u_{x_k} - (4n)^{-1}]^+, \qquad 1 \leqq k \leqq n. \end{aligned}$$

First put $\eta = \bar{u}_{x_k}\phi^2$ in (2.6). By using (2.7)-(2.9) and summing, it follows that (3.2) is valid with $\theta(w) \equiv 1$ and u_{x_k} replaced by $\bar{u}_{x_k}$. Let $\bar{I}_1$ denote the left-hand side of (3.2) in this case. Using (3.2), (3.3), we get (3.4) with $s = 0$ and w replaced by $\tilde{w}$. To estimate, $\int \tilde{w}^{p/2+1}\phi^2 \, dx$, on the right-hand side of (3.4), put $(u-\psi)(\bar{u}_{x_k})^2\phi^2$, $1 \leqq k \leqq n$, in (1.9). Summing, we obtain an inequality analogous to (3.12) with $\tilde{I}_1$, L_1, replaced by $\bar{I}_1$ and

$$\bar{L}_1 = \sum_{k=1}^{n} \int w^{p/2}(\bar{u}_{x_k})^2\phi^2 \, dx.$$

Carrying out the same program for $u^*_{x_k}$, we obtain similar inequalities for L_1^*, I_1^*, which we define as above with $\bar{u}_{x_k}$ replaced by $u^*_{x_k}$. Now if $w > 1$, $0 < \varepsilon < \frac{1}{4}$, then for some k either $\bar{u}_{x_k} > (1/4n)w^{1/2}$ or $u^*_{x_k} > (1/4n)w^{1/2}$. From this fact and the above inequalities it follows that

$$\begin{aligned}\int \tilde{w}^{(p/2+1)}\phi^2 \, dx &\leqq c(R^n + \bar{L}_1 + L_1^*) \\ &\leqq dR^{2\lambda}(\bar{I}_1 + I_1^*) + [dR^{2\lambda}(t-\rho)^{-2} + d]\left(\int_{B(z,t)} \tilde{w}^{p/2} \, dx\right) \\ &\quad + dR^{2\lambda}\left(\int f^2 \tilde{w}^{(s+p/2)}\phi^2 \, dx\right) + dR^n.\end{aligned}$$

Using this inequality in the new versions of (3.2), we find that (3.13) is valid with $s=0$. From Sobolev's inequality, we conclude for properly chosen ϕ and $R=R(d)>0$ small enough that

$$\left[\fint_{B(z,22R)} \tilde{w}^{(2^*p)/4}\,dx\right]^{1/2^*} \leqq d\left[\fint_{B(z,24R)} \tilde{w}^{p/2}\,dx\right]^{1/2}. \tag{3.16}$$

The right-hand side of (3.16) can be estimated using $\eta=(u-\psi)\phi^p$ in (1.9). Doing this we deduce

$$\left(\fint \tilde{w}^{p/2}\,dx\right)^{1/2} \leqq dR^{p(1-\lambda)/2}. \tag{3.17}$$

Clearly (3.17), (3.16), and (3.15) imply (3.11) in case A. A similar proof holds in case B. We omit the details. The proof of Lemma 1 is now complete.

4. Reverse Hölder inequalities. In this section and § 5, we use a modification of the argument in [5, §§ 2–3]. The reader is advised to have this paper on hand, as we will often refer to it. We continue with the same notation as in §§ 2 and 3. Let $m=\max[p/2, 1]$, $s=m+p/2$, and put $v=w^s$. Let

$$b_{ij}(\gamma)=(\varepsilon+|\gamma|^2)^{(1-p/2)}a_{ij}(\gamma), \qquad \gamma\in\mathfrak{R}^n, \tag{4.1}$$

$1\leqq i, j\leqq n$, where a_{ij} is as in (2.8). We also put $\bar{\xi}(x)=\xi(\nabla u)(x)$ as in § 2 and

$$g_0(x)=(2s)[w^m(\bar{\xi}_{x_k}\cdot\nabla u_{x_k})+mw^{m-1}u_{x_k}(\bar{\xi}_{x_k}\cdot\nabla w)],$$

where repeated indices denote summation from 1 to n. Next put

$$g_1(x)=(2s)[b\chi_{\cdot(B_1-K)}+(\nabla\cdot\bar{\xi})\chi_{\cdot K}](u_{x_lx_l}w^m+mu_{x_l}w_{x_l}w^{(m-1)})$$

$$g_{l+1}(x)=(2s)[b\chi_{\cdot(B_1-K)}+(\nabla\cdot\bar{\xi})\chi_{\cdot K}](u_{x_l}w^m), \qquad 1\leqq l\leqq n.$$

Here, $\chi_{\cdot E}$ denotes the characteristic function of E. From (2.9) we see that

$$a^{-1}w^{(s-1)}\left(\sum_i |\nabla u_{x_i}|^2\right)\leqq g_0\leqq aw^{(s-1)}\left(\sum_i |\nabla u_{x_i}|^2\right) \tag{4.2}$$

for some $a=a(p, n)$. We prove the following lemma.

LEMMA 2. *If $\phi\in W_{1,2}(B_1)$ and $\bar{b}_{ij}(x)=b_{ij}(\nabla u(x))$, then*

$$\int \bar{b}_{ij}v_{x_j}\phi_{x_i}\,dx=\int(-g_0+g_1)\phi\,dx+\sum_{k=1}^{n}\int\phi_{x_k}g_{k+1}\,dx. \tag{4.3}$$

To prove (4.3) we put $\eta=u_{x_k}w^m\phi$ in (2.6) and choose $e=e(k)$ as in § 3. Summing, we find that

$$\begin{aligned} J_1&=\int[w^m(\bar{\xi}_{x_k}\cdot\nabla u_{x_k})+mw^{(m-1)}u_{x_k}(\bar{\xi}_{x_k}\cdot\nabla w)]\phi\,dx \\ &=-\int(u_{x_k}w^m)(\bar{\xi}_{x_k}\cdot\nabla\phi)\,dx+\int_{B_1-K}b\left(\sum_k\eta_{x_k}\right)dx+\int_K(\nabla\cdot\bar{\xi})\left(\sum_k\eta_{x_k}\right)dx \\ &=J_2+J_3+J_4. \end{aligned} \tag{4.4}$$

Now,

$$\int g_0\phi\,dx=(2s)J_1. \tag{4.5}$$

Next observe that if $\overline{a_{ij}}(x) = a_{ij}(\nabla u(x))$, then

$$\bar{b}_{ij} v_{x_j} = 2sw^{(s-1)} u_{x_k} (\bar{b}_{ij} u_{x_k x_j}) = 2sw^{(s-p/2)} u_{x_k} (\overline{a_{ij}} u_{x_k x_j}) = 2sw^m u_{x_k} (\bar{\xi}_{x_k})_i,$$

where $(\bar{\xi}_{x_k})_i$ denotes the ith component of $\bar{\xi}_{x_k}$. Using this observation, we obtain

$$-2sJ_2 = \int \bar{b}_{ij} v_{x_j} \phi_{x_i} \, dx. \tag{4.6}$$

Also,

$$\sum_k \eta_{x_k} = u_{x_k x_k} w^m \phi + m u_{x_k} w_{x_k} w^{(m-1)} \phi + u_{x_k} w^m \phi_{x_k}.$$

We put the right-hand side of this equality in the integrals defining J_3, J_4. Using (4.5), (4.6), and rearranging terms, we get Lemma 2. Next let $v_k = u_{x_k} w^{(s-1/2)}$, $1 \leqq k \leqq n$. We shall also need the following lemma.

LEMMA 3. *If* $\phi \in \dot{W}_{1,2}(B_1)$, *then for* $1 \leqq k \leqq n$

$$\left| \int \bar{b}_{ij} (v_k)_{x_j} \phi_{x_i} \, dx \leqq a \int (g_0 + G_0) |\phi| \, dx + a \int |\nabla \phi| G \, dx. \right. \tag{4.7}$$

Here $G_0(x) = (|b| \chi_{\cdot(B_1 - K)} + |\nabla \cdot \bar{\xi}| \chi_{\cdot K})(w^m \sum_{i,j} |u_{x_i x_j}|)$,

$$G(x) = [|b| \chi_{\cdot(B_1 - K)} + |\nabla \cdot \bar{\xi}| \chi_{\cdot K}] w^{m+1/2}.$$

The proof of Lemma 3 is similar to the proof of Lemma 2. We omit the details. Next we prove Lemma 4.

LEMMA 4. *There exists* $\beta = \beta(c)$, $0 < \beta < \frac{1}{4} \min(q/n - 1, 1)$ *with the property*: *If* $0 < R < ((1-\sigma)/1{,}000) R_0$, $z \in B_2$, *then*

$$\fint_{B(z,R)} g_0^{1+\beta} \, dx \leqq c \left(\fint_{B(z,2R)} g_0 \, dx \right)^{1+\beta} + c \fint_{B(z,2R)} E^{1+\beta} \, dx, \tag{4.8}$$

where $E = \tilde{w}^{(m-p/2+1)} \{ |b|^2 + \sum_{i,j} (\psi_{x_i x_j})^2 \}$.

Proof. In Lemma 4, $\tilde{w} = \max(w, 1)$ as usual. We begin by choosing orthonormal vectors $e_1, \cdots, e_{n-1}$ such that

$$\int_{B(z,3R/2)} w^{(m/2+p/4-1/2)} u_{e_k} \, dx = 0, \qquad 1 \leqq k \leqq n-1. \tag{4.9}$$

This choice is possible since if A_k denotes the integral in (4.9) with e_k replaced by x_k, $1 \leqq k \leqq n$, then the solutions to $A_i y_i = 0$ span at least an $(n-1)$-dimensional plane. Now if $e \in \{e_k\}$ we have as in (3.2) and (4.4) for $\eta = u_e w^m \phi^2$, $\phi \in C_0^\infty[B(z, 3R/2)]$,

$$\begin{aligned} H_1 = \int w^m (\bar{\xi}_e \cdot \nabla u_e) \phi^2 \, dx &\leqq \left| \int m w^{(m-1)} u_e (\bar{\xi}_e \cdot \nabla w) \phi^2 \, dx \right| \\ &+ \left| 2 \int \phi u_e w^m (\bar{\xi}_e \cdot \nabla \phi) \, dx \right| + \left| \int_{B_1 - K} b \eta_e \, dx \right| + \left| \int_K (\nabla \cdot \bar{\xi}) \eta_e \, dx \right| = H_2. \end{aligned} \tag{4.10}$$

We estimate H_2 slightly different than I_2 in § 3. Indeed, using Hölder's inequality once again, we deduce for given τ, $0 < \tau < \frac{1}{4}$,

$$\begin{aligned} H_2 - \frac{1}{2} H_1 &\leqq a\tau \int g_0 \phi^2 \, dx + a\tau^{(1-2m)} \int |u_e|^{2m} w^{(p/2-1)} |\nabla u_e|^2 \phi^2 \, dx \\ &+ a \int w^{(s-1)} (u_e)^2 |\nabla \phi|^2 \, dx + c\tau^{-1} \int E \phi^2 \, dx = H_3. \end{aligned} \tag{4.11}$$

Using (4.10), (4.11), and summing, we find that

$$\bar{H}_1 = \sum_{k=1}^{n-1} \int w^{(s-1)} |\nabla u_{e_k}|^2 \phi^2 \, dx \leqq \bar{H}_3, \tag{4.12}$$

where $\bar{H}_3$ is obtained from H_3 by summing over $e \in \{e_k\}$. If e_n is orthogonal to $\{e_k : 1 \leqq k \leqq n-1\}$ and of unit length, then from (4.2) we get

$$\int g_0 \phi^2 \, dx \leqq a\bar{H}_1 + a \int w^{(s-1)} (u_{e_n e_n})^2 \phi^2 \, dx. \tag{4.13}$$

Now for almost every $x \in B_1 - K$ we see from (2.5) that

$$w^{(s-1)} (u_{e_n e_n})^2 \leqq a w^{(m+1-p/2)} |b|^2 + a w^{(s-1)} \sum_{i=1}^{n} \sum_{j=1}^{n-1} (u_{e_i e_j})^2,$$

while if $x \in K$,

$$w^{(s-1)} (u_{e_n e_n})^2 \leqq c \sum_{i,j} (\psi_{x_i x_j})^2.$$

These inequalities and (4.13) imply that

$$\int g_0 \phi^2 \, dx \leqq a\bar{H}_1 + c \int E\phi^2 \, dx.$$

Putting this inequality in (4.12) we see for $\tau = \tau(a) > 0$ small enough that

$$\begin{aligned} \int g_0 \phi^2 \, dx \leqq c\tau^{(1-2m)} \sum_{k=1}^{n-1} \int |u_e|^{2m} w^{(p/2-1)} |\nabla u_e|^2 \phi^2 \, dx \\ \cdot \sum_{k=1}^{n-1} c \int w^{(s-1)} (u_e)^2 |\nabla \phi|^2 \, dx + c\tau^{-1} \int E\phi^2 \, dx. \end{aligned} \tag{4.14}$$

With $\tau = \tau(a)$ now fixed we can estimate the first term on the right-hand side of (4.14) by first putting $\eta = [(u_e)^+]^{(2m+1)} \phi^2$ in (2.6) and then $\eta = [(-u_e)^+]^{(2m+1)} \phi^2$ in (2.6). Using the resulting estimate in (4.14), we conclude

$$\int g_0 \phi^2 \, dx \leqq c \sum_{k=1}^{n-1} \int w^{(s-1)} (u_{e_k})^2 |\nabla \phi|^2 \, dx + c\tau^{-1} \int E\phi^2 \, dx. \tag{4.15}$$

From (4.15), (4.9), it follows as in [5, Lemma 1] that Lemma 4 is true. We omit the details.

5. Hölder continuity of $\nabla u(\cdot, \varepsilon)$. In this section we use Lemmas 1–4 to prove the following lemma for ε, R_0, σ, fixed as in § 1.

LEMMA 5. *There exists $\alpha = \alpha(d) > 0$ such that*

$$|\nabla u(x) - \nabla u(y)| \leqq d|x-y|^{\alpha}, \qquad x, y \in B(x_0, \sigma R_0). \tag{5.1}$$

Proof. Let v_k, $1 \leqq k \leqq n$, be as in Lemma 3, and put $v_0 = v$, v as in Lemma 2. Note from (4.1), (2.9) that the eigenvalues of $(\overline{b_{ij}})$ are uniformly bounded above and below. Suppose $z \in \bar{B}(x_0, \sigma R_0)$, and $0 < r \leqq ((1-\sigma)/10{,}000) R_0$. Then $v_k \in W_{1,2}(B_1)$, so by the usual variational argument there exists $h_k = h_k(\cdot, 4r)$, $0 \leqq k \leqq n$ with $h_k \in W_{1,2}[B(z, 4r)]$, $h_k = v_k$ on $\partial B(z, 4r)$ in the $W_{1,2}$ sense, and

$$Lh_k = (\overline{b_{ij}} h_{k x_j})_{x_i} = 0, \qquad 0 \leqq k \leqq n, \tag{5.2}$$

weakly in $B(z, 4r)$. Since $v_k - h_k \in \dot{W}_{1,2}[B(z, 4r)]$, $0 \leqq k \leqq n$, we have

$$\int_{B(z,4r)} \bar{b}_{ij} (h_k)_{x_i} (v_k - h_k)_{x_j} \, dx = 0,$$

so from (4.1), (2.9),

$$\begin{aligned}\int_{B(z,4r)} |\nabla h_k|^2\,dx &\leqq a\int_{B(z,4r)} |\nabla v_k|^2\,dx \\ &\leqq a\int_{B(z,4r)} \bar{b}_{ij}(v_k)_{x_i}(v_k)_{x_j}\,dx.\end{aligned} \tag{5.3}$$

From (5.3) and well-known ellipticity theory (see [5, (2.12)]), we find for $t<4r$

$$\begin{aligned}t^{2-n}\left(\sum\int_{B(z,t)} |\nabla h_k|^2\,dx\right) &\leqq a\left(\frac{t}{r}\right)^{2\mu}\left[r^{(2-n)}\left(\sum\int_{B(z,4r)} |\nabla h_k|^2\,dx\right)\right] \\ &\leqq a\left(\frac{t}{r}\right)^{2\mu}\left[\sum\int_{B(z,4r)} |\nabla v_k|^2\,dx\right] r^{2-n}\end{aligned} \tag{5.4}$$

for some μ, $0<\mu<1$, depending on the ellipticity constants. If $F_k = h_k - v_k$ in $B(z,4r)$ and $F_k = 0$ otherwise in $\mathfrak{R}^n$, we see from Lemmas 2 and 3,

$$\begin{aligned}\sum_{k=0}^{n}\left(\int |\nabla F_k|^2\,dx\right) &\leqq a\int \overline{b_{ij}}(h_k - v_k)_{x_i}(F_k)_{x_j}\,dx \\ &= -a\int (\overline{b_{ij}})(v_k)_{x_i}(F_k)_{x_j}\,dx \leqq a\int (g_0+G_0)\left(\sum_{k=0}^{n}|F_k|\right)dx \\ &\quad + a\int G\left(\sum_{k=0}^{n}|\nabla F_k|\right)dx \\ &\leqq a\int g_0\left(\sum_{k=0}^{n}|F_k|\right)dx + aM(4r, v_0)\int_{B(z,4r)} E\,dx \\ &\quad + a\int_{B(z,4r)} G^2\,dx + \frac{1}{2}\sum_{k=0}^{n}\left(\int |\nabla F_k|^2\,dx\right),\end{aligned} \tag{5.5}$$

where E is as in Lemma 4 and we have used that fact that $M(4r, F_k)\leqq 2M(4r, v_0)$, to estimate G_0. Let $\tau = (1+\beta)/\beta$ be the conjugate exponent to $1+\beta$ in Lemma 4. Then from Lemma 4 we have for $0\leqq k\leqq n$, $R = 4r$,

$$\begin{aligned}\fint_{B(z,4r)} g_0|F_k|\,dx &\leqq \left[\fint_{B(z,4r)} g_0^{1+\beta}\,dx\right]^{1/1+\beta}\left(\fint |F_k|^\tau\,dx\right)^{1/\tau} \\ &\leqq c\left(\fint_{B(z,8r)} g_0\,dx\right)\left(\fint |F_k|^\tau\,dx\right)^{1/\tau} + cQ_1,\end{aligned} \tag{5.6}$$

where

$$Q_1 \leqq aM(4r, v_0)\left[\fint_{B(z,16r)} E^{1+\beta}\,dx\right]^{1/1+\beta} \leqq aM(4r, v_0)r^{-n/\tilde{q}}\|E\chi\|_{\tilde{q}} \leqq dr^{-n/\tilde{q}}. \tag{5.7}$$

Here, $\tilde{q} = \min(n, q/2)$, $\chi = \chi_{\cdot B_2}$, and we have used Lemma 1. Using (5.6), (5.7), and the definition of G in (5.5), we deduce for $0<t\leqq 4r$,

$$\begin{aligned}\sum_{k=0}^{n}\left(\int_{B(z,t)} |\nabla F_k|^2\,dx\right) &\leqq \sum_{k=0}^{n}\left(\int_{B(z,4r)} |\nabla F_k|^2\,dx\right) \\ &\leqq d\left(\int_{B(z,16r)} g_0\,dx\right)\left(\fint |F_k|^\tau\,dx\right)^{1/\tau} + dr^{(n-2+\delta)},\end{aligned} \tag{5.8}$$

where $\delta = 2-(n/\tilde{q})>0$. Also as in [5, (3.11), (3.12)] we find that

$$M(8r, v_0)^{-1}\left[\sum_{k=0}^{n} \fint |F_k|^\tau \, dx\right]^{1/\tau} \leqq d\left[1-\frac{M(2r, v_0)}{M(8r, v_0)}+\frac{r^{1-n/2\tilde{q}}}{M(8r, v_0)}\right]^{1/\tau}. \tag{5.9}$$

Next let $\delta_1 = \frac{1}{4}\min(\delta, 2\mu)$ and suppose that

$$M(4r, v_0) > d_1 r^{\delta_1}. \tag{5.10}$$

If d_1 is large enough, it follows from ellipticity theory and an argument similar to [5, (3.9)] that

$$M(8r, v_0)\left(\int_{B(z,8r)} g_0 \, dx\right) \leqq a\left(\sum_{k=0}^{n} \int_{B(z,16r)} |\nabla v_k|^2 \, dx\right) + dr^{(n-2+\delta_1)}. \tag{5.11}$$

Using (5.9) and (5.11) in (5.8) we obtain

$$\begin{aligned}\sum_{k=0}^{n} \int_{B(z,t)} |\nabla F_k|^2 \, dx \leqq c&\left[\sum_{k=0}^{n} \int_{B(z,16r)} |\nabla v_k|^2 \, dx\right] \\ &\cdot\left[1-\frac{M(2r, v_0)}{M(8r, v_0)}+dr^{(n-2+\delta_1)}\right]^{1/\tau}+dr^{(n-2+\delta_1)},\end{aligned} \tag{5.12}$$

provided (5.10) is true. Multiplying (5.12) by t^{2-n} and adding the resulting inequality to (5.4) we obtain an inequality which can be iterated as in [5, p. 857], even though we now must also consider the case when (5.10) is false. Doing this and using Poincaré's inequality we find (see [5, 3.22]) that

$$\sum_{k=0}^{n} \fint_{B(z,\rho)} (v_k-\eta_k)^2 \, dx \leqq d\left(\frac{\rho}{R_0}\right)^{\delta_2}, \tag{5.13}$$

for some $\delta_2 = \delta_2(d) > 0$ and $0<\rho<((1-\sigma)/1{,}000)R_0$. Since $z \in B(x_0, \sigma R_0)$ is arbitrary it follows from (5.13) and Lemma 1, in a well-known way, that Lemma 5 is valid.

6. Proof of Theorem 1. We now construct $u(\cdot, \varepsilon)$ for $0<\varepsilon<1/10\times \min[1, \delta(E, \partial\Omega)]$ and show for sufficiently small $R_0>0$ that $\{u(\cdot, \varepsilon)\}$ satisfies (1.9) and (1.11). To do so, let $\phi \in C_0^\infty(\mathfrak{R}^n)$ with $\int \phi \, dx = 1$. Put $\phi_t(x) = t^{-n}\phi(x/t)$, $x \in \mathfrak{R}^n$, and let

$$\psi(x, \varepsilon) = \int_{\mathfrak{R}^n} \hat{\psi}(x-y)\phi_\varepsilon(y) \, dy = \hat{\psi} * \phi_\varepsilon(x), \qquad x \in \mathfrak{R}^n.$$

Then $\psi \in C^\infty(\mathfrak{R}^n)$ and (1.8) is valid. As in Tolksdorf [11, § 5] we let

$$\bar{b}(x, z, h) = \operatorname{sgn} \hat{b}(x, \hat{u}(x), \nabla\hat{u}(x)) \min\{|\hat{b}(x, \hat{u}(x), \nabla\hat{u}(x))|, 2c_1[|h|^p + f(x)]\}$$

for $(x, z, h) \in B \times \mathfrak{R} \times \mathfrak{R}^n$, where c_1, f are as in (1.3). Now put

$$b_\varepsilon(x, z, h) = \operatorname{sgn} \bar{b}(x, z, h) \min\left\{|\bar{b}(x, z, h)|, \frac{1}{\varepsilon}\right\}$$

for $(x, z, h) \in B \times \mathfrak{R} \times \mathfrak{R}^n$. Clearly, (1.10) holds. We claim there exists a solution $u = u(\cdot, \varepsilon)$ to (1.9) with $u - \hat{u} * \phi_\varepsilon \in \dot{W}_{1,p}(B)$, $u \geqq \psi$ in B, almost everywhere, and $u \leqq A < \infty$. The existence of $u(\cdot, \varepsilon)$ follows from standard ellipticity theory using the boundedness of b_ε (see [4, Chap. 10]). Next we prove the following lemma.

LEMMA 6. *There exists* $\rho_1 = \rho_1(c) > 0$, $\mu = \mu(c) > 0$, *depending only on the data such that if* $0 < R_0 \leqq \rho_1$, *then*

$$|u(x)| \leqq c, \qquad x \in B, \tag{6.1}$$

$$\left(\int_B |\nabla u|^p \, dx \right)^{1/p} \leqq c(\|\hat{u}\|_{1,p} + 1), \tag{6.2}$$

$$\operatorname*{osc}_B (u) \leqq c R_0^{\mu}. \tag{6.3}$$

To prove Lemma 6 let $\beta = M(R_0, \hat{u} * \phi_\varepsilon)$ and put $\eta = e^{\gamma(u-\beta)^+} - 1$. Then $\eta \in \dot{W}_{1,p}(B)$ and for this η equality holds in (1.9), since $\hat{u} * \phi_\varepsilon \geqq \psi(\cdot, \varepsilon)$ in B. Hence,

$$\begin{aligned} \gamma \int_{\{u \geqq \beta\}} |\nabla u|^p \, e^{\gamma(u-\beta)^+} \, dx &\leqq \int_B |b_\varepsilon| \eta \, dx + \gamma \int_B (\eta + 1) \, dx \\ &\leqq 2c_1 \int_{\{u \geqq \beta\}} |\nabla u|^p \, e^{\gamma(u-\beta)^+} \, dx \\ &\quad + (2c_1 + \gamma) \int_B \bar{f}(|\eta| + 1) \, dx, \end{aligned} \tag{6.4}$$

where again $\bar{f} = f + 1$. If $\gamma \geqq 4c_1$, then from (6.4) we conclude

$$\begin{aligned} \int_{\{u \geqq \beta\}} |\nabla \, e^{(\gamma/p)(u-\beta)^+}|^p \, dx &\leqq c\gamma^p \int_B \bar{f}(|\eta| + 1) \, dx \\ &\leqq c\gamma^p \left(\int_B \bar{f}(x)^q \, dx \right)^{1/q} \left[\int_B e^{\bar{q}\gamma(u-\beta)^+} \, dx \right]^{1/\bar{q}}, \end{aligned} \tag{6.5}$$

where $\bar{q} = q/(q-1) < n/(n-1)$.

We now consider the following cases (a) $p > n$, (b) $1 < p < n$, (c) $p = n$. In case (a), it follows from Sobolev's theorem and (6.5) that

$$\begin{aligned} \|e^{(\gamma/p)(u-\beta)^+} - 1\|_\infty &\leqq a \left[\int_B |\nabla \, e^{(\gamma/p)(u-\beta)^+}|^p \, dx \right]^{1/p} R_0^{1-n/p} \\ &\leqq c\gamma \|e^{(\gamma/p)(u-\beta^+)}\|_\infty R_0^{\xi}, \end{aligned}$$

where $\xi = 1 - (n/p) + (n/\bar{q}p)$. If $\gamma = 4c_1$ in the above inequality, then we first deduce that $(u-\beta)^+ \leqq c$ for ρ_1 small enough and second for $p > n$ that

$$u(x) \leqq \beta + cR_0^{\xi}, \qquad x \in B. \tag{6.6}$$

If $1 < p < n$, we use Sobolev's theorem again to deduce for $p^* = np/(n-p)$,

$$\left[\int_B (e^{(\gamma/p)(u-\beta)^+} - 1)^{p^*} \, dx \right]^{1/p^*} \leqq c\gamma R_0^{\sigma} \left[\int_B e^{\gamma\bar{q}(u-\beta)^+} \, dx \right]^{1/\bar{q}p}, \tag{6.7}$$

where $\sigma = (n/\bar{q}p) - (n/p^*) > 0$, since $(p^*/p) > \bar{q}$. If $p = n$, then $e^{(\gamma/p)(u-\beta)^+} - 1 \in$ BMO $(\mathfrak{R}^n)$ with BMO norm bounded by the $W_{1,n}$ norm of this function. Using this fact, we see that (6.7) remains valid with $p^* = 2\bar{q}n$. If $w = e^{\bar{q}(u-\beta)^+}$, $\alpha = p^*/(p\bar{q}) > 1$, then from (6.7) we deduce for $1 < p \leqq n$ and $\gamma \geqq 4c_1$,

$$\left[\fint_B w^{\gamma\alpha} \, dx \right]^{1/\alpha\gamma} \leqq c(\gamma^{1/\gamma})^{p\bar{q}} \left[\fint_B w^{\gamma} \, dx \right]^{1/\gamma}.$$

Iterating this inequality we conclude

$$M(R_0, w) \leqq c\left[\fint_B w^{\gamma_0}\, dx\right]^{1/\gamma_0},$$

where $\gamma_0 = 4c_1$. By using (6.7) again, it follows that

$$M(R_0, w)^{1/p} \leqq c\gamma_0 R_0^{\sigma} M(R_0, w)^{1/p} + c,$$

and hence

$$(u-\beta)^+ \leqq c \tag{6.8}$$

for $R_0 \leqq \rho_1$ in cases (b) and (c). To prove an analogue of (6.6) for $1 < p \leqq n$, we use (6.8). If $v = e^{\gamma_0(u-\beta)^+} - 1$, $\eta = v^s$, $s \geqq 1$, then as in (6.4), (6.5), we deduce using (6.8),

$$\int_B |\nabla v^{(s-1)/p+1}|^p\, dx \leqq cs^p \left[\int_B v^{(s-1)\bar{q}}\, dx\right]^{1/\bar{q}}.$$

From this inequality we see that the argument following (6.5) can be repeated to get

$$M(R_0, v) \leqq c\left[\int_B v^{s_0}\, dx\right]^{1/2s_0} R_0^{(1/2)\sigma_1} + cR_0^{\sigma_1} \leqq cM(R_0, v)^{1/2} R_0^{(1/2)\sigma_1} + cR_0^{\sigma_1}$$

for some $s_0 = s_0(c)$, $\sigma_1 = \sigma_1(c) > 0$. Clearly, this inequality implies for $\rho_1 > 0$ small enough that

$$u(x) \leqq \beta + cR_0^{\sigma_1}, \qquad x \in B, \tag{6.9}$$

when $1 \leqq p \leqq n$. Next put $\delta = \min_B (\hat{u} * \phi_\varepsilon)$ and let $\eta = e^{\gamma(\delta-u)^+} - 1 \in \dot{W}_{1,p}(B)$. Using (1.9) it follows that for $\gamma \geqq 4c_1$,

$$\int_B |\nabla e^{(\gamma/p)(\delta-u)^+}|^p\, dx \leqq c\gamma^p \left[\int_B e^{\bar{q}(\delta-u)^+}\, dx\right]^{1/\bar{q}}.$$

Repeating the argument following (6.5) we first get (6.6) with u replaced by $-u$ and β by $-\delta$, for $p > n$. Second, we find $(\delta - u)^+ \leqq c$ for $1 \leqq p \leqq n$. Using this fact, putting $\eta = (e^{(\delta-u)^+} - 1)^s$ in (1.9), and arguing as in the proof of (6.9) we deduce that (6.9) holds with u replaced by $-u$ and β by $-\delta$ when $1 < p \leqq n$. Combining these inequalities and using (1.5), (2.1a) we conclude first that

$$M(R_0, |u|) \leqq M(R_0, |\hat{u} * \phi_\varepsilon|) + cR_0^{\mu} \leqq c,$$

and second that

$$\operatorname*{osc}_B (u) \leqq \operatorname*{osc}_B (\hat{u} * \phi_\varepsilon) + CR_0^{\mu} \leqq cR_0^{\mu},$$

where $\mu = \min(\sigma_1, \bar{\lambda}, \xi)$. Thus (6.1) and (6.3) of Lemma 6 are true for ρ_1 small enough.

To prove (6.2) put $\eta = \hat{u} * \phi_\varepsilon - u$ in (1.9). If $\bar{u} = \hat{u} * \phi_\varepsilon$ it follows from (6.3), (1.5), and the definition of b_ε that

$$\begin{aligned} J_1 &= \int_B (\varepsilon + |\nabla u|^2)^{(p/2-1)} |\nabla u|^2\, dx \\ &\leqq c \int_B (\varepsilon + |\nabla u|^2)^{(p/2-1/2)} |\nabla \bar{u}|\, dx + \int |b_\varepsilon||\eta|\, dx \\ &\leqq \frac{1}{2} J_1 + c \int_B (|\nabla \bar{u}|^p + 1)\, dx + cR^{\mu} \int_B (|\nabla \bar{u}|^p + |\nabla \hat{u}|^p + f)\, dx. \end{aligned}$$

Since, $\int_B |\nabla \bar{u}|^p \leqq a \int_\Omega |\nabla \hat{u}|^p \, dx$, we see from the above inequality that (6.2) also holds when $\rho_1 = \rho_1(c) > 0$ is small enough. This completes the proof of Lemma 6.

We now prove Theorem 1. From Lemma 6, we see there is a subsequence $u_j = u(\cdot, \varepsilon_j)$ of $\{u(\cdot, \varepsilon)\}$ such that u_j converges weakly to a function v with $v - \hat{u}$ in $\dot{W}_{1,p}(B)$. Also from (6.1) we see that Lemmas 1 and 5 can be used to deduce that u_j, ∇u_j, converge uniformly to v, ∇v on compact subsets of B and that these functions are also Hölder continuous on compact subsets of B. We now show there exists $\rho_2 < \rho_1$ depending only on the data such that if $0 < R_0 \leqq \rho_2$, then $v = \hat{u}$. To do this, suppose $v + \zeta \geqq \hat{\psi}$, where $\zeta \in \dot{W}_{1,p}(B)$ is continuous and vanishes in a neighborhood of ∂B. Put

$$\bar{\zeta}_j(x) = \max [\psi(x, \varepsilon_j) - u(x, \varepsilon_j), \zeta(x)], \qquad x \in B,$$

$j = 1, 2, \cdots$. Then supp. $\bar{\zeta}_j \subset$ (supp. $\zeta \cap B$), $j = 1, 2, \cdots$, since $\psi(\cdot, \varepsilon) \leqq u(\cdot, \varepsilon)$ in B. Also, $\bar{\zeta}_j \in \dot{W}_{1,p}(B)$ and (1.9) holds with $\varepsilon = \varepsilon_j$, $\eta = \bar{\zeta}_j$, $j = 1, 2, \cdots$. Using these facts and uniform convergence of ∇u_j, u_j, to ∇v, v on compact subsets of B, we see that

$$\int [|\nabla v|^{(p-2)} \nabla v \cdot \nabla \zeta + \bar{b}(x, v, \nabla v) \zeta] \, dx \geqq 0. \tag{6.10}$$

Suppose now that $\zeta_j \in C_0^\infty(B)$ and $\|\zeta_j + v - \hat{u}\|_{1,p} \to 0$ as $j \to \infty$. Let $\bar{\zeta}_j = \min [\max (\hat{\psi} - v, \zeta_j), 2c]$, where c is chosen so large that $\|v - \hat{u}\|_\infty \leqq c$. Then $\bar{\zeta}_j \to \hat{u} - v$ in the norm of $W_{1,p}(B)$, and $\|\bar{\zeta}_j\|_\infty \leqq c$, $j = 1, 2, \cdots$. Since $\{b_\varepsilon\}$ is Lebesgue dominated by an integrable function, we deduce from these facts and (6.10) with $\zeta = \bar{\zeta}_j, j = 1, 2, \cdots$, that

$$\begin{aligned} &\int_B [|\nabla v|^{(p-2)} \nabla v \cdot \nabla(\hat{u} - v) + \bar{b}(x, v, \nabla v)(\hat{u} - v)] \, dx \\ &\quad = \lim_{j \to \infty} \int [|\nabla v|^{(p-2)} \nabla v \cdot \nabla \bar{\zeta}_j + \bar{b}(x, v, \nabla v) \bar{\zeta}_j] \, dx \geqq 0. \end{aligned} \tag{6.11}$$

Interchanging the roles of $\hat{u}$ and v, we also get

$$\int_B |\nabla \hat{u}|^{(p-2)} \nabla \hat{u} \cdot \nabla(v - \hat{u}) + \hat{b}(x, \hat{u}, \nabla \hat{u})(v - \hat{u}) \, dx \geqq 0. \tag{6.12}$$

By using (6.11), (6.12), it follows that

$$\begin{aligned} &\int_B (|\nabla \hat{u}| + |\nabla v|)^{(p-2)} |\nabla \hat{u} - \nabla v|^2 \, dx \\ &\quad \leqq a \int_B [|\nabla \hat{u}|^{(p-2)} \nabla \hat{u} - |\nabla v|^{(p-2)} \nabla v] \cdot \nabla(\hat{u} - v) \, dx \\ &\quad \leqq a \int_B |\bar{b}(x, v, \nabla v) - \hat{b}(x, \hat{u}, \nabla \hat{u})| \, |\hat{u} - v| \, dx \\ &\quad \leqq c R_0^\mu \int_B |\bar{b}(x, v, \nabla v) - \hat{b}(x, \hat{u}, \nabla \hat{u})| \, dx, \end{aligned} \tag{6.13}$$

where we have used Lemma 6 in the last inequality. Now $\hat{b}(x, \hat{u}(x), \nabla \hat{u}(x)) = \bar{b}(x, v(x), \nabla v(x))$ for $x \in B$ unless

$$2c_1(f(x) + |\nabla v(x)|^p) < \hat{b}(x, \hat{u}(x), \nabla \hat{u}(x)) \leqq c_1(f(x) + |\nabla \hat{u}(x)|^p),$$

thanks to (1.3). Thus

$$f(x)+|\nabla v(x)|^p \leqq \tfrac{1}{2}(f(x)+|\nabla u(x)|^p),$$

and so for $x \in B$,

$$|\hat{b}(x, \hat{u}(x), \nabla \hat{u}(x)) - \bar{b}(x, v(x), \nabla v(x))| \leqq a|\nabla \hat{u}(x) - \nabla v(x)|^2(|\nabla \hat{u}(x)| + (|\nabla v(x)|)^{(p-2)}.$$

Putting this inequality in (6.13), we find that

$$\int_B |\nabla \hat{u}| + |\nabla v|)^{p-2}|\nabla \hat{u} - \nabla v|^2 \, dx \leqq \frac{1}{2}\int_B |\nabla \hat{u} - \nabla v|^2(|\nabla \hat{u}| + |\nabla v|)^{(p-2)} \, dx,$$

provided $0 < R_0 \leqq \rho_2$, and $\rho_2(c) > 0$ is sufficiently small. Hence, $\nabla \hat{u} = \nabla v$ almost everywhere in B. Since $\hat{u} - v \in \dot{W}_{1,p}(B)$ we conclude $\hat{u} \equiv v$. Finally to get Theorem 1 take $R_0 = \rho_2(c)$ and put $\sigma = \frac{1}{2}$ in Lemmas 1 and 5. Then all constants depend only on the data and since $\hat{u}$, $\nabla \hat{u}$, are uniform limits on compact subsets of $u(\cdot, \varepsilon)$, $\nabla u(\cdot, \varepsilon)$, respectively, we see that Theorem 1 holds in $B(x_0, \frac{1}{2}R_0) \cap E$. Since E can be covered by at most $N = N(c)$ such balls, we conclude that Theorem 1 is true.

REFERENCES

[1] E. DI BENEDETTO, *$C^{1+\alpha}$ local regularity of weak solutions of degenerate elliptic equations*, Nonlinear Analysis Theory, Methods, Appl., 7 (1983), pp. 827–850.

[2] M. FUCHS, *Holder continuity of the gradient for degenerate variational inequalities*, to appear.

[3] N. GAROFALO AND J. LEWIS, *A symmetry result related to some over determined boundary value problems*, Amer. J. Math., 111 (1989), pp. 9–33.

[4] D. GILBARG AND N. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1977.

[5] J. LEWIS, *Regularity of the derivatives of solutions to certain degenerate elliptic equations*, Indiana Univ. Math. J., 32 (1983), pp. 849–858.

[6] P. LINDQVIST, *Regularity for the gradient of the solution to a nonlinear obstacle problem with degenerate ellipticity*, Nonlinear Anal., 12 (1988), pp. 1245–1255.

[7] J. MANFREDI, *Regularity for minima of functionals with p-growth*, J. Differential Equations, 76 (1988), pp. 203–212.

[8] J. MICHEL AND W. ZIEMER, *Interior regularity for solutions to obstacle problems*, Nonlinear Anal., 10 (1986), pp. 1427–1448.

[9] T. NORANDO, *$C^{1,\alpha}$ local regularity for a class of quasi linear elliptic variational inequalities*, Boll. U.M.I. Anal. Funzionale Appl., Serievi, 5 (1986), pp. 281–291.

[10] J. SERRIN, *Local behavior of solutions of quasi-linear elliptic equations*, Acta Math., 111 (1964), pp. 247–302.

[11] P. TOLKSDORF, *Regularity for a more general class of quasi-linear elliptic equations*, J. Differential Equations, 51 (1984), pp. 126–150.

[12] N. TRUDINGER, *On Harnack type inequalities and their application to quasilinear elliptic equations*, Comm. Pure Appl. Math., 20 (1967), pp. 721–747.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 639-652, May 1991

005

HOMOGENIZATION WITH SMALL PERFORATIONS OF INCREASINGLY COMPLICATED SHAPES*

ALAIN DAMLAMIAN† AND PATRIZIA DONATO‡

Abstract. The limit case for a domain perturbed with many small and periodically distributed inclusions (holes or cracks) with increasingly complicated shapes is studied for the Poisson problem. On each "perforation" the boundary condition relates the value of the unknown function, assumed to be constant, with the total flux of the same unknown.

The limit problem involves an extra term of order zero whose coefficient depends on all the parameters, in particular the capacity of the normalized inclusion and the measure of its boundary (or their behaviors as the size of the mesh goes to zero). Some examples involving fractal sets and cracks are given.

Key words. Poisson equation, homogenization, small inclusions, perforated domains, fractal cracks

AMS(MOS) subject classifications. 35J20, 35R05, 35J70

Introduction. A well-known application of homogenization theory consists in replacing a partial differential equation which has highly oscillatory coefficients or is posed on an oscillatory perturbation of a fixed domain by a smooth partial differential equation on a fixed domain (see, for example, De Giorgi and Spagnolo [12], Tartar [20], Sanchez-Palencia [19], Bensoussan, Lions, and Papanicolaou [3]).

In this paper we study the limit case for a perforated domain with many small and periodically distributed holes with increasingly complicated shapes. More precisely, we consider a domain Ω, perforated in an ε-periodic fashion by holes or cracks of size r_ε, similar to a reference hole or crack T_ε of unit size.

For simplicity, we consider the Poisson equation with homogeneous Dirichlet condition on the boundary Ω. On each perforation, we consider a boundary condition relating the total flux of the unknown to its value on the perforation (see (1.2)-(1.3)). This is the usual variational formulation on the boundary of a conductor in electrostatics. The purpose here is to consider holes and cracks with very complicated shapes, including some fractal behaviors (see § 6).

We prove that in general the limit problem is posed in the whole of Ω, with an extra term of order zero appearing in the limit equation, whose constant coefficient depends on all the parameters (see Theorem 4.1 and Remark 4.2), in particular on the limit behavior of the capacity of T_ε and of the measure of its boundary.

This problem was first introduced in Kaïzu [15], where a first partial answer is proposed. If the reference hole or crack T_ε does not vary with respect to ε (i.e., $T_\varepsilon \equiv T$), the similar problem with pointwise Dirichlet or Robin condition on the boundary of the perforations has already been studied by many authors (see Marchenko and Hruslov [17], Rauch and Taylor [18], Carbone and Colombini [5], Dal Maso and Longo [9], Cioranescu and Murat [7], Attouch and Picard [2], Conca [8], Cioranescu and Donato [6], Brillard [4], Kaïzu [14]).

* Received by the editors January 24, 1990; accepted for publication (in revised form) April 11, 1990.

† Centre de Mathématiques, Ecole Polytechnique, 91128 Palaiseau Cedex, France, and Université Paris 12, Paris Val-de-Marne, 94010 Créteil Cedex, France.

‡ Dipartimento di Matematica e Applicazioni, Università di Napoli, Via Mezzocannone 8, 80134 Napoli, Italy. The work of this author was partially supported by the Consiglio Nazionale delle Ricerche and the Centre National de la Recherche Scientifique at the Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, Paris, France.

Our method could be generalized to Robin conditions in the case of varying perforations only if the norm of the exterior trace operator on ∂T_ε and the exterior Wirtinger–Poincaré constant associated to T_ε are bounded with respect to ε. This seems to be false in the most interesting case, i.e., in the fractal construction of T_ε.

The plan of this paper is as follows. Section 1 gives notation and position of the problem; § 2 gives an approximation lemma; § 3 gives some auxiliary results; § 4 gives the asymptotic behavior; § 5 is a summary of the results; § 6 gives some fractal examples.

1. Notation and position of the problem. In what follows we will use the following notation:

ε = a positive parameter taking values in a sequence going to zero.
r_ε = a positive parameter such that $r_\varepsilon < \varepsilon/2$.
Ω = a bounded open set of $\mathbb{R}^N$, $N \geqq 2$.
B_r = the ball of $\mathbb{R}^N$ of radius r centered at the origin.
$B^k_{\eta_\varepsilon} = B_{\eta_\varepsilon} + 2\varepsilon k$, $k \in Z^N$, for $0 < \eta_\varepsilon \leqq \varepsilon$.
$\mathcal{K}_\varepsilon = \{k \in Z^N \mid \overline{B^k_{2r_\varepsilon}} \subset \Omega\}$.
$N(\varepsilon) = \operatorname{card} \mathcal{K}_\varepsilon$.
$\mathcal{B}_{\eta_\varepsilon} = \bigcup_{k \in \mathcal{K}_\varepsilon} B^k_{\eta_\varepsilon}$.
T_ε = an open set in B_1 with Lipschitz boundary ∂T_ε.
$T^k_\varepsilon = r_\varepsilon T_\varepsilon + 2\varepsilon k$, $k \in Z^N$.
$\mathcal{T}_\varepsilon = \bigcup_{k \in \mathcal{K}_\varepsilon} T^k_\varepsilon$.
$\Omega_\varepsilon = \Omega \setminus \overline{\mathcal{T}_\varepsilon}$.
χ_A = the characteristic function of A.
$\fint_A f$ denotes the mean value of f over the set A.
We will suppose in the following that

$$r_\varepsilon < \frac{\varepsilon}{2}, \quad \frac{r_\varepsilon}{\varepsilon} \to 0 \quad \text{as } \varepsilon \to 0. \tag{1.1}$$

Thus, Ω_ε denotes the perforated domain obtained removing the set $\mathcal{T}_\varepsilon$ of the "holes" from the domain Ω.

The holes are then periodically distributed in Ω with 2ε-periodicity and, for every ε, each hole is r_ε-homothetic to the unitary hole T_ε. See Fig. 1.

By (1.1) the size of the holes is of order smaller than the period. So, when $\varepsilon \to 0$, the number of the holes $N(\varepsilon) \to \infty$ as ε^{-N} and the N-dimensional measure of $\mathcal{T}_\varepsilon \to 0$ while the shape of the holes varies in general with ε.

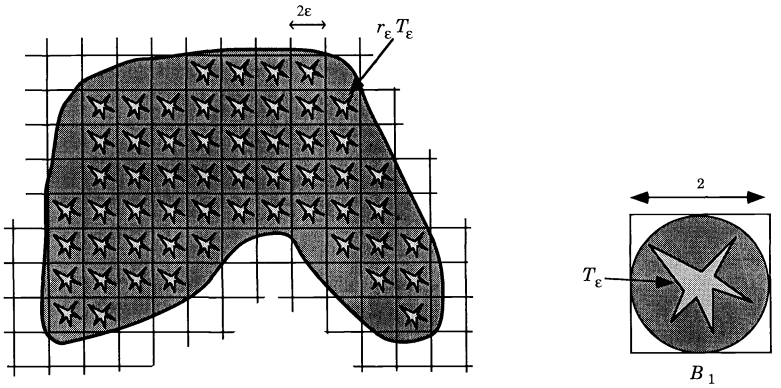


FIG. 1. *The set Ω_ε together with the normalized T_ε.*

Now let $a_\varepsilon \in [0, +\infty[$ and $f \in L^2(\Omega)$ and set for $p \in [1, +\infty]$

$$V^p = W_0^{1,p}(\Omega), \qquad V = V^2,$$

$$V_\varepsilon^p = \{u \in V^p \mid u \text{ constant on each } T_\varepsilon^k,\ k \in \mathcal{K}_\varepsilon \text{ in } T_\varepsilon\}, \qquad V_\varepsilon = V_\varepsilon^2.$$

It is easy to check that V_ε^p is a closed subspace of V^p, and consequently a Banach space when endowed with the $W_0^{1,p}(\Omega)$-norm.

Consider the problem:

$$\begin{aligned} &-\Delta u_\varepsilon = f \quad \text{in } \Omega_\varepsilon, \\ &u_\varepsilon = 0 \quad \text{on } \partial\Omega, \\ &\int_{\partial T_\varepsilon^k} \left(\frac{\partial u_\varepsilon}{\partial n} + a_\varepsilon u_\varepsilon\right) d\sigma = 0 \quad \forall k \in \mathcal{K}_\varepsilon, \\ &u_\varepsilon = V_\varepsilon, \end{aligned} \tag{1.2}$$

where n denotes the outward unit normal to Ω_ε.

The variational formulation of problem (1.2) is

$$\text{Find } u_\varepsilon \in V_\varepsilon \text{ such that}$$
$$\int_{\Omega_\varepsilon} \nabla u_\varepsilon \nabla \varphi \, dx = -\int_{\partial \mathcal{T}_\varepsilon} a_\varepsilon u_\varepsilon \varphi \, d\sigma + \int_{\Omega_\varepsilon} f\varphi \, dx \quad \forall \varphi \in V_\varepsilon. \tag{1.3}$$

The existence and uniqueness of the solution of (1.3) follows from Lax-Milgram's theorem, observing that

$$\|u_\varepsilon\|_{V_\varepsilon} \leqq C(V)\|\nabla u_\varepsilon\|_{L^2(\Omega_\varepsilon)} \leqq C(V)\|u_\varepsilon\|_{V_\varepsilon} \tag{1.4}$$

where $C(V)$ is the Poincaré constant of V.

Our purpose is to study the asymptotic behavior of the solutions u_ε as $\varepsilon \to 0$. It is clear from (1.3), (1.4) that u_ε is bounded in V, so that, by extraction of a subsequence, we can assume that there exists a u in V such that

$$u_\varepsilon \rightharpoonup u \quad \text{weakly in } V. \tag{1.5}$$

We are left with the task of identifying such u's, and determining the corresponding "limit problems," which is the aim of the remaining sections of this paper.

2. An approximate lemma. The following lemma shows that V is the proper variational space for the possible limit problems (see Kaïzu [15]).

LEMMA 2.1. *For every $p < \infty$ and $v \in V^p$, there exists a sequence $\{v_\varepsilon\}$ such that:*

(i) *For all ε, $v_\varepsilon \in V_\varepsilon^p$,*

(ii) *$v_\varepsilon \to v$ strongly in V^p,*

(iii) *If $v \in C^0(\bar{\Omega})$, then $v_\varepsilon \to v$ strongly in $C^0(\bar{\Omega})$.*

Proof. Let $\varphi \in \mathcal{D}(B_2)$ be such that $\varphi \equiv 1$ on B_1, and set

$$v_\varepsilon(x) = v(x) - \sum_{k \in \mathcal{K}_\varepsilon} \varphi\left(\frac{x-k}{r_\varepsilon}\right)\left(v(x) - \fint_{B_{r_\varepsilon}^k} v\right). \tag{2.1}$$

By the definition of $\mathcal{K}_\varepsilon$ we deduce that v_ε belongs to V_ε. We have

$$\begin{aligned} \|\nabla(v_\varepsilon - v)\|_{L^p(\Omega)}^p \leqq \sum_{k \in \mathcal{K}_\varepsilon} \Bigg[&\frac{1}{r_\varepsilon^p}\|\nabla\varphi\|_{L^\infty(B_2)}^p \left\| v - \fint_{B_{r_\varepsilon}^k} v \right\|_{L^p(B_{2r_\varepsilon}^k)}^p \\ &+ \|\varphi\|_{L^\infty(B_2)}^p \|\nabla v\|_{L^\infty(B_{2r_\varepsilon}^k)}^p \Bigg]. \end{aligned} \tag{2.2}$$

A straightforward modification of the classical Wirtinger-Poincaré-Friedrichs inequality for $W^{1,p}(B_2)$ yields a constant $C(B_2, B_1)$ such that

$$\forall \psi \in W^{1,p}(B_2) \quad \int_{B_2} \left[\psi - \fint_{B_1} \psi \right]^p dx \leqq C(B_2, B_1) \int_{B_2} |\nabla \psi|^p \, dx. \tag{2.3}$$

It follows that for every u in V^p satisfying $\fint_{B^k_{r_\varepsilon}} u = 0$ we have

$$\begin{aligned} \int_{B^k_{2r_\varepsilon}} u^p \, dx &= r_\varepsilon^N \int_{k+B_2} u^p(r_\varepsilon y) \, dy \\ &\leqq r_\varepsilon^N C(B_2, B_1) \int_{k+B_2} |\nabla_y u(r_\varepsilon y)|^p \, dy \\ &\leqq r_\varepsilon^p C(B_2, B_1) \int_{B^k_{2r_\varepsilon}} |\nabla u|^p \, dx. \end{aligned} \tag{2.4}$$

By (2.2) and (2.4), it follows that

$$\|\nabla(v_\varepsilon - v)\|^p_{L^p(\Omega)} \leqq C \|\nabla v\|^p_{L^p(\mathscr{B}_{2r_\varepsilon})}. \tag{2.5}$$

A similar calculation shows that

$$\|v_\varepsilon - v\|^p_{L^p(\Omega)} \leqq C r_\varepsilon^p \|\nabla v\|^p_{L^p(\mathscr{B}_{2r_\varepsilon})}. \tag{2.6}$$

Since by (1.1), the measure of $\mathscr{B}_{2r_\varepsilon} \to 0$ as $\varepsilon \to 0$, this implies (i) and (ii) of Lemma 2.1, while (iii) follows easily from the definition of v_ε, making use of the modulus of uniform continuity of v. □

Remark 2.2. In the case where r_ε is of the same order as ε, (2.5) and (2.6) still show the weak convergence of v_ε to v in V^p.

In this case, however, strong convergence does not hold for nonconstant v. Indeed if ∇v_ε converges strongly in $L^p(\Omega)$, then $\chi_{\mathscr{T}_\varepsilon} \nabla v_\varepsilon = 0$ in Ω together with the fact that $\chi_{\mathscr{T}_\varepsilon}$ converges to a nonzero constant weakly star in $L^{p'}(\Omega)$ implies that $\nabla v = 0$.

Remark 2.3. Actually, Lemma 2.1 yields the convergence of V_ε to V in the classical sense of Kuratowski [16].

Remark 2.4. We constructed the v_ε's to be constant in the balls $\mathscr{B}_{r_\varepsilon}$ in order to avoid the difficulty of estimating the Wirtinger-Poincaré constant of T_ε.

3. Some auxiliary results. In this section we introduce an auxiliary problem in order to construct test functions used in the process of passing to the limit in problem (1.2) as $\varepsilon \to 0$.

By (1.1) it follows that $\overline{T_\varepsilon} \subset B_{\varepsilon/r_\varepsilon}$. Consider the problem:

$$\begin{aligned} &-\Delta W_\varepsilon = 0 \quad \text{in } B_{\varepsilon/r_\varepsilon} \setminus \overline{T_\varepsilon}, \\ &W_\varepsilon = 0 \quad \text{on } \partial B_{\varepsilon/r_\varepsilon}, \\ &\int_{\partial T_\varepsilon} \left(\frac{\partial W_\varepsilon}{\partial n} + r_\varepsilon a_\varepsilon (W_\varepsilon - 1) \right) d\sigma = 0, \\ &W_\varepsilon \text{ constant on } T_\varepsilon, \\ &W_\varepsilon \in H^1(B_{\varepsilon/r_\varepsilon}). \end{aligned} \tag{3.1}$$

Set

$$A_\varepsilon = r_\varepsilon a_\varepsilon |\partial T_\varepsilon| \tag{3.2}$$

and

$$\text{cap}_\varepsilon = \text{cap}\,(T_\varepsilon, B_{\varepsilon/r_\varepsilon}) \tag{3.3}$$

where the capacity is in the H^1-sense, i.e.,

For all $A \subset B$ in $\mathbb{R}^N$

$$\text{cap}\,(A, B) = \inf_V \left\{ \int_{\mathbb{R}^N} |\nabla v|^2\, dy,\ \forall v \in H_0^1(B),\ v \equiv 1 \text{ on } A \text{ in the } H^1\text{-sense} \right\}. \tag{3.4}$$

Remark 3.1. If $N \geqq 3$, then

$$\text{cap}_\varepsilon \leqq \text{cap}\,(B_1, B_2) < \infty. \tag{3.5}$$

Moreover, if there exists $\beta > 0$ such that for all ε, T_ε contains a disk (of codimension one) D_β of radius β, then

$$\text{cap}_\varepsilon \geqq \text{cap}\,(D_\beta, \mathbb{R}^N) > 0, \tag{3.6}$$

hence cap_ε is bounded below away from 0.

If $N = 2$, we have

$$\text{cap}_\varepsilon \leqq \text{cap}\,(B_1, B_{\varepsilon/r_\varepsilon}) = \frac{2\pi}{\ln\,(\varepsilon/r_\varepsilon)}, \tag{3.7}$$

so that $\text{cap}_\varepsilon \to 0$ as $\varepsilon \to 0$. However, if there exists $\beta > 0$ such that for all ε, T_ε contains a segment S_β of length β, then (cf. Attouch and Picard [2, Prop. A3]):[1]

$$\text{cap}_\varepsilon \geqq \text{cap}\,(S_\beta, B_{\varepsilon/r_\varepsilon}) \simeq \frac{2\pi}{\ln\,(\varepsilon/\beta r_\varepsilon)} \simeq \frac{2\pi}{\ln\,(\varepsilon/r_\varepsilon)}, \tag{3.8}$$

which proves that, in this case, for $N = 2$, cap_ε is equivalent to $2\pi/\ln\,(\varepsilon/r_\varepsilon)$.

We denote by ψ_ε the capacitory potential of T_ε in $B_{\varepsilon/r_\varepsilon}$, i.e., the function achieving the minimum in (3.4) for $A = T_\varepsilon$ and $B = B_{\varepsilon/r_\varepsilon}$. Then ψ_ε satisfies

$$\begin{aligned} &-\Delta\psi_\varepsilon = 0 \quad \text{in } B_{\varepsilon/r_\varepsilon} \setminus \overline{T_\varepsilon}, \\ &\psi_\varepsilon = 0 \quad \text{on } B_{\varepsilon/r_\varepsilon}, \\ &\psi_\varepsilon = 1 \quad \text{on } \partial T_\varepsilon, \\ &\psi_\varepsilon \in H^1(B_{\varepsilon/r_\varepsilon}). \end{aligned} \tag{3.9}$$

The following lemma holds.

LEMMA 3.2. *There exists a unique solution W_ε for problem* (3.1) *and*

$$W_\varepsilon = \lambda_\varepsilon \psi_\varepsilon \tag{3.10}$$

with $\lambda_\varepsilon \in [0, 1]$ given by

$$\lambda_\varepsilon = \frac{A_\varepsilon}{A_\varepsilon + \text{cap}_\varepsilon} \tag{3.11}$$

where A_ε and cap_ε are defined by (3.2) *and* (3.3).

Proof. By uniqueness of the solution (3.1), there exists λ_ε such that (3.10) holds. To identify λ_ε, observe that

$$\text{cap}_\varepsilon = \int_{B_{\varepsilon/r_\varepsilon}} |\nabla \psi_\varepsilon|^2\, dy = \int_{\partial T_\varepsilon} \frac{\partial \psi_\varepsilon}{\partial n}\, d\sigma. \tag{3.12}$$

[1] The notation $\simeq$ indicates infinitesimal equivalence as $\varepsilon \to 0$.

Using (3.10) in (3.1), we get

$$\lambda_\varepsilon \operatorname{cap}_\varepsilon + r_\varepsilon a_\varepsilon(\lambda_\varepsilon - 1)|\partial T_\varepsilon| = 0 \tag{3.13}$$

where $|\partial T_\varepsilon|$ denotes the $(N-1)$-Lebesgue measure of ∂T_ε. By (3.13), we deduce (3.11) and, consequently, the uniqueness of W_ε. □

For $x \in B_\varepsilon$ set

$$w_\varepsilon(x) = W_\varepsilon\left(\frac{x}{r_\varepsilon}\right), \tag{3.14}$$

and extend w_ε by periodicity of period εZ^N to obtain an element of $H^1_{\text{loc}}(\mathbb{R}^N)$ still denoted as w_ε, which satisfies

$$\begin{aligned} &-\Delta w_\varepsilon = 0 \quad \text{in } \bigcup_{k\in Z^N} \{B_\varepsilon^k \setminus \overline{T_\varepsilon^k}\}, \\ &w_\varepsilon = 0 \quad \text{in } \mathbb{R}^N \Big\backslash \bigcup_{k\in Z^N} B_\varepsilon^k, \\ &\int_{\partial T_\varepsilon^k} \left(\frac{\partial w_\varepsilon}{\partial n} + a_\varepsilon(w_\varepsilon - 1)\right) d\sigma = 0 \quad \forall k \in Z^N, \\ &w_\varepsilon \text{ constant on each } \partial T_\varepsilon^k,\ k \in Z^N. \end{aligned} \tag{3.15}$$

Remark 3.3. Since $0 \leqq \psi_\varepsilon \leqq 1$, Lemma 3.2 with (3.14) implies that

$$0 \leqq W_\varepsilon \leqq \lambda_\varepsilon \leqq 1 \quad \text{a.e. in } B_{\varepsilon/r_\varepsilon},$$

$$0 \leqq w_\varepsilon \leqq \lambda_\varepsilon \leqq 1 \quad \text{a.e. in } \Omega.$$

PROPOSITION 3.4. *Assume that*

$$\frac{r_\varepsilon^{N-2}}{\varepsilon^N} \operatorname{cap}_\varepsilon \lambda_\varepsilon^2 \quad \textit{is bounded as } \varepsilon \to 0; \tag{3.16}$$

then

$$w_\varepsilon \rightharpoonup 0 \quad \textit{weakly in } H^1_{\text{loc}}(\mathbb{R}^N). \tag{3.17}$$

Furthermore, the following equivalence holds:

$$w_\varepsilon \to 0 \quad \textit{strongly in } H^1_{\text{loc}}(\mathbb{R}^N) \Leftrightarrow \frac{r_\varepsilon^{N-2}}{\varepsilon^N} \operatorname{cap}_\varepsilon \lambda_\varepsilon^2 \to 0 \quad \textit{as } \varepsilon \to 0. \tag{3.18}$$

Proof. By Remark 3.3 for any smooth bounded open set $\mathcal{O}$ in $\mathbb{R}^N$ we have

$$\|w_\varepsilon\|_{L^2(\mathcal{O})} \leqq C(\mathcal{O}) \tag{3.19}$$

where $C(\mathcal{O})$ is independent of ε.

On the other hand, Lemma 3.2 together with (3.5) yields

$$\begin{aligned} \int_{\mathcal{O}} |\nabla w_\varepsilon|^2\, dx &\simeq \frac{|\mathcal{O}|}{\varepsilon^N} \int_{B_\varepsilon} |\nabla w_\varepsilon|^2\, dx \\ &\simeq \frac{|\mathcal{O}|}{\varepsilon^N} \int_{B_{\varepsilon/r_\varepsilon}} \frac{1}{r_\varepsilon^2} |\nabla_y W_\varepsilon(y)|^2 r_\varepsilon^N\, dy \\ &\simeq |\mathcal{O}| \frac{r_\varepsilon^{N-2}}{\varepsilon^N} \operatorname{cap}_\varepsilon \lambda_\varepsilon^2. \end{aligned} \tag{3.20}$$

Hence, if (3.16) is satisfied, w_ε is bounded hence weakly relatively compact in $H^1(\mathcal{O})$. Observe that

$$w_\varepsilon \chi_{(\mathcal{O}\setminus \cup_{k\in Z^N} \overline{B_\varepsilon^k})} = 0 \quad \text{a.e. in } \mathcal{O}.$$

Since $\chi_{(\mathcal{O}\setminus \cup_{k\in Z^N} \overline{B_\varepsilon^k})}$ converges weakly star in $L^\infty(\mathcal{O})$ to some strictly positive constant, the whole sequence w_ε converges weakly to zero in $H^1(\mathcal{O})$. Finally, (3.18) is a direct consequence of (3.20). □

We now introduce some measures supported on the boundaries of the $\mathcal{B}_\varepsilon$ and $\mathcal{T}_\varepsilon$. Set

$$\mu_\varepsilon^* = -\left.\frac{\partial w_\varepsilon}{\partial n}\right|_{\partial \mathcal{B}_\varepsilon} \delta_{\partial \mathcal{B}_\varepsilon}, \tag{3.21}$$

$$\mu_\varepsilon = -\left.\frac{\partial w_\varepsilon}{\partial n}\right|_{\partial \mathcal{T}_\varepsilon} \delta_{\partial \mathcal{T}_\varepsilon} \tag{3.22}$$

where n denotes the outward unit normal to $\partial \mathcal{B}_\varepsilon$ and $\partial \mathcal{T}_\varepsilon$ respectively, and where δ denotes the $(N-1)$-dimensional surface measure. We now want to investigate the strong compactness of μ_ε^* in $H^{-1}(\Omega)$.

PROPOSITION 3.5. *If there exists a constant c independent of ε such that*

$$\begin{aligned} &\frac{\lambda_\varepsilon}{\varepsilon^2 \ln(\varepsilon/r_\varepsilon)} \leqq c \quad \text{if } N=2, \\ &\frac{r_\varepsilon^{N-2}}{\varepsilon^N}\lambda_\varepsilon \leqq c \qquad \text{if } N \geqq 3, \end{aligned} \tag{3.23}$$

then $\{\mu_\varepsilon^*\}$ is a compact set of $H^{-1}(\Omega)$.

If further there exists a nonnegative constant α such that

$$\frac{r_\varepsilon^{N-2}}{\varepsilon^N} \operatorname{cap}_\varepsilon \lambda_\varepsilon \to \alpha \quad \text{as } \varepsilon \to 0, \tag{3.24}$$

then

$$\mu_\varepsilon^* \to \alpha\, dx \quad \text{strongly in } H^{-1}(\Omega). \tag{3.25}$$

Remark 3.6. From (3.5) and (3.7) of Remark 3.1 it follows that (3.23) implies the boundedness of $(r_\varepsilon^{N-2}/\varepsilon^N)\operatorname{cap}_\varepsilon \lambda_\varepsilon$. The converse is true for $N \geqq 3$ (respectively, $N=2$) if and only if $\operatorname{cap}_\varepsilon$ (respectively $\operatorname{cap}_\varepsilon \ln(\varepsilon/r_\varepsilon)$) is bounded below away from zero. Thus (3.6) and (3.8) give sufficient conditions for this to occur.

Remark 3.7. If (3.24) holds, using (3.2) and (3.11) we obtain the following explicit formula:

$$\alpha = \lim_{\varepsilon\to 0} \frac{r_\varepsilon^{N-1}}{\varepsilon^N} \frac{a_\varepsilon |\partial T_\varepsilon| \operatorname{cap}_\varepsilon}{a_\varepsilon r_\varepsilon (\partial T_\varepsilon| + \operatorname{cap}_\varepsilon}, \tag{3.26}$$

which shows that all the parameters of the problem appear in the definition of α. In § 6 we will show examples where (3.23) or (3.24) hold. □

Proof of Proposition 3.5. We make use of test functions and a comparison lemma introduced by Cioranescu and Murat in [7].

Let ϖ_ε satisfy

$$\begin{aligned} &-\Delta \varpi_\varepsilon = 0 \quad \text{in } \mathcal{B}_\varepsilon \setminus \overline{\mathcal{B}_{r_\varepsilon}}, \\ &\varpi_\varepsilon = 0 \quad \text{in } \Omega \setminus \overline{\mathcal{B}_\varepsilon}, \\ &\varpi_\varepsilon = 1 \quad \text{in } \mathcal{B}_{r_\varepsilon}, \\ &\varpi_\varepsilon \in H^1(\Omega). \end{aligned} \tag{3.27}$$

By the maximum principle $w_\varepsilon \leqq \lambda_\varepsilon \varpi_\varepsilon$, so

$$0 \leqq -\frac{\partial w_\varepsilon}{\partial n}\,\delta_{\partial \mathscr{B}_\varepsilon} \leqq -\lambda_\varepsilon \frac{\partial \varpi_\varepsilon}{\partial n}\,\delta_{\partial \mathscr{B}_\varepsilon}.$$

By Lemma 2.8 of [7], the relative compactness of μ_ε^* is implied by that of $-\lambda_\varepsilon(\partial \varpi_\varepsilon/\partial n)\delta_{\partial \mathscr{B}_\varepsilon}$. On the other hand, Lemma 2.3 of [7] shows that

$$\lambda_\varepsilon \frac{\partial \varpi_\varepsilon}{\partial n} \simeq \begin{cases} \lambda_\varepsilon \dfrac{r_\varepsilon^{N-2}}{\varepsilon^{N-1}} & \text{if } N \geqq 3 \\ \dfrac{\lambda_\varepsilon}{\varepsilon \ln(\varepsilon/r_\varepsilon)} & \text{if } N = 2, \end{cases}$$

and that $\varepsilon \delta_{\partial \mathscr{B}_\varepsilon}$ is relatively compact in $H^{-1}(\Omega)$. Hence (3.23) implies the relative compactness of μ_ε^*. To show (3.25) let μ^* be a limit point in $H^{-1}(\Omega)$ of $\{\mu_\varepsilon\}$ as $\varepsilon \to 0$. From the ε-periodicity of μ_ε^*, it follows that μ^* is itself invariant under translations, which implies that it is a multiple of the Lebesgue measure.

From (3.2), (3.11), and (3.15), we have

$$\begin{aligned} \fint_\Omega d\mu_\varepsilon^* &= \frac{1}{|\Omega|}\int_{\partial \mathscr{B}_\varepsilon} -\frac{\partial w_\varepsilon}{\partial n}\,d\sigma \simeq \frac{1}{|\Omega|\varepsilon^N}\int_{\partial B_\varepsilon} -\frac{\partial w_\varepsilon}{\partial n}\,d\sigma \\ &= \frac{1}{|\Omega|\varepsilon^N}\int_{\partial T_\varepsilon} -\frac{\partial w_\varepsilon}{\partial n}\,d\sigma = \frac{1}{|\Omega|\varepsilon^N}\int_{\partial T_\varepsilon} a_\varepsilon(1-w_\varepsilon)\,d\sigma \\ &= \frac{r_\varepsilon^{N-1}}{|\Omega|\varepsilon^N} a_\varepsilon(1-\lambda_\varepsilon)|\partial T_\varepsilon| = \frac{r_\varepsilon^{N-2}}{|\Omega|\varepsilon^N}\lambda_\varepsilon \operatorname{cap}_\varepsilon, \end{aligned}$$

which implies that $\mu^* = \alpha\, dx_{|\Omega}$.

Remark 3.8. The above proof shows that if $\operatorname{cap}_\varepsilon$ (respectively, $\operatorname{cap}_\varepsilon \ln(\varepsilon/r_\varepsilon)$) is bounded below away from zero for $N \geqq 3$ (respectively, $N = 2$), then the boundedness of μ_ε^* in the space of measures on Ω implies its relative compactness in $H^{-1}(\Omega)$.

COROLLARY 3.9. *Under* (3.23), *a necessary and sufficient condition for* μ_ε *to converge strongly in* $H^{-1}(\Omega)$ *is that* $(r_\varepsilon^{N-2}/\varepsilon^N)\operatorname{cap}_\varepsilon \lambda_\varepsilon^2 \to 0$ *as* $\varepsilon \to 0$.

Proof. It is enough to note that $\mu_\varepsilon^* - \mu_\varepsilon = \Delta w_\varepsilon$ in $H^{-1}(\Omega)$ and to apply (3.18) with Proposition 3.5. □

4. The asymptotic behavior of problem (1.2). Set

$$X_\varepsilon = \frac{\operatorname{cap}_\varepsilon r_\varepsilon^{N-2}}{\varepsilon^N}, \qquad Y_\varepsilon = \frac{a_\varepsilon r_\varepsilon^{N-1}|\partial T_\varepsilon|}{\varepsilon^N}, \tag{4.1}$$

$$X'_\varepsilon = \begin{cases} \dfrac{r_\varepsilon^{N-2}}{\varepsilon^N} & \text{for } N \geqq 3, \\ \dfrac{2\pi}{\varepsilon^2 \ln(\varepsilon/r_\varepsilon)} & \text{for } N = 2, \end{cases} \tag{4.2}$$

and

$$Y'_\varepsilon = \begin{cases} \dfrac{a_\varepsilon r_\varepsilon^{N-1}|\partial T_\varepsilon|}{\varepsilon^N \operatorname{cap}_\varepsilon} & \text{for } N \geqq 3, \\ \dfrac{2\pi a_\varepsilon r_\varepsilon |\partial T_\varepsilon|}{\varepsilon^2 \ln(\varepsilon/r_\varepsilon)\operatorname{cap}_\varepsilon} & \text{for } N = 2. \end{cases} \tag{4.3}$$

We note that conditions (3.23) and (3.24) are equivalent to $X'_\varepsilon Y'_\varepsilon/(X'_\varepsilon + Y'_\varepsilon)$ bounded, and $\lim_{\varepsilon\to 0} X_\varepsilon Y_\varepsilon/(X_\varepsilon + Y_\varepsilon) = \alpha$, respectively. Our main result can now be stated as follows.

THEOREM 4.1. *Suppose that* (3.23) *and* (3.24) *hold, namely,* $X'_\varepsilon Y'_\varepsilon/(X'_\varepsilon + Y'_\varepsilon)$ *is bounded, and* $\lim_{\varepsilon\to 0} X_\varepsilon Y_\varepsilon/(X_\varepsilon + Y_\varepsilon) = \alpha$, *as* $\varepsilon \to 0$. *Then the solutions* u_ε *of problem* (1.2) *converge weakly in* V *to the unique solution* u *of the following problem*:

$$\begin{aligned} -\Delta u + \alpha u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{4.4}$$

Proof. By the uniqueness of the solution of (4.4) and referring to (1.5), it is enough to show that any weak limit point of $\{u_\varepsilon\}$ is a solution of (4.4).

Let φ be in $\mathcal{D}(\Omega)$ and let $\varphi_\varepsilon \in V_\varepsilon^p$ be the sequence given by Lemma 2.1 such that

$$\forall p \in [1, +\infty[, \quad \varphi_\varepsilon \to \varphi \text{ strongly in } V^p. \tag{4.5}$$

Choosing $\varphi_\varepsilon w_\varepsilon$ as test function in (1.3) yields

$$\int_\Omega w_\varepsilon \nabla u_\varepsilon \nabla \varphi_\varepsilon \, dx + \int_\Omega \varphi_\varepsilon \nabla u_\varepsilon \nabla w_\varepsilon \, dx + \int_{\partial\mathcal{T}_\varepsilon} a_\varepsilon u_\varepsilon \varphi_\varepsilon w_\varepsilon \, d\sigma = \int_\Omega f \varphi_\varepsilon w_\varepsilon \chi_{\Omega_\varepsilon} \, dx. \tag{4.6}$$

Also, multiplying (3.15) by $\varphi_\varepsilon u_\varepsilon$ and integrating by parts gives

$$\begin{aligned} &\int_{\mathcal{B}_\varepsilon \setminus \mathcal{T}_\varepsilon} u_\varepsilon \nabla w_\varepsilon \nabla \varphi_\varepsilon \, dx + \int_{\mathcal{B}_\varepsilon \setminus \mathcal{T}_\varepsilon} \varphi_\varepsilon \nabla w_\varepsilon \nabla u_\varepsilon \, dx \\ &\quad = \int_{\partial\mathcal{B}_\varepsilon} \frac{\partial w_\varepsilon}{\partial n} u_\varepsilon \varphi_\varepsilon \, d\sigma + \int_{\partial\mathcal{T}_\varepsilon} a_\varepsilon (1 - w_\varepsilon) u_\varepsilon \varphi_\varepsilon \, d\sigma. \end{aligned} \tag{4.7}$$

Combining (4.6) and (4.7), and since $w_\varepsilon = 0$ in $\Omega \setminus \mathcal{B}_\varepsilon$, we obtain

$$\begin{aligned} &\int_\Omega w_\varepsilon \nabla u_\varepsilon \nabla \varphi_\varepsilon \, dx - \int_\Omega u_\varepsilon \nabla \varphi_\varepsilon \nabla w_\varepsilon \, dx + \int_{\partial\mathcal{B}_\varepsilon} \frac{\partial w_\varepsilon}{\partial n} u_\varepsilon \varphi_\varepsilon \, d\sigma + \int_{\partial\mathcal{T}_\varepsilon} u_\varepsilon \varphi_\varepsilon a_\varepsilon \, d\sigma \\ &\quad = \int_\Omega \varphi_\varepsilon w_\varepsilon \chi_{\Omega_\varepsilon} \, dx. \end{aligned} \tag{4.8}$$

On the other hand, we can apply Proposition 3.4, which states that w_ε converges weakly to zero in $H^1_{\mathrm{loc}}(\mathbb{R}^N)$ because (3.24) is satisfied and $\lambda_\varepsilon \leqq 1$. In order to go to the limit as $\varepsilon \to 0$ in the first two terms of (4.8), we choose p such that $1/p + 1/q + \frac{1}{2} = 1$, for some $q < 2N/(N-2)$ (respectively, $q < \infty$ for $N = 2$), namely $p > N$, and use the fact that u_ε and w_ε converge strongly in $L^q(\Omega)$, the latter going to zero.

Consequently, because the right-hand side of (4.8) converges to zero, we obtain the following convergence for the remaining terms:

$$\lim_{\varepsilon\to 0} \left| \int_{\partial\mathcal{T}_\varepsilon} a_\varepsilon u_\varepsilon \varphi_\varepsilon \, d\sigma - \langle \mu_\varepsilon^*, u_\varepsilon \varphi_\varepsilon \rangle \right| = 0. \tag{4.9}$$

On the other hand, by Lemma 2.1, φ_ε converges to φ in $C^0(\Omega)$ and $\nabla\varphi_\varepsilon$ converges strongly to $\nabla\varphi$ in $L^p(\Omega)$ so that $u_\varepsilon \varphi_\varepsilon$ converges to $u\varphi$ weakly in V. By Proposition 3.5, μ^* converges strongly to $\alpha\, dx$ in $H^{-1}(\Omega)$, so

$$\lim_{\varepsilon\to 0} \int_{\partial\mathcal{T}_\varepsilon} a_\varepsilon u_\varepsilon \varphi_\varepsilon \, d\sigma = \alpha \int_\Omega u\varphi \, dx. \tag{4.10}$$

Finally, using φ_ε as a test function in (1.3) yields

$$\int_{\Omega_\varepsilon} \nabla u_\varepsilon \nabla \varphi_\varepsilon \, dx = -\int_{\partial \mathcal{T}_\varepsilon} a_\varepsilon u_\varepsilon \varphi_\varepsilon \, d\sigma + \int_{\Omega_\varepsilon} f\varphi_\varepsilon \, dx \quad \forall \varphi \in V_\varepsilon,$$

which, together with (4.10), gives as $\varepsilon \to 0$

$$\int_\Omega \nabla u \nabla \varphi \, dx + \alpha \int_\Omega u\varphi \, dx = \int_\Omega f\varphi \, dx$$

for all $\varphi \in \mathcal{D}(\Omega)$. This is the variational formulation of (4.4), and the proof is complete. □

Remark 4.2. In Theorem 4.1, we assume that the holes are modeled after an open set T_ε. However, it is easy to see that all the definitions and statements of the previous sections remain valid in the case of cracks, i.e., when T_ε is a piecewise C^1 set of codimension 1 (possibly with branchpoints), provided the following changes are performed:

— $\partial T_\varepsilon = T_\varepsilon$, $\partial \mathcal{T}_\varepsilon = \mathcal{T}_\epsilon$,

— The variational formulation (1.3) is preserved and in (1.2) the boundary condition is replaced by

$$\int_{T_\varepsilon^k} \left(-\left[\frac{\partial u_\varepsilon}{\partial n} \right]_n + a_\varepsilon u_\varepsilon \right) d\sigma = 0$$

where $[\partial u_\varepsilon / \partial n]_n$ denotes the jump of $\nabla u_\varepsilon \cdot n$ across the crack along the unit normal n (which is independent of the orientation chosen for n). An example of this situation is given in § 6.

5. Summary of the results. To present the results of the previous paragraph in a synthetic form we will combine them into Table 1, in which we use the following

TABLE 1

	$Y'=0$	$0<Y'<+\infty$	$Y'=+\infty$
$X'=0$	$X=Y=0$ $\alpha=0$	$X=0$ $\alpha=0$	$X=0$ $\alpha=0$
$0<X'<+\infty$	$Y=0$ $\alpha=0$	$\alpha=XY/(X+Y)$	$X<+\infty$ $\alpha=XY/(X+Y)$ Note 1
$X'=+\infty$	$Y=0$ $\alpha=0$	$Y<+\infty$ $\alpha=XY/(X+Y)$	Note 2 If $X=Y=+\infty$ then $\alpha=+\infty$

Note 1. If $\gamma>0$, then $X'=+\infty \Rightarrow X=+\infty$ and $Y'=+\infty \Rightarrow Y=+\infty$ so $\alpha=Y$ or $\alpha=X$, respectively.

Note 2. If $\gamma>0$, then $\alpha=+\infty$ can be obtained by a comparison method which is standard in the theory of Γ-convergence (see Di Giorgi [10], [11], Attouch [1]). Here $\alpha=+\infty$ means that the limit u is identically zero. However, we should note that the case of $\gamma=0$ can occur from two distinct possibilities: either a misjudgment of the values of r_ε (which may have been chosen too big, whereas $\gamma_\varepsilon \to 0$ leading to an indetermination), or a true difficulty for the case where r_ε is adjusted in such a way as to keep T_ε exactly inside a ball of unit radius and inside no smaller ball. In the latter case, the problem remains open.

notation to denote the "normalized" capacity of T_ε:

$$\gamma_\varepsilon = \begin{cases} \operatorname{cap}_\varepsilon & \text{for } N \geqq 3 \\ \dfrac{\operatorname{cap}_\varepsilon}{(2\pi/\ln(\varepsilon/r_\varepsilon))} & \text{for } N=2. \end{cases} \tag{5.1}$$

Note that from (3.5) and (3.7), γ_ε is bounded with respect to ε.

We assume (possibly taking subsequences) that the following convergences occur in $[0, +\infty]$:

$$\lim_{\varepsilon\to 0} X_\varepsilon = X, \qquad \lim_{\varepsilon\to 0} Y_\varepsilon = Y, \tag{5.2}$$

$$\lim_{\varepsilon\to 0} X'_\varepsilon = X', \qquad \lim_{\varepsilon\to 0} Y'_\varepsilon = Y', \tag{5.3}$$

$$\lim_{\varepsilon\to 0} \gamma_\varepsilon = \gamma < +\infty. \tag{5.4}$$

Clearly from (4.1)-(4.3) and provided the following operations make sense in $[0, +\infty]$ we have

$$X = X'\gamma \quad \text{and} \quad Y = Y'\gamma.$$

In this setting, conditions (3.23) and (3.24), which appear as the hypotheses of Theorem 4.1, are respectively equivalent to $\min(X', Y') < +\infty$ and $\alpha = (XY/(X+Y))$. The results of Theorem 4.1 are summarized in Table 1, where the main entries are X' and Y' and the conclusions bear on X, Y, and α. We only remark that the results of Theorem 4.1 can be applied in a more general setting without assuming the existence of the individual limits X, X', Y, and Y'.

6. Some fractal examples. In this section we give some examples involving some fractal sets, namely, the ones which can be defined as limits of recursive sequences of "normal" sets with a specific geometrical operation involved in the inductive step. Therefore, the limit set is self-similar. We refer to Falconer [13, Chap. 8] for a detailed presentation of such sets and the determination of their Hausdorff dimension.

In these constructions, it is more natural to use as a main parameter the induction index, an integer n which goes to ∞, and to consider three sequences $\{\varepsilon_n\}$, $\{a_n\}$, and $\{r_n\}$, in place of ε, a_ε, and r_ε, respectively. The normalized shape is then denoted T_n rather than T_ε.

In each construction, there is a number $\kappa > 0$ which determines the boundary measure of ∂T_n as a function of n:

$$|\partial T_{n+1}| = \kappa |\partial T_n|;$$

this number is intimately connected to the Hausdorff dimension of the boundaries of the limits of the T_n's.

Following the notation of the previous section, we set γ_n, X_n, Y_n, X'_n, and Y'_n in place of γ_ε, X_ε, etc. For simplicity we consider examples for which γ_n converges to a nonzero γ. Then the values of X and Y are obtained as the following limits:

$$X = \gamma \lim_{n\to\infty} \frac{r_n^{N-2}}{\varepsilon_n^N} \quad (N>2), \qquad X = \gamma \lim_{n\to\infty} \frac{2\pi}{\varepsilon_n^2 \ln(\varepsilon_n/r_n)} \quad (N=2),$$

$$Y = \lim_{n\to\infty} \frac{a_n r_n^{N-1} \kappa^n |\partial T_0|}{\varepsilon_n^N}.$$

In these formulas, both the capacity and the coefficient κ appear explicitly to give the limit coefficient α.

For a first example we take the twisted curve in the plane for which $\kappa = 2$ and whose boundary is of Hausdorff dimension 1.5. It is clear that γ is the normalized capacity of the limit of the sequence of sets T_n and is larger than zero. The first four iterations are given in Fig. 2.

Our second example concerns a set T_n in the plane which becomes more and more disconnected (see Fig. 3). It is not very hard to check that in this case also γ is not zero for any scaling factor q, $q \in (0, \frac{1}{2})$. Then, $\kappa = 4q$ and the Hausdorff dimension is $\ln 4/\ln(1/q)$. In Fig. 3, $q = \frac{1}{3}$ so that $\kappa = \frac{4}{3}$ and the Hausdorff dimension is $\ln 4/\ln 3$.

The third example is based on the well-known Koch curve in the plane, for which $\kappa = \frac{4}{3}$ and the Hausdorff dimension is $\ln 4/\ln 3$ (see Fig. 4).

The fourth example (Fig. 5) corresponds to the situation of a fractal cracks, with a similarity factor q and a given number of branches ν. Then $\kappa = 1 + (\nu - 1)q$ and the Hausdorff dimension of the tips is $\ln \nu / -\ln q$.

Each of these examples can be generalized to higher dimensions. In Fig. 6, we show the extension of the Koch curve to three dimensions, for which $\kappa = \frac{3}{2}$ and the Hausdorff dimension is $\ln 6/\ln 2$.

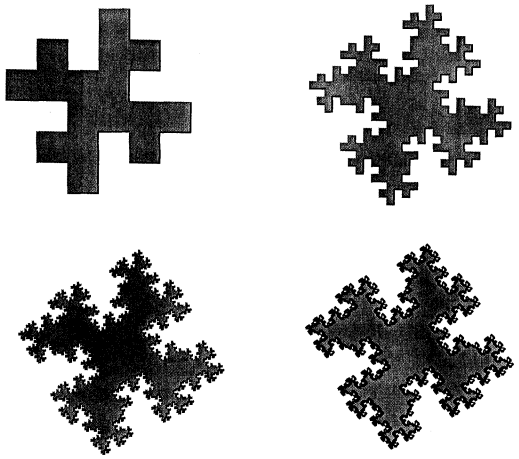


FIG. 2. *Twisted curve. Hausdorff dimension*: 1.5, $\kappa = 2$.

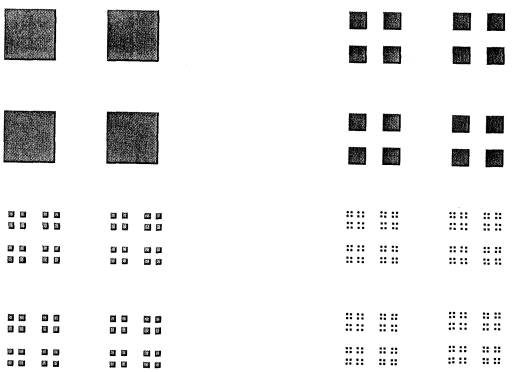


FIG. 3. *Hausdorff dimension*: $\ln 4/\ln 3$, $\kappa = \frac{4}{3}$.

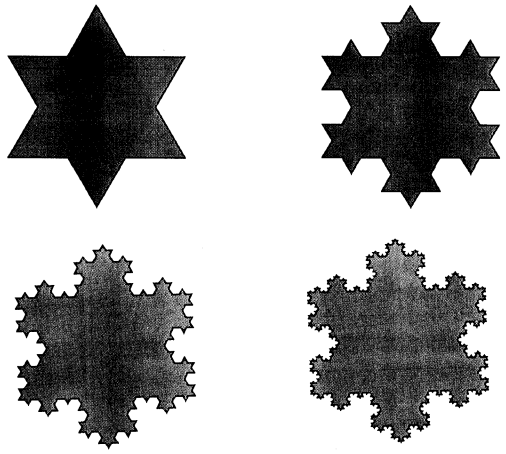


FIG. 4. *The Koch snowflake. Hausdorff dimension*: $\ln 4/\ln 3$, $\kappa = \frac{4}{3}$.

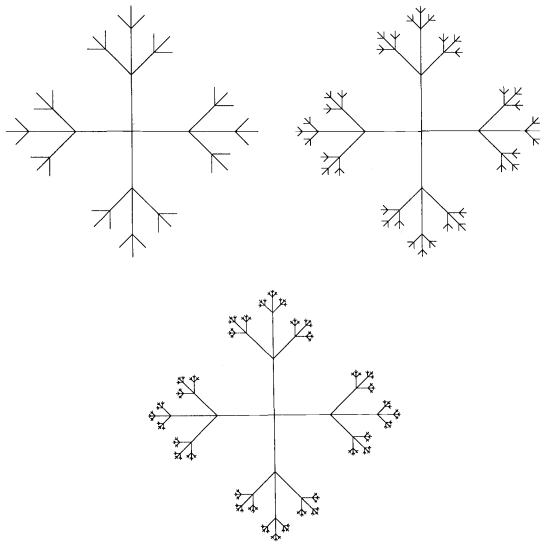


FIG. 5. *A fractal crack. Hausdorff dimension*: $\simeq \ln 3/\ln 2.5$, $\kappa \simeq 1.8$.

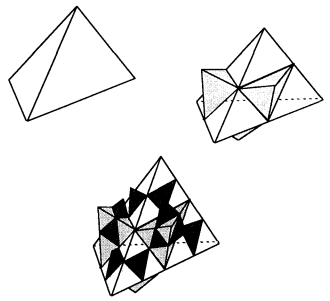


FIG. 6. *A three-dimensional Koch snowflake surface. Hausdorff dimension*: $\ln 6/\ln 2$, $\kappa = \frac{3}{2}$.

REFERENCES

[1] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Applicable Mathematics Series, Pitman, London, 1984.

[2] H. ATTOUCH AND C. PICARD, *Variational inequalities with varying obstacles, the general form of the limit problem*, J. Funct. Anal., 50 (1983), pp. 329-396.

[3] A. BENSOUSSAN, J.-L. LIONS AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1984.

[4] A. BRILLARD, *Asymptotic analysis of two elliptic equations with oscillating terms*, Modél. Math. Anal. Numér., 22 (1988), pp. 187-216.

[5] L. CARBONE AND F. COLOMBINI, *On convergence of functionals with unilateral constraints*, J. Math. Pures Appl., 59 (1980), pp. 465-500.

[6] D. CIORANESCU AND P. DONATO, *Homogénéisation du problème de Neumann non homogène dans des ouverts perforés*, Asymptotic Anal., 1 (1988), pp. 115-138.

[7] D. CIORANESCU AND F. MURAT, *Un terme étrange venu d'ailleurs*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, vol. II, 60, pp. 98-138, Vol. III, 70, pp. 154-178, Res. Notes in Math., Pitman, London, 1981.

[8] C. CONCA, *On the application of the homogenization theory to a class of problems arising in fluid mechanics*, J. Math. Pures Appl., 64 (1985), pp. 31-75.

[9] G. DAL MASO AND P. LONGO, *Γ-limits of obstacles*, Ann. Mat. Pura Appl., 128 (1980), pp. 1-50.

[10] E. DE GIORGI, *Convergence problems for functionals and operators*, in Proc. Internat. Meeting on Recent Methods in Non Linear Analysis, Rome, May 1978, E. De Giorgi, E. Magenes, and U. Mosco, eds., Pitagora Editrice, Bologna, 1979, pp. 131-188.

[11] ———, *G-operators and Γ-convergence*, in Proc. Internat. Congress of Mathematicians, August 1983, Warsaw, P.W.N. Polish Scientific Publishers, Warsaw, and North-Holland, Amsterdam, 1984, pp. 1175-1191.

[12] E. DE GIORGI AND S. SPAGNOLO, *Sulla convergenza degli integrali dell'energia*, Boll. Un. Mat. Ital., 8 (1973), pp. 391-411.

[13] K. J. FALCONER, *The Geometry of the Fractal Sets*, Cambridge Tracts in Mathematics, 85, Cambridge University Press, Cambridge, 1985.

[14] S. KAÏZU, *The Poisson equation with semilinear boundary conditions in domains with many tiny holes*, J. Fac. Sci. Univ. Tokyo Sect. IA, 36 (1989), pp. 43-86.

[15] ———, *Behavior of solutions of the Poisson equation under fragmentation of the boundary of the domain*, Japan J. Appl. Math., 17 (1990).

[16] C. KURATOWSKI, *Topology*, Academic Press, New York, 1966.

[17] V. A. MARCHENKO AND E. J. HRUSHLOV, *Boundary Value Problems in Domains with Finely Granulated Boundary*, Naukova Dumka, Kiev, 1974. (In Russian.)

[18] J. RAUCH AND M. TAYLOR, *Potential and scattering theory on wildly perturbed domains*, J. Funct. Anal., 18 (1975), pp. 27-59.

[19] E. SANCHEZ-PALENCIA, *Nonhomogeneous Media and Vibration Theory*, Lecture Notes in Physics, 127, Springer-Verlag, Berlin, 1980.

[20] L. TARTAR, *Quelques remarques sur l'homogénéisation*, in Functional Analysis and Numerical Analysis, Proc. of the Japan-France Seminar 1976, H. Fujita, ed., Japan Society for the Promotion of Science (1978), pp. 469-482.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 653-679, May 1991

006

THE BIFURCATIONS OF COUNTABLE CONNECTIONS FROM A TWISTED HETEROCLINIC LOOP*

BO DENG†

Abstract. Codimension-two bifurcation phenomena associated with nondegenerate heteroclinic loops are studied. The bifurcation curves of homoclinic orbits in the parameter space are characterized by the twist structure of the heteroclinic loops at the bifurcation points. Among other things, it is shown that heteroclinic orbits with any given winding number around a doubly twisted heteroclinic loop must bifurcate. Applications of these bifurcation phenomena are also discussed.

Key words. twisted heteroclinic orbit, homoclinic orbit, periodic orbit, k-heteroclinic orbit, Sil'nikov's variables, exponential expansions, strong λ-lemmas, entrance sets, exit sets, bifurcation equations

AMS(MOS) subject classifications. 34A34, 34C28, 34C99

1. Introduction. A heteroclinic loop takes place for a vector field F when there exist two heteroclinic orbits $z_1^*(t)$ and $z_2^*(t)$, with z_1^* connecting an equilibrium point a_1, to another one a_2, and z_2^* connecting a_2 to a_1. To be precise,

$$z_i^*(t) \to a_i \quad \text{as } t \to -\infty \quad \text{and} \quad z_i^*(t) \to a_j \quad \text{as } t \to +\infty$$

for $i, j = 1, 2$ and $i \neq j$. Figures 1.1 and 1.2 heuristically illustrate what could happen to two structurally different loops when a planar vector field F is perturbed slightly. In Fig. 1.1, either a homoclinic orbit or a periodic orbit would possibly bifurcate from the loop, while in Fig. 1.2 a heteroclinic orbit winding around the original loop for any finite times before reaching its destinations in both backward and forward evolutions would also be possible under perturbation. The very structure distinguishing the second loop from the first one is that a given heteroclinic orbit arises from and tends to the equilibria from different "sides" of the other heteroclinic orbit. The purpose of this paper is to study the bifurcations of a generic two-parameter family of vector fields in $\mathbb{R}^d$, $d \geqq 2$ which exhibit the above heteroclinic phenomena.

The first obvious generalization is to assume that the equilibria a_i of the equation

$$\dot{z} = F(z) \tag{1.1}$$

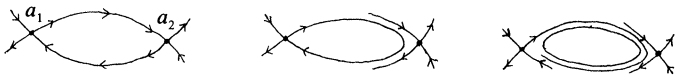


FIG. 1.1

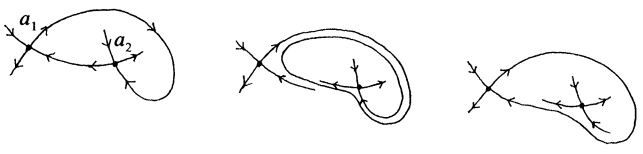


FIG. 1.2

* Received by the editors April 26, 1988; accepted for publication (in revised form) March 23, 1990.
† Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0323.

have the same dimension, $m \geqq 1$, for the stable manifolds W_i^s and the same unstable dimension, $n = d - m \geqq 1$, for the unstable manifolds W_i^u for both $i = 1$ and 2. Moreover, the a_is are simple saddlepoints in the sense that

(1.2) There exist principal eigenvalues $\lambda_i < 0 < \mu_i$ for the linearization $D_z F(a_i)$ and constants $\bar{\lambda}_i < 0 < \bar{\mu}_i$ such that for any other eigenvalue ν of $D_z F(a_i)$ either $\operatorname{Re} \nu < \bar{\lambda}_i < \lambda_i$ or $\operatorname{Re} \nu > \bar{\mu}_i > \mu_i$, for $i = 1$ and 2.

Not as a generalization but as a generic restriction to all the cases, we assume that both equilibria are relatively contractive:

$$\lambda_i + \mu_i < 0 \quad \text{for } i = 1 \text{ and } 2. \tag{1.3}$$

That is, the principal attraction of a_i dominates the principal repelling.

Concerning the structure on the intersection of the unstable manifold W_i^u of a_i and the stable manifold W_j^s of a_j which must be nontransverse along the heteroclinic orbit $\Gamma_i := \{z_i^*(t) : t \in \mathbb{R}\}$, we assume that they are in general position:

$$\operatorname{codim} T_{z_i^*(t)} = 1 \quad \text{for all } t \in \mathbb{R}, \tag{1.4}$$

where

$$T_{z_i^*(t)} := T_{z_i^*(t)} W_i^u + T_{z_i^*(t)} W_j^s$$

and $T_p W$ means the tangent space of a given manifold W at a base point $p \in W$. Also, motivated by the strong λ-lemma from Deng (1989), the following strong inclination property, as another assumption, is also generic:

$$\begin{aligned} \lim_{t \to -\infty} T_{z_i^*(t)} &= T_{a_i} W_i^u + T_{a_i} W_i^{ss}, \\ \lim_{t \to +\infty} T_{z_i^*(t)} &= T_{a_j} W_j^{uu} + T_{a_j} W_j^s. \end{aligned} \tag{1.5}$$

Here, W_i^{ss} and W_i^{uu} are the strong stable and strong unstable manifolds of a_i, respectively. See Fig. 1.3. If the vector field F is C^r, then W^s and W^u are C^r as well (see, e.g., Shub (1987)). But, in general, W^{ss} and W^{uu} are proved to be C^{r-3} instead (see Deng (1989)). Moreover, W^{ss} and W^{uu} are $(m-1)$- and $(n-1)$-dimensional, respectively, characterized by the fact that the limits

$$\begin{aligned} \lim_{t \to -\infty} \frac{z(t) - a}{\|z(t) - a\|} &\neq 0 \quad \text{for } z(0) \in W^u \setminus W^{uu}, \\ \lim_{t \to +\infty} \frac{z(t) - a}{\|z(t) - a\|} &\neq 0 \quad \text{for } z(0) \in W^s \setminus W^{ss} \end{aligned} \tag{1.6}$$

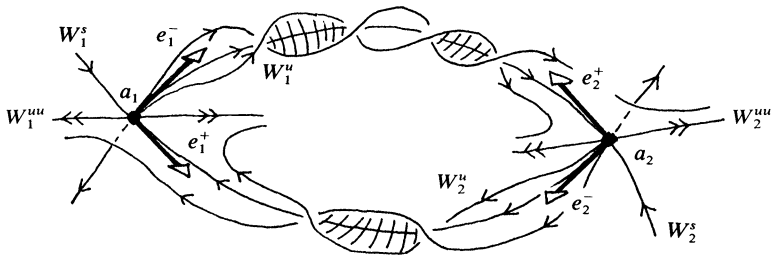
exist and are equal to unit eigenvectors for the principal unstable eigenvalue and principal stable eigenvalue, respectively. The strong inclination property is a generic property provided F is C^r with $r \geqq 7$. See Deng (1989) for the proof.

The last structural assumption reads

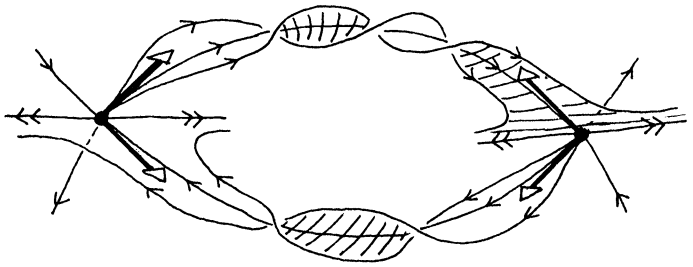
$$\Gamma_i \subset (W_i^u \setminus W_i^{uu}) \cap (W_j^s \setminus W_j^{ss}) \quad \text{for } i, j = 1, 2 \text{ and } i \neq j. \tag{1.7}$$

That is, by virtue of (1.6) this hypothesis says that the heteroclinic orbits arise from and tend to the equilibria along principal eigendirections. It is certainly a generic condition.

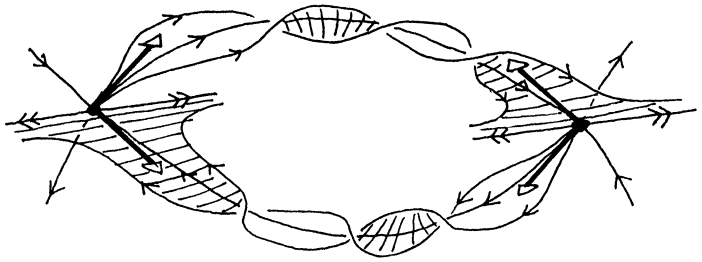
The assumptions (1.4), (1.5), and (1.7) together are referred to as nondegeneracy. They lead to our classifications of heteroclinic orbits into twisted and nontwisted as follows.



(a)



(b)



(c)

FIG. 1.3. (a) *a nontwisted loop,* (b) *a single twisted loop,* (c) *a double twisted loop.*

Let

$$
\begin{aligned}
e_j^+ &= \lim_{t\to+\infty} (z_i^*(t)-a_j)/\|z_i^*(t)-a_j\|, \\
e_i^- &= \lim_{t\to-\infty} (z_i^*(t)-a_i)/\|z_i^*(t)-a_i\|.
\end{aligned} \tag{1.8}
$$

By (1.6) and (1.7), they are unit principal eigenvectors. See Fig. 1.3. Choose $p_i \in \Gamma_i \cap W^u_{i\mathrm{loc}}$ sufficiently close to the equilibrium a_i and $q_j \in \Gamma_i \cap W^s_{j\mathrm{loc}}$ sufficiently close to the other equilibrium a_j. Let $p_i = z_i^*(0)$ and $q_j = z_i^*(T)$ for a large $T \geqq 0$. Because of the strong inclination property (1.5) and the principal asymptotic tangency (1.8), choosing p_i and q_j close enough to a_i and a_j, respectively, implies

$$
\begin{aligned}
&e_i^+ \notin T_{p_i}, \qquad \mathbb{R}^d = T_{p_i} + \mathrm{span}\,(e_i^+), \\
&e_j^- \notin T_{q_j}, \qquad \mathbb{R}^d = T_{q_j} + \mathrm{span}\,(e_j^-).
\end{aligned} \tag{1.9}
$$

Since $T_{z_1^*(t)}$, $0 \leqq t \leqq T$, defines a homotopy from T_{p_i} to T_{q_j}, the following definition is justified (see Fig. 1.3).

DEFINITION 1.1. Let Γ_i be a nondegenerate heteroclinic orbit connecting two simple saddles. Γ_i is said to be twisted if e_i^+ and e_j^- point to opposite sides of T_{p_i} and T_{q_j}, respectively. Otherwise, it is nontwisted.

For the heteroclinic loop, cl $(\Gamma_1 \cup \Gamma_2) = \{a_1, a_2\} \cup \Gamma_1 \cup \Gamma_2$, it is called double twisted if both Γ_1 and Γ_2 are twisted, single twisted if and only if one of them is twisted, and nontwisted if otherwise.

As the last assumption we assume

(1.10) $F: \mathbb{R}^d \times \mathbb{R}^2 \to \mathbb{R}^d$ is a generic C^r vector field with two parameters $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$, having a nondegenerate heteroclinic loop at $\alpha = 0$. Here, the regularity $r \geqq 8$.

By genericity we mean that our results will hold for a residual subset of $C^r(\mathbb{R}^d \times \mathbb{R}^2, \mathbb{R}^d)$ in the weak Whitney topology of C^k-convergence (see, e.g., Hirsch (1976)). To be more precise, we include the following as equivalent conditions for (1.10):

(1.10a) The continuation of Γ_i: Given the fact that the heteroclinic orbit is a codimension 1 object by (1.4), we assume for every $(\alpha_1, 0)$ there exists a heteroclinic orbit $\Gamma_2(\alpha_1) = \{z_2^*(t, \alpha_1): t \in \mathbb{R}\}$ from a_2 to a_1 such that z_2^* is a C^r two-dimensional surface in the phase space. Similarly, $z_1^*(t, \alpha_2)$ forms a C^r surface in $\mathbb{R}^d$ as (t, α_2) takes all values from $\mathbb{R}^2$ and each of the t-curves is a heteroclinic orbit from a_1 to a_2;

(1.10b) The transverse crossing of the stable and unstable manifolds along Γ_i:

$$\lim_{\alpha_1 \to 0} \frac{d_1(\alpha_1, 0)}{|\alpha_1|} \neq 0 \quad \text{and} \quad \lim_{\alpha_2 \to 0} \frac{d_2(0, \alpha_2)}{|\alpha_2|} \neq 0,$$

where $d_1(\alpha_1, \alpha_2)$ denotes the continuously varying distance of $W_1^u(\alpha) \cap \Sigma$ and $W_2^s(\alpha) \cap \Sigma$ with $d_1(0, 0) = 0$ and Σ is an arbitrarily chosen Poincaré cross section to Γ_1. A similar description applies for d_2.

Note that since the Poincaré mapping introduced between any pair of two cross sections is diffeomorphism, the nonzero limiting property in (1.10b) above is independent of the choice of the cross section Σ.

Finally, to state our main theorems we need a few more terms. Let $\mathcal{U}$ be a small tubular neighborhood of the heteroclinic loop cl $(\Gamma_1 \cup \Gamma_2)$. A k-periodic (k-per) orbit is a periodic orbit which is contained in $\mathcal{U}$ and has winding number k in $\mathcal{U}$. Similarly, the closure of a k-homoclinic (k-hom) orbit has winding number k in $\mathcal{U}$. Accordingly, a k-heteroclinic (k-het$_1$) orbit Γ from a_1 to a_2 is such a heteroclinic orbit that cl $(\Gamma \cup \Gamma_2)$ has winding number $k+1$. Similarly, we define a k-heteroclinic (k-het$_2$) orbit from a_2 to a_1. Thus, Γ_1 and Γ_2 themselves are zero-heteroclinic orbits. Note that as long as $\mathcal{U}$ is chosen small enough, the above definition is independent of any particular choice of $\mathcal{U}$. Also, the terminology extends canonically to small perturbations of the vector field $F(\cdot, 0)$.

The first theorem, except for the directions of bifurcation and the k-heteroclinic orbits, is taken from Chow, Deng, and Terman (1990).

THEOREM A. *Suppose F is a generic two-parameter family of vector fields having a nondegenerate heteroclinic loop connecting two relatively contractive and simple saddle equilibria at $\alpha = 0$, i.e.,* (1.2)-(1.5), (1.7), *and* (1.10a, b) *are satisfied. Then there exists a small tubular neighborhood $\mathcal{U}$ of the heteroclinic loop* cl $(\Gamma_1 \cup \Gamma_2)$ *and a neighborhood*

Ω of the bifurcation point $\alpha = 0$ in the parameter space such that up to a nonsingular and differentiable change of the parameters, which leaves the axes invariant as sets, the following is satisfied (cf. the bifurcation diagram Fig. 1.4):

(i) *There exists a C^{r-7} curve $\alpha_2 = \hom_1(\alpha_1)$ with $\alpha_1 > 0$ in Ω such that there exists a homoclinic orbit to a_1 in $\mathcal{U}$ if and only if $\alpha \in \hom_1$. Moreover, $\hom_1$ is asymptotically tangent to the positive α_1-axis as $\alpha_1 \to 0^+$ and the direction of the bifurcation is determined by the twist of the heteroclinic orbit Γ_2 as follows:*

$$\hom_1 \begin{cases} > 0 & \text{if } \Gamma_2 \text{ is twisted,} \\ < 0 & \text{otherwise.} \end{cases}$$

(ii) *There does not exist any k-heteroclinic orbit from a_2 to a_1 for $k \geqq 1$ if Γ_2 is not twisted and there exists at least one 1-heteroclinic orbit from a_2 to a_1 on a C^{r-7} curve; $\alpha_2 = 1\text{-het}_2(\alpha_1)$ for $\alpha_1 > 0$ otherwise. Moreover, 1-het_2 is asymptotically tangent to the α_1-axis as $\alpha_1 \to 0^+$.*

(iii) *Analogous statements hold for homoclinic orbits to a_2 and k-heteroclinic orbits from a_1 to a_2.*

(iv) *Let $\Lambda = \{(\alpha_1, \alpha_2): \alpha_2 > \hom_1(\alpha_1)$ if $\alpha_1 > 0$ or $\alpha_1 > \hom_2(\alpha_2)$ if $\alpha_2 > 0\}$. Then there exists a periodic orbit in $\mathcal{U}$ if and only if $\alpha \in \Lambda$.*

(v) *The homoclinic and periodic orbits do not coexist in $\mathcal{U}$ for a given parameter. They are all unique and are 1-hom and 1-per orbits, respectively.*

Our main result is as follows.

THEOREM B. *In addition to the hypotheses of Theorem A, suppose the stable manifolds of the equilibria a_1 and a_2 are all one-dimensional; then the following is satisfied (cf. Fig. 1.4):*

(i) *If Γ_2 is twisted but Γ_1 is not, then the 1-heteroclinic orbit from a_2 to a_1 is the unique k-heteroclinic orbit for all $k \geqq 1$ and $\alpha \in \Omega$. Moreover,*

$$1\text{-het}_2(\alpha_1) > \hom_1(\alpha_1) \quad \text{for } \alpha_1 > 0.$$

(ii) *If the loop $\operatorname{cl}(\Gamma_1 \cup \Gamma_2)$ is double twisted, then there exist two sequences of C^{r-7} curves $\alpha_2 = k\text{-het}_2(\alpha_1)$ with $\alpha_1 > 0$ and $\alpha_1 = k\text{-het}_1(\alpha_2)$ with $\alpha_2 > 0$ in Ω, respectively, satisfying*

$$0 \leqq k\text{-het}_2 < (k+1)\text{-het}_2 < \hom_1$$

for all $k \geqq 0$ such that there exists a k-heteroclinic orbit from a_2 to a_1 if and only if $\alpha \in k\text{-het}_2$. Moreover, it is a unique heteroclinic orbit in $\mathcal{U}$ with respect to the parameter and $k\text{-het}_2$ is asymptotically tangent to the α_1-axis as $\alpha_1 \to 0^+$. Furthermore, the homoclinic bifurcation curve $\hom_1$ is inaccessible from below in the sense that for every α_1

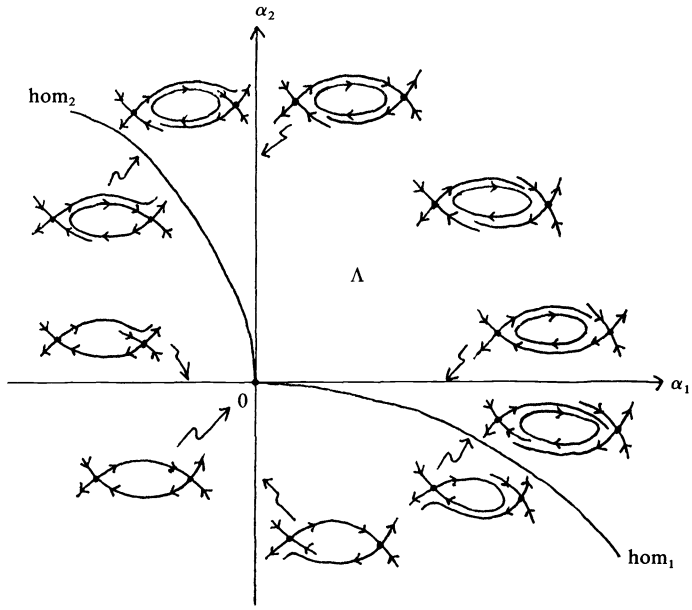
$$k\text{-het}_2(\alpha_1) \to \hom_1(\alpha_1) \quad \text{as } k \to +\infty.$$

An analogous statement also holds for the $k\text{-het}_1$ curves.

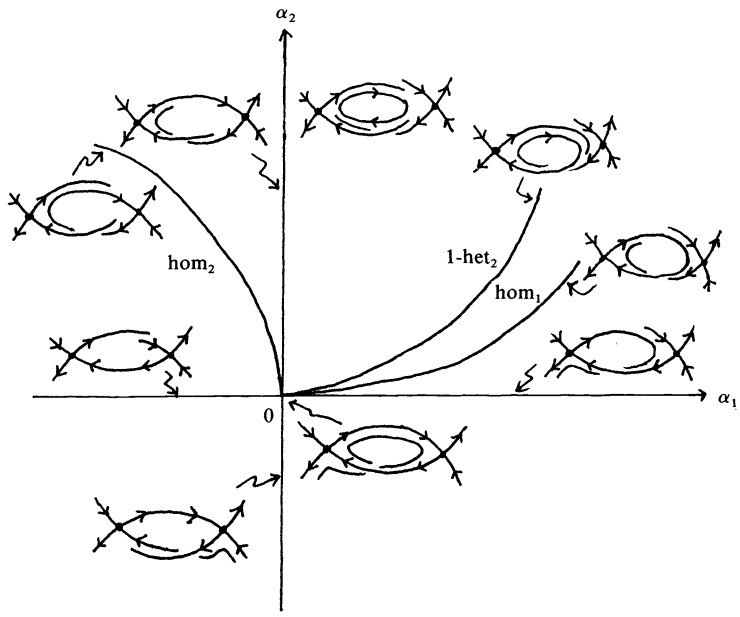
Theorem A provides us with a useful clue to the twist of a given heteroclinic loop: the two zero-heteroclinic continuation curves (which are the parameter axes in our theorems) divide the neighborhood Ω into four sectors. The 1-homoclinic bifurcation curves $\hom_1$ and $\hom_2$ lie in one sector for a double twisted loop, or in two adjacent sectors for a single twisted loop, or in two opposite sectors for a nontwisted loop. Keeping this fact in mind, let us examine the following bifurcation diagram Fig. 1.5 for traveling waves of the FitzHugh–Nagumo equation

$$v_t = v_{xx} + f(v) - w, \quad w_t = \varepsilon(v - \gamma w), \quad \varepsilon, \gamma \geqq 0,$$

where $f(v) = -v + H(v - a)$ with H to be the Heaviside step function and $0 < a < \frac{1}{2}$.

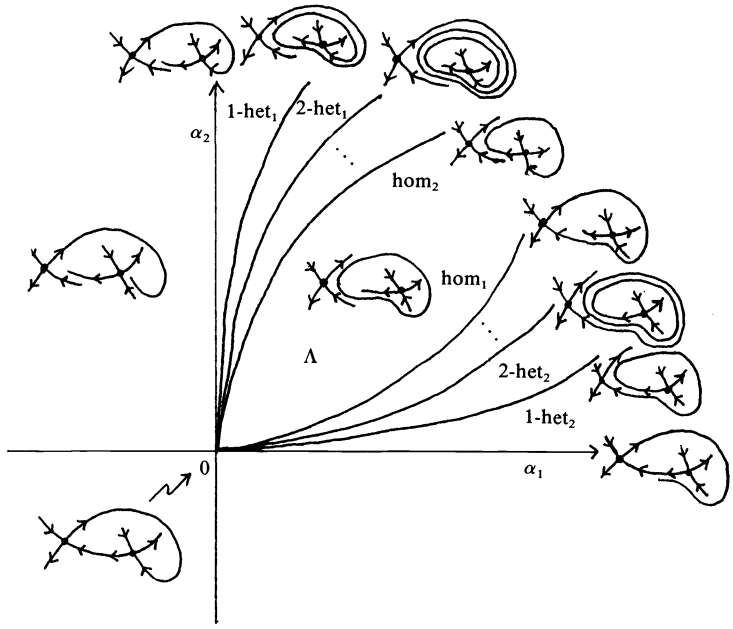


(a)



(b)

FIG. 1.4. *The bifurcation diagrams for* (a) *a nontwisted loop,* (b) *a single twisted loop,* (c) *a double twisted loop.*



(c)

FIG. 1.4.—continued

A traveling wave solution $(v, w)(x, t)$ is a function $(v, w)(z)$ of $z = x + ct$, $c \geqq 0$. Let $u(z) = v'(z)$, then $v_c = (v, u, w)(z)$ satisfies a first-order system of ODE

$$v' = u, \quad u' = cu - f(v) + w, \quad w' = \frac{\varepsilon}{c}(v - \gamma w). \tag{1.11}$$

For fixed $0 < a < \frac{1}{2}$ and $0 < \varepsilon \ll 1$, numerical as well as rigorous arguments from Rinzel and Terman (1982) show that the front curve θ_F, on which there exits a front wave connecting the rest steady state $\mathcal{O}$ to the exitable state $\mathcal{E}$ as shown in Fig. 1.6, crosses transversely the back curve θ_B, on which there is a back wave from $\mathcal{E}$ to $\mathcal{O}$. Thus, at the intersection point θ^* there exists a front wave and a back wave traveling at the same speed. This gives rise to a heteroclinic loop. Their numerical simulation also shows that the impulse curve $\theta_{\mathcal{O}}$, homoclinic to $\mathcal{O}$, and the $\mathcal{E}$-impulse curve $\theta_{\mathcal{E}}$, homoclinic to $\mathcal{E}$, also bifurcate from the loop at θ^*. Note that their bifurcation directions of asymptotic tangency are exactly opposite our bifurcation diagram Fig. 1.4 for Theorems A and B. This is due to the fact that the steady states $\mathcal{O}$ and $\mathcal{E}$ are relatively contractive simple saddles having one-dimensional stable manifolds only for the time reversed $(z \to -z)$ system (1.11). What is most remarkable about these two curves is that both of them lie in the same sector in the parameter space. In fact, this has been rigorously proved (see (3.8) and (3.10) from Rinzel and Terman (1982)). Unfortunately, however, we can only speculate that Theorem A suggests the double twist for the heteroclinic loop. Indeed, we are facing a tantalizing dilemma here: either it is feasible to check the transverse crossing condition (1.10b) and the double twist of the loop due to the piecewise linearity of f but the vector field is not smooth enough, or it becomes a fairly open problem to do so for a smooth vector field, e.g., the usual cubic function $f = v(v-1)(a-v)$. Nevertheless, the implication is interesting: for given

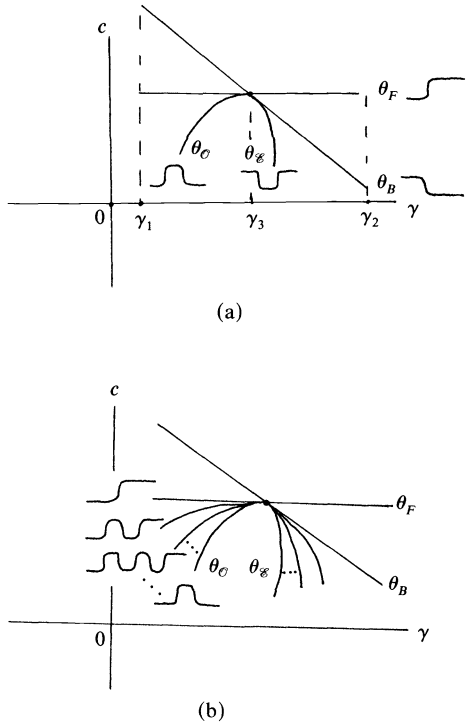


FIG. 1.5. (a) *A heuristic bifurcation diagram produced from Rinzel and Terman* (1982); (b) *The conjectured complete diagram.*

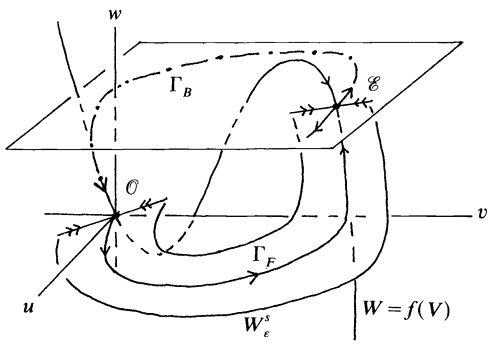


FIG. 1.6. *The conjectured twist for the front-back wave loop.*

parameters ε, $0 < a < \frac{1}{2}$ and $0 < \gamma_3 - \gamma \ll 1$ there would be infinitely many fronts traveling at different speeds. The more "humps" a front were to carry the slower it would travel. If the humps were "too many" (infinity) the traveling wave arising from the rest state would never be able to reach the exitable state $\mathscr{E}$ but would return to itself after a long excursion. Slowing down a little, it would become a traveling train, or periodic orbit for the ODE. On the other hand, push γ slightly to the right of γ_3, the above scenario would repeat for back waves and $\mathscr{E}$-impulses.

By their numerical evidence on the stability of the primary front and back waves with respect to the PDE, as well as other authors' results on somewhat related stability

problems, it has been demonstrated that the stability of a given impulse is closely related to the direction at which the stable and unstable manifolds cross transversely as the speed parameter c varies (see, e.g., Evans (1972), Jones (1984), Kokubu, Nishiura, and Oka (1988)). Thus, we find the second implication is most interesting: there would be infinitely many stable transition waves connecting two stable patterns.

However, all of these phenomena do not appear in the "chaotic" parameter region discussed by Evans, Fenichel, and Feroe (1982) and Hastings (1982), where for a given speed there are infinitely many impulses and traveling trains due to the Sil'nikov saddle-focus homoclinic explosion for (1.11) (see also Sil'nikov (1967)), whereas there would be a unique traveling front, or back, or impulse, or traveling train, except at the bifurcation point θ^* in our case. Indeed, as long as there are two bistable steady states as shown in Fig. 1.6, the equilibria are not saddle-focus. Nevertheless, we would probably not be too surprised by the enormous, stable, yet not "chaotical" transporting capability that a nerve axion would inherit if our conjectures were true.

In contrast to our conjectures above, we will discuss the existence of nontwisted heteroclinic orbits and thus the limited number of connections between two equilibrium states for another type of reaction diffusion systems in § 7. We will also discuss in that section some ways newly discovered by other authors to check all the nondegenerate and generic conditions (1.4), (1.5), (1.7), and (1.10a, b) for their examples to which our theory is immediately applicable.

2. Preliminaries. This section is devoted to introducing the Sil'nikov variables for a Poincaré map around the loop.

Let $0<\delta_0$ be a small number and $B(\delta_0)=\{z\colon z=(z^{(1)},\cdots,z^{(d)}), |z^{(i)}|\leqq\delta_0\}$ be the δ_0-box of the origin. Let the coordinate be locally normalized near the equilibria a_i so that $(x,y)=0$ corresponds to $z=a_i$ and the local stable, unstable manifolds are given by the x-axis and y-axis in $B(\delta_0)$, respectively; i.e., $W^s_{i\,\mathrm{loc}}=\{y=0\}\cap B(\delta_0)$ and $W^u_{i\,\mathrm{loc}}=\{x=0\}\cap B(\delta_0)$. In addition, the directions of the first x-component $x^{(1)}$ and the first y-component $y^{(1)}$ are chosen to be the unit principal eigenvectors e_i^+ and e_i^-, respectively, as in (1.8). Let the points p_i and q_i from (1.9) in the definition of twist be specifically given as $p_i=(x_{i*},0)$ and $q_i=(0,y_{i*})$. We can assume $x^{(1)}_{i*}=\delta_0$ and $y^{(1)}_{i*}=\delta_0$ because of assumption (1.7) for the asymptotic tangency of Γ_i along the principal eigendirections.

Let Σ_i^s and Σ_i^u be two small cross sections, or $(d-1)$-dimensional boxes $B(\delta_1)$ with $0<\delta_1<\delta_0$, centered at p_i and q_i and perpendicular to e_i^+ and e_i^-, respectively (see Fig. 2.1). They are Poincaré cross sections provided δ_0 and δ_1 are sufficiently small. Let σ_i^s be the subset of those initial points $(x_0,y_0)\in\Sigma_i^s$ whose trajectories in $B(\delta_0)$ first hit the exit cross section Σ_i^u at (x_1,y_1) at time $\tau=\tau(x_0,y_0)$. This correspondence gives rise to the local Poincaré map $\Pi_i\colon\sigma_i^s\to\Sigma_i^u$ by $(x_0,y_0)\to(x_1,y_1)$. Similarly, by the continuous dependence on initial data and parameters we can define a global Poincaré map $\Pi_{ij}\colon\Sigma_i^u\to\Sigma_j^s$. Here, without loss of generality Σ_i^u is taken to be the domain of definition for Π_{ij}, whereas σ_i^s is a proper subset of Σ_i^s not containing any point from the stable manifold $W^s_{i\,\mathrm{loc}}$.

Let $(\xi,y)\in\mathbb{R}^{d-1}$ and $(x,\eta)\in\mathbb{R}^{d-1}$ be the normalized local coordinates on Σ_i^s and Σ_i^u so that $(0,0)$ corresponds to the center points p_i and q_i, respectively. Indeed, $\xi=\hat{x}-\hat{x}_{i*}$ and $\eta=\hat{y}-\hat{y}_{i*}$, where $\hat{x}=(x^{(2)},\cdots,x^{(m)})$ and $\hat{y}=(y^{(2)},\cdots,y^{(n)})$ (see Fig. 2.1). Let $s=e^{-\mu_i(\alpha)\tau}$ be the Sil'nikov time near a_i, where $\mu_i(\alpha)$ is the principal unstable eigenvalue for $D_zF(a_i,\alpha)$. Then the Sil'nikov variable for the local map is (s,ξ,η) and the Sil'nikov domain is

$$\Delta_i:=\{(s,\xi,\eta)\colon 0<s<s_0(\alpha),|\xi|,|\eta|\leqq\delta_1\},$$

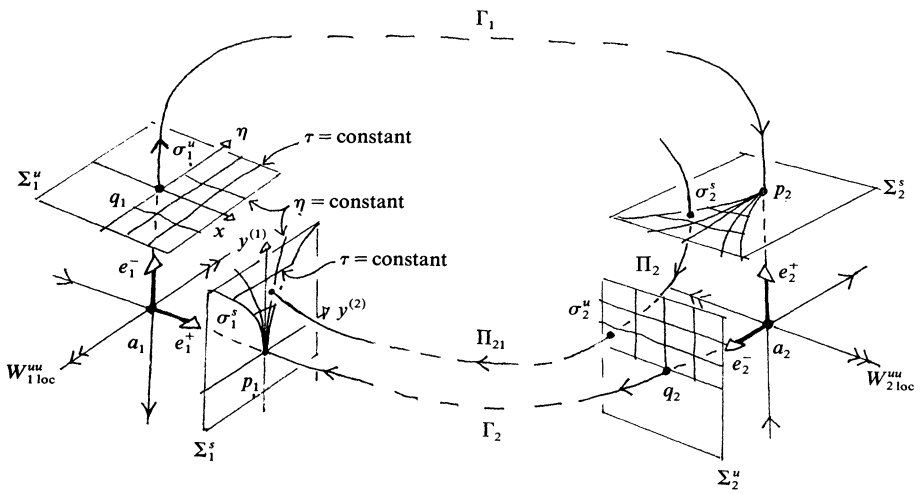


FIG. 2.1. *The cross sections and the corresponding Poincaré map when* $m = 1$, $n = 2$.

where $s_0(\alpha) = e^{-\mu_i(\alpha)\tau_0}$ for some large but fixed τ_0. Note that the dependence on α is suppressed from σ_i^s, σ_i^u, and Δ_i.

It has been proved by Sil'nikov (1967) that for the initial point $(x_0, y_0) \in \sigma_i^s$ and the end point $(x_1, y_1) \in \Sigma_i^u$ with $\tau = \tau(x_0, y_0)$ time units apart, the initial y_0 and the end x_1 components are functions of the Sil'nikov variable:

$$y_0 := Y_i(s, \xi, \eta, \alpha) \quad \text{and} \quad x_1 := X_i(s, \xi, \eta, \alpha).$$

Moreover, it has been observed by Deng (1989) that the maps $\rho_i^s : \Delta_i \to \sigma_i^s$ with $(s, \xi, \eta) \to (\xi, Y_i(s, \xi, \eta, \alpha))$ and $\rho_i^u : \Delta_i \to \sigma_i^s$ with $(s, \xi, \eta) \to (X_i(s, \xi, \eta, \alpha), \eta)$ are actually diffeomorphisms of class C^r. Thus, ρ_i^s gives rise to a smooth change of variables for the local Poincaré map Π_i, which in turn is ρ_i^u under the new Sil'nikov variable. See Fig. 2.2. More important, we have the following exponential expansion result.

PROPOSITION 2.1 (Deng (1988), (1989)). *Let the strong stable manifold and the strong unstable manifold also be normalized such that* $W_i^{ss} = \{x^{(1)} = 0, y = 0\}$ *and* $W_i^{uu} = \{x = 0, y^{(1)} = 0\}$ *locally. Let* $\nu_i(\alpha) = \lambda_i(\alpha)/\mu_i(\alpha) - 1$ *and* $\bar{\lambda}_i(\alpha)$ *and* $\bar{\mu}_i(\alpha)$ *be as in* (1.2). *Then for* δ_0 *sufficiently small there exist* C^{r-7} *functions* $\varphi_i(\xi, \eta, \alpha)$, $\psi_i(\xi, \eta, \alpha)$, $R_{i1}(s, \xi, \eta, \alpha)$, *and* $R_{i2}(s, \xi, \eta, \alpha)$ *over* Δ_i *such that*

$$\begin{aligned} X_i(s, \xi, \eta, \alpha) &= \varphi_i(\xi, \eta, \alpha) s^{1+\nu_i} + R_{i1}(s, \xi, \eta, \alpha), \\ Y_i(s, \xi, \eta, \alpha) &= \psi_i(\xi, \eta, \alpha) s + R_{i2}(s, \xi, \eta, \alpha) \end{aligned} \tag{2.1a}$$

with φ_i *and* ψ_i *satisfying*

$$\begin{aligned} \varphi_i(\xi, \eta, \alpha) &= e_m \delta_0 + O((|\xi| + |\eta| + \delta_0)\delta_0), \\ \psi_i(\xi, \eta, \alpha) &= e_n \delta_0 + O((|\xi| + |\eta| + \delta_0)\delta_0) \end{aligned} \tag{2.1b}$$

and $R_{ij} = R_{ij}(s, \xi, \eta, \alpha)$ *satisfying*

$$\begin{aligned} |D^k_{(\xi,\eta)} R_{i1}| &= O(s^{1+\nu_i+\bar{\nu}_i}), \qquad |D_s R_{i1}| = O(s^{\nu_i+\bar{\nu}_i}), \\ |D^k_{(\xi,\eta)} R_{i2}| &= O(s^{1+\bar{\nu}_i}), \qquad |D_s R_{i2}| = O(s^{\bar{\nu}_i}), \end{aligned} \tag{2.1c}$$

for all $(s, \xi, \eta) \in \Delta_i$, $0 \leqq k \leqq r - 7$, *where*

$$e_m = (1, 0, \cdots, 0)^T \in \mathbb{R}^m,$$

and

$$e_n = (1, 0, \cdots, 0)^T \in \mathbb{R}^n$$

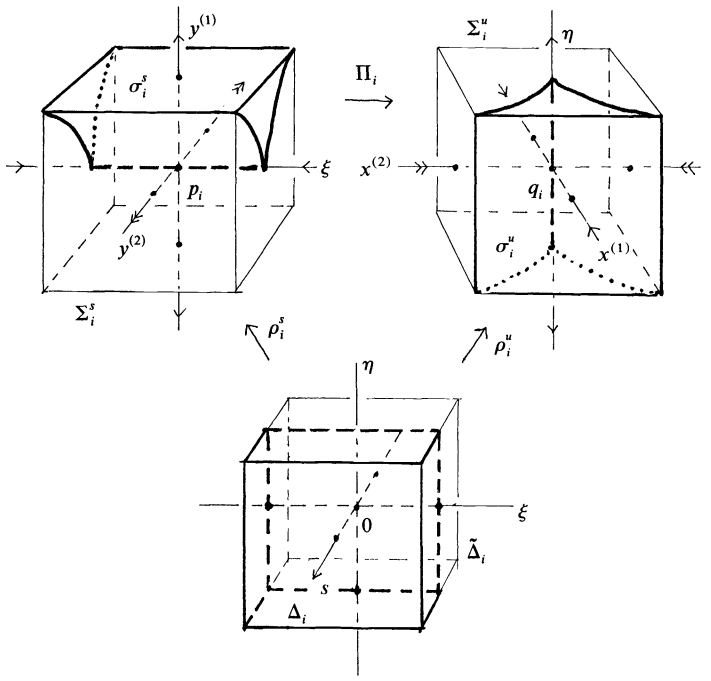


FIG. 2.2. *The Sil'nikov change of variables for the local map when* $m = n = 2$.

and $\bar{\nu}_i > 0$ *is a constant not greater than* $\min\{\bar{\mu}_i(\alpha)/\mu_i(\alpha), (\lambda_i(\alpha) - \bar{\lambda}_i(\alpha))/\mu_i(\alpha)\}$ *for all* $|\alpha| \leqq \delta_0$.

Equation (2.1a) is referred to as the exponential expansion, φ_i and ψ_i the expansion coefficient functions, and R_{ij}, the remainders. From this proposition we immediately have the following proposition.

PROPOSITION 2.2. *For sufficiently small but fixed* s_0, *the domain* σ_i^s *of the local map* Π_i *is contained in the vertical sector* $|\hat{y}| < y^{(1)}$ *on* Σ_i^s. *Moreover, for every boundary point* $(\xi, y) = (\xi, Y_i(s_0, \xi, \eta))$, $y^{(1)} \leqq \delta_0 s_0/2$. (*See Fig.* 2.2.)

It is also easy to see from (2.1c) that the functions X_i and Y_i, thus ρ_i^s and ρ_i^u, can be C^1 extended to $s \leqq 0$. From now on let us use the same notation for the extended functions, but $\tilde{\Delta}_i$ for the extended domain of Δ_i.

Let us conclude this section with two lemmas which will be frequently used later.

LEMMA 2.3. *Let the global map* Π_{ij} *be expressed as* $\xi = P_i(x, \eta, \alpha)$, $y = Q_i(x, \eta, \alpha)$, *under the new coordinates for* Σ_i^u *and* Σ_j^s. *Let*

$$M_i = \begin{bmatrix} D_\eta P_i(0,0,0) & 0 & -I \\ D_\eta Q_i(0,0,0) & e & 0 \end{bmatrix}_{(d-1)\times(d-1)} \quad \text{and} \quad \hat{M}_i = \begin{bmatrix} D_\eta P_i(0,0,0) & -I \\ D_\eta Q_i(0,0,0) & 0 \end{bmatrix}_{(d-2)\times(d-2)},$$

where $e = \psi_j(0,0,0)$, *or* e_n. *Then both* M_i *and* $\hat{M}_i$ *are nonsingular for sufficiently small* δ_0.

Proof. Note first that all the column vectors of M_i except the middle one $(0, e)$ span the linear subspace $T_{p_j}(W_i^u \cap \Sigma_j^s) + T_{p_j}(W_j^s \cap \Sigma_j^s)$, which has dimension $d-2$ by the hypothesis (1.4) and the choices of Σ_j, which are transverse to the flow. On the other hand, the remaining column vector is approximately parallel to the principal unstable eigenvector e_j^- by the exponential expansion property (2.1b). Thus, M_i achieves its maximal rank by (1.9). Moreover, the strong inclination property also implies $D_\eta \hat{Q}_i(0,0)$ is a diffeomorphism, thus the truncated square submatrix $\hat{M}_i$ achieves its maximal rank $d-2$ as well. □

LEMMA 2.4. *Let M_i be the same as in Lemma 2.3 above with $e=\psi_j(0,0,0)$ and*

$$N_i=\begin{bmatrix} D_\eta P_i(0,0,0) & 0 & -I \\ D_\eta Q_i(0,0,0) & f & 0 \end{bmatrix}_{(d-1)\times(d-1)}$$

with $f=(0,1,\cdots,0)^T,\cdots,(0,0,\cdots,1)^T$. Then there is a constant m_0 so that

$$\lim_{\delta_0\to 0}\frac{|\det M_i|}{\delta_0}\geqq m_0>0 \tag{2.2a}$$

and

$$\lim_{\delta_0\to 0}\det N_i=0. \tag{2.2b}$$

Proof. (2.2a) is true because of $\psi_j(0,0)/\delta\to e_n$ as $\delta_0\to 0$ and the strong inclination property (1.5) and (1.9). Since $(0,f)$ is contained in $T_{a_j}W_j^{uu}$, (2.2b) is also true for the same reason. □

3. Entrance and exit sets and their extensions. In this section we only consider the heteroclinic connections from a_2 to a_1. Analogous analysis and result can be immediately extended to the a_1 to a_2 connections. Again, the parameter α is suppressed from the text if no confusions arise.

Let $\mathrm{Ex}_2^0:=W_{2\mathrm{loc}}^u\cap\Sigma_2^u$ denote the intersection of the local unstable manifold of a_2 with the exit cross section Σ_2^u given in the previous section. Ex_2^0 is referred to as the initial exit set of W_2^u. It is obvious that there is a heteroclinic orbit (from a_2 to a_1) if and only if there is a solution of the initial exit set which also lies on the local stable manifold of the other equilibrium a_1. Thus, we need to closely follow the images of the initial exit set under those successive local and global Poincaré maps. To be precise, if we set the image of an empty set under a given map to be empty, then all the following sets are well defined:

$$\mathrm{En}_1^k:=\Pi_{21}(\mathrm{Ex}_2^{k-1}),\qquad \mathrm{En}_2^k:=\Pi_{12}(\mathrm{Ex}_1^k),$$
$$\mathrm{Ex}_i^k:=\Pi_i(\mathrm{En}_i^k\cap\rho_i^s(\Delta_i)),\qquad i=1,2,\quad k=1,2,\cdots,$$

where $\rho_i^s:\Delta_i\to\sigma_i^s$ is the Sil'nikov change of variables and $\rho_i^s(\Delta_i)$ is contained in the domain σ_i^s of Π_i. En_i^k and Ex_i^k are referred to as the kth entrance set and the kth exit set of W_2^u near a_i, respectively. They might be empty except for the initial exit set Ex_2^0 and the first entrance set En_1^1 near a_1. Nevertheless, we have the following.

PROPOSITION 3.1. *There exists a $(k-1)$-heteroclinic orbit from a_2 to a_1 sufficiently close to the loop $\Gamma_1\cup\Gamma_2$ if and only if $\mathrm{En}_i^j,\ \mathrm{Ex}_i^j\neq\varnothing$ for $1\leqq j\leqq k-1$, $i=1,2$, and $\mathrm{En}_1^k\cap W_{1\mathrm{loc}}^s\neq\varnothing$.*

By definition, for every point $(\xi,y)\in\mathrm{En}_1^k\neq\varnothing$ there exist $(0,\eta)\in\mathrm{Ex}_2^0$ and $z_i^j\in\sigma_i^s$ with $z_i^j=(\xi_i^j,y_i^j)$, $1\leqq j\leqq k-1$ and $i=1,2$ from the orbit of $(0,\eta)$ such that $\Pi_{21}(0,\eta)=z_1^1$, $\Pi_{12}\circ\Pi_1(z_1^1)=z_2^1,\cdots,\Pi_{21}\circ\Pi_2(z_2^{k-1})=(\xi,y)$. Using the "pull back," we have a unique $\zeta_i^j=(s_i^j,\xi_i^j,\eta_i^j)\in\Delta_i$ satisfying $z_i^j=\rho_i^s(\zeta_i^j)$. Thus, the following proposition is valid.

PROPOSITION 3.2. (3.2a) *The kth entrance set to a_1, En_1^k, is nonempty if and only if the following system of $l_1=(2k-1)(d-1)$ equations has solutions for the l_1+n-1 unknown variables $\eta,\zeta_1^1,\cdots,\zeta_2^{k-1},\xi,y$ with $\zeta_i^j=(s_i^j,\xi_i^j,\eta_i^j)\in\Delta_i$ satisfying the constraints $s_i^j>0$ for all $i=1,2$ and $1\leqq j\leqq k-1$:*

$$\Pi_{21}(0,\eta)=\rho_1^s(\zeta_1^1)$$
$$\vdots$$
$$\Pi_{21}(\rho_2^u(\zeta_2^{k-1}))=(\xi,y);$$

(3.2b) *The kth exit set from* a_1, Ex_1^k, *is nonempty if and only if after replacing* (ξ, y) *by* $\rho_1^s(\zeta_1^k)$ *with* $\zeta_1^k \in \Delta_1$ *the same statement of* (3.2a) *holds true*;

(3.2c) *The kth entrance set to* a_2, En_2^k, *is nonempty if and only if the following system of* $l_2 = 2k(d-1)$ *equations has solutions for the* $l_2 + n - 1$ *unknown variables*

$$\eta, \zeta_1^1, \cdots, \zeta_1^k, \xi, y$$

with $\zeta_i^j = (s_i^j, \xi_i^j, \eta_i^j) \in \Delta_i$ *satisfying the constraints* $s_i^j > 0$ *for all* $i = 1, 2$ *and* $1 \leqq j \leqq k$:

$$\Pi_{21}(0, y) = \rho_1^s(\zeta_1^1)$$
$$\vdots$$
$$\Pi_{12}(\rho_1^u(\zeta_1^k)) = (\xi, y);$$

(3.2d) *The kth exit set from* a_2, Ex_2^k, *is nonempty if and only if after replacing* (ξ, y) *by* $\rho_2^s(\zeta_2^k)$ *with* $\zeta_2^k \in \Delta_2$ *the same statement of* (3.2c) *holds true.*

Solving equations (3.2a-d) is more difficult with the constraints $s_i^j > 0$ than without them when those maps ρ_i^s and ρ_i^u are considered as the extended maps on the extended domain $\tilde{\Delta}_i$ introduced in the previous section. Let us now study the extended equations and leave the consideration of the constraints to the next section.

Note that each system of the extended equations (3.2a-d) might locally define an $(n-1)$-dimensional manifold near the origin of the l-dimensional Euclidean space $\mathbb{R}^l$, where $l = l_i + n - 1$. In fact, we have the following.

LEMMA 3.3. *Suppose the equilibria are simple saddle, relatively contractive, and the heteroclinic loop is nondegenerate. Then there exists a small constant* $\delta > 0$ *independent of k such that in the* δ*-box* $B(\delta)$ *of the origin in* $\mathbb{R}^l$ *each of the systems of the extended equations* (3.2a-d) *defines an* $(n-1)$*-dimensional manifold* $\mathcal{M}$ *in* $B(\delta)$ *which contains the origin and can be written as the graph of a* C^1 *vector-valued function of the last* $n-1$ *components, i.e., either* $\hat{y} = (y^{(2)}, \cdots, y^{(n)})$ *or* η_i, *where* $l = l_1 + n - 1$, *or* $l_2 + n - 1$, *accordingly.*

Proof. We prove the lemma by the implicit function theorem for equation (3.2a) only since the other cases are identical. By using the notation $\Pi_{ij} = (P_i, Q_i)^T$, $\rho_i^s = (\xi_i, Y_i)$ and $\rho_i^u = (X_i, \eta_i)$ from the last section, we see that solving equation (3.2a) is equivalent to solving the zero of the following equation:

$$\Phi(\zeta, \hat{y}) = 0,$$

where

$$\Phi(\zeta, \hat{y}) = \begin{bmatrix} -\xi_1^1 + P_2(0, \eta) \\ -Y_1^1 + Q_2(0, \eta) \\ -\xi_2^1 + P_1(X_1^1, \eta_1^1) \\ -Y_2^1 + Q_1(X_1^1, \eta_1^1) \\ \vdots \\ -\xi + P_2(X_2^{k-1}, \eta_2^{k-1}) \\ -y + Q_2(X_2^{k-1}, \eta_2^{k-1}) \end{bmatrix},$$

$$\zeta = (\eta, s_1^1, \xi_1^1, \eta_1^1, \cdots, s_2^{k-1}, \xi_2^{k-1}, \eta_2^{k-1}, \xi, y^{(1)}) \in \mathbb{R}^{l_1},$$

$X_i^j = X_i(s_i^j, \xi_i^j, \eta_i^j)$ and $Y_i^j = Y_i(s_i^j, \xi_i^j, \eta_i^j)$. It is obvious that the existence of the heteroclinic loop $\Gamma_1 \cup \Gamma_2$ (at $\alpha = 0$!) implies $\Phi(0, 0) = 0$. A simple calculation yields that the $l_1 \times l_1$ square Jacobian matrix $\partial\Phi/\partial\zeta$ at $(\zeta, \hat{y}) = (0, 0)$ has the following diagonal property

$$\left| \det \frac{\partial \Phi}{\partial \zeta}(0, 0) \right| = |\det \operatorname{diag}(M_2, M_1, \cdots, M_1, M_2)|,$$

where the diagonal blocks have the forms of Lemma 2.3 with all the blocks M_i except the last M_2 taking $e=-\psi_j(0,0,0)$. Therefore, the Jacobian matrix is nonsingular. Hence, by the implicit function theorem there exists a $\delta>0$ such that $\Phi(\zeta,\hat{y})=0$ defines in $B(\delta)$ an $(n-1)$-dimensional manifold $\mathcal{M}$ (containing the origin at $\alpha=0$) which is the graph of a C^1 function of the variable $\hat{y}$. Moreover, because of the diagonal block structure it is not difficult to see that δ can be chosen to be independent of the number of the equations $l_1=(2k-1)(d-1)$. □

Let $\mathbf{P}$ be the canonical projection from $\mathbb{R}^l=\mathbb{R}^{l-(d-1)}\times\mathbb{R}^{d-1}$ onto the last $d-1$ components. Then Lemma 3.3 implies that the projection $\mathbf{P}\mathcal{M}$ of the manifold $\mathcal{M}$ is also the graph of a C^1 function of the last $n-1$ coordinates. Therefore, we have the following.

DEFINITION 3.4. $\tilde{\mathrm{En}}_i^k:=\mathbf{P}\mathcal{M}$ and $\tilde{\mathrm{Ex}}_i^k:=\rho_i^u(\mathbf{P}\mathcal{M})$ are called the extended kth entrance set and the extended kth exit set (of W_2^u) near a_i, respectively, according to whether $\mathcal{M}$ is taken to be the manifold defined by the extended equations (3.2a–d) for the entrance sets or the exit sets in Lemma 3.3.

Since all the extended entrance and exit sets exist in some small but fixed δ-box $B(\delta)$ of the center points on the entrance and exit cross sections, respectively, we can easily construct a tubular neighborhood $\mathcal{U}$ of the heteroclinic loop $\mathrm{cl}(\Gamma_1\cup\Gamma_2)$ such that the intersections of $\mathcal{U}$ with the entrance and exit cross sections (Σ_i^s and Σ_i^u) are exactly those δ-boxes. Thus, we only need to consider the real entrance and exit sets (of W_2^u) in $B(\delta)$ and rename $E\beta^k:=E\beta^k\cap B(\delta)$ for simplicity of notation, where $\beta=n_i$ or x_i. Now we are ready to compare these sets with their extensions. Because the kth extended entrance set $\tilde{\mathrm{En}}_1^k$ near a_1 is a graph over the last $n-1$ coordinates $\hat{y}$ on Σ_1^s and the nonemptiness of its intersection with the local stable manifold W_1^s of a_1 forces $\hat{y}=0$, we have proved the following.

COROLLARY 3.5. *If a $(k-1)$-heteroclinic orbit from a_2 to a_1 exists in $\mathcal{U}$, then it must be unique (for the corresponding parameter).*

COROLLARY 3.6. *Let $\tilde{\mathrm{En}}_i^k=\mathrm{graph}(G,H)$ with $\xi=G(\hat{y})$ and $y^{(1)}=H(\hat{y})$. If $\dim W_1^s=\dim W_2^s=1$, $|H(\hat{y})|\leqq\delta_0 s_0/4$ and the derivative $|DH(\hat{y})|\leqq\frac{1}{2}$ for all $|\hat{y}|<\delta$, then $\tilde{\mathrm{En}}_i^k\cap\sigma_i^s\neq\varnothing$ if and only if $0<H(0)$, where δ_0 and s_0 are as in Proposition 2.2.*

Proof. Since $\dim W_1^s=\dim W_2^s=1$, $\xi=G(y)=\delta_0$, the x-component of the center of the entrance section Σ_i^s. Since the boundary point $(\delta_0,y_0)=(\delta_0,Y(s_0,0))\in\partial\sigma_1^s$ satisfies $y_0^{(1)}\geqq\delta_0 s_0/2>\max|H|$ by Proposition 2.2 and our assumption, the two boundary points (δ_0,y_0) and $(\delta_0,0)$ must be in different sides of $\tilde{\mathrm{En}}_i^k$ if $0<H(0)$. The path connectedness of σ_1^s implies there must be a point $(\delta_0,Y(s_1,\delta_0,0))\in\tilde{\mathrm{En}}_i^k$. This shows $0<H(0)$ is sufficient. To show that it also necessary, suppose it false, i.e., $H(0)\leqq 0$. Since $\tilde{\mathrm{En}}_1^k=\mathrm{graph}(H)$, then $y_0^{(1)}=H(0)\leqq 0$. Thus, $y^{(1)}-y_0^{(1)}\leqq|DH||\hat{y}|\leqq|\hat{y}|/2$ by our assumption. Let $y_1\in\tilde{\mathrm{En}}_i^k\cap\sigma_i^s\neq\varnothing$; then $|\hat{y}_1|<y_1^{(1)}$ holds true by Proposition 2.2. It follows that $|\hat{y}_1|-y_0^{(1)}<y_1^{(1)}-y_0^{(1)}\leqq|\hat{y}_1|/2$ and $0\leqq|\hat{y}_1|/2<y_0^{(1)}$, a contradiction. □

Our main result of this section is as follows.

THEOREM 3.7. *Let $\tilde{\mathrm{En}}_i^k=\mathrm{graph}(G,H)$ with $\xi=G(\hat{y})$ and $y^{(1)}=H(\hat{y})$. If the derivative of H satisfies $|DH|\leqq\frac{1}{2}$ for all k, $i=1,2$ and $E\beta^k$ is nonempty, then $E\beta^k=\tilde{E}\beta^k$, where $\beta=n_i$ or x_i.*

Proof. Suppose it is false; then there exists a first $E\beta^k$ such that $E\beta^k\neq\tilde{E}\beta^k$. We claim first $\beta=x_i$. If $\beta=n_i$ then there exists a point $p_0\in\tilde{\mathrm{En}}_i^k-\mathrm{En}_i^k$. Hence, there exists a point $(\zeta,\hat{y})\in\mathcal{M}$ with $p_0=\mathbf{P}(\zeta,\hat{y})$, and $(\zeta,\hat{y})$ has at least one $s_i^j\leqq 0$, where $\mathcal{M}$ and the projection $\mathbf{P}$ are as in Definition 3.4. To be precise, say $s_1^j\leqq 0$.

Let us first note the following: Denote $\mathcal{M}=\mathscr{E}\beta^k$ according to whether it is obtained by the extended equations for the kth entrance set when $\beta=n_i$ or the exit set when $\beta=x_i$. Now it is not difficult to see that if $p\in\mathscr{E}\beta^k$ then the point q, whose components consist of the first $(2j-1)(d-1)+n-1$ components of p, belongs to $\mathscr{E}x_1^j$ since q satisfies

the first $(2j-1)(d-1)$ equations (cf. Proposition 3.2). Similarly, if q is obtained by keeping the first $2j(d-1)+n-1$ components of p, then q is in $\mathscr{E}x_2^j$.

Now, resume our assumption $s_1^j \leqq 0$ and let q_0 be such a truncated point of p_0 which belongs to $\mathscr{E}x_1^j$. Then $\mathbf{P}(q_0) \in \tilde{E}x_1^j - \mathrm{Ex}_1^j$ since $s_1^j \leqq 0$. This contradicts our assumption for $\tilde{\mathrm{En}}_i^k$. Hence, the claim holds true.

Since we have $\varnothing \neq \mathrm{Ex}_i^k \neq \tilde{\mathrm{Ex}}_i^k$, it follows that there exist $\zeta_j = (s_j, \xi_j, \eta_j) \in \tilde{\Delta}_i$ with $j = 1, 2$, $s_1 \leqq 0$, and $s_2 > 0$ such that $\rho_i^s(\zeta_1) \in \tilde{\mathrm{En}}_i^k - \mathrm{En}_i^k$ and $\rho_i^s(\zeta_2) \in \mathrm{En}_i^k \cap \sigma_i^s$ for the same reason as above on the truncated point q. Since $\tilde{\mathrm{En}}_i^k$ is path connected, being a graph over the path connected set $|\hat{y}| < \delta$, there exists a $\zeta_0 = (s_0, \xi_0, \eta_0) \in \tilde{\Delta}_i$ with $s_0 = 0$. Hence, $\rho_i^s(\zeta_0) = (\xi_0, Y_i(0, \xi_0, \eta_0)) = (\xi_0, 0) \in \partial\sigma_i^s \cap W_{1\mathrm{loc}}^s$. That is, $(\xi_0, 0) \in \tilde{\mathrm{En}}_i^k =$ graph (G, H). This implies $|y^{(1)}| \leqq |DH||\hat{y}| \leqq |\hat{y}|/2$ by our assumption for all $(\xi, y) \in \tilde{\mathrm{En}}_i^k$. It follows that $\tilde{\mathrm{En}}_i^k \cap \sigma_i^s = \varnothing$, since by Proposition 2.2 $|\hat{y}| < y^{(1)}$ for all $(\xi, y) \in \sigma_i^s$. This contradicts $\rho_i^s(\zeta_2) \in \mathrm{En}_i^k \cap \sigma_i^s \subset \tilde{\mathrm{En}}_i^k \cap \sigma_i^s$. □

4. Bifurcation equations. From now on, we shall spell out the parameter explicitly wherever it is necessary. In this section we consider the constraints $s_i^j > 0$ in terms of their sign changes with the parameter, in particular the sign changes of $y^{(1)} = H(0)$. Here (G, H) gives the graph of the extended entrance set. For this reason, we consider the following equations:

$$\Pi_{21}(0, \eta, \alpha) = (\xi, y), \tag{4.1a}$$

$$\Pi_{21}(0, \eta, \alpha) = \rho_1^s(\zeta_1, \alpha), \tag{4.1b}$$

$$\Pi_{21}(\rho_2^u(\zeta_2, \alpha), \alpha) = (\xi, y), \tag{4.1c}$$

$$\Pi_{21}(\rho_2^u(\zeta_2, \alpha), \alpha) = \rho_1^s(\zeta_1, \alpha), \tag{4.1d}$$

$$\Pi_{12}(\rho_1^u(\zeta_1, \alpha), \alpha) = (\xi, y), \tag{4.1c'}$$

$$\Pi_{12}(\rho_1^u(\zeta_1, \alpha), \alpha) = \rho_2^s(\zeta_2, \alpha), \tag{4.1d'}$$

which introduce every new s_i^j or $y^{(1)}$ into our recursive construction of the entrance and exit sets in the last section, where $\zeta_i = (s_i, \xi_i, \eta_i)$.

Note that each of the systems above defines a system of $d-1$ equations with $l_1 = (d-1)+(n-1)+2$ unknown variables for (4.1a-b) or $l_2 = 2(d-1)+2$ variables for (4.1c-d'), including the parameters α_1 and α_2. Thus, presumably, each of them defines an $(l_i-(d-1))$-dimensional manifold in $\mathbb{R}^{l_i}$ accordingly. Indeed, we have the following lemma.

LEMMA 4.1. *Suppose the conditions of Theorem* A *are satisfied. Then there exists a small constant* $\delta > 0$ *such that in the* δ*-box* $B(\delta)$ *of the origin in* $\mathbb{R}^{l_i}$ *each of the extended equations* (4.1a-d) *defines an* $(l_i-(d-1))$*-dimensional differential manifold in* $B(\delta)$ *which can be written as the graph of a* C^1 *function of the last* $(n-1)+2$ *variables* $(\hat{y}, \alpha)$ *or* (η, α) *when* $l_i = l_1$ *or the first* m *and the last* $(n-1)+2$ *variables* $(s_i, \xi_i, \hat{y}, \alpha)$ *or* $(s_i, \xi_i, \eta, \alpha)$ *when* $l_i = l_2$, *accordingly. Moreover, up to only one nonsingular and differentiable change of the parameter for all the equations* (4.1a-d') *considered, the following bifurcation equations are satisfied for solutions to* (4.1a-d') *with the corresponding alphabetical order*:

$$m_1 y^{(1)} + \cdots + m_n y^{(n)} = \alpha_2 + O(|\hat{y}|^2), \tag{4.2a}$$

$$s_1 = \alpha_2 + O(|\eta_1||s_1| + |s_1|^{1+\bar{\nu}_1}), \tag{4.2b}$$

$$m_1 y^{(1)} + \cdots + m_n y^{(n)} = \alpha_2 + \tau_2(\alpha) s_2^{1+\nu_2} + O((|\xi_2| + |s_2|^{\bar{\nu}_2})|s_2|^{1+\nu_2} + |\hat{y}|^2), \tag{4.2c}$$

$$s_1 = \alpha_2 + \tau_2(\alpha) s_2^{1+\nu_2} + O((|\xi_2| + |s_2|^{\bar{\nu}_2})|s_2|^{1+\nu_2} + (|\eta_1| + |s_1|^{\bar{\nu}_1})|s_1|), \tag{4.2d}$$

$$\hat{m}_1 y^{(1)} + \cdots + \hat{m}_n y^{(n)} = \alpha_1 + \tau_1(\alpha) s_1^{1+\nu_1} + O((|\xi_1| + |s_1|^{\bar{\nu}_1})|s_1|^{1+\nu_1} + |\hat{y}|^2), \tag{4.2c'}$$

(4.2d′) $$s_2 = \alpha_1 + \tau_1(\alpha)s_1^{1+\nu_1} + O((|\xi_1| + |s_1|^{\bar{\nu}_1})|s_1|^{1+\nu_1} + (|\eta_2| + |s_2|^{\bar{\nu}_2})|s_2|),$$

where ν_i, $\bar{\nu}_i$ are as in Proposition 2.1. *Here $m_i = m_i(\alpha)$ are differentiable functions of α satisfying $1/(2\delta_0) < m_1 < 2/\delta_0$, $m_i/m_1 = o(1)$ as $\delta_0 \to 0$ for $i \neq 1$ and the analogous properties also hold for $\hat{m}_i$. Moreover, the scalar functions, $\tau_i = \tau_i(\alpha)$, called twist functions, are nonzero, differentiable, and satisfy*

(4.3) $$\tau_i(0)\begin{cases} <0 & \text{if } \Gamma_i \text{ is twisted,} \\ >0 & \text{otherwise.} \end{cases}$$

Furthermore, the change of the parameters leaves the parameter axes, as sets, invariant, but may reverse their directions.

The basic framework for the proof of this lemma, in particular, the derivation of the bifurcation equations through a modified Lyapunov-Schmidt reduction, has much in common with the spirit of Chow, Deng, and Terman (1990). Thus, we will prove it in the Appendix with necessary modifications given to the twist terms $\tau_i(\alpha)s_i^{1+\nu_i}$ and the order estimates on the higher-order terms.

The following corollaries concern the conditions of Corollary 3.6 and Theorem 3.7 when the parameter is taken into consideration.

COROLLARY 4.2. *Let $\tilde{\text{En}}_i^k = \text{graph}\,(G(\cdot, \alpha), H(\cdot, \alpha))$ with $\xi = G(\hat{y}, \alpha)$ and $y^{(1)} = H(\hat{y}, \alpha)$. Then δ_0 and δ can be chosen sufficiently small but fixed such that $|H(\hat{y}, \alpha)| \leqq \delta_0 s_0/4$ and $|DH(\hat{y}, \alpha)| \leqq \frac{1}{2}$ for all $|\hat{y}|$, $|\alpha| < \delta$ and $k \geqq 1$, where D is the differentiation operator in $\hat{y}$.*

Proof. Using (4.2c) or (4.2c′), we have

$$|y^{(1)}| \leqq \left[\max_{i \neq 1} |m_i(\alpha)|\delta + O(\delta)\right] \Big/ m_1(\alpha) = o(1)\delta + 2\delta_0 O(\delta),$$

and

$$|DH| \leqq \left[\max_{i \neq 1} |m_i(\alpha)| + O(\delta^{\nu_0})\right] \Big/ m_1(\alpha) = o(1) + 2\delta_0 O(\delta^{\nu_0}),$$

where $\nu_0 = \min\{\nu_1, \nu_2, \bar{\nu}_1, \bar{\nu}_2\}$. Choosing δ_0 and δ so small but fixed implies the desired estimates. □

COROLLARY 4.3. *δ can be chosen small so that if $\tilde{\text{En}}_i^k = \text{graph}\,(G(\cdot, \alpha), H(\cdot, \alpha))$ crosses the stable manifold $W^s_{i\,\text{loc}} \cap \Sigma_i^s$, then it does so transversely in α_j in the sense that $\partial H(0, \alpha)/\partial \alpha_j > 0$ for $|\alpha| < \delta$. In other words, if $H(0, \alpha^0) = 0$ then for the fixed ith component $\alpha_i = \alpha_i^0$, $H(0, \alpha) > 0$ if and only if $\alpha_j > \alpha_j^0$.*

Proof. Using (4.2c) or (4.2c′) again, we have $m_1(\alpha)H(0, \alpha) = \alpha_j + O(|\alpha|^{1+\nu_0})$ with $\nu_0 = \min\{\nu_1, \nu_2, \bar{\nu}_1, \bar{\nu}_2\}$. Thus

$$\partial H(0, \alpha)/\partial \alpha_j \geqq \left(1 - \left[\left|\frac{\partial m_1}{\partial \alpha_j} H(0, \alpha)\right| + O(|\alpha|^{\nu_0})\right]\right) \Big/ m_1(\alpha) > 0$$

for $|\alpha| < \delta$ by an appropriately chosen small δ. □

COROLLARY 4.4. *The first entrance set En_1^1 to a_1 intersects the domain σ_1^s of the local map nonempty if and only if $\alpha_2 > 0$.*

Proof. By (4.2b), $s_1 = \alpha_2 + O(|\eta_1||s_1| + |s_1|^{1+\nu_1})$, implying $s_1 > 0$ if and only if $\alpha_2 > 0$. □

5. Proof of Theorem A. As we mentioned earlier, Theorem A has been proved in Chow, Deng, and Terman (1990), except for the directions of homoclinic bifurcations and the k-heteroclinic orbits. Thus, we are going to outline the proof from that paper and provide the necessary details for the other part of the proof.

Consider the Poincaré map $\Pi_{11}=\Pi_{21}\circ\Pi_2\circ\Pi_{12}\circ\Pi_1$ from a subset of σ_1^s into the entrance section Σ_1^s near a_1. Using the Sil'nikov changes of variables, we can similarly reduce the problem of finding periodic points of Π_{11} into solving a system of equations for the unknown Sil'nikov variables with the constraints $s_i>0$. The conditions of the nondegenerate heteroclinic loop and the relative contraction for the simple saddle equilibria imply that the extended system has a unique solution parametrized by α by the implicit function theorem. This uniqueness allows us to consider the simple homoclinic, periodic orbits only. Thus, by the implicit function theorem, we solve $(\zeta_1^*, \xi_2^*)(\alpha)=(s_1^*, \xi_1^*, \eta_1^*, s_2^*, \xi_2^*, \eta_2^*)(\alpha)$ as the solution for the extended equations $\Pi_{12}(\rho_1^u(\zeta_1, \alpha), \alpha)=\rho_2^s(\zeta_2, \alpha)$ and $\Pi_{21}(\rho_2^u(\zeta_2, \alpha), \alpha)=\rho_1^s(\zeta_1, \alpha)$, where $\zeta_i=(s_i, \xi_i, \eta_i)$ with $|\zeta_i^*(\alpha)|=O(|\alpha|)$. Now, substituting ζ_1^* and ζ_2^* into the bifurcation equation (4.2d') and (4.2d), we have $s_2^*=\alpha_1+O(|\alpha|^{1+\nu_0})$ and $s_1^*=\alpha_2+O(|\alpha|^{1+\nu_0})$. It follows that the map $\alpha\to(s_1^*, s_2^*)$ is a diffeomorphism. Thus, the sector Λ for the periodic orbits is given by $s_1^*>0$ and $s_2^*>0$ and the curve $\hom_1$ for 1-homoclinic orbits from a_1 to a_1 is given as a piece of the boundary $\partial\Lambda$ with $s_1^*=0$ but $s_2^*>0$. Substituting $s_1^*=0$ and $s_2^*>0$ into (4.2d') and (4.2d), again we have $s_2^*=\alpha_1+O(|\alpha|)|s_2^*|^{1+\nu_0}$, implying $\alpha_1>0$, and

$$0=\alpha_2+\tau_2(\alpha)s_2^{*1+\nu_2}+O(|\alpha|)|s_2^*|^{1+\nu_2}, \tag{5.1}$$

implying $\alpha_2=\hom_1(\alpha_1)=[-\tau_2(\alpha)+O(|\alpha|)]|s_2^*|^{1+\nu_2}$. Therefore, the bifurcation directions in (i) hold true because of (4.3). Finally, to complete the proof, we only need to prove (ii).

Let us consider the $3(d-1)$ equations $\Pi_{21}(0, \eta, \alpha)=\rho_1^s(\zeta_1, \alpha)$,

$$\Pi_{12}(\rho_1^u(\zeta_1, \alpha), \alpha)=\rho_2^s(\zeta_2, \alpha), \text{ and } \Pi_{21}(\rho_2^u(\zeta_2, \alpha))=(\xi, 0) \quad \text{for } 3(d-1)+1 \text{ variables,}$$

including the two parameters α_1 and α_2, and assume Γ_2 is twisted first. We solve the following $3(d-1)-1$ equations first:

$$\Phi(\zeta, \alpha)=\begin{bmatrix} -\xi_1+P_2(0, \eta, \alpha) \\ -Y_1+Q_2(0, \eta, \alpha) \\ -\xi_2+P_1(X_1, \eta_1, \alpha) \\ -Y_2+Q_1(X_1, \eta_1, \alpha) \\ -\xi+P_2(X_2, \eta_2, \alpha) \\ \hat{Q}_2(X_2, \eta_2, \alpha) \end{bmatrix}=0, \tag{5.2}$$

with $\hat{Q}_2=(Q_2^{(2)}, \cdots, Q_2^{(n)})$, and solve the leftover equation $Q_2^{(1)}(X_2, \eta_2, \alpha)=0$ later, where $\zeta=(\eta, s_1, \xi_1, \eta_1, s_2, \xi_2, \eta_2, \xi)$. The existence of the heteroclinic loop implies $\Phi(0, 0)=0$ and a simple calculation shows the Jacobian square matrix satisfies

$$\left|\det\frac{\partial\Phi}{\partial\zeta}(0, 0)\right|=|\det \operatorname{diag}(M_2, M_1, \hat{M}_2)|,$$

where M_i and $\hat{M}_2$ are the same as in Lemma 2.3 with e being $\psi_j(0, 0)$ in M_i. Hence, it follows from Lemma 2.3 that the Jacobian $\partial\Phi/\partial\zeta(0, 0)$ is nonsingular and ζ can be solved as a C^1 function ζ^* of α satisfying $\zeta^*(0)=0$ by the implicit function theorem. Substituting $\zeta=\zeta^*(\alpha)$ into the remaining equation $Q_2^{(1)}(X_2^*, \eta_2^*, \alpha)=0$, we find it equivalent to solving α from

$$\begin{aligned} &\xi^*=P_2(X_2^*, \eta_2^*, \alpha), \\ &0=Q_2(X_2^*, \eta_2^*, \alpha). \end{aligned} \tag{5.3}$$

Notice that this equation has the form of the connecting equation (4.2c) and the corresponding bifurcation equation (4.2c) of Lemma 4.1 applies. Thus, it is equivalent to

$$0=\alpha_2+\tau_2(\alpha)|s_2^*|^{1+\nu_2}+O((|\xi_2^*|+|s_2^*|^{\bar\nu_2})|s_2^*|^{1+\nu_2}). \tag{5.4}$$

Since s_2^*, $\xi_2^*=O(|\alpha|)$, this equation always has a unique solution $\alpha_2:=1\text{-het}_2^{\wedge}(\alpha_1)$ for every α_1 by the implicit function theorem. Since Γ_2 is twisted, then $\tau_2(\alpha)<0$. This implies $1\text{-het}_2^{\wedge}>0$. Also $1\text{-het}_2^{\wedge}=O(|\alpha_1|^{1+\nu_2})$. To see if the constraint $s_1^*(\alpha)>0$ and $s_2^*(\alpha)>0$ are satisfied at $\alpha\in 1\text{-het}_2^{\wedge}$ we need to consider the other two bifurcation equations (4.2b) and (4.2d′) corresponding to the first two connections:

$$s_1^*=\alpha_2+O(|\eta_1^*||s_1^*|+|s_1^*|^{1+\bar\nu_1}), \tag{5.5a}$$

$$s_2^*=\alpha_1+\tau_1(\alpha)|s_1^*|^{1+\nu_1}+O((|\xi_1^*|+|s_1^*|^{\bar\nu_1})|s_1^*|^{1+\nu_1}+(|\eta_2^*|+|s_2^*|^{\bar\nu_2})|s_2^*|). \tag{5.5b}$$

From (5.5a) and $\alpha_2=1\text{-het}_2^{\wedge}(\alpha_1)>0$ it is obvious to see that $s_1^*>0$ is automatically satisfied. Moreover, $s_1^*=O(\alpha_2)=O(|\alpha_1|^{1+\nu_2})$. Substituting this order for s_1^* into (5.5b) yields $s_2^*>0$ if and only if $\alpha_1>0$ since α_1 is the leading term in the right-hand side when s_1^* and α_2 are of order $O(|\alpha_1|^{1+\nu_2})$. Let $1\text{-het}_2:=1\text{-het}_2^{\wedge}|_{\alpha_1>0}$ be the desired curve.

To show the nonexistence of k-heteroclinic orbits for $k\geqq 1$ under the nontwist assumption for Γ_2, let us solve a system of equations similar to (5.2). Analogously, it is equivalent to solving those α from the following bifurcation equations so that $s_i^{j^*}(\alpha)>0$ for all i,j:

$$\begin{aligned}
&s_1^{1^*}=\alpha_2+O((|\eta_1^{1^*}|+|s_1^{1^*}|^{\bar\nu_1})|s_1^{1^*}|)\\
&s_2^{1^*}=\alpha_1+\tau_1|s_1^{1^*}|^{1+\nu_1}+O((|\xi_1^{1^*}|+|s_1^{1^*}|^{\bar\nu_1})|s_1^{1^*}|^{1+\nu_1}+(|\eta_2^{1^*}|+|s_2^{1^*}|^{\bar\nu_2})|s_2^{1^*}|)\\
&\qquad\vdots\\
&0=\alpha_2+\tau_2|s_2^{k^*}|^{1+\nu_2}+O((|\xi_2^{k^*}|+|s_2^{k^*}|^{\bar\nu_2})|s_2^{k^*}|^{1+\nu_2}).
\end{aligned} \tag{5.6}$$

Since $\tau_2>0$, the last equation implies $\alpha_2<0$. This forces $s_1^{1^*}<0$ from the first equation. This completes the proof. □

6. Proof of Theorem B. To prove Theorem B we need the following three lemmas.

When $\dim W_i^s=1$, two given graphs $E_j=\text{graph}\,(H_j)$ over $\hat y$ on the entrance set Σ_i^s are denoted by $E_1\leqq E_2$ if $H_1(\hat y)\leqq H_2(\hat y)$, or $E_1<E_2$ if $H_1(\hat y)<H_2(\hat y)$ for all common $\hat y$. A point $p=(\delta_0,y)\in\Sigma_i^s$ is said to satisfy $p\leqq(<)E=\text{graph}\,(H)$ if $y^{(1)}\leqq(<)H(\hat y)$. In what follows, let $\text{Fx}_1^0:=W_{1\text{loc}}^u\cap\Sigma_1^u$ and Fn_1^1 be the corresponding first entrance set of W_1^u to a_1. Their definitions are analogous to Ex_2^0 and En_1^1, respectively. Now we have Lemma 6.1 (cf. Fig. 6.1).

LEMMA 6.1. *If* $\dim W_1^s=\dim W_2^s=1$ *and all the entrance sets* Fn_i^1 *and* En_i^j *to* a_i *up to a given number* $k\geqq j$ *are nonempty graphs over the* $\hat y$*-axis, then*

If Γ_2 *is twisted but* Γ_1 *is not, then*

$$\text{En}_1^2<\text{En}_i^j<\text{Fn}_1^1<\text{En}_1^1 \quad \textit{for } 3\leqq j\leqq k; \tag{6.1a}$$

If both Γ_1 *and* Γ_2 *are twisted, then*

$$\text{Fn}_1^1<\text{En}_i^k<\text{En}_i^{k-1}<\cdots<\text{En}_1^2<\text{En}_1^1. \tag{6.1b}$$

Proof. The proof is based on the following two simple observations: (1) The range σ_i^u of the local map Π_i contains all the exit sets and lies to one side of the corresponding initial exit graph (Fx_1^0 or Ex_1^0). The images of the exit sets under the global map Π_{ij}

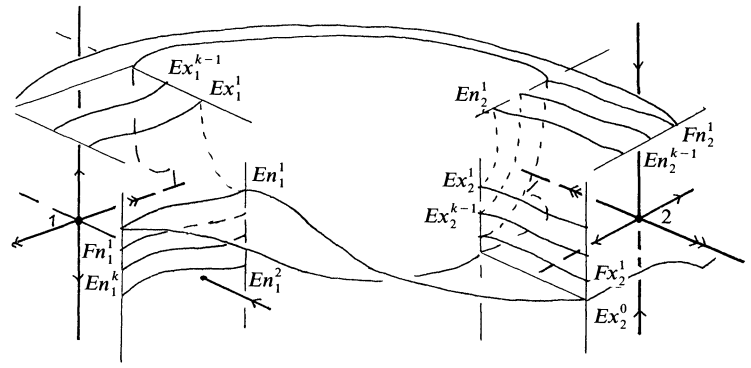
lie below (above), i.e., $<(>)$, the corresponding first entrance graph to the other equilibrium a_j if the connection Γ_i from a_i to a_j is twisted (nontwisted); (2) For a given pair of entrance graphs near a given equilibrium a_i ordered by $<$, the ordering $(<)$ for the consecutive entrance graphs near the other equilibrium a_j is (not) to be reversed if the connection Γ_i from a_i to a_j is (not) twisted.

To show (6.1(a)), we have $\mathrm{Ex}_2^j \subset \sigma_2^u$, thus $\mathrm{En}_1^{j+1} < \mathrm{En}_1^1$ for all $j \geqq 1$ by (1). By (2), the single twist and $\mathrm{En}_1^j < \mathrm{En}_1^1$ imply $\mathrm{En}_1^2 < \mathrm{En}_1^{j+1}$ for all $j \geqq 2$. Since $\mathrm{Fn}_2^1 < \mathrm{En}_2^j$ for $j \geqq 1$ by (1), we have $\mathrm{En}_1^{j+1} < \mathrm{Fn}_1^2$ by (2) (see Fig. 6.1(a)). To show (6.1(b)), we have $\mathrm{En}_2^j < \mathrm{Fn}_2^1$ and $\mathrm{En}_1^{j+1} < \mathrm{En}_1^1$ for all $j \geqq 1$ by (1). The twist of Γ_2 and $\mathrm{En}_2^j < \mathrm{Fn}_2^1$ imply $\mathrm{Fn}_1^1 < \mathrm{En}_1^{j+1}$ for all $j \geqq 0$ by (2). The double twists and $\mathrm{En}_1^2 < \mathrm{En}_1^1$ imply $\mathrm{En}_1^3 < \mathrm{En}_1^2$ by (2). Last, a simple inductive argument shows $\mathrm{En}_1^{j+1} < \mathrm{En}_1^j$ for all $j \geqq 1$ (see Fig. 6.1(b)). □

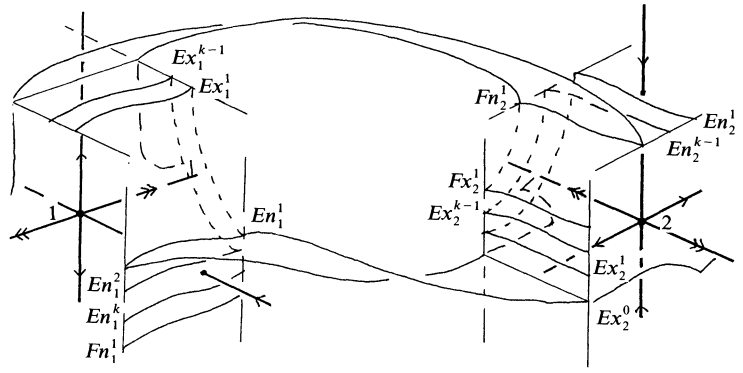
LEMMA 6.2. *If* $\dim W_1^s = \dim W_2^s = 1$ *and* $|\alpha_2| < \alpha_1$, *then the kth exit set* Ex_1^k *from* a_1 *is nonempty if all the previous* $k-1$ *exit sets* Ex_1^j *from* a_1 *are nonempty and the kth extended entrance set* $\tilde{\mathrm{En}}_1^k$ *near* a_1 *has nonempty intersection with the domain* σ_1^s *of the local map* Π_1.

Proof. We first claim that the $(k-1)$st exit set Ex_2^{k-1} from a_2 is nonempty. Since $\mathrm{Ex}_1^{k-1} \neq \varnothing$, the $(k-1)$st entrance set En_2^{k-1} to a_2 is nonempty. By $|DH(\cdot, \alpha)| \leqq \frac{1}{2}$ of Corollary 4.2 and Theorem 3.7, we have $\mathrm{En}_2^{k-1} = \tilde{\mathrm{En}}_2^{k-1} = \mathrm{graph}\,(G(\cdot, \alpha), H(\cdot, \alpha))$. Using the bifurcation equation (4.2c′) we have

$$H(0, \alpha) = [\alpha_1 + O(|\alpha|^{1+\nu_0})]/m_1(\alpha) > 0,$$



(a)



(b)

FIG. 6.1. (a) Γ_2 *is twisted but* Γ_1 *is not twisted*; (b) *Both* Γ_1 *and* Γ_2 *are twisted.*

since α_1 is dominating by our assumption $\alpha_1 > |\alpha_2|$. It follows from $|H(\hat{y}, \alpha)| \leqq \delta_0 s_0/4$ of Corollary 4.2 and Corollary 3.6 that $\tilde{\mathrm{En}}_2^{k-1} \cap \sigma_2^s \neq \emptyset$. Hence,

$$\mathrm{Ex}_2^{k-1} = \Pi_2(\mathrm{En}_2^{k-1} \cap \sigma_2^s) \neq \emptyset.$$

The claim is proved. Obviously, this claim implies the kth entrance set En_1^k to a_1 is nonempty and equal to its extension $\tilde{\mathrm{En}}_1^k$ by Corollary 4.2 and Theorem 3.7. Thus, the condition $\tilde{\mathrm{En}}_1^k \cap \sigma_1^s \neq \emptyset$ implies the kth exit set Ex_1^k near a_1 is nonempty. □

Combining Corollary 3.6, Theorem 3.7, and Corollaries 4.2, 4.3 above we have the following important result.

LEMMA 6.3. *If* $\dim W_1^s = \dim W_2^s = 1$ *and* $p_i = W_{i\,\mathrm{loc}}^s(\alpha) \cap \Sigma_i^s \in \tilde{\mathrm{En}}_i^k$ *at some* $\alpha = \alpha^0 = (\alpha_1^0, \alpha_2^0)$ *with* $k \geqq 1$, *then for the fixed ith component* $\alpha_i = \alpha_i^0$, $\tilde{\mathrm{En}}_i^k \cap \sigma_i^s \neq \emptyset$ *if and only if* $\alpha_j > \alpha_j^0$, *where* $|\alpha| < \delta$. *Moreover, if* En_i^k *is nonempty then the same statement holds true for* En_i^k.

Proof. Let $\tilde{\mathrm{En}}_i^k = \mathrm{graph}\,(H(\cdot, \alpha))$. Corollary 4.3 implies $H(0, \alpha) > 0$ for $\alpha_i = \alpha_i^0$ if and only if $\alpha_j > \alpha_j^0$. Since Corollary 4.2 implies the conditions $|DH| < \frac{1}{2}$ and $|H| \leqq \delta_0 s_0/4$ of Corollary 3.6, we conclude from Corollary 3.6 that $\tilde{\mathrm{En}}_i^k \cap \sigma_i^s \neq \emptyset$ for $\alpha_i = \alpha_i^0$ if and only if $\alpha_j > \alpha_j^0$. By the condition $|DH| < \frac{1}{2}$ of Corollary 4.2 and Theorem 3.7, we have $\mathrm{En}_i^k = \tilde{\mathrm{En}}_i^k$. □

Proof of Theorem B. (i) Corollary 4.2 implies the conditions of Corollary 3.6. The nonemptiness of the kth entrance set En_1^k with $k \geqq 3$ implies $\mathrm{En}_1^2 \cap \sigma_1^s \neq \emptyset$. Hence, by Corollary 3.6, $0 < H(0)$, where $\mathrm{En}_1^2 = \mathrm{graph}\,(H)$, implying the center point $p_1 = W_{1\,\mathrm{loc}}^s(\alpha) \cap \Sigma_1^s$ lies below the second entrance set En_1^2. It follows from (6.1a) that the existence of k-heteroclinic orbits from a_2 to a_1 is impossible for $k \geqq 2$. This together with Theorem A(ii) proves the uniqueness of the 1-heteroclinic orbit from a_2 to a_1.

Let 1-het$_2$ be as in Theorem A. To show 1-het$_2(\alpha_1) > \mathrm{hom}_1(\alpha_1)$ for $\alpha_1 > 0$, we notice that Fn_1^1 is nonempty in the spirit of Corollary 4.4. By (6.1a) we have $\mathrm{En}_1^2 < \mathrm{Fn}_1^1 < \mathrm{En}_1^1$. It follows that with α_2 increasing, the homoclinic connection $p_1 \in \mathrm{Fn}_1^1$ takes place before the single heteroclinic connection $p_1 \in \mathrm{En}_1^2$ does by the transverse crossing property of Corollary 4.3 and Lemma 6.3.

(ii) First we zoom in the region of the parameter where $(k-1)$-heteroclinic orbits (from a_2 to a_1) can take place. We first claim there exists a $(k-1)$-heteroclinic orbit only if $0 \leqq \alpha_2 < \mathrm{hom}_1(\alpha_1)$ with $\alpha_1 > 0$, where hom_1 is the bifurcation curve for the 1-homoclinic orbit at a_1. If $\alpha_1 < 0$ then $\mathrm{En}_2^k < \mathrm{Fn}_2^1 < p_2$ for all $k \geqq 1$ by Corollary 4.4 and the twist of Γ_1. Thus, $\mathrm{En}_2^k \cap \sigma_2^s = \emptyset$ by Corollary 3.6. Hence, $\mathrm{En}_1^{k+1} = \emptyset$ for $k \geqq 1$. If $\alpha_2 > \mathrm{hom}_1(\alpha_1)$ then $p_1 < \mathrm{Fn}_1^1$ by Corollary 4.3. By Lemma 6.1 $p_1 < \mathrm{Fn}_1^1 < \mathrm{En}_1^k$ for all $k \geqq 1$. This proves our claim.

The existence of these heteroclinic bifurcation curves $(k-1)$-het$_2$ in the region $0 \leqq \alpha_2 < \mathrm{hom}_1(\alpha_1)$ is an immediate consequence of the following claim: in $0 \leqq \alpha_2 < \mathrm{hom}_1(\alpha_1)$ there exists a unique $\alpha_2 = (k-1)$-het$_2(\alpha_1)$ for every $k \geqq 1$ such that $p_1 < \mathrm{En}_1^k$ if and only if $\alpha_2 > (k-1)$-het$_2(\alpha_1)$. We proceed by induction. When $k = 1$, it is trivial by the existence of the primary heteroclinic orbits and the transverse crossing property of Corollary 4.4. Suppose the claim holds true for $k-1$. Then, by Lemma 6.3 we have $\mathrm{En}_1^{k-1} \cap \sigma_1^s = \emptyset$ for $\alpha_2 < (k-2)$-het$_2(\alpha_1)$; hence, $(k-1)$-heteroclinic orbits do not exist. Again, by Lemma 6.2 and 6.3 we have $\mathrm{En}_1^k \neq \emptyset$ for $(k-2)$-het$_2(\alpha_1) < \alpha_2 \leqq \mathrm{hom}_1(\alpha_1) < \alpha_1$. On the other hand, Corollary 4.4 implies $\mathrm{Fn}_2^1 \cap \sigma_2^s \neq \emptyset$ and thus $\mathrm{Fn}_1^1 \neq \emptyset$ for these parameters. The transverse crossing properties of Lemma 6.3 and Corollary 3.6 imply $\mathrm{Fn}_1^1 < p_1$ for $\alpha_2 < \mathrm{hom}_1(\alpha_1)$. Lemma 6.1 implies $\mathrm{Fn}_1^1 < \mathrm{En}_1^k$. It follows from the continuity of En_1^1 on the parameter that when $0 < \alpha_2 - (k-2)$-het$_2(\alpha_1) \ll 1$, En_1^{k-1} is sufficiently close to the stable manifold W_1^s, and Ex_1^{k-1} is close to Fx_1^0. Therefore, $\mathrm{Fn}_1^1 < \mathrm{En}_1^k < p_1$ holds. Since $p_1 \in \mathrm{Fn}_1^1 < \mathrm{En}_1^k$ at $\alpha_2 = \mathrm{hom}_1(\alpha_1)$, there must be an α_2^{k-1}

such that $p_1 \in \mathrm{En}_1^k$ at $(\alpha_1, \alpha_2^{k-1})$. Because Lemma 6.3 implies this crossing is transverse, $p_1 < \mathrm{En}_1^k$ if and only if $\alpha_2 > \alpha_2^{k-1}$. Let $\alpha_2^{k-1} := (k-1)\text{-het}_2(\alpha_1)$. This completes the claim. Furthermore, by the property of transverse crossing, the function $(k-1)$-het_2 is also differentiable by the implicit function theorem.

Finally, to show the inaccessibility of the homoclinic bifurcation curve hom_1 from below, we suppose to the contrary that $(k-1)\text{-het}_2(\alpha_1^0) \to \alpha_2^0 < \mathrm{hom}_1(\alpha_1^0)$ for some α_1^0. Then, at $\alpha^0 = (\alpha_1^0, \alpha_2^0)$ $\mathrm{En}_1^k \neq \varnothing$ for all k. Let $\lim_{k\to\infty} \mathrm{En}_1^k = E^\infty$. Then $E^\infty \geqq p_1 > \mathrm{Fn}_1^1$ at α^0. Thus, $p_1 \notin E^\infty$ because otherwise $\mathrm{Fn}_1^1 < \mathrm{En}_1^k < p_1$ for sufficiently large k for En_1^k would be sufficiently close to the stable manifold W_1^s and En_1^{k+1} would be empty. Therefore, $p_1 < E^\infty$. But, in this case, by moving α_2 down a little, i.e., $0 < \alpha_2^0 - \alpha_2 \ll 1$, $p_1 < E^\infty \leqq \mathrm{En}_1^k$ would still hold true for all k and $\alpha = (\alpha_1^0, \alpha_2)$. Thus, there would exist a k such that $\alpha_2 < (k-1)\text{-het}_2(\alpha_1)$. This would imply $\mathrm{En}_1^k < p_1$ by our second claim above. This is a contradiction. □

7. Remarks. (a) It seems that the bifurcation equations (5.6) are solvable for $s_i^{j^*}(\alpha) > 0$ if the twist functions τ_1 and τ_2 are all negative. Incidentally, by neglecting the higher-order terms, the truncated recursive formulas do give rise to a monotone increasing set $s_2^{i^*}$ for $i = 1, \cdots, k-1$ with $s_2^{1^*} = \alpha_1 + \tau_1 \alpha_2^{1+\nu_1} > 0$ and a monotone decreasing set $s_1^{i^*}$ for $i = 1, \cdots, k$ with $s_1^{k^*} := \alpha_2 + \tau_2 |s_2^{k-1^*}|^{1+\nu_2} = 0$. Furthermore, the $(k-1)$-het_2 curve is then defined by the recursive formulas. Unfortunately, this argument fails when those fuzzy error terms are taken into consideration. A similar situation of our losing control over the full system appears in the homoclinic bifurcations with resonant principal eigenvalues (i.e., $\lambda_1(0) = \mu_1(0)$ in our notation) studied by Chow, Deng, and Fiedler (1990). This is the reason we impose the condition $\dim W^s = m = 1$ and approach our problem topologically by considering the entrance and exit sets and their extensions. We feel that this restriction may not be merely technical, since without it the position of Fn_1^1 relative to En_1^k may behave in an unpredictable way.

(b) It can be easily seen from the bifurcation equation (5.1) that the asymptotic tangency of the homoclinic curve hom_i is completely determined by the sign $\lambda_j + \mu_j$ for $i \neq j$. That is, hom_i is asymptotically tangent to the α_i-axis at $\alpha = 0$ for a relatively contractive a_j, or the α_j-axis for a relatively repelling a_j, or tangent to none of them for an a_j with principal resonant eigenvalues, i.e., $\lambda_j + \mu_j = 0$ at $\alpha = 0$. In all cases, however, k-heteroclinic orbits are expected to bifurcate at a twisted heteroclinic loop. In particular, as in the homoclinic doubling bifurcation for a twisted homoclinic orbit (see, e.g., Chow, Deng, and Fiedler (1990)), a double homoclinic bifurcation will also probably take place at a single twisted loop with the resonant eigenvalues, i.e., $(1+\nu_1)(1+\nu_2) = 1$ at $\alpha = 0$, or in the case where the heteroclinic loop is degenerate (see Yanagida (1986)).

Relaxing the equal dimensionality $\dim W_1^s = \dim W_2^s$ assumption will also lead to countable k-heteroclinic connections which, in contrast to Theorem B, take place in an open set of the parameter space. This was observed by Deng (1989). Also, as the principal eigenvalue (either stable or unstable) becomes a pair of complex, the system itself at $\alpha = 0$ becomes rather chaotically complicated (see Tresser (1984) for the case where $\dim W_1^s = \dim W_2^s$, and Bykov (1980) where $\dim W_1^s \neq \dim W_2^s$).

(c) The heteroclinic loop gives us another new bifurcation point which an oriented homoclinic path can hit globally in the parameter space (cf. Fig. 7.1). The orientation of a given homoclinic path is determined by the nonzero orbit index of the periodic orbits nearby. See Mallet-Paret and Yorke (1982), Fiedler (1985), and Chow, Deng, and Fiedler (1990) for more details on the orientation relative to the orbit index. Let us suppose the homoclinic paths hom_1 and hom_2 are oriented as shown in the figure.

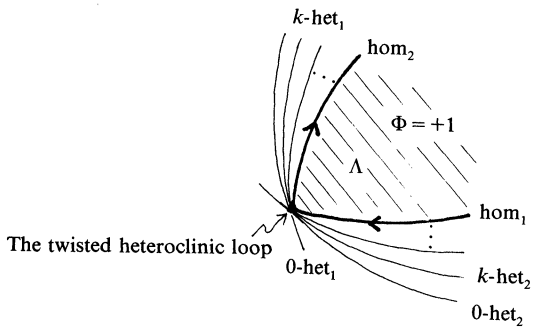


FIG. 7.1. *Φ is the orbit index. If it is assumed to be* 1 *then the region Λ is on the right of* hom$_i$ *curves.*

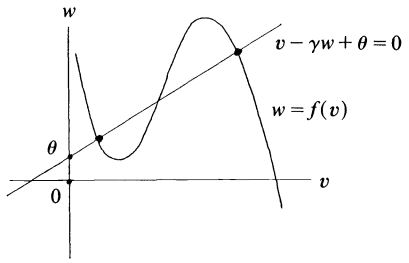


FIG. 7.2

Following hom$_1$, it will hit and terminate at the heteroclinic loop bifurcation point $\alpha = 0$. But, right at this point, the homoclinic orbit to a_1 trades itself to a homoclinic orbit to a_2 and the curve hom$_2$ arises from $\alpha = 0$ as if it is the continuation of hom$_1$. For this reason, we may also call our heteroclinic loop bifurcation the homoclinic trading bifurcation. However, if we "follow" (actually we do not know how at this moment) a heteroclinic path in the double twisted case, we will find doubly infinite heteroclinic trading partners at the homoclinic trading place $\alpha = 0$. Certainly, this immediately complicates any "global heteroclinic path following" attempt. But it also gives us one more hope that a global homoclinic path following result seems on its way (see a detailed discussion from Chow, Deng, and Fiedler (1990)).

(d) While writing these remarks, I received a preprint by Kokubu, Nishiura, and Oka (1988). I found that our notion of twisted heteroclinic loop has been propagating faster than I could finish writing this paper. Their work demonstrates that the non-degenerate conditions (1.4), (1.5), (1.7), and (1.10a, b) are verifiable. Indeed, motivated by the idea for the Mel'nikov function and the method of singular limit eigenvalue problem developed by Nishiura and Fujii (1987) and Nishiura (1989), they derive not only an analogous function to detect the transverse crossing of the stable and unstable manifolds (also see, e.g., Kokubu (1988)), but also a computable twist function to detect the strong inclination property and the twist of a given heteroclinic orbit at the same time for the system of ODE for the traveling waves, $(v, w)(x, t) = (v, w)(x + ct)$. The reaction diffusion system they consider is as follows:

$$\begin{aligned} \varepsilon\tau v_t &= \varepsilon^2 v_{xx} + f(v) - w, \\ w_t &= w_{xx} + v - \gamma w + \theta, \end{aligned} \tag{7.1}$$

where $f = -v^3 + v$. Starting at a standing front wave and a standing back wave which forms a heteroclinic loop (i.e., at $c = 0$), they manage to obtain the local codimension-three bifurcation unfoldings with c, θ, γ being the relevant parameters and globally

extend the local bifurcation diagram. Among the most interesting is the nontwistedness of all the resulting heteroclinic loops involved due to various symmetries exhibited by the system. According to Theorem A, this implies that there are no multiple heteroclinic connections other than the persistence of the zero-heteroclinic orbits from the loops. They also show that some of the finite connections are actually unstable.

Perhaps some comparisons between their system and the FitzHugh-Nagumo equation considered in the Introduction are worthwhile. First of all they model systems of different worlds—chemical reactions, predator and prey populations for the former while nerve impulses for the latter. Theoretically speaking, however, they are the same system but at different values of the diffusion parameter for the w dynamics. Indeed, if we move the origin to the left equilibrium state in Fig. 7.2, the parameter θ is the same as the parameter a in the FitzHugh-Nagumo equation (1.11). Rescaling the time and the space variables in (7.1) yields

$$v_t = v_{xx} + f(v) - w, \qquad w_t = \delta w_{xx} + \varepsilon(v - \gamma w),$$

where we renamed $\delta := \tau/\varepsilon$ and $\varepsilon := \varepsilon\tau$. Thus, it is the same FitzHugh-Nagumo equation except for a large diffusion coefficient δ for w. Since both systems have the same symmetries, I think the appearance of the second diffusion simply "untwists" the twisting structure somewhere. Thus, it is natural to ask whether it happens at some $\delta_0 > 0$ or just at $\delta_0 = 0$. Indeed, there are two types of bifurcations involved. When $\delta_0 = 0$ the system of the ODE is singularly perturbed. When $\delta_0 > 0$, however, to untwist a heteroclinic loop a heteroclinic orbit must be degenerate in general and, in particular, it must violate the strong inclination property. None of these bifurcations problems has ever been fully investigated. Nevertheless, the idea developed in this paper offers more hope for solving the bifurcation of twists than the problem of singular perturbation.

Appendix.

Proof of Lemma 4.1. The proof for the first half part of the lemma is identical to the proof of Lemma 3.3 in § 3. Thus, we omit it here. Using the notation from § 2, we write equations (4.1a-4.1d) in the following equivalent forms:

$$\Phi(\zeta, z, \alpha) = 0,$$

where $\Phi = \Phi_\iota$ with the subindex $\iota = a, b, c, d$ in correspondence with the alphabets in the equation numbers. Here $\zeta = (\eta_2, \xi_1)$ and z varies with equations as follows.

$$\Phi_a(\zeta, z, \alpha) = \begin{pmatrix} P_2(0, \eta_2, \alpha) - \xi_1 \\ Q_2(0, \eta_2, \alpha) - y_1 \end{pmatrix}, \qquad z = y_1,$$

$$\Phi_b(\zeta, z, \alpha) = \begin{pmatrix} P_2(0, \eta_2, \alpha) - \xi_1 \\ Q_2(0, \eta_2, \alpha) - Y_1 \end{pmatrix}, \qquad z = (s_1, \eta_1),$$

$$\Phi_c(\zeta, z, \alpha) = \begin{pmatrix} P_2(X_2, \eta_2, \alpha) - \xi_1 \\ Q_2(X_2, \eta_2, \alpha) - y_1 \end{pmatrix}, \qquad z = (s_2, \xi_2, y_1),$$

$$\Phi_d(\zeta, z, \alpha) = \begin{pmatrix} P_2(X_2, \eta_2, \alpha) - \xi_1 \\ Q_2(X_2, \eta_2, \alpha) - Y_1 \end{pmatrix}, \qquad z = (s_2, \xi_2, s_1, \eta_1),$$

where $X_2 = X_2(s_2, \xi_2, \eta_2, \alpha)$ and $Y_1 = Y_1(s_1, \xi_1, \eta_1, \alpha)$. Delete the mth component of Φ and let $\hat{\Phi} = (\Phi^{(1)}, \cdots, \Phi^{(m-1)}, \Phi^{(m+1)}, \cdots, \Phi^{(d-1)})^T$. We solve $\hat{\Phi} = 0$ first by the implicit function theorem for $\zeta = \zeta^*(z, \alpha)$ and then solve the reduced remaining equation $\Phi^{(m)}(\zeta^*, z, \alpha) = 0$ later.

Because of the existence of the heteroclinic loop we have $\hat{\Phi}(0, 0, 0) = 0$. Moreover, the square Jacobian $\partial\hat{\Phi}/\partial\zeta(0, 0, 0) = \hat{M}_2$ is nonsingular by Lemma 2.3. (Note that this

also implies $\partial\hat{Q}_2/\partial\eta(0,0,0)$ nonsingular.) Hence, by the implicit function theorem there exists a differentiable function ζ^* of $|z|, |\alpha| < \delta$ satisfying $\zeta^*(0,0)=0$ and $|\zeta^*| < \delta$ such that $\hat{\Phi}(\zeta, z, \alpha)=0$ if and only if $\zeta=\zeta^*(z,\alpha)$. To solve the reduced equation $\Phi^{(m)}(\zeta^*, z, \alpha)=0$, we need some important facts about ζ^*. Since ζ^* usually are not the same for different indexes $\iota = a, b, c,$ or d, we denote it by ζ_ι^*, accordingly. Since $\hat{\Phi}_a$ and $\hat{\Phi}_c$ do not depend on the first y-component $y_1^{(1)}$, ζ_a^* and ζ_c^* are functions of $\hat{y}_1$ and α only. Moreover, since $X_2(0,\xi_2,\eta_2,\alpha)=0$ and $Y_1(0,\xi_1,\eta_1,\alpha)=0$ by the property (2.1a) of exponential expansions, it is easy to see that when equations $\hat{\Phi}_\iota=0$ for all ι are restricted to $s_1=0$, $s_2=0$, and $\hat{y}_1=0$ (whichever applies) they are all reduced to the same equations as follows:

$$\text{(A.0)} \qquad \begin{cases} P_2(0,\eta_2,\alpha)-\xi_1=0, \\ Q_2(0,\eta_2,\alpha)=0, \end{cases} \qquad \zeta=(\eta_2,\xi_1).$$

Thus, the solution ζ depends only on α. Therefore, the following functions of restrictions $\zeta_a^*|_{\hat{y}_1=0}$, $\zeta_b^*|_{s_1=0}$, $\zeta_c^*|_{s_2=0,\hat{y}_1=0}$, and $\zeta_d^*|_{s_1=s_2=0}$ are in fact equal to the same function of α, say $(u,v)(\alpha)$, which is the solution to (A.0). Note also that $|\zeta_\iota^*-(u,v)^T|=O(|\hat{y}_1|+|s_2|)$ or $O(|s_1|+|s_2|)$, accordingly.

As another preparation, we need the following procedures one way or another. Expand $(P_2, Q_2)(x,\eta,\alpha)$ at $(x,\eta)=(0,u)$,

$$\text{(A.1)} \qquad \begin{pmatrix} P_2 \\ Q_2 \end{pmatrix}(x,\eta,\alpha)=\begin{pmatrix} P_{2\alpha} \\ Q_{2\alpha} \end{pmatrix}+\frac{\partial(P_2,Q_2)}{\partial(x,\eta)}(0,u,\alpha)\begin{pmatrix} x \\ \eta-u \end{pmatrix}+O(|x|^2+|\eta-u|^2),$$

where $(P_{2\alpha}, Q_{2\alpha})=(P_2,Q_2)(0,u,\alpha)$. Expand $\varphi(\xi,\eta,\alpha)$ at $(\xi,\eta)=(0,u)$ and $\psi(\xi,\eta,\alpha)$ at $(\xi,\eta)=(v,0)$, respectively:

$$\text{(A.2)} \qquad X(s_2,\xi_2,\eta_2,\alpha)=\varphi_{2\alpha}s_2^{1+\nu_2}+O[(|\xi_2|+|\eta_2-u|)|s_2|^{1+\nu_2}+|s_2|^{1+\nu_2+\bar{\nu}_2}],$$

$$\text{(A.3)} \qquad Y(s_1,\xi_1,\eta_1,\alpha)=\psi_{1\alpha}s_1+O[(|\xi_1-v|+|\eta_1|)|s_1|+|s_1|^{1+\bar{\nu}_1}],$$

where $\varphi_{2\alpha}=\varphi(0,u,\alpha)$ and $\psi_{1\alpha}=\psi(v,0,\alpha)$. Let

$$L_2(\alpha)=\begin{bmatrix} \dfrac{\partial P_2}{\partial\eta}(0,u,\alpha) & -I \\ \dfrac{\partial Q_2}{\partial\eta}(0,u,\alpha) & 0 \end{bmatrix}_{(d-1)\times(d-2)}$$

and

$$\hat{M}_2(\alpha)=\begin{bmatrix} \dfrac{\partial P_2}{\partial\eta}(0,u,\alpha) & -I \\ \dfrac{\partial \hat{Q}_2}{\partial\eta}(0,u,\alpha) & 0 \end{bmatrix}_{(d-2)\times(d-2)}.$$

Then, by the continuity of u on α and Lemma 2.3 we may assume that, without loss of generality, $L_2(\alpha)$ has the maximal rank $d-2$ which is attained by the submatrix $\hat{M}_2(\alpha)$. Note that this also implies $\partial\hat{Q}_2/\partial\eta(0,u,\alpha)$ nonsingular. Also, up to $m-1$ permutations we may still call

$$M_2(\alpha)=\left[L_2(\alpha),\begin{pmatrix} 0 \\ \psi_{1\alpha} \end{pmatrix}\right]_{(d-1)\times(d-1)}=\begin{bmatrix} \dfrac{\partial P_2}{\partial\eta}(0,u,\alpha) & -I & 0 \\ \dfrac{\partial Q_2}{\partial\eta}(0,u,\alpha) & 0 & \psi_{1\alpha} \end{bmatrix}$$

and we have $|\det M_2(\alpha)| > m_0\delta_0$ for $|\alpha|<\delta$ by Lemma 2.3 and Lemma 2.4.

Now we are ready to solve $\Phi^{(m)}(\zeta^*, z, \alpha) = 0$. Since $\hat{\Phi}(\zeta^*, z, \alpha) \equiv 0$, $\Phi(\zeta^*, z, \alpha) = (0, \cdots, \Phi^{(m)}, 0, \cdots, 0)(\zeta^*, z, \alpha)$. This implies

$$\det [L_2(\alpha), \Phi(\zeta^*, z, \alpha)] = (-1)^{(d-1)+m} \det \hat{M}_2(\alpha) \Phi^{(m)}(\zeta^*, z, \alpha).$$

Hence, $\Phi^{(m)}(\zeta^*, z, \alpha) = 0$ is equivalent to

$$\det [L_2(\alpha), \Phi(\zeta^*, z, \alpha)] = 0. \tag{A.4}$$

Since the simplifications for these equations are all identical we will only treat two typical cases $\iota = a$ and $\iota = c$ here, with emphasis on how the nonsingular change of parameters and the functions m_i and τ_i are obtained for all the bifurcation equations.

When $\iota = a$, substituting $(x, \eta) = (0, \eta_2^*)$ into the Taylor expansion (A.1) and using $|\eta_2^* - u| = O(|\hat{y}_1|)$, we have

$$\Phi_a(\xi^*, z, \alpha) = \begin{pmatrix} P_{2\alpha} \\ Q_{2\alpha} \end{pmatrix} + \frac{\partial (P_2, Q_2)}{\partial \eta} (0, u, \alpha)(\eta_2^* - u) + O(|\hat{y}_1|^2) - \begin{pmatrix} \xi_1^* \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ y_1 \end{pmatrix}.$$

Since the second and the fourth terms all belong to the range of $L_2(\alpha)$, they will disappear in (A.4). This implies

$$\det \left[L_2(\alpha), \begin{pmatrix} 0 \\ y_1 \end{pmatrix} \right] = \det \left[L_2, \begin{pmatrix} P_{2\alpha} \\ Q_{2\alpha} \end{pmatrix} \right] + O(|\hat{y}_1|^2).$$

Dividing this equation by $\det M_2(\alpha)$ and expressing the left-hand side in terms of a homogeneous linear combination in $y_1^{(i)}$ we have

$$m_1(\alpha) y_1^{(1)} + \cdots m_n(\alpha) y_1^{(n)} = c_2(\alpha_1, \alpha_2) + O(|\hat{y}_1|^2),$$

where

$$m_i = \det \left[L_2(\alpha), \begin{pmatrix} 0 \\ e_i \end{pmatrix} \right] \Big/ \det M_2(\alpha),$$

and

$$c_2 = \det \left[L_2(\alpha), \begin{pmatrix} P_{2\alpha} \\ Q_{2\alpha} \end{pmatrix} \right] \Big/ \det M_2(\alpha), \tag{A.5}$$

and $e_i \in \mathbb{R}^n$ has zero components except for the ith component of 1. This has the form of (4.2a). Let us show that the functions m_i satisfy the required properties and postpone the discussion of c_2 until later.

Since $(0, \psi_j(0, 0))/\delta_0 \to (0, e_1)$ as $\delta_0 \to 0$ by the exponential expansion property (2.1b), $\det M_2(\alpha)$ is approximately the product of δ_0 and the numerator for m_1 as $\delta_0 \to 0$. Hence, for small but fixed δ_0 we have $1/2\delta_0 < m_1 < 2/\delta_0$. Also, it follows from Lemma 2.4 that

$$\frac{m_i}{m_1} = \det \left[L_2(\alpha), \begin{pmatrix} 0 \\ e_i \end{pmatrix} \right] \Big/ \det \left[L_2(\alpha), \begin{pmatrix} 0 \\ e_1 \end{pmatrix} \right] = o(1) \quad \text{as } \delta_0 \to 0.$$

Before we check the properties for c_2, let us first obtain the bifurcation equation (4.2d). Substitute $(x, \eta) = (X_2^*, \eta_2^*)$ with $X_2^* = X(s_2, \xi_2, \eta_2^*, \alpha)$ into (A.1) and use $|X_2^*| = O(|s_2|^{1+\nu_2})$ and $|\eta_2^* - u| = O(|s_1| + |s_2|)$. Then substitute the exponential expansion (A.2) for X_2^*. Finally, substitute the obtained (A.1) and (A.3) with $\xi_1 = \xi_1^*$ into the

function $\Phi_d(\zeta^*, z, \alpha)$. We have

$$\Phi_d(\zeta^*, z, \alpha) = \begin{pmatrix} P_{2\alpha} \\ Q_{2\alpha} \end{pmatrix} + \frac{\partial(P_2, Q_2)}{\partial x}(0, u, \alpha)\varphi_{2\alpha} s_2^{1+\nu_2} + \frac{\partial(P_2, Q_2)}{\partial \eta}(0, u, \alpha)(\eta_2^* - u) - \begin{pmatrix} 0 \\ \psi_{1\alpha} \end{pmatrix} s_1 + O((|\xi_2| + |s_2|^{\bar{\nu}_2})|s_2|^{1+\nu_2} + (|\eta_1| + |s_1|^{\bar{\nu}_1})|s_1|).$$

Similarly, the third and fourth terms belong to the range of $L_2(\alpha)$; hence, they disappear in equation (A.4). This yields

$$\det\left[L_2(\alpha), \begin{pmatrix} 0 \\ \psi_{1\alpha} \end{pmatrix}\right] s_1 = \det\left[L_2(\alpha), \begin{pmatrix} P_{2\alpha} \\ Q_{2\alpha} \end{pmatrix}\right] + \det\left[L_2(\alpha), \frac{\partial(P_2, Q_2)}{\partial x}(0, u, \alpha)\varphi_{2\alpha}\right] s_2^{1+\nu_2} + (\text{the same form of higher order}).$$

Dividing both sides by $\det M_2(\alpha) = \det[L_2(\alpha), \binom{0}{\psi_{1\alpha}}]$, we obtain the desired form

$$s_1 = c_2 + \tau_2 s_2^{1+\nu_2} + (\text{the same form of higher order}),$$

where the function c_2 of α is the same as (A.5), and

$$\tau_2 = \det\left[L_2(\alpha), \frac{\partial(P_2, Q_2)}{\partial x}(0, u, \alpha)\varphi_{2\alpha}\right] \Big/ \det M_2(\alpha).$$

Now we show $c_2(\alpha_1, 0) = 0$ and $\partial c_2/\partial\alpha_2(0, 0) \neq 0$ and (4.3). Recall that $\hat{Q}_{2\alpha} \equiv 0$. Thus, from (A.5) we obtain

$$c_2(\alpha_1, \alpha_2) = (-1)^{(m-1)m} \frac{\det \partial\hat{Q}_2/\partial\eta(0, u, \alpha)}{\det M_2(\alpha)} Q_{2\alpha}^{(1)}.$$

Because $\partial\hat{Q}_2/\partial\eta(0, u, \alpha)$ is nonsingular it suffices to show $Q_{2\alpha}^{(1)} = 0$ when $\alpha = (\alpha_1, 0)$ and $\partial Q_{2\alpha}^{(1)}/\partial\alpha_2 \neq 0$ at $\alpha = (0, 0)$ by the product rule of differentiation. It is trivial to check $Q_{2\alpha}^{(1)} = 0$ at $\alpha = (\alpha_1, 0)$ because of the existence of the primary heteroclinic connections from a_2 to a_1 on the α_1-axis. Also, since $(P_{2\alpha}, Q_{2\alpha})^T = (P_2, Q_2)^T(0, u, \alpha)$ is on the unstable manifold $W_2^u(\alpha) \cap \Sigma_1^s$ for $(0, u) \in W_{2\text{loc}}^u(\alpha) \cap \Sigma_2^u$, by (1.10b) for the distance between W_2^u and W_1^s on Σ_1^s and $\hat{Q}_{2\alpha} \equiv 0$ we have

$$0 \leqq d_2(\alpha_1, \alpha_2) \leqq \min_{(\xi, 0) \in W_{1\text{loc}}^s} |(P_{2\alpha}, Q_{2\alpha}) - (\xi, 0)| = |Q_{2\alpha}^{(1)}|.$$

This implies $Q_{2\alpha}^{(1)}$ at $\alpha = (0, \alpha_2)$ has a constant sign for $\alpha_2 > 0$, say > 0, since $0 < d_2(0, \alpha_2)$ by our assumptions. Therefore,

$$0 < \frac{d_2(0, \alpha_2)}{\alpha_2} \leqq \frac{Q_{2\alpha}^{(1)}}{\alpha_2} \quad \text{for } \alpha = (0, \alpha_2).$$

Passing the limit $\alpha_2 \to 0^+$ above, we have $\partial Q_{2\alpha}^{(1)}/\partial\alpha_2 > 0$ at $\alpha = (0, 0)$ by (1.10b). This completes the proof for c_2.

To show (4.3), we notice first that $u(0) = 0$ and the set of all the column vectors of $M_2(0)$ forms a base for $T_{p_1}\Sigma_1^s$ and

$$\frac{\partial(P_2, Q_2)}{\partial x}(0, 0, 0)\varphi_{20} = \frac{\partial(P_2, Q_2)}{\partial(x, \eta)}(0, 0, 0)\begin{pmatrix} \varphi_{20} \\ 0 \end{pmatrix}.$$

Project this vector onto the one-dimensional linear space span $\{(0, \psi_{10})^T\}$, which is complementary to the range of $L_2(0)$; namely, span $\{T_{p_1}(W_2^u \cap \Sigma_1^s), T_{p_1}(W_{1\text{loc}}^s \cap \Sigma_1^s)\}$. We obtain $\partial(P_2, Q_2)/\partial x(0,0,0)\varphi_{20} = \tilde{\tau}_2(0, \psi_{10})^T + h$ with $h \in \text{range } L_2(0)$. Hence,

$$\det[L_2(0), \partial(P_2, Q_2)/\partial x(0,0,0)\varphi_{20}] = \tilde{\tau}_2 \det M_2(0)$$

and $\tilde{\tau}_2 = \tau_2(0)$ follows. Of course, $\tau_2(0) > 0$ if and only if Γ_2 is not twisted by our Definition 1.1. This completes the proof. □

Acknowledgments. The author is indebted to the reviewers for their many useful suggestions. He also has benefited from many conversations with S. N. Chow, J. K. Hale, and J. Mallet-Paret. Special thanks go to D. Terman, who corrected the author's misunderstanding of his work with J. Rinzel, which was the key motivation for beginning this work.

REFERENCES

V. V. BYKOV, *Bifurcation of dynamical systems close to systems with a separatrix contour containing a saddle-focus, Methods of qualitative theory of differential euqations*, Gor'kov, Gos. University, Gorki, Soviet Union, 1980, pp. 44–72.

S.-N. CHOW, B. DENG, AND B. FIEDLER, *Homoclinic bifurcations at resonant eigenvalues*, J. Dynamical Systems and Differential Equations, 2 (1990), pp. 177–244.

S.-N. CHOW, B. DENG, AND D. TERMAN, *The bifurcation of homoclinic and periodic orbits from two heteroclinic orbits*, SIAM J. Math. Anal., 21 (1990), pp. 179–204.

B. DENG, *The Sil'nikov problem, exponential expansion, strong λ-lemma, C^1-linearization and homoclinic bifurcation*, J. Differential Equations, 79 (1989), pp. 189–231.

———, *Exponential expansion with principal eigenvalues*, preprint, 1988.

J. W. EVANS, *Nerve axon equations. III. stability of the nerve impulse*, Indiana Univ. Math. J., 22 (1972), pp. 577–594.

J. W. EVANS, N. FENICHEL, AND J. A. FEROE, *Double impulse solutions in nerve axon equations*, SIAM J. Appl. Math., 42 (1982), pp. 219–234.

B. FIEDLER, *An index for global Hopf bifurcation in parabolic systems*, J. Reine Angew. Math., 359 (1985), pp. 1–36.

S. P. HASTINGS, *Single and multiple pulse waves for the FitzHugh–Nagumo equations*, SIAM J. Appl. Math., 42 (1982), pp. 247–260.

M. W. HIRSCH, *Differential Topology*, Springer-Verlag, New York, 1976.

K. R. T. JONES, *Stability of the traveling wave solution of the FitzHugh–Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 439–469.

H. KOKUBU, *Homoclinic and heteroclinic bifurcation of vector fields*, Japan J. Appl. Math., 5 (1988), pp. 455–501.

H. KOKUBU, Y. NISHIURA, AND H. OKA, *Heteroclinic and homoclinic bifurcation in bistable reaction diffusion systems*, preprint KSU/ICS 88-08, 1988.

J. MALLET-PARET AND J. A. YORKE, *Snakes, oriented families of periodic orbits, their sources, sinks, and continuation*, J. Differential Equations 43 (1982), pp. 419–450.

Y. NISHIURA, *Singular limit approach to stability and bifurcation for bistable reaction diffusion systems*, to appear in *Proc. Workshop on Nonlinear* PDE's, *March* 1987, *Provo, Utah*, P. Bates and P. Fife, eds., Springer-Verlag, New York, 1989.

Y. NISHIURA AND H. FUJII, *Stability of singularly perturbed solutions to systems of reaction-diffusion equations*, SIAM J. Math. Anal. 18 (1987), pp. 1726–1770.

J. RINZEL AND D. TERMAN, *Propagation phenomena in a bistable reaction-diffusion system*, SIAM J. Appl. Math., 42 (1982), pp. 1111–1137.

M. SHUB, *Global Stability of Dynamical Systems*, Springer-Verlag, New York, 1987.

L. P. SIL'NIKOV, *The existence of a countable set of periodic motions in the neighborhood of a homoclinic curve*, Soviet Math. Dokl., 8 (1967), pp. 102–106.

C. TRESSER, *About some theorems by L. P. Sil'nikov*, Ann. Inst. H. Poincaré, 40 (1984), pp. 440–461.

E. YANAGIDA, *Branching of double pulse solutions from single pulse solutions in nerve axon equations*, J. Differential Equations, 66 (1986), pp. 243–262.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 680–716, May 1991

007

LAYERED VELOCITY INVERSION: A MODEL PROBLEM FROM REFLECTION SEISMOLOGY*

WILLIAM W. SYMES†

Abstract. A simple model problem in exploration seismology requires that a depth-varying sound velocity distribution be estimated from reflected sound waves. For various physical reasons, these reflected signals or echoes have very small Fourier coefficients at both very high and very low frequencies. Nonetheless, both geophysical practice, based on heuristic considerations, and recent numerical evidence indicate that a spectrally complete estimate of the velocity distribution is often achievable. We prove a theorem to this effect, showing that "sufficiently rough" velocity distributions may be recovered from reflected waves under some restrictions, independently of the very low- or high-frequency content of the data. The main restriction is that the velocity depend only on a single (depth) variable; only in this case are sufficiently refined propagation-of-singularity results available. The proof is based on a novel variational principle, from which numerical algorithms have been derived. These algorithms have been implemented and used to estimate velocity distributions from both synthetic and field reflection seismograms.

Key words. inverse problems, hyperbolic partial differential equations, sound velocity, reflection seismology

AMS(MOS) subject classification. 35R25

1. Introduction.

A simple model of the physical setting for reflection seismology is constant-density *linear acoustics*, in which the sound velocity field $c(x)$ ($x \in \mathbf{R}^3$) is connected to the pressure field $u(x,t)$ via the wave equation

$$\frac{1}{c^2(x)} \frac{\partial^2 u(x,t)}{\partial t^2} - \Delta u(x,t) = f(t)\delta(x),$$
$$u \equiv 0, \qquad t < 0.$$

The right-hand side represents an isotropic point dilatational energy source radiating with time-varying (transient) intensity $f(t)$ ("the *source wavelet*"). The *seismogram* is a sampling of the pressure u at a number of "receiver" points. We adopt the idealization that these points form the continuum $\{x_3 =: z = 0\}$ ("the surface (of the earth)") and that the measurement of u is also continuous in time for some time interval $0 \le t \le t_{\max}$. Regarding the source (i.e., $f(t)$) as known, the pressure field, hence the seismogram, becomes a function of the sound velocity:

$$s[c] := u|_{z=0}.$$

In this simple model, the fundamental problem of reflection seismology is to estimate c from $s[c]$, i.e., to solve a functional equation of the form

$$s[c] = s_{\text{data}}$$

possibly in some least-error sense accommodating the possibility (virtual certainty!) of inconsistent data error.

This model is grossly inadequate for some practical purposes, as it ignores significant physics of seismic wave generation and propagation. Nonetheless, it forms

* Received by the editors February 21, 1989; accepted for publication (in revised form) April 20, 1990. This work was partially supported by National Science Foundation grant DMS 8603614 and by Office of Naval Research contract N00014-85-0725.

† Department of Mathematical Sciences, Rice University, Houston, Texas 77251.

the basis for most contemporary seismic data processing (see, for instance, Yilmaz (1987)), and it exhibits two fundamental features of the "real" problem:

(i) $s[c]$ is nonlinear in c;

(ii) f should be chosen to suppress Fourier components in $s[c]$ at "low" and "high" frequencies.

Item (i) is simply the nonlinearity of solutions of linear equations as functions of their coefficients. Item (ii) is required by observations of the spectra of reflection seismograms: for various physical reasons, Fourier components at very low (<4Hz) and very high (> 80 Hz) temporal frequencies are essentially missing from real reflection seismograms.

The suppression of high-frequency components simply means that $c \mapsto s[c]$ is a smoothing operator. Techniques for management of the resulting high-frequency instability are well known (see, e.g., Tikhonov and Arsenin (1977), Miller (1970), Payne (1975)).

In contrast, the instability resulting from the lack of low-frequency data has been little discussed in the mathematical literature on inverse problems, even though it is nearly ubiquitous in real-world parameter-estimation problems based on wave propagation. This *low-frequency lacuna* is a striking feature of reflection seismology, in particular, and the possible ambiguities resulting from spectral incompleteness of data have sparked considerable discussion within the geophysical research community.

The present paper is devoted to the proof of a uniqueness and continuous dependence result for a restricted version of the inverse problem described above, in which the estimates are independent of the behaviour of $\hat{f}(\omega)$ near $\omega = 0$. We shall show that c is well determined by $s[c]$ when c is *sufficiently nonsmooth.*

This rather strange sounding requirement is natural in view of the application to reflection seismology: rapid changes in the mechanical properties of rock are entirely responsible for the return of substantial echoes to the surface, hence for the information content of seismic reflection data. The times of arrival of these echoes—or rather, the signals which simulate them in the model described above—carry information about the slowly varying components of c, completely independently of the low-frequency behavior of f. On the other hand, the identifiability of these echoes depends on the wave nature of the seismic disturbance, in other words on the propagation of singularity (or regularity), according to geometric optics. This propagation property clearly requires some smoothness of the coefficients; thus the conditions necessary for the creation of strong reflected signals are in tension with those necessary for their propagation.

We show in this paper that this tension can be resolved at least in a special case. The resolution requires certain estimates concerning propagation of regularity currently available only under the additional constraint:

> The sound velocity c is a function only of the "depth" variable $x_3(= z)$.

That is, the result detailed here applies to *layered* constant density fluid models only. The necessary technical results for this class of models were established in previous papers of Symes ((1981), (1983), (1985), (1986a), (1986b)). These are essentially energy estimates and are related to earlier work of Rauch and Taylor (1974) and Kreiss (1970) on mixed problems for linear hyperbolic systems in two independent variables. Other authors basing results about one-dimensional inverse problems for hyperbolic equations on the same ideas include Fawcett (1984) and Suzuki (1988).

As mentioned above the constant-density fluid model is severely oversimplified. For example, in reality the density of sedimentary rock is also variable, though much

less dramatically so than the compressional wave velocity. Also, density and compressional velocity are somewhat correlated in real rocks. Nonetheless the variation of density must be taken into account in explaining the details of real reflection seismograms. A natural extension of the problem considered here is therefore the determination of density and velocity profiles in a layered acoustic model. While treatment of this problem goes beyond the intention of the present paper, several conjectures seem plausible. First, based on the treatment of the linearized acoustic problem in Santosa and Symes (1988), we expect to recover in addition to the velocity, the oscillatory or rough part of the density profile (technically, we expect a Gårding-type estimate for small perturbations in density, similar to that stated in Theorem 4.1 below for the velocity). Second, we do *not* expect to gain control over the slowly varying part of the density, since the mechanism active in determination of the slowly varying part of the velocity is *inactive* for the density (that is, we do *not* expect analogues of Theorem 2.1 on Theorem 6.1 to hold for the density). The only hope for density trends appears to be that the above-mentioned correlation with velocity is accurate on long length scales. Third, we expect the inclusion of variable density to be quite important in accurate estimation of velocities. This last conclusion is based on the experience of Symes and others in velocity estimation from field seismograms.

More complicated (and so realistic) layered models (elastic, anelastic, ...) can doubtless be studied from the present point of view, but it is difficult to foresee the results. Of course the most interesting question is the possibility of extension of our results to *nonlayered* models—deviations from layering being the most dramatic way in which the real earth differs from our simple model. A few remarks concerning the prospects for weakening the layered medium assumption in these arguments may be found in the next section.

The paper is organized as follows. Section 2 gives a precise statement of the main results, a brief review of the literature, and discussion of related issues. Section 3 introduces the *plane-wave decomposition* and estimates for the plane-wave problems from Symes's previous work; this material forms the technical basis for the rest of the paper. As we are concerned mostly with the information independent of the low-frequency behaviour of f, we introduce in §4 the (temporary) assumption that f is a compactly supported measure, defining an elliptic convolution operator bounded on $L^2(\mathbf{R})$. Under this assumption, we prove an estimate of Gårding type for the derivative of the plane-wave seismogram map. In order to do more, it is necessary to consider the various plane-wave problems simultaneously. Each plane-wave *model* is parameterized by the vertical plane-wave velocity, viewed as a function of (its own) travel-time. These models are a priori independent. We derive a *coherency condition* in §5, equivalent to the existence of a (single) velocity profile $c(z)$ from which all of the plane-wave models are derived (i.e., "every (plane-wave) experiment sees the same earth"). In §6 we show that a least-squares version of the inverse problem stated above, posed in terms of the plane-wave models and augmented by the coherency condition (as a so-called penalty term) has a positive-definite Hessian (second derivative) operator at a consistent (zero-residual) solution provided that the corresponding velocity profile is sufficiently rough, as stated above. Our main result follows immediately via the implicit function theorem. So far we have maintained the elliptic assumption concerning f; this is dropped in §7, for the usual price, paid for the solution of compact operator equations, of a priori constraints on the smoothness of c.

2. Statement of main result and discussion. In this section we give a precise statement of the major result of this paper, followed by a brief review of the literature and a conceptual overview of the problem described above.

Since the principal goal of the present work is the production of a "solution" of the inverse problem stated above with continuity properties independent of the *low-frequency* behavior of the source wavelet f, we make the temporary assumption that:

> $f \in \mathcal{E}'(\mathbf{R})$ is a Borel measure, defining a bounded elliptic convolution operator on $H^1(\mathbf{R})$: i.e., for positive K_0, K_1, K^*, $\phi \in H^1(\mathbf{R})$

$$K_1 \left\| \frac{\partial \phi}{\partial t} \right\|_{L^2(\mathbf{R})} - K_0 \|\phi\|_{L^2(\mathbf{R})} \leq \left\| f * \frac{\partial \phi}{\partial t} \right\|_{L^2(\mathbf{R})} \leq K^* \left\| \frac{\partial \phi}{\partial t} \right\|_{L^2(\mathbf{R})}. \tag{2.1}$$

> Also, $\operatorname{supp} f \subset \{t \in \mathbf{R} : t \geq 0\}$.

Such a distribution necessarily has a finite *first moment*

$$m_f^1 = \sup_{\phi \in C^\infty} \frac{\langle f, t\phi \rangle}{\|\phi\|_{L^\infty}}.$$

The "elliptic" assumption (2.1) will be weakened, to the extent possible, in §7. Simple examples of wavelets (i.e., kernels) f having the elliptic property (2.1) are obtained by subtracting from a slightly shifted Dirac delta function a smooth approximation.

The conventional, though perhaps ill-founded, choice of measure for the seismogram error is some weighted version of the L^2-norm (see Tarantola (1987, Chap. 6)). We shall adopt this choice also. The elliptic nature of f precludes square-integrability of $s[c]$, however, as can easily be seen, so we choose what amounts to a very singular "weight": we define, for suitably small *slowness* p,

$$S[c](t,p) = \int\int dx_1 dx_2 \frac{\partial}{\partial t} s[c](x_1, x_2, t + px_1).$$

That is, S is a version of the Radon transform of s, which we shall further restrict to a rectangle $\{(t,p)\colon 0 < t \leq T_1, P_1 \leq p \leq P_2\} =: R_1$. It is easily seen that, for smooth c, any $T_1 > 0$, and P_2 sufficiently small, S is square integrable (Santosa and Symes (1988)). Thus the basic data set of this paper will be a member of $L^2(R_1)$.

Another piece of notation needed to state our main result is used in our method of measuring "roughness": as mentioned in the Introduction, a stable solution of the inverse problem can only be expected for sufficiently nonsmooth coefficient c.

The "roughness" measure depends on an arbitrary Dirac kernel $h_1 \in C_0^\infty(\mathbf{R})$ satisfying

$$h_1 \geq 0, \quad h_1(0) > 0, \quad \int h_1 = 1.$$

Set

$$h_\epsilon(z) = \frac{1}{\epsilon} h_1\left(\frac{z}{\epsilon}\right).$$

For $Z_0, \epsilon, \Delta > 0$, $c \in H^1_{\text{loc}}$, define

$$\bar{r}[c](Z_0, \Delta) = \sup_{0 \leq z \leq Z_0} \frac{1}{\Delta} \int_{z-\Delta/2}^{z+\Delta/2} |c'|^2,$$

$$r_*[c](Z_0,\epsilon,\Delta) = \inf_{0\le z\le Z_0} \frac{\epsilon^2}{\Delta}\int_{z-\Delta/2}^{z+\Delta/2} |h'_\epsilon * c'|^2,$$

$$r^*[c](Z_0,\epsilon,\Delta) = \sup_{0\le z\le Z_0} \frac{\epsilon^2}{\Delta}\int_{z-\Delta/2}^{z+\Delta/2} |h'_\epsilon * c'|^2.$$

These are local average measures of fluctuation. For example, r_* is a sizable fraction of $\bar{r}, r^*$ when c has significant Fourier components at frequencies proportional to $\frac{1}{\epsilon}$, locally uniformly on the length scale Δ.

The geometry of the plane-wave problem, which will occupy most of this paper, is determined by the travel-time function

$$\tau(z,p) = 2\int_0^z \sqrt{\frac{1}{c^2} - p^2},$$

which gives the time necessary for a point on a planar wavefront at slowness p, fixed horizontal coordinates, to travel to depth z and back to the surface.

The earlier results of Symes (1986a, 1986b) imply that S extends to a bounded, continuous map on the bounded set $\sum_c \subset H^1_{\text{loc}}(\mathbf{R})$ parameterized by positive numbers T_0, P_1, c_0, c_1, c^*, according to

$$\begin{aligned}\Sigma_c \quad = \quad & \{c \in H^1_{\text{loc}}(\mathbf{R}) : c(z) = c_0, z < 0; \quad \text{for} \\ & Z_0 > 0 \quad \text{so that} \quad T_0 = \tau(Z_0, P_1), \\ & \|\log c\|_{H^1[0,Z_0]} \le c^*; \\ & c(z) = c_1 \quad \text{for} \quad z \ge Z_0\}.\end{aligned}$$

We have also shown, however, that this extension is *not* locally Lipschitz-continuous, and certainly not differentiable, in these metrics. S does become differentiable when the domain is metricized more strictly (H^3), but then the derivative fails to have a lower bound. Thus the implicit function theorem does not apply to the solution of least-squares problems for S. The computational consequences of this pathology are also striking (Santosa and Symes 1989). Proofs of the above assertions, together with an incisive discussion of their consequences, are given in the 1989 thesis of R. M. Lewis in the Department of Mathematical Sciences at Rice University.

A suitable family of "rough" subsets Σ'_c of Σ_c depends on positive parameters $M_1, M_2, \bar{\epsilon}$, and $\bar{\Delta}$ according to

$$\begin{aligned}\Sigma'_c \quad = \quad & \{c \in \Sigma_c : \quad \text{for} \quad Z_0 > 0 \quad \text{such that} \\ & T_0 = \tau(Z_0, P_1) \quad \text{and some} \quad 0 < \epsilon \le \bar{\epsilon}, \quad 0 < \Delta \le \bar{\Delta}, \\ & \text{the following inequalities hold:} \\ & M_1 \le r_*(Z_0,\epsilon,\Delta), \\ & M_2 r_*(Z_0,\epsilon,\Delta) \ge \max(\bar{r}(Z_0,\Delta), r^*(Z_0,\epsilon,\Delta))\}.\end{aligned}$$

We shall verify that Σ'_c is nonempty for suitable choices of parameters.

The main result of our paper is Theorem 2.1.

THEOREM 2.1. *Suppose that* $0, T_0 < T_1$, $0 \leq P_1 < P_2$, $0 \leq K_0$, $0 < K_1 \leq K^*$ $0 < c_0, c_1, c^*$ *are given. Then there exist constants* $M_1, M_2, \bar{\epsilon}, \bar{\Delta}, \bar{m}$, *and* L^* *depending on* T_0, T_1, P_1, P_2, K_0, K_1, K^*, c_0, c_1, *and* c^* *so that if* $f \in \mathcal{E}'(\mathbf{R})$ *satisfies* (2.1) *and*

$$m_f^1 \leq \bar{m}$$

then Σ_c' *is nonempty and there exists an open neighborhood* U *of the set*

$$\{S[c] \in L^2(R_1) : c \in \Sigma_c'\}$$

and a map

$$I : U \longrightarrow L^2_{\text{loc}}(\mathbf{R})$$

so that

(i) *For* $c \in \Sigma_c'$, $IS[c] = c$;
(ii) *For* $D_1, D_2 \in U$, $Z_1 > 0$

$$\|I(D_1) - I(D_2)\|_{L^2[0,Z_1]} \leq L^* \|D_1 - D_2\|_{L^2(R_1)}.$$

Thus we obtain a continuous left inverse for S, under various constraints. The requirement that c be constant for large z is simply a way of controlling c at depths below the zone influencing the seismogram. That the zone of constancy begins at Z_0 such that $T_0 = \tau(Z_0, P_1)$, rather than $T_1 = \tau(Z_0, P_1)$, is an unfortunate side effect of the "width" of the source wavelet f: since supp f is not a point, the depth interval in which the seismogram gives sure control over the velocity coefficient is strictly smaller than the depth interval needed to compute the seismogram. Given an arbitrary velocity profile with mean c_1 near $z = Z_0$, we can, of course, truncate it to a member of Σ_c, i.e., by setting the velocity constant ($= c_1$) for $z > Z_0$, as in the definition. The corresponding seismograms are then different only in the "gap" ($T_0 \leq t \leq T_1$). If the original profile obeys the uniform roughness conditions as in the definition of Σ_c', then it follows from arguments similar to those in §§3 and 4 that the L^2-norm of the difference of seismograms is $O(m_f^1)$. The theorem then gives the same qualitative estimate for the error due to application of I. We leave to the reader the formulation of a theorem embodying this extension of our results.

A more subtle consequence of this gap is that the value of c in this "basement" region must be specified a priori, i.e., the condition that $c(z) = c_1$ for $z \geq Z_0$. It seems clear that this additional piece of data should have little influence on the values of c at shallower depths—and, to the extent that it does, should be determined by S as well. This may be a fruitful subject for further work; some related ideas are discussed in Sacks and Santosa (1987). In any case the author does not see at present how to formulate a convenient theorem without such a restriction.

It is easy to see that the Lipschitz estimate (ii) cannot be strengthened much. See Symes (1986b), for instance. In particular it is not possible to replace L^2 by H^1 on the left-hand side.

Estimates of the sort presented in Theorem 2.1 are only of qualitative importance. Numerical evidence (Symes (1988), (1990); Symes and Carazzone (1989), (1990)) indicates that typical values of the Lipschitz constant L^* are very large. On the other

hand, restriction to a submanifold of Σ_c' of small codimension diminishes L^* to a useful magnitude, while leaving enough freedom in the model that some information about c is still obtained from the data. Analysis of an approximation to S in Symes (1988), (1990) and Symes and Carazzone (1989) illustrates this feature. A full understanding of the need for this "residual regularization" is still lacking at this writing.

The significance of the present result lies in its reliance on the implicit function theorem: i.e., the stability follows directly from linearization stability, and any residual ill-conditioning can be improved by the straightforward addition of linear constraints. This is surprising—and, perhaps, of practical importance—since the implicit function theorem cannot be applied to S directly, as noted above, for elliptic f.

The left inverse I will be produced via the solution of an auxiliary least-squares problem, developed in §5. In §7 we remove the elliptic requirement on f, to a certain extent: for suitable $f \in C_0^\infty$, we obtain an approximate left inverse to replace I, which depends on a choice of regularization.

While a few previous papers (Symes (1983), (1986a), (1986b), Fawcett (1984), Suzuki (1988)) provide rigorous treatment of the central uniqueness, existence, and continuous dependence issues, none treat the "bandlimited" problem described in §1 in a satisfactory way: without exception these works either assume the low-frequency content problem away, or ignore it. The problem is well known to geophysical researchers, and its severity is explicitly illustrated in Pao, Santosa, and Symes (1984) and Gray and Symes (1985).

Note that even the known uniqueness theorems for several-dimensional inverse problems in wave propagation either require data in a frequency interval $[0, \omega)$ (Sacks and Symes (1985), Sun (1988), Ramm (1986)) or do not allow the reflection configuration (sources and receivers separated from the target region by a hyperplane (Nachman (1987), Rakesh and Symes (1988)). Thus the present paper is the only instance, to the author's knowledge, in which any version of the reflection inverse problem has been shown to be well posed in the presence of a low-frequency lacuna.

Nonetheless, evidence of two sorts indicates that velocities, at least, are quite well determined by bandlimited data. Conventional seismic data processing, as practiced by academic and industrial reflection seismologists, appears to produce such information. While based on numerous drastic approximations, the so-called "velocity analysis" procedures incorporate a great deal of data-driven insight (see, for example, Yilmaz (1987), Chap. 3). A side effect of our work is to provide a partial but rigorous mathematical basis for these important techniques.

A universal feature of "velocity analysis," as currently practiced, is the implicit separation of velocity (and other parameters) into slowly and rapidly varying parts. The influence of the rapid parameter variations is treated as a perturbation about the slowly varying trends. (We warn the reader that this linearization is seldom stated explicitly in the applied literature, but is simply assumed without comment.) To the author, one of the most gratifying features of the treatment given below is the *natural* emergence of this scale separation in the course of our analysis (Theorem 4.1, arguments in §6).

Other indications that velocities are well determined are to be found in recent numerical investigations of the nonlinear least-squares problem

$$\min_c \|s[c] - s_{\text{data}}\|_{L^2}^2 . \tag{2.2}$$

These have turned up more direct evidence that bandlimited seismograms determine velocity profiles (see Kolb, Collino, and Lailly (1986), McAulay (1985), Gauthier,

Tarantola, and Virieux (1986), and Mora (1987)). All of these papers also reveal that any reasonable setting of (2.2) results in a very difficult optimization problem. The reasons for both the success and the difficulty of this so-called "least-squares inversion" are noted briefly in §3, and explained in great detail in Santosa and Symes (1989), to which we refer the reader for extensive discussion. In any case, the computational difficulty of (2.2) was the main motivation for the work reported here, which relies on a different, "relaxed" least-squares problem (§5).

In this paper, we give only a qualitative analysis of this "relaxed" problem, which we call the *coherency optimization problem*, leading to Theorem 2.1. A quantitative analysis of an approximation appears in Symes (1988), (1990), and numerical experiments are reported there and in Symes and Carazzone (1989) establishing the feasibility of the optimization, its relative insensitivity to noise and its favorable comparison to (2.2) regarding computational efficiency. In Symes and Carazzone (1989), (1990) the technique is applied to field reflection seismograms with quite satisfactory results.

A very important remaining question concerns the extension of these results to other models, notably to nonlayered velocities (i.e., c depending on all space variables). As shown in Symes and Carazzone (1989), for example, an approximation to the coherency optimization problem can be formulated for the general nonlayered fluid model. Preliminary numerical experiments with an implementation for two-dimensional constant density acoustics will appear in Symes (1991). A full-blown extension of our results awaits better understanding of propagation of regularity for hyperbolic equations with nonsmooth coefficients, and implications for the relation between solutions and coefficients, analogous to the results for problems in two independent variables detailed in §3.

3. Preliminary considerations: Reduction to plane waves, properties of the one-dimensional forward map. We assume that seismograms are given on an open set Ω of the space-time boundary of cylinder form:

$$\Omega = \Omega' \times [0, t_{\max}]$$

with Ω' a neighborhood of the "source" point $x = 0$. We shall also assume that all velocity profiles $c : \mathbf{R} \to \mathbf{R}^+$ satisfy

$$0 < c_{\min} \leq c(z) \leq c_{\max}\,, \qquad z \geq 0 \tag{3.1}$$

for a priori fixed $c_{\min}, c_{\max}$. Whenever convenient we will also think of (3.1) as an $L^\infty(\mathbf{R})$-bound on $\log c$.

The principal technical device of this paper is the introduction of the Radon transformed field

$$U(p, z, t) := \int_{\mathbf{R}^2} dx\, u(x, z, t + p \cdot x_1), \qquad p \in \mathbf{R}.$$

Standard arguments show that U is well defined for small p, t under the assumptions made so far. For an attempt to maximize the domain of definition of U, see Santosa and Symes (1988). We remind the reader that the Radon transform is defined on natural classes of distributions as a *push-forward* (Duistermaat (1973, Prop. 1.3.4, p. 3), for example). It is commonplace to retain the integral notation even for distribution arguments, however, and we shall follow this practice.

A straightforward calculation shows that for suitably small $p \geq 0$ so that $c_{\max} p < 1$,

$$\begin{aligned} \frac{1}{v^2(z,p)} \frac{\partial^2 U}{\partial t^2}(z,t) - \frac{\partial^2 U}{\partial z^2}(z,t) &= \delta(z) f(t), \\ U \equiv 0, \qquad t << 0 & \end{aligned} \tag{3.2}$$

where the *vertical wave velocity* $v(z,p)$ (or $v[c]$, to emphasize the dependence on c) is defined by

$$v[c](z,p) = \frac{c(z)}{\sqrt{1 - c^2(z)p^2}}.$$

Because of the a priori bounds (3.1), the support of u, hence of s, is contained in a cone

$$c_{\max} t \geq \sqrt{|x|^2 + z^2}.$$

Therefore, for sufficiently small $p_{\max} (< 1/c_{\max})$ there exists $\tau_{\max} \geq 0$ so that for

$$R = \{(t,p) : 0 \leq t \leq \tau_{\max}, |p| \leq p_{\max}\}$$

we have for $(t,p) \in R$

$$\begin{aligned} S[c](t,p) &:= \frac{\partial U}{\partial t}(p,0,t) - f(t) \\ &= \int_{\mathbf{R}^2} dx \left(\mathbf{1}_\Omega \frac{\partial u}{\partial t}(x,z,t+px_1) - f(t+px_1)\delta(x) \right), \end{aligned} \tag{3.3}$$

i.e., the domains of integration of the Radon integrals intersect the support of u inside Ω. We assume tacitly in the sequel that all (t,p) domains satisfy this constraint.

Now we recall some facts about the one-dimensional seismogram map $S[c](\cdot,p) =: S_0[v]$ (fixed p) for which references are Symes (1986a), (1986b). It is convenient to include explicitly the source wavelet temporarily in the notation. That is, write $S_0[v,f]$ for the map defined by the solution of (3.2), (3.3) followed by restriction to fixed p. Recall that f is assumed to define an *elliptic convolution operator of order zero*. For the choice $f = \delta$, S_0 defines a bounded continuous map from

$$H^{1,+}_{\mathrm{loc}}(\mathbf{R}, v_0) = \left\{ v \in H^1(\mathbf{R}) : v \equiv v_0 \text{ for } z < 0, \log v \in H^1_{\mathrm{loc}}(\mathbf{R}) \right\}$$

into $L^2[0,T]$ for any $T > 0$, but $S_0[\cdot,\delta]$ is not locally uniformly continuous (Symes (1986b)). In order to recover the necessary degree of regularity for the arguments to follow, we introduce the "travel-time velocity" $\tilde{v}[c]$ defined by

$$\tilde{v} \circ \tau = v$$

where

$$\tau(z) = 2 \int_0^z \frac{1}{v}$$

is the (one-way) travel time. More discussion of the map $c \mapsto \tilde{v}[c]$ appears in § 5 (see also Symes (1986a)). A short calculation shows that $\tilde{U}$, defined by

$$\tilde{U}(\tau(z),t) = U(z,t),$$

satisfies

$$\frac{1}{\tilde{v}(x)}\frac{\partial^2 \tilde{U}}{\partial t^2}(x,t) - \frac{\partial}{\partial x}\frac{1}{\tilde{v}(x)}\frac{\partial \tilde{U}}{\partial x}(x,t) = \frac{1}{v_0} f(t)\delta(x).$$

Set

$$\tilde{S}_0[\tilde{v},f](t) = \frac{\partial \tilde{U}}{\partial t}(0,t).$$

We recapitulate a number of properties of $\tilde{S}_0$. With exceptions noted below, all of these may be found in Symes (1986a). Note that $\tilde{S}_0[\tilde{v},f] \equiv S_0[v,f]$ if $\tilde{v}\circ\tau = v$. $\tilde{S}_0$ also defines a bounded map: $H^{1,+}_{\text{loc}}(\mathbf{R},v_0) \to L^2_{\text{loc}}(\mathbf{R})$ for any $v_0 > 0$. Moreover, $\tilde{S}_0$ is actually of class C^2, viewed as a map: $H^{1,+}([0,T/2],v_0) \to L^2[0,T]$. The derivative is given by the formal perturbation ($\delta\tilde{v} \in H^1_{\text{loc}}(\mathbf{R})$, $\delta\tilde{v} \equiv 0, z < 0$)

$$\begin{aligned}\left(\frac{1}{\tilde{v}(x)}\right)\frac{\partial^2 \delta\tilde{U}}{\partial t^2}(x,t) - \frac{\partial}{\partial x}\left(\frac{1}{\tilde{v}(x)}\frac{\partial \delta\tilde{U}}{\partial x}\right)(x,t)&\\ = \frac{\partial}{\partial x}\left(\frac{\delta\tilde{v}}{\tilde{v}^2}\frac{\partial \tilde{U}}{\partial x}(x,t)\right), \qquad x \geq 0,&\\ \delta\tilde{U} = 0, \qquad t << 0,&\end{aligned}$$

$$D\tilde{S}_0[\tilde{v},f]\delta\tilde{v} = \left.\frac{\partial \delta\tilde{U}}{\partial t}\right|_{x=0},$$

so that

$$\begin{aligned}\|\tilde{S}_0[\tilde{v}+\delta\tilde{v},f] - \tilde{S}_0[\tilde{v},f] - D\tilde{S}_0[\tilde{v},f]\cdot\delta\tilde{v}\|_{L^2[0,T]}&\\ = o\left(\|\delta\tilde{v}\|_{H^1([0,T/2])}\right).&\end{aligned}$$

For $f = \delta$, more is true: for constants $C_-, C_+ > 0$ depending on $\|\log\tilde{v}\|_{H^1[0,T/2]}$ and on T,

$$\begin{aligned} C_-\left\|\frac{\partial\delta\tilde{v}}{\partial x}\right\|_{L^2[0,T/2]} &\leq \left\|D\tilde{S}_0[\tilde{v},\delta]\delta\tilde{v}\right\|_{L^2[0,T]} \\ &\leq C_+\left\|\frac{\partial\delta\tilde{v}}{\partial x}\right\|_{L^2[0,T/2]}. \end{aligned} \tag{3.4}$$

Also, for $f = H$ (the Heaviside function), there exists C_0 depending on $\|\log\tilde{v}\|_{H^1[0,T/2]}$ and on T so that

$$\|D\tilde{S}_0[\tilde{v},H]\delta\tilde{v}\|_{L^2[0,T]} \leq C_0\|\delta\tilde{v}\|_{L^2[0,T/2]}. \tag{3.5}$$

Note that for $\delta\tilde{v} \in H^1_{\text{loc}}$, $D\tilde{S}_0[\tilde{v},H]\delta\tilde{v} \in H^1_{\text{loc}}$ and

$$D\tilde{S}_0[\tilde{v},\delta] = \frac{\partial}{\partial t}D\tilde{S}_0[\tilde{v},H].$$

Also

$$\begin{aligned} \tilde{S}_0[\tilde{v}, f] &= f * \tilde{S}_0[\tilde{v}, \delta], \\ D\tilde{S}_0[\tilde{v}, f] &= f * D\tilde{S}_0[\tilde{v}, \delta]. \end{aligned}$$

All of these results are based on simple local energy estimates; with the exception of (3.5) they are stated explicitly in Symes (1986a). The Heaviside estimate (3.5) is not given there, but the proof presents no novelties.

4. Ellipticity for the one-dimensional forward map. We have assumed (until §7) that $f \in \mathcal{E}'(\mathbf{R})$ defines an elliptic convolution operator of order zero. Since $D\tilde{S}_0[\tilde{v}, \delta]$ for $\log \tilde{v} \in H^1_{\text{loc}}$ is invertible, ((3.6), (3.7)), it seems clear that $D\tilde{S}_0[\tilde{v}, f] = f * \tilde{S}_0[\tilde{v}, \delta]$ should be "elliptic" as well. The purpose of this section is to formulate and prove a precise result along these lines, keeping track of the dependence of various constants on the H^1-norm of $\tilde{v}$, the time and depth intervals used, etc. The result is a Gårding-type estimate for $D\tilde{S}_0$, which requires that $D\tilde{S}_0$ be given on $[0, T_1]$, $T_1 > T_0$, to estimate $\delta\tilde{v}$ on $[0, T_0]$. It is clear that a little "extra" data is required, since the support of f is not assumed to be a point. In fact, the principal constant intervening in the estimates is the first moment of f, which measures its "spread," and is related in the estimates to the size of the necessary "margin" $T_1 - T_0$.

Select a cutoff function $\psi \in C^\infty(\mathbf{R})$ with $\psi(t) \equiv 0$ for $t > T_1$ and $\psi(t) \equiv 1$ for $t \leq T_0$, and define the smoothly cutoff version of $\tilde{S}_0$:

$$\tilde{S}_\psi := \psi \tilde{S}_0$$

so also

$$D\tilde{S}_\psi := \psi D\tilde{S}_0.$$

In the following, we will apply the estimates of the previous section on various t-intervals. Since the constants C_-, C_+, etc. depend on the length of the interval, we will include the length explicitly in the notation, for the moment. Thus for estimates on the interval $[0, T]$, C_- becomes $C_-[T]$, etc. These constants depend on the H^1-norm of $\log \tilde{v}$ on the appropriate intervals.

Recall that "elliptic," applied to f, means the inequalities (2.1), which we recall here for convenience: for $\phi \in H^1(\mathbf{R})$,

$$\begin{aligned} K^* \left\| \frac{\partial \phi}{\partial x} \right\|_{L^2(\mathbf{R})} &\geq \left\| f * \frac{\partial \phi}{\partial x} \right\|_{L^2(\mathbf{R})} \\ &\geq K_1 \left\| \frac{\partial \phi}{\partial x} \right\|_{L^2(\mathbf{R})} - K_0 \|\phi\|_{L^2(\mathbf{R})}, \\ \text{supp} f \subset \mathbf{R}^+. \end{aligned} \tag{4.1}$$

From (3.5) and (3.6), for $\log \tilde{v} \in H^1_{\text{loc}}(\mathbf{R})$, $\delta\tilde{v} \in H^1_{\text{loc}}(\mathbf{R})$, $\tilde{v}(x) = v_0$ for $x < 0$:

$$\begin{aligned} \|D\tilde{S}_\psi[\tilde{v}, \delta]\delta\tilde{v}\|_{L^2(\mathbf{R})} &\leq \|D\tilde{S}_0[\tilde{v}, \delta]\delta\tilde{v}\|_{L^2[0,T_1]} \\ &\leq C_+[T_1] \left\| \frac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2[0,T_1/2]} \end{aligned} \tag{4.2}$$

while

$$\begin{aligned} \|D\tilde{S}_\psi[\tilde{v},\delta]\delta\tilde{v}\|_{L^2(\mathbf{R})} &\geq \|D\tilde{S}_0[\tilde{v},\delta]\delta\tilde{v}\|_{L^2[0,T_0]} \\ &\geq C_-[T_0]\left\|\frac{\partial\delta\tilde{v}}{\partial x}\right\|_{L^2[0,T_0/2]}. \end{aligned} \tag{4.3}$$

Also

$$\|D\tilde{S}_\psi[\tilde{v},H]\delta\tilde{v}\|_{L^2(\mathbf{R}^+)} \leq C_0[T_1]\|\delta\tilde{v}\|_{L^2[0,T_1/2]}. \tag{4.4}$$

Now

$$\begin{aligned} D\tilde{S}_\psi[\tilde{v},f]\delta\tilde{v} &= \psi(f * D\tilde{S}_0[\tilde{v},\delta]\delta\tilde{v}) \\ &= f * D\tilde{S}_\psi[\tilde{v},\delta]\delta\tilde{v} + e \end{aligned}$$

where for e we have the standard commutator estimate

$$\|e\|_{L^2(\mathbf{R})} \leq \|\psi'\|_{L^\infty} m_f^1 \|D\tilde{S}_0[\tilde{v},\delta]\delta\tilde{v}\|_{L^2[0,T_1]}.$$

Here m_f^1 denotes the measure norm of the distribution

$$\phi \longmapsto \langle |f|, t\phi\rangle,$$

i.e., the first moment of f, as explained in §2.

Combine this estimate with (4.1)–(4.4) to get

$$\begin{aligned} \|D\tilde{S}_0[v,f]\delta\tilde{v}\|_{L^2[0,T_1]} &\geq \|D\tilde{S}_\psi[\tilde{v},f]\delta\tilde{v}\|_{L^2(\mathbf{R})} \\ &\geq \|f * \psi\tfrac{\partial}{\partial t} D\tilde{S}_0[\tilde{v},H]\delta\tilde{v}\|_{L^2(\mathbf{R})} - \|e\|_{L^2(\mathbf{R})} \\ &\geq \|f * \tfrac{\partial}{\partial t}\psi D\tilde{S}_0[\tilde{v},H]\delta\tilde{v}\|_{L^2(\mathbf{R})} \\ &\quad -\|f * \tfrac{\partial\psi}{\partial t} D\tilde{S}_0[\tilde{v},H]\delta\tilde{v}\|_{L^2(\mathbf{R})} - \|e\|_{L^2(\mathbf{R})} \\ &\geq K_1\|\tfrac{\partial}{\partial t}\psi D\tilde{S}_0[\tilde{v},H]\delta\tilde{v}\|_{L^2(\mathbf{R})} - K_0\|\psi D\tilde{S}_0[\tilde{v},H]\delta\tilde{v}\|_{L^2(\mathbf{R})} \\ &\quad -K^*\|\tfrac{\partial\psi}{\partial t} D\tilde{S}_0[\tilde{v},H]\delta\tilde{v}\|_{L^2(\mathbf{R})} - \|e\|_{L^2(\mathbf{R})} \\ &\geq K_1\|D\tilde{S}_0[\tilde{v},\delta]\delta\tilde{v}\|_{L^2[0,T_0]} - K_0\|D\tilde{S}_0[\tilde{v},H]\delta\tilde{v}\|_{L^2[0,T_1]} \\ &\quad -K^*\|D\tilde{S}_0[\tilde{v},H]\delta\tilde{v}\|_{L^2[T_0,T_1]}\left\|\tfrac{\partial\psi}{\partial t}\right\|_{L^\infty(\mathbf{R})} \\ &\quad -m_f^1\|\tfrac{\partial\psi}{\partial t}\|_{L^\infty(\mathbf{R})}\|D\tilde{S}_0[\tilde{v},\delta]\delta\tilde{v}\|_{L^2[0,T_1]} \\ &\geq K_1 C_-[T_0]\|\tfrac{\partial\delta\tilde{v}}{\partial x}\|_{L^2[0,T_0/2]} \\ &\quad -C_0[T_1]\left(K_0 + K^*\left\|\tfrac{\partial\psi}{\partial t}\right\|_{L^\infty(\mathbf{R})}\right)\|\delta\tilde{v}\|_{L^2[0,T_1/2]} \\ &\quad -m_f^1\|\tfrac{\partial\psi}{\partial t}\|_{L^\infty(\mathbf{R})} C_+[T_1]\|\tfrac{\partial\delta\tilde{v}}{\partial x}\|_{L^2[0,T_1/2]}. \end{aligned}$$

The upshot of all of this is the inequality

$$\begin{aligned}\|D\tilde{S}_0[\tilde{v},f]\delta\tilde{v}\|_{L^2[0,T_1]} \quad \geq \quad & K_1 C_-[T_0] \left\| \tfrac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2[0,T_0/2]} \\ & - m_f^1 \left\| \tfrac{\partial \psi}{\partial t} \right\|_{L^\infty(\mathbf{R})} C_+[T_1] \left\| \tfrac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2[0,T_1/2]} \\ & - C_0[T_1] \left(K_0 + K^* \left\| \tfrac{\partial \psi}{\partial t} \right\|_{L^\infty(\mathbf{R})} \right) \|\delta\tilde{v}\|_{L^2[0,T_1/2]}.\end{aligned}$$

Remark. As noted in Symes (1986a), in general the constants $C_+[T]$, $C_0[T]$ increase with T, while $C_-[T]$ decreases.

Remark. It is worth noting the relation of the various bounds involving f to its Fourier transform. Indeed, obviously

$$K^* \geq \|\hat{f}\|_{L^\infty(\mathbf{R})}$$

whereas K_0, K_1 are related to the detailed behaviour of the Fourier transform near $\omega = 0$. Suppose that, for some K_*, $\omega > 0$,

$$|\hat{f}(\omega)| \geq K_* \quad \text{for } |\omega| > \omega \tag{4.5}$$

(i.e., $[\omega, \infty)$ constitutes the "passband" of $\hat{f}$, measured with tolerance K_*). Then for $\phi \in H^1(\mathbf{R})$, it is easy to see that

$$\left\| f * \frac{\partial \phi}{\partial t} \right\|_{L^2(\mathbf{R})} \geq K_* \left[\left\| \frac{\partial \phi}{\partial t} \right\|_{L^2(\mathbf{R})} - \omega \|\phi\|_{L^2(\mathbf{R})} \right]$$

which gives the relations

$$K_1 \geq K_*, \qquad K_0 \leq \omega K_*.$$

Note also the effect of scaling: if f satisfies (4.5), then

$$f_\epsilon(t) := \frac{1}{\epsilon} f\left(\frac{t}{\epsilon}\right)$$

satisfies (4.5) with $\omega_{\min}/\epsilon = \omega_{\min}/\epsilon$, while $m^1_{f_\epsilon} = \epsilon m^1_f$. Thus $K_1 = \mathcal{O}(1)$, $K_0 = \mathcal{O}(\frac{1}{\epsilon})$, and $m^1_{f_\epsilon} = \mathcal{O}(\epsilon)$ as $\epsilon \to 0$.

It is clear from the preceding discussion and the form of (4.5) that, for any prescribed $T_1 > T_0, \psi$ as above, and any "base" source wavelet f, a scaled version of f will have small enough first moment that

$$m_f^1 \left\| \frac{\partial \psi}{\partial t} \right\|_{L^\infty(\mathbf{R})} C_+[T_1] \leq \frac{1}{2} K_1 C_-[T_0]$$

with fixed K_1 (independent of scaling). Note that we can certainly choose ψ so that

$$\left\| \frac{\partial \psi}{\partial x} \right\|_{L^\infty(\mathbf{R})} \leq \frac{2}{T_1 - T_0}.$$

These obserations establish the nonvacuousness of Theorem 4.1.

THEOREM 4.1. *Choose* $T_1 > T_0 > 0$, *and* $K^* > K_1 > 0$. *Then for any* $\log \bar{v} \in H^1_{\text{loc}}(\mathbf{R})$, *there exists* $\bar{m}, L_2, L_1, L_0 > 0$ *depending on* $\tilde{v}$ *and on* T_1, T_0, K^*, K_1, *and so that if* f *satisfies* (4.1) *with some* $K_0 > 0$ *and*

$$m_f^1 \leq \bar{m}$$

then for $\delta\tilde{v} \in H^1_{\text{loc}}(\mathbf{R})$,

$$\|D\tilde{S}_0[\tilde{v}, f]\delta\tilde{v}\|_{L^2[0,T_1]} + m_f^1 L_2 \left\| \frac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2[T_0/2, T_1/2]} \geq L_1 \left\| \frac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2[0,T_0/2]} - L_0 \|\delta\tilde{v}\|_{L^2[0,T_1/2]}.$$

5. The optimum coherency principle. While we have shown that the linearized one-dimensional forward map is elliptic under the circumstances which concern us, it is certainly not boundedly invertible—or rather, the hypotheses concerning f do not imply any uniform bound on the inverse. This circumstance is widely remarked in the literature (for a sample, see Santosa and Symes (1989)), where numerical examples are also given (see especially Chapters 6 and 7). Recall, however, the provenance of the one-dimensional problem: it governs the propagation of a plane wave, the surface data for which are identical to the Radon transform of the point source surface data at fixed slowness (or angle). The possibility remains that the collection of *all* (precritical) plane-wave data might constrain the velocity estimate more severely than does a single plane-wave component.

In this and the next section we confirm this possibility. Since we will consider the data in an interval of slownesses $P_1 \leq p \leq P_2$, we will work with a suite of travel-time velocity models $\{\tilde{v}(t,p) : 0 \leq t \leq T,\ P_1 \leq p \leq P_2\}$. Recall from §3 that these are derived from velocity profiles $c(z)$; accordingly we begin with the question: what condition must $\tilde{v}(t,p)$ satisfy in order that $\tilde{v} = \tilde{v}[c]$ for some c? That is, we seek an operator whose kernel is identical to the range of $c \to \tilde{v}[c]$. We will call membership in the null space of the required operator (or in the range of $c \to \tilde{v}[c]$) the *coherency condition*, since coherence of the travel-time velocities $\tilde{v}$ is then forced across various values of p: all are representations of the same mechanical model, in different coordinate systems.

Denote by $\zeta(t,p)$ the inverse of the two-way travel time function $\tau(z,p)$, i.e.,

$$t = 2\int_0^{\zeta(t,p)} \frac{dz}{v(z,p)}.$$

Then clearly

$$\zeta(t,p) = \int_0^{t/2} d\tau \tilde{v}(\tau, p).$$

We will now regard ζ as being *defined* by this formula, hence as a functional of $\tilde{v}$. So, given $\tilde{v}$ we can compute ζ, whether $\tilde{v} = \tilde{v}[c]$ for some c, or not. Thus we can compute the quantity

$$\gamma[\tilde{v}] = \left[\frac{1}{(\tilde{v} \circ \zeta^{-1})^2} + p^2 \right].$$

(Here $\tilde{v} \circ \zeta^{-1}(z,p) = \tilde{v}(\zeta^{-1}(z,p),p)$.) Referring to the definitions (§3), we see that if $\tilde{v} = \tilde{v}[c]$, then

$$\gamma[\tilde{v}] = \frac{1}{c^2}$$

and it is thus independent of p: that is,

$$\tilde{v} = \tilde{v}[c] \Rightarrow \frac{\partial}{\partial p}\gamma[\tilde{v}] \equiv 0. \tag{5.1}$$

This last condition still involves the travel-time change of variables, so does not define a sufficiently regular function of $\tilde{v}$ (recall the discussion in §3). Define instead

$$\begin{aligned} \tilde{Q}[\tilde{v}] &= -\frac{\tilde{v}^3}{2}\left[\frac{\partial}{\partial p}\gamma[\tilde{v}]\right] \circ \zeta \\ &= -\frac{\tilde{v}^3}{2}\left(\left[\frac{\partial}{\partial p}(\tilde{v} \circ \zeta^{-1})^{-2}\right] \circ \zeta\right) - p\tilde{v}^3 \\ &= \left\{\frac{\partial \tilde{v}}{\partial p} + \left(\frac{\partial}{\partial p}\zeta^{-1}\right) \circ \zeta \frac{\partial \tilde{v}}{\partial t}\right\} - p\tilde{v}^3. \end{aligned} \tag{5.2}$$

A short chain-rule calculation gives

$$\left(\frac{\partial}{\partial p}\zeta^{-1}\right) \circ \zeta = -\frac{2}{\tilde{v}}\int_0^{t/2} \frac{\partial \tilde{v}}{\partial p},$$

which relation allows us to view $\tilde{Q}[\tilde{v}]$ as a functional of $\tilde{v}$. Clearly, from (5.1)

$$\tilde{v} = \tilde{v}[c] \Rightarrow \tilde{Q}[\tilde{v}] = 0.$$

It will be important to define the coherency condition in such a way as to have the largest possible domain contained in $H^{1,0}_{\text{loc}}(\mathbf{R} \times [P_1, P_2])$ (which will be the natural domain for the forward map, defined below). An obvious choice is H^1, but $\tilde{Q}$ is not continuous in that topology. As it happens, we can replace $\tilde{Q}$ with another operator having the same kernel, but which is continuous (even C^∞) in the H^1 sense.

Suppose temporarily that $\tilde{v} = \tilde{v}[c]$ for some c. Then

$$\begin{aligned} \frac{1}{\tilde{v}}\tilde{Q}[\tilde{v}] &= \frac{1}{\tilde{v}}\frac{\partial \tilde{v}}{\partial p} - \frac{2}{\tilde{v}^2}\left(\int_0^{t/2} \frac{\partial \tilde{v}}{\partial p}\right)\frac{\partial \tilde{v}}{\partial t} - p\tilde{v}^2 \\ &= \frac{\partial}{\partial t}\left(\frac{2}{\tilde{v}}\int_0^{t/2} \frac{\partial \tilde{v}}{\partial p} - 2p\int_0^{t/2} \tilde{v}^2\right) \\ &= 0. \end{aligned}$$

Thus

$$\tilde{v} = \tilde{v}[c] \Rightarrow \frac{2}{\tilde{v}}\int_0^{t/2} \frac{\partial \tilde{v}}{\partial p} = 2p\int_0^{t/2} \tilde{v}^2. \tag{5.3}$$

The map

$$Q[\tilde{v}] := \frac{\partial \tilde{v}}{\partial p} - 2p\left(\int_0^{t/2} \tilde{v}^2\right)\frac{\partial \tilde{v}}{\partial t} - p\tilde{v}^3$$

therefore also satisfies

$$\tilde{v} = \tilde{v}[c] :\Rightarrow Q[\tilde{v}] = 0. \tag{5.4}$$

On the other hand, for any $T > 0$, $R = [0,T] \times [P_1, P_2]$, Q obviously defines a C^2-map

$$Q : H^1(R) \to L^2(R).$$

The following converse to (5.4) shows that $Q = 0$ is an adequate coherency condition.

LEMMA 5.1. *Suppose that* $\log \tilde{v} \in H^1(R)$ *and* $Q[\tilde{v}] = 0$. *Then for some* $Z > 0$ *and some* $c \in H^1[0,Z]$,

$$\tilde{v} = \tilde{v}[c]|_R.$$

Proof. Integrate in t:

$$\begin{aligned} 0 &= \int_0^{t/2} dt' Q[\tilde{v}](t',p) \\ &= \int_0^{t/2} dt' \left\{ \frac{\partial \tilde{v}}{\partial p}(t',p) - 2p \left[\int_0^{t'/2} dt'' \tilde{v}^2(t'',p) \right] \frac{\partial \tilde{v}}{\partial t}(t',p) - p\tilde{v}^3(t',p) \right\} \\ &= \int_0^{t/2} dt' \frac{\partial \tilde{v}}{\partial p}(t',p) - 2p\tilde{v}(t,p) \int_0^{t/2} dt' \tilde{v}^2(t',p) \end{aligned}$$

after integration by parts, so we once again recover the relation

$$\frac{1}{\tilde{v}} \int_0^{t/2} \frac{\partial \tilde{v}}{\partial p} = 2p \int_0^{t/2} \tilde{v}^2.$$

It follows immediately that $\tilde{Q}[\tilde{v}] = 0$ as well, which is equivalent to

$$\frac{\partial}{\partial p} \gamma[\tilde{v}] = 0.$$

Set

$$\begin{aligned} R_\zeta &= \{(z,p) : P_1 \le p \le P_2,\ 0 \le z \le \zeta(T,p)\}, \\ v &= \tilde{v} \circ \zeta^{-1}. \end{aligned}$$

It is easily checked that $\log v \in H^1(R_\zeta)$. On the other hand, with

$$\begin{aligned} c(z) : &= \left(\frac{1}{v(z,p)^2} + p^2 \right)^{-1/2} \\ &= \gamma[\tilde{v}]^{-1/2}(z,p) \end{aligned}$$

we have $c \in H^1[0,Z]$, $Z = \sup_{P_1 \le p \le P_2} \zeta(T,p)$, and

$$v(z,p) = \frac{c(z)}{\sqrt{1 - c^2(z)p^2}},$$

$$\zeta^{-1}(z,p) = 2\int_0^z \frac{1}{v},$$

whence $\zeta^{-1} = \tau$, and the conclusion follows. □

We now turn to the definition of the multi-plane-wave forward map. For simplicity, define seismograms on the *data rectangle*

$$R_1 = [0,T_1] \times [P_1,P_2].$$

Choose $c_0, c_1 > 0$ with $c_0P_2, c_1P_2 < 1$, $T_2 \geq T_0$, and $\ell^* > 0$ and set

$$\sum = \{\tilde{v} \in H^1_{\text{loc}}(\mathbf{R} \times [P_1,P_2]) : \tilde{v}(x,p) = c_0/\left(1 - c_0p^2\right)^{1/2}, \quad x < 0,$$

$$\tilde{v}(x,p) = c_1/\left(1 - c_1^2p^2\right)^{1/2},$$

$$x \geq \tfrac{1}{2}T_2$$

$$\|\log \tilde{v}\|_{H^1([0,1/2T_2]\times[P_1,P_2])} < \ell^*\}$$

and its "tangent space"

$$T\sum = \{\delta\tilde{v} \in H^1_{\text{loc}}(\mathbf{R} \times [P_1,P_2]) :$$
$$\delta v(x,p) = 0, \quad x \leq 0 \text{ or } x \geq \tfrac{1}{2}T_2\}.$$

Only a finite interval in t is needed for the arguments which follow. With $T_2 > T_1$ to be determined below, set

$$\tilde{R}_2 = [0,T_2/2] \times [P_1,P_2].$$

Identify elements of $\sum$ and $T\sum$ with their restrictions to $\tilde{R}_2$, and topologize $\sum$ and $T\sum$ as subsets of $H^1(\tilde{R}_2)$.

With these conventions, define the forward map

$$\tilde{S} : \sum \longrightarrow L^2(R_1)$$

by

$$\tilde{S}[\tilde{v}](t,p) = \tilde{S}_0[\tilde{v}(\cdot,p), f](t)$$

with $\tilde{S}_0$ as in §4. From the results stated there and in §3, it follows that $\tilde{S}$ is of class C^2, with derivatives bounded in terms of c_0, T_1, f, and ℓ^*.

The derivative $D\tilde{S}[\tilde{v}]$ also obeys an "elliptic" estimate. To state this, set

$$\tilde{R}_0 = [0,T_0/2] \times [P_1,P_2], \qquad \tilde{R}_1 = [0,T_1/2] \times [P_1,P_2].$$

Then from Theorem 4.1 follows Theorem 5.2.

THEOREM 5.2. *Given $P_2 > P_1 \geq 0$, $T_1 > T_0 > 0$, $K^* \geq K_1 > 0$, and $\ell^* > 0$, there exist $\bar{m}, L_0, L_1, L_2, L_1 > 0$, so that for $K_0 > 0$ and $f \in \mathcal{E}'(\mathbf{R})$ satisfying* (4.1) *and*

$$m_f^1 \leq \bar{m}$$

so that for $\tilde{v} \in \sum, \delta\tilde{v} \in T\sum$

$$\|D\tilde{S}[\tilde{v}]\delta\tilde{v}\|_{L^2(R_1)} + m_f^1 L_2 \left\| \frac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2(\tilde{R}_1 \setminus \tilde{R}_0)} \geq L_1 \left\| \frac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2(\tilde{R}_0)} - L_0 \|\delta\tilde{v}\|_{L^2(\tilde{R}_1)}.$$

Moreover, for each $p \in [P_1, P_2]$,

$$\|D\tilde{S}[\tilde{v}]\delta\tilde{v}(\cdot, p)\|_{L^2[0,T_1]} + m_f^1 L_2 \left\| \frac{\partial \delta\tilde{v}}{\partial x}(\cdot, p) \right\|_{L^2[T_0/2, T_1/2]} \geq L_1 \left\| \frac{\partial \delta\tilde{v}}{\partial x}(\cdot, p) \right\|_{L^2[0,T_0/2]} - L_0 \|\delta\tilde{v}(\cdot, p)\|_{L^2[0,t_1/2]}.$$

In view of Lemma 5.1 and the obvious relation

$$\tilde{S}[\tilde{v}[c]] = S[c],$$

we can now state a version of the inverse problem closely related to the least-squares problem (2.2), as:

$$\begin{array}{ll} \min & \|\tilde{S}[\tilde{v}] - D\|^2_{L^2(R_1)} \quad \text{over } \tilde{v} \in \sum \\ \text{subject to} & Q[\tilde{v}] = 0. \end{array} \tag{5.5}$$

In fact, a solution to this problem clearly yields a solution to (2.2) on a suitable depth interval. On the other hand, this problem would appear to have the advantage of regularity: both the objective and constraint functions are of class $C^{2,1}$. Moreover, it is possible to show that, under the circumstances described in Theorem 6.1 below, the Hessian operator of the objective function is positive definite on the null space of the linearized constraints, at a consistent data set, i.e., when

$$D = S[c^*]$$

for suitable c^*.

Unfortunately these properties are insufficient to yield a stable local-existence result. To motivate the next step in the development, we digress, with a brief review of Lagrangian theory for constrained optimization, and a simple but closely related example.

Suppose that X, Y are Hilbert spaces, $f : U \to \mathbf{R}, g : U \to Y$ smooth on an open set $U \subset X$. As is well known (Luenberger (1973)), a local solution of the constrained optimization problem

$$\begin{array}{ll} \min\limits_{x \in U} & f(x) \\ \text{subject to} & g(x) = 0 \end{array}$$

is a critical point of the Lagrangian

$$\mathcal{L}(x, \lambda) = f(x) + \langle \lambda, g(x) \rangle.$$

We wish to apply the implicit function theorem, to assure that a solution is stable against $(C^{2,1})$ perturbations in f. The conditions necessary to apply the implicit function theorem to the critical-point problem

$$\operatorname{grad}_{x,\lambda} \mathcal{L}(x,\lambda) = 0$$

are also sufficient to ensure the convergence of Newton's method (and, generally, of its computationally efficient quasi-Newton relatives). These amount to

(i) $Dg(x)\delta x = 0 \Longrightarrow \langle \delta x, [\operatorname{Hess} f(x) + \langle \lambda, \operatorname{Hess} g(x)\rangle], \delta x\rangle \geq \ell_1 \|\delta x\|^2$,

(ii) $\|Dg(x)^*\lambda\| \geq \ell_2 \|\lambda\|$

for constants $\ell_1, \ell_2 > 0$. The so called "second-order sufficiency" condition (i) may be verified for the problem (5.3), as was mentioned above. The so-called "constraint qualification," however, fails, in the manner illustrated by the following simple example, which captures the main features of (5.5).

Let $X = \{u \in H^1(R) : \int \int_R u = 0\}$, $Y = L^2(R)$, $R = [0,1]^2$ the unit square in $\mathbf{R}^2$. Set for $u \in X$, $v \in Y$ given

$$f(u) = \left\| \frac{\partial u}{\partial x_1} - v \right\|^2_{L^2(R)}, \qquad g(u) = \frac{\partial u}{\partial x_2}$$

and let us suppose that $v = v(x_1)$, so that the problem

$$\begin{array}{ll} \min\limits_{u \in X} & \operatorname{grad} f(u) \\ \text{subject to} & g(u) = 0 \end{array}$$

has the unique, zero-residual solution

$$u(x_1, x_2) = \int_0^{x_1} dx\, v(x) + \text{const.}$$

The second-order sufficiency condition (i) is an immediate consequence of a form of Poincaré's inequality (Nečas (1967)). The adjoint $Dg^*(u)$ is given by

$$Dg^*(u) = -\mathcal{N} \frac{\partial}{\partial x_2}$$

where $\mathcal{N}$ is the solution operator of the Neumann problem: $\mathcal{N}b = w$, where

$$\begin{aligned} (I - \Delta)w &= b \quad \text{in } R, \\ \frac{\partial w}{\partial n} &= 0 \quad \text{on } \partial R. \end{aligned}$$

Set $\lambda_k(x) = \cos k_1 \pi x_1 \cos k_2 \pi x_2$ for $k \in \mathbf{Z}^2$. Then

$$\|Dg^*(u)\lambda_k\|_{H^1(R)} = \frac{2k_2}{\sqrt{1 + k_1^2 + k_2^2}},$$

which can be made as small as we like by taking k_1 large. Thus the constraint qualification (ii) fails.

For similar but nonlinear problems such as (5.5), the nonzero singular values of Dg are proportional to the instantaneous radii of curvature of the constraint set. The

failure of the constraint qualification could conceivably be associated with the presence of arbitrarily small radii of curvature—indeed, this must be the case for uniformly nonlinear problems like (5.5). In essence, the constraint set for such a problem has a cusp at every point! No reasonable stability properties can be expected for the solutions of such a problem.

On the other hand, the Hessian operator of the "penalized" cost functional

$$f(u) + \sigma^2 \|g(u)\|^2_{L^2(R)}$$

is positive-definite for any $\sigma > 0$—again, this is simply Poincaré's inequality. This observation motivates the following construction.

For choices of time and slowness intervals $T_2 > T_1 > T_0 > 0$, $P_2 > P_1 \geq 0$, "data set" $D \in L^2(R_1)$ and "tuning" parameters $\sigma, \lambda \geq 0$, define for $\tilde{v} \in \sum$

$$J_{\sigma,\lambda}[\tilde{v}] = \frac{1}{2}\left\{ \left\| \tilde{S}[\tilde{v}] - D \right\|^2_{L^2(R_1)} + \sigma^2 \|Q\tilde{v}\|^2_{L^2(\tilde{R}_2)} + \lambda^2 \left\| \frac{\partial \tilde{v}}{\partial x} \right\|_{L^2(\tilde{R}_2 \setminus \tilde{R}_0)} \right\}. \tag{5.6}$$

The first two summands in the definition of $J_{\sigma,\lambda}$ are motivated by analogy with the "Dirichlet" problem discussed above. The form of the "elliptic" estimate (Theorem 5.2) and the need, explained in the next section, to choose $T_2 > T_0$ motivate the last term. For example, the elliptic estimate gives a bound on $\delta\tilde{v}$ only on the shorter interval $[0, T_0]$, so $\delta\tilde{v}$ must be bounded a priori for $t > T_0$.

We shall call the three terms on the right-hand side of (5.6) "data," "coherency," and "extension" terms. Minimization of (5.6) is the "coherency optimization problem."

6. Proof of the main theorem. We begin this section with the proof that the Hessian operator of $J_{\sigma,\lambda}$ as defined in § 5 is *positive-definite* at a sufficiently rough, coherent $\tilde{v}$. The main idea is that the Hessian quadratic form consists of the "data" term dominating the high frequencies in $\delta\tilde{v}$, and a "coherency" term essentially consisting of the product of the indefinite integral (in x) of $\delta\tilde{v}$ and a derivative of $\tilde{v}$. This latter term is thus the product of a smooth factor (depending on $\delta\tilde{v}$) and a rough factor (derivative of $\tilde{v}$). When the rough factor is rough enough, uniformly on the length scale of significant change in the smooth factor, then the product dominates the smooth factor. Lemma 6.2 below makes this heuristic reasoning precise, and establishes a mathematical meaning for "sufficiently rough." The "smooth factor" discussed here is the indefinite (t-) integral of $\delta\tilde{v}$; the estimate for it, together with the elliptic estimate from § 5 and an interpolation argument, give a bound on $\delta\tilde{v}$ in terms of the "data," "coherency," and "extension" terms of the Hessian of $J_{\sigma,\lambda}$.

Recall from the preceding sections the geometry of the coherency optimization problem:

$$\begin{aligned} &0 < T_0 < T_1 < T_2 && \text{time limits} \\ &0 \leq P_1 < P_2 && \text{slowness limits,} \end{aligned}$$

$$\begin{aligned} R_\nu &= [0, T_\nu] \times [P_1, P_2], \\ \tilde{R}_\nu &= [0, T_\nu/2] \times [P_1, P_2], \end{aligned} \qquad \nu = 0, 1, 2$$

(data, model rectangles);

the function spaces involved in its setting:

$$\sum = \left\{ \tilde{v} \in H^1(\tilde{R}_2) : \tilde{v}(0,p) = \frac{c_0}{\sqrt{1-c_0^2p^2}}, \tilde{v}(T_2,p) = \frac{c_1}{\sqrt{1-c_1^2p^2}}, \| \log \tilde{v} \|_{H^1(\tilde{R}_2)} < \ell^* \right\},$$

$$\dot{\sum} = \left\{ \delta\tilde{v} \in H^1(\tilde{R}_2) : \delta\tilde{v}(0,p) = 0 = \delta\tilde{v}(T_2,p) \right\}$$

and the maps:

$$\begin{aligned} \tilde{S} : &\quad \sum \longrightarrow L^2(R_1), \\ D\tilde{S} : &\quad \sum \times \dot{\sum} \longrightarrow L^2(R_1), \\ Q : &\quad \sum \longrightarrow L^2(\tilde{R}_2), \\ DQ : &\quad \sum \times \dot{\sum} \longrightarrow L^2(\tilde{R}_2). \end{aligned}$$

It will be essential in the arguments given below that T_2 be related appropriately to T_0, as follows. For $\tilde{v} \in \sum$, the L^∞-bound on $\tilde{v}$ implies that

$$\begin{aligned} \zeta(T_0, P_2) &= \sup_{P_1 \le p \le P_2} \zeta(T_0, p) \\ &\le \frac{1}{2} T_0 e^{\ell^*} := Z \end{aligned}$$

whence

$$\begin{aligned} \tau(Z, P_1) &= \sup_{P_1 \le p \le P_2} \tau(Z, p) \\ &\le 2Ze^{\ell^*} = T_0 e^{2\ell^*}. \end{aligned}$$

Set $T_2 = T_0 e^{2\ell^*}$. Then it follows that for any $p_1, p_2 \in [P_1, P_2]$,

$$\tau(\zeta(T_0, p_1), p_2) \le T_2.$$

Define

$$R_Z = [0, Z] \times [P_1, P_2].$$

Then for $\delta\tilde{v} \in H^1_{\text{loc}}(\mathbf{R})$, $\tilde{v} \in \sum$

$$\begin{aligned} \|\delta\tilde{v}\|_{H^1(\tilde{R}_0)} &\le C \|\delta\tilde{v} \circ \tau\|_{H^1(R_Z)} \\ &\le C' \|\delta\tilde{v}\|_{H^1(\tilde{R}_2)} \end{aligned}$$

and similarly for L^2-norms.

Here we have introduced the habit, to which we shall uniformly adhere, of denoting by $C, C', \cdots$ constants which may be chosen uniformly over $\sum$. We shall make no attempt to identify optimum choices of such constants, so that the end result of our argument is only qualitative in nature.

Recall from §2 that the determination of c from bandlimited data requires that c be "rough" in some sense. For convenience, we restate the basic "roughness" criterion, phrased in terms of an (arbitrary) Dirac kernel

$$h_\epsilon(z) = \frac{1}{\epsilon} h_1\left(\frac{z}{\epsilon}\right)$$

where h_1 satisfies the usual requirements:

$$h_1 \in C_0^\infty(\mathbf{R}), \quad h_1(0) > 0, \quad h_1 \geq 0\,, \quad \int h_1 = 1\,.$$

Then $\{\epsilon h_\epsilon'\}$ is bounded in L^1, and defines a family of "low-cut" filters, i.e., convolution with $\epsilon h_\epsilon'$ suppresses Fourier components at frequencies less than or equal to $\mathcal{O}(1/\epsilon)$.

For $c \in H^1_{\text{loc}}(\mathbf{R})$, $\epsilon > 0, \Delta > 0$ define

$$\begin{aligned} r[c](z,\epsilon,\Delta) &= \frac{1}{\Delta}\int_{z-\Delta/2}^{z+\Delta/2} |\epsilon h_\epsilon' * c'|^2, \\ \bar{r}[c](z,\Delta) &= \frac{1}{\Delta}\int_{z-\Delta/2}^{z+\Delta/2} |c'|^2 \end{aligned}$$

and for any $Z_0 > 0$,

$$\begin{aligned} r_*[c](Z_0,\epsilon,\Delta) &= \inf_{0\leq z\leq Z_0} r[c](z,\epsilon,\Delta), \\ r^*[c](Z_0,\epsilon,\Delta) &= \sup_{0\leq z\leq Z_0} r[c](z,\epsilon,\Delta), \\ \bar{r}^*[c](Z_0,\Delta) &= \sup_{0\leq z\leq Z_0} \bar{r}[c](z,\Delta). \end{aligned}$$

The main step in the proof of Theorem 2.1 is embodied in Theorem 6.1.

THEOREM 6.1. *There exist constants $\bar{m}, M_1, M_2, \epsilon_0$, and $\Delta_0 > 0$ depending on $T_0, T_1, T_2, \ell^*, K_0, K_1$, and K^* so that if*

(i) *$f \in \mathcal{E}'$ satisfies (4.1) and $m_f^1 \leq \bar{m}$;*

(ii) *$\tilde{v} \in \sum$ is consistent with $D \in L^2(R_1)$, i.e., $S[\tilde{v}] = D$*

(iii) *$\tilde{v} \in \sum$ is coherent, $\tilde{v} = \tilde{v}[c]$ for $c \in H^1_{\text{loc}}(\mathbf{R})$;*

(iv) *For some ϵ, Δ, $Z_0 > 0$ with $\epsilon \leq \epsilon_0$, $\Delta \leq \Delta_0$, and $\tau(Z_0, P_1) = 2\int_0^{Z_0} dz \left(1/c^2(z) - p_1^2\right)^{1/2} \geq T_0$ the following inequalities hold: $r_*[c](Z_0,\epsilon,\Delta) \geq M_1$, $M_2 r_*[c](Z_0,\epsilon,\Delta) \geq \max(r^*[c](Z_0,\epsilon,\Delta),\ \bar{r}[c](Z_0,\Delta))$.*

Then there exists $\mu > 0$ depending on $T_0, T_1, T_2, \ell^*, K_0, K_1, \sigma$, and λ so that for $\delta\tilde{v} \in \dot{\sum}$,

$$\langle \delta\tilde{v}, \text{Hess}\, J_{\sigma,\lambda}\delta\tilde{v}\rangle_{H_1(\tilde{R}_2)} \geq \mu\|\delta\tilde{v}\|^2_{H^1(\tilde{R}_2)}.$$

Before giving the proof of Theorem 6.1, we digress to demonstrate the meaning of condition (iv) in the theorem, and show that the set of $\tilde{v}$ satisfying it is nonempty.

Let $c_0, \delta c_1 \in C^\infty(\mathbf{R}) \cap L^\infty(\mathbf{R})$, so that

$$0 < r_* \leq \int_{z-1}^{z+1} |h_1' * \delta c_1'|^2 \leq r^*, \qquad -\infty \leq z \leq \infty$$

and take $\bar{r} = \sup_{z \in \mathbf{R}} \int_{z-1}^{z+1} |\delta c_1'|^2$. Clearly we can choose δc_1 so that $r^* \geq r_*$, $\bar{r}$ achieve any prescribed values. On the other hand, set

$$\delta c_\epsilon(z) = 2\epsilon \delta c_1(\frac{z}{\epsilon}), \qquad c_\epsilon(z) = c_0(z) + \delta c_\epsilon(z)$$

for $0 \leq z \leq Z_0$, cut off to constants elsewhere. Then $\{\log c_\epsilon\}_{\epsilon \leq \epsilon_0}$ is bounded in $H^1[0, Z_0]$, for suitable ϵ_0, but

$$r_* - \mathcal{O}(\epsilon) \leq \frac{1}{2\epsilon} \int_{z-\epsilon}^{z+\epsilon} |\epsilon h_\epsilon' * c_\epsilon'|^2 \leq r^* + \mathcal{O}(\epsilon),$$

$$\frac{1}{2\epsilon} \int_{z-\epsilon}^{z+\epsilon} |c_\epsilon'| = \bar{r} + \mathcal{O}(\epsilon).$$

Thus with the choice $\Delta = 2\epsilon$, the quantities

$$r_*[c_\epsilon](Z_0, \epsilon, 2\epsilon),$$
$$r^*[c_\epsilon](Z_0, \epsilon, 2\epsilon), \qquad \bar{r}[c_\epsilon](Z_0, 2\epsilon)$$

stand in more or less fixed proportions throughout the set $\{c_\epsilon\}$. We conclude that, whatever the values of the constants $M_1, M_2, \bar{\epsilon}, \bar{\Delta}$ specified in Theorem 6.1, the set of travel-time velocities satisfying condition (iv) is nonempty—simply take $\tilde{v} = \tilde{v}[c_\epsilon]$ for sufficiently small ϵ, and a suitable choice of $c_0, \delta c_1$.

Also the meaning of condition (iv) is clear from this construction. As $\epsilon \to 0$, the perturbation δc_ϵ becomes smaller, but its derivative has uniformly bounded (above and below) mean square over intervals of length ϵ, and this even after convolution with the oscillatory kernel $\epsilon h_\epsilon'$. Thus c_ϵ has significant oscillation everywhere on the scale ϵ, i.e., c_ϵ is "uniformly rough."

In the estimates which follow, we will write for convenience

$$\delta Q = DQ[\tilde{v}]\delta\tilde{v}, \qquad \delta\tilde{S} = D\tilde{S}[\tilde{v}]\delta\tilde{v}.$$

Also, for any function u of (t,p) or (z,p) we shall denote by $\int u$ the function

$$(t,p) \mapsto \int_0^t dt' u(t',p) \quad \text{or} \quad (z,p) \mapsto \int_0^z dz' u(z',p).$$

To begin the proof of Theorem 6.1, note that

$$\langle \delta\tilde{v}, \operatorname{Hess} J_{\sigma,\lambda}[\tilde{v}] \cdot \delta\tilde{v} \rangle = \|\delta\tilde{S}\|^2_{L^2(R_1)} + \sigma^2 \|\delta Q\|^2_{L^2(\tilde{R}_2)} + \lambda^2 \left\| \frac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2(\tilde{R}_2 \setminus \tilde{R}_0)} \tag{6.1}$$

because of the consistency and coherency assumptions.

The second term in (6.1) is the most interesting. It follows immediately from the definition of Q that

$$\begin{aligned}\delta Q &= \frac{\partial \delta\tilde{v}}{\partial p} - 2p\left(\int \tilde{v}^2\right)\frac{\partial \delta\tilde{v}}{\partial x} - 4p\left(\int \tilde{v}\delta\tilde{v}\right)\frac{\partial \tilde{v}}{\partial x} - 3p\tilde{v}^2\delta\tilde{v} \\ &= \left(\frac{\partial}{\partial p}(\delta\tilde{v}\circ\tau)\right)\circ\zeta - 4p\left(\int \tilde{v}\delta\tilde{v}\right)\frac{\partial \tilde{v}}{\partial x} - 3p\tilde{v}^2\delta\tilde{v} \\ &= \left\{\frac{\partial \delta v}{\partial p} - \frac{p}{2}\left(\int dz\,\delta v\right)\frac{\partial v^2}{\partial z} - 3pv^2\delta v\right\}\circ\zeta\end{aligned}$$

where we have written for convenience $\delta v = \delta\tilde{v}\circ\tau$. Note that δv is *not* the perturbation in v resulting from a perturbation in c.

Choose a test kernel $g \in C_0^\infty(\mathbf{R}_z)$ with $\|g\|_{L^1(\mathbf{R})} = 1$ and $\operatorname{supp} g \subset [0,\infty)$. Then

$$g * (\delta Q \circ \tau) = \frac{\partial}{\partial p} g * \delta v - \frac{1}{2} p\left(\int dz\,\delta v\right) g * \frac{\partial v^2}{\partial z} + E_1 - 3pg * (v^2\delta v).$$

The error term E_1 is the commutator of a multiplication operator and convolution with g:

$$E_1 = p\left\{\left(\int dz\,\delta v\right) g * \frac{\partial v^2}{\partial z} - g * \left(\left(\int dz\,\delta v\right)\frac{\partial v^2}{\partial z}\right)\right\}$$

for which a standard estimate gives

$$\|E_1(\cdot,p)\|_{L^2[0,T_2]} \le C m_g^1 \|\delta v(\cdot,p)\|_{L^\infty(\mathbf{R})}.$$

(We have used the notation

$$m_g^k = \int_{\mathbf{R}} dz\, z^k |g(z)|$$

for the moments of $|g|$ as explained before.)

Set

$$K_g = -\frac{1}{2} p\left(\int dz\,\delta v\right) g * \frac{\partial v^2}{\partial z}.$$

The next goal is an L^2 estimate for K_g on the domain $R_Z = [0,Z] \times [P_1, P_2]$. Recall that Z is chosen so that

$$T_0 \le \tau(Z,P_2) < \tau(Z,P_1) \le T_2.$$

First note that for its indefinite p-integral,

$$\int_{p_1}^{p_2} dp\, K_g = \int_{p_1}^{p_2} g * (\delta Q \circ \tau) - g * \delta v\Big|_{p=p_1}^{p_2} - \int_{p_1}^{p_2} dp\, E_1.$$

Since composition with τ and ζ, the indefinite p-integral, and z-convolution with g are all bounded operators on $L^2_{\mathrm{loc}}(\mathbf{R} \times [p_1,p_2])$,

$$\begin{aligned}\left\|\int_{p_1}^{p_2} dp K_g\right\|_{L^2[0,Z]} \le\ & C\Big\{\|g * \delta v(\cdot,p_1)\|_{L^2[0,Z]} + \|g * \delta v(\cdot,p_2)\|_{L^2[0,Z]} \\ & + \|\delta Q\|_{L^2(\tilde{R}_2)} + m_g^1\left\|\frac{\partial \delta v}{\partial z}\right\|_{L^2(R_Z)}\Big\}.\end{aligned}$$

Now assume that $\int g = 0$ and choose $\psi \in C_0^\infty(\mathbf{R})$ with $\psi \equiv 1$ on $[0, Z]$. For $p \in [P_1, P_2]$, denote by $\delta\tilde{v}$ the extension δv by a constant for $z > Z$. Then (since $\operatorname{supp} g \subset \mathbf{R}^+$)

$$\begin{aligned}
\|g * \delta v(\cdot, p)\|^2_{L^2[0,Z]} &\leq \|g * \psi\delta\bar{v}(\cdot, p)\|^2_{L^2[\mathbf{R}]} \\
&= \frac{1}{\sqrt{2\pi}} \int dk |\hat{g}(k)|^2 \widehat{\psi\delta\bar{v}}(k)|^2 \\
&= \frac{1}{\sqrt{2\pi}} \int dk \frac{|\hat{g}(k)|^2}{|k|^2} |ik\widehat{\psi\delta\bar{v}}(k)|^2 \\
&\leq \left(\sup_k \frac{|\hat{g}(k)|}{|k|}\right)^2 \left\| \frac{\partial}{\partial z}(\psi\delta\bar{v}(\cdot, p)) \right\|^2_{L^2(\mathbf{R})} \\
&\leq C \left(\sup_k \frac{|\hat{g}(k)|}{|k|}\right)^2 \left\| \frac{\partial \delta v}{\partial z}(\cdot, p) \right\|^2_{L^2([0,Z])},
\end{aligned}$$

since $\delta\bar{v}$ is constant for $z > Z$ and $\delta v(0, p) \equiv 0$.

Define

$$\epsilon_g := \sup_k \frac{|\hat{g}(k)|}{|k|}.$$

From the bound

$$\left\| \frac{\partial \delta v}{\partial z}(\cdot, p) \right\|_{L^2[0,Z]} \leq C \left\| \frac{\partial \delta\tilde{v}}{\partial x}(\cdot, p) \right\|_{L^2[0,T_2]}$$

valid for $\delta\tilde{v} \in \dot{\sum}$, we get for any $P_1 \leq p_1 \leq p_2 \leq P_2$,

$$\begin{aligned}
\left\| \int_{p_1}^{p_2} dp\, K_g \right\|^2_{L^2[0,Z]} \leq C \Bigg\{ \epsilon_g \Bigg[&\left\| \frac{\partial \delta\tilde{v}}{\partial x}(\cdot, p_1) \right\|^2_{L^2[0,T_2/2]} + \left\| \frac{\partial \delta\tilde{v}}{\partial x}(\cdot, p_2) \right\|^2_{L^2[0,T_2/2]} \Bigg] \\
&+ \|\delta Q\|^2_{L^2(\tilde{R}_2)} + m_g^1 \left\| \frac{\partial \delta\tilde{v}}{\partial x} \right\|^2_{L^2(\tilde{R}_2)} \Bigg\}. \qquad (6.2)
\end{aligned}$$

Next we estimate the p-derivative of K_g:

$$\begin{aligned}
\frac{\partial K_g}{\partial p} = -\frac{1}{2} \Bigg[&\left(\int dz\, \delta v \right) g * \frac{\partial v^2}{\partial z} \\
&+ p \left(\int dz \frac{\partial \delta v}{\partial p} \right) g * \frac{\partial v^2}{\partial z} + p \left(\int dz\, \delta v \right) g * \frac{\partial^2 v^2}{\partial p\, dz} \Bigg].
\end{aligned}$$

The first term is clearly dominated in $L^2[0, Z]$ for each p by a ($\dot{\sum}$-dependent) multiple of

$$\|\delta\tilde{v}(\cdot, p)\|_{L^2[0,T_2/2]}.$$

From the definition of δQ,

$$\int dz \frac{\partial \delta v}{\partial p} = \int dz \left\{ \delta Q \circ \tau - \frac{1}{2} p \left(\int dz\, \delta v \right) \frac{\partial v^2}{\partial z} - 3p \int v^2 \delta v \right\},$$

which is bounded by a ($\sum$-dependent) multiple of

$$\|\delta Q(\cdot, p)\|_{L^2[0,T_2]} + \|\delta \tilde{v}(\cdot, p)\|_{L^2[0,T_2/2]}.$$

Finally, we note that

$$v = \frac{c}{\sqrt{1 - c^2 p^2}}$$

for some c with

$$\|c\|_{H^1[0,Z]} \leq C$$

so that for any $p \in [P_1, P_2]$,

$$\left\| \frac{\partial^2 v^2}{\partial z \partial p}(\cdot, p) \right\|_{L^2[0,Z]} \leq C.$$

The upshot is the estimate

$$\left\| \frac{\partial K_g}{\partial p}(\cdot, p) \right\|_{L^2[0,Z]} \leq C \left\{ \|\delta Q(\cdot, p)\|_{L^2[0,T_2/2]} + \|\delta \tilde{v}(\cdot, p)\|_{L^2[0,T_2/2]} \right\}. \tag{6.3}$$

Integrating in p, we also get

$$\left\| \frac{\partial K_g}{\partial p} \right\|_{L^2(R_Z)} \leq C(\|\delta Q\|_{L^2(\tilde{R}_2)} + \|\delta \tilde{v}\|_{L^2(\tilde{R}_2)}). \tag{6.4}$$

Now we combine (6.2) and (6.4) to estimate K_g via a simple interpolation inequality, which is a special case of Gilbarg and Trudinger (1983, Theorem 7.28):

For $u \in H^2[a,b], \alpha > 0$:

$$\|u'\|^2_{L^2[a,b]} \leq \alpha \|u''\|^2_{L^2[a,b]} + \frac{C}{\alpha} \|u\|^2_{L^2[a,b]} \tag{6.5}$$

where $C = C(|b-a|)$.

Apply (6.5) to $u(p) = \int_{p_0}^{p} dp' K_g(z,p)$ for arbitrary $p_0 \in [P_1, P_2]$, integrate the result in z, and use (6.2) and (6.4) to obtain

$$\begin{aligned} \|K_g\|^2_{L^2(R_Z)} \leq\ & \frac{C}{\alpha} \left\{ \epsilon_g^2 \left([P_2 - P_1] \left\| \tfrac{\partial \delta \tilde{v}}{\partial x}(\cdot, p_0) \right\|^2_{L^2[0,T_2/2]} + \left\| \tfrac{\partial \delta \tilde{v}}{\partial x} \right\|^2_{L^2(\tilde{R}_2)} \right) \right. \\ & \left. + (m_g^1)^2 \left\| \tfrac{\partial \delta \tilde{v}}{\partial x} \right\|^2_{L^2(\tilde{R}_2)} + \|\delta Q\|^2_{L^2(\tilde{R}_2)} \right\} \\ & + \alpha C' \left\{ \|\delta Q\|^2_{L^2(\tilde{R}_2)} + \|\delta \tilde{v}\|^2_{L^2(\tilde{R}_2)} \right\}. \end{aligned}$$

With the choice

$$\alpha^* = \max(\epsilon_g, m_g^1), \qquad \alpha_* = \min(\epsilon_g, m_g^1),$$

this becomes after integration in p_0 from P_1 to P_2

$$\begin{aligned} ||K_g||^2_{L^2(R_Z)} &\leq C\tfrac{(\alpha^*)^2}{\alpha_*}\left\{\left\|\tfrac{\partial \delta\tilde{v}}{\partial x}\right\|^2_{L^2(\tilde{R}_2)} + \|\delta\tilde{v}\|^2_{L^2(\tilde{R}_2)}\right\} \\ &\quad + C'\left(\alpha^* + \tfrac{1}{\alpha_*}\right)\|\delta Q\|^2_{L^2(\tilde{R}_2)}. \end{aligned} \tag{6.6}$$

The next step is to show that K_g actually dominates the indefinite integral of $\delta\tilde{v}$. It is at this point that some constraints on $\tilde{v}$ (hence on c), other than coherency and membership in $\sum$, become necessary. Recall that

$$K_g = -\frac{1}{2}p\left(\int dz\, \delta v\right) g * \frac{\partial v^2}{\partial z}.$$

Thus K_g is the product of a relatively smooth factor (the indefinite integral) and a relatively rough factor (the derivative). Clearly, some estimate of the smooth factor must be possible, provided that the rough factor is sufficiently *uniformly* rough. The following simple lemma gives a crude criterion of this type.

LEMMA 6.2. *Suppose that* $u, \Phi \in C^\infty(\mathbf{R})$. *Set for* $\Delta > 0$, $a < b$

$$r(x,\Delta) = \frac{1}{\Delta}\int_{x-\Delta/2}^{x+\Delta/2} |u|^2,$$

$$r_*(\Delta) = \inf_{x\in[a,b]} r(x,\Delta), \qquad r^*(\Delta) = \sup_{x\in[a,b]} r(x,\Delta).$$

Then for any $\Delta > 0$ (L^2-*norms*) :

$$\|\Phi u\|^2_{L^2[a,b]} \geq \frac{r_*(\Delta)}{2}\|\Phi\|^2_{L^2[a,b]} - \frac{16}{9}(r_*(\Delta) + r^*(\Delta))\Delta^2\|\Phi'\|^2_{L^2[a,b]}.$$

Proof. Set

$$\bar{\Phi}_\Delta(x) = \frac{1}{\Delta}\int_{x-\Delta/2}^{x+\Delta/2} \Phi.$$

Then the Cauchy–Schwarz inequality gives

$$|\Phi(y) - \overline{\Phi}_\Delta(x)| \leq \frac{4}{3}\Delta^{1/2}\int_{x-\Delta/2}^{x+\Delta/2} |\Phi'|^2$$

for $x - \Delta/2 \leq y \leq x + \Delta/2$. Thus

$$\begin{aligned} \int_{x-\Delta/2}^{x+\Delta/2} |\Phi u|^2 &\geq \left[|\bar{\Phi}_\Delta(x)^2 - \frac{16}{9}\Delta\int_{x-\Delta/2}^{x+\Delta/2} |\Phi'|^2\right]\int_{x-\Delta/2}^{x+\Delta/2} |u|^2 \\ &\geq \Delta|\bar{\Phi}_\Delta(x)|^2 r_*(\Delta) - \frac{16}{9}\Delta^2 r^*(\Delta)\int_{x-\Delta/2}^{x+\Delta/2} |\Phi'|^2. \end{aligned}$$

Similarly,

$$\Delta|\bar{\Phi}_\Delta(x)|^2 \geq \frac{1}{2}\int_{x-\Delta/2}^{x+\Delta/2} |\Phi|^2 - \frac{16}{9}\Delta^2\int_{x-\Delta/2}^{x+\Delta/2} |\Phi'|^2.$$

Thus

$$\int_{x-\Delta/2}^{x+\Delta/2} |\Phi u|^2 \geq \frac{r_*(\Delta)}{2} \int_{x-\Delta/2}^{x+\Delta/2} |\Phi|^2 - \frac{16}{9}(r_*(\Delta) + r^*(\Delta))\Delta^2 \int_{x-\Delta/2}^{x+\Delta/2} |\Phi'|^2.$$

Now sum both sides over $x = (k + \frac{1}{2})\Delta$, $k = 0, \ldots [(b-a)/\Delta]$ to obtain the result. □

We shall apply Lemma 6.2 to K_g, with the identifications

$$\Phi \sim -\frac{1}{2}p \int \delta v(\cdot, p), \qquad u \sim g * \frac{\partial v^2}{\partial z}(\cdot, p).$$

Note that

$$\frac{\partial v^2}{\partial z} = (1 - c^2p^2)^{-2} c \frac{\partial c}{\partial z}$$

so that

$$g * \frac{\partial v^2}{\partial z} = (1 - c^2p^2)^{-2} cg * \frac{\partial c}{\partial z} + E_2.$$

We will assume that the length scale Δ is chosen so that supp $g \subset [\Delta/2, \Delta/2]$. Then a slight refinement of the standard error estimate gives

$$\frac{1}{\Delta} \int_{z-\Delta/2}^{z+\Delta/2} |E_2| \leq C\Delta^2 r_\delta^*(\Delta) a_g(\Delta)$$

where we have written, for any compactly supported measure h, and $Z_0 > 0$ to be determined:

$$\begin{aligned}
r_h^*(\Delta) &= \sup_{z \in [0, Z_0]} \frac{1}{\Delta} \int_{z-\Delta/2}^{z+\Delta/2} \left| h * \frac{dc}{dz} \right|^2, \\
r_{h,*}(\Delta) &= \inf_{z \in [0, Z_0]} \frac{1}{\Delta} \int_{z-\Delta/2}^{z+\Delta/2} \left| h * \frac{dc}{dz} \right|^2, \\
a_h(\Delta) &= \sup_{z \in [0, Z_0]} \frac{1}{\Delta} \int_{z-\Delta/2}^{z+\Delta/2} \left(|h| * \left| \frac{dc}{dz} \right| \right)^2.
\end{aligned}$$

For various $\sum$-dependent constants C_1–C_4, any $z \in [0, Z_0]$,

$$\begin{aligned}
C_1 r_{g,*}(\Delta) - C_2\Delta^2 r_\delta^*(\Delta) a_g(\Delta) &\leq \frac{1}{\Delta} \int_{z-\Delta/2}^{z+\Delta/2} \left| g * \frac{\partial v^2}{\partial z} \right|^2 \\
&\leq C_3 r_g^*(\Delta) + C_4\Delta^2 r_g^*(\Delta) a_g(\Delta).
\end{aligned}$$

Accordingly Lemma 6.2 implies, so long as $Z_0 \leq Z$, for $p_0 \in [P_1, P_2]$

$$\begin{aligned}
\|K_g(\cdot, p)\|^2_{L^2[0,Z]} &\geq Cp^2 r_{g,*}(\Delta) \left\| \int dz\, \delta v(\cdot, p) \right\|^2_{L^2[0,Z_0]} \\
&\quad - C'(r_{g,*}(\Delta) + r_g^*(\Delta) + r_g^*(\Delta) a_g(\Delta))\Delta^2 \|\delta v(\cdot, p)\|^2_{L^2[0,Z_0]}.
\end{aligned} \tag{6.7}$$

These estimates have no force unless the quantities $r_{g,*}(\Delta)$ and

$$r_t^*(\Delta) = (r_{g,*}(\Delta) + r_g^*(\Delta) + r_g^*(\Delta)a_g(\Delta))$$

are comparable. These quantities depend on the parameters Δ and Z_0 and on the kernel g (all of which are still to be chosen)—and, of course, on the velocity profile c.

To further manipulate the norms of $\delta v, \delta\tilde{v}$, we require an estimate on the p-derivative: after some algebra,

$$\begin{aligned}
&\frac{d}{dp}\left\|\int dz\, \delta v(\cdot,p)\right\|^2_{L^2[0,Z_0]} \\
&\quad = 2\left\langle \int dz\, \delta v(\cdot,p), \int dz \left(\delta C \circ \tau + \frac{1}{2}p\left(\int dz\, \delta v\right)\frac{\partial v^2}{\partial z} + 3pv^2\delta v\right)\right\rangle_{L^2[0,Z_0]} \\
&\quad \le C\left\|\int dz\, \delta v(\cdot,p)\right\|^2_{L^2[0,Z_0]} + C'\left\|\delta Q \circ \tau(\cdot,p)\right\|^2_{L^2[0,Z_0]}.
\end{aligned}$$

In the following lemma we have employed an estimate on norms of products.

LEMMA 6.3. *For $f \in H^1[a,b]$, $g \in L^2[a,b]$, define*

$$\int_a^z fg = h(z), \qquad a \le z \le b.$$

Then

$$\|h\|_0 \le C\|f\|_1 \left\|\int_a g\right\|_0$$

for a constant C depending on a, b.

Proof. Since

$$h(z) = f(z)\int_a^z g - \int_a^z dz' f'(z')\int_a^{z'} g$$

we have

$$\begin{aligned}
\|h\|_0 &\le \|f\|_{L^\infty}\left\|\int g\right\|_0 + \left(\int_a^b dz\left[\int_a^z dz' f'(z')\int_0^{z'} g\right]^2\right) \\
&\le C\|f\|_1\left\|\int g\right\|_0 + \int_a^b dz\left[\int_a^z dz'(f'(z'))^2\right]\left[\int_a^z dz'\left[\int_0^{z'} g\right]^2\right],
\end{aligned}$$

whence the required inequality follows. □

Thus Gronwall's inequality gives

$$\left|\left\|\int dz\, \delta v(\cdot,p)\right\|^2_{L^2[0,Z_0]} - \left\|\int dz\, \delta v(\cdot,p_0)\right\|^2_{L^2[0,Z_0]}\right| \le C\|\delta Q\|^2_{L^2(\tilde{R}_2)}$$

for any $p_0 \in [P_1, P_2]$. Thus

$$\int_{P_1}^{P_2} dp\, p^2 \left\| \int dz\, \delta v(\cdot,p) \right\|^2_{L^2[0,Z_0]} \geq \left(\left\| \int dz\, \delta v(\cdot,p_0) \right\|^2_{L^2[0,Z_0]} - C\|\delta Q\|^2_{L^2(\tilde{R}_2)} \right) \frac{P_2^3 - P_1^3}{3}$$

so that (6.7) implies, after integrating the preceding inequality in p_0 from P_1 to P_2,

$$\begin{aligned} C\|K_g\|^2_{L^2(R_Z)} \;\geq\; & C' r_{g,*}(\Delta) \| \int dz\, \delta v\|^2_{L^2([0,Z_0]\times[P_1,P_2])} \\ & - C'' \left\{ \|\delta Q\|^2_{L^2(\tilde{R}_2)} + r_t^*(\Delta)\Delta^2 \|\delta v\|^2_{L^2([0,Z_0]\times[p_1,p_2])} \right\}. \end{aligned}$$

Now concatenate the above inequality with (6.6) to get

$$\begin{aligned} & r_{g,*}(\Delta) \left\| \int \delta v \right\|^2_{L^2([0,Z_0]\times[P_1,P_2])} - C'' \left\{ \|\delta Q\|^2_{L^2(\tilde{R}_2)} + r_t^*(\Delta)\Delta^2 \|\delta v\|^2_{L^2([0,Z_0]\times[P_1,P_2])} \right\} \\ & \leq C \frac{(\alpha^*)^2}{\alpha_*} \|\delta\tilde{v}\|_{H^1(\tilde{R}_2)} + C' \left(\alpha^* + \frac{1}{\alpha_*} \right) \|\delta Q\|^2_{L^2(\tilde{R}_2')}. \end{aligned}$$

Next recall that the L^2-norms of δv and $\delta\tilde{v}$ are related by

$$\|\delta v\|_{L^2(R_Z)} \leq C\|\delta\tilde{v}\|_{L^2(\tilde{R}_2)}$$

as noted at the beginning of the section, which allows us to simplify the above inequality to

$$\begin{aligned} r_{g,*}(\Delta) \| \int dz\, \delta v\|^2_{L^2([0,Z_0]\times[P_1,P_2])} \leq\; & C \left(\frac{(\alpha^*)^2}{\alpha_*} + r_t^*(\Delta)\Delta^2 \right) \|\delta\tilde{v}\|^2_{H^1(\tilde{R}_2)} \\ & + C' \left(1 + \alpha^* + \frac{1}{\alpha_*} \right) \|\delta Q\|^2_{L^2(\tilde{R}_2)}. \end{aligned} \tag{6.8}$$

It is convenient at this point to put the left-hand side in terms of $\delta\tilde{v}$ also, by means of Lemma 6.3. Since

$$\int_0^z dz' \delta v(z',p) = \int_0^{\tau(z,p)} dx\, \tilde{v}(x,p)\delta\tilde{v}(x,p)$$

we can apply Lemma 6.3 with $f = (\tilde{v}(\cdot,p))^{-1}$, $g = \tilde{v}(\cdot,p)\delta\tilde{v}(\cdot,p)$, to get

$$\left\| \int \delta\tilde{v}(\cdot,p_0) \right\|_{L^2[0,\tau(Z_0,p_0)]} \leq C \left\| \int \delta v(\cdot,p_0) \right\|_{L^2[0,Z_0]}$$

whence (6.8) implies, after integration in p_0 from P_1 to P_2,

$$\begin{aligned} r_{g,*}(\Delta) \left\| \int \delta\tilde{v} \right\|^2_{L^2(A_3)} \;\leq\; & C \left(\frac{(\alpha^*)^2}{\alpha_*} + r_t^*(\Delta)\Delta^2 \right) \|\delta\tilde{v}\|^2_{H^1(\tilde{R}_2)} \\ & + C' \left(1 + \alpha^* + \frac{1}{\alpha_*} \right) \|\delta Q\|^2_{L^2(\tilde{R}_2)}. \end{aligned} \tag{6.9}$$

Here A_3 is the region

$$\{(x,p) : P_1 \leq p \leq P_2, \quad 0 \leq x \leq \tau(Z_0,p)\}.$$

At this point, we determine Z_0: set

$$Z_0 = \zeta\left(\frac{1}{2}T_0, P_1\right) = \inf_{p \in P_1, P_2} \zeta\left(\frac{1}{2}T_0, p\right).$$

Thus

$$A_3 \subset \tilde{R}_0.$$

The interpolation inequality (6.5) and (6.9) now yield

$$\begin{aligned} \|\delta\tilde{v}\|^2_{L^2(A_3)} \quad \leq \quad & \left[\beta + \frac{C}{\beta r_{g,*}(\Delta)}\left(\frac{(\alpha^*)^2}{\alpha_*} + r_t^*(\Delta)\Delta^2\right)\right] \|\delta\tilde{v}\|^2_{H^1(\tilde{R}_2)} \\ & + \frac{C'(1 + \alpha^* + (1/\alpha_*))}{\beta r_{g,*}(\Delta)} \|\delta Q\|^2_{L^2(\tilde{R}_2)}. \end{aligned} \tag{6.10}$$

It follows immediately from the Dirichlet condition at $x = T_2$ in the definition of $\sum$ that

$$\|\delta\tilde{v}\|^2_{H^1(\tilde{R}_2)} \leq \|\delta\tilde{v}\|^2_{H^1(\tilde{R}_0)} + C\left\|\frac{\partial\delta\tilde{v}}{\partial x}\right\|^2_{L^2(\tilde{R}_2\setminus\tilde{R}_0)}.$$

It is only slightly more difficult to estimate

$$\|\delta\tilde{v}\|^2_{L^2(\tilde{R}_1\setminus A_3)} \leq C\left\{\left\|\frac{\partial\delta\tilde{v}}{\partial x}\right\|^2_{L^2(\tilde{R}_2\setminus\tilde{R}_0)} + \|\delta Q\|^2_{L^2(\tilde{R}_2)}\right\}. \tag{6.11}$$

In fact, from the inequality

$$\left|\frac{\partial\delta v}{\partial p}\right|^2 + \left|\frac{\partial\delta v}{\partial x}\right|^2 \leq C\left(\left|\frac{\partial\delta v}{\partial x}\right|^2 + |\delta Q|^2\right),$$

which holds at every point in $\tilde{R}_2$, and the already-used inequality

$$\|\delta\tilde{v}\|_{L^2(\tilde{R}_2\setminus\tilde{R}_0)} \leq \sqrt{T_2 - T_0}\left\|\frac{\partial\delta v}{\partial x}\right\|_{L^2(\tilde{R}_2\setminus\tilde{R}_0)}$$

(Dirichlet condition at $x = T_2$!) follows the estimate

$$\|\delta\tilde{v}\|_{H^1(\tilde{R}_2\setminus\tilde{R}_0)} \leq C\left(\left\|\frac{\partial\delta v}{\partial x}\right\|_{L^2(\tilde{R}_2\setminus\tilde{R}_0)} + \|\delta Q\|_{L^2(\tilde{R}_2)}\right)$$

whence from the trace theorem

$$\|\delta\tilde{v}(\cdot, P_1)\|_{L^2[T_0,T_2]} \leq C\left\{\left\|\frac{\partial\delta\tilde{v}}{\partial x}\right\|_{L^2(\tilde{R}_2\setminus\tilde{R}_0)} + \|\delta Q\|_{L^2(\tilde{R}_2)}\right\}.$$

Any point in $\tilde{R}_1\setminus A_3$ lies on a curve of the family

$$p \mapsto \tau(z, p),$$

which joins it to the segment $[T_0, T_2] \times \{P_1\}$, and δQ is a first-order partial differential operator (acting on $\delta\tilde{v}$) whose principal part is a tangent vector field to this family of curves. Now (6.11) follows from the standard energy estimate for hyperbolic systems.

Combining (6.11) with (6.10), we get

$$\begin{aligned} \|\delta\tilde{v}\|^2_{L^2(\tilde{R}_1)} \quad \leq \quad & \left[\beta + \frac{C}{\beta r_{g,*}(\Delta)}\left(\frac{(\alpha^*)^2}{\alpha_*} + r_t^*(\Delta)\Delta^2\right)\right] \|\delta\tilde{v}\|^2_{H^1(\tilde{R}_0)} \\ & + C'\left[\|\delta Q\|^2_{L^2(\tilde{R}_2)} + \left\|\frac{\partial \delta\tilde{v}}{\partial x}\right\|^2_{L^2(\tilde{R}_2 \setminus \tilde{R}_0)}\right] \end{aligned}$$

where C' depends on $\beta, \alpha_*, \alpha^*$, and $r_{g,*}$ as well. Now recall the conclusion of Theorem 5.2 which we rewrite in a suggestive way:

$$\|\delta\tilde{S}\|^2_{L^2(R_1)} + (m_g^1)^2 L_2^2 \left\|\frac{\partial \delta\tilde{v}}{\partial x}\right\|^2_{L^2(\tilde{R}_1 \setminus \tilde{R}_0)} \geq L_1^2 \|\delta\tilde{v}\|^2_{H^1(\tilde{R}_0)} - (L_0^2 + L_1^2)\|\delta\tilde{v}\|^2_{L^2(\tilde{R}_1)}.$$

Evidently, Theorem 6.1 has been proved if we can make the various choices deferred above so that

$$(L_0^2 + L_1^2)\left[\beta + \frac{C}{\beta r_{g,*}(\Delta)}\left(\frac{(\alpha^*)^2}{\alpha_*} + r_t^*(\Delta)\Delta^2\right)\right] < L_1^2. \tag{6.12}$$

To understand the left-hand side of (6.12) we first examine $\alpha = \max(\epsilon_g, m_g^1)$. We relieve the reader's suspense by identifying the test kernel g with the kernel $\epsilon h'_\epsilon$ appearing in the statement of the theorem, so that the issue becomes one of choosing $\epsilon > 0$. Note that with this identification

$$m_g^1 = \int dz |z\epsilon h'_\epsilon(z)| = \mathcal{O}(\epsilon),$$

and similarly,

$$\epsilon_g = \mathcal{O}(\epsilon).$$

On the other hand, from the definitions,

$$\begin{aligned} r_g^*(\Delta) &= r^*(Z_0, \Delta, \epsilon), \\ r_{g,*}(\Delta) &= r_*(Z_0, \Delta, \epsilon), \\ r_g^*(\Delta) &= \bar{r}(Z_0, \Delta) \end{aligned}$$

while Young's inequality gives

$$a_g(\Delta) \leq C\bar{r}(Z_0, \Delta)$$

provided that $\epsilon \leq \Delta$, say, so that the support of h_ϵ is contained within a fixed number of Δ-intervals. In fact, we shall adopt the convention that $\epsilon \leq C\Delta^2$ and assume that Δ is sufficiently small that $\epsilon \leq \Delta$ as well. Then

$$r_{t,*}(\Delta) \leq C(r_*(Z_0, \epsilon, \Delta) + r^*(Z_0, \epsilon, \Delta) + \bar{r}(Z_0, \Delta)^2)$$

where the constant now depends on the test wavelet h_1 as well as on $\sum$—this will be the default dependence for constants "C" for the rest of the argument.

Note the uniform (over $c \in H^1_{\mathrm{loc}}$ corresponding to $\tilde{v} \in \sum$) constraints, for all $\epsilon, \Delta > 0$:

$$r_*(Z_0, \epsilon, \Delta) \le r^*(Z_0, \epsilon, \Delta) \le C,$$
$$\bar{r}(Z_0, \Delta) \le C,$$
$$r^*(Z_0, \epsilon, \Delta) \le C\bar{r}^*(Z_0, \Delta).$$

Choose a lower bound M_1 and a relative lower bound M_2, in the statement of Theorem 6.1, subject to these constraints, and assume that c satisfies condition (iv) of the theorem (recall that we have already met the constraint on Z_0, viz.

$$T_0 = \tau(Z_0, P_1) \ ).$$

That is,

$$\begin{aligned} &M_1 \le r_*(Z_0, \epsilon, \Delta), \\ &\bar{r}(Z_0, \Delta) \le M_2 r_*(Z_0, \epsilon, \Delta), \\ &r^*(Z_0, \epsilon, \Delta) \le M_2 r_*(Z_0, \epsilon, \Delta). \end{aligned} \tag{6.13}$$

Then the ratio

$$\frac{C}{r_{g,*}(\Delta)} \left(\frac{(\alpha^*)^2}{\alpha_*} + r_t^*(\Delta)\Delta^2 \right)$$

is $\mathcal{O}(\epsilon + \Delta^2) = \mathcal{O}(\Delta^2)$ for coherent $\tilde{v}\epsilon \sum$ for which the corresponding velocity profile c satisfies (6.13). Then we can choose $\beta = \mathcal{O}(\Delta)$ so that the left-hand side of (6.12) becomes $\mathcal{O}(\Delta)$ as well. In particular, for sufficiently small Δ, other choices as above, the left-hand side of (6.12) is

$$\le C(L_0^2 + L_1^2)\Delta \quad < \frac{1}{2}L_1^2$$

whence finally

$$\frac{1}{2}L_1^2 \|\delta\tilde{v}\|^2_{H^1(\tilde{R}_0)} \le \|\delta\tilde{S}\|^2_{L^2(R_1)} + C\|\delta Q\|^2_{L^2(\tilde{R}_2)} + C' \left\| \frac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2(\tilde{R}_2 \setminus \tilde{R}_0)}. \tag{6.14}$$

The right-hand side of (6.14) is bounded by a multiple of the Hessian quadratic form, the factor now depending on the penalty constants σ and λ as well. This completes the proof of Theorem 6.1. □

The proof of Theorem 2.1 is now immediate. Given appropriate $T_0, T_1, T_2, K_0, K_1, K^*$, P_1, P_2, c_0, c_1, and c^*, let $\sum_c$ denote the collection of $c \in H^1_{\mathrm{loc}}(\mathbf{R})$ satisfying

(i) $c(z) = c_0, \qquad z < 0,$

(ii) $\|\log c\|_{L^\infty(\mathbf{R})} \le c^* \qquad (\le -\log P_2),$

(iii) $c(z) = c_1 \quad \text{if} \quad 2\int_0^z \sqrt{\frac{1}{c^2} - P_1^2} \ge T_0.$

For some $\ell^* = \ell(c^*)$,

$$\|\log c\|_{H^1[0,Z]} \le c^* \longrightarrow \|\log \tilde{v}[c](\cdot, p)\|_{H^1[0,T_0/2]} \le \ell^*.$$

Then $c \in \sum_c \Rightarrow \tilde{v}[c] \in \sum$, as defined before the statement of Theorem 6.1.

Let $\bar{m}$ be as in the statement of Theorem 5.2 and assume that $f \in \mathcal{E}'(\mathbf{R})$ is chosen to satisfy

$$K^* \left\| \frac{\partial \phi}{\partial x} \right\|_{L^2(\mathbf{R})} \geq \left\| f * \frac{\partial \phi}{\partial x} \right\|_{L^2(\mathbf{R})} \geq K_1 \left\| \frac{\partial \phi}{\partial x} \right\|_{L^2(\mathbf{R})} - K_0 \|\phi\|_{L^2(\mathbf{R})}$$

for $\phi \in H^1(\mathbf{R})$,

$$m_f^1 \leq \bar{m}.$$

Choose a test wavelet h_1 as described above. Then choose $M_1, M_2, \bar{\Delta}$, and $\bar{\epsilon}$ as in the statement of Theorem 6.1 and define

$$\begin{aligned} \sum_c' \quad = \quad & \{c \in \textstyle\sum_c : \text{for } \Delta, \epsilon \text{ with } \ 0 < \Delta \leq \bar{\Delta}, \quad 0 < \epsilon \leq \bar{\epsilon}, \\ & \text{and } Z_0 \text{ satisfying } 2 \textstyle\int_0^{Z_0} \left(1/c_2 - p_1^2\right)^{1/2} = T_0, \\ & \text{the following inequalities hold: } M_1 \leq r_*(Z_0, \epsilon, \Delta) \\ & M_2 r_*(Z_0, \epsilon, \Delta) \geq \max(\bar{r}(Z_0, \Delta), r^*(Z_0, \epsilon, \Delta))\}. \end{aligned}$$

Then c is constant for $z > Z_0$, so $\tilde{v}[c]$ is constant (for each $p \in [P_1, P_2]$) for $t \geq \tau(Z_0, p)$, whence a fortiori for $t \geq T_0$.

The remarks after the statement of Theorem 6.1 show that, for arbitrary (but consistent) choices of the various parameters, the set $\sum_c'$ is nonempty.

Finally, assume that in the definition of $J_{\sigma,\lambda}$,

$$D = S[c] = \tilde{S}[\tilde{v}[c]]$$

for $c \in \sum_c'$. Then

$$J_{\sigma,\lambda}[\tilde{v}[c]] = 0$$

while Theorem 6.1 shows that the Hessian of $J_{\sigma,\lambda}$ at $\tilde{v}[c]$ is positive definite.

Therefore the implicit function theorem implies the existence of

(1) An open neighborhood U in $L^2(R_1)$ of the set

$$\{S[c] : c \in \textstyle\sum_c'\},$$

(2) An open neighborhood V in $\sum$ of the set

$$\{\tilde{v}[c] : c \in \textstyle\sum_c'\},$$

so that for each $D \in U$, the problem

$$\min_{\tilde{v} \in \sum} J_{\sigma,\lambda}[\tilde{v}]$$

has a unique solution $\tilde{v} = \tilde{I}[D] \in V$, which is moreover a Lipschitz continuous function of the data D.

Define the averaging operator

$$A : \sum \longrightarrow L^2_{\text{loc}}(\mathbf{R})$$

by

$$A\tilde{v}(z) = \frac{1}{P_2 - P_1}\int_{P_1}^{P_2} dp \left[\left(\frac{1}{\tilde{v}(\tau(z,p),p)}\right)^2 + p^2\right]^{-1/2}.$$

Remark. A performs a version of the operation "normal moveout correction, stack" from the reflection seismic data processing stream (see, e.g., Yilmaz (1987, §1.4)).

Then for $c \in \sum_c$,

$$A\tilde{v}[c] = c.$$

Also, A is Lipschitz continuous in the topologies indicated in its definition.

Set

$$I = A \circ \tilde{I} : U \to L^2_{\text{loc}}.$$

Then I has all of the properties indicated in the statement of Theorem 2.1. In particular, for $D_1, D_2 \in U$,

$$\|I(D_1) - I(D_2)\|_{L^2[0,Z]} \leq L^* \|D_1 - D_2\|_{L^2(R_1)}$$

for suitable L^*, depending on the various parameters defining $\sum_c'$ and on f. This completes the proof of Theorem 2.1. □

7. Nonelliptic sources. In this section we give a very brief sketch of the state of affairs when f is smooth. The necessary regularization arguments have become quite commonplace, so we shall concentrate on the steps necessary to modify the proof of Theorem 6.1.

Thus, suppose that $f \in C_0^\infty(\mathbf{R})$: then the best "near-elliptic" estimate might have the form

$$\left\| f * \frac{\partial \phi}{\partial x} \right\|_{L^2(\mathbf{R})} \geq K_1 \left\| \frac{\partial \phi}{\partial x} \right\|_{L^2(\mathbf{R})} - K_0 \|\phi\|_{L^2(\mathbf{R})} - K_2 \left\| \frac{\partial^2 \phi}{\partial x^2} \right\|_{L^2(\mathbf{R})} \tag{7.1}$$

for $\phi \in H^2(\mathbf{R})$. The size of K_2 measures the "passband" of f: i.e., if $|\hat{f}(\omega)|$ is uniformly large in an interval $\Omega_\ell \leq |\omega| \leq \Omega_h$, then $K_2 = \mathcal{O}(1/\Omega_h)$.

The analogue of Theorem 5.2 is Theorem 7.1.

THEOREM 7.1. *Given $P_2 > P_1 \geq 0$, $T_1 > T_0 > 0$, $K^* \geq K_1 > 0$, $K_0, K_2 \geq 0$, and $\ell^* > 0$, there exist $\bar{m}, L_0, L_1, L_2, L_3, > 0$, so that if $f \in C_0^\infty(\mathbf{R})$ satisfies (7.1) and $m_f^1 \leq \bar{m}$, then for $\tilde{v} \in \sum \cap H^2_{\text{loc}}(\mathbf{R})$, $\delta\tilde{v} \in \sum \cap H^2_{\text{loc}}(\mathbf{R})$*

$$\|D\tilde{S}[\tilde{v}]\delta\tilde{v}\|_{L^2(R_1)} \geq L_1 \left\| \frac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2(\tilde{R}_0)} - L_0 \|\delta\tilde{v}\|_{L^2(\tilde{R}_1)}$$
$$- m_f^1 L_3 \left\| \frac{\partial \delta\tilde{v}}{\partial x} \right\|_{L^2(\tilde{R}_1 \setminus \tilde{R}_0)} - L_2 \left\| \frac{\partial^2 \delta\tilde{v}}{\partial x^2} \right\|_{L^2(\tilde{R}_1)}.$$

The principal new ingredient in the proof is the higher-order estimate for the planewave problem

$$\|DS_0[\tilde{v}, \delta']\delta\tilde{v}\|_{L^2[0,T]} \leq C_2 \|\delta\tilde{v}\|_{H^2[0,\, T/2]}$$

for $\log \tilde{v} \in H^2_{\text{loc}}(\mathbf{R})$, $\delta\tilde{v} \in H^2_{\text{loc}}(\mathbf{R})$. See, for instance, Suzuki (1988) for similar estimates.

Most of the proof of Theorem 6.1 goes through as before, except that now the smoothness constraint implicit in Theorem 7.1 conflicts with the roughness conditions. For example, for a coherent $\tilde{v} \in \sum \cap H^2_{\text{loc}}(\mathbf{R})$, its corresponding $c \in H^2_{\text{loc}}(\mathbf{R})$ satisfies

$$\begin{aligned} \frac{1}{\Delta}\int_{z-\Delta/2}^{z+\Delta/2} |\epsilon h'_\epsilon * c'|^2 &= \frac{1}{\Delta}\int_{z-\Delta/2}^{z+\Delta/2} |\epsilon h_\epsilon * c''|^2 \\ &\leq C\frac{\epsilon^2}{\Delta}\|c''\|^2_{L^2[z-2\Delta,z+2\Delta]}. \end{aligned}$$

For the constraint $\epsilon = 0(\Delta)$, which we were bound to impose, this gives

$$\leq C\Delta\|c''\|_{L^2[0,Z_0]}.$$

So over any bounded set in $H^2(\tilde{R}_2)$, r_* is $\mathcal{O}(\Delta)$ over coherent travel-time velocities. Thus estimates of the sort developed in § 6 can only succeed if

(i) f is sufficiently "broadband" that (7.1) holds with small K_2 relative to $\sum, K_0$, K_1, K^*;

(ii) Target velocities exist in the intersection of $\sum'_c$ and a sufficiently large ball in $H^2(\tilde{R}_2)$, for which the regularized cost functional

$$\begin{aligned} J_{\sigma,\lambda,\rho}[\tilde{v}]: &= \frac{1}{2}\left\{\|\tilde{S}[\tilde{v}] - D\|^2_{L^2(R_1)} + \sigma^2\|C[\tilde{v}]\|^2_{L^2(\tilde{R}_1)}\right. \\ &\qquad \left. +\lambda^2\left\|\frac{\partial\tilde{v}}{\partial x}\right\|_{L^2(\tilde{R}_2\setminus\tilde{R}_0)} + \rho^2\left\|\frac{\partial\tilde{v}}{\partial x}\right\|_{L^2(\tilde{R}_2)}\right\} \end{aligned}$$

takes a sufficiently small value, with $\rho = \mathcal{O}(K_2)$.

Then the Hessian of $J_{\sigma,\lambda,\rho}$ will be positive-definite at target velocities as described in (ii), while the *value* of $J_{\sigma,\lambda,\rho}$ will be small enough to conclude the existence of a nearby local minimizer. Since it will no longer be possible to have $J = 0$, only an approximation will be obtained even for data corresponding exactly to $c \in \sum'_c \cap H^2_{\text{loc}}(\mathbf{R})$.

The reader is referred to Symes (1986b) for details of a similar construction.

REFERENCES

J. Duistermaat (1973), *Fourier Integral Operators*, Courant Institute Lecture Notes, Courant Institute of Mathematical Sciences, New York University, New York.

J. Fawcett (1984), *On the stability of inverse scattering problems*, Wave Motion, 6, pp. 489–499.

O. Gauthier, A. Tarantola, and J. Virieux (1986), *Two-dimensional nonlinear inversion of seismic waveforms*, Geophysics, 51, pp. 1387–1403.

D. Gilbarg and N. Trudinger (1983), *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, Heidelberg, Tokyo.

S. Gray and W. Symes (1985), *Stability considerations for one-dimensional inverse problems*, Geophys. J. Roy. Astr. Soc., 80, pp. 149–163.

S. Helgason (1980), *The Radon Transform*, Birkhauser, Boston, Basel, Stuttgart.

P. Kolb, F. Collino, and P. Lailly (1986), *Prestack inversion of a* 1D *medium*, Proc. IEEE, 74, pp. 498–506.

H-O. Kreiss (1970), *Initial-boundary value problems for hyperbolic systems*, Comm. Pure and Appl. Math., 23, pp. 277–298.

D. G. Luenberger (1973), *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA.

A. McAulay (1985), *Prestack inversion with plane-layer point sourcemodeling*, Geophysics, 50, pp. 77–89.

K. MILLER (1970), *Least-squares methods for ill-posed problems with prescribed bounds*, SIAM J. Math. Anal., 1, pp. 52–74.

P. MORA (1987), *Nonlinear 2-d elastic inversion of multi-offset seismic data*, Geophysics, 52, pp. 1211–1228.

A. NACHMAN (1987), *Reconstructions from boundary measurements*, preprint.

J. NECAS (1967), *Les methodes directes en theorie des equations elliptiques*, Maison, Paris.

Y-H. PAO, F. SANTOSA, AND W. SYMES (1984), *Inverse Problems of Acoustic and Elastic Waves*, in Inverse Problems of Acoustic and Elastic Waves, F. Santosa, Y-H. Pao, W. Symes, and C. Holland, eds., Society for Industrial and Applied Mathematics, Philadelphia.

L. PAYNE (1975), *Improperly Posed Problems in Partial Differential Equations*, CBMS–NSF Regional Conference Series in Applied Mathematics 22, Society for Industrial and Applied Mathematics, Philadelphia.

RAKESH AND W. SYMES (1988), *Uniqueness for an inverse problem for the wave equation*, Comm. Partial Differential Equations, 13, pp. 87–96.

A. RAMM (1986), *Scattering by Obstacles*, Reidel, Dordrecht.

J. RAUCH AND M. TAYLOR (1974), *Exponential decay of solutions to hyperbolic equations in bounded domains*, Indiana J. Math., 24, pp. 79–86.

P. SACKS AND F. SANTOSA (1987), *A simple computational scheme for determining the sound speed of an acoustic medium from the surface values of its impulse response*, SIAM J. Sci. Statist. Comput., 3, pp. 501–520.

P. SACKS AND W. SYMES (1985), *Uniqueness and continuous dependence for a multidimensional hyperbolic inverse problem*, Comm. Partial Differential Equations, 10, pp. 635–676.

F. SANTOSA AND W. SYMES (1988), *High frequency perturbational analysis for the point-source response of a layered acoustic medium*, J. Comput. Phys., 74, pp. 318–381.

_____ (1989), *An Analysis of Least-Squares Velocity Inversion*, Geophysical Monograph Number 4, Society of Exploration Geophysicists, Tulsa, OK.

Z. SUN (1988), *On the uniqueness of a multidimensional hyperbolic inverse problem*, Comm. Partial Differential Equations, 13, pp. 1189–1208.

T. SUZUKI (1988), *Stability of an inverse hyperbolic problem*, Inverse Problems, 4, pp. 273–290.

W. SYMES (1981), *The inverse reflection problem for a smoothly stratified elastic medium*, SIAM J. Math. Anal., 12, pp. 421–453.

_____ (1983), *Impedance Profile Inversion via the First Transport Equation*, J. Math. Anal. Appl., 94, pp. 435–453.

_____ (1985), *Stability and instability results for inverse problems in several-dimensional wave propagation*, in Computing Methods in Applied Science and Engineering VII, R. Glowinski and J. L. Lions, eds., North-Holland, Amsterdam.

_____ (1986a), *On the relation between coefficient and boundary values for solutions of Webster's horn equation*, SIAM J. Math. Anal., 17, pp. 1400–1420.

_____ (1986b), *Stability properties for the velocity inversion problem*, in Proc. SIAM Conference on Seismic Exploration and Reservoir Modeling, W. Fitzgibbon, ed., Society for Industrial and Applied Mathematics, Philadelphia.

_____ (1988), *Velocity inversion by coherency optimization*, Technical Report 88-4, Department of Mathematical Sciences, Rice University, Houston, Texas.

_____ (1990), *Velocity inversion: a case study in infinite-dimensional optimization*, Math. Programming, 48, pp. 71–102.

_____ (1991), *Estimation of the Marmousi Velocity Model by Differential Semblance Optimization: Initial Attempts*, in The Marmousi Experience: Practical Aspects of Seismic Inversion, R. Versteeg and G. Grau, eds., Institut Francais du Petrole - Technip.

W. SYMES AND J. J. CARAZZONE (1989), *Velocity inversion by coherency optimization*, Proc. Workshop on Geophysical Inversion, J.B. Bednar, ed., Society for Industrial and Applied Mathematics, Philadelphia (1991).

W. SYMES AND J. J. CARAZZONE (1990), *Velocity inversion by differential semblance optimization*, Geophysics, to appear.

A. TARANTOLA (1987), *Inverse Problem Theory*, Elsevier, New York.

A.N. TIKHONOV AND V. Y. ARSENIN (1977), *Solution of Ill-Posed Problems*, Winston, New York.

O. YILMAZ (1987), *Seismic Data Processing*, Society of Exploration Geophysicists, Investigations in Geophysics No. 2, Tulsa, OK.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 717–731, May 1991

008

SOLUTION OF THE INVERSE SCATTERING PROBLEM FOR THE THREE-DIMENSIONAL SCHRÖDINGER EQUATION USING A FREDHOLM INTEGRAL EQUATION*

TUNCAY AKTOSUN† AND CORNELIS VAN DER MEE‡

Abstract. It is shown that the inverse scattering problem for the three-dimensional Schrödinger equation with a potential having no spherical symmetry can be solved using a Fredholm integral equation. The integral operator studied here is shown to be compact and self-adjoint with its spectrum in $[-1,1]$. The relationship between solutions of this Fredholm equation and of a related Riemann–Hilbert problem is also clarified, and it is shown that the Fredholm integral equation is uniquely solvable if and only if the Riemann–Hilbert problem is uniquely solvable.

Key words. inverse scattering, three-dimensional Schrödinger equation, Fredholm integral equation

AMS(MOS) subject classifications. 81U40, 35P25, 35Q15, 35R30, 47A40

1. Introduction.

Consider the Schrödinger equation in three dimensions

$$\Delta\psi(k,x,\theta) + k^2\psi(k,x,\theta) = V(x)\psi(k,x,\theta), \tag{1.1}$$

where Δ is the Laplacian, $k^2 \in \mathbf{R}$ is energy, $x \in \mathbf{R}^3$ is the space coordinate, and $\theta \in S^2$ is a unit vector in $\mathbf{R}^3$. We assume that the potential $V(x)$ is real and decreases to zero sufficiently fast as $|x| \to \infty$. However, we do not assume any spherical symmetry on the potential. As $|x| \to \infty$, the wavefunction $\psi(k,x,\theta)$ satisfies

$$\psi(k,x,\theta) = e^{ik\theta\cdot x} + \frac{e^{ik|x|}}{|x|} A\left(k, \frac{x}{|x|}, \theta\right) + o\left(\frac{1}{|x|}\right) \tag{1.2}$$

where $A(k,\theta,\theta')$ is the scattering amplitude. The scattering operator $S(k,\theta,\theta')$ is then defined by

$$S(k,\theta,\theta') = \delta(\theta-\theta') - \frac{k}{2\pi i} A(k,\theta,\theta'), \tag{1.3}$$

where δ is the Dirac delta distribution on S^2. In operator notation (1.3) is written as

$$S(k) = \mathbf{I} - \frac{k}{2\pi i} A(k),$$

where the operators are all defined on $L^2(S^2)$, the Hilbert space of complex valued square integrable functions on the unit sphere S^2 in $\mathbf{R}^3$ with the usual inner product.

In this article we study the inverse scattering problem, which consists of recovering $V(x)$ when $S(k)$ is known. Since the main source of information about molecular, atomic, and subatomic particles consists of collision experiments, solving the inverse

* Received by the editors August 16, 1989; accepted for publication (in revised form) May 1, 1990.

†Department of Mathematics, Southern Methodist University, Dallas, Texas 75275. The work of this author was supported in part by National Science Foundation grant DMS-9096268.

‡Department of Physics and Astronomy, Free University, Amsterdam, the Netherlands. The work of this author was supported in part by National Science Foundation grant DMS-8823102.

scattering problem is equivalent to determining the forces among particles from scattering data.

For one-dimensional and radial Schrödinger equations, the inverse scattering problem is fairly well understood (at least for certain classes of potentials) [1]. In higher dimensions, however, the situation is quite different. The solution methods developed in higher dimensions include the Newton–Marchenko method [2]–[4], the generalized Gel'fand-Levitan method [2]–[5], the $\bar{\partial}$ method [6]–[9], a method that only relies on backward scattering data [10]–[13], a method that uses the Green's function of Faddeev [14]–[16], and the Wiener–Hopf factorization method [17]. There are still many open problems in multidimensional inverse scattering, and the methods developed are still far from being complete. Newton's recent book [18] gives a comprehensive review of the methods and related open problems in three-dimensional inverse scattering prior to 1989.

The main idea behind both the Newton–Marchenko and generalized Gel'fand–Levitan methods is to formulate the inverse scattering problem as a Riemann–Hilbert boundary value problem, to transform this Riemann–Hilbert problem into a nonhomogeneous integral equation where the kernel contains the Fourier transform of the scattering data, and to obtain the potential from the solution of the resulting integral equation. In this paper we give the solution of the three-dimensional inverse scattering problem by generalizing a method by Muskhelishvili and Vekua [19], [20] developed to solve Riemann–Hilbert problems with several unknown functions. In the radial case, Newton and Jost used this method to construct potentials from an $n \times n$ scattering matrix for a system of n ordinary differential equations [21]. Here we generalize the Muskhelishvili-Vekua method (and hence the Newton–Jost method) to solve an operator Riemann–Hilbert problem and thus to obtain the solution of the inverse scattering problem for the three-dimensional Schrödinger equation. In this method, the kernel of the key integral equation is an $n \times n$ matrix valued function whereas in our case we deal with an integral equation whose kernel is an operator valued function. In the Newton–Jost method the inverse scattering problem pertains to a system of n ordinary differential equations with an $n \times n$ scattering matrix; however, in this paper, we deal with the inverse scattering problem for a partial differential equation where the kernel of the key integral equation is an operator valued function. Contrary to the three-dimensional Newton–Marchenko and generalized Gel'fand–Levitan inversion methods, we do not use any Fourier transform in our solution of the inverse scattering problem.

The present paper is organized as follows. In §2 we identify the class of potentials for which all of the results in this paper are valid, and we state the key Riemann–Hilbert problem which helps to solve the inverse scattering problem for the three-dimensional Schrödinger equation. In §3, using the Riemann–Hilbert problem, we obtain our fundamental Fredholm integral equation (3.10). In §4, we show that the Fredholm integral operator of (3.10) is compact and self-adjoint and its spectrum is confined to $[-1, 1]$. In §5, we study the relationship between solutions of our Fredholm integral equation and of the Riemann–Hilbert problem and relate the unique solvability of the Fredholm equation to that of the Riemann–Hilbert problem. In §6, utilizing the solution of the fundamental Fredholm integral equation, we give the solution of the inverse scattering problem for the three-dimensional Schrödinger equation. Finally in §7, the conclusion is given.

2. Riemann–Hilbert problem. We first identify the class of potentials for which all of the results in this paper are valid. Except for the third condition given in

the following definition, these conditions are standard assumptions on the potential [18]. These conditions were instrumental in proving the Hölder continuity of the scattering operator and the existence of a Wiener–Hopf factorization of the scattering operator [17].

DEFINITION 2.1. A potential $V(x)$ is said to belong to the *Newton class* if $V(x)$ is real valued and measurable and satisfies

1. $\exists a, b > 0$ such that

$$\int_{\mathbf{R}^3} dx\, |V(x)| \left(\frac{|x| + |y| + a}{|x - y|} \right)^2 \leq b, \quad \forall y \in \mathbf{R}^3, \tag{2.1}$$

2. $\exists c > 0, s > \frac{1}{2}$ such that

$$|V(x)| \leq \frac{c}{(1 + |x|^2)^s}, \quad \forall x \in \mathbf{R}^3, \tag{2.2}$$

3. $\exists \alpha > 0$ and some $\beta \in (0, 1]$ such that

$$\int_{\mathbf{R}^3} dx\, |x|^\beta |V(x)| \leq \alpha, \tag{2.3}$$

4. $k = 0$ is not an exceptional point [22]. This condition is satisfied if at zero energy there are neither bound states nor half bound states.

In the Schrödinger equation (1.1), k appears as k^2 and hence $\psi(-k, x, \theta)$ is a solution whenever $\psi(k, x, \theta)$ is. These two solutions are related to each other as [2]

$$\psi(k, x, \theta) = \int_{S^2} d\theta'\, S(k, -\theta, \theta') \psi(-k, x, \theta'). \tag{2.4}$$

Define

$$f(k, x, \theta) = e^{-ik\theta \cdot x} \psi(k, x, \theta). \tag{2.5}$$

If the potential satisfies (2.1) and if there are no bound states, for fixed x and θ, the function $f(k, x, \theta)$ has an analytic extension in k to the upper half complex plane $\mathbf{C}^+$ and $f(k, x, \theta) = 1 + O(\frac{1}{|k|})$ as $|k| \to \infty$ there [2]. Similarly $f(-k, x, \theta)$ has an analytic extension in k to the lower half complex plane $\mathbf{C}^-$. Hence, using (2.4), we obtain the Riemann–Hilbert problem

$$f(k, x, \theta) = \int_{S^2} d\theta'\, e^{-ik\theta \cdot x} S(k, -\theta, \theta') e^{-ik\theta' \cdot x} f(-k, x, \theta'), \qquad k \in \mathbf{R}. \tag{2.6}$$

Let us define

$$G(k, x, \theta, \theta') = e^{-ik\theta \cdot x} S(k, -\theta, -\theta') e^{ik\theta' \cdot x}, \tag{2.7}$$

$$X_\pm(k) = f(\pm k, x, \pm\theta) - 1, \tag{2.8}$$

where $f(k, x, \theta)$ is as in (2.5). Then we can write (2.6) in vector form as

$$X_+(k) = G(k) X_-(k) + [G(k) - \mathbf{I}]\hat{1}, \qquad k \in \mathbf{R}, \tag{2.9}$$

where $G(k)$ is the operator on $L^2(S^2)$ with its kernel given in (2.7), $\mathbf{I}$ is the identity operator on this space, and $\hat{1}$ is the function on $L^2(S^2)$ defined as $\hat{1}(\theta) = 1$, $\forall \theta \in S^2$. Note that, since x enters (2.9) only as a parameter, we have suppressed the x-dependence of all the operators and vectors in (2.9).

If there are bound states, the extension of $f(k, x, \theta)$ in k to $\mathbf{C}^+$ becomes meromorphic with simple poles on the imaginary axis. A pole at $k = i\kappa$ corresponds to

a bound state of the Hamiltonian with energy $-\kappa^2$. It is possible to remove these simple poles from the Riemann–Hilbert problem by a reduction method [4]. Assume there is a bound state corresponding to a pole at $k = i\kappa$. Using a suitable orthogonal projection $\mathbf{B}$, we form the rational function

$$\Pi(k) = \mathbf{I} - \mathbf{B} + \frac{k + i\kappa}{k - i\kappa}\mathbf{B}$$

and define the corresponding reduced quantities

$$G^{\rm red}(k) = \Pi(k)^{-1} G(k) \Pi(k)$$

$$X_+^{\rm red}(k) = \Pi(k)^{-1} X_+(k) + [\Pi(k)^{-1} - \mathbf{I}]\hat{1} \tag{2.10}$$

$$X_-^{\rm red}(k) = \Pi(k) X_-(k) + [\Pi(k) - \mathbf{I}]\hat{1}. \tag{2.11}$$

As a result, $X_+^{\rm red}(k)$ does not have a pole at $k = i\kappa$ and $X_-^{\rm red}(k)$ does not have a pole at $k = -i\kappa$. If there is more than one bound state, this procedure must be repeated to remove the finitely many poles corresponding to the bound states; the details can be found in [4]. This eventually leads to the reduced Riemann–Hilbert problem

$$X_+^{\rm red}(k) = G^{\rm red}(k) X_-^{\rm red}(k) + [G^{\rm red}(k) - \mathbf{I}]\hat{1}, \qquad k \in \mathbf{R}. \tag{2.12}$$

Once the reduced Riemann–Hilbert problem (2.12) is solved, the solution of the original Riemann–Hilbert problem (2.9) can easily be obtained using (2.10) and (2.11). Hence, in the following sections, without any loss of generality, we will obtain the solution of the Riemann–Hilbert problem assuming that $X_+(k)$ and $X_-(k)$ have analytic extensions to $\mathbf{C}^+$ and $\mathbf{C}^-$, respectively, and vanish in the norm of $L^2(S^2)$ as $k \to \infty$ from that half plane.

3. Fredholm integral equation. In this section we show that the Riemann–Hilbert problem posed in (2.9) leads to a Fredholm integral equation, which will be the key equation to solve the inverse scattering problem.

From the Cauchy integral formula we have

$$X_+(k) = \frac{1}{\pi i}{\rm CPV}\int_{-\infty}^{+\infty} dt\, \frac{X_+(t)}{t - k}, \tag{3.1}$$

$$X_-(k) = -\frac{1}{\pi i}{\rm CPV}\int_{-\infty}^{+\infty} dt\, \frac{X_-(t)}{t - k}, \tag{3.2}$$

where CPV denotes the Cauchy principal value. Operating on (3.2) by $G(k)$ and adding the result to (3.1), we obtain

$$X_+(k) + G(k)X_-(k) = \frac{1}{\pi i}{\rm CPV}\int_{-\infty}^{+\infty} \frac{dt}{t - k}[X_+(t) - G(k)X_-(t)]. \tag{3.3}$$

Using (2.9) in (3.3) we obtain

$$\begin{aligned} &2X_+(k) + [\mathbf{I} - G(k)]\hat{1} \\ &\quad = \frac{1}{\pi i}{\rm CPV}\int_{-\infty}^{+\infty} \frac{dt}{t - k}\left([\mathbf{I} - G(k)G(t)^{-1}]X_+(t) + G(k)[\mathbf{I} - G(t)^{-1}]\hat{1}\right). \end{aligned} \tag{3.4}$$

Define the integral operator K whose kernel is given by

$$K(k,t) = \frac{1}{2\pi i}\frac{\mathbf{I} - G(k)G(t)^{-1}}{t - k}. \tag{3.5}$$

Then we can write (3.4) as

$$X_+(k) - \mathrm{CPV}\int_{-\infty}^{+\infty} dt\, K(k,t)X_+(t) = H(k), \tag{3.6}$$

where $H(k)$ is given by

$$H(k) = [G(k) - \mathbf{I} + \int_{-\infty}^{\infty} dt\, K(k,t)]\hat{1}.$$

If the potential $V(x)$ belongs to the Newton class defined in §2, the operator $G(k)^{-1}$ is Hölder continuous [17]. Hence the integral in (3.6) is no longer singular and we can drop CPV in front of this integral. Thus, we obtain the regular Fredholm integral equation of the second kind

$$X_+(k) - \int_{-\infty}^{+\infty} dt\, K(k,t)X_+(t) = H(k). \tag{3.7}$$

The Möbius transformation $k \to \xi = (k-i)/(k+i)$ maps the extended real axis onto the unit circle $\mathbf{T}$, the upper half complex plane $\mathbf{C}^+$ onto the unit disk $\mathbf{T}^+$, and the lower half complex plane $\mathbf{C}^-$ onto the exterior of the unit disk $\mathbf{T}^-$ where ∞ is considered to be a point of $\mathbf{T}^-$. Let $\tilde{S}(\xi) = S(k)$ under this transformation, and let us adopt this notation and use the tilde to denote the Möbius transformed quantity for other functions and operator valued functions throughout the paper.

Let $k \to \xi = (k-i)/(k+i)$ and $t \to \eta = (t-i)/(t+i)$ under this Möbius transformation. Defining

$$Y(\xi) = \frac{\tilde{X}_+(\xi)}{1-\xi} = \frac{k+i}{2i}X_+(k) \tag{3.8}$$

and

$$L(\xi) = \frac{\tilde{H}(\xi)}{1-\xi} = \frac{k+i}{2i}H(k), \tag{3.9}$$

we can transform (3.7) into

$$Y(\xi) - \int_{\mathbf{T}} d\eta\, \tilde{K}(\xi,\eta)Y(\eta) = L(\xi), \tag{3.10}$$

where the kernel of the integral operator $\tilde{K}$ is given by

$$\tilde{K}(\xi,\eta) = \frac{1}{2\pi i}\frac{\mathbf{I} - \tilde{G}(\xi)\tilde{G}(\eta)^{-1}}{\eta - \xi}. \tag{3.11}$$

Comparing (3.11) with (3.5) we see that $\tilde{K}$ is the Möbius transformed operator for K.

4. Properties of the integral operator. In this section we show that the integral operator $\tilde{K}$ in (3.10) is compact and self-adjoint and its spectrum lies in $[-1,1]$.

For $\alpha \in (0,1]$ let $\mathcal{H}_\alpha(\mathbf{T}; L^2(S^2))$ be the Banach space of Hölder continuous functions $W : \mathbf{T} \to L^2(S^2)$; i.e., the Banach space of all (strongly) continuous functions $W : \mathbf{T} \to L^2(S^2)$ which are bounded with respect to the norm

$$|||W|||_\alpha = \max_{\xi\in\mathbf{T}} \|W(\xi)\| + \sup_{\xi\neq\eta\in\mathbf{T}} \frac{\|W(\xi) - W(\eta)\|}{|\xi - \eta|^\alpha}.$$

Here and in the following $\|\cdot\|$ without any subscript denotes the operator norm on $L^2(S^2)$. Let $\mathcal{C}(\mathbf{T};L^2(S^2))$ denote the Banach space of (strongly) continuous functions $W:\mathbf{T}\to L^2(S^2)$ with norm $\|W\|_\infty=\max_{\xi\in\mathbf{T}}\|W(\xi)\|$. Finally, for $1\le p\le\infty$ let $L^p(\mathbf{T};L^2(S^2))$ denote the Banach space of all strongly measurable functions $W:\mathbf{T}\to L^2(S^2)$ such that $\|W(\cdot)\|$ belongs to $L^p(\mathbf{T})$ [23].

Let Γ be the singular integral operator on $L^2(S^2)$ defined by

$$(\Gamma f)(\xi)=\frac{1}{\pi i}\mathrm{CPV}\int_{\mathbf{T}}d\eta\,\frac{f(\eta)}{\eta-\xi}. \tag{4.1}$$

Then, from (3.10) and (3.11) it is seen that we can write $\tilde{K}$ as

$$\tilde{K}=\tfrac{1}{2}(\Gamma-\tilde{G}\Gamma\tilde{G}^{-1}), \tag{4.2}$$

where $\tilde{G}$ and $\tilde{G}^{-1}$ are operators of multiplication by the respective functions. The space $L^2(\mathbf{T};L^2(S^2))$ is a Hilbert space which allows the decomposition into the orthogonal closed subspaces $\mathcal{E}_+(\mathbf{T};L^2(S^2))$ and $\mathcal{E}_-(\mathbf{T};L^2(S^2))$. Here $\mathcal{E}_+(\mathbf{T};L^2(S^2))$ is the subspace of all L^2-functions which allow for an analytic continuation to $\mathbf{T}^+$, while $\mathcal{E}_-(\mathbf{T};L^2(S^2))$ is the subspace of L^2-functions which allow for an analytic continuation to $\mathbf{T}^-$ that vanishes at infinity. If $f(\xi)$ is an L^2-function defined on the unit circle $\mathbf{T}$, then using its Fourier series, we have the decomposition

$$f(\xi)=\sum_{n=-\infty}^{-1}f_n\xi^n+\sum_{n=0}^{\infty}f_n\xi^n$$

as a sum of elements in $\mathcal{E}_-(\mathbf{T};L^2(S^2))$ and $\mathcal{E}_+(\mathbf{T};L^2(S^2))$, and this decomposition is orthogonal. If we denote the two summations in the above decomposition as f_- and f_+, respectively, we obtain

$$\Gamma f=f_+-f_-. \tag{4.3}$$

Thus, Γ is self-adjoint and has unit norm on $L^2(\mathbf{T};L^2(S^2))$. More generally, Γ is a bounded linear operator on $\mathcal{E}(\mathbf{T};L^2(S^2))$, where $\mathcal{E}$ represents L^p with $1<p<\infty$ or $\mathcal{H}_\gamma$ with $0<\gamma<1$. This result can be derived from the boundedness of Γ on the spaces $\mathcal{E}(\mathbf{T})$ of scalar functions (Theorems I2.1 and I6.1 of [24]) and the density of the linear subspace $\{\sum_{j=1}^n\varphi_j(\cdot)h_j:n\in\mathbf{N},\varphi_j\in\mathcal{E}(\mathbf{T}),h_j\in L^2(S^2)\}$ in $\mathcal{E}(\mathbf{T};L^2(S^2))$ [23].

For potentials identified in Definition 2.1, we will prove the following three propositions.

PROPOSITION 4.1. *Let $V(x)$ belong to the Newton class. Then the Fredholm integral operator $\tilde{K}$ in (3.10) is compact on $L^p(\mathbf{T};L^2(S^2))$ for $1\le p\le\infty$, on $\mathcal{C}(\mathbf{T};L^2(S^2))$, and on $\mathcal{H}_\gamma(\mathbf{T};L^2(S^2))$ for $0<\gamma<\mu$. Here $\mu=\beta/2(1+\beta)$ for $s>1$ in (2.2) and $\mu=\beta(1-s)/(\beta-s+\frac{3}{2})$ for $\frac{1}{2}<s<1$ in (2.2), where $\beta\in(0,1]$ is the constant in (2.3)* [17].

Proof. Define the integral operator M acting on $L^2(S^2)$ as

$$(MZ)(\xi)=\frac{1}{2\pi i}\int_{\mathbf{T}}d\eta\,\frac{\tilde{G}(\eta)-\tilde{G}(\xi)}{\eta-\xi}Z(\eta). \tag{4.4}$$

Then $(\tilde{K}Y)(\xi)=(M\tilde{G}^{-1}Y)(\xi)$. Due to the fact that the scattering operator $S(k)$ is unitary, the operator $\tilde{G}(\xi)^{-1}$ is bounded with norm $\|\tilde{G}^{-1}\|=1$. Hence, to prove that $\tilde{K}$ is compact, it is sufficient to prove that M is compact, and we will prove this by showing that M can be approximated by a sequence of finite rank operators.

It is known [17] that $\tilde{G}(\xi) - \mathbf{I}$ is a compact operator on $L^2(S^2)$ depending continuously on $\xi \in \mathbf{T}$. Hence it can be approximated by a sequence of finite rank operators $\{\tilde{G}_n(\xi) - \mathbf{I}\}_{n=1}^{\infty}$ given by

$$\tilde{G}_n(\xi) - \mathbf{I} = \sum_{j=-n}^{n} \xi^j A_j,$$

where $\{A_j\}_{j=-\infty}^{\infty}$ is a sequence of mutually orthogonal finite rank operators on $L^2(S^2)$. We then obtain

$$\frac{\tilde{G}_n(\eta) - \tilde{G}_n(\xi)}{\eta - \xi} = \sum_{j=1}^{n} \sum_{l=0}^{j-1} (\xi^l \eta^{j-l-1} - \xi^{l-j} \eta^{-l-1}) A_j. \tag{4.5}$$

Using (4.5) in (4.4) we obtain a sequence of operators $\{M_n\}_{n=1}^{\infty}$ given by

$$(M_n Y)(\xi) = \frac{1}{2\pi i} \sum_{j=1}^{n} \left(\sum_{l=0}^{j-1} \xi^l - \sum_{l=j}^{2j-1} \xi^{l-2j} \right) A_j \int_{\mathbf{T}} d\eta\, \eta^{j-l-1} Y(\eta). \tag{4.6}$$

Hence, for each n, M_n is a finite rank operator. We will complete the proof of our proposition by showing that M_n converges to M as $n \to \infty$ in the function spaces $L^p(\mathbf{T}; L^2(S^2))$, $\mathcal{C}(\mathbf{T}; L^2(S^2))$, and $\mathcal{H}_\gamma(\mathbf{T}; L^2(S^2))$.

From (4.2) and (4.4), it is seen that $M = \frac{1}{2}(\Gamma\tilde{G} - \tilde{G}\Gamma)$ and $M_n = \frac{1}{2}(\Gamma\tilde{G}_n - \tilde{G}_n\Gamma)$. Hence, using the boundedness of Γ on $\mathcal{E}(\mathbf{T}; L^2(S^2))$, where $\mathcal{E}$ is equal to L^p with $1 < p < \infty$ or equal to $\mathcal{H}_\gamma$ with $0 < \gamma < \mu$, we have in the operator norm on $\mathcal{E}(\mathbf{T}; L^2(S^2))$

$$||M - M_n||_{\mathcal{E}} \leq ||\Gamma||_{\mathcal{E}}\, ||\tilde{G} - \tilde{G}_n||_{\mathcal{E}} \leq c\, |||\tilde{G} - \tilde{G}_n|||_\gamma$$

for some constant $c > 0$. Thus, $M_n \to M$ in the norm of $L^p(\mathbf{T}; L^2(S^2))$ for $1 < p < \infty$ and in the norm of $\mathcal{H}_\gamma(\mathbf{T}; L^2(S^2))$ for $0 < \gamma < \mu$.

It remains to prove the convergence of M_n to M as $n \to \infty$ in the operator norm on $\mathcal{E}(\mathbf{T}; L^2(S^2))$ where $\mathcal{E}$ is L^1, L^∞, or $\mathcal{C}$. First, from (4.4) we obtain

$$||(M - M_n)Y||_p \leq m_\gamma |||\tilde{G} - \tilde{G}_n|||_\gamma ||Y||_p, \qquad p = 1, \infty, \tag{4.7}$$

where

$$m_\gamma = \frac{1}{2\pi} \int_{\mathbf{T}} \frac{|d\eta|}{|\eta - \xi|^{1-\gamma}} = \frac{1}{2\pi} \int_0^{2\pi} \frac{dt}{(2 - 2\cos t)^{(1-\gamma)/2}}$$

is a constant independent of $\xi \in \mathbf{T}$. Then, from (4.7) it follows that $M_n \to M$ in the operator norm on $L^1(\mathbf{T}; L^2(S^2))$ and on $L^\infty(\mathbf{T}; L^2(S^2))$. Next, as a result of the dominated convergence theorem for Bochner integrals (Theorem II.3 of [23]), if $\xi \to \xi_0$ in $\mathbf{T}$, then for every $Y \in L^\infty(\mathbf{T}; L^2(S^2))$, $[(M - M_n)Y](\xi)$ converges to $[(M - M_n)Y](\xi_0)$ in the norm of $L^2(S^2)$. Therefore, M and M_n map $L^\infty(\mathbf{T}; L^2(S^2))$ into $\mathcal{C}(\mathbf{T}; L^2(S^2))$. Hence, we obtain from (4.7) that the convergence of M_n to M also holds in the operator norm on $\mathcal{C}(\mathbf{T}; L^2(S^2))$. Thus, the proof is complete. □

If λ is a nonzero eigenvalue of $\tilde{K}$, then, by the compactness of $\tilde{K}$ on $\mathcal{E}(\mathbf{T}; L^2(S^2))$, there exists some integer $N \geq 1$ such that the kernel of $(\tilde{K} - \lambda\mathbf{I})^n$ coincides with the kernel of $(\tilde{K} - \lambda\mathbf{I})^N$ if $n \geq N$. The finite dimension of this subspace is called the algebraic multiplicity of λ. The dimension of the kernel of $\tilde{K} - \lambda\mathbf{I}$ is called the geometric multiplicity of λ.

As a result of the compactness of $\tilde{K}$, we have the following result.

COROLLARY 4.2. *Let $V(x)$ belong to the Newton class. Then the nonzero eigenvalues of $\tilde{K}$ as well as their (algebraic) multiplicities do not depend on the choice of the function space on which they are defined. As a result, the eigenvectors and generalized eigenvectors of $\tilde{K}$ corresponding to its nonzero eigenvalues belong to each of the spaces $\mathcal{C}(\mathbf{T}; L^2(S^2))$, $L^p(\mathbf{T}; L^2(S^2))$ for $1 \leq p \leq \infty$, and $\mathcal{H}_\gamma(\mathbf{T}; L^2(S^2))$ for $0 < \gamma < \mu$, where μ is the constant specified in Proposition* 4.1.

Proof. Note that for $0 < \beta \leq \gamma$ and $1 \leq p \leq q < \infty$ we have

$$\mathcal{H}_\gamma \subset \mathcal{H}_\beta \subset \mathcal{C} \subset L^\infty \subset L^q \subset L^p.$$

Consider any of the four pairs of spaces $\{L^\infty, L^p\}$, $\{L^q, L^p\}$, $\{\mathcal{C}, L^p\}$, and $\{\mathcal{H}_\gamma, \mathcal{C}\}$. In each pair, let $\mathcal{E}_1$ denote the first space and $\mathcal{E}_2$ denote the second space for functions in $L^2(S^2)$. For example, for the first pair, we let $\mathcal{E}_1 = L^\infty(\mathbf{T}; L^2(S^2))$ and $\mathcal{E}_2 = L^p(\mathbf{T}; L^2(S^2))$. Then for all the pairs, $\mathcal{E}_1$ is continuously and densely imbedded in $\mathcal{E}_2$. Since $\tilde{K}$ is compact in $\mathcal{E}_1$ and in $\mathcal{E}_2$, for each nonzero complex number λ and natural number n, the closure in $\mathcal{E}_2$ of the image of $\mathcal{E}_1$ under $(\tilde{K} - \lambda \mathbf{I})^n$ coincides with the image of $\mathcal{E}_2$ under $(\tilde{K} - \lambda \mathbf{I})^n$. As a result, the complements of the ranges of $(\tilde{K} - \lambda \mathbf{I})^n$ in both $\mathcal{E}_1$ and in $\mathcal{E}_2$ have the same finite dimension. Since $(\tilde{K} - \lambda \mathbf{I})^n$ is a Fredholm operator of index 0, its kernels in $\mathcal{E}_1$ and $\mathcal{E}_2$ also have the same finite dimension. As a consequence, the dimensions of the (generalized) eigenspaces for each nonzero eigenvalue of $\tilde{K}$ are the same in $\mathcal{C}(\mathbf{T}; L^2(S^2))$, in $L^p(\mathbf{T}; L^2(S^2))$ for $1 \leq p \leq \infty$, and in $\mathcal{H}_\gamma(\mathbf{T}; L^2(S^2))$ for $0 < \gamma < \mu$, where μ is the constant specified in Proposition 4.1. □

PROPOSITION 4.3. *Let the potential $V(x)$ belong to the Newton class. Then $\tilde{K}$ is self-adjoint on $L^2(\mathbf{T}; L^2(S^2))$. As a result, all eigenvalues of $\tilde{K}$ are real and their algebraic and geometric multiplicities coincide.*

Proof. Due to the unitarity of the scattering operator $S(k)$, the operators $\tilde{G}$ and $\tilde{G}^{-1}$ are unitary. Because of (4.3), the singular integral operator Γ is self-adjoint. Thus, from (4.2) it follows that $\tilde{K}$ is self-adjoint on $L^2(\mathbf{T}; L^2(S^2))$. □

PROPOSITION 4.4. *Let the potential $V(x)$ belong to the Newton class. Then the norm of $\tilde{K}$ in $L^2(\mathbf{T}; L^2(S^2))$ is bounded above by* 1, *and all eigenvalues of $\tilde{K}$ belong to* $[-1, 1]$.

Proof. From (4.3) it is seen that the singular integral operator Γ has unit norm on $L^2(\mathbf{T}; L^2(S^2))$. Due to their unitarity, the multiplication operators $\tilde{G}$ and $\tilde{G}^{-1}$ each have unit norm. Thus, as seen from (4.2), $\tilde{K}$ has at most unit norm on $L^2(\mathbf{T}; L^2(S^2))$. Furthermore, by Proposition 4.3, $\tilde{K}$ is self-adjoint and hence its spectrum is real. Thus, the spectrum of $\tilde{K}$ lies in $[-1, 1]$, and by Corollary 4.2, the spectrum does not depend on the function space used. □

If ± 1 are not eigenvalues of $\tilde{K}$, the Fredholm integral equation (3.10) has a unique solution which can be obtained by iteration. Since we have shown in Corollary 4.2 that the spectral radius of $\tilde{K}$ does not depend on the function space used, the iteration converges in the norm of any of the spaces mentioned in Proposition 4.1.

5. Relationship between solutions of the Fredholm integral equation and of the Riemann–Hilbert problem. In this section we study the relationship between solutions of the Fredholm integral equation (3.10) and solutions of the Riemann–Hilbert problem on $\mathcal{E}(\mathbf{T}; L^2(S^2))$, where $\mathcal{E}$ is either L^p with $1 < p < \infty$ or $\mathcal{H}_\gamma$ with $0 < \gamma < \mu$, μ being the constant specified in Proposition 4.1. We also investigate the relationship between the partial indices of $\tilde{G}(\xi)$ [17] and the existence and uniqueness of the solution of (3.10).

Using $F(\xi) = (1/(1-\xi))[\tilde{G}(\xi) - \mathbf{I}]\hat{1}$, we can transform (2.9) into the Riemann–Hilbert problem on the unit circle $\mathbf{T}$ to obtain

$$Y_+(\xi) = \tilde{G}(\xi)Y_-(\xi) + F(\xi), \qquad \xi \in \mathbf{T}. \tag{5.1}$$

Let us write (3.10) as

$$(\mathbf{I} - \tilde{K})Y = L. \tag{5.2}$$

Note that the nonhomogeneous terms in (5.1) and in (5.2) are related to each other by

$$L = \tfrac{1}{2}\tilde{G}(\mathbf{I} + \Gamma)\tilde{G}^{-1}F. \tag{5.3}$$

We will relate the solutions of (5.1) with $F \in \mathcal{E}(\mathbf{T}; L^2(S^2))$ and $Y_\pm \in \mathcal{E}_\pm(\mathbf{T}; L^2(S^2))$ to the solutions of (5.2) with $L, Y \in \mathcal{E}(\mathbf{T}; L^2(S^2))$.

As mentioned prior to Proposition 4.1, the singular integral operator Γ defined in (4.1) is bounded on $\mathcal{E}(\mathbf{T}; L^2(S^2))$. Then $\frac{1}{2}(\mathbf{I}+\Gamma)$ and $\frac{1}{2}(\mathbf{I}-\Gamma)$ are complementary bounded projections onto the subspace $\mathcal{E}_+(\mathbf{T}; L^2(S^2))$ of all functions in $\mathcal{E}(\mathbf{T}; L^2(S^2))$ with an analytic continuation to $\mathbf{T}^+$ and onto the subspace $\mathcal{E}_-(\mathbf{T}; L^2(S^2))$ of all functions in $\mathcal{E}(\mathbf{T}; L^2(S^2))$ with an analytic continuation to $\mathbf{T}^-$ vanishing at infinity.

THEOREM 5.1. *Let $F \in \mathcal{E}(\mathbf{T}; L^2(S^2))$. If $Y_\pm$ is a solution of* (5.1) *in $\mathcal{E}_\pm(\mathbf{T}; L^2(S^2))$, then $Y = Y_+$ is a solution of* (5.2). *Conversely, if Y is a solution of* (5.2) *and $Y \in \mathcal{E}_+(\mathbf{T}; L^2(S^2))$, then $Y_\pm$, where $Y_+ = Y$ and $Y_- = \tilde{G}^{-1}(Y - F)$, is a solution of* (5.1) *with $Y_\pm \in \mathcal{E}_\pm(\mathbf{T}; L^2(S^2))$.*

Proof. Let $Y_\pm \in \mathcal{E}_\pm(\mathbf{T}; L^2(S^2))$ be a solution of (5.1). Then using (4.2), (5.3), and $(\mathbf{I} \pm \Gamma)Y_\mp = 0$, we have

$$\begin{aligned}(\mathbf{I} - \tilde{K})Y_+ &= \tfrac{1}{2}(\mathbf{I} - \Gamma)Y_+ + \tfrac{1}{2}\tilde{G}(\mathbf{I} + \Gamma)\tilde{G}^{-1}Y_+ \\ &= \tfrac{1}{2}\tilde{G}(\mathbf{I} + \Gamma)(Y_- + \tilde{G}^{-1}F) = \tfrac{1}{2}\tilde{G}(\mathbf{I} + \Gamma)\tilde{G}^{-1}F = L.\end{aligned}$$

Conversely, let $Y \in \mathcal{E}_+(\mathbf{T}; L^2(S^2))$ be a solution of (5.2). Clearly $Y_+ = Y$ and $Y_- = \tilde{G}^{-1}(Y - F)$ satisfy (5.1), provided $Y_- \in \mathcal{E}_-(\mathbf{T}; L^2(S^2))$. This follows from

$$(\mathbf{I} + \Gamma)Y_- = (\tilde{G}^{-1} + \Gamma\tilde{G}^{-1})Y - (\mathbf{I} + \Gamma)\tilde{G}^{-1}F = 2\tilde{G}^{-1}L - (\mathbf{I} + \Gamma)\tilde{G}^{-1}F = 0,$$

where we have used (4.2) and (5.3). □

From (4.2) and $\Gamma^2 = \mathbf{I}$ it is immediate that $\tilde{K}^2$ and Γ commute. Hence, $\tilde{K}^2$ maps $\mathcal{E}_+(\mathbf{T}; L^2(S^2))$ into $\mathcal{E}_+(\mathbf{T}; L^2(S^2))$ and $\mathcal{E}_-(\mathbf{T}; L^2(S^2))$ into $\mathcal{E}_-(\mathbf{T}; L^2(S^2))$. Thus, using the compactness of $\tilde{K}^2$, we can decompose the kernel and range of $\mathbf{I} - \tilde{K}^2$ as

$$\operatorname{Ran}(\mathbf{I} - \tilde{K}^2) = \{\operatorname{Ran}(\mathbf{I} - \tilde{K}^2) \cap \mathcal{E}_+(\mathbf{T}; L^2(S^2))\} \oplus \{\operatorname{Ran}(\mathbf{I} - \tilde{K}^2) \cap \mathcal{E}_-(\mathbf{T}; L^2(S^2))\} \tag{5.4}$$

and

$$\operatorname{Ker}(\mathbf{I} - \tilde{K}^2) = \{\operatorname{Ker}(\mathbf{I} - \tilde{K}^2) \cap \mathcal{E}_-(\mathbf{T}; L^2(S^2))\} \oplus \{\operatorname{Ker}(\mathbf{I} - \tilde{K}^2) \cap \mathcal{E}_-(\mathbf{T}; L^2(S^2))\}.$$

PROPOSITION 5.2. *Let $F \in \mathcal{E}(\mathbf{T}; L^2(S^2))$. Then there exists at least one solution of* (5.2) *if and only if there exists at least one solution of the equation*

$$(\mathbf{I} - \tilde{K}^2)Z = (\mathbf{I} + \tilde{K})L. \tag{5.5}$$

Moreover, if it exists, it is possible to choose the solution Z of (5.5) *in $\mathcal{E}_+(\mathbf{T}; L^2(S^2))$, but this solution may not satisfy* (5.2).

Proof. If Y is a solution of (5.2) in $\mathcal{E}(\mathbf{T}; L^2(S^2))$, then clearly it is also a solution of (5.5). To prove the converse, let us first take $\mathcal{E} = L^2$. The solution of (5.5) exists

provided $(\mathbf{I}+\tilde{K})L$ is orthogonal to $\mathrm{Ker}(\mathbf{I}-\tilde{K}^2)$. Let Z_0 be such that $(\mathbf{I}-\tilde{K}^2)Z_0 = 0$. Writing $Z_0 = Z_1+Z_2$ where $\tilde{K}Z_1 = Z_1$ and $\tilde{K}Z_2 = -Z_2$, and using the self-adjointness of $\tilde{K}$, we obtain

$$\langle(\mathbf{I}+\tilde{K})L, Z_2\rangle = \langle L, (\mathbf{I}+\tilde{K})Z_2\rangle = 0$$

and

$$\langle(\mathbf{I}+\tilde{K})L, Z_1\rangle = \langle L, (\mathbf{I}+\tilde{K})Z_1\rangle = 2\langle L, Z_1\rangle.$$

Hence, $(\mathbf{I}+\tilde{K})L$ is orthogonal to $\mathrm{Ker}(\mathbf{I}-\tilde{K}^2)$ if and only if L is orthogonal to $\mathrm{Ker}(\mathbf{I}-\tilde{K})$. Thus, a solution of (5.5) exists if and only if a solution of (5.2) exists.

Furthermore, from (4.2) and (5.3) we obtain

$$(\mathbf{I}+\tilde{K})L = \frac{1}{4}(\mathbf{I}+\Gamma)\tilde{G}(\mathbf{I}+\Gamma)\tilde{G}^{-1}F \in \mathcal{E}_+(\mathbf{T}; L^2(S^2)).$$

Then, since $N := \frac{1}{4}(\mathbf{I}+\Gamma)\tilde{G}(\mathbf{I}+\Gamma)\tilde{G}^{-1}$ has a closed range in each $\mathcal{E}(\mathbf{T}; L^2(S^2))$, the image of $L^2(\mathbf{T}; L^2(S^2))$ under N is the closure in $L^2(\mathbf{T}; L^2(S^2))$ of the image of $\mathcal{E}(\mathbf{T}; L^2(S^2))$ under N. Similarly, the image of $L^2(\mathbf{T}; L^2(S^2))$ under $\mathbf{I}+\tilde{K}$ is the closure in $L^2(\mathbf{T}; L^2(S^2))$ of the image of $\mathcal{E}(\mathbf{T}; L^2(S^2))$ under $\mathbf{I}+\tilde{K}$. Hence, if

$$N[L^2(\mathbf{T}; L^2(S^2))] \subset (\mathbf{I}+\tilde{K})[L^2(\mathbf{T}; L^2(S^2))]$$

as proven above, we also have

$$N[\mathcal{E}(\mathbf{T}; L^2(S^2))] \subset (\mathbf{I}+\tilde{K})[\mathcal{E}(\mathbf{T}; L^2(S^2))]$$

for $\mathcal{E} = L^p$ with $2 < p < \infty$ or $\mathcal{E} = \mathcal{H}_\gamma$ with $0 < \gamma < \mu$. The same conclusion may be drawn if $\mathcal{E} = L^p$ with $1 < p < 2$, but this time we use the fact that $L^2(\mathbf{T}; L^2(S^2))$ is continuously and densely imbedded in $\mathcal{E}(\mathbf{T}; L^2(S^2))$.

Finally, from (5.5) it is seen that a solution of (5.5) exists provided $(\mathbf{I}+\tilde{K})L \in \mathrm{Ran}(\mathbf{I}-\tilde{K}^2)$. Then using (5.4) and the invertibility of $(\mathbf{I}-\tilde{K}^2)$ on $\mathrm{Ran}(\mathbf{I}-\tilde{K}^2)$, it appears the solution of (5.5) can be chosen in $\mathrm{Ran}(\mathbf{I}-\tilde{K}^2)\cap\mathcal{E}_+$. □

The number of linearly independent solutions of the homogeneous Riemann–Hilbert problem

$$Y_+(\xi) = \tilde{G}(\xi)Y_-(\xi), \qquad \xi \in \mathbf{T} \tag{5.6}$$

is the sum of the positive partial indices of $\tilde{G}(\xi)$. Let $\{\rho_j\}$ denote the set of partial indices of $\tilde{G}(\xi)$. These partial indices arise in the Wiener–Hopf factorization of $\tilde{G}(\xi)$ [17].

A special case of Theorem 5.1 with $L = 0$ concludes that any solution Y_+ of (5.6) corresponds to a solution of the homogeneous Fredholm integral equation $(\mathbf{I}-\tilde{K})Y = 0$ in $\mathcal{E}_+(\mathbf{T}; L^2(S^2))$. As a result,

$$\sum_{\rho_j>0} \rho_j = \dim\{\mathrm{Ker}(\mathbf{I}-\tilde{K})\cap\mathcal{E}_+(\mathbf{T}; L^2(S^2))\}. \tag{5.7}$$

To obtain an expression for the sum of the negative indices, we consider the homogeneous Riemann–Hilbert problem which is adjoint to (5.6)

$$Z_+(\xi) = \tilde{G}(\xi)^{-1}Z_-(\xi), \quad \xi \in \mathbf{T} \tag{5.8}$$

where $Z_\pm \in \mathcal{E}_\pm(\mathbf{T}; L^2(S^2))$. Then the number of linearly independent solutions of (5.8) is the sum of the positive partial indices of $\tilde{G}(\xi)^{-1}$. Due to the unitarity of $\tilde{G}(\xi)$,

these indices coincide with the negatives of the negative partial indices of $\tilde{G}(\xi)$ [17]. Using (4.2) and (5.8) we have

$$(\mathbf{I}+\tilde{K})Z_- = (\mathbf{I}+\tfrac{1}{2}\Gamma-\tfrac{1}{2}\tilde{G}\Gamma\tilde{G}^{-1})Z_- = \tfrac{1}{2}(Z_- - \tilde{G}\Gamma Z_+) = 0.$$

Conversely, if $Z_- \in \mathcal{E}_-(\mathbf{T};L^2(S^2))$ and $(\mathbf{I}+\tilde{K})Z_- = 0$, then $\tilde{G}^{-1}Z_- = \tilde{G}^{-1}Z_- - \tilde{G}^{-1}(\mathbf{I}+\tilde{K})Z_- = \tilde{G}^{-1}Z_- - \tfrac{1}{2}\tilde{G}^{-1}(\mathbf{I}+\Gamma)Z_- - \tfrac{1}{2}(\mathbf{I}-\Gamma)\tilde{G}^{-1}Z_- = \tfrac{1}{2}(\mathbf{I}+\Gamma)\tilde{G}^{-1}Z_- \in \mathcal{E}_+(\mathbf{T};L^2(S^2))$. As a result, we find the expression

$$-\sum_{\rho_j<0}\rho_j = \dim\{\mathrm{Ker}(\mathbf{I}+\tilde{K})\cap\mathcal{E}_-(\mathbf{T};L^2(S^2))\}. \tag{5.9}$$

The norm of $\tilde{K}$ in $L^2(\mathbf{T};L^2(S^2))$, i.e., its spectral radius, can be expressed in terms of the gap between certain subspaces [25], [26] For Hilbert spaces the gap between two closed subspaces $\mathcal{M}_1$ and $\mathcal{M}_2$ equals $||P_1-P_2||$ where P_1 and P_2 are the orthogonal projections onto $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Now notice that $\Gamma_\pm = \frac{1}{2}(\mathbf{I}\pm\Gamma)$ are the orthogonal projections of $\mathcal{E}(\mathbf{T};L^2(S^2))$ onto $\mathcal{E}_\pm(\mathbf{T};L^2(S^2))$ and $\Lambda_\pm = \frac{1}{2}\tilde{G}(\mathbf{I}\pm\Gamma)\tilde{G}^{-1}$ are the orthogonal projections of $\mathcal{E}(\mathbf{T};L^2(S^2))$ onto $\tilde{G}[\mathcal{E}_\pm(\mathbf{T};L^2(S^2))]$. In terms of these projections, from (4.2) we obtain

$$\tilde{K} = \Gamma_+ - \Lambda_+ \qquad \text{and} \qquad -\tilde{K} = \Gamma_- - \Lambda_-. \tag{5.10}$$

Hence, for $\mathcal{E} = L^2$ we have $||\tilde{K}|| = \mathrm{gap}\left(\mathcal{E}_\pm(\mathbf{T};L^2(S^2)), \tilde{G}[\mathcal{E}_\pm(\mathbf{T};L^2(S^2))]\right)$.

Using the projections $\Gamma_\pm$ and $\Lambda_\pm$, we will now derive more convenient expressions for the sums of the positive and negative partial indices of $\tilde{G}(\xi)$ than (5.7) and (5.9). Observe from (5.10) that

$$\mathbf{I}-\tilde{K} = \Gamma_- + \Lambda_+ \qquad \text{and} \qquad \mathbf{I}+\tilde{K} = \Gamma_+ + \Lambda_-. \tag{5.11}$$

Then we easily find from (5.7)

$$\sum_{\rho_j>0}\rho_j = \dim\{\mathcal{E}_+(\mathbf{T};L^2(S^2))\cap\tilde{G}[\mathcal{E}_-(\mathbf{T};L^2(S^2))]\}, \tag{5.12}$$

while we obtain from (5.9)

$$-\sum_{\rho_j<0}\rho_j = \dim\{\mathcal{E}_-(\mathbf{T};L^2(S^2))\cap\tilde{G}[\mathcal{E}_+(\mathbf{T};L^2(S^2))]\}. \tag{5.13}$$

We can now prove the following.

THEOREM 5.3. *The following statements are equivalent:*

(1) ±1 *are not eigenvalues of the Fredholm integral operator* $\tilde{K}$ *of* (5.2).

(2) *The Riemann–Hilbert problem* (5.1) *has a unique solution* $Y_\pm \in \mathcal{E}_\pm(\mathbf{T};L^2(S^2))$ *for every* $F \in \mathcal{E}(\mathbf{T};L^2(S^2))$.

(3) *There exists a right canonical Wiener–Hopf factorization*

$$\tilde{G}(\xi) = \tilde{G}_+(\xi)\tilde{G}_-(\xi), \qquad \xi\in\mathbf{T},$$

of $\tilde{G}(\xi)$ *where* $\tilde{G}_\pm(\xi)$ *and* $\tilde{G}_\pm(\xi)^{-1}$ *belong to* $\mathcal{H}_\gamma(\mathbf{T};\mathcal{L}(L^2(S^2)))$ *and have an analytic continuation to* $\mathbf{T}^\pm$, *where* γ *is the constant specified in Proposition* 4.1.

(4) *There exists a left canonical Wiener–Hopf factorization*

$$\tilde{G}(\xi) = \tilde{\mathcal{G}}_-(\xi)\tilde{\mathcal{G}}_+(\xi), \quad \xi\in\mathbf{T},$$

of $\tilde{G}(\xi)$ *where* $\tilde{\mathcal{G}}_\pm(\xi)$ *and* $\tilde{\mathcal{G}}_\pm(\xi)^{-1}$ *belong to* $\mathcal{H}_\gamma(\mathbf{T};\mathcal{L}(L^2(S^2)))$ *and have an analytic continuation to* $\mathbf{T}^\pm$.

(5) *The operator function* $\tilde{G}(\xi)$, *or equivalently the operator function* $G(k)$ *given in* (2.7), *has no partial indices* [17].

(6) *The three-dimensional Jost operator* [5] *exists and is unique* [17].

Proof. The equivalence of (2)–(6), as well as the existence of the left and right Wiener–Hopf factorizations of $\tilde{G}(\xi)$, has been proven in [17]. First, let us show that (1) implies (2). Note that from (5.11) we have

$$(\mathbf{I}-\tilde{K})[\mathcal{E}_{\pm}(\mathbf{T};L^2(S^2))]\subset\tilde{G}[\mathcal{E}_{\pm}(\mathbf{T};L^2(S^2))].$$

Since $\mathbf{I}-\tilde{K}$ is boundedly invertible in the absence of eigenvalues ± 1, it must act as a boundedly invertible operator from $\mathcal{E}_{\pm}(\mathbf{T};L^2(S^2))$ onto $\tilde{G}[\mathcal{E}_{\pm}(\mathbf{T};L^2(S^2))]$. As a result, the unique solution of (5.2) with right-hand side (5.3) belongs to $\mathcal{E}_{+}(\mathbf{T};L^2(S^2))$ for every $F\in\mathcal{E}(\mathbf{T};L^2(S^2))$. Then by Theorem 5.1, we can conclude that (1) implies (2).

To complete the proof of our theorem, it suffices to prove that (3) and (4) together imply (1). Indeed, let (3) and (4) be true. The canonical Wiener–Hopf factorization of $\tilde{G}(\xi)$ exists only when all the partial indices are zero. Thus, from (5.11) and (5.12) we obtain

$$\mathcal{E}_{\pm}(\mathbf{T};L^2(S^2))\cap\tilde{G}[\mathcal{E}_{\mp}(L^2(S^2))]=\{0\}. \tag{5.14}$$

Hence, if $Y\in\mathrm{Ker}(\mathbf{I}-\tilde{K})$, then using (5.11) one obtains

$$\Gamma_{-}Y=-\Lambda_{+}Y\in\mathcal{E}_{-}(\mathbf{T};L^2(S^2))\cap\tilde{G}[\mathcal{E}_{+}(\mathbf{T};L^2(S^2))].$$

and hence, by virtue of (5.14),

$$Y\in\mathcal{E}_{+}(\mathbf{T};L^2(S^2))\cap\tilde{G}[\mathcal{E}_{-}(\mathbf{T};L^2(S^2))],$$

which proves that $\mathrm{Ker}(\mathbf{I}-\tilde{K})=\{0\}$. Hence ± 1 are not eigenvalues of $\tilde{K}$, and our proof is complete. □

The following corollary gives a sufficient condition on the scattering operator so that the Fredholm integral equation (3.10) is uniquely solvable.

COROLLARY 5.4. *The Fredholm integral equation* (3.10) *and the Riemann–Hilbert problem* (2.9) *are uniquely solvable when the scattering operator satisfies* $\sup_{k\in\mathbf{R}}||S(k)-\mathbf{I}||<1$.

Proof. From (4.2) we have $\tilde{K}=\frac{1}{2}(\mathbf{I}-\tilde{G})\Gamma-\frac{1}{2}\tilde{G}\Gamma\tilde{G}^{-1}(\mathbf{I}-\tilde{G})$. Since Γ, $\tilde{G}$, and $\tilde{G}^{-1}$ have unit norms, we then obtain $||\tilde{K}||\leq||\mathbf{I}-\tilde{G}||$. However, we have $||\mathbf{I}-\tilde{G}(\xi)||=||\mathbf{I}-G(k)||=||S(k)-\mathbf{I}||$. Thus, if $\sup_{k\in\mathbf{R}}||S(k)-\mathbf{I}||<1$, we have $||\tilde{K}||<1$, and by Theorem 5.3 both the Riemann–Hilbert problem and the Fredholm integral equation are uniquely solvable.

6. Solution of the inverse problem. Once the Riemann–Hilbert problem posed in (2.9) is solved by solving the Fredholm integral equation (3.10), we obtain $f(k,x,\theta)$ given in (3.2) using (3.8) and (2.8). From the Schrödinger equation (1.1) we then obtain the potential as

$$V(x)=\frac{(\Delta+2ik\theta\cdot\nabla)X_{+}(k,x,\theta)}{1+X_{+}(k,x,\theta)}. \tag{6.1}$$

Note that the right-hand side of this equation contains θ and k whereas these two variables are absent from the left-hand side. Hence the solution of the Riemann–Hilbert problem will lead to a potential only if the right-hand side of (6.1) is independent of θ and k. Below we show that if the so-called miracle condition [2] occurs and the Riemann–Hilbert problem has a unique solution, then the right-hand side of (6.1) is

independent of θ and k and becomes equal to a potential function of x. In the absence of bound states the proof has been given in [17], and here we give the proof when the bound states are present.

From (2.10) we have

$$X_+(k) = X_+^{\mathrm{red}}(k) + [\Pi(k) - \mathbf{I}]\hat{1} + [\Pi(k) - \mathbf{I}]X_+^{\mathrm{red}}(k), \tag{6.2}$$

where we have again suppressed the x-dependence of $X_+(k)$, $X_+^{\mathrm{red}}(k)$, and $\Pi(k)$. Defining

$$\eta(\alpha, x, \theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, X_+(k)\, e^{-ik\alpha}, \tag{6.3}$$

$$\eta^{\mathrm{red}}(\alpha, x, \theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, X_+^{\mathrm{red}}(k)\, e^{-ik\alpha},$$

$$\Omega(\alpha, x, \theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, [\Pi(k) - \mathbf{I}]\, e^{-ik\alpha},$$

we can take the Fourier transform of (6.2) to obtain

$$\eta(\alpha, x, \theta) = \eta^{\mathrm{red}}(\alpha, x, \theta) + \Omega(\alpha, x, \theta)\hat{1} + \int_{-\infty}^{\infty} d\beta\, \Omega(\alpha - \beta, x, \theta)\, \eta^{\mathrm{red}}(\beta, x, \theta). \tag{6.4}$$

Note that $X_+(k) = O(\frac{1}{k})$ as $k \to \pm\infty$ and is analytic in $\mathbf{C}^+$; $\Pi(k) - \mathbf{I} = O(\frac{1}{k})$ as $k \to \pm\infty$ and is analytic in $\mathbf{C}^-$. As a result, $\eta^{\mathrm{red}}(\alpha, x, \theta) = 0$ for $\alpha < 0$ and $\Omega(\alpha, x, \theta) = 0$ for $\alpha > 0$. Thus, we can write (6.4) as

$$\eta(\alpha, x, \theta) = \eta^{\mathrm{red}}(\alpha, x, \theta) + \int_{\alpha}^{\infty} d\beta\, \Omega(\alpha - \beta, x, \theta)\, \eta^{\mathrm{red}}(\beta, x, \theta), \qquad \alpha > 0, \tag{6.5}$$

$$\eta(\alpha, x, \theta) = \Omega(\alpha, x, \theta)\hat{1} + \int_{0}^{\infty} d\beta\, \Omega(\alpha - \beta, x, \theta)\eta^{\mathrm{red}}(\beta, x, \theta), \qquad \alpha < 0. \tag{6.6}$$

In the Newton–Marchenko inversion theory [4], the potential $V(x)$ is obtained from (6.5) and (6.6) as

$$\begin{aligned} V(x) &= -2\theta \cdot \nabla \lim_{\alpha \to 0^+} [\eta(\alpha, x, \theta) - \eta(-\alpha, x, \theta)] \\ &= -2\theta \cdot \nabla \lim_{\alpha \to 0^+} [\eta^{\mathrm{red}}(\alpha, x, \theta) - \Omega(\alpha, x, \theta)]\hat{1}, \end{aligned} \tag{6.7}$$

provided the right-hand side is independent of θ; this θ-independence is known as the "miracle" identity of Newton [4]. If the miracle occurs and the Riemann–Hilbert problem (2.9) has a unique solution, $\eta(\alpha, x, \theta)$ satisfies the equation

$$\left[\Delta - 2\frac{\partial}{\partial \alpha}\theta \cdot \nabla - V(x)\right] \eta(\alpha, x, \theta) = 0. \tag{6.8}$$

Then we would like to show that (6.7) and (6.8) imply (6.1). To see this, note that from (6.3) we have

$$ikX_+(k) = \lim_{\alpha \to 0^+} [\eta(-\alpha, x, \theta) - \eta(\alpha, x, \theta)] - \int_{-\infty}^{\infty} d\alpha\, e^{ik\alpha} \frac{\partial \eta(\alpha, x, \theta)}{\partial \alpha}.$$

Thus, using (6.7) and (6.8) we obtain

$$[\Delta + 2ik\theta \cdot \nabla - V(x)]X_+(k) = V(x) + \int_{-\infty}^{\infty} d\alpha\, e^{ik\alpha} \left[\Delta - 2\frac{\partial}{\partial \alpha}\theta \cdot \nabla - V(x)\right] \eta(\alpha, x, \theta).$$

From this last equation, it is seen that (6.1) is implied by (6.7) and (6.8); hence, whenever the miracle condition of Newton is satisfied and the Riemann–Hilbert problem has a unique solution, the potential of the Schrödinger equation is given by (6.1).

Remark that whenever the scattering operator $S(k)$ is known to have a corresponding potential, it is guaranteed that the right-hand side of (6.1) is independent of k and θ. As a consequence, any of the statements in Theorem 5.3 is sufficient to guarantee that the right-hand side of (6.1) is independent of k and θ.

7. Concluding remarks. The results of this article remain true for any real measurable potential $V(x)$ on $\mathbf{R}^3$ without real exceptional points that leads to a scattering matrix $S(k)$ such that $S(k) - \mathbf{I}$ is compact for all $k \in \mathbf{R}$ and $\tilde{S}(\xi) = S(i(1+\xi)/(1-\xi))$ is Hölder continuous in ξ on $\mathbf{T}$. In that case we may generalize our results here to potentials on $\mathbf{R}^n$ with $n \geq 2$.

Acknowledgments. The authors are indebted to Roger Newton for his comments.

REFERENCES

[1] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, Second ed., Springer-Verlag, New York, 1989.

[2] R. G. Newton, *Inverse scattering.* II. *Three dimensions*, J. Math. Phys., 21 (1980), pp. 1698–1715; 22 (1981), p. 631; 23 (1982), p. 693.

[3] ———, *Inverse scattering.* III. *Three dimensions, continued*, J. Math. Phys., 22 (1981), pp. 2191–2200; 23 (1982), p. 693.

[4] ———, *Inverse scattering.* IV. *Three dimensions: generalized Marchenko construction with bound states*, J. Math. Phys., 23 (1982), pp. 2257–2265.

[5] ———, *The Gel'fand–Levitan method in the inverse scattering problem in quantum mechanics*, Scattering Theory in Mathematical Physics, J.A. Lavita and J.-P. Marchand (Eds.), Reidel, Dordrecht, 1974, pp. 193–225.

[6] A. I. Nachman and M. J. Ablowitz, *A multidimensional inverse scattering method*, Studies in Appl. Math., 71 (1984), pp. 243–250.

[7] R. Beals and R. R. Coifman, *Multidimensional inverse scattering and nonlinear* P.D.E.*'s*, Proc. Sympos. Pure Math., 43 (1985), pp. 45–70.

[8] ———, *The D-bar approach to inverse scattering and nonlinear evolutions*, Phys. D, 18 (1986), pp. 242–249.

[9] R. G. Novikov and G. M. Henkin, *Solution of a multidimensional inverse scattering problem on the basis of generalized dispersion relations*, Sov. Math. Dokl., 35 (1987), pp. 153–157. Dokl. Akad. Nauk. SSSR, 292 (1987), pp. 814–818. (In Russian.).

[10] R. T. Prosser, *Formal solution of inverse scattering problems*, J. Math. Phys., 10 (1969), pp. 1819–1822.

[11] ———, *Formal solution of inverse scattering problems.* II, J. Math. Phys., 17 (1976), pp. 1775–1779.

[12] ———, *Formal solution of inverse scattering problems.* III, J. Math. Phys., 21 (1980), pp. 2648–2653.

[13] ———, *Formal solution of inverse scattering problems.* IV, J. Math. Phys., 23 (1982), pp. 2127–2130.

[14] L. D. Faddeev, *Increasing solutions of the Schrödinger equation,*, Sov. Phys. Dokl. 10 (1965), pp. 1033–1035. Dokl. Akad. Nauk. SSSR, 165 (1965), pp. 514–517. (In Russian.)

[15] *Three-dimensional inverse problem in the quantum theory of scattering*, J. Sov. Math., 5 (1976), pp. 334–396. Itogi Nauki i Tekhniki, 3 (1974), pp. 93–180. (In Russian.).

[16] R. G. Newton, *A Faddeev-Marchenko method for inverse scattering in three dimensions*, Inverse Problems, 1 (1985), pp. 371–380.

[17] T. AKTOSUN AND C. VAN DER MEE, *Inverse scattering problem for the three-dimensional Schrödinger equation and Wiener–Hopf factorization of the scattering operator*, J. Math. Phys., 31 (1990), pp. 2172–2180.

[18] R. G. NEWTON, *Inverse Schrödinger Scattering in Three Dimensions*, Springer-Verlag, New York, 1989.

[19] N. I. MUSKHELISHVILI AND N. P. VEKUA, *The Riemann boundary problem for several unknown functions and its application to systems of singular integral equations,* Trudy Tbilissk. Mat. Inst., 12 (1943), pp. 1–46. (In Russian.)

[20] N. I. MUSKHELISHVILI, *Singular Integral Equations* , Noordhoff, Groningen, 1953. Nauka, Moscow, 1946. (In Russian.)

[21] R. G. NEWTON AND R. JOST, *The construction of potentials from the S-matrix for systems of differential equations*, Nuovo Cimento, 1 (1955), pp. 590–622.

[22] R. G. NEWTON, *Noncentral potentials: The generalized Levinson theorem and the structure of the spectrum*, J. Math. Phys., 18 (1977), pp. 1348–1357.

[23] J. DIESTEL AND J. J. UHL, JR., *Vector Measures*, Math. Surveys 15, Amer. Math. Soc., Providence, 1977.

[24] I. GOHBERG AND N. KRUPNIK, *Einführung in die Theorie der eindimensionalen singulären Integraloperatoren,* Birkhäuser, Basel, 1979. Stiinca, Kishinev, 1973. (In Russian.)

[25] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, 1966.

[26] I. C. GOHBERG AND A. S. MARKUS, *Two theorems on the gap between subspaces of a Banach space,* Uspekhi Mat. Nauk., 14 (1959), pp. 135–140. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 732–753, May 1991

009

THE INVERSE EIGENVALUE PROBLEM WITH FINITE DATA*

DAVID C. BARNES†

Abstract. This work is concerned with the inverse eigenvalue problem for ordinary differential equations such as $y'' + (\lambda - q(x))y = 0$ and with some higher-order generalizations. A classical, well-known inverse problem is to reconstruct the coefficient function $q(x)$ in the differential equation using only spectral data, say $\lambda_n(q)$. Most treatments of this subject require an infinite amount of data which, of course, requires asymptotic formulae for the eigenvalues. Unfortunately, such formulae are sometimes very difficult or impossible to obtain. This work considers a different kind of inverse problem which ignores asymptotic formulae and uses only a *fixed and finite* amount of spectral data. The central problem considered in this paper is that of extracting from the finite amount of spectral data as much information as possible about the coefficient function $q(x)$.

It turns out that to understand such problems, it is necessary to examine the topological foundations of the continuity properties of the eigenvalues. This is done in §1. Based on these continuity theorems, a general theorem is given proving the convergence of some numerical approximations to the solution of the finite inverse problem.

In §2 a numerical algorithm for approximating a coefficient function using only a finite amount of spectral data is given. The method works by minimizing the l_2 norm of the difference between the eigenvalues $\lambda_i(q)$, $i = 1, 2, 3, \cdots, N$ and the spectral data.

Key words. inverse problem, eigenvalue problem, continuous dependence

AMS(MOS) subject classifications. 35B25, 35J25

1. Formulation of the finite inverse problem.

1.1. Introduction.

Consider the eigenvalue problem

$$y'' + (\lambda - q(x))y = 0, \qquad \begin{aligned} a_1y(0) + a_2y'(0) + a_3y(1) + a_4y'(1) &= 0, \\ b_1y(0) + b_2y'(0) + b_3y(1) + b_4y'(1) &= 0. \end{aligned} \tag{1}$$

We suppose that this is a self-adjoint problem and we denote the eigenvalues of (1) by $\lambda_n(q)$. Let $q^*(x)$ represent the unknown coefficient function and suppose spectral data $\boldsymbol{\lambda} = (\Lambda_1, \Lambda_2, \cdots)$ are given so that $\lambda_n(q^*) = \Lambda_n$.

There are (at least) two inverse problems which could be considered for this equation. The first, we call the *infinite inverse problem*; it consists of determining the function $q(x)$ given either an infinite amount of spectral data $\lambda_n(q)$, $n = 1, 2, \cdots \infty$, or at least enough data together with an asymptotic formula which may be used to approximate an infinite amount of data.

Various works have considered the problem of reconstructing $q(x)$ in such cases. The review article by McLaughlin [7] and the references given there provide many examples of such inverse problems. For example, Hald [4] has shown that if two functions q and q^* are close enough together and if both q and q^* are symmetric, $q(x) = q(1 - x)$, then

$$\|q - q^*\|_\infty \leq 2 \cdot 10^{8+38M+11M^2} \sum_{k=0}^{\infty} |\lambda_k - \lambda_k^*|. \tag{2}$$

* Received by the editors July 7, 1989; accepted for publication (in revised form) March 27, 1990. This work was partially supported by the ATT Corporation.

† Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington 99164-2930. E-mail address, BARNES@WSUMATH.BITNET.

Therefore, if the eigenvalues $\lambda_n(q)$ are close enough to those of q^* then the functions q and q^* are close in the $\|\cdot\|_\infty$ norm. This shows that (for symmetric functions) the information contained in an infinite amount of spectral data does, in fact, contain a uniform approximation to the unknown function $q^*(x)$. We will see that the situation is very different for the finite inverse problem. Although (2) is an interesting result, it is not useful for actual numerical estimates since for even small values of M (say $M \approx 1$) the constant in (2) is of order 10^{57}.

Another result of this type does not require the symmetry condition on $q(x)$ but uses the norming constants for (1) which are defined by $\rho_n(q) = \|y_n\|_2^2/y'_n(0)$. If q is close enough to q^*, then McLaughlin [8] has shown

$$\int_0^1 (q(x) - q^*(x))^2\,dx \le K \sum_{n=1}^{\infty} [(\lambda_n(q) - \lambda_n(q^*)]^2 + n^6\,[\rho_n(q) - \rho_n(q^*)]^2\,.$$

This result shows that, even without the symmetry condition on $q(x)$, an infinite amount of spectral data together with all of the norming constants still contains enough information to reconstruct an L_2 approximation to $q^*(x)$. The result of Sacks [10], which proves the L_2 convergence of certain numerical approximations to $q^*(x)$, is also of this general type.

These results can only be used if we have an infinite amount of data or, at least, an asymptotic formula and a sufficient amount of spectral data to make good approximations to the infinite data. In some cases, neither is available. This leads us to consider a second kind of inverse problem which we call the *finite inverse problem.* Here, we assume that only a finite amount of data has been collected and that asymptotic formulae are not available. Clearly, such an inverse problem cannot be solved uniquely since the set of all coefficient functions $q(x)$ is infinite-dimensional, whereas the set of vectors $\boldsymbol{\Lambda} = (\Lambda_1, \Lambda_2, \cdots, \Lambda_N)$ is finite dimensional. Therefore, we understand that to solve such a finite inverse problem simply means that we must produce a function $q(x)$ which has the correct spectral behavior. That is, $\lambda_n(q) = \lambda_n(q^*)$ for $n = 1, 2, 3, \cdots, N$. We must then provide a proper mathematical foundation and interpretation of the results. In particular, we need to know if this process produces an approximation to $q^*(x)$ in some suitable topology. If it does, we need to know something about the topology used and how accurate the approximation is. Theorem 1.2 provides a partial answer to this need. It shows that under certain conditions, as $N \to \infty$, the approximating function $q(x)$ converges to $q^*(x)$ in a certain topology. In addition, Theorem 1.1 indicates what kind of variation in the solution of such an inverse problem is still possible even after all of the requirements of a finite amount of spectral data have been satisfied. These results provide a proper mathematical foundation for the finite inverse problem.

Seidman [11] has studied a similar kind of finite inverse problem. He showed that if a sequence q_N satisfies $\lambda_n(q_k) = \Lambda_k$ for $k = 1, 2, 3, \cdots, N$ and if a certain norm $\|q_N\|_*$ is minimized over some class of functions, then $q_N \to q^*$ in a certain topology on the class of functions. This is analogous to the classical spline interpolation problem for real functions. Theorem 1.2 is also of this type. However, using the compactness of the class $\mathcal{C}(H)$ gives a much easier proof and it does not require the minimizing of a norm to achieve convergence of the interpolating sequence. Of course, such a minimizing condition could be tacked on to Theorem 1.2 if desired. We will also show that the topology used is stronger than that used in [11] so that Theorem 1.2 gives a better approximation to $q^*(x)$.

One of the tools we will need is the following variational characterization of the

eigenvalues using the Rayleigh quotient [3], [12]. Let $\mathcal{L}$ be a self-adjoint operator defined on a dense subspace $\mathbf{D}$ of a separable Hilbert space. Suppose that the lower part of the spectrum of $\mathcal{L}$ consists of isolated eigenvalues $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots$, each having finite multiplicity. Let y_i denote the eigenfunction corresponding to λ_i and let $\mathcal{U}_n$ be the subspace spanned by the first n eigenfunctions $y_1, y_2, \cdots, y_n$. Let $\mathcal{V}_n$ be any other n-dimensional subspace of $\mathbf{D}$. It follows that

$$\lambda_n \leq \max_{u \in \mathcal{V}_n} \frac{(\mathcal{L}u, u)}{(u, u)} \quad \text{and} \quad \lambda_n = \max_{u \in \mathcal{U}_n} \frac{(\mathcal{L}u, u)}{(u, u)}. \tag{3}$$

1.2. The topology of the finite inverse eigenvalue problem. A crucial part of the theory of the finite inverse problem is concerned with a norm $\|\cdot\|_{2L_1}$ which we define by

$$\|q\|_{2L_1} = \left|\int_0^1 q(x)\,dx\right| + \left|\int_0^1\int_0^x q(x_1)\,dx_1\,dx\right| + \int_0^1 \left|\int_0^x\int_0^{x_2} q(x_1)\,dx_1\,dx_2\right| dx. \tag{4}$$

The link between the $2L_1$ norm and eigenvalue problems is provided by the following theorem.

THEOREM 1.1. *Let $\lambda_n(q)$ denote the nth eigenvalue of* (1). *There is a constant $K_n(H)$ (which depends only on n and H) such that for any $q_1, q_2 \in C(H)$ we have*

$$|\lambda_n(q_1) - \lambda_n(q_2)| \leq K_n(H)\,\|q_1(s) - q_2(s)\|_{2L_1}. \tag{5}$$

We will give the proof in §1.3 below. From this, it follows that small changes in the $2L_1$ norm will produce small changes in the eigenvalues. However, it is vital to notice that the modulus of continuity depends on n. Roughly speaking then, the only information about q which can be extracted numerically from a finite number of eigenvalues is an L_1 approximation to the expression $\int_0^x \int_0^t q(s)ds\,dt$. Thus any numerical method that attempts to reconstruct, as its final product, a pointwise approximation to $q(x)$ using a fixed amount of spectral data will, implicitly, hinge on taking two derivatives of an L_1 approximation to $\int_0^x \int_0^t q(s)\,ds\,dt$. This stands in sharp contrast to the results dealing with the L_2 norm (and even the L_∞ norm) quoted above, which use an infinite amount of data.

To see how very different all of these norms really are and just how sharp this contrast is, consider a perturbation in $q^*(x)$ of the form $\Delta q(x) = A\cos Bx$ so that $q(x) = q^*(x) + \Delta q^*(x)$. A simple calculation shows, for large A and B, that $\|\Delta q^*(x)\|_{2l_1} = O(A/B^2)$. Thus, even if A is quite large, the size of the perturbation $A\cos Bx$ will be very small in the $2L_1$ norm if only B is also large. This is somewhat like the Riemann-Lebesgue lemma, in that, if the positive and negative oscillations tend to balance out, then the overall effect will be small. On the other hand, if A is large, then $\Delta q^*(x)$ will be large in the L_2 (or especially the L_∞) norms no matter what B is. Unfortunately, any candidate for $q^*(x)$ (perhaps obtained using some numerical approximation method) could be perturbed in this way and it would still be a good solution to the finite inverse problem. The penalty for dealing with only a finite amount of data and ignoring the infinite amount of high frequency spectral data is that we must put up with such high frequency noise in the determination of $q^*(x)$.

We will show later that, depending on how smooth q is, the eigenvalues may even be continuous in topologies much weaker than that defined by the $2L_1$ norm. Thus we may actually know even less about $q^*(x)$ than a $2L_1$ approximation. Of course, we should use the weakest topology in which all the eigenvalues are continuous since,

otherwise, we would be attempting to extract more information from the spectral data than it contains. This may very well result in ill-conditioned numerical procedures. However, this weak topology seems to be very difficult to determine exactly. In this work, we give some approximations to it.

Next, we will present a general convergence theorem which applies to the numerical solution of both the finite and infinite inverse eigenvalue problems. We first discuss the infinite inverse problem. Given a constant H, let $\mathcal{C}(H)$ be the set of all functions $q(x)$ which satisfy $|q(x)| \leq H$ and let $\mathcal{S}$ be the set of all infinite sequences $(\lambda_1, \lambda_2, \lambda_3, \cdots)$ of numbers. Define a mapping

$$\Phi : \mathcal{C}(H) \to \mathcal{S} \quad \text{by } \Phi(q) = (\lambda_1(q), \lambda_2(q), \lambda_3(q), \cdots),$$

and use the brief notation $\boldsymbol{\lambda}(q) = (\lambda_1(q), \lambda_2(q), \lambda_3(q), \cdots)$. Let $\mathcal{S}(H) \subset \mathcal{S}$ be the range of Φ. Define a topology on $\mathcal{S}(H)$ using the component-wise convergence criteria $\boldsymbol{\lambda}(q_j) \to \boldsymbol{\lambda}(q^*)$ if and only if, for each n, $\lambda_n(q_j) \to \lambda_n(q^*)$ as $j \to \infty$. We now topologize $\mathcal{C}(H)$ using the weakest topology in which all the eigenvalues are continuous. This simply means that $q_j \to q^*$ as $j \to \infty$ if and only if $\lambda_n(q_j) \to \lambda_n(q^*)$ for each n.

We say that a sequence q_N interpolates to the spectral data $\lambda_n(q^*)$ if, for each N, there is an ϵ_N such that

$$|\lambda_i(q_N) - \lambda_i(q^*)| < \epsilon_N \quad \text{for } i = 1, 2, \cdots, N \quad \text{and} \quad \epsilon_N \to 0 \quad \text{as } N \to \infty.$$

The theorem uses the $1Max$ norm which is defined by

$$\|q\|_{1Max} = \max_x \left| \int_0^x q(t)\, dt \right|. \tag{6}$$

THEOREM 1.2. *Given an eigenvalue problem and some spectral data $\lambda_n(q^*)$, suppose that the inverse eigenvalue problem has a unique solution. That is, given any two coefficient functions q_1 and q_2, if $\lambda_n(q_1) = \lambda_n(q_2)$ for all of the data then it follows that $q_1 = q_2$. Let q_N be any sequence of functions which interpolates to the data $\lambda_n(q^*)$. Then*

$$\|q_N - q^*\|_{1Max} \to 0 \quad \text{as } N \to \infty.$$

Furthermore, the range space $\mathcal{S}(H)$ is compact in the topology of component-wise convergence.

Proof. Consider the set of all functions $Q(x) = \int_0^x q(t)\, dt$ for $q \in \mathcal{C}(H)$. Such functions $Q(x)$ will satisfy the Lipschitz condition $|Q(x) - Q(y)| \leq H|x - y|$. Applying the Ascoli theorem shows that the space $\mathcal{C}(H)$ is compact in the $1Max$ norm. We will show below that the eigenvalues are continuous functionals in the $1Max$ norm. Let q_{N_k} denote a convergent subsequence, say $q_{N_k} \to \widehat{q}$. Then, if $N_k \geq n$, it follows that $|\lambda_n(q_{N_k}) - \lambda_n(q^*)| < \epsilon_{N_k}$. Letting $k \to \infty$ gives $\lambda_n(\widehat{q}) = \lambda_n(q^*)$. Uniqueness implies that $q^* = \widehat{q}$, so every convergent subsequence converges to q^*. Thus the original sequence q_N must converge to q^*.

We will now show that $\mathcal{S}(H)$ is compact. Unique solvability of the inverse problem implies that the mapping Φ has an inverse—say $\Psi : \mathcal{S}(H) \to \mathcal{C}(H)$. Since Φ and Ψ are one-to-one, the weak topology induced on $\mathcal{C}(H)$ by Φ and the strong topology induced on $\mathcal{C}(H)$ by Ψ are the same and both Φ and Ψ are homeomorphisms [2]. Since $\mathcal{C}(H)$ is compact in the $1Max$ topology, which is stronger than the weak topology, it follows that $\mathcal{C}(H)$ is compact in the weak topology. Thus $\mathcal{S}(H)$ is also compact. □

The theorem shows that an interpolating sequence converges. In fact, it shows that an infinite amount of spectral data contains a uniform approximation to $\int_0^x q^*(t)\,dt$ having arbitrarily high accuracy, however, the results [8], [4] show that such data actually contains much more than this. Our previous remarks regarding the information content of a finite amount of spectral data are, of course, still valid, since this theorem may require an arbitrarily large amount of data to produce such an approximation to $\int_0^x q(t)\,dt$. Even with such a large amount of data, this uniform approximation must still be differentiated in order to construct a pointwise approximation to $q(x)$.

Consider again the perturbation $\Delta q^*(x) = A\cos Bx$. It is easy to see that by choosing appropriate constants A and B, an example can be constructed which is small in the $2L_1$ norm but is not small in the $1Max$ norm. Even after the spectral data has been satisfied, a good $1Max$ approximation to $q^*(x)$ may not be obtained. However, Theorem 1.2 provides a (perhaps rough) $1Max$ approximation to q^* while Theorem 1.1 shows how much variability is still left in the unknown function $q^*(x)$ after using a finite amount of the data. In this sense, we have upper and lower bounds on the information content of the spectral data. Some of these compactness and continuity ideas have been used by Krein [5] to prove the existence of certain functions which maximize an eigenvalue.

We will now consider the finite inverse problem. Since the solution of this inverse problem is not unique, we must consider the isospectral equivalence classes on $\mathcal{C}(H)$. Call this set of classes $\mathcal{C}_N(H)$. Theorem 1.2 shows that if the infinite inverse problem has a unique solution then the diameter (as measured in the $1Max$ norm) of each equivalence class tends to zero as $N\to\infty$. Now $\mathcal{C}_N(H)$ inherits a natural topology from $\mathcal{C}(H)$, and there are natural extensions of Φ and Ψ to this context. These extensions will be both continuous and one-to-one. Thus, $\mathcal{C}_N(H)$ is homeomorphic to a subset of ordinary N-dimensional Euclidian space. This determines the weakest topology on $\mathcal{C}_N(H)$ in which the eigenvalues are continuous. Unfortunately, knowing the topology on $\mathcal{C}_N(H)$ still does not easily translate into usable conditions on the coefficient functions q which, of course, belong to $\mathcal{C}(H)$, not to $\mathcal{C}_N(H)$. The following section considers this very difficult problem.

1.3. Approximating the weak topology. We are now interested in extracting from the spectral data whatever information it contains about $q^*(x)$. Theorem 1.2 may provide only a rough approximation to $\int_0^x q(t)\,dt$ since we may not have the data to take N large enough. It would be best, of course, if we could completely characterize the weak topology on $\mathcal{C}(H)$ in terms of conditions on the functions $q(x)$. This seems to be very difficult to do, but we will give some partial answers. First, we will give the proof of Theorem 1.1.

Proof. Let $\mathcal{U}_n(q)$ denote the space spanned by the first n-eigenfunctions of (1) and use the notation $\Delta q(x) = q_1(x) - q_2(x)$, $\Delta\lambda_n = \lambda_n(q_1) - \lambda_n(q_2)$. Integrating by parts, we see that

$$\begin{aligned}\int_0^1 \Delta q(x)y^2(x)\,dx = y^2(1)\int_0^1 \Delta q(x_1)\,dx_1 \\ -y^{2\prime}(1)\int_0^1\int_0^{x_2} \Delta q(x_1)\,dx_1\,dx_2 \\ +\int_0^1 \left(y^2(x)\right)''\int_0^x\int_0^{x_2} \Delta q(x_1)\,dx_1\,dx_2\,dx.\end{aligned}$$

We will show below that there is a constant $K_n(H)$ such that for all $x \in [0,1]$, for all functions $q \in \mathcal{C}(H)$, and for all functions $y \in \mathcal{U}_n(q)$, it follows that y^2, y'^2, and y''^2 are uniformly[1] bounded by some term of the form $K_n(H)\int_0^1 y^2\,dx$. Substituting these bounds into the integration by parts formula and using (3) to compute the Rayleigh quotient for (1), we find that

$$\frac{-yy'\Big|_{x=0}^{x=1} + \int_0^1 y'^2 - q_2(x)y^2\,dx}{\int_0^1 y^2\,dx} \le \frac{-yy'\Big|_{x=0}^{x=1} + \int_0^1 y'^2 - q_1(x)y^2\,dx}{\int_0^1 y^2\,dx} + K_n(H)\,\|\Delta q\|_{2L_1}.$$

Next, take the maximum over all functions $y \in \mathcal{U}_n(q_1)$ and use the inequality (3). We find that $\lambda_n(q_2) \le \lambda_n(q_1) + K_n(H)\,\|\Delta q\|_{2L_1}$. This inequality, together with the one obtained by reversing the roles of q_1 and q_2, gives (5).

We need to prove the uniform boundedness property of y, y', and y'' for all $y \in \mathcal{U}_n(q)$. Let y_n be an eigenfunction of (1) and suppose that $(y,y) = (y_n,y_n) = 1$. We see that

$$(7) \qquad |y_n'(x)| = \left|y_n'(\xi) + \int_\xi^x y_n''(s)\,ds\right| \le |y_n'(\xi)| + \int_0^1 |(\lambda - q(s))y_n(s)|\,ds.$$

If y_n is not a monotone function, then we will select ξ to be some point for which $y_n'(\xi) = 0$. Otherwise, we may assume that y_n is an increasing function and there is a number m for which $y'(x) \ge m > 0$. It follows that there is some number c for which $y^2(x) \ge m^2(x-c)^2$. Thus, $1 = \int_0^1 y_n^2\,dx \ge \int_0^1 m^2(x-\frac{1}{2})^2\,dx = m^2/12$ with equality when $y_n = m(x-\frac{1}{2})$. Therefore, we must have $m < 4$. In this event, we select ξ to be some point for which $y_n'(\xi) < 4$. In either case, we obtain $|y_n'(x)| \le 4 + \int_0^1 |(\lambda - q(x))y_n(s)|\,ds$. Since $|q(x)| \le H$, it follows that $\lambda_n(q)$ is uniformly bounded, so the Schwarz inequality shows that $|y_n'(x)|$ is uniformly bounded. Thus, $|y_n|$ is uniformly bounded and (1) shows that y_n'' is also uniformly bounded. Since y is normalized and is a linear combination of the functions y_n, the theorem follows. □

It is well known (see [7] and the references given there) that two different sets of eigenvalues corresponding to different boundary conditions are sufficient to determine $q^*(x)$ uniquely. Furthermore, the results of McLaughlin and Rundell [9] show that if a fixed eigenvalue—say $\lambda_2(q^*)$—is known for an infinite number of distinct boundary conditions of the form $y(0) = y'(1) + \beta y(1) = 0$, then the function $q^*(x)$ is uniquely determined. However, the constant $K_n(H)$ of Theorem 1.1 is independent of the boundary conditions. It follows that no finite amount of spectral data corresponding to any number of different boundary conditions can reproduce an accurate pointwise approximation to the coefficient function $q(x)$. At most, the only information contained in such data is an L_1 approximation to $\int_0^x \int_0^s q(t)\,dt\,ds$ together with approximations of the two quantities $\int_0^1 q(x)\,dx$ and $\int_0^1 \int_0^x q(s)\,ds\,dx$.

Incidentally, it is easy to see that approximating these two quantities is equivalent to approximating the two Hausdorff moments $\int_0^1 q(x)\,dx$ and $\int_0^1 xq(x)\,dx$. That is, the norm

$$(8) \qquad \|q\|_{2H_1} = \left|\int_0^1 q(x)\,dx\right| + \left|\int_0^1 xq(x)\,dx\right| + \int_0^1 \left|\int_0^x \int_0^{x_2} q(x_1)\,dx_1\,dx_2\right| dx$$

[1] Throughout this work, "uniformly" refers to a property which holds for all functions $q \in \mathcal{C}(H)$ and for all of the functions $y \in \mathcal{U}_n(q)$, not just uniformly for $x \in [0,1]$.

is topologically equivalent to the norm $\|q\|_{2L_1}$. Furthermore, if boundary conditions for which either y or y' vanish at the endpoints are used, then the boundary terms in the integration by parts formula drop out. Thus, the first two terms in the $2L_1$ norm may be negelected and the much simpler norm

$$\|q\|^*_{2L_1} = \int_0^1 \left| \int_0^x \int_0^{x_2} q(x_1)\, dx_1\, dx_2 \right| dx$$

may be used.

Next, suppose that the coefficient functions q are all of uniformly bounded variation. This is only a slightly stronger assumption and still includes most of the applications of the inverse problem. It follows that y_n'' will also be of bounded variation and that we may now integrate by parts a third time to obtain

$$\begin{aligned}\int_0^1 \Delta q(x) y^2(x)\, dx = {} & y^2(1) \int_0^1 \Delta q(x_1)\, dx_1 - y^{2\prime}(1) \int_0^1 \int_0^{x_2} \Delta q(x_1)\, dx_1\, dx_2 \\ & + y^{2\prime\prime}(1) \int_0^1 \int_0^{x_3} \int_0^{x_2} \Delta q(x_1)\, dx_1\, dx_2\, dx_3 \\ & - \int_0^1 \int_0^x \int_0^{x_3} \int_0^{x_2} \Delta q(x_1)\, dx_1\, dx_2\, dx_3\, d(y^{2\prime\prime}).\end{aligned}$$

We will now use the norm defined by

$$\|q\|_{3Max} = \left| \int_0^1 q(x)\, dx \right| + \left| \int_0^1 \int_0^x q(x_1)\, dx_1\, dx \right| + \max_x \left| \int_0^x \int_0^{x_3} \int_0^{x_2} \Delta q(x_1)\, dx_1\, dx_2\, dx_3 \right|.$$

Note that $\|q\|_{3Max} \le \|q\|_{2L_1}$. It is then easy to prove the following theorem.

THEOREM 1.3. *Let V be a constant and let $\mathcal{C}(H, V)$ denote the class of functions $q \in \mathcal{C}(H)$ which have total variation at most V. Then there is a constant $K_n(H, V)$ such that, for any $q_1,\ q_2 \in \mathcal{C}(H, V)$, it follows that*

$$|\lambda_n(q_1) - \lambda_n(q_2)| \le K_n(H, V) \|q_1 - q_2\|_{3Max}. \tag{9}$$

Thus, we see that a finite amount of spectral data can yield, at most, an approximation to the first two Hausdorff moments and a uniform approximation to the quantity $\int_0^x \int_0^s \int_0^t q(z)\, dz\, dt\, ds$. Generating a pointwise approximation to $q(x)$ requires taking three derivatives of this uniform approximation.

If we use the subset of $\mathcal{C}(H)$ consisting of coefficient functions q which have bounded derivatives—say $|q'(x)| \le H_1$—then we can prove that

$$|\lambda_n(q_1) - \lambda_n(q_2)| \le K_n(H, H_1, V) \|q_1 - q_2\|_{3L_1}$$

where

$$\begin{aligned}\|q\|_{3L_1} = {} & \left| \int_0^1 q(x)\, dx \right| + \left| \int_0^1 \int_0^{x_2} q(x_1)\, dx_1\, dx_2 \right| \\ & + \left| \int_0^1 \int_0^{x_3} \int_0^{x_2} q(x_1)\, dx_1\, dx_2\, dx_3 \right| + \int_0^1 \left| \int_0^{x_4} \int_0^{x_3} \int_0^{x_2} q(x_1)\, dx_1\, dx_2\, dx_3 \right| dx_4.\end{aligned}$$

Generating a pointwise approximation to $q(x)$ requires taking three derivatives of an L_1 approximation to $\int_0^x \int_0^s \int_0^t q(z)\, dz\, dt\, ds$. Thus, we see a hierarchy of continuity conditions building. As q becomes smoother, we can build weaker topologies. We will investigate this idea further.

1.4. Some connections with weak convergence and other topologies. Define a norm similar to (4) but using only one integration by parts,

$$\|q\|_{1L_1} = \left|\int_0^1 q(x)\,dx\right| + \int_0^1 \left|\int_0^x q(x_1)\,dx_1\right| dx.$$

It is easy to show that $\|q\|_{2L_1} \le 2\|q\|_{1L_1}$. Thus, the norm $\|q\|_{2L_1}$ is weaker than $\|q\|_{1L_1}$. More generally, if the coefficient functions have $N-2$ uniformly bounded derivatives, then we can integrate by parts N times. Considering an appropriate norm $\|q\|_{(N+2)L_1}$ will then give continuity of the eigenvalues in a whole sequence of weaker topologies. If q is analytic, the process can be continued forever to give an infinite sequence of ever weaker topologies.

The norms used in these continuity results can be related to the concept of weak convergence using some fundamental facts of measure theory. For example, the following theorem holds.

THEOREM 1.4. *Let q, $q_n \in \mathcal{C}(H)$. The sequence $\int_0^x q_n(t)\,dt$ converges pointwise to the function $\int_0^x q(t)\,dt$ if and only if $(q_n, g)\to(q, g)$ for all functions $g \in L_2[0,1]$.*

The proof of this theorem hinges on taking $g(x)$ to be the characteristic function of the interval $[0, x]$. Therefore, the topology of weak convergence of $q_n(x)$ fits between the topologies defined by the two norms $\int_0^1 |\int_0^x q(t)\,dt|\,dx$ and $\max_x |\int_0^x q(t)\,dt|$.

Construct a topology τ on $\mathcal{C}(H)$ so that convergence of the sequence $q_n(x)\to q(x)$ means that $(q_n, G)\to(q, G)$ for every function G which satisfies $\int_0^1 G^2 + G'^2\,dx < \infty$. One of Seidman's [11] results is that the eigenvalues are continuous in the topology τ. We will now show that the $2L_1$ norm defines a topology weaker than τ and that the $1Max$ norm is stronger that τ.

Let $g \in L_2[0,1]$ and set $G(x) = \int_x^1 g(t)\,dt$. We see that $(q_n, G) = (\int_0^x q_n(t)\,dt, g)$. Thus, if $q_n\to q$ in the topology τ, then $\int_0^x q_n(t)\,dt$ converges weakly to $\int_0^x q(t)\,dt$. By Theorem 1.4, $\int_0^x \int_0^{x_1} q_n(t)\,dt\,dx_1$ converges pointwise to $\int_0^x \int_0^{x_1} q(t)\,dt\,dx_1$. Pointwise convergence implies L_1 convergence in $\mathcal{C}(H)$. Thus convergence in τ implies convergence in the norm $\int_0^1 |\int_0^x \int_0^{x_1} q_n(t)\,dt\,dx_1|\,dx$. Furthermore, we may take $G = 1$ and $G = x$ to show that

$$\int_0^1 q_n(x)\,dx \to \int_0^1 q(x)\,dx \quad \text{and} \quad \int_0^1\int_0^{x_2} q_n(x_1)dx_1\,dx_2 \to \int_0^1\int_0^{x_2} q(x_1)dx_1\,dx_2.$$

We now make use of (8) to finish the proof. Thus, Theorem 1.1 shows continuity of the eigenvalues in a topology weaker than τ. Theorem 1.4 can be easily adapted to show that $1Max$ is stronger than τ.

1.5. More general second-order equations. We may also apply these methods to the self-adjoint equation

$$y'' + \lambda p(x) y = 0, \qquad \begin{aligned} &a_1 y(0) + a_2 y'(0) + a_3 y(1) + a_4 y'(1) = 0,\\ &b_1 y(0) + b_2 y'(0) + b_3 y(1) + b_4 y'(1) = 0. \end{aligned} \tag{10}$$

We will need to use a slightly different class of coefficient functions $p(x)$ that are defined by

$$\mathcal{C}(H_1, H_2) = \{p(x) \mid H_1 \le p(x) \le H_2\} \qquad \text{where } H_1 > 0.$$

We have the following theorem.

THEOREM 1.5. *Let $\lambda_n(p)$ denote the nth eigenvalue of* (10). *There is a constant $K_n(H)$ which depends only on n, H_1, and H_2 such that for any $p_1, p_2 \in \mathcal{C}(H_1, H_2)$ we have*

$$|\lambda_n(p_1) - \lambda_n(p_2)| \leq K_n(H)\, \|p_1(s) - p_2(s)\|_{2L_1} \ . \tag{11}$$

Proof. First, we consider the Rayleigh quotient, $\mathcal{Q}(p, y) = (Ay, y)/(y, y)$, for (10). A short calculation yields

$$\mathcal{Q}(p, y) = \frac{\displaystyle\int_0^1 y'^2\, dx - yy'\Big|_{x=0}^{x=1}}{\displaystyle\int_0^1 p(x) y^2\, dx}.$$

A further calculation yields

$$\mathcal{Q}(p_1, y) = \mathcal{Q}(p_2, y) \left[1 + \frac{\displaystyle\int_0^1 \Delta p(x) y^2\, dx}{\displaystyle\int_0^1 p_1(x) y^2\, dx} \right] \qquad \text{where } \Delta p(x) = p_2(x) - p_1(x).$$

Next we develop an inequality for $\Delta p(x)$ just like we did for $\Delta q(x)$ in §1.3. It follows that there exists a constant A such that, uniformly for all functions $y \in \mathcal{U}_n$,

$$\int_0^1 \Delta p(x) y^2\, dx \leq A \|\Delta p(x)\|_{2L_1} \int_0^1 y^2\, dx.$$

Using this and the relation $p_1(x) \geq H_1$, we see that

$$\mathcal{Q}(p_1, y) \leq \mathcal{Q}(p_2, y) \left[1 + \frac{A}{H_1} \|\Delta p(x)\|_{2L_1} \right].$$

Applying the variational relation (3), we find that

$$\lambda_n(p_1) \leq \lambda_n(p_2) \left[1 + \frac{A}{H_1} \|\Delta p(x)\|_{2L_1} \right].$$

This result together with the one obtained by reversing the roles of p_2 and p_1 can be easily manipulated to finish the proof. □

Next, we will apply these methods to the equation

$$(r(x)y')' + \lambda y = 0, \qquad \begin{aligned} &a_1 y(0) + a_2 r(0) y'(0) + a_3 y(1) + a_4 r(1) y'(1) = 0, \\ &b_1 y(0) + b_2 r(0) y'(0) + b_3 y(1) + b_4 r(1) y'(1) = 0. \end{aligned} \tag{12}$$

We assume that this is a self-adjoint system. Integration by parts may be used on the expression $\int_0^1 y^2 \Delta r(x)\, dx$ just as in Theorem 1.2. In this case, y will be continuous but y' will have discontinuities if r does. We must also take account of the values of r at the endpoints since they occur in the boundary conditions. Given constants H, h, V, A, B, let $\mathcal{R}$ be the class of functions $r(x)$ which satisfy the following conditions:

1. $H \geq r(x) \geq h > 0$.
2. $r(x)$ has total variation at most V.
3. $r(x)$ has given values at the endpoints, $\lim_{x \downarrow 0} r(x) = A$, $\lim_{x \uparrow 1} r(x) = B$.

Using the $1Max$ norm we obtain the following theorem.

THEOREM 1.6. *Let $\lambda_n(r)$ denote the nth eigenvalue of* (12). *There is a constant $K = K_n(H, V, h)$ such that for any r_1, $r_2 \in \mathcal{R}(H, h, V)$, it follows that*

$$|\lambda_n(r_1) - \lambda_n(r_2)| \leq K\|r_1 - r_2\|_{1Max}. \tag{13}$$

Proof. Let y_n be the nth eigenfunction of (12). For any x, $\xi \in [0,1]$ we see that

$$r(x)y_n'(x) = r(\xi)y_n'(\xi) + \int_\xi^x (r(s)y_n'(s))'\,ds = r(\xi)y_n'(\xi) - \int_\xi^x \lambda y_n(s)\,ds. \tag{14}$$

Either the term $r(\xi)y_n'(\xi)$ is of one sign or not. If not, we may assume $r(\xi)y_n'(\xi) = 0$ for some ξ. If it is of one sign, we may assume that $r(\xi)y_n'(\xi) \geq m > 0$ for some constant m. As in Theorem 1.1, it follows that $m^2 \leq 12H^2$. The uniform boundedness properties follow from this and (14). We note that the differential equation shows

$$y'(x) = \frac{r(0)y'(0) - \int_0^x \lambda y(s)\,ds}{r(x)}.$$

Since $r(x)$ is of uniformly bounded variation, this equation shows that y_n' is also of uniformly bounded variation. So the following Stieltjes integral formula is valid:

$$\int_0^1 y_n'^2(\Delta r(x))\,dx = (y_n'^2)(1)\int_0^1 (\Delta r(x))\,dx - \int_0^1\int_0^x \Delta r(x)\,d(y_n'^2).$$

We now use this formula and the bounds on y to construct the Rayleigh quotient for (12) and obtain the relationship

$$\frac{-yr_1y'\Big|_{x=0}^{x=1} + \int_0^1 r_1(x)y'^2\,dx}{\int_0^1 y^2\,dx} \leq \frac{-yr_2y'\Big|_{x=0}^{x=1} + \int_0^1 r_2(x)y'^2\,dx}{\int_0^1 y^2\,dx} + K\,\|r_1(x) - r_2(x)\|_{1Max}.$$

The theorem follows, using (3). □

It is clear that these three examples can be extended to obtain theorems dealing with the most general Sturm–Liouville equation $(r(x)y')' + (\lambda p(x) - q(x))y = 0$. It is also clear that Theorem 1.2 can be generalized to include these equations.

1.6. Higher-order equations. Consider the fourth-order equation

$$y'''' - (\lambda - q(x))y = 0. \tag{15}$$

Suppose that four boundary conditions are given and that the resulting problem is self-adjoint. Since the eigenfunctions of such an equation are necessarily smoother than those for the second-order case, we may integrate by parts more times and

obtain continuity in much weaker topologies. The required formula is

$$\begin{aligned}\int_0^1 y^2\Delta q(x)\,dx &= y^2(1)\int_0^1 \Delta q(x_1)\,dx_1 - y^{2\prime}(1)\int_0^1\int_0^{x_2} \Delta q(x_1)\,dx_1\,dx_2\\ &\quad + y^{2\prime\prime}(1)\int_0^1\int_0^{x_3}\int_0^{x_2} \Delta q(x_1)\,dx_1\,dx_2\,dx_3\\ &\quad - y^{2\prime\prime\prime}(1)\int_0^1\int_0^{x_4}\int_0^{x_3}\int_0^{x_2} \Delta q(x_1)\,dx_1\,dx_2\,dx_3\,dx_4\\ &\quad + \int_0^1\int_0^{x}\int_0^{x_4}\int_0^{x_3}\int_0^{x_2} y^{2\prime\prime\prime\prime}\Delta q(x_1)\,dx_1\,dx_2\,dx_3\,dx_4\,dx.\end{aligned}$$

The new norm is given by

$$\begin{aligned}\|q\|_{4L_1} &= \left|\int_0^1 q(x)\,dx\right| + \left|\int_0^1\int_0^{x_2} q(x_1)\,dx_1\,dx_2\right| + \left|\int_0^1\int_0^{x_3}\int_0^{x_2} q(x_1)\,dx_1\,dx_2\,dx_3\right|\\ &\quad + \left|\int_0^1\int_0^{x_4}\int_0^{x_3}\int_0^{x_2} q(x_1)\,dx_1\,dx_2\,dx_3\,dx_4\right|\\ &\quad + \int_0^1\left|\int_0^{x}\int_0^{x_4}\int_0^{x_3}\int_0^{x_2} q(x_1)\,dx_1\,dx_2\,dx_3\,dx_4\right| dx.\end{aligned}$$

Methods like those used for Theorem 1.1 yield the following theorem.

THEOREM 1.7. *Let $\lambda_n(q)$ denote the nth eigenvalue of* (15) *subject to self-adjoint boundary conditions. Then there is a constant $M_n(H)$ such that, for any q_1, $q_2 \in C(H)$, we have*

$$|\lambda_n(q_1) - \lambda_n(q_2)| \le M_n(H)\|q_1 - q_2\|_{4L_1}. \tag{16}$$

The constant $M_n(H)$ is independent of the boundary conditions.

If we are willing to assume that the coefficients are of bounded variation, then we can integrate once more and imitate the procedure given in Theorem 1.2 above. Thus, the only information contained in a finite amount of spectral data for this fourth-order problem is a uniform approximation to the quantity $\int_0^x \int\int\int\int q$ together with approximations to the first four Hausdorff moments. Reconstructing a pointwise approximation to the coefficient function implicitly requires the numerical computation of five derivatives of this uniform approximation.

A generalization of Theorem 1.2 can also be obtained for higher-order equations.

2. Solving the finite inverse problem. We will now provide a numerical method for solving the finite inverse problem. We will also give a few examples and an analysis of the method. We will only consider the second-order equation (1) in any detail, although the methods will work for the general equation Sturm–Louville equation.

$$(r(x)y')' + (\lambda p(x) - q(x))y = 0, \qquad \begin{aligned} a_1y(0) + a_2y'(0) + a_3y(1) + a_4y'(1) &= 0,\\ b_1y(0) + b_2y'(0) + b_3y(1) + b_4y'(1) &= 0.\end{aligned} \tag{17}$$

The spectral data need not consist of the eigenvalues corresponding to one set of boundary conditions listed in increasing order. The data may, for example, consist of eigenvalues corresponding to several different sets of boundary conditions interlaced

in some manner or, perhaps, data like that considered by McLaughlin and Rundell [9].

This method works by minimizing the norm $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(q)\|_2$ so that, if the inverse problem has nonunique solutions, then we can still produce an approximation to one of them. Even if there is no solution at all, then our method may produce a function which minimizes the norm $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(q)\|_2$. We will need the following variational formula for the functionals $\boldsymbol{\lambda}(p, q, r)$ given by Barnes [1].

THEOREM 2.1. *For coefficient functions p, q, r and p^*, q^*, r^*, let y_i, λ_i and y_i^*, λ_i^* be corresponding eigenpairs for* (17). *Suppose that y_i^* is normalized so that $\int_0^1 p^* y_i^{*2}\, dx = 1$. For two sets of coefficient functions p^*, q^*, r^* and p, q, r define a functional $J(p, q, r)$ on the set $C(H)$ and define a boundary term BT by*

$$
\begin{aligned}
J(p, q, r) &= \int_0^1 [\lambda_i^* p y_i^{*2} + q y_i^{*2} - r y_i^{*'2}]\, dx, \\
BT &= r^* y_i^{*'} y_i - r y_i' y_i^* - r^* y_i^{*'} y_i^* \Big|_{x=0}^{x=1}.
\end{aligned} \tag{18}
$$

Then

$$
\lambda_i(p, q, r) = \lambda_i(p^*, q^*, r^*) + BT - J(p, q, r) + O_2, \tag{19}
$$

where the term O_2 is defined by

$$
O_2 = \int_0^1 [p^* y_i^* \Delta\lambda_i \Delta y_i + \lambda_i^* y_i^* \Delta p \Delta y_i + y_i^* \Delta q \Delta y_i - y_i^{*'} \Delta r \Delta y_i']\, dx.
$$

This theorem shows that the two functionals $\lambda_i(p, q, r)$ and $\lambda_i(p^*, q^*, r^*) + BT - J(p, q, r)$ are tangent to each other at $p = p^*, q = q^*$, and $r = r^*$ when considered as functionals depending on (p, q, r).

2.1. Minimizing the norm $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(q)\|_2$. In order to simplify the presentation, we will restrict the following study to the simple equation (1) together with boundary conditions for which either y or y' vanish at both ends of the interval. However, it would be easy to generalize to other cases. Theorem 1.2 requires a sequence $q_N(x)$ which interpolates to the data. We will look for one in the class $C(M, H)$ of all functions of the general form

$$
q(x) = \sum_{k=1}^{M} \beta_k h_k(x) \quad \text{with } |q(x)| \le H. \tag{20}
$$

Here, the $h_k(x)$ can be any conveniently chosen basis functions and the β_k are constants. With this kind of representation we will sometimes use the notation $\boldsymbol{\lambda}(q) = \boldsymbol{\lambda}(\boldsymbol{\beta})$ where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_M)$.

For a given value of N, we will select a value of M (usually $M > N$) and take q_N^* to be a solution to the problem

$$
\min_{q \in C(M,H)} \|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(q)\|_2 = \|\boldsymbol{\lambda} - \boldsymbol{\lambda}(q_N^*)\|_2 \qquad \text{where} \quad q_N^*(x) = \sum_{k=1}^{M} \beta_k^* h_k(x). \tag{21}
$$

Such a function always exists since there are only a finite number of the β_k. If the inverse problem does in fact have a unique solution, and if $N \to \infty$, then the sequence of minimums must approach zero. It follows that the functions q_N^* interpolate to the data. Therefore, $q_N^* \to q^*$ in the $1Max$ norm. If the solution of the inverse problem exists but is not unique, then the compactness of $\mathcal{C}(M, H)$ shows that a convergent subsequence of q_N^* can still be selected. Even if there is no solution of the inverse problem, there will still be a subsequence which will converge to some function which gives (at least a local) minimum to the norm $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(q)\|_2$. Under the circumstances, this gives the best possible solution of the inverse problem.

The choice of M should be made large enough to allow $q_N^*(x)$ to approximate $q^*(x)$, which will make the norms (21) small. On the other hand, M should not be so large that it becomes difficult just to evaluate $q_N^*(x)$. According to Theorem 1.2, it does not make any real difference how large M is since we will always have at least a subsequence of q_N^* converging to q^*. In the numerical examples given below, selecting a value of M in the range $N \leq M \leq 2N$ seems to work well. We will give additional reasons for this choice later.

We will now show how to minimize the norm $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(q)\|_2$. We first use (20) and (21) in the variational formula given by Theorem 2.1 to obtain the principal linear part of the eigenvalue functional $\lambda_i(q)$ in the form

$$
\begin{aligned}
&\lambda_i(q) = \lambda_i(q^*) - \int_0^1 q^*(x) y_i^{*2}\, dx + \int_0^1 q(x) y_i^{*2}\, dx + O_2, \\
&\lambda_i(q) = \lambda_i(q^*) - \sum_{k=1}^{M} \beta_k^* \int_0^1 h_k(x) y_i^{*2}\, dx + \sum_{k=1}^{M} \beta_k \int_0^1 h_k(x) y_i^{*2}\, dx + O_2.
\end{aligned} \tag{22}
$$

We now recast these equations in matrix form. Let data $\boldsymbol{\Lambda} = (\Lambda_1, \Lambda_2, \cdots, \Lambda_N)$ be given and let y_i^{*2} be the eigenfunctions corresponding to $\lambda_i(\boldsymbol{\beta}_N^*)$ and define a matrix Γ and a vector $\boldsymbol{R}$ by

$$
\Gamma = \left(\int_0^1 h_k(x) y_i^{*2}\, dx \right)_{N \times M}, \qquad \boldsymbol{R} = \boldsymbol{\Lambda} - \boldsymbol{\lambda}(\boldsymbol{\beta}_N^*) + \Gamma \boldsymbol{\beta}_N^*. \tag{23}
$$

We see that

$$
\boldsymbol{\lambda}(\boldsymbol{\beta}) = \boldsymbol{\lambda}(\boldsymbol{\beta}^*) - \Gamma \boldsymbol{\beta}^* + \Gamma \boldsymbol{\beta} + \boldsymbol{O}_2. \tag{24}
$$

We will now substitute this equation into the norm function. This converts the problem of minimizing a quantity involving the difficult functional $\lambda_i(q)$ into a sequence of simple least squares problems. The mathematical details of this conversion are based on the following theorem.

THEOREM 2.2. *For fixed values of N, M, let $\boldsymbol{\beta}_N^*$ be an approximate solution of the inverse problem in the sense that it minimizes the norm $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(\boldsymbol{\beta})\|_2$. Then the choice $\boldsymbol{\beta} = \boldsymbol{\beta}_N^*$ also minimizes the norm $\|\boldsymbol{R} - \Gamma \boldsymbol{\beta}\|_2$. That is,*

$$
\|\boldsymbol{R} - \Gamma \boldsymbol{\beta}_N^*\|_2 \leq \|\boldsymbol{R} - \Gamma \boldsymbol{\beta}\|_2. \tag{25}
$$

Conversely, if some vector $\boldsymbol{\beta}_N^*$ satisfies the minimum condition (25), *then it must be an extremal of the norm $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(\boldsymbol{\beta})\|_2$.*

Proof. Using (24), we find that

$$
\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(\boldsymbol{\beta})\|_2 = \|\boldsymbol{R} - \Gamma \boldsymbol{\beta}\|_2 + O_2. \tag{26}
$$

Equation (26) shows that the two norm functions are tangent to each other at $\boldsymbol{\beta} = \boldsymbol{\beta}_N^*$. The converse part of the theorem follows easily. Assuming that $\boldsymbol{\beta}_N^*$ is an extremal of the norm $\|\boldsymbol{\Lambda} - \boldsymbol{\lambda}(\boldsymbol{\beta})\|_2$, it follows that $\boldsymbol{\beta}_N^*$ is also an extremal of the norm $\|\boldsymbol{R} - \Gamma\boldsymbol{\beta}\|_2$. However, this norm is a nonnegative quadratic form in the components of $\boldsymbol{\beta}$. It can not have a maximum, so $\boldsymbol{\beta}_N^*$ must be a minimum for it. □

Except for the fact that both Γ and $\boldsymbol{R}$ depend on $\boldsymbol{\beta}_N^*$, it would follow that (25) is a simple least squares problem. As it is, finding a vector $\boldsymbol{\beta}_N^*$ which satisfies the minimizing condition is more difficult. Even so, we will set up an iterative scheme which can solve for such a vector having the required properties.

First, we pick an approximation to $\boldsymbol{\beta}_N^*$—call it $\boldsymbol{\beta}_{N,1}^*$—and solve the eigenvalue problem for $\lambda_i(\boldsymbol{\beta}_{N,1}^*)$, and the corresponding eigenfunctions. These quantities are then approximations to $\lambda_i(\boldsymbol{\beta}_N^*)$ and y_i^*, so we use them in (23) to compute approximations to Γ and to $\boldsymbol{C}$. Call them Γ_1 and $\boldsymbol{C}_1$. We then let $\boldsymbol{\beta}_{N,2}^*$ be the solution of the linear least squares problem of minimizing $\|\boldsymbol{C}_1 - \Gamma_1\boldsymbol{\beta}\|_2$ over all $\boldsymbol{\beta}$. This gives a better approximation to $\boldsymbol{\beta}^*$. Continue this iteration process.

In view of the compactness of $\mathcal{C}(H)$, the method must at least produce a convergent subsequence. It may not, however, produce a truly convergent sequence since there will always be infinitely many solutions $q(x)$ of the equations $\lambda_i(q) = \Lambda_i$ for $i = 1, 2, \cdots, N$, and it is possible that the sequence may oscillate among them. Note, however, that the convergence of the method comes from Theorem 1.2 by letting $N \to \infty$. Although it may seem strange, it is not necessary for the intermediate iterations producing the values of $\boldsymbol{\beta}_{N,k}^*$ to converge.

2.2. On the numerical conditioning of the finite inverse problem. The fundamental difficulty with the inverse eigenvlaue problem is caused by the fact that the eigenvalues are continuous functionals in some very weak topologies. However, we will now show that the inverse problem is, in fact, well conditioned if the validity of the solution is measured in terms of the proper topology. In order to simplify the presentation, we will deal only with boundary conditions for which either $y = 0$ or $y' = 0$ at both ends of the interval, although the methods can be generalized to cover the case of arbitrary self-adjoint boundary conditions.

The proper topology really means the smallest topology in which the eigenvalues are continuous functions. It seems very difficult to characterize this topology; however, a good approximation to it is the one defined by the $2L_1$ norm. As noted above, the only information contained in a finite number of eigenvalues is a $2L_1$ approximation to $q^*(x)$. Therefore, based only on spectral data, we should not even attempt to reconstruct an approximation using the $1Max$ norm, much less a pointwise approximation. Rather, we should only attempt to reconstruct an L_1 approximation to the function $Q(x)$ given by $Q(x) = \int_0^x \int_0^s q(z)\, dz\, ds$.

We may then change our viewpoint a little bit and consider the eigenvalues to depend on the function $Q(x)$ instead of $q(x)$. We therefore write $\lambda_i = \lambda_i(Q)$ rather than $\lambda_i = \lambda_i(q)$ to indicate this change in perspective. The direct eigenvalue problem now consists of calculating the eigenvalues $\boldsymbol{\lambda}$, given the function $Q(x)$, while the inverse problem is that of finding $Q^*(x)$, given spectral data $\boldsymbol{\Lambda} = \boldsymbol{\lambda}(Q^*)$. We will now show that the numerical behavior of this new problem based on $Q(x)$ is quite different from the behavior of the original problem based on $q(x)$. Both the direct and the inverse problem for $Q(x)$ are numerically well conditioned!

In order to investigate the conditioning of the inverse problem, we will suppose

that the functions $Q(x)$ and $q(x)$ are both expressed in the same general form:

$$Q(x) = \sum_{k=1}^{M} \alpha_k B_k(x), \qquad q(x) = \sum_{k=1}^{M} \beta_k B_k(x).$$

Any convenient set of functions $B_k(x)$ could be used for the basis; however, in order to avoid generalized functions (which, of course is *not* absolutely necessary), the functions $Q(x)$ must have piecewise continuous second derivatives. We will now consider the variation in the eigenvalues with small changes in the parameters α_k and β_k. We first integrate by parts twice in (22) to obtain the following relation:

$$\begin{aligned} \lambda_i(Q) &= \lambda_i(q^*) - \int_0^1 q^*(x) y_i^{*2}\, dx + \int_0^1 Q(x)(y_i^{*2})''\, dx + O_2, \\ \lambda_i(Q) &= \lambda_i(q^*) - \int_0^1 q^*(x) y_i^{*2}\, dx + \sum_{k=1}^{N} \alpha_k \int_0^1 B_k(x)(y_i^2)''\, dx + O_2. \end{aligned} \tag{27}$$

We now appeal to a version of (22) which uses the basis $B_k(x)$ and to (27) to take derivatives. We find that

$$\frac{\partial \lambda_i(Q)}{\partial \alpha_k} = \int_0^1 B_k(x)(y_i{}^2)''\, dx, \qquad \frac{\partial \lambda_i(q)}{\partial \beta_k} = \int_0^1 B_k(x)(y_i{}^2)\, dx. \tag{28}$$

These relations say a lot about the solution of the inverse problem. The asymptotic law for eigenvalues and eigenfunctions shows that the derivatives of $\lambda_i(Q)$ are larger than the derivatives of $\lambda_i(q)$ roughly by a multiple of i^2. This means that a small perturbation in $Q(x)$ will produce a much larger change in the eigenvalues than will the same perturbation applied to $q(x)$. Now consider this observation from the perspective of the inverse problem. It implies that the spectral data can be used to reconstruct the function $Q(x)$ much more easily than it can be used to find $q(x)$. The inverse problem for $Q(x)$ is much better conditioned than that for $q(x)$.

Suppose that i is large enough so that $\lambda_i(q) > |q(x)|$. The differential equation shows that the zeros of y_i and y_i'' coincide. Therefore, if the basis function $B_k(x)$ is positive and has its support concentrated in a very small interval, then some of the derivatives (28) will have a better chance of being large because it is less likely that the positive and negative parts of the integral involving $(y_i{}^2)''$ will cancel out. Also the large values of $B_k(x)$ will more readily match large values of the eigenfunctions. This is a good thing since it makes the derivatives larger and improves the conditioning of the inverse problem both for $\boldsymbol{\lambda}(q)$ and for $\boldsymbol{\lambda}(Q)$. This provides an intuitive foundation for the rule of thumb $N \le M \le 2N$ given in §2.1. This analysis also shows that the conditioning of a given inverse problem may depend on the location of the support of the basis functions and provides some guidelines for selection of the specific form of $B_k(x)$.

On the other hand, the direct problem of finding $\boldsymbol{\lambda}$ given $Q(x)$ would seem to require the computation of two numerical derivatives of $Q(x)$ in order to find $q(x)$ and use the differential equation. Perhaps the direct problem of using $Q(x)$ to calculate $\boldsymbol{\lambda}$ will now be ill-conditioned. However, Theorem 1.1 shows that this is not the case, in that *any* function which approximates $Q(x)$ well enough in the L_1 norm will give about the same eigenvalues. It simply does not matter if $Q''(x)$ is a good pointwise approximation to $q(x)$ or not. So we now have the best of both worlds in that the inverse problem and the direct problem connecting $Q(x)$ and $\boldsymbol{\lambda}(Q)$ are well

conditioned. Of course, the real difficulty with this point of view is that there are no well-known numerical algorithms for calculating $\lambda_n(q)$ when given only $Q(x)$.

Another way to think about the inverse problem for $\boldsymbol{\lambda}(q)$ is to break it into two separate steps. The first step is to use the spectral data to find an L_1 approximation to $Q(x)$. The second step is to use some kind of numerical differentiation procedure to compute $q(x) = Q''(x)$. It is only the second step which is ill-conditioned. The inverse eigenvalue problem is only ill-conditioned if we attempt to extract more information from the spectral data than it contains.

If it is known that the function $q(x)$ is differentiable, then Theorem 1.2 can be replaced with a stronger statement which involves three integrations by parts. The function $Q(x)$ can then be taken to be $Q(x) = \int_0^x \int_0^t \int_0^s q(z)\,dz\,ds\,dt$ and the above argument can be carried through using three derivatives rather than two.

2.3. Some numerical examples. To construct an example, we will use (1) with

$$q^*(x) = \begin{cases} 1 & \text{if } 0 \le x < \frac{1}{2}, \\ -1 & \text{if } \frac{1}{2} \le x \le 1, \end{cases} \qquad \text{and} \quad \boldsymbol{\Lambda} = \boldsymbol{\lambda}(q^*). \tag{29}$$

We choose as spectral data the first five eigenvalues of (1) using the boundary conditions $y(0) = y(1) = 0$ together with the first five eigenvalues of (1) using the boundary conditions $y(0) = y'(1) = 0$ for a total of 10 data values. We choose the representation (20) using the simple basis functions given by

$$h_k(x) = \begin{cases} 1 & \text{if } x \le \dfrac{k}{M}, \\ 0 & \text{otherwise,} \end{cases} \qquad k = 1, 2, \cdots, M.$$

Given $N = 10$, we choose $M = 15$ and we now take, as the first approximation to $q^*(x)$, the function $q_{15,1} \equiv .25$. Using (23) we compute $\boldsymbol{R}_1$ and Γ_1 and let $q_{15,2}(x)$ be the result of solving the least squares problem $\|\boldsymbol{R}_1 - \Gamma_1\boldsymbol{\beta}\| = \text{minimum}$.

A little computation shows that

$$\|q^*(x) - q_{15,2}(x)\|_{1Max} \approx .008\cdots, \qquad \int_0^1 |q^*(x) - q_{15,2}(x)|\,dx \approx 0.144\cdots,$$

$$\|q^*(x) - q_{15,2}(x)\|_{2L_1} \approx .002\cdots, \qquad \|\boldsymbol{\lambda}(q_{15,2}) - \boldsymbol{\Lambda}\|_2 \approx .041\cdots.$$

Thus, we have very good $1Max$ and $2L_1$ approximations but a very poor L_1 approximation. Graphs of the various functions are given in Fig. 1 where the following labels are used.

- The curve ① represents the function $q^*(x)$.
- The curve ④ represents the function $q_{15,2}(x)$.
- The two triangular shape curves ② and ⑤ represent, respectively, the functions $\int_0^x q^*(z)\,dz$ and $\int_0^x q^*_{15,2}(z)\,dz$.
- Graphically the three functions $\int_0^x\int_0^s q^*(z)\,dz\,ds\,dx$, $\int_0^x\int_0^s q{\downarrow_{15,2}}(z)\,dz\,ds\,dx$, and $\int_0^x\int_0^s q_{15,2}(z)\,dz\,ds\,dx$ are indistinguishable and the S-shape curve ③ represents all of them. Only a slight thickening of the curve can be detected.
- The curve ⑥ represents the function $q{\downarrow_{15,2}}(x)$ which will be defined in §2.4 below.

We may now use the function $q_{15,2}$ just obtained as a new approximation to $q^*(x)$ and repeat the minimization process. However, the new minimizing function $q_{15,3}$ obtained this way is, essentially, the same as $q_{15,2}$, at least as far as the $1Max$ and

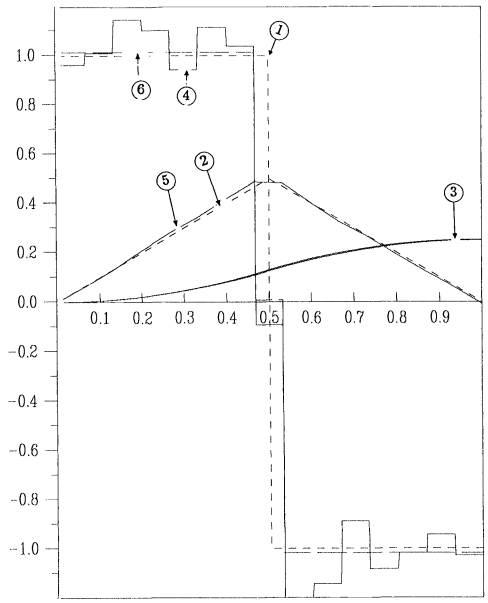


FIG. 1. *Graphs of the coefficient functions and some first and second integrals.*

$2L_1$ norms are concerned. Only some random roundoff error distinguishes them and the method converges in only one iteration step. This is typical of the method in that, depending on the choice of the first guess, it usually converges with only two or three steps. This is due to the fact that, for any choice of coefficient function $q(x)$, the eigenfunctions corresponding to the same set of boundary conditions look very much alike so that $\boldsymbol{R}$ and Γ are quite insensitive to changes in $q(x)$.

2.4. Constrained solutions using Hanson–Haskel's program LSEI. In view of the remarks in §§1.2 and 1.3, it seems impossible to extract pointwise information about $q(x)$ from a finite amount of spectral data. However, if it is possible to take advantage of some additional information about the unknown function, then we may very well be able to obtain good pointwise information. Consider, for example, the inverse problem associated with the free vibrations of the earth [6]. It seems quite reasonable to assume that the density of the earth is an increasing function of depth below the surface.

Such constraints can be used to eliminate the possibility of large oscillations like those given above by the example $\Delta q(x) = A \cos Bx$. In fact, there are many conditions which can be used to eliminate the oscillations. If, for example, the functions are also required to be monotone or convex or concave or even just unimodal, such oscillations will not be possible. Of course, the idea is to identify such an a priori constraint which will prevent the oscillations and, at the same time, still allow us to interpolate to the spectral data.

The CMLIB computing package contains a very useful FORTRAN program, written by R. J. Hanson and K. H. Haskel, called LSEI, which can solve constrained least

squares problems for matrices. Specifically, it solves the following problem:

$$\text{Minimize} \quad \|\boldsymbol{C} - \Gamma\boldsymbol{\beta}\|_2 \quad \text{subject to constraints}$$
$$E\boldsymbol{\beta} = \boldsymbol{F} \quad \text{and} \quad G\boldsymbol{\beta} \geq \boldsymbol{K}. \tag{30}$$

Other kinds of constraints on the coefficient functions, such as symmetry $q(x) = q(1-x)$, are sometimes used when dealing with inverse problems. Or the function could be given only for $0 \leq x \leq \frac{1}{2}$, and we might be required to find its unknown values for $\frac{1}{2} \leq x \leq 1$. Fortunately, this iterative least squares method can be easily adapted to take advantage of any such additional information.

We will now give an example which uses LSEI to solve a constrained inverse problem. We will use the same function $q^*(x)$ given in (29); however, in addition to the 10 spectral data values, we will now assume that $q^*(x)$ is also known to be a decreasing function of x. This condition is reflected in the constraint $\beta_k \geq 0$. Incidentally, this shows one good reason for using a local basis like $h_k(x)$. It would be very difficult to take advantage of the decreasing condition on $q(x)$ using a global basis such as a Fourier series. We now use LSEI to solve the constrained least squares problem with the same iterative procedure as before. Since the constraint eliminates the oscillations, we now get better results.

Using the previous solution $q_{15,2}(x)$ for the first guess, we solve the least squares problem subject to the decreasing condition to obtain a new function which we call $q{\downarrow}_{15,2}\ (x)$. Its graph, given in Fig. 1, clearly shows that it is a much better approximation. It even gives a fairly good pointwise approximation to $q^*(x)$ in that $|q^*(x) - q{\downarrow}_{15,2}\ (x)| < .019$ except for the short interval about the midpoint. Further computations show that

$$\|q^* - q{\downarrow}_{15,2}\|_{1Max} \approx .0034, \qquad \int_0^1 |q^*(x) - q{\downarrow}_{15,2}(x)|\,dx \approx .083,$$
$$\|q^* - q{\downarrow}_{15,2}\|_{2L_1} \approx .00092, \qquad \|\boldsymbol{\lambda}(q{\downarrow}_{15,2}) - \boldsymbol{\Lambda}\|_2 \approx .039.$$

Computing additional terms $q{\downarrow}_{15,n}$ in the sequence did not produce much better approximations, at least as far as the $2L_1$ norm is concerned.

2.5. Jump discontinuities in the coefficient functions. In the previous example, we obtained a solution which was very good at all points except near the jump discontinuity $x = \frac{1}{2}$. We will now develop a method of allowing the break points of the functions $h_k(x)$ to move around so as to better approximate a discontinuous function $q^*(x)$. To do this, we introduce the function $h_t(x)$ defined for any number t by

$$h_t(x) = \begin{cases} 1 & \text{if } x \leq t, \\ 0 & \text{if } t < x. \end{cases} \tag{31}$$

We then select points $t_k,\ t_k^* \in [0,1]$ and suppose that $q(x)$ and $q^*(x)$ are any two functions of the form

$$q^*(x) = \sum_{k=1}^{M} \beta_k^* h_{t_k^*}(x) \quad \text{and} \quad q(x) = \sum_{k=1}^{M} \beta_k h_{t_k}(x). \tag{32}$$

We suppose that both t_k and β_k are variables and introduce the new vector $\boldsymbol{\delta} = (\boldsymbol{\beta}, \boldsymbol{t})$. We then substitute the representation (32) into (22) and compute the principal linear

part of $\lambda_i(q)$ as a function of $\boldsymbol{\delta}$. Finally, we will apply the iterative process used previously to the new vector $\boldsymbol{\delta}$.

First, we derive the linear approximations to $\lambda_n(q)$. Putting (32) into (22) shows that

$$\lambda_i(q) = \lambda_i(q^*) - \int_0^1 q^*(x) y_i^{*2}\,dx + \sum_{k=1}^{M} \beta_k \int_0^1 h_{t_k}(x) y_i^{*2}\,dx + O_2. \tag{33}$$

We now use the relation

$$\beta_k \int_0^1 h_{t_k}(x) y_i^2\,dx = \beta_k^* \int_0^{t_k^*} y_i^{*2}\,dx + \Delta\beta_k \int_0^{t_k^*} y_i^{*2}\,dx + \Delta t_k \beta_k^* y_i^2(t_k^*) + O_2 \tag{34}$$

in (33) to obtain

$$\begin{aligned}\lambda_i(q) =& \lambda_i(q^*) - \int_0^1 q^*(x) y_i^{*2}\,dx - \sum_{k=1}^{M} t_k^* \beta_k^* y_i^{*2}(t_k^*) \\ &+ \sum_{k=1}^{M} \beta_k \int_0^1 h_{t_k^*}(x) y_i^{*2}\,dx + \sum_{k=1}^{M} t_k \beta_k^* y_i^{*2}(t_k^*) + O_2.\end{aligned}$$

We recast this equation in matrix form as $\boldsymbol{\lambda}(q) = -\boldsymbol{S} + \Gamma\boldsymbol{\delta} + O_2$ where

$$\begin{aligned}\boldsymbol{S}_i &= -\lambda_i(q^*) + \int_0^1 q^*(x) y_i^{*2}\,dx + \sum_{k=1}^{M} t_k^* \beta_k^* y_i^{*2}(t_k^*), \\ \Gamma_2 &= \left(\int_0^1 h_{t_k^*}(x) y_i^{*2}\,dx \right), \quad \Gamma_3 = \left(\beta_k^* y_i^{*2}(t_k^*) \right), \quad \Gamma = (\Gamma_2\ \Gamma_3).\end{aligned} \tag{35}$$

We may now take $\boldsymbol{R} = \boldsymbol{\Lambda} - \boldsymbol{S}$ and iterate in the minimum problem as before to solve for $\boldsymbol{\delta}$. However, a few important details need to be considered first.

In the process of allowing the values of t_k to move about, we may also use LSEI to impose constraints. One obvious constraint to use is that $t_1 \le t_2 \le \cdots \le t_M$, which serves to keep the indexing of the t's straight. However, there are other problems with using such a basis. If, for example, some of the t_k's are the same, then the "basis" will not be linearly independent. Even if the t_k's are allowed to come very close to each other, then the basis will become numerically ill-conditioned. To avoid this kind of problem, we will use LSEI to impose the restriction[2] $|t_k - \frac{k}{M}| \le \frac{1}{3M}$, which implies that $|t_k - t_j| \ge \frac{1}{3M}$.

However, problems may still arise with variable values for t. To see why this is so, consider the variation in the eigenvalues with changes in β and t. Taking derivatives as we did for (27) we find that at $\beta_k = \beta_k^*$ and $t_k = t_k^*$,

$$\frac{\partial\lambda_i(q)}{\partial\beta_k} = \int_0^1 h_{t_k^*}(x) y_i^2\,dx = \int_0^{t_k^*} y_i^2\,dx \quad \text{and} \quad \frac{\partial\lambda_i(q)}{\partial t_k} = \beta_k^* y_i^2(t_k^*). \tag{36}$$

These two relations explain a great deal about the behavior of the inverse problem in general. First notice that a small change in β_k^* will produce a change in λ_i^* which depends on a global quantity involving an integral over an entire interval of

[2] Actually, the example of §2.3 was implemented using the functions (31) and LSEI with the restriction $t_k = k/M$.

values. Such quantities are numerically well behaved. Now notice that a variation in t_k^* produces a variation in λ_i^* which depends on the local pointwise value of the eigenfunction $y_i^2(t_k^*)$. So small changes in t_k^* may or may not produce small changes in λ_i^*. It depends on whether $\beta_k^* y_i^2(t_k^*)$ is large or small, and this is unpredictable because of the rapid oscillations of the eigenfunctions.

If we attempt to solve an inverse eigenvalue problem and, by random bad luck, it happens that the discontinuity of the unknown coefficient $q^*(x)$ lies on or near some nodal point for each of the eigenfunctions corresponding to all of the known spectral data, then changes in t_k^* will produce very small changes in the eigenvalues. From the perspective of the inverse problem, we see that it would be impossible to locate such a discontinuity with any degree of accuracy using only spectral data. On the other hand, it is fortunate that the nodal points of any eigenfunction are almost uniformly spread out through the interval. Even with only a finite amount of data, we expect that at least some of the eigenfunctions will have their maximum values close to the point of discontinuity. Therefore, *even using only a finite amount of spectral data, we should still be able to accurately locate discontinuities in the coefficient function.* We only need to know a somewhat uniform sample of spectral data and to have a jump discontinuity for which β_k^*, the size of the jump, is not especially small.

We now give a numerical example of this process. We will use the same function $q^*(x)$ and the same spectral data used before. The solution $q{\downarrow}_{15,2}(x)$, obtained using the decreasing restriction, strongly suggests that $q^*(x)$ has a point of discontinuity in the interval $[t_7, t_8]$ where $t_7 = .466666$ and $t_8 = .533333$. We choose to let t_8 become a variable, but we will impose on $\boldsymbol{t}$ and $\boldsymbol{\beta}$ the restrictions

$$t_k = \frac{k}{15} \quad \text{for } k \neq 8 \quad \text{but} \quad .47 \leq t_8 \leq .54 \quad \text{and} \quad \beta_k \geq 0. \tag{37}$$

These constraints will insure that the basis is numerically well conditioned, but at the same time they will allow the discontinuity to be located.

Using the solution $q{\downarrow}_{15,2}(x)$ obtained in §2.4 as the first guess in the least squares iteration procedure, subject to (37), gives a sequence of functions—call them $\widehat{q}_n$—for which the points of discontinuity, called T_n, converge to .500. Some other important numbers are listed in Table 1.

TABLE 1

Concerning the sequence of iterations $\widehat{q}_n$.

n	T_n	$\|q^* - \widehat{q}_n\|_{1Max}$	$\|q^* - \widehat{q}_n\|_{2L_1}$	$\|\boldsymbol{\lambda}(\widehat{q}_n) - \boldsymbol{\Lambda}\|_2$
1	.53333333	$2.7457058{\cdot}10^{-02}$	$1.0778840{\cdot}10^{-03}$	$2.5811850{\cdot}10^{-03}$
2	0.4878601	$3.3297241{\cdot}10^{-02}$	$3.0311858{\cdot}10^{-03}$	$3.9687917{\cdot}10^{-02}$
3	0.5150741	$2.3708206{\cdot}10^{-03}$	$7.3498831{\cdot}10^{-04}$	$8.4827110{\cdot}10^{-02}$
4	0.5001864	$8.6408332{\cdot}10^{-03}$	$3.3640519{\cdot}10^{-03}$	$2.1405114{\cdot}10^{-02}$
5	0.5000984	$2.1051883{\cdot}10^{-03}$	$7.7785080{\cdot}10^{-05}$	$2.8585272{\cdot}10^{-02}$
6	0.5004199	$8.2593737{\cdot}10^{-04}$	$1.0909934{\cdot}10^{-04}$	$1.1293867{\cdot}10^{-03}$

At this point we must, subjectively, decide whether the unknown coefficient does or does not have a discontinuity. If it does, then we accept the solution given in Table 1. If we decide that there is no discontinuity, then we should construct a continuous approximation to the function $q{\downarrow}_{15,2}$ as the solution. Unfortunately, the size of $\|\boldsymbol{\lambda}(q) - \boldsymbol{\Lambda}\|_2$ cannot be used to make this decision because the norm can be

made small using many different functions q. Perhaps the strongest evidence for the existence of a discontinuity is the clear convergence of the sequence T_n indicated by the values in the table.

In this way, we have found a very good pointwise approximation to $q^*(x)$ using the spectral data and the additional decreasing information about $q^*(x)$. There is nothing special about the fact that there was only one discontinuity in this example. Several discontinuities could have been dealt with the same way.

These methods have also been used to solve some inverse problems for the equations $y'' + \lambda q(x)y = 0$ and $(r(x)y')' + \lambda y = 0$, and the numerical behavior was in most respects similar to that reported for (1). Theorem 2.1 must be used to obtain new approximation formulae, replacing (22). Otherwise, the procedure is identical. More generally, the method could be easily modified to study the equation $(r(x)y')' + (\lambda p(x) - q(x))y = 0$, together with arbitrary self-adjoint boundary conditions.

3. Summary and conclusions. We have provided an extensive mathematical analysis of the finite inverse problem, proving a convergence theorem for some approximations to its numerical solution. The most important tool used was the compactness of the set $\mathcal{C}(H)$ in the $1Max$ topology. Another major part of the analysis was introducing the idea of the weak topology on $\mathcal{C}(H)$ and approximating it using the $2L_1$ norm. The weak topology determines how much information about the coefficient function is contained in the spectral data. Finally, the application of Theorem 2.1 of [1] to develop (25), (28), and (36) yielded a very good intuitive understanding of the finite inverse eigenvalue problem.

Perhaps one of the most striking results of this analysis was to show the great disparity between the very weak approximations which must be used when given only a finite amount of data (a $2L_1$ approximation or worse) and the much stronger approximation (using the norms $\|\cdot\|_\infty$ or $\|\cdot\|_2$) which are possible when using an infinite amount of data [4], [8]. However, even if the approximation is quite weak in the finite case, it is still very a usable result. It is especially useful when combined with additional information about the coefficient function, such as monotonicity or convexity, which may frequently be available simply from the physics of the problem. In such cases, it may still be possible to generate good pointwise approximations using only finite data. This is an especially important development for equations of the form (10) and (12) which may have discontinuous coefficient functions since asymptotic formula are difficult to obtain in such cases. It would be interesting, and probably not especially difficult, to develop a set of theorems to the effect that "If $q_n(x)$ converges in the $2L_1$ norm and if $q_n(x)$ is a decreasing function of x or a convex function of x or $\cdots$ then $q_n(x)$ converges in a much stronger topology."

REFERENCES

[1] D. C. Barnes, *Some approximation formula for stochastic eigenvalues*, SIAM J. Math. Anal., 18 (1987), pp. 933–940.
[2] H. F. Cullen, *Introduction to General Topology*, D. C. Heath, Boston, 1968.
[3] N. Dunford and J. T. Schwartz, *Linear Operators*, Interscience, New York, 1958.
[4] O. Hald, *The inverse Sturm–Liouville problem with symmetric potentials*, Acta Math., 141 (1978), pp. 263–291.
[5] M. G. Krein, *On certain problems on the maximum and minimum of characteristic values and on Lyapunov zones of stability*, Trans. Amer. Math. Soc., 2 (1955), pp. 163–187.

[6] E. R. LAPWOOD AND T. USAMI, *Free Oscillations of the Earth*, Cambridge University Press, Cambridge, 1981.
[7] J. R. MCLAUGHLIN, *Analytical methods for recovering coefficients in differential equations from spectral data*, SIAM Rev., 28 (1986), pp. 53–72.
[8] ———, *Stability theorems for two inverse spectral problems*, Inverse Problems, 4 (1988), pp. 529–540.
[9] J. R. MCLAUGHLIN AND W. RUNDELL, *A uniqueness theorem for an inverse Sturm–Liouville problem*, J. Math. Phys., 28 (1987), pp. 1471–1472.
[10] P. SACKS, *An iterative method for the inverse Dirichlet problem*, Inverse Problems, 4 (1988), pp. 1055–1069.
[11] T. SEIDMAN, *A convergent approximation scheme for the inverse Sturm–Liouville problem*, Inverse Problems, 1 (1985), pp. 251–262.
[12] W. STENGER, *On the variational principles for eigenvalues for a class of unbounded operators*, J. Math. Mech., 17 (1968), pp. 641–648.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 754–768, May 1991

010

HALF-BOUND STATES AND LEVINSON'S THEOREM FOR DISCRETE SYSTEMS*

D. B. HINTON†, M. KLAUS‡, AND J. K. SHAW‡

Abstract. A second-order difference equation $-y_{n+1}+2y_n-y_{n-1}+q_ny_n=\lambda y_n$, $n=1,2,\cdots$ is considered. The perturbation terms q_n are assumed to satisfy the scattering condition $\sum_{n=1}^{\infty} n|q_n|<\infty$. A boundary condition $y_0+\alpha y_1=0$ is imposed, and a formula is derived for the number of eigenvalues of the associated self-adjoint operator. This formula, known as Levinson's theorem, is in terms of the change of phase of the complex amplitude function for solutions of the difference equation and of other factors whose value depends on the existence of so-called *half-bound states*, which are defined herein. An application is given to the equations of motion of a semi-infinite chain of masses connected by springs. It is established how the large-t asymptotics depend on the existence of half-bound states.

Key words. difference equation, Titchmarsh–Weyl theory, spectrum, half-bound state

AMS(MOS) subject classifications. 47E05, 34B25

1. Introduction. The spectral analysis of difference equations arises in many contexts. One such context is the dynamics of differential-difference equations. A simple example of this is an infinite chain of masses, each of which is connected by springs to its nearest neighbors. The same system of differential equations models certain large electrical systems and disordered crystals. In the case of a semi-infinite chain with identical springs, the difference equation

$$-y_{n+1}+2y_n-y_{n-1}+q_ny_n=\lambda y_n, \qquad n=1,2,\cdots, \tag{1.1}$$

arises in the computation of the vibrational frequencies or energy spectrum. Associated with (1.1) is a boundary condition

$$y_0+\alpha y_1=0 \tag{1.2}$$

where α is a real number. The coefficients q_n are real.

First we associate with (1.1)-(1.2) a self-adjoint operator T_α in the Hilbert space l_2 of square summable sequences $\{y_n\}_1^\infty$. Define T_α in l_2 by

$$(T_\alpha y)_n=-y_{n+1}+2y_n-y_{n-1}+q_ny_n, \qquad n=1,2,\cdots$$

with $y_0=-\alpha y_1$. Note that T_0 is associated with the boundary condition $y_0=0$. Furthermore, the domain of T_α is the set of all $y\in l_2$ such that $T_\alpha y\in l_2$. The operator T_α is known to be self-adjoint, e.g., [1], [2].

We consider T_α under the scattering condition

$$\sum_{n=1}^{\infty} n|q_n|<0, \tag{1.3}$$

which implies that T_α is a bounded operator on l_2. Under (1.3), it turns out that the essential spectrum of T_α is $[0,4]$ and that T_α may have finitely many eigenvalues in $(-\infty,0)\cup(4,\infty)$. In § 2 we derive a formula which counts the number of eigenvalues or bound states. This formula, of the type known as Levinson's theorem, is in terms of the change of phase of the complex amplitude function associated with a solution

* Received by the editors January 22, 1990; accepted for publication May 1, 1990.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996.

‡ Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

of (1.1)-(1.2). Additional terms of $-\pi$ must be added when *half-bound states* are present; these are defined below (2.14). A good deal of work has recently been done on Levinson-type theorems in different contexts. The first such result was proved by Levinson for the Schrödinger operator [9]. Results have been obtained more recently for the Schrödinger equation [8], the Dirac equation [5], [12], and four-dimensional Hamiltonian systems [6], [11]. In [5], [6], and [8] the condition placed on the potentials is the scattering condition. For difference equations of the type considered here, a Levinson theorem has been formulated in [4]. But a proof under condition (1.3) was not given and the half-bound state case was not considered.

In § 3 the theory is applied to the equations of motion of a semi-infinite chain of masses connected by springs. It is found that the long-time behavior of a displacement is generically either of order $t^{-3/2}$ or $t^{-1/2}$ according to whether or not a half-bound state exists. (See the definition below (2.14).) Thus the influence of a half-bound state is exhibited in the rate of return of each mass to its equilibrium position.

2. Asymptotics of (1.1) and Levinson's theorem. For the sake of clarity we treat only the case where $\alpha = 0$ in (1.2). A basic role will be played by the two solutions $\theta(\lambda) = \{\theta_n(\lambda)\}$, $\phi(\lambda) = \{\phi_n(\lambda)\}$ of (1.1) defined by the initial conditions

$$\theta_0(\lambda) = \theta_1(\lambda) = 1, \quad \phi_0(\lambda) = 0, \quad \phi_1(\lambda) = 1. \tag{2.1}$$

For the unperturbed case, i.e., $q_n = 0$ for all n, we designate θ_n, ϕ_n by θ_n^0, ϕ_n^0, respectively. These sequences satisfy

$$y_{n+1} + (\lambda - 2)y_n + y_{n-1} = 0, \tag{2.2}$$

and we now calculate them. Setting $y_n = \tau^n$ in (2.2) gives that $\tau^2 + (\lambda - 2)\tau + 1 = 0$ so that

$$2 - \lambda = \tau + \frac{1}{\tau}, \qquad \tau = \frac{2 - \lambda \pm \sqrt{\lambda^2 - 4\lambda}}{2}. \tag{2.3}$$

We choose, for $\mathcal{I}\lambda \geqq 0$, a branch of square root in the following manner. Let

$$D_\gamma = \{\gamma \in \mathbb{C}: 0 \leqq \mathcal{R}\gamma \leqq \pi,\ 0 \leqq \mathcal{I}\gamma < \infty\},$$
$$D_\tau = \{\tau \in \mathbb{C}: |\tau| \leqq 1,\ \mathcal{I}\tau \geqq 0,\ \tau \neq 0\},$$
$$D_\lambda = \{\lambda \in \mathbb{C}: \mathcal{I}\lambda \geqq 0\}.$$

Then it is easily seen that the maps $\gamma \to \tau = e^{i\gamma} \to \lambda = 2 - \tau - 1/\tau$ are one to one and onto from $D_\gamma \to D_\tau \to D_\lambda$, respectively; furthermore, interiors are mapped onto interiors of these regions. Hence we see that for $\mathcal{I}\lambda > 0$ we may choose τ satisfying (2.3) so that $|\tau| < 1$ and $\mathcal{I}\tau > 0$. The map $\lambda = 2 - \tau - 1/\tau$ maps $(-1, 0)$ onto $(4, \infty)$ and $(0, 1)$ onto $(-\infty, 0)$, so we may continue the map analytically (by conjugation) from $\{|\tau| < 1: \tau \neq 0\}$ onto $\mathbb{C}\backslash[0, 4]$. This is the way we choose the solution τ of $\tau^2 + (\lambda - 2)\tau + 1 = 0$ for $\lambda \in \mathbb{C}\backslash[0, 4]$. The corresponding γ satisfies $-\pi < \mathcal{R}\gamma < \pi$, $\mathcal{I}\gamma > 0$.

Then for fixed $\lambda \in \mathbb{C}\backslash[0, 4]$, we have two solutions of the free problem, namely, $\tau^n = e^{in\gamma}$ and $\tau^{-n} = e^{-in\gamma}$. Note that $\tau^n \to 0$ as $n \to \infty$ since $\mathcal{I}\gamma > 0$. To compute $\{\theta_n^0(\lambda)\}$ and $\{\phi_n^0(\lambda)\}$ we express them as linear combinations of $\{e^{in\gamma}\}$ and $\{e^{-in\gamma}\}$ and substitute into (2.1). This yields, after some computation, that for $n = 0, 1, \cdots$,

$$\theta_n^0(\lambda) = \frac{\cos(n - \frac{1}{2})\gamma}{\cos \gamma/2} \quad \text{and} \quad \phi_n^0(\lambda) = \frac{\sin n\gamma}{\sin \gamma}.$$

To compute $\{\theta_n(\lambda)\}$ and $\{\phi_n(\lambda)\}$ we apply variation of constants to (1.1) written in the form

$$-y_{n+1} + (2 - \lambda)y_n - y_{n-1} = f_n, \qquad f_n = -q_n y_n. \tag{2.4}$$

This gives, for the solution of (2.4), that

$$y_n = c_1 \frac{\sin n\gamma}{\sin \gamma} + c_2 \frac{\cos (n-\frac{1}{2})\gamma}{\cos (\gamma/2)} + \sum_{k=1}^{n-1} \frac{\sin (n-k)\gamma}{\sin \gamma} q_k y_k$$

where a summation is zero when the upper limit is less than the lower limit. Using (2.1) now to compute c_1 and c_2, we obtain for $n = 1, 2, \cdots$

$$\phi_n(\lambda) = \frac{\sin n\gamma}{\sin \gamma} + \sum_{k=1}^{n-1} \frac{\sin (n-k)\gamma}{\sin \gamma} q_k \phi_k(\lambda), \tag{2.5}$$

$$\theta_n(\lambda) = \frac{\cos (n-\frac{1}{2})\gamma}{\cos \gamma/2} + \sum_{k=1}^{n-1} \frac{\sin (n-k)\gamma}{\sin \gamma} q_k \theta_k(\lambda). \tag{2.6}$$

Now we introduce the Jost functions and the asymptotics of solutions. Multiplying (2.5) by $e^{in\gamma}$, we get

$$e^{in\gamma}\phi_n(\lambda) = \frac{e^{2in\gamma} - 1}{2i \sin \gamma} + \sum_{k=1}^{n-1} \frac{e^{2i(n-k)\gamma} - 1}{2i \sin \gamma} q_k e^{ik\gamma} \phi_k(\lambda). \tag{2.7}$$

Since $\mathcal{I}\gamma > 0$ for $\lambda \in \mathbb{C}\backslash[0, 4]$, equation (2.7) implies that if $v_k = |e^{ik\gamma}\phi_k(\lambda)|$, then for $n = 1, 2, \cdots$,

$$v_n \leqq K\left(1 + \sum_{k=1}^{n-1} |q_k| v_k\right), \qquad K = \frac{1}{|\sin \gamma|}. \tag{2.8}$$

Application of the Gronwall inequality for difference equations to (2.8) yields that

$$\begin{aligned} v_n = |e^{in\gamma}\phi_n(\lambda)| &\leqq K \exp\left\{K \sum_{k=1}^{n-1} |q_k|\right\} \\ &\leqq K \exp\left\{K \sum_{k=1}^{\infty} |q_k|\right\}. \end{aligned} \tag{2.9}$$

Using the bound (2.9) in (2.7) it follows that

$$e^{in\gamma}\phi_n(\lambda) = -(2i \sin \gamma)^{-1}\left[1 + \sum_{k=1}^{\infty} e^{ik\gamma} q_k \phi_k(\lambda)\right] + o(1) \tag{2.10}$$

as $n \to \infty$. Similar calculations with $\theta_n(\lambda)$ in (2.6) yield that as $n \to \infty$,

$$e^{in\gamma}\theta_n(\lambda) = (2i \sin \gamma)^{-1}\left[(ie^{i\gamma/2} \sin \gamma)/(\cos \gamma/2) - \sum_{k=1}^{\infty} e^{ik\gamma} q_k \theta_k(\lambda)\right] + o(1). \tag{2.11}$$

Thus for $\lambda \in \mathbb{C}\backslash[0, 4]$, we define the Jost functions by

$$\begin{aligned} f_\phi(\lambda) &= 1 + \sum_{k=1}^{\infty} e^{ik\gamma} q_k \phi_k(\lambda), \\ f_\theta(\lambda) &= -(i e^{i\gamma/2} \sin \gamma)/(\cos \gamma/2) + \sum_{k=1}^{\infty} e^{ik\gamma} q_k \theta_k(\lambda). \end{aligned} \tag{2.12}$$

As in the differential equations case, the Jost function is the coefficient of the dominant asymptotic term. Its zeros therefore constitute the λ values corresponding to decaying solutions. Closely related to the Jost function is the Titchmarsh-Weyl function $m(\lambda)$ defined by

$$m(\lambda) = -\lim_{n \to \infty} \frac{\theta_n(\lambda)}{\phi_n(\lambda)}, \qquad \mathcal{I}\lambda \neq 0. \tag{2.13}$$

A general proof for the existence of the limit in (2.13) may be found in the text of Atkinson [1] and the papers of Asahi [2] and Bennewitz [3]; in our case the asymptotics of (2.10) and (2.11) give that

$$m(\lambda) = -f_\theta(\lambda)/f_\phi(\lambda), \tag{2.14}$$

provided $f_\phi(\lambda) \neq 0$ for $\mathcal{I}\lambda \neq 0$. That $f_\phi(\lambda) \neq 0$ for $\mathcal{I}\lambda \neq 0$ follows from the asymptotics of solutions of (1.1). Under the condition (1.3), the asymptotic methods of differential equations applied to difference equations yields solutions $\{z_n(\lambda)\}$, $\{w_n(\lambda)\}$ of (1.1) such that for $\lambda \in \mathbb{C}\backslash[0,4]$, as $n \to \infty$,

$$z_n(\lambda) = e^{in\gamma}[1+o(1)], \qquad w_n(\lambda) = e^{-in\gamma}[1+o(1)].$$

Since $\{z_n(\lambda)\}$, $\{w_n(\lambda)\}$ is a basis, it follows that $f_\phi(\lambda) = 0$ if and only if $\{z_n(\lambda)\}$ and $\{\phi_n(\lambda)\}$ are linearly dependent. This is equivalent to $\{\phi_n(\lambda)\} \in l_2$ since $\{z_n(\lambda)\} \in l_2$ and $\{w_n(\lambda)\} \notin l_2$. Thus $f_\phi(\lambda) = 0$ for $\lambda \in \mathbb{C}\backslash[0,4]$ if and only if λ is an eigenvalue for the self-adjoint operator T_α with $\alpha = 0$. Since T_α has no complex eigenvalues, we have another proof that $f_\phi(\lambda) \neq 0$ for $\mathcal{I}\lambda \neq 0$.

For $\lambda \in (0,4)$, $\theta_n(\lambda)$ and $\phi_n(\lambda)$ are asymptotically oscillatory, and we can show that this corresponds to the continuous spectrum of T_α (but we will not need this fact). Thus a point $\lambda \in \mathbb{C}\backslash[0,4]$ is in the resolvent set unless it is an eigenvalue, in which case $f_\phi(\lambda) = 0$ (for the $\alpha = 0$ case). From (2.9) and (2.12) we see that $f_\phi(\lambda) \to 1$ as $|\lambda| \to \infty$ so there are no eigenvalues sufficiently far out. We will presently show that the behaviour of $f_\phi(\lambda)$ as $\lambda \to 0$ rules out a clustering of eigenvalues at $\lambda = 0$, and similarly at $\lambda = 4$.

In particular, there will be only finitely many eigenvalues. This leaves only the points $\lambda = 0$ and $\lambda = 4$. We say that $\lambda = 0$ is a *half-bound state* (HBS) provided that $\{\phi_n(0)\}$ is a bounded sequence, and a non-half-bound state (non-HBS) otherwise. A similar definition may be given at $\lambda = 4$ in terms of $\{\phi_n(4)\}$. Sometimes it will be more convenient to say that $\phi(0)$ or $\phi(4)$ is a half-bound state. If instead $\{\theta_n(0)\}$ is bounded, we say that we have a θ-half-bound state (θ-HBS) at $\lambda = 0$, and similarly at $\lambda = 4$. It is the analysis of the Jost functions f_θ and f_ϕ at HBS that comprises the rest of this paper. We begin with f_ϕ near $\lambda = 0$.

THEOREM 2.1. *If* $\lambda = 0$ *is an* HBS, *then*

$$f_\phi(\lambda) = i\gamma a_\phi^{(1)} + o(|\gamma|) \quad \textit{where } a_\phi^{(1)} = \sum_{k=1}^{\infty} kq_k\phi_k(0) \neq 0. \tag{2.15}$$

If $\lambda = 0$ *is a non-*HBS, *then*

$$f_\phi(\lambda) = a_\phi^{(0)} + o(1) \quad \textit{where } a_\phi^{(0)} = 1 + \sum_{k=1}^{\infty} q_k\phi_k(0) \neq 0. \tag{2.16}$$

Proof. For $\lambda = 0$ we have $\gamma = 0$ and the unperturbed problem $-y_{n+1} + 2y_n - y_{n-1} = 0$ has solutions

$$\phi_n^0(0) = n, \quad \theta_n^0(0) = 1, \quad n = 0, 1, \cdots.$$

Equations (2.5) and (2.6) are replaced by

$$\begin{aligned} \phi_n(0) &= n + \sum_{k=1}^{n-1} (n-k)q_k\phi_k(0) \\ &= n\left[1 + \sum_{k=1}^{n-1} q_k\phi_k(0)\right] - \sum_{k=1}^{n-1} kq_k\phi_k(0), \end{aligned} \tag{2.17a}$$

$$\theta_n(0) = 1 + \sum_{k=1}^{n-1} (n-k) q_k \theta_k(0)$$

$$= \left[1 - \sum_{k=1}^{n-1} q_k \theta_k(0) \right] + n \sum_{k=1}^{n-1} k q_k \theta_k(0) \tag{2.17b}$$

for $n = 1, 2, \cdots$. Under the condition (1.3) with $\lambda = 0$, the methods of differential equations yield a pair of linearly independent solutions $\{z_n(0)\}$, $\{w_n(0)\}$ of (1.1) such that as $n \to \infty$, $z_n(0) = [1 + o(1)]$ and $w_n(0) = [n(1 + o(1)]$. Returning to $\{\phi_n(0)\}$, we see that if $\{\phi_n(0)\}$ is bounded, then by (2.17a) $1 + \sum_{k=1}^{\infty} q_k \phi_k(0) = 0$. On the other hand, since $\{\phi_n(0)\}$ is a linear combination of $\{z_n(0)\}$ and $\{w_n(0)\}$, it follows from (2.17a) that $\phi_n(0) = o(n)$ as $n \to \infty$ if $1 + \sum_{k=1}^{\infty} q_k \phi_k(0) = 0$. This then implies that $\{\phi_n(0)\}$ and $\{z_n(0)\}$ are linearly dependent, so that $\{\phi_n(0)\}$ is bounded. Thus $\{\phi_n(0)\}$ is bounded if and only if $1 + \sum_{k=1}^{\infty} q_k \phi_k(0) = 0$ in which case $\sum_{k=1}^{\infty} k q_k \phi_k(0) \neq 0$, for otherwise (2.17a) implies that as $n \to \infty$,

$$|\phi_n(0)| \leqq n \left| \sum_{k=n}^{\infty} q_k \phi_k(0) \right| + o(1)$$

$$\leqq \sum_{k=n}^{\infty} k |q_k| |\phi_k(0)| + o(1) = o(1),$$

contrary to $\{\phi_n(0)\}$ being a nonzero multiple of $\{z_n(0)\}$.

In a similar manner it follows that $\{\theta_n(0)\}$ is bounded if and only if $\sum_{k=1}^{\infty} q_k \theta_k(0) = 0$, in which case $1 - \sum_{k=1}^{\infty} k q_k \theta_k(0) \neq 0$.

We now examine the behaviour of $f_\phi(\lambda)$ as $\lambda \to 0$ in $\mathbb{C} \backslash [0, 4]$. First consider $f_\phi(\lambda)$ when $\phi(0)$ is not an HBS. Then

$$f_\phi(\lambda) = 1 + \sum_{k=1}^{\infty} e^{ik\gamma} q_k \phi_k(\lambda)$$

$$= \left[1 + \sum_{k=1}^{\infty} q_k \phi_k(0) \right] + \sum_{k=1}^{\infty} e^{ik\gamma} q_k [\phi_k(\lambda) - \phi_k(0)] + \sum_{k=1}^{\infty} (e^{ik\gamma} - 1) q_k \phi_k(0). \tag{2.18}$$

When $\phi(0)$ is an HBS we have

$$f_\phi(\lambda) = i\gamma \sum_{k=1}^{\infty} k q_k \phi_k(0) + \sum_{k=1}^{\infty} e^{ik\gamma} q_k [\phi_k(\lambda) - \phi_k(0)]$$

$$+ \sum_{k=1}^{\infty} (e^{ik\gamma} - 1 - ik\gamma) q_k \phi_k(0). \tag{2.19}$$

In both cases we need estimates on the differences $\phi_k(\lambda) - \phi_k(0)$. The difficult case is when ϕ is an HBS and we consider this estimate first.

From (2.5) and (2.17a) we have that for $n \geqq 1$,

$$\phi_n(\lambda) - \phi_n(0) = \left[\frac{\gamma}{\sin \gamma} - 1 \right] \phi_n(0) + \frac{\gamma}{\sin \gamma} D(\lambda) \tag{2.20}$$

where

$$
\begin{aligned}
D(\lambda) &= \frac{\sin\gamma}{\gamma}\phi_n(\lambda) - \phi_n(0) \\
&= \left[\frac{\sin n\gamma}{\gamma} - n\right]\left[1 + \sum_{k=1}^{n-1} q_k\phi_k(0)\right] \\
&\quad + \sum_{k=1}^{n-1} \gamma^{-1}\sin(n-k)\gamma\, q_k[\phi_k(\lambda) - \phi_k(0)] \\
&\quad + \sum_{k=1}^{n-1} \gamma^{-1}[-\sin n\gamma + k\gamma + \sin(n-k)\gamma] q_k\phi_k(0).
\end{aligned} \tag{2.21}
$$

For the sine function we have for some constant c,

$$
|\sin z - z| \leqq c|z|g(|z|)\, e^{\mathcal{I}z}, \qquad |\cos z - 1| \leqq cg(|z|)\, e^{\mathcal{I}z} \tag{2.22}
$$

on $\mathcal{I}z \geqq 0$, where $g(t) = t^2/(1+t^2)$. Also, for $\mathcal{I}\gamma \geqq 0$,

$$
|-\sin n\gamma + k\gamma + \sin(n-k)\gamma| = \left|\gamma \int_{n-k}^{n} [1-\cos\gamma t]\, dt\right| \leqq c|\gamma| kg(n|\gamma|)\, e^{\mathcal{I}n\gamma} \tag{2.23}
$$

by application of (2.22). Also by application of (2.22) and $g(t) \leqq 1$ we have

$$
\left|\frac{\sin n\gamma}{\gamma} - n\right| \leqq cng(n|\gamma|)\, e^{\mathcal{I}n\gamma}, \tag{2.24}
$$

$$
\left|\frac{\sin(n-k)\gamma}{\gamma}\right| \leqq n + cn\, e^{\mathcal{I}(n-k)\gamma} \leqq n(1+c)\, e^{\mathcal{I}n\gamma}. \tag{2.25}
$$

Define $\tilde{\phi}_n(\lambda) = e^{in\gamma}\phi_n(\lambda)$ and $\tilde{\phi}_n(0) = e^{in\gamma}\phi_n(0)$. Further note that for λ sufficiently small, $\lambda \in \mathbb{C}\backslash[0,4]$, we have for some constant c_1,

$$
|\gamma^{-1}\sin\gamma| \leqq c_1, \qquad |\gamma(\sin\gamma)^{-1} - 1| \leqq c_1|\gamma|^2. \tag{2.26}
$$

Multiplying (2.20) by $e^{in\gamma}$ and using (2.23)–(2.26) gives that

$$
\begin{aligned}
|\tilde{\phi}_n(\lambda) - \tilde{\phi}_n(0)| &\leqq c_1|\gamma|^2|\tilde{\phi}_n(0)| + c_1 cng(n|\gamma|)\left|1 + \sum_{k=1}^{n-1} q_k\phi_k(0)\right| \\
&\quad + c_1 \sum_{k=1}^{n-1} n(1+c)|q_k||\tilde{\phi}_k(\lambda) - \tilde{\phi}_k(0)| \\
&\quad + c_1 \sum_{k=1}^{n-1} ckg(n|\gamma|)|q_k||\tilde{\phi}_k(0)|.
\end{aligned} \tag{2.27}
$$

Suppose now $\phi(0)$ is an HBS. Then $\{\tilde{\phi}_n(0)\}$ is bounded and from (2.27) there is a constant d_1, independent of λ, such that (recall that $n[1+\sum_{k=1}^{n-1} q_k\phi_k(0)]$ is bounded from (2.17a))

$$
|\tilde{\phi}_n(\lambda) - \tilde{\phi}_n(0)| \leqq d_1\left\{|\gamma|^2 + g(n|\gamma|) + n\sum_{k=1}^{n-1} |q_k||\tilde{\phi}_k(\lambda) - \tilde{\phi}_k(0)|\right\}. \tag{2.28}
$$

To solve (2.28), let $\alpha_n = \sum_{k=1}^{n} |q_k||\tilde{\phi}_k(\lambda) - \tilde{\phi}_k(0)|$ for $n \geqq 1$. Then

$$
\alpha_n - \alpha_{n-1} = |q_n||\tilde{\phi}_n(\lambda) - \tilde{\phi}_n(0)| \leqq a_n + b_n\alpha_{n-1}
$$

where $a_n = |q_n|d_1[|\gamma|^2 + g(n|\gamma|)]$ and $b_n = d_1 n|q_n|$. Hence

$$
\alpha_n \leqq a_n + (1+b_n)\alpha_{n-1},
$$

and it follows by induction since $\alpha_1 = 0$ that for $n \geqq 2$,

$$\alpha_n \leqq a_n + \sum_{k=2}^{n-1} a_k(1+b_{k+1}) \cdots (1+b_n).$$

Since $(1+b_k) \leqq e^{b_k}$, it follows that

$$\alpha_n \leqq c_2 \sum_{k=1}^{n} a_k, \qquad c_2 = \exp\left\{ d_1 \sum_{n=1}^{\infty} n|q_n| \right\}$$

and thus

$$|\tilde{\phi}_n(\lambda) - \tilde{\phi}_n(0)| \leqq d_1 \left\{ |\gamma|^2 + g(n|\gamma|) + c_2 n \sum_{k=1}^{n-1} a_k \right\}. \tag{2.29}$$

We return now to (2.19). Using (2.29), we obtain

$$\left| \sum_{k=1}^{\infty} e^{ik\gamma} q_k [\phi_k(\lambda) - \phi_k(0)] \right| \leqq \sum_{k=1}^{\infty} |q_k| |\tilde{\phi}_k(\lambda) - \tilde{\phi}_k(0)|$$

$$\leqq d_1 \sum_{k=1}^{\infty} |q_k| [|\gamma|^2 + g(k|\gamma|)] + d_1 c_2 \sum_{k=1}^{\infty} k|q_k| \sum_{s=1}^{k} a_s.$$

The sum

$$\sum_{k=1}^{\infty} |q_k| q(k|\gamma|) = |\gamma| \sum_{k=1}^{\infty} (k|q_k|) \frac{k|\gamma|}{1+k^2|\gamma|^2} = o(|\gamma|)$$

as $|\gamma| \to 0$ by the Lebesgue dominated convergence theorem. The sum

$$\sum_{k=1}^{\infty} k|q_k| \sum_{s=1}^{k} a_s \leqq \sum_{s=1}^{\infty} d_1 |q_s| [|\gamma|^2 + g(s|\gamma|)] \sum_{k=1}^{\infty} k|q_k| = o(|\gamma|)$$

as $|\gamma| \to 0$ for the same reason. Hence

$$\sum_{k=1}^{\infty} e^{ik\gamma} q_k [\phi_k(\lambda) - \phi_k(0)] = o(|\gamma|)$$

as $|\gamma| \to 0$. Since $\mathcal{I}\gamma > 0$,

$$|e^{ik\gamma} - 1 - ik\gamma| \leqq d \frac{k^2|\gamma|^2}{1+k|\gamma|}$$

for some constant d independent of γ. Again by the Lebesgue dominated convergence theorem, as $|\gamma| \to 0$,

$$\sum_{k=1}^{\infty} (e^{ik\gamma} - 1 - ik\gamma) q_k \phi_k(0) = o(|\gamma|).$$

Substituting these estimates into (2.19) yields (2.15) for $\lambda = 0$ an HBS, as $\lambda \to 0$.

For $\lambda = 0$ not an HBS, then equation (2.28) is replaced by

$$|\tilde{\phi}_n(\lambda) - \tilde{\phi}_n(0)| \leqq d_1 \left\{ |\gamma|^2 + ng(n|\gamma|) + n \sum_{k=1}^{n-1} |q_k| |\tilde{\phi}_k(\lambda) - \tilde{\phi}_k(0)| \right\}.$$

An analysis like that above applied to (2.18) shows that as $\lambda \to 0$, (2.16) holds. This completes the proof of Theorem 2.1.

Remarks. A similar analysis applies to the solution $\theta(\lambda)$ at $\lambda = 0$, and we omit the details. The result is if $\theta(0)$ is an HBS, then as $\lambda \to 0$,

$$f_\theta(\lambda) = i a_\theta^{(1)} \gamma + o(|\gamma|), \qquad a_\theta^{(1)} = -1 + \sum_{k=1}^{\infty} k q_k \theta_k(0) \neq 0 \tag{2.30}$$

while if $\theta(0)$ is not an HBS, then

$$f_\theta(\lambda) = a_\theta^{(0)} + o(1), \qquad a_\theta^{(0)} = \sum_{k=1}^{\infty} q_k \theta_k(\lambda) \neq 0. \tag{2.31}$$

Note that as $\lambda \to 0$, series expansions of $i\gamma = \log \tau \approx (\tau - 1)$ and $-\lambda = (\tau - 1) + (\tau^{-1} - 1) \approx (\tau - 1)^2$ show that $\gamma/\sqrt{\lambda} \to 1$ as $\lambda \to 0$ in $\mathbb{C}\backslash[0, 4]$. Thus γ may be replaced by $\sqrt{\lambda}$ in the equations (2.15)-(2.16) and (2.30)-(2.31) where $\sqrt{\lambda}$ denotes the branchcut along the positive real axis. At $\lambda = 4$, the same analysis as above holds where γ is replaced in (2.15)-(2.16) and (2.30)-(2.31) by $\sqrt{\lambda - 4}$ and the branchcut for $\sqrt{}$ is along the negative real axis.

We are now ready to state Levinson's theorem for (1.1)-(1.3), and we do so for the general operator T_α. Let $y_\alpha(\lambda)$ be the solution of (1.1) satisfying $y_0 = -\alpha$ and $y_1 = 1$. Then $y_\alpha(\lambda) = \{y_n(\lambda)\}$ satisfies (1.2); furthermore, $y_\alpha(\lambda) = -\alpha\theta(\lambda) + (1+\alpha)\phi(\lambda)$ and $e^{in\gamma} y_n(\lambda) \to (1+\alpha) f_\phi(\lambda) - \alpha f_\theta(\lambda)$ as $n \to \infty$ for $\lambda \in \mathbb{C}\backslash[0, 4]$. Inspection of the bound (2.9) for $\phi(\lambda)$, and analogously for $\theta(\lambda)$, shows that the bound holds as $\lambda \to \lambda_0 \in (0, 4)$, $\mathcal{I}\lambda \neq 0$. Since the $\phi_k(\lambda)$ and $\theta_k(\lambda)$ are polynomials in λ, this means that the formulas (2.12) have a continuous extension from $\mathcal{I}\lambda > 0$ to include the interval $(0, 4)$.

Recall from the discussion above equations (2.15) that $f_\phi(\lambda) \to 1$ as $|\lambda| \to \infty$, $\lambda \in \mathbb{C}\backslash[0, 4]$.

THEOREM 2.2. *Let $f_\alpha(\lambda) = (1+\alpha) f_\phi(\lambda) - \alpha f_\theta(\lambda)$. Under the condition* (1.3), *the number N of eigenvalues of T_α is given by*

$$2\pi N = 2[\arg f_\alpha(4) - \arg f_\alpha(0)] + \Delta_0 + \Delta_4$$

where $f_\alpha(4)$, $f_\alpha(0)$ is the limiting argument of $f_\alpha(\lambda)$ as $\lambda \to 4$, 0, respectively, in $(0, 4)$ *($f_\alpha(\lambda)$ in* $(0, 4)$ *is the continuous extension of $f_\alpha(\lambda)$ from $\mathcal{I}\lambda > 0$) and*

$$\Delta_j = \begin{cases} -\pi & y_\alpha \text{ has an HBS at } j = \lambda, \\ 0 & y_\alpha \text{ does not have an HBS at } j = \lambda. \end{cases}$$

Proof. For simplicity we consider the $\alpha = 0$ case; the general proof is similar. We apply the argument to the contour $\Gamma = \Gamma_1 + \Gamma_2$, where Γ_1 is a circle about the origin of radius R, and Γ_2 consists of two circles of radius ε about $\lambda = 0, 4$ and joined by two intervals above and below the slit $0 < \lambda < 4$; orientation of Γ_1 and Γ_2 is counterclockwise. Since T_0 is bounded, for R sufficiently large, all eigenvalues are interior to Γ_1. Also the eigenvalues are exterior to Γ_2 for ε sufficiently small. The eigenvalues, which are zeros of $f_\phi(\lambda)$, must be finite in number, i.e., cannot have zero or four as an accumulation point, by the asymptotics (2.15)-(2.16). Let S_ε be the circle of radius ε at zero. By the asymptotics (2.15)-(2.16), with γ replaced by $\sqrt{\lambda}$, it is clear that [*change of argument in $f_\phi(\lambda)$ on S_ε*] $\to \pi$ or 0 as $\varepsilon \to 0$ according to whether or not $\phi(0)$ is an HBS. Similar calculations apply at $\lambda = 4$. By the argument principle

$$2\pi N = \int_{\Gamma_1} \frac{f_\phi'(\lambda)}{f_\phi(\lambda)}\, d\lambda - \int_{\Gamma_2} \frac{f_\phi'(\lambda)}{f_\phi(\lambda)}\, d\lambda.$$

Since $f_\phi(\lambda) \to 1$ as $\lambda \to \infty$, $\int_{\Gamma_1} f_\phi'/f_\phi \to 0$ as $R \to \infty$. Letting $R \to \infty$ as $\varepsilon \to 0$ completes the proof by recalling $f_\phi(\lambda)$ is analytic on $\mathbb{C}\backslash[0, 4]$ with $f_\phi(\bar{\lambda}) = \overline{f_\phi(\lambda)}$.

A simple example is obtained by taking one q_n nonzero, say $q_m \neq 0$ and $q_n = 0$ for $n \neq m$. Then from (2.12) we have

$$f_\phi(\lambda) = 1 + e^{im\gamma} q_m \phi_m(\lambda) = 1 + e^{im\gamma} q_m \sin m\gamma / \sin \gamma.$$

So for $\alpha = 0$ in (1.2), the eigenvalues of (1.1)-(1.2) are the solutions of $f_\phi(\lambda) = 0$. For

$m=1$ this is the simple equation $1+e^{i\gamma}q_1=1+\tau q_1=0$ or $\tau=-1/q_1$. Since $|\tau|<1$ we have a single eigenvalue given by $\lambda=2-\tau-1/\tau=2+q_1+1/q_1$ when $|q_1|>1$.

The asymptotics (2.15)-(2.16) may be substituted into (2.14) to yield asymptotics for $m(\lambda)$ at an endpoint of the essential spectrum. At $\lambda=0$ we obtain that as $\lambda\to 0$,

$$m(\lambda)\sim\begin{cases}-a_\theta^{(0)}/a_\phi^{(0)} & \text{neither } \phi(0) \text{ nor } \theta(0) \text{ an HBS},\\ -a_\theta^{(0)}/(ia_\phi^{(1)}\sqrt{\lambda}) & \phi(0) \text{ an HBS},\\ -ia_\theta^{(1)}\sqrt{\lambda}/a_\phi^{(0)} & \theta(0) \text{ an HBS}.\end{cases}$$

3. An application. In this section we consider the effect of half-bound states on the large-t asymptotics of a spring-mass system. Such a system is described by displacements $u_n(t)$ of the nth mass, where

$$M_n\ddot{u}_n=u_{n+1}-2u_n+u_{n-1},\quad n=1,2,\cdots,\quad \ddot{u}=d^2u/dt^2 \tag{3.1}$$

with $u_0\equiv 0$. Physically this represents a semi-infinite linear chain of springs (with spring constants equal to 1) and masses M_n where the first mass u_0 is attached on the left to a wall. The horizontal displacement at time t of the mass M_k from its rest position is $u_k(t)$. Let H denote the operator $(Hy)_n=y_{n+1}-2y_n+y_{n-1}$, $y_0=0$, $n=1,2,\cdots$, and let M be the operator $(My)_n=M_ny_n$. Assume $M_n=1+\delta_n$, where $\delta_n>-1$, and

$$\sum_{n=1}^{\infty} n|\delta_n|<\infty. \tag{3.2}$$

In l_2, H is the self-adjoint operator $-T_0$ where T_0 is as in § 2. Then H has continuous spectrum on $[-4,0]$. Let $l_{2,M}$ be the Hilbert space of sequences with inner product $\langle\{y_n\},\{z_n\}\rangle_M=\langle\{y_n\},M\{z_n\}\rangle$ where $\langle\cdot,\cdot\rangle$ is the inner product in l_2. Under (3.2) the norm $l_{2,M}$ is equivalent to the norm in l_2, and the operator

$$A=-M^{-1}H \tag{3.3}$$

is a bounded self-adjoint operator in $l_{2,M}$; furthermore, $A\geqq 0$, so $A^{1/2}$ exists. Since (3.1) is the same as $\ddot{u}+Au=0$, it can be solved by [10]

$$u(t)=(\cos A^{1/2}t)u(0)+(A^{-1/2}\sin A^{1/2}t)\dot{u}(0) \tag{3.4}$$

where $u(0)$ and $\dot{u}(0)$ denote the initial values. For simplicity we assume that

$$\dot{u}(0)=\{0\},\quad u(0)=\{\delta_{ns}\},\quad \delta_{ns}=\text{Kronecker delta}, \tag{3.5}$$

for some fixed s. Let $a^{(m)}$ denote the vectors with components $a_n^{(m)}=M_m^{-1}\delta_{nm}$, and let E_λ denote the spectral family for A in $l_{2,M}$. Then we have

$$u_m(t)=\langle a^{(m)},u(t)\rangle_M=\int_0^\infty \cos\sqrt{\lambda}\,t\,d_\lambda\langle a^{(m)},E_\lambda u(0)\rangle_M. \tag{3.6}$$

By Stone's formula, for a closed λ-interval Δ,

$$\langle h,E_\Delta g\rangle_M=\lim_{\varepsilon\to 0}\frac{1}{2\pi i}\int_\Delta\langle h,[(A-(\mu+i\varepsilon))^{-1}-(A-(\mu-i\varepsilon))^{-1}]g\rangle_M\,d\mu. \tag{3.7}$$

The resolvent of A can be constructed as follows. If $\Im\lambda\neq 0$ and

$$(A-\lambda)y=g, \tag{3.8}$$

then

$$(-H-\lambda\delta-\lambda)y=Mg \tag{3.9}$$

where δ is multiplication by δ_n in each component. The results of § 2 carry over in that the solution $\phi(\lambda)$ of the homogeneous equation (3.9) ($g=0$) is given by (2.5) with q_n replaced by $-\lambda\delta_n$. That is, we have

$$\phi_n(\lambda)=\frac{\sin n\gamma}{\gamma}-\sum_{k=1}^{n-1}\frac{\sin(n-k)\gamma}{\sin\gamma}\lambda\delta_k\phi_k(\lambda). \tag{3.10}$$

Also from (2.12),

$$f_\phi(\lambda)=1-\lambda\sum_{k=1}^{\infty}e^{ik\gamma}\delta_k\phi_k(\lambda). \tag{3.11}$$

The Jost solution $F(\lambda)$ of the homogeneous equation (3.9), i.e., the solution satisfying $F_n(\lambda)=e^{in\gamma}+o(1)$ as $n\to\infty$, may be calculated by the variation of constants formula to satisfy

$$F_n(\lambda)=e^{in\gamma}+\sum_{k=n+1}^{\infty}\frac{\sin(n-k)\gamma}{\sin\gamma}\lambda\delta_kF_k(\lambda). \tag{3.12}$$

Using the solutions $\phi(\lambda)$ and $F(\lambda)$ as a basis, we use variation of constants to solve (3.9), obtaining

$$y_n=-\phi_n(\lambda)W(\lambda)^{-1}\sum_{k=n}^{\infty}F_k(\lambda)M_kg_k-F_n(\lambda)W(\lambda)^{-1}\sum_{k=1}^{n-1}\phi_k(\lambda)M_kg_k \tag{3.13}$$

where $W(\lambda)$ is the Wronskian (independent of n),

$$W(\lambda)=\phi_n(\lambda)F_{n+1}(\lambda)-\phi_{n+1}(\lambda)F_n(\lambda). \tag{3.14}$$

Using $\phi_n(\lambda)=e^{-in\gamma}(2i\sin\gamma)^{-1}[f_\phi(\lambda)+o(1)]$ and $F_n(\lambda)=e^{in\gamma}[1+o(1)]$ as $n\to\infty$ gives that $W(\lambda)=-f_\phi(\lambda)$. Also note that $W(\lambda)=-F_0(\lambda)$. Suppose now that $\lambda\to\mu\pm i0$ with $0\leqq\mu\leqq4$. Using $F_n(\mu-i0)=\overline{F_n}(\mu+i0)$, $\phi_n(\mu-i0)=\phi_n(\mu+i0)$ (recall $\phi_n(\lambda)$ is a polynomial in λ [1, p. 97]) and $W(\mu-i0)=\bar{W}(\mu+i0)$, we have from (3.13) that

$$\begin{aligned}([(A-\lambda-i0)^{-1}-(A-\lambda+i0)^{-1}]g)_n&=-\phi_n(\mu)\sum_{k=n}^{\infty}\left[\frac{F_k(\mu)}{W(\mu)}-\frac{\overline{F_k}(\mu)}{\bar{W}(\mu)}\right]M_kg_k\\&\quad-\left[\frac{F_n(\mu)}{W(\mu)}-\frac{\overline{F_n}(\mu)}{\bar{W}(\mu)}\right]\sum_{k=1}^{n-1}\phi_k(\mu)M_kg_k\\&=\frac{2i\sin\gamma}{|W(\mu)|^2}\phi_n(\mu)\sum_{k=1}^{\infty}\phi_k(\mu)M_kg_k\end{aligned} \tag{3.15}$$

since by (3.14) and a calculation,

$$\begin{aligned}\bar{W}(\mu)F_n(\mu)-W(\mu)\overline{F_n}(\mu)&=\phi_n(\mu)[\bar{F}_{n+1}(\mu)F_n(\mu)-F_{n+1}(\mu)\overline{F_n}(\mu)]\\&=-2i\phi_n(\mu)\sin\gamma.\end{aligned}$$

The last equality follows from letting $n\to\infty$ in the Wronskian $\bar{F}_{n+1}(\mu)F_n(\mu)-F_{n+1}(\mu)\overline{F_n}(\mu)$ and using (3.12). For $\mu>4$, $f_\phi(\mu)$ is real and the arguments of § 2 show that the spectrum of A on $(4,\infty)$ consists of at most finitely many eigenvalues. Substitution of (3.15) into (3.7) for $0\leqq\lambda\leqq4$ gives

$$\langle h,E_\lambda g\rangle_M=\int_0^\lambda\frac{\sin\gamma}{\pi|W(\mu)|^2}\sum_{k=1}^{\infty}h_kM_k\phi_k(\mu)\sum_{k=1}^{\infty}g_kM_k\phi_k(\mu)\,d\mu$$

and thus (3.6) becomes

$$u_m(t) = M_s \int_0^4 \cos\sqrt{\lambda}\, t\, (\pi|W(\lambda)|^2)^{-1} \sin\gamma\phi_m(\lambda)\phi_s(\lambda)\, d\lambda + M_s \sum_{k=1}^{N} \cos\sqrt{\lambda_k}\, t\, \phi_s(\lambda_k)\phi_m(\lambda_k)\|\phi(\lambda_k)\|_M^{-2} \tag{3.16}$$

where the sum represents the contribution due to the possible eigenvalues in $(4, \infty)$ in (4.8). Below we will show that this sum never vanishes for any m and s when eigenvalues are present. First we determine the large-t asymptotics of the integral on the right of (3.16). In the following we only consider the case where $\{\delta_n\}$ has finite support, i.e.,

$$\delta_n = 0 \quad \text{for } n > N, \tag{3.17}$$

for some $N \geqq 1$.

Note that $\lambda = 0$ is not an HBS since $\phi_n(0) = n$. However $\lambda = 4$ may be an HBS and the analysis of § 2 shows that $W(\lambda) = -f_\phi(\lambda) \to W(4) \neq 0$ if $\lambda = 4$ is not an HBS and $(\lambda - 4)^{-1/2} W(\lambda) \to C \neq 0$ if $\lambda = 4$ is an HBS.

First suppose that $\lambda = 4$ is not an HBS. Suppose also that $\phi_m(4)\phi_s(4) \neq 0$. Let η_1, η_2 be nonnegative C^∞ functions such that $\eta_1(x) + \eta_2(x) = 1$, $\eta_1(x) = 1$ in a neighborhood of zero, and $\eta_1(x) = 0$ in a neighborhood of 4. Then $u_m(t) = I_1(t) + I_2(t)$ in (3.16), where

$$I_k(t) = M_s \int_0^4 \cos\sqrt{\lambda}\, t\, (\pi|W(\lambda)|^2)^{-1} \sin\gamma\, \phi_m(\lambda)\phi_s(\lambda)\eta_k(\lambda)\, d\lambda. \tag{3.18}$$

In I_1 put $u = \sqrt{\lambda}$ so that

$$\sin\gamma = \frac{(\tau - \tau^{-1})}{2i} = \frac{(\lambda^2 - 4\lambda)^{1/2}}{2i} = \frac{u(4-u^2)^{1/2}}{2}. \tag{3.19}$$

Then I_1 is of the form (for a certain function g)

$$I_1(t) = \int_0^2 (\cos ut) u^2 \eta_1(u^2) g(u)\, du. \tag{3.20}$$

Under condition (3.17), it follows from (3.11) and (3.19) that $W(u^2)$ is C^∞, in fact analytic, for $0 \leqq u < 2$. This allows us to integrate by parts twice in (3.20) with the result that $I_1(t) = O(t^{-2})$ as $t \to \infty$.

In $I_2(t)$ put $u = \sqrt{\lambda}$ again to obtain I_2 of the form

$$I_2(t) = \int_0^2 (\cos ut)(2-u)^{1/2} \eta_2(u) h(u)\, du \tag{3.21}$$

where $h(2) \neq 0$, because $\phi_m(4)\phi_s(4) \neq 0$ and h is C^∞ under (3.17). The substitution $s = 2 - u$ gives

$$I_2(t) = \int_0^2 (\cos(2-s)t) s^{1/2} g(s)\, ds, \qquad g(s) = \eta_2(2-s)h(2-s),$$

which after integration by parts reduces to

$$I_2(t) = \int_0^2 t^{-1}(\sin(2-s)t)(2^{-1}s^{-1/2}g(s) + s^{1/2}g'(s))\, ds. \tag{3.22}$$

In (3.22) write $g(s)=g(0)+s\tilde{g}(s)$ and then integrate by parts the terms with $\tilde{g}$ and g'. This gives

$$I_2(t)=\int_0^2 t^{-1}(\sin(2-s)t)2^{-1}s^{-1/2}g(0)\,ds+O(t^{-2})$$
$$=\left(\frac{1}{2t}\right)g(0)\left[\sin 2t\int_0^2 s^{-1/2}\cos st\,ds-\cos 2t\int_0^2 s^{-1/2}\sin st\,ds\right]+O(t^{-2}).$$

Since

$$\int_0^2 s^{-1/2}\cos st\,ds=t^{-1/2}\int_0^{2t} w^{-1/2}\cos w\,dw$$
$$=t^{-1/2}\left[\int_0^{\infty} w^{-1/2}\cos w\,dw-\int_{2t}^{\infty} w^{-1/2}\cos w\,dw\right]$$
$$=t^{-1/2}\left[\left(\frac{\pi}{2}\right)^{1/2}+O(t^{-1/2})\right]$$

with a similar expression for $\int_0^2 s^{-1/2}\sin st\,ds$, we have that

$$I_2(t)=\frac{g(0)\pi^{1/2}}{2t^{3/2}}\sin\left(2t-\frac{\pi}{4}\right)+O(t^{-2}),\qquad g(0)\neq 0, \tag{3.23}$$

as $t\to\infty$.

Now suppose $\lambda=4$ is an HBS. Then near $\lambda=4$, $W(\lambda)$ is of the form $|W(\lambda)|^2=\alpha|\lambda-4|+o(\lambda-4)$, $\alpha\neq 0$. Near $\lambda=0$ the situation is the same as in the non-HBS case and $I_1(t)=O(t^{-2})$ under (3.17). For I_2 however, we get in place of (3.21),

$$I_2(t)=\int_0^2(\cos ut)(2-u)^{-1/2}\eta_2(u)h(u)\,du$$

where again $h(2)\neq 0$ if $\phi_m(4)\phi_s(4)\neq 0$. With $s=2-u$ and $h(2-s)=h(2)+\tilde{h}(s)$ an analysis as in the non-HBS case shows that as $t\to\infty$

$$I_2(t)=h(2)\left[\cos 2t\int_0^2 s^{-1/2}\cos st\,ds+\sin 2t\int_0^2 s^{-1/2}\sin st\,ds\right]+O(t^{-1})$$
$$=\frac{h(2)\pi^{1/2}}{t^{1/2}}\sin\left(2t+\frac{\pi}{4}\right)+O(t^{-1}),\qquad h(2)\neq 0. \tag{3.24}$$

In both cases $u_m(t)\sim I_2(t)$ as $t\to\infty$.

The case where $\phi_m(4)\phi_s(4)=0$ can be handled in a similar manner. First note that $\phi_n(\lambda)$ is a polynomial and its zeros are simple [1, p. 100]. Thus $\phi_m(\lambda)\phi_s(\lambda)$ is $O(4-\lambda)$ or $O((4-\lambda)^2)$ near 4. Then two more integrations by parts are possible in $I_1(t)$ to show that $I_1(t)=O(t^{-4})$. The integral I_2 has the form

$$I_2(t)=\int_0^2(\cos ut)(2-u)^p\eta_2(u)h(u)\,du,\qquad p=\frac{3}{2}\text{ or }\frac{5}{2}, \tag{3.25}$$

in the non-HBS case and

$$I_2(t)=\int_0^2(\cos ut)(2-u)^p\eta_2(u)h(u)\,du,\qquad p=\frac{1}{2}\text{ or }\frac{3}{2}, \tag{3.26}$$

in the HBS case. This implies that $I_2(t)\sim t^{-5/2}$ or $I_2(t)\sim t^{-7/2}$ in the non-HBS case and that $I_2(t)\sim t^{-3/2}$ or $I_2(t)\sim t^{-5/2}$ in the HBS case. Again the large-t asymptotics of $u_m(t)$ are determined by $I_2(t)$.

We now know in what ways the continuous spectrum may affect the large-t asymptotics of $u_m(t)$. In the following discussion we make our results more precise. For this we need a lemma. Its proof uses an idea from [13].

LEMMA 3.1. (a) *Suppose A has $N \geqq 1$ eigenvalues in $(4, \infty)$ and let λ_1 denote the largest such eigenvalue. Then*

$$(-1)^{n+1}\phi_n(\lambda_1) > 0, \qquad n = 1, 2, \cdots. \tag{3.27}$$

(b) *Suppose A has no eigenvalue in $(4, \infty)$. Then*

$$(-1)^{n+1}\phi_n(4) > 0, \qquad n = 1, 2, \cdots. \tag{3.28}$$

Proof. Put $w(t) = e^{At}w(0)$ where $w(0) = \{\delta_{ns}\}$ (see (3.5)). Let $a^{(m)}$ be as in (3.6) so that $w_m(t) = \langle a^{(m)}, w(t)\rangle_M$. We proceed to determine the large-t behavior of $w_m(t)$. Of course, this can be done by essentially repeating the steps following (3.6) but with $\cos\sqrt{\lambda}\, t$ replaced by $e^{\lambda t}$. We therefore omit the details and just state the results. In case (a) we have

$$w_m(t) \sim M_s\, e^{\lambda_1 t}\phi_s(\lambda_1)\phi_m(\lambda_1)\|\phi(\lambda_1)\|_M^{-2}, \qquad t \to \infty, \tag{3.29}$$

again provided $\phi_s(\lambda_1)\phi_m(\lambda_1) \neq 0$. Now we introduce the unitary (in $l_{2,M}$) transformation $(Uy)_n = (-1)^n y_n$, and let $\tilde{A} = UAU^{-1}$. Then

$$(\tilde{A}y)_n = M_n^{-1}(y_{n+1} + 2y_n + y_{n-1}) \qquad (y_0 = 0). \tag{3.30}$$

Moreover,

$$w_m(t) = \langle a^{(m)}, w(t)\rangle_M = \langle Ua^{(m)}, e^{\tilde{A}t}Uw(0)\rangle_M = (-1)^{m+s}\langle a^{(m)}, e^{\tilde{A}t}w(0)\rangle_M. \tag{3.31}$$

It is clear from (3.30) that $\tilde{A}$ leaves invariant the cone of vectors with nonnegative components. Thus $\langle a^{(m)}, e^{\tilde{A}t}w(0)\rangle_M \geqq 0$ for $t > 0$ (in fact, we have strict inequality since $e^{\tilde{A}t}y$ has strictly positive components if y has nonnegative ones). Put $s = 1$, then $\phi_s(\lambda_1) = 1$, and compare (3.31) and (3.29). Clearly, if $\phi_m(\lambda_1) \neq 0$ then $(-1)^{m+1}\phi_m(\lambda_1) > 0$. If $\phi_m(\lambda_1) = 0$ then $\phi_{m-1}(\lambda_1) = -\phi_{m+1}(\lambda_1)$ $(\neq 0)$ since $A\phi = \lambda_1\phi$ and so $(-1)^m\phi_{m-1}(\lambda_1) = (-1)^{m+1}\phi_{m+1}(\lambda_1) = -(-1)^{m+2}\phi_{m+1}(\lambda_1) < 0$, a contradiction. In other words, $\phi_m(\lambda_1) = 0$ is impossible. This proves (3.27).

In case (b) we find that

$$w_m(t) \sim c_1\phi_m(4)\phi_s(4)t^{-3/2}\, e^{4t}, \quad \lambda = 4 \text{ is an HBS} \tag{3.32}$$

and

$$w_m(t) \sim c_2\phi_m(4)\phi_s(4)t^{-1/2}\, e^{4t} \qquad \lambda = 4 \text{ is not an HBS} \tag{3.33}$$

where $c_1, c_2 > 0$. Comparing (3.32) and (3.33) with (3.31) and arguing as in part (a) yields (3.28). Lemma 3.1 is proved.

The important conclusion from this lemma is that $\phi_n(\lambda_1)$ and $\phi_n(4)$ do not vanish in the situations described in part (a) and (b), respectively. Therefore we can say the following.

THEOREM 3.2. *Suppose* (3.5) *and* (3.17) *hold. Then*:

(a) *If A has $N \geqq 1$ eigenvalues $4 < \lambda_N < \cdots < \lambda_1$, then*

$$u_m(t) = \sum_{k=1}^{N} c_k \cos\sqrt{\lambda_k}\, t + \varepsilon_m(t) \tag{3.34}$$

*where $c_1 \neq 0$ and $\varepsilon_m(t) = O(t^{-3/2})$ in the non-*HBS *case and $\varepsilon_m(t) = O(t^{-1/2})$ in the* HBS *case.*

(b) *If A has no eigenvalues, then*

$$u_m(t) = \alpha_1 t^{-3/2} \sin\left(2t - \frac{\pi}{4}\right) + O(t^{-2}), \qquad \alpha_1 \neq 0, \tag{3.35}$$

*in the non-*HBS *case and*

$$u_m(t) = \alpha_2 t^{-1/2} \sin\left(2t - \frac{\pi}{4}\right) + O(t^{-1}), \qquad \alpha_2 \neq 0, \tag{3.36}$$

in the HBS *case.*

As an example, let $\delta_n = (\mu - 1)\delta_{nm}$ in (3.1). Then by direct calculation, or letting $\gamma \to \pi$ in (3.10), we have

$$\phi_n(4) = (-1)^{n+1} n, \qquad n \leqq m,$$

$$\begin{aligned}\phi_n(4) &= (-1)^{n+1} n - (-1)^{n-m+1}(n-m)4(\mu-1)(-1)^{m+1} m \\ &= (-1)^{n+1}(n + 4(n-m)m(\mu-1)), \qquad n > m.\end{aligned}$$

Thus we have an HBS at $\lambda = 4$ if $4m(\mu - 1) = -1$, i.e., $M_m = \mu = 1 - 1/4m$. Then $\phi_n(4) = (-1)^{n+1} n$ for $n \leqq m$ and $\phi_n(4) = (-1)^{n+1} m$ for $n > m$. Also from (3.11)

$$\begin{aligned}f_\phi(\lambda) &= 1 - \lambda(\mu - 1)\, e^{im\gamma}(-1)^{m+1} m \\ &= 1 + (1-\mu)(\tau - 1)^2 (-1)^m m \tau^{m-1},\end{aligned}$$

using $\tau = e^{i\gamma}$ and $\lambda\tau = -(\tau - 1)^2$. As τ goes from -1 to 0, λ goes from 4 to ∞. Hence there is a zero of $f_\phi(\lambda)$ in $(4, \infty)$, i.e., eigenvalue, if $f_\phi(4) = 1 - 4m(1-\mu) < 0$ since $f_\phi(\infty) > 0$. Thus there is an eigenvalue if $\mu < 1 - 1/4m$. For $m = 1$ this happens if $M_1 < \frac{3}{4}$. For $M_1 = \frac{1}{2}$ we have $\phi_n(4) = (-1)^{n+1}(n - 2(n-1)) = (-1)^{n+1}(2 - n)$ for $n \geqq 2$. Thus $\phi_2(4) = 0$ indeed happens, but we also have an eigenvalue in this case.

The $t^{-1/2}$ decay of the oscillations of a single mass also occurs in (3.1) with $n = 0, \pm 1, \pm 2, \cdots$, i.e., a doubly infinite chain of masses and springs. With no perturbation terms, this system is solved exactly in [7, pp. 385-387] where the solution $u_n(t) = J_{2n}(2t)$, J_{2n} the $2n$th order Bessel function, also exhibits $t^{-1/2}$ time decay. This is to be expected since $\lambda = 4$ is an HBS for the full line unperturbed problem ($\phi_n(4) = (-1)^{n+1}$). We refer to Chapter III of [14, §§ 2.1-2.4] for a discussion of a similar problem in the continuous case.

REFERENCES

[1] F. V. ATKINSON, *Discrete and Continuous Boundary Value Problems*, Academic Press, New York, 1964.

[2] T. ASAHI, *Spectral theory of the difference equation*, Supplement Progr. Theoret. Phys., 36 (1966), pp. 55-96.

[3] C. BENNEWITZ, *Spectral theory for Sturm-Liouville equations*, Proc. London Math. Soc. (3), 59 (1989), pp. 294-338.

[4] K. M. CASE AND M. KAC, *A discrete version of the inverse scattering problem*, J. Math. Phys., 14 (1973), pp. 594-603.

[5] D. HINTON, M. KLAUS, AND J. K. SHAW, *Levinson's theorem and Titchmarsh-Weyl theory for Dirac systems*, Proc. Roy. Soc. Edinburgh, Sect. A, 109 (1988), pp. 173-186.

[6] ———, *Titchmarsh-Weyl-Levinson theory for a four-dimensional Hamiltonian system*, Proc. London Math. Soc. (3), 59 (1989), pp. 339-372.

[7] J. P. KEENER, *Principles of Applied Mathematics*, Addison-Wesley, Reading, MA, 1988.

[8] M. KLAUS, *On the variation-diminishing property of Schrödinger operators*, C. M. S. Conference Proceedings 8, American Mathematical Society, Providence, RI, 1986, pp. 199-204.

[9] N. LEVINSON, *On the uniqueness of the potential in a Schrödinger equation for a given asymptotic phase*, Mat. Fys. Medd., 25 (1949), pp. 3-29.

[10] R. LEIS, *Initial Boundary Value Problems in Mathematical Physics*, John Wiley, New York, and B. G. Teubner, Stuttgart, 1986.

[11] Z. Q. MA, *Levinson's theorem for a Dirac particle moving in a background magnetic monopole field*, Phys. Rev. D (3), 32 (1985), pp. 2203-2212.

[12] Z. Q. MA AND G. NI, *Levinson theorem for Dirac particles*, Phys. Rev. D (3), 31 (1985), pp. 1482-1488.

[13] M. MURATA, *Positive solutions and large time behaviors of Schrödinger semigroups, Simon's problem*, J. Funct. Anal., 56 (1989), pp. 300-310.

[14] A. G. RAMM, *Scattering by Obstacles*, D. Reidel, Dordrecht, the Netherlands, 1986.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 769–773, May 1991

011

CONDITIONS FOR OSCILLATION OF DIFFERENCE EQUATIONS WITH APPLICATIONS TO EQUATIONS WITH PIECEWISE CONSTANT ARGUMENTS*

I. GYÖRI†, G. LADAS‡, AND L. PAKULA‡

Abstract. By using z-transforms it is proved that every solution of a homogeneous system of linear difference equations with constant coefficient matrices oscillates componentwise if and only if its characteristic equation has no positive roots. This result is also applied to obtain necessary and sufficient conditions for the oscillation of all solutions of a linear system with piecewise constant arguments.

Key words. oscillation, difference equations, equations with piecewise constant arguments

AMS(MOS) subject classification. 39A12

1. Introduction.

Consider the homogeneous system of linear difference equations with constant coefficient matrices

$$y_{n+k} + P_1 y_{n+k-1} + \cdots + P_k y_n = 0 \tag{1}$$

for $n = 0, 1, \cdots$ where k is some positive integer, the coefficients P_i are real $r \times r$ matrices, and $\{y_n\}_{n=0}^{\infty}$ is a sequence of points in $\mathbf{R}^r$.

We associate with (1) its characteristic equation

$$\det(\lambda^k I + \lambda^{k-1} P_1 + \cdots + P_k) = 0. \tag{2}$$

Our aim in this paper is to establish the following result and then apply it to obtain necessary and sufficient conditions for the oscillation of all solutions of a linear system with piecewise constant arguments.

THEOREM 1. *Let k be a positive integer and let $P_1, \cdots, P_k$ be real $m \times m$ matrices. Then the following are equivalent:*

(a) *Every solution $\{y_n\}_{n=0}^{\infty}$ of equation (1) oscillates.*

(b) *The characteristic equation (2) has no positive roots.*

A sequence of real numbers $\{r_n\}_{n=0}^{\infty}$ is said to *oscillate* if the terms r_n are not eventually positive or eventually negative. Otherwise the sequence is called *nonoscillatory.*

For scalar equations, Theorem 1 was stated and proved (for $k \leq 3$) in [15], the proof being based on a lengthy analysis of the solution space of (1). A simple proof for scalar equations with positive coefficients was given in [12]. See also [11] for several applications of this theorem to scalar equations.

Our proof of Theorem 1 uses the method of z-transforms, the discrete analogue of the Laplace transform which has been recently used in [3] and [8] to establish the continuous analogue of Theorem 1 for delay differential equations. For more information about z-transforms, see, e.g., [14].

*Received by the editors January 23, 1989; accepted for publication (in revised form) May 14, 1990.

†This author is on leave from the Computing Centre, Szeged University of Medicine, 6720 Szeged, Pecsi u. 4/a, Hungary.

‡Department of Mathematics, The University of Rhode Island, Kingston, Rhode Island 02881.

2. Proof of Theorem 1. The proof of (a) $\Rightarrow$ (b) is simple. If (a) holds, (2) cannot have a positive root since if λ_0 were such a root then there would be a nonzero vector $\xi \in \mathbf{R}^m$ such that

$$(\lambda^k I + \lambda^{k-1}P_1 + \cdots + P_k)\xi = 0.$$

But then, $y_n = {\lambda_0}^n \xi$ is a solution of (1) with at least one nonoscillatory component.

The proof of (b) $\Rightarrow$ (a) uses the z-transform. Assume that (b) holds and, for the sake of contradiction, assume that (1) has a solution $y_n = [y_n^1, \cdots, y_n^r]^T$ with at least one nonoscillatory component. With no loss of generality we assume that $\{y_n^1\}$ is eventually positive. As (1) is autonomous, we will assume in fact that $y_n^1 > 0$ for $n \geq 0$.

Clearly, there exists a $c \in (0, \infty)$ such that $\| y_n \| < c^n$. Then the z-transform of $\{y_n\}$,

$$Y(z) = \sum_{n=0}^{\infty} y_n z^{-n}$$

exists for $|z| > c$. From (1) we find that

$$F(z)Y(z) = \Phi(z) \tag{3}$$

holds for $|z| > c$ where, with $P_0 = I$,

$$F(z) = \sum_{i=0}^{k} P_i z^{k-i} \tag{4}$$

and

$$\Phi(z) = \sum_{i=0}^{k} P_i \sum_{j=0}^{k-i-1} z^{k-i-j} y_j. \tag{5}$$

By hypothesis

$$\det[F(z)] \neq 0 \tag{6}$$

for $z \in (0, \infty)$.

Let $Y_1(z)$ be the z-transform of y_n^1 and let M be the modulus of the largest zero of $\det[F(z)]$. Then by Cramer's rule, for $|z| > \max\{c, M\}$,

$$\det[F(z)]Y_1(z) = \det[D(z)] \tag{7}$$

where $D(z)$ has components of $F(z)$ and $\Phi(z)$ as its entries. Clearly, the determinants in (7) are polynomials in z.

Let $W(z) = Y_1(1/z)$ so that $W(z)$ is a power series with positive coefficients having radius of convergence $\rho > 0$. Equation (7) holds for $|z| > 1/\rho$, equivalently, $\det[F(1/z)]W(z) = \det[D(1/z)]$ for $0 < |z| < \rho$. Now it is known (see, e.g.,[9, p. 133]) that a power series with positive coefficients having radius of convergence $\rho < \infty$ has a singularity (in the sense of analytic continuation) at $z = \rho$. But since $\det[F(1/\rho)] \neq 0$ we see that $\det[D(1/z)]/\det[F(1/z)]$ is analytic in a disk centered at ρ and agrees with $W(z)$ on that part of the disk where $|z| < \rho$. This contradiction shows that we

must have $\rho = \infty$ and it follows that (7) holds for $|z| > 0$. But then $y_n^1 = 0$ for all sufficiently large n since otherwise the left side of (7), but not the right side, would have an essential singularity at $z = 0$. This contradicts the assumption that $\{y_n^1\}$ is nonoscillatory and the proof is complete.

Remark 1. It can easily be seen that the conclusion of Theorem 1 is also true when k is a negative integer and $\det P_k \neq 0$.

Remark 2. It should be observed that we proved a little more than we stated in Theorem 1. Namely, if (1) has a solution $\{y_n\}$ with one component eventually zero, then the characteristic equation (2) has a real root.

3. Application to linear systems with piecewise constant arguments. Let $[\cdot]$ denote the greatest-integer function, $\mathbf{N}$ the set of nonnegative integers, and $\mathbf{R}^{r\times r}$ the set of all $r \times r$ matrices with real components.

In this section we will apply Theorem 1 to obtain necessary and sufficient conditions for the oscillation of all solutions of the system of linear differential equations with piecewise constant arguments

$$\dot{x}(t) + \sum_{j=-k}^{\ell} Q_j x([t+j]) = 0, \qquad t \geq 0 \tag{8}$$

where

$$k,\ \ell \in \mathbf{N} \quad \text{and} \quad Q_j \in \mathbf{R}^{r\times r} \quad \text{for} \quad j = -k, \cdots, \ell. \tag{9}$$

With (8) we associate initial conditions of the form

$$x(j) = a_j \quad \text{for} \quad j \in \{-k, \cdots, 0\} \cup \{0, \cdots, \ell - 1\} \tag{10}$$

with the convention that if $\ell = 0$ the last set in (10) is empty.

By a solution of (8) we mean a function x defined on the set $\{-k, \cdots, 0\} \cup (0, \infty)$ with values in $\mathbf{R}^r$ and which satisfies the following properties:

(i) x is continuous on $[0, \infty)$.

(ii) The derivative $\dot{x}(t)$ exists at each point $t \in [0, \infty)$ with the possible exception of the points $t \in \mathbf{N}$ where finite one-sided derivatives exist.

(iii) Equation (2) is satisified on each interval $[n, n+1)$ for $n \in N$.

Let $a_j \in \mathbf{R}^r$ for $\{-k, \cdots, 0\} \cup \{0, \cdots, \ell - 1\}$ be given. Then, as we will show in Lemma 1, the intial value problem (8) and (10) has a unique solution provided that

$$\begin{aligned} &\text{if} \quad \ell = 1 \quad \text{then} \quad \det(Q_1 + I) \neq 0, \\ &\text{if} \quad \ell \geq 2 \quad \text{and} \quad k = 0 \quad \text{then} \quad \det Q_\ell \neq 0 \end{aligned} \tag{11}$$

with no restrictions for other values of k and ℓ, where I is the $r \times r$ identity matrix.

A solution $x(t) = [x_1(t), \cdots, x_r(t)]^T$ of (8) is called *oscillatory* if every component $x_i(t)$ has arbitrarily large zeros.

During the last few years there has been a lot of activity concerning the oscillation and asymptotic behavior of differential equations with piecewise constant arguments. See, for example, [1], [2], [5]–[8], and [10].

Among other things, equations with piecewise constant arguments provide the simplest examples of differential equations capable of displaying chaotic behavior. For example, as remarked in [10] the unique solution of the initial value problem

$$\begin{cases} \dot{y}(t) = 3y([t]) - y^2([t]), \\ y(0) = a_0 \end{cases}$$

has the property that

$$y(n+1) = 4y(n) - y^2(n), \qquad n = 0, 1, \cdots.$$

If we choose $a_0 = 4\sin^2(\pi/9)$ then the unique solution of this difference equation is $y(n) = 4\sin^2(2^n\pi/9)$ which has period 3 and hence exhibits chaotic behavior by the results of Li and Yorke [13]. See also [4].

The following lemma is the basic existence and uniqueness result for the initial value problem (8) and (10).

LEMMA 1. *Assume that* (9) *and* (11) *hold. Then the initial value problem* (8) *and* (10) *has a unique solution* $x(t)$. *Furthermore,* $x(t)$ *is given by*

$$x(t) = a_n - \left(\sum_{j=-k}^{\ell} Q_j a_{n+j}\right)(t-n) \quad \text{for} \quad n \le t < n+1 \quad \text{and} \quad n \in \mathbf{N} \tag{12}$$

where $\{a_n\}$ *is a sequence of vectors in* $\mathbf{R}^r$ *which satisfies the difference equation.*

$$a_{n+1} + a_n - \sum_{j=-k}^{\ell} Q_j a_{n+j} \quad \text{for} \quad n = 0, 1, \cdots. \tag{13}$$

Proof. Let $x(t)$ be a solution of (8) and (10). Then in the interval $n \le t < n+1$, for any $n \in \mathbf{N}$, (8) becomes

$$\dot{x}(t) + \sum_{j=-k}^{\ell} Q_j a_{n+j} = 0, \qquad n \le t < n+1 \tag{14}$$

where we use the notation

$$a_n = x(n) \quad \text{for} \quad n \in \{-k, \cdots, 0, 1, \cdots\}.$$

Clearly, the solution of (14) with $x(n) = a_n$ is given by (12). By continuity of the solution, as $t \to n+1$, (12) yields (13). So far we have proved that if $x(t)$ is a solution of (8) and (10) then $x(t)$ is given by (12) where the sequence $\{a_n\}$ satisfies (13).

Conversely, let $\{a_n\}$ be the solution of (13) with initial values $a_{-k}, \cdots, a_0, \cdots, a_{\ell-1}$. Note that this solution exists and is unique provided that (11) holds. Now define $x(t)$ by (10) and (12). The it can easily be shown by direct substitution that $x(t)$ satisfies (8). The proof is complete.

The characteristic equation associated with the difference equation (13) is

$$\det\left(\lambda I - I + \sum_{j=-k}^{\ell} Q_j \lambda^j\right) = 0. \tag{15}$$

The main result in this section is the following theorem. For the scalar case see [12] and [15].

THEOREM 2. *Assume that conditions* (9) *and* (11) *hold. Then the following statements are equivalent:*

(a) *Every solution of* (8) *oscillates.*

(b) *Every solution of* (13) *oscillates.*

(c) *The characteristic equation* (15) *has no positive roots.*

Proof. The fact that (b) is equivalent to (c) follows from Theorem 1. The proof of (b) $\Rightarrow$ (a) is an obvious consequence of the fact that $x(n) = a_n$. It remains to show that (a) $\Rightarrow$ (b). To this end, assume that every component of every solution of (8) oscillates but, for the sake of contradiction, assume that (13) has a solution $\{a_n\}$ which is nonoscillatory. Let

$$a_n = [a_n^1, \cdots, a_n^r].$$

Then one of the components of a_n is eventually positive or eventually negative. Without loss of generality we will assume that the first component a_n^1 is eventually positive; that is, there exists an n_0 such that $a_n^1 > 0$ for $n \geq n_0$. From (12) and (13) and the continuity of $x(t)$ we see that for $n \leq t \leq n+1$ and $n \in \mathbf{N}$

$$x(t) = a_n - (a_n - a_{n+1})(t-n) = [1-(t-n)]a_n + (t-n)a_{n+1}.$$

Hence the first component of $x(t) = [x^1(t), \cdots, x^r(t)]$ is such that

$$x^1(t) = [1-(t-n)]a_n^1 + (t-n)a_{n+1}^1 > 0.$$

This contradicts the hypothesis that every component of $x(t)$ oscillates, and the proof is complete.

REFERENCES

[1] A. R. Aftabizadeh, J. Wiener, and Jian-Ming Xu, *Oscillatory and periodic properties of delay differential equations with piecewise constant argument*, Proc. Amer. Math. Soc., 99 (1987), pp. 673–679.

[2] A. R. Aftabizadeh and J. Wiener, *Oscillatory and periodic solutions for systems of two first order linear differential equations with piecewise constant arguments*, Applicable Anal., 26 (1988), pp. 327–333.

[3] O. Arino and I. Győri, *Necessary and sufficient condition for oscillation of neutral differential systems with several delays*, J. Differential Equations, to appear.

[4] L. A. V. Carvalho and K. L. Cooke, *A nonlinear equation with piecewise continuous argument*, Differential and Integral Equations, 1 (1988), pp. 354–367.

[5] K. L. Cooke and J. Wiener, *An equation alternately of retarded and advanced type*, Proc. Amer. Math. Soc., 99 (1987), pp. 726–732.

[6] ———, *Neutral differential equations with piecewise constant argument*, Boll. Un. Mat. Ital., 1-B (1987), pp. 321–346.

[7] I. Györi and G. Ladas, *Linearized oscillations for equations with piecewise constant arguments*, Differential and Integral Equations, 1 (1988), pp. 281–286.

[8] I. Győri, G. Ladas, and L. Pakula, *Oscillation theorems for delay differential equations via Laplace transforms*, Canadian Math. Bull., to appear.

[9] E. Hille, *Analytic Function Theory*, Ginn and Co., Boston, 1959.

[10] G. Ladas, *Oscillations of equations with piecewise constant mixed arguments*, in Differential Equations and Applications, Ohio University Press, Athens, OH, 1988, pp. 64–69.

[11] G. Ladas, *Explicit conditions for the oscillation of difference equations*, J. Math. Anal. Appl., to appear.

[12] G. Ladas, Ch. G. Philos, and Y. G. Sficas, *Necessary and sufficient conditions for the oscillation of difference equations*, Libertas Math. 9 (1989), pp. 121–125.

[13] T.-Y. Li and J. A. Yorke, *Period three implies chaos*, Amer. Math. Monthly, 82 (1975), pp. 985–992.

[14] R. E. Mickens, *Difference Equations*, Van Nostrand Reinhold Co., New York, 1987.

[15] E. G. Partheniadis, *Stability and oscillation of neutral delay differential equations with piecewise constant argument*, Differential and Integral Equations, 1 (1988), pp. 459–472.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 774-784, May 1991

012

ATKINSON'S SUPERLINEAR OSCILLATION THEOREM FOR MATRIX DIFFERENCE EQUATIONS*

ALLAN C. PETERSON† AND JERRY RIDENHOUR‡

Abstract. The concept of a generalized zero of a prepared solution of a superlinear matrix difference equation is introduced. Riccati techniques are used to establish necessary and sufficient conditions for all prepared solutions to be oscillatory.

Key words. generalized zero, oscillatory solution, superlinear equation, matrix difference equation, Riccati techniques

AMS(MOS) subject classifications. primary 34C10; secondary 34A34

1. Introduction. In 1955, Atkinson [3] proved that the second-order superlinear scalar ordinary differential equation

$$y''+f(t)y^{2n+1}=0, \qquad t\geqq 0$$

with $f(t)>0$ and continuous for each $t\geqq 0$ is oscillatory if and only if

$$\int_0^\infty tf(t)\,dt=\infty.$$

Versions of this result for the $m\times m$ matrix differential equation

$$Y''+(Y^nQ(t)Y^{*n})Y=0 \tag{1}$$

have recently been obtained by Kura [12], Butler and Erbe [4], and Ahlbrandt, Ridenhour, and Thompson [2]. In [2], it is shown that, when $Q(t)$ is Hermitian, positive definite, and continuous for each $t\geqq 0$, a necessary and sufficient condition for all prepared solutions of (1) which extend to infinity to be oscillatory is that

$$\int_0^\infty t\cdot\lambda_{\max}[Q(t)]\,dt=\infty.$$

Here, $\lambda_{\max}[Q(t)]$ denotes the maximum eigenvalue of $Q(t)$.

In the past few years, several papers (e.g., [8], [9], [11], [14], and [15]) have appeared on the oscillation theory of second-order scalar difference equations. Mingarelli [13] has shown that Atkinson's superlinear oscillation theorem is valid for second-order real scalar difference equations. The oscillation of second-order linear matrix differential equations has been studied extensively (see [5]–[7] and the several references therein). Ahlbrandt and Hooker [1] have studied the principal solutions of second-order matrix difference equations.

Our results show that Atkinson's theorem also carries over to the case of second-order superlinear matrix difference equations. In particular, we study the difference equation

$$\Delta^2Y(t-1)+[Y^n(t)Q(t)Y^{*n}(t)]Y(t)=0, \tag{2}$$

where $\mathbb{Z}^+$ is the set of positive integers, $t\in\mathbb{Z}^+$, $Y(t)$ and $Q(t)$ are $m\times m$ complex-valued matrices, $Y^*(t)$ is the Hermitian adjoint (i.e., conjugate transpose) of $Y(t)$, and $Q(t)$

* Received by the editors March 20, 1989; accepted for publication May 14, 1990.
† Department of Mathematics and Statistics, University of Nebraska, Lincoln, Nebraska 68588-0323.
‡ Department of Mathematics and Statistics, Utah State University, Logan, Utah 84322-3900.

is assumed to be Hermitian and positive definite for each $t \in \mathbb{Z}^+$. Since solutions are defined only at discrete points, the first consideration will be in arriving at an appropriate definition of zeros and generalized zeros of solutions. Once this is settled and a corresponding definition of an oscillatory prepared solution is made, the main result establishes that equation (2) is oscillatory if and only if

$$\sum_{t=1}^{\infty} t\lambda_{\max}[Q(t)] = \infty. \tag{3}$$

Although we employ Riccati techniques similar to those in [2], the arguments are different and more subtle. The difficulty presented by the generalized zeros and an important interplay between two different Riccati variables have no counterpart in the differential equations case.

2. Preliminaries and definitions. We see from (2) that specifying the values of $Y(t)$ at two consecutive positive integers leads iteratively to a unique solution $Y(t)$ defined for all $t \in \mathbb{Z}^+$. Hence, solutions always exist on all of $\mathbb{Z}^+$ even though (2) is nonlinear.

It is well known (see [2] and the references therein) that a satisfactory oscillation theory in the matrix case can be given only for the class of prepared solutions. A solution $Y(t)$ of (2) is said to be *prepared* if and only if

$$Y^*(t)\Delta Y(t) = \Delta Y^*(t)\, Y(t), \qquad t \in \mathbb{Z}^+.$$

It is easy to see that, for any solution $Y(t)$ of (2),

$$\Delta[\, Y^*(t)\Delta Y(t) - \Delta Y^*(t)\, Y(t)] = K,$$

and the solution $Y(t)$ is prepared only when K is the zero matrix.

Given a solution $Y(t)$ of (2), we introduce the Riccati functions $W(t)$ and $V(t)$, defined by

$$W(t) := \Delta Y(t-1)\, Y^{-1}(t-1)$$

and

$$V(t) := \Delta Y(t-1)\, Y^{-1}(t).$$

Obviously, $W(t)$ and $V(t)$ are defined only at integers t, where $Y^{-1}(t-1)$ and $Y^{-1}(t)$, respectively, exist. We have the following important identities relating Y, W, and V:

$$W(t) + I = Y(t)\, Y^{-1}(t-1), \tag{4}$$

$$I - V(t) = Y(t-1)\, Y^{-1}(t) = [\, W(t) + I]^{-1}. \tag{5}$$

Here, as elsewhere, I denotes the $m \times m$ identity matrix. We note that (4) follows from

$$W(t) = \Delta Y(t-1)\, Y^{-1}(t-1) = [\, Y(t) - Y(t-1)]\, Y^{-1}(t-1) = Y(t)\, Y^{-1}(t-1) - I$$

and (5) is proved in a similar fashion.

An absolutely essential fact in what follows is that, for a prepared solution of (2), Riccati functions $W(t)$ and $V(t)$ are Hermitian at points where they exist. The fact that $W(t)$ is Hermitian follows immediately from the definition of $W(t)$ and the fact that $Y(t)$ is prepared. We then see, from (5), that $V(t)$ is also Hermitian.

Let $Y(t)$ be a solution of (2). Using (2), (4), and the elementary formulas

$$\Delta[A(t)B(t)] = A(t+1)\Delta B(t) + \Delta A(t)B(t) = \Delta A(t)B(t+1) + A(t)\Delta B(t)$$

and

$$\Delta[A^{-1}(t)] = -A^{-1}(t)\Delta A(t)A^{-1}(t+1) = -A^{-1}(t+1)\Delta A(t)A^{-1}(t),$$

we obtain on intervals of integers where $W(t)$ exists that

$$\begin{aligned}\Delta W(t) &= \Delta^2 Y(t-1)Y^{-1}(t) - \Delta Y(t-1)Y^{-1}(t)\Delta Y(t-1)Y^{-1}(t-1)\\ &= -Y^n(t)Q(t)Y^{*n}(t) - \Delta Y(t-1)Y^{-1}(t-1)Y(t-1)Y^{-1}(t)W(t)\end{aligned}$$

or

$$\Delta W(t) = -Y^n(t)Q(t)Y^{*n}(t) - W(t)[W(t)+I]^{-1}W(t). \tag{6}$$

We will refer to (6) as the Riccati equation associated with (2). It will play a central role in the analysis to follow. We could just as well have derived a Riccati equation for $V(t)$ that is very similar to (6) and have based our analysis on it; however, either one of these equations together with (4) and (5) yields all the necessary information.

We are ready to define what we mean by generalized zeros and oscillatory solutions.

DEFINITION. We say a prepared solution $Y(t)$ of (2) has a *generalized zero* at $t_0 \in \mathbb{Z}^+$ if at least one of the eigenvalues of $W(t_0)+I$ is less than or equal to zero. A prepared solution $Y(t)$ of (2) is said to be *oscillatory* if, for any $n \in \mathbb{Z}^+$, $Y(t)$ has a generalized zero at some t_0 with $t_0 > n$; otherwise, we say the prepared solution $Y(t)$ is *nonoscillatory*. Equation (2) is said to be *oscillatory* if all prepared solutions are oscillatory and is said to be *nonoscillatory* if there exists at least one prepared solution which is *nonoscillatory*.

At first glance, our definition of a generalized zero seems unnatural. We justify it below, collecting our reasons under four headings.

I. *Real scalar equations.* Following Hartman [8], a discrete solution $y(t)$ of a real scalar difference equation is said to have a generalized zero at t if either $y(t)=0$ or $y(t-1)$ and $y(t)$ are of opposite signs. By (4), this occurs precisely when $w(t)+1 \leqq 0$.

II. *Complex scalar equations.* In this case, it follows from (4) and (2) that the function values of a prepared solution determine a set of points in the complex plane which all lie on a fixed line through the origin. When $w(t_0)+1<0$, the successive solution values $y(t_0-1)$ and $y(t_0)$ lie on opposite sides of the origin, a situation that agrees geometrically with the notion of a generalized zero.

III. *Hermitian solutions.* Suppose $Y(t)$ is a prepared Hermitian solution of (2) and t_0-1 is a positive integer for which $\det[Y(t_0-1)] \neq 0$. From matrix analysis [10, p. 229], if A and B are Hermitian with $AB=BA$, then there exists a unitary matrix U which simultaneously diagonalizes both A and B (i.e., $UAU^*=D$ and $UBU^*=\Lambda$, where D and Λ are diagonal matrices). Since $Y(t_0)=[W(t_0)+I]Y(t_0-1)$ is Hermitian, it follows that there is a unitary matrix $U(t_0)$ such that

$$U(t_0)[W(t_0)+I]U^*(t_0) = D(t_0) = \operatorname{diag}(d_1(t_0), \cdots, d_m(t_0))$$

and

$$U(t_0)Y(t_0-1)U^*(t_0) = \Lambda(t_0) = \operatorname{diag}(\lambda_1(t_0), \cdots, \lambda_m(t_0))$$

with $U(t_0)U^*(t_0)=I$. Therefore

$$\begin{aligned}Y(t_0) &= U^*(t_0)D(t_0)U(t_0)U^*(t_0)\Lambda(t_0)U(t_0)\\ &= U^*(t_0)\operatorname{diag}(d_1(t_0)\lambda_1(t_0), \cdots, d_m(t_0)\lambda_m(t_0))U(t_0).\end{aligned}$$

This shows that the eigenvalues of $Y(t_0)$ are products of eigenvalues of $W(t_0)+I$ with eigenvalues of $Y(t_0-1)$. That is, the unitary transformation $U(t_0)$ allows "tracking"

of eigenvalues in that the eigenvalue $\lambda_i(t_0)$ of $Y(t_0-1)$ is transformed into the eigenvalue $d_i(t_0)\lambda_i(t_0)$ of $Y(t_0)$. If some eigenvalue $d_i(t_0)$ of $W(t_0)+I$ is zero or negative, then the iteration from $Y(t_0-1)$ to $Y(t_0)$ either produces a singular matrix $Y(t_0)$ or an eigenvalue that changes sign. In a continuous evolution, an eigenvalue that changes sign must pass through zero yielding a point at which the matrix is singular. Therefore, a zero or negative eigenvalue of $W(t_0)+I$ implies the presence of a singularity of $Y(t)$ in passing from $Y(t_0-1)$ to $Y(t_0)$, thus supporting our definition of a generalized zero.

We now give an example which further illustrates the correctness of our definition in the case of Hermitian solutions. If there exists a constant unitary matrix U which diagonalizes $Q(t)$ for all $t\in\mathbb{Z}^+$, say, $UQ(t)U^*=D(t)=\operatorname{diag}(d_1(t),\cdots,d_m(t))$, then it is easy to see that $Y(t)=U^*Z(t)U$ is a prepared Hermitian solution of (2) provided that $Z(t)$ is a diagonal solution of

$$\Delta^2 Z(t-1)+Z^n(t)D(t)Z^{n+1}(t)=0.$$

If $Q(t)$ is the constant matrix given by

$$Q=\begin{bmatrix} 1.64 & -0.48 \\ -0.48 & 1.36 \end{bmatrix},$$

this technique can be used to find that $Y_0(t)$ given by

$$Y_0(t)=\left\{\begin{bmatrix} -0.28 & 0.96 \\ 0.96 & 0.28 \end{bmatrix},\begin{bmatrix} 0.56 & -1.92 \\ -1.92 & -0.56 \end{bmatrix},\begin{bmatrix} -5.96 & 6.72 \\ 6.72 & -2.04 \end{bmatrix},\begin{bmatrix} 1681.5 & -1275.4 \\ -1275.4 & 937.5 \end{bmatrix},\cdots\right\} \tag{7}$$

is a prepared Hermitian solution of

$$\Delta^2 Y(t-1)+Y(t)QY^*(t)Y(t)=0. \tag{8}$$

The sequence of eigenvalue pairs of $Y_0(t)$ is

$$\{(1,-1),(2,-2),(3,-1),(2638,-19),\cdots\}.$$

Since the eigenvalues of Q are $\lambda=1$ and $\lambda=2$ and (3) is consequently satisfied, all prepared solutions of (8), including $Y_0(t)$, ought to be oscillatory. Yet, simply by looking at $Y_0(t)$ as given in (7), or at the sequence of eigenvalue pairs, or perhaps at something like the sequence $\{\det(Y_0(t))\}$, it is difficult to detect why we ought to consider $Y_0(t)$ oscillatory. However, the sequence $\{W_0(t)+I\}$ has

$$\{(-2,-2),(-5.5,-1.5),(-239.82,-6.33),\cdots\}$$

as the corresponding sequence of eigenvalue pairs indicating that *both* eigenvalues of $Y_0(t)$ actually change sign at each iteration; hence, $Y_0(t)$ has infinitely many generalized zeros and is oscillatory by our definition.

IV. *The general case.* We will prove in the next section that (2) is oscillatory if and only if (3) holds. Therefore, the above definitions are appropriate for non-Hermitian as well as Hermitian solutions if the principal aim is to construct an oscillation theory for matrix difference equations which parallels the known results for matrix differential equations.

3. Main results. For any $m\times m$ Hermitian matrix H, we let $\lambda_1(H)\leqq\lambda_2(H)\leqq\cdots\leqq\lambda_m(H)$ denote the eigenvalues. Also, $\lambda_{\max}(H)$ and $\lambda_{\min}(H)$ will denote the maximum and minimum eigenvalues of H, respectively. We now mention three eigenvalue inequalities that will be used below. For the purpose of describing these inequalities,

assume H_1 and H_2 are any $m \times m$ Hermitian matrices, P is any $m \times m$ matrix that is Hermitian and positive definite, and A is any $m \times m$ matrix. The three inequalities are

$$\lambda_k(H_1)+\lambda_1(H_2) \leqq \lambda_k(H_1+H_2) \leqq \lambda_k(H_1)+\lambda_m(H_2),$$
$$\lambda_k(APA^*) \geqq \lambda_k(P)\lambda_1(AA^*),$$
$$\lambda_k(APA^*) \geqq \lambda_k(AA^*)\lambda_1(P),$$

all valid for $1 \leqq k \leqq m$. The first is due to Weyl [10, p. 181], the second is due to Ostrowski [10, pp. 224–225], and the third is an immediate consequence of the Courant–Fischer min-max theorem [10, p. 179].

We begin by proving a lemma which gives detailed information about the Riccati functions.

LEMMA. *Suppose $Y(t)$ is a nonoscillatory prepared solution of* (2). *Then there exists $t_0 \in \mathbb{Z}^+$ such that $W(t)$ and $V(t)$ are both positive definite and decreasing for $t \geqq t_0$ with*

$$\lim_{t\to\infty} W(t) = \lim_{t\to\infty} V(t) = 0. \tag{9}$$

Furthermore, multiplication of $W(t)$ and $V(t)$ is commutative at points where both exist and $W(t)V(t) = V(t)W(t)$ is positive definite for $t \geqq t_0$.

Proof. Since $Y(t)$ is nonoscillatory and prepared, we begin by choosing $t_0 \in \mathbb{Z}^+$ so that $W(t)+I>0$ for $t \geqq t_0$. Since $Q(t)$ is also positive definite, we see from the Riccati equation, Ostrowski's inequality, and Weyl's inequality that $\Delta W(t)<0$ for $t \geqq t_0$. Hence, by Weyl's inequality, each eigenvalue $\lambda_k[W(t)]$, $1 \leqq k \leqq m$, is a decreasing function of t for $t \geqq t_0$. Furthermore, each $\lambda_k[W(t)]$ is bounded below for $t \geqq t_0$ since $W(t)+I>0$. Therefore, $\lim_{t\to\infty} \lambda_k[W(t)]$ exists for $1 \leqq k \leqq m$. Since the eigenvalues of $W(t)+I$ decrease but remain positive, the eigenvalues of $[W(t)+I]^{-1}$ are positive and increasing for $t \geqq t_0$.

From the Riccati equation and the above-mentioned eigenvalue inequalities we obtain

$$\begin{aligned} \lambda_k[-\Delta W(t)] &> \lambda_k\{W(t)[W(t)+I]^{-1}W(t)\} \\ &\geqq \lambda_k[W^2(t)]\lambda_{\min}([W(t)+I]^{-1}) \\ &\geqq \{\lambda_k[W(t)]\}^2\lambda_{\min}([W(t_0)+I]^{-1}), \end{aligned} \tag{10}$$

valid for $1 \leqq k \leqq m$ and $t \geqq t_0$. We claim that

$$\lim_{t\to\infty} \lambda_k[W(t)] = 0 \tag{11}$$

holds for $1 \leqq k \leqq m$. Suppose not. We choose k_0 with $1 \leqq k_0 \leqq m$ such that

$$\lim_{t\to\infty} \lambda_{k_0}[W(t)] = \lambda_0 \neq 0. \tag{12}$$

In view of (10) and (12), we then choose an integer t_1 with $t_1 \geqq t_0$ and a positive number δ such that

$$\lambda_{k_0}[-\Delta W(t)] > \delta \quad \text{for } t \geqq t_1. \tag{13}$$

But this leads to (letting tr (A) denote the trace of A):

$$\begin{aligned} \lambda_{\max}[-W(t)+W(t_1)] &= \lambda_{\max}\left[\sum_{\tau=t_1}^{t-1} -\Delta W(\tau)\right] \geqq \frac{1}{m}\operatorname{tr}\left[\sum_{\tau=t_1}^{t-1} -\Delta W(\tau)\right] \\ &= \frac{1}{m}\sum_{\tau=t_1}^{t-1} \operatorname{tr}[-\Delta W(\tau)] \geqq \frac{1}{m}\sum_{\tau=t_1}^{t-1} \lambda_{k_0}[-\Delta W(\tau)]. \end{aligned}$$

By (13), this implies $\lambda_{\max}[-W(t)+W(t_1)]\to\infty$ and $\lambda_{\min}[W(t)]\to-\infty$ as $t\to\infty$. But this contradicts the fact that the eigenvalues of $W(t)$ are bounded below for $t\geqq t_0$.

This proves that (11) holds and, consequently, that $\lim_{t\to\infty} W(t)=0$. Therefore, $W(t)$ is positive for $t\geqq t_0$, $t\in\mathbb{Z}^+$. From (5),

$$V(t)=I-[W(t)+I]^{-1},$$

so the eigenvalues of $V(t)$ are also positive and decreasing for $t\geqq t_0$, with $\lim_{t\to\infty}\lambda_k[V(t)]=0$ for $1\leqq k\leqq m$ showing that (9) holds.

Finally, from (5), we see that

$$[W(t)+I][I-V(t)]=I=[I-V(t)][W(t)+I],$$

from which it follows that

$$W(t)V(t)=V(t)W(t)=W(t)-V(t)$$

at all integers t where both $W(t)$ and $V(t)$ exist. Since

$$\begin{aligned}V(t)[W(t)+I]V(t)&=V(t)W(t)V(t)+V^2(t)\\&=[W(t)-V(t)]V(t)+V^2(t)\\&=W(t)V(t),\end{aligned}$$

we see that $W(t)V(t)=V(t)W(t)$ is positive for $t\geqq t_0$ completing the proof of the lemma. □

The main result is the following theorem.

THEOREM 1. *Suppose $Q(t)$ is Hermitian and positive definite for all $t\in\mathbb{Z}^+$. Then equation* (2) *is oscillatory if and only if* (3) *holds.*

Proof. We first assume $n=1$ (a cubic nonlinearity) in (2). The general case will be treated later. Suppose (3) holds but (2) has a nonoscillatory prepared solution $Y(t)$. Applying the lemma, we choose $t_0\in\mathbb{Z}^+$ so that $Y(t)$ is invertible and the matrices $W(t)$, $V(t)$, and $W(t)V(t)=V(t)W(t)$ are all positive definite for $t\geqq t_0-1$, $t\in\mathbb{Z}^+$.

In the next few steps, we derive an identity relating $tQ(t)$ to $Y(t)$, $W(t)$, and $V(t)$. First,

$$\begin{aligned}\Delta[Y^{-1}(t-1)Y^{*-1}(t-1)]&=-Y^{-1}(t)\Delta Y(t-1)Y^{-1}(t-1)Y^{*-1}(t)\\&\quad-Y^{-1}(t-1)Y^{*-1}(t)\Delta Y^{*}(t-1)Y^{*-1}(t-1)\\&=-Y^{-1}(t)W(t)Y^{*-1}(t)\\&\quad-Y^{-1}(t-1)V(t)Y^{*-1}(t-1).\end{aligned}\tag{14}$$

Using the product rule

$$\begin{aligned}\Delta[A(t)B(t)C(t)D(t)E(t)]&=A(t+1)B(t+1)C(t+1)D(t+1)\Delta E(t)\\&\quad+A(t+1)\Delta B(t)C(t+1)D(t+1)E(t)\\&\quad+A(t+1)B(t)\Delta C(t)D(t+1)E(t)\\&\quad+A(t+1)B(t)C(t)\Delta D(t)E(t)\\&\quad+\Delta A(t)B(t)C(t)D(t)E(t),\end{aligned}$$

we obtain

$$
\begin{aligned}
&\Delta[(t-1)Y^{-1}(t)\Delta Y(t-1)Y^{-1}(t-1)Y^{*-1}(t)] \\
&\quad = -tY^{-1}(t+1)\Delta Y(t)Y^{-1}(t)Y^{*-1}(t)\Delta Y^{*}(t)Y^{*-1}(t+1) \\
&\qquad - tY^{-1}(t)\Delta Y(t)Y^{-1}(t+1)\Delta Y(t)Y^{-1}(t)Y^{*-1}(t) - tQ(t) \\
&\qquad - tY^{-1}(t)\Delta Y(t-1)Y^{-1}(t)\Delta Y(t-1)Y^{-1}(t-1)Y^{*-1}(t) \\
&\qquad + Y^{-1}(t)\Delta Y(t-1)Y^{-1}(t-1)Y^{*-1}(t) \\
&\quad = -tY^{-1}(t+1)W^{2}(t+1)Y^{*-1}(t+1) \\
&\qquad - tY^{-1}(t)V(t+1)W(t+1)Y^{*-1}(t) \\
&\qquad - tQ(t) - tY^{-1}(t)V(t)W(t)Y^{*-1}(t) + Y^{-1}(t)W(t)Y^{*-1}(t).
\end{aligned} \tag{15}
$$

Applying the following slightly different form of the product rule,

$$
\begin{aligned}
&\Delta[A(t)B(t)C(t)D(t)E(t)] \\
&\quad = A(t+1)B(t+1)C(t+1)\Delta D(t)E(t+1) \\
&\qquad + A(t+1)B(t+1)\Delta C(t)D(t)E(t+1) + A(t+1)\Delta B(t)C(t)D(t)E(t+1) \\
&\qquad + A(t+1)B(t)C(t)D(t)\Delta E(t) + \Delta A(t)B(t)C(t)D(t)E(t),
\end{aligned}
$$

we find that

$$
\begin{aligned}
&\Delta[(t-1)Y^{-1}(t-1)\Delta Y(t-1)Y^{-1}(t)Y^{*-1}(t-1)] \\
&\quad = -tY^{-1}(t)W(t+1)V(t+1)Y^{*-1}(t) - tQ(t) \\
&\qquad - tY^{-1}(t)W(t)V(t)Y^{*-1}(t) - tY^{-1}(t-1)V^{2}(t)Y^{*-1}(t-1) \\
&\qquad + Y^{-1}(t-1)V(t)Y^{*-1}(t-1).
\end{aligned} \tag{16}
$$

Combining (14)-(16), we have the identity

$$
\begin{aligned}
\Delta[Y^{-1}(t-1)Y^{*-1}(t-1)] = &-tH(t) - 2tQ(t) - \Delta[(t-1)Y^{-1}(t)W(t)Y^{*-1}(t)] \\
&- \Delta[(t-1)Y^{-1}(t-1)V(t)Y^{*-1}(t-1)],
\end{aligned} \tag{17}
$$

where

$$
\begin{aligned}
H(t) = &\, Y^{-1}(t+1)W^{2}(t+1)Y^{*-1}(t+1) + Y^{-1}(t-1)V^{2}(t)Y^{*-1}(t-1) \\
&+ Y^{-1}(t)[2V(t+1)W(t+1) + 2V(t)W(t)]Y^{*-1}(t).
\end{aligned}
$$

Summing both sides of (17) yields

$$
\begin{aligned}
2\sum_{\tau=t_0}^{t-1}\tau Q(\tau) = &-\sum_{\tau=t_0}^{t-1}\tau H(\tau) - Y^{-1}(t-1)Y^{*-1}(t-1) \\
&-(t-1)[Y^{-1}(t)W(t)Y^{*-1}(t) + Y^{-1}(t-1)V(t)Y^{*-1}(t-1)] + C,
\end{aligned} \tag{18}
$$

where C is a constant Hermitian matrix.

For $t \geqq t_0+1$, all terms except C on the right-hand side of (18) are negative definite and consequently there is a real constant K such that

$$
\lambda_{\max}\left(\sum_{\tau=t_0}^{t-1}\tau Q(\tau)\right) \leqq K \quad \text{for } t \geqq t_0+1.
$$

By Weyl's inequality, there is another constant K_1 so that

$$
\lambda_{\max}\left(\sum_{\tau=1}^{t-1}\tau Q(\tau)\right) \leqq K_1 \quad \text{for } t \geqq t_0+1. \tag{19}
$$

However,

$$\lambda_{\max}\left(\sum_{\tau=1}^{t-1}\tau Q(\tau)\right) \geqq \frac{1}{m}\operatorname{tr}\left(\sum_{\tau=1}^{t-1}\tau Q(\tau)\right)$$

$$= \frac{1}{m}\sum_{\tau=1}^{t-1}\operatorname{tr}[\tau Q(\tau)] \tag{20}$$

$$\geqq \frac{1}{m}\sum_{\tau=1}^{t-1}\lambda_{\max}[\tau Q(\tau)].$$

By (3), $\sum_{\tau=1}^{t-1}\lambda_{\max}[\tau Q(\tau)]\to\infty$ as $t\to\infty$, contradicting (19).

This proves that (3) is a sufficient condition for (2) to be oscillatory in the case $n=1$. We now prove the same in the case of a general positive integer n. Suppose to the contrary that (3) holds but there is a positive integer n such that (2) has a prepared nonoscillatory solution $Y_0(t)$. First, since $Y_0(t)$ is nonoscillatory, we choose $t_0\in\mathbb{Z}^+$ such that $Y_0(t)$ is invertible for $t\geqq t_0$ and the Riccati functions $W_0(t)$ and $V_0(t)$ are positive definite for $t\geqq t_0$. We let $Q_0(t)$ be defined by

$$Q_0(t) = Y_0^{n-1}(t)Q(t)Y_0^{*n-1}(t), \qquad t\in\mathbb{Z}^+.$$

We note that $Y_0(t)$ is also a nonoscillatory prepared solution of

$$\Delta^2 Y(t-1) + Y(t)Q_0(t)Y^*(t)Y(t) = 0, \qquad t\in\mathbb{Z}^+. \tag{21}$$

Now $Y_0^*(t)Y_0(t)$ and $Y_0(t)Y_0^*(t)$ have the same eigenvalues; furthermore, for $t\geqq t_0$ we have

$$\Delta[Y_0^*(t)Y_0(t)] = Y_0^*(t+1)\Delta Y_0(t) + Y_0^*(t)\Delta Y_0(t)$$

$$= Y_0^*(t+1)V_0(t+1)Y_0(t+1) + Y_0^*(t)W_0(t+1)Y_0(t). \tag{22}$$

From (22), we see that $\Delta[Y_0^*(t)Y_0(t)]>0$ for $t\geqq t_0$ so the eigenvalues of $Y_0(t)Y_0^*(t)$ are increasing; hence, we choose a positive real number δ so that $\lambda_{\min}[Y_0(t)Y_0^*(t)]>\delta$ for $t\geqq t_0$, $t\in\mathbb{Z}^+$. By Ostrowski's inequality,

$$\lambda_{\max}[Q_0(t)]\geqq\lambda_{\max}[Q(t)]\cdot\delta^{n-1} \quad \text{for } t\geqq t_0.$$

Hence, $Q_0(t)$ is Hermitian and positive definite for $t\geqq t_0$ with

$$\sum_{t=1}^{\infty} t\lambda_{\max}[Q_0(t)] = \infty.$$

Even though $Q_0(t)$ may only be positive semi-definite rather than positive definite for $t<t_0$ and $t\in\mathbb{Z}^+$, it is clear from the first part of the proof that (21) is oscillatory. Since $Y_0(t)$ is a nonoscillatory solution, we have reached a contradiction.

This completes the proof that (2) is oscillatory if (3) holds. Now we prove that (3) is a necessary condition if (2) is to be oscillatory.

Suppose that

$$\sum_{t=1}^{\infty} t\lambda_{\max}[Q_0(t)] < \infty. \tag{23}$$

We need to show that there is at least one nonoscillatory prepared solution of (2). Atkinson's original proof [3] in this direction can be recast as an application of the contraction mapping principle; our proof is in the same vein with some adjustment necessary because of the different matrix norms.

We begin by recalling some facts from elementary matrix analysis [10, Chap. 5]. Let $\mathcal{M}_m$ denote the set of $m \times m$ complex matrices, let $|z|$ denote the modulus of the complex number z, and let A_{ij} denote the entry in the ith row and jth column of a matrix A. Further, let $|||\cdot|||_\infty$, $|||\cdot|||_1$, and $|||\cdot|||_2$ be the matrix norms on $\mathcal{M}_m$ induced by the l_∞, l_1, and l_2 vector norms, respectively. Then $|||\cdot|||_\infty$ is the maximum row sum norm, $|||\cdot|||_1$ is the maximum column sum norm, and $|||\cdot|||_2$ is the spectral norm with $|||A|||_2 = [\lambda_{\max}(AA^*)]^{1/2}$ for $A \in \mathcal{M}_m$ and $|||H|||_2 = \lambda_{\max}(H)$ when H is Hermitian and positive semi-definite. The inequalities

$$|||A|||_\infty \leqq \sqrt{m}\, |||A|||_2, \qquad |||A|||_1 \leqq \sqrt{m}\, |||A|||_2, \quad \text{and} \quad |||A|||_1 \leqq m|||A|||_\infty$$

hold for all $A \in \mathcal{M}_m$. In addition, the submultiplicativity of matrix norms will play a key role.

If t_0 is a fixed positive integer, then a direct calculation shows that any $m \times m$ complex matrix-valued function $Y(t)$ defined for $t \geqq t_0$ and satisfying the equation

$$Y(t) = I - \sum_{s=t+1}^{\infty} (s-t)\, Y^n(s) Q(s) Y^{*n}(s) Y(s) \tag{24}$$

also satisfies (2) for $t \geqq t_0$. We will apply the contraction mapping principle to produce a solution of (24). First, we choose $t_0 \in \mathbb{Z}^+$ so large that

$$\sum_{s=t_0}^{\infty} s\lambda_{\max}[Q(s)] < m^{-5/2} 2^{-2n} (2n+1)^{-1}. \tag{25}$$

We let $\mathcal{X}_{t_0}$ denote the set of all $m \times m$ complex matrix-valued functions $U(t)$ defined for integers $t \geqq t_0$ and such that $\lim_{t\to\infty} U(t)$ exists as a finite matrix. For $U \in \mathcal{X}_{t_0}$, let

$$|||U||| = \sup_{t \geqq t_0} |||U(t)|||_\infty.$$

Then $\mathcal{X}_{t_0}$ equipped with this norm is a Banach space. Let $\mathcal{A} = \{U \in \mathcal{X}_{t_0} : |||U - I||| \leqq 1\}$. Then $\mathcal{A}$ is a closed and nonempty subset of $\mathcal{X}_{t_0}$. For $Y \in \mathcal{A}$, define TY by

$$TY(t) = I - \sum_{s=t+1}^{\infty} (s-t)\, Y^n(s) Q(s) Y^{*n}(s) Y(s). \tag{26}$$

Take $Y \in \mathcal{A}$. Then, for $s \geqq t$,

$$\begin{aligned} |[(s-t)\, Y^n(s) Q(s) Y^{*n}(s) Y(s)]_{ij}| &\leqq |||(s-t)\, Y^n(s) Q(s) Y^{*n}(s) Y(s)|||_\infty \\ &\leqq s|||Y(s)|||_\infty^{n+1} |||Y^n(s)|||_1 |||Q(s)|||_\infty \\ &\leqq sm|||Y|||^{n+1} \cdot m|||Y^n(s)|||_\infty \cdot \sqrt{m}\, |||Q(s)|||_2 \\ &\leqq 2^{2n+1} m^{3/2} s\lambda_{\max}[Q(s)]. \end{aligned} \tag{27}$$

From (27), we see that the series on the right-hand side of (26) is convergent for all $t \in \mathbb{Z}^+$. Moreover, from (25) and (26), we see that, for $t \geqq t_0$,

$$|||TY(t) - I|||_\infty \leqq 2^{2n+1} m^{5/2} \sum_{s=t_0+1}^{\infty} s\lambda_{\max}[Q(s)] < 1.$$

Therefore $|||TY - I||| \leqq 1$. Since $\lim_{t\to\infty} TY(t) = I$, we see that T is a mapping from $\mathcal{A}$ into $\mathcal{A}$.

For Y and Z in $\mathscr{A}$ and $t \geqq t_0$, we have

$$|[TY(t)-TZ(t)]_{ij}| \leqq \sum_{s=t_0+1}^{\infty} s \| Y^n(s)Q(s)Y^{*n}(s)Y(s) - Z^n(s)Q(s)Z^{*n}(s)Z(s) \|_\infty. \tag{28}$$

Shortening the notation in a self-evident way,

$$\begin{aligned} &\| Y^nQY^{*n}Y - Z^nQZ^{*n}Z \|_\infty \\ &\quad \leqq \| Y^nQY^{*n}Y - Y^nQY^{*n}Z \|_\infty \\ &\qquad + \| Y^nQY^{*n}Z - Y^nQZ^{*n}Z \|_\infty + \| Y^nQZ^{*n}Z - Z^nQZ^{*n}Z \|_\infty \\ &\quad \leqq \| Y \|_\infty^n \cdot \sqrt{m}\, \lambda_{\max}(Q) \cdot m \| Y^n \|_\infty \| Y - Z \|_\infty \\ &\qquad + \| Y^n \|_\infty \cdot \sqrt{m}\, \lambda_{\max}(Q) \cdot m \| Y^n - Z^n \|_\infty \| Z \|_\infty \\ &\qquad + \| Y^n - Z^n \|_\infty \cdot \sqrt{m}\, \lambda_{\max}(Q) \cdot m \| Z \|_\infty^{n+1}. \end{aligned} \tag{29}$$

Also,

$$\begin{aligned} \| Y^n - Z^n \|_\infty &\leqq \| Y^n - Y^{n-1}Z \|_\infty + \| Y^{n-1}Z - Y^{n-2}Z^2 \|_\infty + \cdots + \| YZ^{n-1} - Z^n \|_\infty \\ &\leqq \| Y^{n-1}(Y-Z) \|_\infty + \| Y^{n-2}(Y-Z)Z \|_\infty + \cdots + \| (Y-Z)Z^{n-1} \|_\infty \\ &\leqq n \cdot 2^{n-1} \| Y - Z \|_\infty. \end{aligned} \tag{30}$$

Combining (29) and (30) yields

$$\| Y^nQY^{*n}Y - Z^nQZ^{*n}Z \|_\infty \leqq (2n+1) \cdot 2^{2n} \cdot m^{3/2} \lambda_{\max}(Q) \| Y - Z \|_\infty. \tag{31}$$

From (28) and (31), we then find

$$\| TY - TZ \| \leqq \left[m^{5/2}(2n+1)2^{2n} \sum_{s=t_0+1}^{\infty} s \lambda_{\max}[Q(s)] \right] \| Y - Z \|.$$

Therefore, from (25), it follows that $T: \mathscr{A} \to \mathscr{A}$ is a contraction mapping.

Consequently, there is a solution $Y(t)$ of (24) which is also a solution of (2) for $t \geqq t_0$. Extending this solution backward to $t = 1$, we obtain a solution for all t satisfying (24) for $t \geqq t_0$. Since

$$\lim_{t \to \infty} Y(t) = I \quad \text{and} \quad \lim_{t \to \infty} \Delta Y(t) = 0,$$

it follows that $Y(t)$ is a prepared solution of (2). Finally

$$\lim_{t \to \infty} W(t) = \lim_{t \to \infty} \Delta Y(t-1) Y^{-1}(t-1) = 0,$$

so $W(t) + I \to I$ as $t \to \infty$, making $Y(t)$ a nonoscillatory solution of (2).

This completes the proof of Theorem 1. □

4. Extensions to more general equations. If equation (2) is altered slightly, the analysis changes somewhat. Let $H_n(t)$ denote a general nth degree product of $Y(t)$ and $Y^*(t)$; that is,

$$H_n(t) = Z_1(t)Z_2(t) \cdots Z_n(t),$$

where each $Z_i(t)$, $1 \leqq i \leqq n$, is either $Y(t)$ or $Y^*(t)$. Additionally, take $H_0(t) = I$. Consider the equations

$$\Delta^2 Y(t-1) + [Y(t)H_{n-1}(t)Q(t)H_{n-1}^*(t)Y^*(t)]Y(t) = 0, \tag{32}$$

$$\Delta^2 Y(t-1) + [Y^*(t)H_{n-1}(t)Q(t)H_{n-1}^*(t)Y(t)]Y(t) = 0. \tag{33}$$

Then (2) is a special case of (32) with $Z_i(t) = Y(t)$ for $1 \leqq i \leqq n-1$ and

$$\Delta^2 Y(t-1) + Y^{*n}(t) Q(t) Y^{n+1}(t) = 0$$

is a special case of (33) with $Z_i(t) = Y^*(t)$ for $1 \leqq i \leqq n-1$.

The definition of a prepared solution, the Riccati functions $W(t)$ and $V(t)$, and the definition of generalized zeros is the same as above. Following a development along the lines of that in [2] for matrix differential equations, we are led to Theorem 2 below. Since all the ideas that set the difference equations apart from the differential equations are already presented in the development of Theorem 1, we omit the proof.

THEOREM 2. *Suppose $Q(t)$ is Hermitian and positive definite for all $t \in \mathbb{Z}^+$. Then equation* (32) *is oscillatory if and only if* (3) *holds. Equation* (33) *is oscillatory provided that*

$$\sum_{t=1}^{\infty} t\lambda_{\min}[Q(t)] = \infty$$

and is nonoscillatory when

$$\sum_{t=1}^{\infty} t\lambda_{\max}[Q(t)] < \infty.$$

REFERENCES

[1] C. D. AHLBRANDT AND J. W. HOOKER, *Recessive solutions of symmetric three term recurrence relations*, in Oscillation, Bifurcation and Chaos, Canadian Mathematical Society Conference Proc., Ottawa, Ont., Vol. 8, 1987, pp. 3-42.

[2] C. D. AHLBRANDT, J. RIDENHOUR, AND R. C. THOMPSON, *Oscillation of superlinear matrix differential equations*, Proc. Amer. Math. Soc., 105 (1989), pp. 141-148.

[3] F. V. ATKINSON, *On second-order non-linear oscillations*, Pacific J. Math., 5 (1955), pp. 643-647.

[4] G. J. BUTLER AND L. H. ERBE, *Oscillation theory for second order differential systems with functionally commutative matrix coefficients*, Funkcial. Ekvac., 28 (1985), pp. 47-55.

[5] G. J. BUTLER, L. H. ERBE, AND A. B. MINGARELLI, *Riccati techniques and variational principles in oscillation theory for linear systems*, Trans. Amer. Math. Soc., 302 (1987), pp. 263-282.

[6] R. BYERS, B. J. HARRIS, AND M. K. KWONG, *Weighted means and oscillation conditions for second order matrix differential equations*, J. Differential Equations, 61 (1986), pp. 164-177.

[7] S. GOFF AND D. F. ST. MARY, *The Bohl transformation and oscillation of linear differential systems*, SIAM J. Math. Anal., 20 (1989), pp. 215-221.

[8] P. HARTMAN, *Difference equations: disconjugacy, principal solutions, Green's functions, complete monotonicity*, Trans. Amer. Math. Soc., 246 (1978), pp. 1-30.

[9] J. W. HOOKER AND W. T. PATULA, *Riccati type transformations for second-order linear difference equations*, J. Math. Anal. Appl., 82 (1981), pp. 451-462.

[10] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[11] M. K. KWONG, J. W. HOOKER, AND W. T. PATULA, *Riccati type transformations for second-order difference equations*, J. Math. Anal. Appl., 107 (1985), pp. 182-196.

[12] T. KURA, *A matrix analogue of Atkinson's oscillation theorem*, Funkcial. Ekvac., 25 (1982), pp. 223-226.

[13] A. B. MINGARELLI, *Volterra-Stieltjes integral equations and generalized ordinary differential expressions*, Lecture Notes in Math. 989, Springer-Verlag, Berlin, 1983.

[14] W. T. PATULA, *Growth and oscillation properties of second-order linear difference equations*, SIAM J. Math. Anal., 10 (1979), pp. 55-61.

[15] ———, *Growth, oscillation and comparison theorems for second-order linear difference equations*, SIAM J. Math. Anal., 10 (1979), pp. 1272-1279.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 785–791, May 1991

013

AN H_0^m INTERPOLATION RESULT*

S. JENSEN†

Abstract. This paper presents a proof of an interpolation result related to the approximation theory for higher-order finite element or spectral methods when C^1 (or higher) regularity is convenient for the finite-dimensional subspaces. This can be a natural choice, for example, for the Stokes problem, the biharmonic problem, or higher-order plate and shell models. It is shown that the same intermediate spaces are obtained whether one (1) interpolates between two Sobolev spaces defined on a domain with nonsmooth boundary first and then enforces the homogeneous boundary conditions afterwards or (2) interpolates between two Sobolev spaces where the homogeneous boundary conditions are enforced throughout the interpolation process.

Key words. interpolation, Peetre, boundary conditions, nonsmooth domains, small angle elliptic regularity

AMS(MOS) subject classifications. 65N30, 46E35, 35J40, 35B65

1. Introduction. The aim of this note is to prove an interpolation result for domains in $\mathbf{R}^2$ with finitely many corners and otherwise smooth boundary. We consider a bounded open set Ω of $\mathbf{R}^2$, whose boundary is a curvilinear polygon of class C^∞ (see [6]). We denote each of the C^∞ curves which constitute the boundary by $\overline{\Gamma_j}$ for some j ranging from 1 to N. The curve $\overline{\Gamma_{j+1}}$ follows $\overline{\Gamma_j}$ according to the positive orientation, on each connected component of Γ. We denote by C_j the vertex which is the end point of $\overline{\Gamma_j}$ and by α_j the measure of the angle at C_j (toward the interior of Ω). By a corner we mean a vertex C_j with an angle α_j not in the set $\{0, \pi, 2\pi\}$. The result is an extension to H_0^m of the one in [2], [1] for H_0^1 which would be useful in approximation theory for Sobolev spaces; see [8, Remark 2.2.9], [17, Remark 4.2], [7], and [14, the line following (III.26) in the proof of Thm. III.2]. For example, consider solving the Stokes problem via the p version of the finite element method or a polynomial spectral method. Then the discrete velocity $\vec{U}_p$ is an elliptic projection onto a finite-dimensional subspace Z_p of $Z = [H_0^1(\Omega)]^2 \cap \mathrm{Ker}(\mathrm{div})$ centering interest on the approximation problem. Introducing stream functions ($\vec{U} = \mathrm{rot}\phi, \vec{U}_p = \mathrm{rot}\phi_p$) will translate this approximation problem to $H_0^2(\Omega)$. Now an energy estimate is obtained for free — $\|\phi - \phi_p\|_2$ bounded when $\phi \in H_0^2$ only — and there exist constructive approximation estimates, $\|\phi - \phi_p\|_2 \leq Cp^{2-t}\|\phi\|_t$, when $\phi \in H^t(\Omega) \cap H_0^2(\Omega)$ for $t > 7/2$; see [17]. Now, one wishes to interpolate between these spaces and hopes to get spaces that coincide in some sense with the ones predicted by regularity theory, but the trace constraints on a nonsmooth boundary makes this identification nontrivial. If the boundary is smooth, the result can be deduced from [13]. In general such an identification is useful for higher-order finite element or spectral methods when C^1 (or higher) regularity is convenient for the finite-dimensional subspaces. This can be a natural choice, for example, also for the biharmonic problem or higher-order plate and shell models.

Let $H^s(\Omega)$ be the standard Sobolev space of order s based on L_2 with corresponding norm $\|\cdot\|_s$. For $m \in \mathbf{Z}_+$, $H_0^m(\Omega)$ is the set of functions in $H^m(\Omega)$ for which the traces of the function and its normal derivatives up to order $m-1$ vanish on $\partial\Omega$.

* Received by the editors November 20, 1989; accepted for publication (in revised form) May 14, 1990.

† Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland 21228. This work was supported in part by Office of Naval Research contract N00014-87-K-0427.

We shall use the interpolation spaces of Peetre (see, e.g., [3]) in the cases $1 \leq q \leq \infty$ and $0 < \theta < 1$ where we define $[H^t(\Omega) \cap H_0^m(\Omega), H_0^m(\Omega)]_{\theta,q}$ explicitly: For $u \in H^t \cap H_0^m$, we set

$$K(u,s) = \inf_{\substack{u = v + w \\ v \in H_0^m,\ w \in H^t \cap H_0^m}} (\|v\|_m + s\|w\|_t)$$

and we define the norm

$$\|u\|_{[\cdot,\cdot]_{\theta,q}} = \|s^{-1/q-\theta} K(u,s)\|_{L_q(0,\infty)}.$$

Then

$$[H^t(\Omega) \cap H_0^m(\Omega), H_0^m(\Omega)]_{\theta,q} = \{u \in H_0^m(\Omega) : \|u\|_{[\cdot,\cdot]_{\theta,q}} < \infty\}.$$

$[H^t(\Omega), H^m(\Omega)]_{\theta,q}$ is defined similarly. Note that this space will be a Sobolev space if we choose $q = 2$ and in general a Besov space.

2. The interpolation result. We state and prove the following proposition.

PROPOSITION 1. *Let $\Omega \subseteq \mathbf{R}^2$ be piecewise C^∞ with finitely many corners of angles in $(0, 2\pi) \setminus \{\pi\}$. Then the following identity holds for all $\theta \in (0,1)$, $1 \leq q \leq \infty$ and $t \geq m$, $t \notin m + \{\frac{1}{2}, \ldots, m - \frac{1}{2}\}$, $m \in \mathbf{Z}_+$:*

$$[H^t(\Omega) \cap H_0^m(\Omega), H_0^m(\Omega)]_{\theta,q} = [H^t(\Omega), H^m(\Omega)]_{\theta,q} \cap H_0^m(\Omega).$$

Proof. We follow the main ideas of [2] but have weights be unity for simplicity; see also [18] and [1].

The inclusion from left to right follows directly from the definition.

The proof of the reverse inclusion can through a partition of unity be reduced to considering a domain Ω with one corner of angle $\alpha \in (0, 2\pi) \setminus \{\pi\}$. We shall distinguish between two cases: whether $\alpha \in (0, \pi)$ or $(\pi, 2\pi)$.

Case 1. $\alpha \in (0, \pi)$. Then there exists a linear transformation from Ω to $\widetilde{\Omega}$ with a corner of angle $\tilde{\alpha} < \min\{\omega_0, \pi\}$ where ω_0 will be introduced in the next section as a sufficiently small angle that a certain shift theorem will hold. Let L be the associated map of functions defined on $\widetilde{\Omega}$ to functions defined on Ω. If $w \in H^t(\widetilde{\Omega})$, then we let $\tilde{v} = \Pi_{\widetilde{\Omega}} w$ denote the solution to

$$(2.1) \qquad \begin{aligned} (-\triangle)^m \tilde{v} + \tilde{v} &= (-\triangle)^m w + w \text{ in } \widetilde{\Omega}, \\ \tilde{v} &\in H_0^m(\widetilde{\Omega}). \end{aligned}$$

Thus $\Pi_{\widetilde{\Omega}}$ is a projection from $H^m(\widetilde{\Omega})$ to $H_0^m(\widetilde{\Omega})$. As proven in the next section on regularity, there exists ω_0, dependent on m and t, such that the following shift theorem holds provided $\tilde{\alpha} < \omega_0$:

$$\|\tilde{v}\|_{H^t} \leq C\|(-\triangle)^m w + w\|_{H^{t-2m}}.$$

In particular, $P_\Omega = L \circ \Pi_{\widetilde{\Omega}} \circ L^{-1} \in \mathcal{B}(H^m(\Omega), H_0^m(\Omega))$ and $P_\Omega \in \mathcal{B}(H^t(\Omega), H^t(\Omega) \cap H_0^m(\Omega))$. Thus, by interpolation,

$$P_\Omega \in \mathcal{B}([H^t(\Omega), H^m(\Omega)]_{\theta,q}, [H^t(\Omega) \cap H_0^m(\Omega), H_0^m(\Omega)]_{\theta,q}).$$

Since $P_\Omega|_{H_0^m(\Omega)} = I$ (the identity),

$$[H^t(\Omega), H^m(\Omega)]_{\theta,q} \cap H_0^m(\Omega) \subseteq [H^t(\Omega) \cap H_0^m(\Omega), H_0^m(\Omega)]_{\theta,q}. \tag{2.2}$$

Case 2. $\alpha \in (\pi, 2\pi)$. Let B be a ball centered at the corner and containing Ω. By E_Ω, we denote the Stein extension [16, Chap. VI, §3] of functions on Ω to functions on B vanishing at ∂B. Let $\Omega^c = B \setminus \Omega$ and let E_{Ω^c} be the Stein extension of functions on Ω^c to all of B. Theorem 5 in [16] states that $E_\Omega \in \mathcal{B}(H^k(\Omega), H^k(B))$ and $E_{\Omega^c} \in \mathcal{B}(H^k(\Omega^c), H^k(B))$, for all $k \in \mathbf{N}$. Now define

$$P_\Omega = E_{\Omega^c} \circ P_{\Omega^c} \circ E_\Omega + (I - E_{\Omega^c} \circ E_\Omega)$$

with P_{Ω^c} being the the same operator as P_Ω was in Case 1. Then $P_\Omega|_{H_0^m(\Omega)} = I$ and

$$P_\Omega \in \mathcal{B}([H^t(\Omega), H^m(\Omega)]_{\theta,q}, [H^t(\Omega) \cap H_0^m(\Omega), H_0^m(\Omega)]_{\theta,q})$$

which ends the proof of the proposition. □

Remark 1. We have explicitly excluded vertices of angles $0, \pi$, or 2π. In these cases it is not possible to map linearly onto a domain of sufficiently small angle. In case Ω is a polygon the exclusion only amounts to 2π.

Remark 2. The theorem and proof hold in $\mathbf{R}^3$ for conical points with smooth cross section almost verbatim.

3. Regularity for small angles. In [12] it is stated that, given $k \in \mathbf{N}$, if the domain contains only corners of sufficiently small angles and $f \in H_0^k$, then the solution (u) of a Dirichlet problem with zero boundary conditions and a $2m$ order, elliptic operator $(Lu = f, u \in H_0^m)$ belongs to H^{k+2m}. We present a proof here following and expanding upon the ideas in [12, pp. 292–294] and [4] extending to the case where $f \in H^k$, $k \geq -m$. We use the notation of [12].

Let L be strongly elliptic of order $2m$ with C^∞ coefficients and $u \in H_0^m$ be the solution of

$$Lu = f$$

in a plane sector with opening $\tilde{\alpha} > 0$ and the corner translated to the origin (a case of C^∞ sides may be reduced to this by a C^∞ diffeomorphism). In [12] a technique is used that involves a combination of (1) looking at $\mathcal{L}(0, \partial/\partial x)$: the principal part of the operator $L(x, \partial/\partial x)$ with coefficients fixed at the origin, (2) changing to polar variables (r, ω), so that $\mathcal{L}u = f$ takes the form:

$$\sum_{0 \leq i_2 \leq i_1 \leq 2m} \frac{a_{i_1 i_2}(\omega)}{r^{i_1}} \frac{\partial^{2m-i_2} u}{\partial r^{2m-i_1} \partial \omega^{i_1-i_2}} = f,$$

(3) making the change of the radial variable $(\varrho = \ln 1/r)$ so that $\mathcal{L}u = f$ now takes the form

$$\sum_{1 \leq k_1+k_2 \leq 2m, 0 \leq k_1, k_2} \breve{a}_{k_1 k_2}(\omega) \frac{\partial^{k_1+k_2} u}{\partial \varrho^{k_1} \partial \omega^{k_2}} = f \cdot e^{-2m\varrho} = F,$$

and (4) taking the Fourier transform with respect to the "radial" variable (ϱ). The domain then consists of angles $\omega \in \widetilde{D}$ — an interval on S^1 (for $\mathbf{R}^2$ — in dimension n a domain on S^{n-1} with smooth boundary). The final form of $\mathcal{L}u = f$ is

$$L_0\left(\omega, i\lambda, \frac{\partial}{\partial \omega}\right)\hat{u} = \hat{F}.$$

The boundary conditions undergo similar transformations. Steps 3 and 4 constitute a partial Mellin transform with respect to the r variable. Steps 2 through 4 are sometimes called the Kondrat'ev transform [9]. Let $R(\lambda)$ be the resolvent operator — a meromorphic function of λ — associated with the resulting boundary value problem. In [12], [11], and [4] it is shown that if $f \in H^k$, then $u \in H^m$ has the expansion

$$u = \sum_{h_0 < \mathrm{Im}\lambda_j < h} \sum_{s=0}^{n_j} \alpha_{js} r^{-i\lambda_j} \log^s r \sum_{q=0}^{s} P_{jsq}(r \log^q r) + w \tag{3.1}$$

where $h_0 = -1 + m$, $h = -1 + k + 2m$, λ_j are the poles of $R(\lambda)$ of multiplicity n_j, P_{jsq} are polynomials of degree $[h - \mathrm{Im}\lambda_j]$, whose coefficients are C^∞ functions of ω, and $w \in H^{k+2m}$. From this expansion we see that the smoothness of u depends on the poles λ_j of the function $R(\lambda)$ which lie above the straight line $\mathrm{Im}\lambda = -1 + m$. We will show that the following lemma holds.

LEMMA 1. *Given any positive h, there exists ω_0 such that, if the angle of the corner is smaller than ω_0, then the strip $-1 + m < \mathrm{Im}\lambda < h$ contains no poles of $R(\lambda)$.*

Proof. Let h be given and λ_0 be a pole of $R(\lambda)$ lying in the strip $-1+m < \mathrm{Im}\lambda < h$. When $\lambda = \lambda_0$, there exists a nonzero solution $u_0(\omega)$ to

$$\begin{aligned} L_0\left(\omega, i\lambda, \frac{\partial}{\partial\omega}\right) u_0 &= 0 \quad \text{in } \widetilde{D}, \\ u_0 = \frac{\partial u_0}{\partial\omega} = \cdots = \frac{\partial^{m-1} u_0}{\partial\omega^{m-1}} &= 0 \quad \text{on } \partial\widetilde{D}. \end{aligned}$$

Now, $L_0 u_0 = L_1 u_0 + \lambda L_2 u_0$, where the operator L_2 contains derivatives of order less than $2m$. Since this system is strongly elliptic for all real λ, it is strongly elliptic for $\lambda = 0$. So L_1 is strongly elliptic. Let

$$I(u) = \int_{\widetilde{D}} (L_0 u)\overline{u}\, d\omega$$

which at u_0 is zero: integrate by parts

$$\begin{aligned} &I_1(u_0) + I_2(u_0) + I_3(u_0) \\ &\quad = \int_{\widetilde{D}} \Bigg\{ \Bigg[\tilde{a}_{mm}(\omega) \frac{\partial^m u_0}{\partial\omega^m} \frac{\partial^m \overline{u_0}}{\partial\omega^m} \\ &\qquad + \sum_{0<i+j<2m, 0\le i,j\le m} \lambda^{i+j} \tilde{a}_{m-i,m-j}(\omega) \frac{\partial^{m-i} u_0}{\partial\omega^{m-i}} \frac{\partial^{m-j}\overline{u_0}}{\partial\omega^{m-j}} \Bigg] \\ &\qquad + \lambda^{2m} \tilde{a}_{00}(\omega) u_0 \overline{u_0} \Bigg\}\, d\omega \\ &\quad = 0. \end{aligned}$$

By strong ellipticity of L_1,

$$|\mathrm{Re} I_1(u_0)| \ge \gamma_0 \int_{\widetilde{D}} \left| \frac{\partial^m u_0}{\partial\omega^m} \right|^2 d\omega - C_0 \int_{\widetilde{D}} |u_0|^2 d\omega$$

where γ_0 and C_0 do not depend on u_0 or σ — the diameter of $\widetilde{D}$ $(= \tilde{\alpha})$.

$$|I_2(u_0)| \le \epsilon \int_{\widetilde{D}} \left| \frac{\partial^m u_0}{\partial\omega^m} \right|^2 d\omega + C(\epsilon)|\lambda|^{2(m-1)} \int_{\widetilde{D}} |u_0|^2 d\omega$$

by intermediate derivative inequalities as sketched in [12] and

$$|I_3(u_0)| \le C|\lambda|^{2m} \int_{\widetilde{D}} |u_0|^2 d\omega$$

where $C > 0$ does not depend on σ. Also

$$\int_{\widetilde{D}} \left| \frac{\partial^m u_0}{\partial \omega^m} \right|^2 d\omega \ge \frac{\gamma}{\sigma^{2m}} \int_{\widetilde{D}} |u_0|^2 \, d\omega.$$

Upon contracting like terms, substituting this last inequality, and cancelling $\int_{\widetilde{D}} |u_0|^2 d\omega$, we get from $|\mathrm{Re} I_1| \le |I_2| + |I_3|$

$$\gamma_1 \sigma^{-2m} \le 1 + |\lambda|^{2(m-1)} + |\lambda|^{2m}$$

for some $\gamma_1 > 0$ independent of σ. Since $\mathrm{Im}\lambda \in (-1+m, h)$, if $\lambda = \rho e^{i\theta}$, $\theta \in [-\pi/2, 3\pi/2)$, given any $\epsilon > 0$, there exists σ sufficiently small so that

$$\theta \in \left(-\frac{\epsilon}{2m}, \frac{\epsilon}{2m} \right) \cup \left(\pi - \frac{\epsilon}{2m}, \pi + \frac{\epsilon}{2m} \right)$$

and thus $2m\theta \in (-\epsilon, \epsilon) \cup (2m\pi - \epsilon, 2m\pi + \epsilon) = (-\epsilon, \epsilon)$ such that $\mathrm{Re}(\lambda^{2m}) > 0$. Thus $|\mathrm{Re}\ \lambda^{2m}|/|\lambda|^{2m} = \cos(2m\theta) \ge \cos\epsilon \ge \frac{1}{2}$ for $\epsilon < \pi/3$ so that for $\gamma_2 > 0$ independent of σ and u_0,

$$|\mathrm{Re} I_3| \ge \frac{\gamma_2}{2} |\lambda|^{2m} \int_{\widetilde{D}} |u_0|^2 d\omega.$$

If we again contract, substitute, and cancel as before, but now in $|\mathrm{Re}I_1 + \mathrm{Re}I_3| \le |I_2|$ (using that $\tilde{a}_{00}$ and $\tilde{a}_{mm}$ are of the same sign and $\mathrm{Re}(\lambda^{2m}) > 0$), we get with positive C_1, C_2

$$\sigma^{-2m} + C_1|\lambda|^{2m} \le C_2|\lambda|^{2(m-1)}$$

admitting no solutions λ for sufficiently small σ. □

Remark 3. Another way of proving this lemma for the biharmonic operator is by checking that the angle $\tilde{\alpha}$ can be chosen such that the roots of the equation

$$\sinh^2(\tau\omega) = \tau^2 \sin^2 \omega$$

see [6], (7,2,2,1), except for $-i$ and 0 all have sufficiently small (negative, large absolute value) imaginary value; cf. [5]. Note $1 + i\tau = -i\lambda$.

We finally employ a recent regularity theorem in [4].

LEMMA 2. *Assume that* $k \ge -m$, $k \notin -m + \{\frac{1}{2}, \frac{3}{2}, \cdots, m - \frac{1}{2}\}$. *Let* L *be a strongly elliptic partial differential operator of order* $2m$ *with* C^∞ *coefficients and let* $f \in H^k(\widetilde{\Omega})$ ($\widetilde{\Omega}$ *is defined above*). *Then, for sufficiently small* $\tilde{\alpha}$, *the solution* $u \in H_0^m(\widetilde{\Omega})$ *to*

$$Lu = f \text{ in } \widetilde{\Omega} \tag{3.2}$$

belongs to $H^{k+2m}(\widetilde{\Omega})$ *and* $\|u\|_{H^{k+2m}} \le C\|f\|_{H^k}$.

Proof. The statement is really a corollary of Lemma 1 and a recent shift theorem in [4, Cor. (5.2)] proven in [4, §10]. It is stated for a cone in Theorem (1.11). In order

to apply this result we select the angle $\tilde{\alpha}$ sufficiently small that Lemma 1 ensures that $R(\lambda)$ has no poles in the strip $\mathrm{Im}\lambda \in [m-1, t-2]$. This in turn implies Dauge's condition (C2*) ($\sim$ (R2)) as follows: If $-i\lambda \in \mathbf{N}$, then [4, Corollary (4.6′)] yields (C2*) and if $-i\lambda \notin \mathbf{N}$, then Corollary (4.9) along with Corollary (4.15) and the fact that $a = 2$ (see [4, p. 39]) concludes the proof. □

For the interpolation result we use this lemma with $L = (-\triangle)^m + I$.

Remark 4. For the H_0^2 interpolation to hold, it would have sufficed to quote [6, Theorem 7.2.2.3]. The H_0^2 interpolation result is thus essentially — along with the reasoning of the proof of the proposition and a localization of the poles of $R(\lambda)$ for the biharmonic — a consequence of the analysis in Grisvard's monograph [6]. Such an analysis was first done in [10] and [15] for the Stokes problem (which via the stream function connects to the biharmonic problem).

Remark 5. For $m > 2$ it is possible to prove Lemma 2 directly from [12] when $k \geq -1$, $k \in \mathbf{Z}$; [12] has the result for $f \in H_0^k$ and $k \in \mathbf{N}$. Then it is possible to prove for $f \in H^k$ when $k \geq -1$ by generalizing the trace Theorem 7.2.2.3 in [6] for a domain $\widetilde{\Omega}$ with one corner C of sufficiently small angle $\tilde{\alpha}$ between the two linear pieces Γ_j, $j = 1, 2$. First, by this, one finds $v \in H^{k+2m}(\widetilde{\Omega}) \cap H_0^m(\widetilde{\Omega})$ such that

$$(-\triangle)^m v + v - f \in H_0^k(\widetilde{\Omega}). \tag{3.3}$$

Then one applies to $w = u - v$ Kondrat'evs result (when $f \in H_0^k$) with Lemma 1 choosing $\tilde{\alpha}$ sufficiently small that $w \in H^{k+2m}$ where $v \in H^{k+2m}$ is the solution to (3.3). In [6] the generalization of Kondrat'ev's weighted spaces is given for $k = -1$. It seems difficult, however, to go to all the remaining negative integers.

Acknowledgments. The author wishes to expressly thank a number of people at the University of Maryland, College Park for some very helpful discussions: Michael Vogelius, Ivo Babuška, Bruce Kellogg, John Osborn, and Tobias von Petersdorff. Also thanks to the Institute for Physical Science and Technology, College Park, for giving me the opportunity for such discussions.

REFERENCES

[1] I. Babuška and M. Dorr, *Error estimates for the combined h and p versions of the finite element method*, Numer. Math., 37 (1981), pp. 257–277.

[2] I. Babuška, R. Kellogg, and J. Pitkäranta, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math., 33 (1979), pp. 447–471.

[3] J. Bergh and J. Löfstrom, *Interpolation Spaces*, Springer, New York, 1976.

[4] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains*, Lecture Notes in Math. 1341, Springer, New York, 1988.

[5] ———, *Stationary Stokes and Navier-Stokes systems on two- or three- dimensional domains with corners. Part* I: *Linearized equations*, SIAM J. Math. Anal., 20 (1989), pp. 74–97.

[6] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 24, Pitman, Boston, 1985.

[7] S. Jensen, *On computing the pressure by the p-version of the finite element method for Stokes problem.* Tech. Note, University of Maryland Baltimore County, Baltimore, MD, 1990.

[8] I. N. Katz and D. W. Wang, *The p-version of the finite element method for problems requiring C^1-continuity*, SIAM J. Numer. Anal., 22 (1985), pp. 1082-1106.

[9] R. Kellogg, *Notes on piecewise smooth elliptic boundary value problems.* private communication, 1989.

[10] R. Kellogg and J. Osborn, *A regularity result for the Stokes problem in a convex polygon*, J. Funct. Anal., 21 (1976), pp. 397–431.

[11] V. A. Kondrat'ev, *Boundary problems for parabolic equations in closed domains*, Trans. Moscow Math. Soc., 15 (1966), pp. 450–504.

[12] ———, *Boundary problems for elliptic equations in domains with conical or angular points*, Trans. Moscow Math. Soc., 16 (1967), pp. 227–313.

[13] J. L. LIONS AND E. MAGENES, *Non-homogeneous boundary value problems and applications*, 1, Grundlehren 181, Springer, New York, 1972.

[14] Y. MADAY, *Analysis of spectral projections in multi-dimensional domains.* private communication, May 1989.

[15] J. OSBORN, *Regularity of solutions of the Stokes problem in a polygonal domain*, in Numerical Solution of Partial Differential Equations - III, B. Hubbard, ed., Academic Press, New York, 1976, pp. 393–411.

[16] E. M. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton Mathematical Series 30, Princeton University Press, NJ, 1970.

[17] M. SURI, *The p-version of the finite element method for elliptic equations of order 2l*, Modél. Math. Anal. Numér., 24 (1990), pp. 265–304.

[18] M. VOGELIUS AND G. PAPANICOLAOU, *A projection method applied to diffusion in a perodic structure*, SIAM J. Appl. Math., 42 (1982), pp. 1302–1322.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 792-795, May 1991

014

AN EXTREMAL PROBLEM CONCERNING A MARKOV-TYPE INEQUALITY*

P. DÖRFLER†

Abstract. For any polynomial f with complex coefficients let $\|f\|$ be the norm in $L^2[0,\infty)$ with the Laguerre weight function $w(t)=e^{-t}$. Let P_n be the set of all complex polynomials whose degree does not exceed n and $\gamma_n := \sup_{f\in P_n}(\|f'\|/\|f\|)$. We show that $\gamma_n/n \to 2/\pi$ as $n\to\infty$.

Key words. Markov inequality, L^2 norm, Laguerre weight

AMS(MOS) subject classifications. 33A65, 41A17, 41A44

1. For any polynomial f with complex coefficients we define the norm

$$\|f\| := \left\{\int_0^\infty |f(t)|^2 e^{-t}\,dt\right\}^{1/2}.$$

Let P_n denote the set of all complex polynomials whose degree does not exceed n and consider

$$\gamma_n := \sup_{f\in P_n}\frac{\|f'\|}{\|f\|}, \qquad n\in\mathbb{N}. \tag{1}$$

In [4] Schmidt obtained estimates for γ_n that are asymptotically sharp. Some years later, Turán [5] found the exact value of γ_n:

$$\gamma_n = \left(2\sin\frac{\pi}{4n+2}\right)^{-1}, \qquad n\in\mathbb{N}.$$

In the present paper we show the convergence of γ_n/n as $n\to\infty$ and determine the limit, which turns out to be $2/\pi$. This result is an immediate consequence of the above-mentioned results obtained by Schmidt and Turán. Now, it is the purpose of this paper to present a quite different approach to this problem by using a new method. This method is based on results developed in [1] and on some function-theoretic considerations derived from [2].

If P_n is restricted to certain smaller classes of polynomials, there exist several results concerning γ_n [6], [7]. For the class of polynomials with nonnegative coefficients, Milovanović [3] computed the exact value of γ_n even for the generalized Laguerre weight.

2. In [1] Dörfler proved that γ_n is the largest singular value of the $n\times(n+1)$ matrix

$$A_n^{(1)} = \begin{bmatrix} 0 & -1 & \cdots & -1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 \end{bmatrix},$$

i.e., the square root of the largest eigenvalue of $(A_n^{(1)})^t A_n^{(1)}$.

* Received by the editors October 10, 1988; accepted for publication (in revised form) April 13, 1990.
† Institut für Mathematik und Angewandte Geometrie, Montanuniversität Leoben, A-8700 Leoben, Austria.

Since all the elements in the first row and the first column of $(A_n^{(1)})^t A_n^{(1)}$ are zero, it suffices to consider the eigenvalues of the $n \times n$ matrix

$$B_n := \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 2 \\ \vdots & \vdots & & \vdots \\ 1 & 2 & \cdots & n \end{bmatrix}, \qquad n \in \mathbb{N}.$$

First we derive a recurrence formula for the characteristic polynomial of B_n

$$Q_n(t) := \det(B_n - tI_n), \qquad n \in \mathbb{N}, \tag{2}$$

where I_n denotes the $n \times n$ identity matrix.

LEMMA 1. $Q_{n+1} = (1-2t)Q_n - t^2 Q_{n-1}$, $n \geqq 2$.

Proof. Consider $B_{n+1} - tI_{n+1}$. If we multiply the nth row by (-1) and then add it to the last row and, afterwards, multiply the nth column by (-1) and add it to the last column, we obtain

$$Q_{n+1} = \det \begin{bmatrix} 1-t & 1 & \cdots & 1 & 0 \\ 1 & 2-t & \cdots & 2 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & 2 & \cdots & n-t & t \\ 0 & 0 & \cdots & t & 1-2t \end{bmatrix}.$$

By expanding this determinant by the last row and expanding one of the subdeterminants by its last column, the above assertion follows instantly. □

With the aid of Lemma 1 we can compute Q_n explicitly.

LEMMA 2.

$$Q_n(t) = \sum_{k=0}^{n} \binom{n+k}{n-k} (-t)^{n-k}, \qquad n \in \mathbb{N}.$$

Proof. We use induction on n. By definition (2), $Q_1 = 1-t$ and $Q_2 = 1-3t+t^2$, so that the assertion is obviously true for $n=1$ and $n=2$. Suppose now that the assertion is also true for all n smaller than or equal to a fixed $N \geqq 2$. By substituting for Q_N and Q_{N-1} in the recurrence formula provided by Lemma 1, a lengthy but straightforward computation gives the desired result for Q_{N+1}. □

Remark. γ_n^2 is the largest root of Q_n. Although we know Q_n explicitly, it is difficult to determine γ_n from it. Its asymptotic behavior, however, can be studied by using function-theoretic methods.

3. First of all we introduce the rational functions

$$R_n(z) := \frac{Q_n(n^2 z)}{(-n^2 z)^n}, \qquad n \in \mathbb{N}, \tag{3}$$

which are well defined and analytic for all $z \in \mathbb{C} \setminus \{0\}$. By Lemma 2 we have

$$R_n(z) = \sum_{k=0}^{n} n^{-2k} \binom{n+k}{n-k} (-z)^{-k}, \qquad n \in \mathbb{N}. \tag{4}$$

LEMMA 3. *Let $n \in \mathbb{N}$ be arbitrary but fixed and let $n^2 w = z$. Then*

$$R_n(w) = 0 \Leftrightarrow Q_n(z) = 0.$$

Proof. This is clear by definition (3) and the fact that $Q_n(0) \neq 0$ for all $n \in \mathbb{N}$. □

LEMMA 4. *Let $G \subseteq \mathbb{C}$ be a region, $0 \notin G$. Then*

$$\lim_{n\to\infty} R_n(z) = \cos(z^{-1/2})$$

in G and this convergence is uniform on compact subsets of G.

Proof. (a) Let $k \in \mathbb{N} \cup \{0\}$ be fixed. Then

$$\lim_{n\to\infty} n^{-2k}\binom{n+k}{n-k} = \frac{1}{(2k)!}. \tag{5}$$

For $k = 0$ this is obviously true. Thus consider $k > 0$. If $n < k$, then $\binom{n+k}{n-k} = 0$ by definition. For $n \geqq k$ we have

$$n^{-2k}\binom{n+k}{n-k} = \frac{n+k}{n}\,\frac{n+k-1}{n}\cdots\frac{n-k+1}{n}\,\frac{1}{(2k)!}, \tag{6}$$

which implies (5) because k is fixed. Since

$$\cos(z^{-1/2}) = \sum_{k=0}^{\infty} \frac{1}{(2k)!}(-z)^{-k},$$

it follows from (4) and (5) that $R_n(z)$ converges to $\cos(z^{-1/2})$ in G as $n \to \infty$.

(b) Let $k \in \mathbb{N} \cup \{0\}$ and $n \in \mathbb{N}$ be arbitrary. Then

$$n^{-2k}\binom{n+k}{n-k} \leqq \frac{1}{k!}. \tag{7}$$

For $k = 0$ and $n < k$, this is evidently true. For $0 < k \leqq n$, equation (6) holds, from which we derive

$$n^{-2k}\binom{n+k}{n-k} = \left[1+\frac{k}{n}\right]\left[1-\left(\frac{k-1}{n}\right)^2\right]\cdots\left[1-\left(\frac{1}{n}\right)^2\right]\frac{1}{(2k)!},$$

and finally (7) follows.

Let $K \subseteq G$ be compact. Then by (4) and (7)

$$|R_n(z)| \leqq \sum_{k=0}^{\infty} \frac{1}{k!}|z|^{-k} \leqq \max_{z\in K} e^{1/|z|} =: C_K < \infty$$

for all $z \in K$ and all $n \in \mathbb{N}$, where C_K is a constant depending on K only. Hence, $\{R_n(z)\}$ is locally bounded and, by Vitali's theorem, the convergence of $R_n(z)$, shown in (a), is uniform on compact subsets of G. □

Now we are ready to state and prove our theorem.

THEOREM. *Let γ_n be defined as in* (1). *Then*

$$\lim_{n\to\infty} \frac{\gamma_n}{n} = \frac{2}{\pi}.$$

Proof. Let $z_{1n}, \cdots, z_{nn}$ be the roots of Q_n, the characteristic polynomial of B_n. Since $(A_n^{(1)})^t A_n^{(1)}$ is positive semidefinite, its eigenvalues are real and nonnegative. Hence, because $Q_n(0) \neq 0$, for all $n \in \mathbb{N}$ we can introduce the ordering $z_{1n} \geqq z_{2n} \geqq \cdots \geqq z_{nn} > 0$. Let, by Lemma 3, $w_{1n} \geqq w_{2n} \geqq \cdots \geqq w_{nn} > 0$ be the zeros of R_n corresponding to $z_{1n}, \cdots, z_{nn}$, respectively. Then, since $\gamma_n^2 = z_{1n}$,

$$w_{1n} = \left(\frac{\gamma_n}{n}\right)^2, \qquad n \in \mathbb{N}. \tag{8}$$

Let w_j denote the (positive) zeros of $\cos(z^{-1/2})$, that is,

$$w_j = \frac{4}{\pi^2(2j-1)^2}, \qquad j \in \mathbb{N},$$

which, in particular, implies $w_1 > w_2 > \cdots > w_j > \cdots > 0$. Then, by Lemma 4 and the theorem of Hurwitz we may conclude that for every fixed j

$$\lim_{n\to\infty} w_{jn} = w_j, \qquad j = 1, 2, 3, \cdots.$$

The case $j = 1$, combined with (8), proves the assertion of our theorem. □

REFERENCES

[1] P. DÖRFLER, *New inequalities of Markov type*, SIAM J. Math. Anal., 18 (1987), pp. 490-494.

[2] E. HILLE, G. SZEGÖ, AND J. D. TAMARKIN, *On some generalizations of a theorem of A. Markoff*, Duke Math. J., 3 (1937), pp. 729-739.

[3] G. V. MILOVANOVIĆ, *An extremal problem for polynomials with nonnegative coefficients*, Proc. Amer. Math. Soc., 94 (1985), pp. 423-426.

[4] E. SCHMIDT, *Über die nebst ihren Ableitungen orthogonalen Polynomensysteme und das zugehörige Extremum*, Math. Ann., 119 (1944), pp. 165-204.

[5] P. TURÁN, *Remark on a theorem of Erhard Schmidt*, Mathematica (Cluj), 2 (1960), pp. 373-378.

[6] A. K. VARMA, *Some inequalities of algebraic polynomials having real zeros*, Proc. Amer. Math. Soc., 75 (1979), pp. 243-250.

[7] ———, *Derivatives of polynomials with positive coefficients*, Proc. Amer. Math. Soc., 83 (1981), pp. 107-112.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 796-809, May 1991

015

MINIMAL EXTRAPOLATIONS OF FILTERS*

BENJAMIN B. WELLS†

Abstract. In this paper the following question is answered for the line and for the circle. When is a trapezoid t the graph of a function whose Fourier norm is smallest among functions whose graphs coincide with t on its intervals of constancy? When such functions are viewed as frequency responses of low-pass filters, they are optimally stable. In the periodic case this means that their Fourier transforms have the property that when implemented as digital filters, the least upper bound of all ratios of norms of output sequences to norms of corresponding input sequences is as small as possible.

Key words. absolutely convergent Fourier series, A-norm, quotient norm, pseudomeasure, low-pass filter

AMS(MOS) subject classifications. 42A16, 42A28

1. Introduction and preliminaries. We denote by $\mathbb{R}$ the group of real numbers under addition, by T the circle group of real numbers under addition modulo 2π, and by Z the group of integers. By an A-function we mean either a continuous complex-valued function f defined on $\mathbb{R}$ and expressed as the Fourier transform of an absolutely integrable function, or else a function f defined on the circle T and having absolutely convergent Fourier series. In the first case the A-norm of f is defined to be the L^1-norm of the absolutely integrable function giving rise to f; in the latter case the A-norm of f is defined to be the sum of the absolute values of the terms of the Fourier series of f.

Let E denote a fixed closed subset of the reals $\mathbb{R}$ or of the circle T, and let $I(E)$ denote those A-functions that are identically zero on E. An A-function f is said to be of *minimal extrapolation* from E provided

$$\|f\|_A = \inf\{\|f+g\|_A \mid g \text{ is in } I(E)\}. \tag{1}$$

The concept of minimal extrapolation was introduced by Beurling [2].

The expression on the right-hand side of (1) defines a norm on the quotient algebra $A(E) = A/I(E)$. The dual space of $A(E)$ may be identified with those pseudomeasures that annihilate $I(E)$. For the closed sets E considered in this paper, this class coincides with the pseudomeasures supported by E. Most of the pseudomeasures treated in this paper are discrete measures. To emphasize the duality with A, however, we shall use the term "pseudomeasure" throughout.

PROPOSITION 1.1. *The function f is of minimal extrapolation from E if and only if there is a pseudomeasure S of norm 1 whose support is contained in E such that*

$$(S, f) = \|f\|_A.$$

Proof. Suppose the existence of S. Thus,

$$(S, f+g) = (S, f) = \|f\|_A \tag{2}$$

for all g belonging to $I(E)$. An immediate consequence of (2) is that $\|f+g\|_A \geqq \|f\|_A$ for all g in $I(E)$, i.e., f is of minimal extrapolation.

* Received by the editors January 29, 1990; accepted for publication May 3, 1990.
† Science Applications International Corporation, 803 West Broad Street, Falls Church, Virginia 22046.

Conversely assume that f is of minimal extrapolation. Thus, the A-norm of f equals its quotient norm. By duality, we may select a pseudomeasure S, annihilating $I(E)$, such that $(S,f) = \|f\|_A$. This completes the proof of the proposition.

Let E denote the points of an equipartition of the circle T. A result of Herz (cf. [6, p. 58]) is that any continuous function that is piecewise linear on the intervals defined by E is of minimal extrapolation from E. Reference [1, Prop. 1] contains another proof of the Herz criterion. If the points of the partition E are not uniformly spaced, we might expect that a piecewise linear function need not be of minimal extrapolation relative to E. Indeed, it is no longer clear whether minimal extrapolation functions exist.

To guarantee the existence of minimal extrapolation functions, it is necessary to make some assumption about the set E. To see this, let E be a countable independent set having zero as an accumulation point. Let J be an interval containing zero such that $J \cap E$ and $J^c \cap E$ are both nonempty. Furthermore, let f be a function defined to be equal to 1 on $J \cap E$ and zero on $J^c \cap E$. The function f has no minimal extrapolation from the set E. This is a consequence of the fact that the quotient norm of f relative to E is equal to 1. It is easy to see that there can be no absolutely convergent Fourier series representing f and having A-norm equal to 1.

Esseen [3] pointed out that Lemma 1 of [2] implies that a minimal extrapolation always exists if E is the closure of its interior. By a standard convergence argument it is easy to demonstrate that a sufficient condition for the existence of minimal extrapolations relative to a set E is that every portion of E (i.e., every nonvoid intersection of E with an open interval) have positive measure. Thus, certainly, if E is a finite union of nondegenerate, closed intervals this condition would be met.

The question that led us to this study is the following. Among functions whose graphs are trapezoids symmetrical about the y-axis, which ones are of minimal extrapolation from their intervals of constancy? This problem is relevant to the subject of filter theory. For a detailed account of this subject, the reader is referred to the monograph [5]. A function whose graph is a trapezoid is an example of the amplitude of a bandpass filter. When such a filter is used as a multiplier, function values are "passed" in the pass band (where the amplitude of the filter is equal to 1), and function values in the stop band (where the values of the filter are equal to zero) are made equal to zero.

A function is said to define a *stable* filter if the operation defined by convolution with its Fourier coefficients is a bounded operator from the space of bounded sequences into itself. The stability of a filter is determined by what is happening in the transition bands. In the case of the trapezoid, these are the intervals where the function values give rise to the sides of the trapezoid. If a function defining a filter is an A-function, its A-norm is equal to the norm of the operator from the space of bounded sequences to itself defined by convolution with its Fourier coefficients. Indeed, the A-norm of the filter is exactly equal to the least upper bound of all ratios of norms of output sequences to norms of corresponding input sequences. Reference [1] contains estimates for lower bounds on the A-norms of filters having specified pass, stop, and transition bands.

For a given pass band and a given stop band, it is desirable to ensure that the least upper bound of all ratios of norms of output sequences to norms of corresponding input sequences is as small as possible. That is, we would like to know the minimal extrapolation of a filter from its intervals of constancy. For the case of trapezoids on the line, it is shown in § 2 that a necessary and sufficient condition is that the length of the pass band be an integral multiple of the length of the transition band. In § 3

the circle group is treated. It is shown that the only trapezoids that are minimal extrapolations from their intervals of constancy are those satisfying the Herz criterion. In the final section of the paper it is shown that in the case of the circle T, the trapezoid is never a unique minimal extrapolation from its intervals of constancy.

In the remainder of the paper we shall be dealing with extrapolations from the intervals of constancy of a function. The following definition will help make our statements more concise.

DEFINITION. We shall say that a function f has the M property if it is of minimal extrapolation from its intervals of constancy.

2. Filters defined by trapezoids on $\mathbb{R}$. In this section we prove that a necessary and sufficient condition that a trapezoid on $\mathbb{R}$ be a minimal extrapolation from its intervals of constancy is that the length of its pass band be an integral multiple of the length of its transition band.

For $0 \leqq \varepsilon \leqq \omega \leqq \omega + \varepsilon < \infty$, $t_{\omega,\varepsilon}$ will denote the function whose graph has the shape of a trapezoid of height 1, which is equal to 1 on the interval $[-\omega + \varepsilon, \omega - \varepsilon]$ and is equal to zero outside the interval $[-\omega - \varepsilon, \omega + \varepsilon]$:

$$\begin{aligned} t_{\omega,\varepsilon}(x) &= 1, \qquad -\omega + \varepsilon \leqq x \leqq \omega - \varepsilon, \\ &= 0, \qquad \omega + \varepsilon \leqq |x|. \end{aligned} \tag{3}$$

In the present setting, the Fourier transform of the function $t_{\omega,\varepsilon}$ is in $L^1(\mathbb{R})$, and its L^1-norm is the A-norm of $t_{\omega,\varepsilon}$. Fourier transforms of pseudomeasures are functions which belong to $L^\infty(\mathbb{R})$.

A straightforward computation yields that the Fourier transform of $t_{\omega,\varepsilon}$ is given by

$$\begin{aligned} t^{\wedge}_{\omega,\varepsilon}(y) &= 2(\sin \omega y \sin \varepsilon y)/\varepsilon y^2, \qquad y \neq 0, \\ &= 2\omega, \qquad y = 0. \end{aligned} \tag{4}$$

THEOREM 2.1. *The function $t_{\omega,\varepsilon}$ has the M property if and only if ω is an integral multiple of ε.*

LEMMA 2.2. *For positive α, $\|t_{\omega,\varepsilon}\|_{A(\mathbb{R})} = \|t_{\alpha\omega,\alpha\varepsilon}\|_{A(\mathbb{R})}$.*

Proof. The proof is an immediate consequence of the relation

$$t_{\alpha\omega,\alpha\varepsilon}(x) = t_{\omega,\varepsilon}(x/\alpha),$$

and the fact that $\|f(x/\alpha)\|_{A(\mathbb{R})} = \|f\|_{A(\mathbb{R})}$ for any $f \in A(\mathbb{R})$.

LEMMA 2.3. *For positive α, $t_{\omega,\varepsilon}$ has the M property if and only if $t_{\alpha\omega,\alpha\varepsilon}$ does.*

Proof. Apply Lemma 2.2.

We first prove that $t_{\omega,\varepsilon}$ has the M property when ω is an integral multiple of ε. By Lemma 2.3 there is no loss of generality in assuming that $\varepsilon/2\pi = 1$ and that $\omega/2\pi$ is an integer. Define the periodic function f by

$$f(y) = \operatorname{sign}(\sin \omega y).$$

A straightforward calculation yields that the Fourier coefficients of f are given by

$$\begin{aligned} f^{\wedge}(\omega n/2\pi) &= 4i/\omega n \quad \text{for } n \text{ odd}, \\ &= 0 \quad \text{for } n \text{ even}. \end{aligned}$$

Define the pseudomeasure S on $\mathbb{R}$ by setting its Fourier transform to be the periodic function

$$S^{\wedge}(y) = \operatorname{sign}(\sin \omega y \sin \varepsilon y). \tag{5}$$

It is evident from the expressions (4) and (5) of the Fourier transforms that S imparts its A-norm to $t_{\omega,\varepsilon}$. From the calculation of the last paragraph, we see that the nth Fourier coefficient of $S^\wedge$ is given by

$$S(n) = -\frac{8}{\pi\omega} \sum_{j\in\Lambda} 1/j(n - j\omega/2\pi), \tag{6}$$

where Λ is the set of odd integers j such that $n - j\omega/2\pi$ is an odd multiple of $\varepsilon/2\pi$. Since $\varepsilon/2\pi = 1$, this implies that $S(n)$ can be nonzero only when n and $\omega/2\pi$ have opposite parity.

It is easy to see from the definition of Λ that the support of S misses the intervals $(-\omega - \varepsilon, -\omega + \varepsilon)$ and $(\omega - \varepsilon, \omega + \varepsilon)$. For, suppose that $2\pi n = j\omega + k\varepsilon$ lies in the interval $(\omega - \varepsilon, \omega + \varepsilon)$ with j and k odd and $\omega = q\varepsilon$, then $-\varepsilon < \{(j-1)q + k\}\varepsilon < \varepsilon$. Since $(j-1)q + k$ is odd, this is not possible. It follows from Proposition 1.1 that $t_{\omega,\varepsilon}$ has the M property.

To complete the proof of the theorem, it is necessary to show that $t_{\omega,\varepsilon}$ fails to have the M property whenever ω is not an integral multiple of ε. Suppose, therefore, that $\omega/2\pi$ and $\varepsilon/2\pi$ are relatively prime positive integers, $\varepsilon \leqq \omega$, with $\varepsilon/2\pi \neq 1$. At least one of $\omega/2\pi$ and $\varepsilon/2\pi$ is odd. If $\varepsilon/2\pi$ is odd, it suffices to show that (6) is nonzero at $\omega/2\pi + 1$. A similar argument shows that (6) is nonzero at $\omega/2\pi$ in the case where $\omega/2\pi$ is odd and $\varepsilon/2\pi$ is even. Suppose that $\omega/2\pi + 1 = j_0\omega/2\pi + k_0\varepsilon/2\pi$ for odd integers j_0 and k_0. Except for the multiplicative constant in front, the sum in (6) may be written as

$$\sum_{\lambda=-\infty}^{\infty} 1/\{(j_1 + 2\lambda\varepsilon/2\pi)(\omega/2\pi + 1 - (j_1 + 2\lambda\varepsilon/2\pi)\omega/2\pi)\},$$

where j_1 is an odd integer such that $1 \leqq j_1 < 2\varepsilon/2\pi$, and $j_1 \equiv j_0$ modulo $2\varepsilon/2\pi$. It is clear that each of the terms of the above sum is negative, except possibly the $\lambda = 0$ term. Now, j_1 cannot be equal to 1, for if $j_0 \equiv 1$ modulo $2\varepsilon/2\pi$, it would follow that $\varepsilon/2\pi = 1$, contrary to assumption. Therefore, the term corresponding to $\lambda = 0$ is negative as well, and it follows that S has nonzero support in the transition intervals $(-\omega - \varepsilon, -\omega + \varepsilon)$ and $(\omega - \varepsilon, \omega + \varepsilon)$.

Since $t^\wedge_{\omega,\varepsilon}$ is zero on a discrete set, $S^\wedge$ is undefined only on a set of Lebesgue measure zero. If it were the case that $t_{\omega,\varepsilon}$ had the M property, by Proposition 1.1 there would exist a pseudomeasure S_0 such that $S_0^\wedge(x) = S^\wedge(x)$ almost everywhere with respect to Lebesgue measure and such that its support is disjoint from the transition intervals. That is to say, as L^∞-functions $S_0^\wedge$ and $S^\wedge$ are identical. Therefore, $t_{\omega,\varepsilon}$ fails to have the M property.

It follows from the last paragraph and Lemma 2.3 that $t_{\omega,\varepsilon}$ fails to enjoy the M property whenever the quotient ω/ε is rational and nonintegral. Suppose now that this quotient is irrational. By Lemma 2.3 there is no loss of generality in assuming that $\omega = 2\pi$. Any function in $L^\infty(\mathbb{R})$ of norm 1 which imparts its A-norm to $t_{\omega,\varepsilon}$ must be equal almost everywhere to the function $S^\wedge(y)$ defined by (5). Therefore, it only remains to check that the support of S is not disjoint from the transition intervals $(-\omega - \varepsilon, -\omega + \varepsilon)$ and $(\omega - \varepsilon, \omega + \varepsilon)$.

Recall that the Fourier coefficients of the $2\pi/\varepsilon$ periodic function

$$f(y) = \text{sign}\,(\sin \varepsilon y)$$

are given by

$$f^\wedge(\varepsilon n/2\pi) = 4i/\varepsilon n \quad \text{for } n \text{ odd},$$
$$= 0 \quad \text{for } n \text{ even}.$$

Therefore, the support of the pseudomeasure S will be at the points of the form $\{\varepsilon n/2\pi + k\}$ where n and k are odd integers. In fact, since $\varepsilon/2\pi$ is irrational, the sum $\varepsilon n/2\pi + k$ uniquely determines n and k, therefore

$$S(\varepsilon n/2\pi + k) = -8/(\pi\varepsilon nk), \qquad n \text{ and } k \text{ odd}.$$

Again using the irrationality of $\varepsilon/2\pi$, we see that the set of numbers of the form $\{\varepsilon n/2\pi + k\}$ as n and k range over odd integers is dense in $\mathbb{R}$, and certainly cannot be disjoint from the transition intervals $(-\omega - \varepsilon, -\omega + \varepsilon)$ and $(\omega - \varepsilon, \omega + \varepsilon)$. This completes the proof of Theorem 2.1.

3. Filters defined by trapezoids on T. In this section we shall characterize those trapezoids on the circle group that have the M property. The class of such trapezoids turns out to be a more narrow class than that defined on the line $\mathbb{R}$ and enjoying the same property. Indeed, beyond those guaranteed by the criterion of Herz, there are none.

Suppose that $0 \leqq \varepsilon \leqq \omega \leqq \omega + \varepsilon \leqq \pi$. The exact statement of the theorem is the following.

THEOREM 3.1. *The function $t_{\omega,\varepsilon}$ has the M property on T if and only if either $\omega = \varepsilon$ or there are positive integers j_0 and N, such that $\omega = \pi j_0/N$ and $\varepsilon = \pi/N$.*

The sufficiency of the condition is the Herz criterion applied to $t_{\omega,\varepsilon}$ and an equipartition of $(-\pi, \pi]$ having mesh equal to $2\pi/N$. If zero is a point of the equipartition, j_0 is odd, otherwise j_0 is even. The remainder of the section is devoted to a proof of Theorem 3.1.

Let S denote a pseudomeasure defined on $\mathbb{R}$. A pseudomeasure S_0 corresponding to S is defined on T by specifying that its Fourier transform be the restriction of $S^\wedge$ to Z. The pseudomeasure S_0 is defined by

$$S_0(x) = \sum_{k=-\infty}^{\infty} S((x/2\pi) + k). \tag{7}$$

Convergence of (7) is understood to mean pointwise convergence on evaluation at A-functions.

Let the pseudomeasure S be defined on $\mathbb{R}$ by the equality

$$S^\wedge(y) = \operatorname{sign}(\sin(\omega y)).$$

Recall from § 2 that S is given by

$$S(\omega n/2\pi) = 4i/\omega n \quad \text{for } n \text{ odd},$$
$$= 0 \quad \text{for } n \text{ even}.$$

If $\omega/2\pi$ is irrational, the pseudomeasure S_0 is given by

$$S_0(\omega n) = 4i/\omega n \quad \text{for } n \text{ odd}$$
$$= 0 \quad \text{for } n \text{ even}, \tag{8}$$

where the argument of S_0 is taken modulo 2π. If $\omega = \pi k_0/N$ for relatively prime integers k_0 and N, using (7) we obtain

$$S_0(\omega n) = \sum_{k=-\infty}^{\infty} \frac{4i}{\omega(n + 2Nk)} \quad \text{for } n \text{ odd}, \tag{9}$$
$$= (2i/k_0)\cot(n\pi/2N).$$

The last equality comes about from recognizing the Weierstrass partial fractions decomposition of the cotangent function.

We now let S be the pseudomeasure defined on $\mathbb{R}$ by the equality

$$S^{\wedge}(y) = \operatorname{sign}(\sin(\omega y)\sin(\varepsilon y)). \tag{10}$$

We first consider the case of ω/π irrational and $\varepsilon/\pi = k_0/N$ in lowest terms. In this case $S^{\wedge}$ is defined everywhere except at the points NZ where $\sin(\varepsilon y)$ vanishes. The points $\{j\omega + k\varepsilon\}$ modulo 2π as j and k range over odd integers are distinct and dense in T. From (8) and (9) we have for odd j and k:

$$S_0(j\omega + k\varepsilon) = -(8/\omega j k_0)\cot(k\pi/2N). \tag{11}$$

It is clear that the right-hand side of (11) is nonzero, except if N is odd and $k = \xi N$ for odd ξ. If μ is a pseudomeasure with spectrum contained in NZ, then μ is $2\pi/N$ periodic, i.e., $\mu(x) = \mu(x + 2\pi/N)$. It must be shown that no pseudomeasure of the form $S_0 - \mu$, where the spectrum of μ is contained in NZ, will have support disjoint from the transition intervals $(-\omega - \varepsilon, -\omega + \varepsilon)$ and $(\omega - \varepsilon, \omega + \varepsilon)$. However, this follows easily, since if $\zeta \in (\omega - \varepsilon, \omega + \varepsilon)$, $\zeta + 2\lambda\varepsilon$ belongs to the interval $(-\omega - \varepsilon, -\omega + \varepsilon)$ for some integer λ. The value of μ is the same at ζ as at $\zeta + 2\lambda\varepsilon$, although the values of S_0 are distinct at these two points. Therefore, by Proposition 1.1 $t_{\omega,\varepsilon}$ does not have the M property.

We next consider the case where ω/π and ε/π are both irrational with $\varepsilon \neq \omega$. Since $S_0^{\wedge}(n)$ is defined at *every* value of n, it follows that it is only necessary to show that the pseudomeasure S_0 has nonvoid support in the transition intervals $(-\omega - \varepsilon, -\omega + \varepsilon)$ and $(\omega - \varepsilon, \omega + \varepsilon)$. As a first subcase, assume that ε is not a rational multiple of ω. Then the points $\{j\omega + k\varepsilon\}$ modulo 2π as j and k range over odd integers are distinct and dense in T. We have from (8) that

$$S_0(j\omega + k\varepsilon) = -4^2/(j\omega k\varepsilon) \quad \text{for } j \text{ and } k \text{ odd integers,}$$

and therefore, the support of S_0 is nonvoid in the intervals $(-\omega - \varepsilon, -\omega + \varepsilon)$ and $(\omega - \varepsilon, \omega + \varepsilon)$.

Next consider the subcase where $\varepsilon = p\omega/q$ for relatively prime integers p and q. Then the points $\{j\omega + k\varepsilon\}$ modulo 2π are not distinct, but are still dense in T, since ω/π and ε/π are irrational. By using (8), it follows that the value of S_0 at $\zeta = j\omega + k\varepsilon$ for j and k odd is given by

$$\begin{aligned} S_0(j\omega + k\varepsilon) &= -4^2 \sum \frac{1}{\omega(j + 2np)} \frac{1}{\varepsilon(k - 2nq)}, \\ &= \frac{-4^2}{\zeta}\left[\sum \frac{1}{\omega j + 2n\omega p} + \sum \frac{1}{\varepsilon k - 2n\omega p}\right], \end{aligned} \tag{12}$$

where the sums are taken over all integers n.

Recognizing the Weierstrass decomposition of the cotangent function, we may rewrite (12) as

$$S_0(\zeta) = \frac{-4^2\pi}{2\zeta\omega p}\left[\cot\left(\frac{j\pi}{2p}\right) + \cot\left(\frac{k\pi}{2q}\right)\right]. \tag{13}$$

Note that (13) is equal to zero only when $jq + kp \equiv 0$ modulo $2pq$, i.e., when ζ is a multiple of $2p\omega$. Of course, if $\omega = \varepsilon$, then $p = q = 1$ and $S_0(\zeta) = 0$ for all such ζ except $\zeta = 0$. Thus, in the present case when $\varepsilon \neq \omega$, the support of S_0 intersects the intervals $(-\omega - \varepsilon, -\omega + \varepsilon)$ and $(\omega - \varepsilon, \omega + \varepsilon)$. Again, by Proposition 1.1 we conclude that $t_{\omega,\varepsilon}$ does not have the M property.

We now consider the final case, namely, where ω/π and ε/π are both rational numbers. In this case the support of S_0 is contained in the finite group $T_N = \{\pi k/N \mid 0 \leqq k < 2N\}$ for some integer N. For the remainder of the section, the appropriate setting for computations of the Fourier transform is the group T_N.

LEMMA 3.2. *For relatively prime positive integers p and N, set $\omega/\pi = p/N$, and let S be the pseudomeasure defined by the equality $S\hat{}(n) = \operatorname{sign}(\sin \omega n)$, for n not in NZ.*

(a) *If p is odd and $S\hat{}(Nk) = (-1)^k$ for $k \in Z$, S is given by*

$$S\left(\frac{\pi k}{N}\right) = \frac{4}{1 - e^{i\pi kp'/N}} \quad \text{for } k \text{ odd},$$

$$= 0 \quad \text{for } k \text{ even},$$

where p' denotes the multiplicative inverse of p modulo 2N.

(b) *If p is even and $S\hat{}(Nk) = 1$ for $k \in Z$, S is given by*

$$S\left(\frac{\pi k}{N}\right) = 4\left[\frac{1 - e^{i\pi kp'([N/2]+1)/N}}{1 - e^{i\pi kp'/N}}\right] \quad \text{for } k \text{ even},$$

$$= 0 \quad \text{for } k \text{ odd},$$

where p' denotes the multiplicative inverse of p/2 modulo N.

Proof. The result is a straightforward calculation starting with the Fourier transform of $S\hat{}$ on the finite group T_N. For the proof of part (a) we begin with the expression

$$\begin{aligned} S\left(\frac{\pi k}{N}\right) &= \sum_{n=0}^{2N-1} \operatorname{sign}\left(\sin\frac{\pi pn}{N}\right) e^{i\pi kn/N} \\ &= \sum_{m=0}^{2N-1} \operatorname{sign}\left(\sin\frac{\pi m}{N}\right) e^{i\pi kmp'/N} \\ &= \sum_{m=0}^{N-1} e^{i\pi kmp'/N} - \sum_{m=N}^{2N-1} e^{i\pi kmp'/N} \\ &= 2\sum_{m=0}^{N-1} e^{i\pi kmp'/N} \quad \text{for } k \text{ odd}, \\ &= 0 \quad \text{for } k \text{ even}. \end{aligned}$$

This may easily be expressed in the form of the conclusion of part (a).

For the proof of part (b) note that

$$S\left(\frac{\pi k}{N}\right) = \sum_{n=0}^{2N-1} \operatorname{sign}\left(\sin\frac{\pi pn}{N}\right) e^{i\pi kn/N} \quad \text{becomes}$$

$$S\left(\frac{\pi k}{N}\right) = 2\sum_{n=0}^{N-1} \operatorname{sign}\left(\sin\frac{\pi pn}{N}\right) e^{i\pi kn/N} \quad \text{when } k \text{ is even},$$

$$= 0 \quad \text{for } k \text{ odd}.$$

For k even, the substitution $n = mp'$ allows this to be expressed as

$$S\left(\frac{\pi k}{N}\right) = 2\sum_{m=0}^{[N/2]} e^{i\pi kmp'/N} - 2\sum_{m=[N/2]+1}^{N-1} e^{i\pi kmp'/N},$$

which is easily rewritten to conform to the conclusion of part (b).

We first assume that $\omega/\pi = p/N$, and $\varepsilon/\pi = q/N$, where p and q are odd and relatively prime to N. Later, we shall indicate changes in the proof to accommodate

the case of either p or q even and provide a proof for the case when p and q have common factors with N. Recall that the nth Fourier coefficient of the trapezoid $t_{\omega,\varepsilon}$ is given by

$$\begin{aligned} t^{\wedge}_{\omega,\varepsilon}(n) &= (\sin \omega n \sin \varepsilon n)/\pi\varepsilon n^2 \quad \text{for } n \neq 0, \\ &= \omega/\pi \quad \text{for } n = 0. \end{aligned} \tag{14}$$

LEMMA 3.3. *For λ and N positive integers, let the pseudomeasure S_0 be defined on $T_{\lambda N}$. The pseudomeasure S defined by the relation*

$$S\left(\frac{\pi n}{N}\right) = \frac{1}{\lambda} \sum_{\nu=0}^{\lambda-1} S_0\left\{\frac{\pi}{\lambda}\left(\frac{n}{N}+2\nu\right)\right\}, \qquad 0 \leqq n < 2N,$$

satisfies

$$S^{\wedge}(n) = S_0^{\wedge}(\lambda n), \qquad n \in Z.$$

Proof. A straightforward calculation gives the result. The Fourier transform of S is given by

$$\begin{aligned} S^{\wedge}(n) &= \frac{1}{N} \sum_{k=0}^{2N-1} \frac{1}{\lambda} \sum_{\nu=0}^{\lambda-1} S_0\left(\frac{\pi}{\lambda}\left(\frac{k}{N}+2\nu\right)\right) e^{-in\pi\lambda k/\lambda N} \\ &= \frac{1}{\lambda N} \sum_{k=0}^{2N-1} \sum_{\nu=0}^{\lambda-1} S_0\left(\frac{\pi}{\lambda N}(k+2\nu N)\right) e^{-i(k+2\nu N)\pi\lambda n/\lambda N} \\ &= S_0^{\wedge}(\lambda n). \end{aligned}$$

Take S to be the pseudomeasure defined by

$$\begin{aligned} S^{\wedge}(n) &= \operatorname{sign}(\sin \pi p n/N \sin \pi q n/N) \quad \text{for } n \neq 0, \\ &= 1 \quad \text{for } n = 0. \end{aligned} \tag{15}$$

Replace n by nq' in the right-hand side of (15), where q' denotes the multiplicative inverse of q modulo $2N$. Let $p^{\sim}$ denote the representative of pq' modulo $2N$, $1 \leqq p^{\sim} < 2N$. If $1 \leqq p^{\sim} < N$ set $r = p^{\sim}$. If $N < p^{\sim} < 2N$, set $r = 2N - p^{\sim}$. In the computations below we shall assume the former case. If the latter holds, except for the zeroth coefficient, the sign of each Fourier coefficient is reversed. This change of sign does not alter the arguments presented below.

The case where $r = 1$ corresponds to $p = q$ when the condition of the theorem holds. Therefore, assume that $1 < r$, and let r', $1 < r' < N$, denote the inverse of r modulo N, so that $rr' = 1 + \lambda N$ for some positive integer λ.

Under the above assumptions, for n even we have

$$\begin{aligned} S\left(\frac{\pi n}{N}\right) &= \sum_{\nu=0}^{2N-1} \operatorname{sign}\left\{\sin\left(\frac{\pi r\nu}{N}\right) \sin\left(\frac{\pi\nu}{N}\right)\right\} e^{i\pi\nu n/N} \\ &= 2\left[\sum_{\nu=0}^{N-1} \operatorname{sign}\left(\sin\frac{\pi r\nu}{N}\right) e^{i\pi\nu n/N}\right]. \end{aligned} \tag{16}$$

The constant coefficient of (16) is defined to be 1. Before proceeding to the evaluation of $S(\pi n/N)$, we replace N by λN, and evaluate $S_0(\pi n/\lambda N)$. For n even it is given by

$$\begin{aligned} 1 + \sum_{\nu=1}^{2\lambda N-1} \operatorname{sign}\left\{\sin\left(\frac{\pi r\nu}{\lambda N}\right) \sin\left(\frac{\pi\nu}{\lambda N}\right)\right\} e^{i\pi\nu n/\lambda N} \\ = 2\left[1 + \sum_{\nu=1}^{\lambda N-1} \operatorname{sign}\left(\sin\frac{\pi r\nu}{\lambda N}\right) e^{i\pi\nu n/\lambda N}\right]. \end{aligned} \tag{17}$$

The term inside the brackets of the right-hand side of (17) may be expressed in blocks with alternating sign:

$$\sum_{\nu=0}^{r'-1} e^{\pi i\nu n/\lambda N} - \sum_{\nu=r'}^{2r'-1} e^{\pi i\nu n/\lambda N} + \cdots + (-1)^{r-1} \sum_{\nu=(r-1)r'}^{rr'-1} e^{\pi i\nu n/\lambda N}. \tag{18}$$

This may be simplified to arrive at the expression

$$\left[\frac{1+e^{\pi in/\lambda N}}{1-e^{\pi in/\lambda N}}\right]\left[\frac{1-e^{\pi inr'/\lambda N}}{1+e^{\pi inr'/\lambda N}}\right] \quad \text{for } n \text{ even}, \quad n \neq 0,$$

which in turn is equal to

$$\cot(\pi n/2\lambda N)\tan(\pi r'n/2\lambda N). \tag{19}$$

By Lemma 3.3 and (19) it follows that for n even and not zero, except for a multiplicative constant, $S(\pi n/N)$ is given by

$$\frac{1}{\lambda}\sum_{j=0}^{\lambda-1} \cot\left\{\frac{\pi}{2\lambda}\left(\frac{n}{N}+2j\right)\right\}\tan\left\{\frac{\pi r'}{2\lambda}\left(\frac{n}{N}+2j\right)\right\}. \tag{20}$$

Set $x=\pi n/N$ and let φ denote the function defined by (19). The derivative of φ is given by

$$\varphi'(x)=\frac{r'\sin(x/\lambda)-\sin(r'x/\lambda)}{2^2\lambda\sin^2(x/2\lambda)\cos^2(r'x/2\lambda)}.$$

When $\lambda < r'$, $\varphi' > 0$ for $0 < x < \lambda\pi$. Since (20) is a sum of terms like (19), it follows that it defines an increasing function Ψ on intervals whose endpoints are points of discontinuity of at least one of the terms.

Under present assumptions p and q are odd integers. Furthermore, assume for the moment that $q \geqq 5$. In this case we may assume that $1 < r' < N/2$, for, if not, we replace r' by $N-r'$, so that $r(N-r')=-1+(r-\lambda)N$. Then proceed with the computations outlined above. In formulas (19) and (20) λ is replaced by $(r-\lambda)$, and a 1 is added.

From (16) and Proposition 1.1 it follows that to prove the theorem in the present case, it is enough to show that the polynomial $\sum_{\nu=0}^{N-1} \operatorname{sign}(\sin(\pi r\nu/N))X^{\nu}$ (the constant coefficient is defined to be 1) does not have each of the four roots of unity $e^{\pi i(p\pm1)q'/N}$ and $e^{\pi i(p\pm3)q'/N}$ as roots. Assume the contrary. These roots of unity are primitive roots of unity for some integers $N_1, \cdots, N_4$. Since this polynomial has integer coefficients and q' is relatively prime to N, it follows that it has $e^{\pi i(p\pm1)/N}$ and $e^{\pi i(p\pm3)/N}$ as additional roots, since they are primitive roots of unity for the same respective integers $N_1, \cdots, N_4$ (cf. [8, pp. 203–208]). Therefore (20) is zero for $n = p \pm 1$ and for $n = p \pm 3$. In view of the monotonicity of Ψ on the intervals of continuity of its terms, and since (20) is zero at each of these points, there exist odd integers, ξ_1, ξ_2, ξ_3 corresponding to endpoints of intervals of continuity of Ψ such that

$$\frac{p-3}{N} < \frac{\xi_3}{r'} < \frac{p-1}{N} < \frac{\xi_1}{r'} < \frac{p+1}{N} < \frac{\xi_2}{r'} < \frac{p+3}{N}.$$

Hence $4/r' \leqq (\xi_2-\xi_3)/r' \leqq 6/N$, which implies that $2N/3 < r'$, contrary to the assumption that $1 < r' < N/2$. Hence S is nonzero at one of the four points $\pi(p\pm1)/N$, $\pi(p\pm3)/N$, and by Proposition 1.1 it follows that $t_{\omega,\varepsilon}$ does not have the M property.

Assume now that $q=3$. It must be shown that (20) is not zero at one of $p\pm1$. Assume otherwise. By the same reasoning applied above, there exists an odd integer ξ such that either

(i) $(p-1)/N < \xi/r' < p/N$, or

(ii) $p/N < \xi/r' < (p+1)/N$.

Suppose that $1 < 3p' < N$. Set $r' = 3p'$, and suppose that case (i) holds. Now, by definition $pp' = 1 + \tau 2N$ for some integer τ. Therefore, (i) becomes

$$3(1 + \tau 2N) - r' < \xi N < 3(1 + \tau 2N).$$

This is clearly impossible in view of the fact that ξ is odd and $1 < r' < N$. In the case where (ii) holds, we have

$$3(1 + \tau 2N) < \xi N < 3(1 + \tau 2N) + r'.$$

If $N < 3p' < 2N$, set $r' = 2N - 3p'$; if $2N < 3p' < 3N$, set $r' = 3p' - 2N$; and if $3N < 3p' < 4N$, set $r' = 4N - 3p'$. In the first and third instances, λ is replaced by $2r - \lambda$ and $4r - \lambda$, respectively, and the proof proceeds as before. In each case the above inequalities (i) and (ii) result in contradiction, except possibly in the case where $3 + r' > N$. The cases where $r' = N - 1$ and $r' = N - 2$ must be examined separately.

In the case where $r' = N - 1$, we have that $r = N - 1$; in the case where $r' = N - 2$, we have that N must be odd and $r = (N - 1)/2$. In both cases the expression (17) with $\lambda = 1$ may be evaluated directly. In the first case (17) is constant and equal to 2 for n even and not zero. In the second case the expression inside the brackets of (17) is equal to

$$1 + 2\cos\frac{\pi n}{N}\frac{|1 + e^{i\pi n/N}|^2}{|1 + e^{2\pi i n/N}|^2}. \tag{21}$$

It is straightforward to check that (21) is zero only for $n/N = \pm\frac{2}{3}$, and therefore is zero for at most one of $n = p \pm 1$. By Proposition 1.1 it follows that $t_{\omega,\varepsilon}$ does not have the M property when p and q are odd and relatively prime to N.

If exactly one of p and q is even, the proof outlined above remains unchanged, except that formulas (16)–(21) are to be evaluated for n odd. If both p and q are even, but still relatively prime to N, we replace them by $N - p$ and $N - q$, respectively. Note that (15) is unchanged by this substitution. The proof then proceeds as above, except that the roots of unity of the new polynomial are $e^{\pi i(p\pm 1)(N-q)'/N}$ and $e^{\pi i(p\pm 3)(N-q)'/N}$. As in the previous paragraph, we conclude that $e^{\pi i(p\pm 1)/N}$ and $e^{\pi i(p\pm 3)/N}$ are additional roots and proceed as before. This completes the proof of the theorem for the case where p and q are relatively prime to N.

We now address the case when p and q have nontrivial factors common with N. Suppose that $\omega/\pi = p/N_1$, and $\varepsilon/\pi = q/N_2$ are written in lowest terms, and let N_3 denote the greatest common divisor of N_1 and N_2, so that $N_1 = cN_3$ and $N_2 = dN_3$ with c and d relatively prime and not both equal to 1. In this case the imaginary part of $S(\gamma)$ is nonzero. Let λ denote the least common multiple of p and q. We shall first evaluate S when $\omega/\pi = p/\lambda N_1$ and $\varepsilon/\pi = q/\lambda N_2$ and later apply Lemma 3.3 to get S when $\omega/\pi = p/N_1$ and $\varepsilon/\pi = q/N_2$. Set $\omega/\pi = M_1 = \lambda N_1/p$ and $M_2 = \lambda N_2/q$. By Lemma 3.2(a) (where p of the lemma is taken to be 1, and N is replaced by M_1 and M_2), the desired convolution is given by

$$S(\gamma) = \sum_{\substack{j \text{ even} \\ 0 \leqq j < 2N_3 - 1}} \frac{4}{1 - \exp(i\pi(r_1 + j\alpha)/M_1)} \frac{4}{1 - \exp(i\pi(r_2 - j\beta)/M_2)},$$

where $\gamma = r_1/M_1 + r_2/M_2$ (r_1 and r_2 odd integers), $\alpha = c\lambda/p$, $\beta = d\lambda/q$, and $S(\gamma) = 0$ for other γ. This may be rewritten in the form

$$4 \sum_{u=0}^{M_2-1} \sum_{\nu=0}^{M_1-1} \exp\left(i\pi\left\{\frac{r_1}{M_1}\nu + \frac{r_2}{M_2}u\right\}\right) \sum_{\substack{j \text{ even} \\ 0 \leqq j < 2N_3 - 1}} \exp(i\pi j(\nu - u)/N_3). \tag{22}$$

Make the substitutions $\nu = v + \xi_1 N_3$ and $u = w + \xi_2 N_3$, where $0 \leqq v, w < N_3$, and $0 \leqq \xi_1 \leqq \alpha - 1$ and $0 \leqq \xi_2 \leqq \beta - 1$. Expression (22) becomes

$$4N_3 \sum_{\xi_1=0}^{\alpha-1} \sum_{\xi_2=0}^{\beta-1} \sum_{\nu=0}^{N_3-1} e^{i\pi\gamma\nu} \exp\left(i\pi \frac{r_1}{\alpha} \xi_1\right) \exp\left(i\pi \frac{r_2}{\beta} \xi_2\right). \tag{23}$$

The real part of (23) is

$$\begin{aligned} &4N_3 \left[\frac{\sin \pi \dfrac{r_1}{\alpha}}{\left(1 - \cos \pi \dfrac{r_1}{\alpha}\right)} + \frac{\sin \pi \dfrac{r_2}{\beta}}{\left(1 - \cos \pi \dfrac{r_2}{\beta}\right)} \right] \frac{\sin \pi\gamma}{(1 - \cos \pi\gamma)} \\ &= 4N_3 \left[\cot \frac{\pi}{2} \frac{r_1}{\alpha} + \cot \frac{\pi}{2} \frac{r_2}{\beta} \right] \cot \frac{\pi\gamma}{2}. \end{aligned} \tag{24}$$

The imaginary part of (23) is

$$4N_3 \left[\cot \frac{\pi}{2} \frac{r_1}{\alpha} + \cot \frac{\pi}{2} \frac{r_2}{\beta} \right]. \tag{25}$$

We now consider $\omega/\pi = p/N_1$ and $\varepsilon/\pi = q/N_2$, where both fractions have been written in lowest terms. Suppose that $c \neq 1$. It follows from Lemma 3.3 and (25) that, except for a multiplicative constant, the imaginary part of $S(\gamma)$ is given by

$$\frac{1}{\lambda} \sum_{\nu=0}^{\lambda-1} \left[\cot \frac{\pi}{2} \frac{r_1}{\lambda} \frac{p}{c} + \cot \frac{\pi}{2} \frac{(r_2 q + 2\nu N_2)}{\lambda d} \right], \tag{26}$$

where $\gamma = r_1 p/N_1 + r_2 q/N_2$ (r_1 and r_2 are odd, and $r_1 p$ and $r_2 q$ are taken modulo N_1 and N_2, respectively, $-1 \leqq \gamma < 1$). The real part of $S(\gamma)$ has a similar expression. Furthermore, $S(\gamma) = 0$ for $\gamma = n/cdN_3$, when n has parity opposite to that of $d + c$.

Note that $S\hat{}(n)$ is not defined for n belonging to the union of the cosets $N_1 Z$ and $N_2 Z$. Therefore, to complete the proof of Theorem 3.1, by Proposition 1.1 it is necessary to show that for no two pseudomeasures μ_1 and μ_2 with spectra contained in $N_1 Z$ and $N_2 Z$, respectively, is it the case that $S - \mu_1 - \mu_2$ has support disjoint from the transition intervals. It suffices to show that either the real part of S or the imaginary part of S has no such representation on the transition intervals. Consider the imaginary part of S, and assume on the contrary that it may be expressed as $\mu_1(\gamma) + \mu_2(\gamma)$ for γ belonging to the transition intervals. Since the spectra of μ_1 and μ_2 are assumed to be contained in $N_1 Z$ and $N_2 Z$, respectively, it follows that μ_1 is $2\pi/N_1$-periodic and μ_2 is $2\pi/N_2$-periodic.

Except in the case where $c = 1$ and $q = 1$ (where the condition of the theorem is true), it is easy to show that for each ξ/N_1 ($-N_1 \leqq \xi < N_1$), except for $\xi \equiv p$ modulo $2c$ if $q = 1$, there are pairs of integers η_0, η_e (the subscripts indicate parity) such that $\pi[\xi/N_1 + \eta/N_2] \in (w - \varepsilon, \omega + \varepsilon)$, where the symbol η stands for those of either odd index "o" or even index "e." To see this, perform the division

$$pd - \xi d = Qc + r, \tag{27}$$

where the remainder r satisfies $0 < r < c$. If Q is even, set $\eta_e = Q$ and $\eta_o = Q + 1$, otherwise, set $\eta_o = Q$ and $\eta_e = Q + 1$. It is immediate that $\pi[\xi/N_1 + \eta/N_2] \in (\omega - \varepsilon, \omega + \varepsilon)$.

We shall assume in the following argument that p and q are both odd. Minor modifications to the proof allow for either p or q to be even. Let $\varphi(\gamma)$ denote the function defined by (26). Let ξ denote an odd integer. Now, by assumption,

$$\varphi\left(\frac{\xi}{N_1}+\frac{\eta_o}{N_2}\right)=\mu_2\left(\frac{\xi}{c}\right)+\mu_1\left(\frac{\eta_o}{d}\right). \tag{28}$$

If c and d are both odd, the sum $(\xi/N_1+\eta_e/N_2)$ does not have the form $r_1/N_1+r_2/N_2$ for r_1 and r_2 both odd integers. Therefore, $\mu_1(\eta_e/d)+\mu_2(\xi/c)=0$. Hence, (28) becomes

$$\varphi\left(\frac{\xi}{N_1}+\frac{\eta_o}{N_2}\right)=-\mu_1\left(\frac{\eta_e}{d}\right)+\mu_1\left(\frac{\eta_o}{d}\right). \tag{29}$$

Since $\pi[\xi/N_1+\eta_o/N_2]$ and $\pi[\xi/N_1+\eta_e/N_2]\in(\omega-\varepsilon,\omega+\varepsilon)$, it follows that $\pi[(\xi-2p)/N_1+\eta_o/N_2]$ and $\pi[(\xi-2p)/N_1+\eta_e/N_2]\in(-\omega-\varepsilon,-\omega+\varepsilon)$. By assumption,

$$\varphi\left(\frac{\xi-2p}{N_1}+\frac{\eta_o}{N_2}\right)=\mu_2\left(\frac{\xi-2p}{c}\right)+\mu_1\left(\frac{\eta_o}{d}\right). \tag{30}$$

Since c and d are both odd, it follows that $(\xi-2p)/N_1+\eta_e/N_2$ does not have the form $r_1/N_1+r_2/N_2$ for r_1 and r_2 both odd integers. Therefore, $\mu_1(\eta_e/d)+\mu_2((\xi-2p)/c)=0$. Hence, (30) becomes

$$\varphi\left(\frac{\xi-2p}{N_1}+\frac{\eta_o}{N_2}\right)=-\mu_1\left(\frac{\eta_e}{d}\right)+\mu_1\left(\frac{\eta_o}{d}\right). \tag{31}$$

Comparing (29) and (31), we obtain

$$\varphi\left(\frac{\xi}{N_1}+\frac{\eta_o}{N_2}\right)=\varphi\left(\frac{\xi-2p}{N_1}+\frac{\eta_o}{N_2}\right).$$

This is the equality

$$\begin{aligned}&\frac{1}{\lambda}\sum_{\nu=0}^{\lambda-1}\left[\cot\frac{\pi}{2}\frac{\xi}{\lambda c}+\cot\frac{\pi}{2}\frac{(\eta_o+2\nu N_2)}{\lambda d}\right]\\&\quad=\frac{1}{\lambda}\sum_{\nu=0}^{\lambda-1}\left[\cot\frac{\pi}{2}\frac{(\xi-2p)}{\lambda c}+\cot\frac{\pi}{2}\frac{(\eta_o+2\nu N_2)}{\lambda d}\right].\end{aligned} \tag{32}$$

Therefore, $\cot(\pi/2)\xi/\lambda c=\cot(\pi/2)(\xi-2p)/\lambda c$, which is clearly impossible.

We finally consider the case where one of c and d is even. If c is odd and d is even, the arguments are the same as in the preceding paragraph, since it is still true that $(\xi/N_1+\eta_e/N_2)$ does not have the form $r_1/N_1+r_2/N_2$ for ξ, r_1, and r_2 odd integers. However, if c is even and d is odd, $(\xi/N_1+\eta_e/N_2)$ does have the form $r_1/N_1+r_2/N_2$ for ξ, r_1, and r_2 odd integers, namely, $(\xi/N_1+\eta_e/N_2)=(\xi-c)/N_1+(\eta_e+d)/N_2$. Therefore, for the remainder of the argument assume that c is even and d is odd.

The functional expression analogous to (31) is

$$\varphi\left(\frac{\xi}{N_1}+\frac{\eta_e}{N_2}\right)-\varphi\left(\frac{\xi}{N_1}+\frac{\eta_o}{N_2}\right)=\mu_1\left(\frac{\eta_e}{d}\right)-\mu_1\left(\frac{\eta_o}{d}\right),$$

which leads to

$$\varphi\left(\frac{\xi}{N_1}+\frac{\eta_e}{N_2}\right)-\varphi\left(\frac{\xi}{N_1}+\frac{\eta_o}{N_2}\right)=\varphi\left(\frac{\xi-2p}{N_1}+\frac{\eta_e}{N_2}\right)-\varphi\left(\frac{\xi-2p}{N_1}+\frac{\eta_o}{N_2}\right),$$

since $(\xi-2p)/N_1+\eta/N_2$ belongs to the interval $(-\omega-\varepsilon, -\omega+\varepsilon)$. Using the definition of φ, this is the equality:

$$\begin{aligned}
&\frac{1}{\lambda}\sum_{\nu=0}^{\lambda-1}\left[\tan\frac{\pi}{2}\frac{\xi}{\lambda c}+\tan\frac{\pi}{2}\frac{(\eta_e+2\nu N_2)}{\lambda d}\right]\\
&\quad+\frac{1}{\lambda}\sum_{\nu=0}^{\lambda-1}\left[\cot\frac{\pi}{2}\frac{\xi}{\lambda c}+\cot\frac{\pi}{2}\frac{(\eta_o+2\nu N_2)}{\lambda d}\right]\\
&=\frac{1}{\lambda}\sum_{\nu=0}^{\lambda-1}\left[\tan\frac{\pi}{2}\frac{(\xi-2p)}{\lambda c}+\tan\frac{\pi}{2}\frac{(\eta_e+2\nu N_2)}{\lambda d}\right]\\
&\quad+\frac{1}{\lambda}\sum_{\nu=0}^{\lambda-1}\left[\cot\frac{\pi}{2}\frac{(\xi-2p)}{\lambda c}+\cot\frac{\pi}{2}\frac{(\eta_o+2\nu N_2)}{\lambda d}\right].
\end{aligned}\tag{33}$$

This reduces to

$$\tan\frac{\pi}{2}\frac{\xi}{\lambda c}+\cot\frac{\pi}{2}\frac{\xi}{\lambda c}=\tan\frac{\pi}{2}\frac{(\xi-2p)}{\lambda c}+\cot\frac{\pi}{2}\frac{(\xi-2p)}{\lambda c}$$

and hence to $\sin(\pi\xi)/\lambda c=\sin(\pi(\xi-2p)/\lambda c)$. The latter equality is easily seen to be false. This concludes the proof of Theorem 3.1.

4. Nonuniqueness of minimal extrapolations on T. In this section we will show that the trapezoids $t_{\omega,\varepsilon}$ on the circle T having the M property are not unique minimal extrapolations from their intervals of constancy. The proof of Theorem 3 of [4] applies to show uniqueness of the $t_{\omega,\varepsilon}$ for the case of the line $\mathbb{R}$.

Let φ be the function defined by

$$\begin{aligned}
\varphi(x)&=\sin(\pi x/\varepsilon), \qquad -\varepsilon\leqq x\leqq\varepsilon,\\
&=0, \qquad \varepsilon<|x|\leqq\pi.
\end{aligned}$$

Now define

$$f(x)=\varphi(x-\omega)-\varphi(x+\omega).$$

It is clear that f is supported on the intervals $(-\omega-\varepsilon, -\omega+\varepsilon)$ and $(\omega-\varepsilon, \omega+\varepsilon)$. Furthermore, the nth Fourier coefficient of f is given by

$$\begin{aligned}
f\hat{}(n)&=-(2\varepsilon/(\varepsilon^2n^2-\pi^2))\sin\varepsilon n\sin\omega n, \qquad n\neq\pm\pi/\varepsilon,\\
&=0, \qquad n=\pm\pi/\varepsilon.
\end{aligned}$$

Because n assumes only discrete values, it is easy to check that for sufficiently small positive λ

$$\lambda|f\hat{}(n)|<|t^{\wedge}_{\omega,\varepsilon}(n)| \quad \text{for all } n.$$

Hence

$$\begin{aligned}
\operatorname{sign}(t^{\wedge}_{\omega,\varepsilon}(n)+\lambda f\hat{}(n))&=\operatorname{sign}(t^{\wedge}_{\omega,\varepsilon}(n))\\
&=\operatorname{sign}(\sin\omega n\sin\varepsilon n) \quad \text{for all } n.
\end{aligned}$$

Since $t_{\omega,\varepsilon}$ has the M property, by Proposition 1.1 there exists a pseudomeasure S with support disjoint from the intervals $(-\omega-\varepsilon, -\omega+\varepsilon)$ and $(\omega-\varepsilon, \omega+\varepsilon)$ such that

$$\begin{aligned}
S\hat{}(n)&=\operatorname{sign}(\sin\omega n\sin\varepsilon n) \quad \text{for } n\neq 0,\\
&=1 \quad \text{for } n=0.
\end{aligned}$$

Therefore,

$$\|t_{\omega,\varepsilon}\|_A = (S, t_{\omega,\varepsilon}) = (S, t_{\omega,\varepsilon} + \lambda f)$$
$$= \|t_{\omega,\varepsilon} + \lambda f\|_A.$$

Therefore, both $t_{\omega,\varepsilon}$ and $t_{\omega,\varepsilon} + \lambda f$ have the M property and, of course, have the same intervals of constancy.

REFERENCES

[1] G. BENKE AND B. WELLS, *Estimates for the stability of low-pass filters*, IEEE Trans. Acoustics Speech Signal Process., 33 (1985), pp. 98–105.

[2] A. BEURLING, *Sur les intégrales de Fourier absolument convergentes et leur application à une transformation fonctionnelle*, Neuvième congrès des Math. Scand., Helsinki, 1938, pp. 345–366.

[3] C. G. ESSEEN, *Fourier analysis of distribution functions*, Acta Math., 77 (1945), pp. 6–32.

[4] Y. DOMAR, *On the uniqueness of minimal extrapolations*, Ark. Math., (1959), pp. 19–29.

[5] H. DYM AND H. MCKEAN, *Fourier Series and Integrals*, Academic Press, New York, 1972.

[6] J.-P. KAHANE, *Séries de Fourier absolument convergentes*, Springer-Verlag, Berlin, New York, 1970.

[7] Y. KATZNELSON, *An Introduction to Harmonic Analysis*, Dover, New York, 1976.

[8] S. LANGE, *Algebra*, Addison-Wesley, Reading, MA, 1965.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 810-820, May 1991

016

ASYMPTOTIC EXPANSION OF A CLASS OF FERMI-DIRAC INTEGRALS*

J. BOERSMA† AND M. L. GLASSER‡

Abstract. A procedure is presented for obtaining the complete asymptotic expansion of a class of fractional integrals (of Riemann-Liouville type), in which the integrand contains the product of two derivatives of the Fermi-Dirac integral. The procedure uses two-sided Laplace transforms and Abelian asymptotics of the inverse Laplace transform. The fractional integrals considered arise in various problems from statistical mechanics and solid state physics.

Key words. Fermi-Dirac integral, asymptotic expansion, Riemann-Liouville fractional integral, Laplace transform, Abelian asymptotics

AMS(MOS) subject classifications. 41A60, 33A70, 44A10, 26A33, 82

1. Introduction. This paper is concerned with the asymptotic expansion, as $\eta \to \infty$, of the class of integrals

$$G_{\mu,p}^{(m,n)}(\eta) = \frac{1}{\Gamma(\mu)} \int_{-\infty}^{n} (\eta - t)^{\mu-1} F_p^{(m)}(t) F_p^{(n)}(t)\, dt \qquad (\mu > 0). \tag{1.1}$$

Here, m and n are nonnegative integers, and $F_p^{(m)}(t)$ denotes the mth derivative of the Fermi-Dirac integral $F_p(t)$ defined by [1]

$$F_p(t) = \frac{1}{\Gamma(p+1)} \int_0^{\infty} \frac{x^p\, dx}{1 + e^{x-t}} \qquad (p > -1). \tag{1.2}$$

The class of Riemann-Liouville fractional integrals (1.1) is important in a number of areas of statistical mechanics and solid state physics. Two examples are the exchange energy of a d-dimensional electron gas [6] ($\mu = (d-1)/2$, $p = -\frac{1}{2}$, $m = n = 0$) and the temperature-dependent gradient expansion coefficients for the interaction functional of an inhomogeneous electron gas [5] ($\mu = 2$, $p = -\frac{1}{2}$, $m = n = 2$).

In the special case $p = -\frac{1}{2}$, $m = n = 0$, the asymptotics of the integral (1.1) has been treated by Glasser and Boersma [6]. Their procedure, which uses the two-sided Laplace transform, is generalized in the present paper to accommodate the additional parameters p, m, and n. The Laplace transform method is explained in § 2, where it is also shown that the asymptotic analysis may be restricted to the case $m = n$. Let the Laplace transform of $[F_p^{(m)}(t)]^2$ be denoted by $g(s)$ (with transform variable s), then the asymptotic expansion of $G(\eta)$ as $\eta \to \infty$ can be found by applying Abelian asymptotics to the series expansion of $g(s)$ around $s = 0$. By starting from a suitable integral representation for $g(s)$ as derived in § 3, the expansion of $g(s)$ around $s = 0$ is determined in §§ 4 and 5. In the final section, § 6, the corresponding complete asymptotic expansion of $G(\eta)$, as given by (1.1), is presented.

2. Laplace transform method. Following the procedure of § 3 of [6], we first determine the two-sided Laplace transform of (1.1):

$$\bar{G}_{\mu,p}^{(m,n)}(s) = \int_{-\infty}^{\infty} e^{-s\eta} G_{\mu,p}^{(m,n)}(\eta)\, d\eta = s^{-\mu} g_p^{(m,n)}(s) \tag{2.1}$$

* Received by the editors February 26, 1990; accepted for publication May 1, 1990.

† Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, the Netherlands.

‡ School of Science, Clarkson University, Potsdam, New York 13676.

where

$$g_p^{(m,n)}(s) = \int_{-\infty}^{\infty} e^{-st} F_p^{(m)}(t) F_p^{(n)}(t)\, dt. \tag{2.2}$$

From the known asymptotic behaviour [1]

$$F_p(t) = O(e^t), \quad t \to -\infty, \qquad F_p(t) = O(t^{p+1}), \quad t \to \infty, \tag{2.3}$$

it follows that the Laplace transforms $\bar{G}_{\mu,p}^{(m,n)}(s)$ and $g_p^{(m,n)}(s)$ are defined in the strip $0 < \operatorname{Re} s < 2$.

Assuming that $m \leqq n$ in (2.2) and integrating by parts, we establish the recurrence relations

$$g_p^{(m,m+1)}(s) = \tfrac{1}{2} s g_p^{(m,m)}(s), \tag{2.4}$$

$$g_p^{(m,n)}(s) = s g_p^{(m,n-1)}(s) - g_p^{(m+1,n-1)}(s), \qquad n \geqq m+2. \tag{2.5}$$

By repeated application of these relations we are led to

$$\begin{aligned} g_p^{(m,m+2)}(s) &= \tfrac{1}{2} s^2 g_p^{(m,m)}(s) - g_p^{(m+1,m+1)}(s), \\ g_p^{(m,m+3)}(s) &= \tfrac{1}{2} s^3 g_p^{(m,m)}(s) - \tfrac{3}{2} s g_p^{(m+1,m+1)}(s), \\ g_p^{(m,m+4)}(s) &= \tfrac{1}{2} s^4 g_p^{(m,m)}(s) - 2 s^2 g_p^{(m+1,m+1)}(s) + g_p^{(m+2,m+2)}(s). \end{aligned} \tag{2.6}$$

The coefficients in (2.4) and (2.6) are now used to form the polynomials

$$\begin{aligned} &p_0(s) = 1, \quad p_1(s) = \tfrac{1}{2} s, \quad p_2(s) = \tfrac{1}{2} s^2 - 1, \\ &p_3(s) = \tfrac{1}{2} s^3 - \tfrac{3}{2} s, \qquad p_4(s) = \tfrac{1}{2} s^4 - 2 s^2 + 1, \cdots, \end{aligned} \tag{2.7}$$

which, by (2.5), satisfy the recurrence relation

$$p_k(s) = s p_{k-1}(s) - p_{k-2}(s), \qquad k \geqq 2. \tag{2.8}$$

The latter recurrence relation is identical to that of the Chebyshev polynomials $T_k(s/2)$ (cf. [3, § 10.11]). Thus we find

$$p_k(s) = T_k\left(\frac{s}{2}\right) = \frac{1}{2} k \sum_{l=0}^{[k/2]} \frac{(-1)^l (k-l-1)!}{l!(k-2l)!} s^{k-2l}, \qquad k \geqq 1, \tag{2.9}$$

whereupon the results in (2.6) generalize to

$$g_p^{(m,m+k)}(s) = \frac{1}{2} k \sum_{l=0}^{[k/2]} \frac{(-1)^l (k-l-1)!}{l!(k-2l)!} s^{k-2l} g_p^{(m+l,m+l)}(s), \qquad k \geqq 1. \tag{2.10}$$

Consequently, without loss of generality we can restrict our further asymptotic analysis to the case $m = n$. Accordingly, we simplify the notation by setting $G_{\mu,p}^{(m,m)}(\eta) \equiv G_{\mu,p}^{(m)}(\eta)$ and $g_p^{(m,m)}(s) \equiv g_p^{(m)}(s)$.

In the Laplace transform method the asymptotic expansion of $G_{\mu,p}^{(m)}(\eta)$ as $\eta \to \infty$ is obtained by applying Abelian asymptotics [2, Kap. 7] to the series expansion of $g_p^{(m)}(s)$ around $s = 0$. To determine the latter expansion, we rewrite the integral (2.2) in a more convenient form by means of Parseval's formula:

$$g_p^{(m)}(s) = \int_{-\infty}^{\infty} e^{-st} [F_p^{(m)}(t)]^2\, dt = \int_{-\infty}^{\infty} f(u) f(-u)\, du \tag{2.11}$$

where

$$f(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-st/2} F_p^{(m)}(t)\, e^{iut}\, dt = \frac{(s/2 - iu)^m}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-st/2} F_p(t)\, e^{iut}\, dt. \tag{2.12}$$

The second integral in (2.12) is evaluated by inserting the integral representation (1.2) for $F_p(t)$, interchanging the order of integration, and applying the substitution $y = e^{x-t}$ in the t-integral:

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-st/2} F_p(t)\, e^{iut}\, dt &= \frac{1}{\Gamma(p+1)} \int_0^{\infty} x^p\, e^{-(s/2-iu)x}\, dx \int_0^{\infty} \frac{y^{s/2-iu-1}}{1+y}\, dy \\ &= \frac{\pi(s/2-iu)^{-p-1}}{\sin[\pi(s/2-iu)]}.\end{aligned} \tag{2.13}$$

The result for $f(u)$ thus found is inserted into (2.11), and we have

$$g_p^{(m)}(s) = \pi \int_{-\infty}^{\infty} \frac{(s^2/4+u^2)^{m-p-1}}{\cosh(2\pi u) - \cos(\pi s)}\, du. \tag{2.14}$$

Obviously, $g_p^{(m)}(s)$ depends only on the difference $m-p$; this was to be expected from the basic recursion formula $F_p'(t) = F_{p-1}(t)$. Finally, for brevity we introduce

$$\nu = m - p - \tfrac{1}{2}, \qquad g_\nu(s) = g_p^{(m)}(s); \tag{2.15}$$

then the representation (2.14) becomes

$$g_\nu(s) = \pi \int_{-\infty}^{\infty} \frac{(s^2/4+u^2)^{\nu-1/2}}{\cosh(2\pi u) - \cos(\pi s)}\, du. \tag{2.16}$$

3. Integral representation for $g_\nu(s)$. The representation (2.16) for $g_\nu(s)$ is further reduced by another application of Parseval's formula. It is convenient to distinguish three cases.

Case i. $\nu < \frac{1}{2}$. From [4, Formulas 1.9(6), 1.3(7)] we quote the Fourier cosine transforms

$$\int_0^{\infty} \frac{\cos(xu)}{\cosh(2\pi u) - \cos(\pi s)}\, du = \frac{1}{2\sin(\pi s)} \frac{\sinh[(1-s)x/2]}{\sinh(x/2)}, \tag{3.1}$$

$$\int_0^{\infty} \left(\frac{s^2}{4}+u^2\right)^{\nu-1/2} \cos(xu)\, du = \frac{\pi^{1/2}}{\Gamma(\frac{1}{2}-\nu)} \left(\frac{x}{s}\right)^{-\nu} K_\nu\left(\frac{sx}{2}\right) \tag{3.2}$$

where we used that $K_{-\nu}(z) = K_\nu(z)$ by [7, Formula 3.71(8)]. Next, by means of Parseval's formula applied to (2.16) we are led to

$$g_\nu(s) = \frac{2\pi^{1/2}}{\Gamma(\frac{1}{2}-\nu)} \frac{s^\nu}{\sin(\pi s)} \int_0^{\infty} \frac{\sinh[(1-s)x/2]}{\sinh(x/2)}\, x^{-\nu} K_\nu\left(\frac{sx}{2}\right) dx. \tag{3.3}$$

It is easily seen that the integral (3.3) is convergent if $\nu < \frac{1}{2}$.

Case ii. $\nu > \frac{1}{2}$, $\nu - \frac{1}{2} \notin \mathbb{N}$. Let k be the smallest integer greater than or equal to ν; then we set $\nu = k - q$, where $0 \leqq q < 1$ and $q \neq \frac{1}{2}$. To apply Parseval's formula in (2.16), we need the Fourier cosine transform

$$\int_0^{\infty} \frac{(s^2/4+u^2)^k}{\cosh(2\pi u) - \cos(\pi s)} \cos(xu)\, du = \frac{1}{2\sin(\pi s)} \left(\frac{s^2}{4} - \frac{d^2}{dx^2}\right)^k \left\{\frac{\sinh[(1-s)x/2]}{\sinh(x/2)}\right\} \tag{3.4}$$

obtainable from (3.1), and the transform (3.2) with ν replaced by $-q$. As a result it is found that the representation (2.16) passes into

$$g_\nu(s) = \frac{2\pi^{1/2}}{\Gamma(q+\frac{1}{2})} \frac{s^{-q}}{\sin(\pi s)} \int_0^{\infty} \left(\frac{s^2}{4} - \frac{d^2}{dx^2}\right)^k \left\{\frac{\sinh[(1-s)x/2]}{\sinh(x/2)}\right\} x^q K_{-q}\left(\frac{sx}{2}\right) dx. \tag{3.5}$$

To further reduce (3.5), we would like to integrate by parts so that the differential operator acts on $x^q K_{-q}(sx/2)$. Here a difficulty comes up, since the resulting integral is divergent at the lower limit $x=0$ and the intermediate endpoint contributions at $x=0$ become infinite. To overcome this, we introduce the "finite part" (in the sense of Hadamard) of the resulting integral and end point contributions, defined as follows.

For $\delta \geqq 0$, let $f(\delta)$ have an asymptotic expansion as $\delta \downarrow 0$, that consists of terms $\delta^r(\log \delta)^j$ with real r and integer j. Suppose the expansion contains a finite number of singular terms (i.e., terms with $r<0$ or with $r=0$, $j \geqq 1$), and let $f_\infty(\delta)$ denote the sum of the singular terms. Then we define the finite part of $f(\delta)$ as $\delta \downarrow 0$ by

$$\operatorname*{fin}_{\delta \downarrow 0} f(\delta) = \lim_{\delta \downarrow 0} [f(\delta) - f_\infty(\delta)]. \tag{3.6}$$

Likewise, if $\int_0^\infty h(x)\,dx$ is divergent or convergent at $x=0$, we define the finite part of the integral as

$$\fint_0^\infty h(x)\,dx = \operatorname*{fin}_{\delta \downarrow 0} \int_\delta^\infty h(x)\,dx. \tag{3.7}$$

When integrating by parts in (3.5), the finite part of a typical endpoint contribution looks like

$$\operatorname*{fin}_{\delta \downarrow 0} \left(\frac{d}{dx}\right)^j \left\{\frac{\sinh[(1-s)x/2]}{\sinh(x/2)}\right\} \left(\frac{d}{dx}\right)^l \left\{x^q K_{-q}\left(\frac{sx}{2}\right)\right\}\Bigg|_{x=\delta} \tag{3.8}$$

where j and l are nonnegative integers with $j+l$ odd. We expand this in a power series in powers of $x=\delta$. Then the expansion is found to contain terms δ^{2n-j-l}, $\delta^{2n+2q-j-l}$ and, if $q=0$, also $\delta^{2n-j-l}\log\delta$, whereby $n=0,1,2,\cdots$. Because $q \neq \frac{1}{2}$ and $j+l$ is odd, none of the exponents $2n-j-l$ or $2n+2q-j-l$ is zero and the finite part (3.8) vanishes. In this way we find, through integration by parts in (3.5), that

$$g_\nu(s) = \frac{2\pi^{1/2}}{\Gamma(q+\frac{1}{2})} \frac{s^{-q}}{\sin(\pi s)} \fint_0^\infty \frac{\sinh[(1-s)x/2]}{\sinh(x/2)} \left(\frac{s^2}{4} - \frac{d^2}{dx^2}\right)^k \left\{x^q K_{-q}\left(\frac{sx}{2}\right)\right\} dx. \tag{3.9}$$

Setting $t=sx/2$, by repeated use of the recurrence formula [7, § 3.71]

$$\left(1-\frac{d^2}{dt^2}\right)\frac{K_\nu(t)}{t^\nu} = \frac{2\nu+1}{t}\frac{d}{dt}\left(\frac{K_\nu(t)}{t^\nu}\right) = -(2\nu+1)\frac{K_{\nu+1}(t)}{t^{\nu+1}} \tag{3.10}$$

we find

$$\left(\frac{s^2}{4} - \frac{d^2}{dx^2}\right)^k \left\{x^q K_{-q}\left(\frac{sx}{2}\right)\right\} = \frac{\Gamma(q+\frac{1}{2})}{\Gamma(q-k+\frac{1}{2})} s^k x^{q-k} K_{k-q}\left(\frac{sx}{2}\right). \tag{3.11}$$

Inserting (3.11) into (3.9) and restoring the notation $\nu = k-q$, we are led to the integral representation

$$g_\nu(s) = \frac{2\pi^{1/2}}{\Gamma(\frac{1}{2}-\nu)} \frac{s^\nu}{\sin(\pi s)} \fint_0^\infty \frac{\sinh[(1-s)x/2]}{\sinh(x/2)} x^{-\nu} K_\nu\left(\frac{sx}{2}\right) dx. \tag{3.12}$$

This result is identical to the corresponding representation (3.3) for Case i, except that now the finite part of the divergent integral is to be taken (as indicated by the notation $\fint$). From the original representation (2.16) it is clear that $g_\nu(s)$ is an analytic function of ν in the whole complex ν-plane. Therefore the finite part integral (3.12) is also the analytic continuation of the integral (3.3) which is analytic for $\operatorname{Re}\nu < \frac{1}{2}$.

Case iii. $\nu = n+\frac{1}{2}$, $n = 0, 1, 2, \cdots$. In this case the integral (2.16) can be evaluated by means of (3.4), viz.

$$g_{n+1/2}(s) = \pi \int_{-\infty}^{\infty} \frac{(s^2/4+u^2)^n}{\cosh(2\pi u)-\cos(\pi s)}\,du = \frac{\pi}{\sin(\pi s)}\left(\frac{s^2}{4}-\frac{d^2}{dx^2}\right)^n \left\{\frac{\sinh[(1-s)x/2]}{\sinh(x/2)}\right\}\Bigg|_{x=0}. \tag{3.13}$$

Thus for $n=0$, $\nu=\frac{1}{2}$, we have

$$g_{1/2}(s) = \frac{\pi}{\sin(\pi s)}(1-s). \tag{3.14}$$

To evaluate the derivative in (3.13), we substitute

$$\frac{\sinh[(1-s)x/2]}{\sinh(x/2)} = \frac{e^{(1-s/2)x}-e^{sx/2}}{e^x-1} = -2\sum_{k=0}^{\infty} B_{2k+1}\left(\frac{s}{2}\right)\frac{x^{2k}}{(2k+1)!} \tag{3.15}$$

where $B_{2k+1}(s/2)$ is the Bernoulli polynomial [3, § 1.13]. Then we find

$$g_{n+1/2}(s) = -\frac{2\pi}{\sin(\pi s)}\sum_{k=0}^{n}(-1)^k\binom{n}{k}\left(\frac{s^2}{4}\right)^{n-k}\frac{B_{2k+1}(s/2)}{2k+1}. \tag{3.16}$$

4. Expansion of $g_\nu(s)$ if $2\nu \notin \mathbb{Z}$. To determine the series expansion of $g_\nu(s)$ around $s=0$, we start from the integral representation (3.12) which includes the representation (3.3) as a special case. For convenience it is assumed that 2ν is not integral. By substitution of

$$\frac{\sinh[(1-s)x/2]}{\sinh(x/2)} = e^{-sx/2} - 2\sinh\left(\frac{sx}{2}\right)\frac{e^{-x}}{1-e^{-x}} \tag{4.1}$$

the representation (3.12) is rewritten as

$$g_\nu(s) = \frac{2\pi^{1/2}}{\Gamma(\frac{1}{2}-\nu)}\frac{s^\nu}{\sin(\pi s)}\int_0^\infty e^{-sx/2}x^{-\nu}K_\nu\left(\frac{sx}{2}\right)dx - \frac{4\pi^{1/2}}{\Gamma(\frac{1}{2}-\nu)}\frac{s^\nu}{\sin(\pi s)}\int_0^\infty \sinh\left(\frac{sx}{2}\right)K_\nu\left(\frac{sx}{2}\right)\frac{x^{-\nu}e^{-x}}{1-e^{-x}}\,dx. \tag{4.2}$$

From [4, Formula 6.8(28)] we have

$$\int_0^\infty e^{-sx/2}x^{-\nu}K_\nu\left(\frac{sx}{2}\right)dx = \frac{\pi^{1/2}\Gamma(1-2\nu)}{\Gamma(\frac{3}{2}-\nu)}s^{\nu-1}, \tag{4.3}$$

valid for $\operatorname{Re}\nu < \frac{1}{2}$. By analytic continuation the result (4.3) also holds for $\operatorname{Re}\nu \geqq \frac{1}{2}$, $2\nu \notin \mathbb{N}$, provided that the finite part of the integral is taken as in (4.2). To evaluate the second integral in (4.2), we expand the product $\sinh(sx/2)K_\nu(sx/2)$ in a power series. Starting from the definition

$$K_\nu(z) = \frac{\pi}{2\sin(\nu\pi)}[I_{-\nu}(z)-I_\nu(z)] \qquad (\nu \notin \mathbb{Z}) \tag{4.4}$$

we employ Watson's expansion [7, Formula 5.41(1)] for the products $J_\mu(z)J_{\pm\nu}(z)$ with $\mu=\frac{1}{2}$, $z=isx/2$. As a result, we obtain

$$\sinh\left(\frac{sx}{2}\right)K_\nu\left(\frac{sx}{2}\right) = \frac{\pi^{1/2}}{2\sin(\nu\pi)}\left[\sum_{k=0}^{\infty}\frac{\Gamma(2k-\nu+\frac{3}{2})}{(2k+1)!\Gamma(2k-2\nu+2)}(sx)^{2k-\nu+1} - \sum_{k=0}^{\infty}\frac{\Gamma(2k+\nu+\frac{3}{2})}{(2k+1)!\Gamma(2k+2\nu+2)}(sx)^{2k+\nu+1}\right]. \tag{4.5}$$

The latter expansion is inserted into the second integral in (4.2) and we apply a term-by-term integration using the auxiliary integral

$$\int_0^\infty x^\alpha \frac{e^{-x}}{1-e^{-x}}\,dx = \Gamma(\alpha+1)\zeta(\alpha+1) \qquad (\mathrm{Re}\,\alpha > 0) \tag{4.6}$$

where $\zeta(\alpha+1)$ denotes Riemann's zeta function [3, § 1.12]. By analytic continuation the result (4.6) also holds for $\mathrm{Re}\,\alpha \leqq 0$, $\alpha \notin \mathbb{Z}$, provided that the finite part of the integral is taken.

Finally, by compiling the previous results we are led to the desired expansion

$$g_\nu(s) = \frac{2\pi}{\Gamma(\frac{1}{2}-\nu)\sin(\nu\pi)}\frac{s}{\sin(\pi s)}\left[\sum_{k=0}^{\infty}\frac{\Gamma(2k+\nu-\frac{1}{2})}{\Gamma(2k+2\nu)}\zeta(2k)s^{2k+2\nu-2} - \sum_{k=0}^{\infty}\frac{\Gamma(2k-\nu+\frac{3}{2})}{(2k+1)!}\zeta(2k-2\nu+2)s^{2k}\right], \tag{4.7}$$

valid if $2\nu \notin \mathbb{Z}$. It is readily seen that the expansion (4.7) is convergent for $0<|s|<1$.

5. Expansion of $g_\nu(s)$ if $2\nu \in \mathbb{Z}$. Since $g_\nu(s)$ is a continuous function of the parameter ν, the series expansion of $g_\nu(s)$ when $2\nu = N \in \mathbb{Z}$ can be found by taking limits in (4.7) as $\nu \to N/2$. We distinguish four cases.

Case i. $\nu = n$, $n = 1, 2, 3, \cdots$. Rewrite the expansion (4.7) as

$$g_\nu(s) = \frac{2\pi}{\Gamma(\frac{1}{2}-\nu)}\frac{s}{\sin(\pi s)}\frac{1}{\sin(\nu\pi)} \cdot\left[-\sum_{k=0}^{n-2}\frac{\Gamma(2k-\nu+\frac{3}{2})}{(2k+1)!}\zeta(2k-2\nu+2)s^{2k} + \sum_{k=n-1}^{\infty}\left\{\frac{\Gamma(2k-2n+\nu+\frac{3}{2})}{\Gamma(2k-2n+2\nu+2)}\zeta(2k-2n+2)s^{2k-2n+2\nu} - \frac{\Gamma(2k-\nu+\frac{3}{2})}{(2k+1)!}\zeta(2k-2\nu+2)s^{2k}\right\}\right] \tag{5.1}$$

where it is noted that the terms of the finite sum and of the infinite series vanish when $\nu = n$. By properly taking limits as $\nu \to n$, the expansion (5.1) passes into

$$g_n(s) = \frac{4}{\pi}\Gamma\left(n+\frac{1}{2}\right)\frac{s}{\sin(\pi s)} \cdot\left[\sum_{k=0}^{n-2}\frac{\Gamma(2k-n+\frac{3}{2})}{(2k+1)!}\zeta'(2k-2n+2)s^{2k} + \sum_{k=n-1}^{\infty}\frac{\Gamma(2k-n+\frac{3}{2})}{(2k+1)!}\zeta(2k-2n+2)s^{2k} \cdot\left\{\log s + \psi\left(2k-n+\frac{3}{2}\right) - \psi(2k+2) + \frac{\zeta'(2k-2n+2)}{\zeta(2k-2n+2)}\right\}\right], \tag{5.2}$$

valid for $n = 1, 2, 3, \cdots$; here, $\psi(z)$ denotes the logarithmic derivative of the Γ-function, i.e., $\psi(z) = \Gamma'(z)/\Gamma(z)$. By means of [3, Formula 1.12(23)] we have

$$\zeta'(2k-2n+2) = (-1)^{n-k-1}\frac{(2n-2k-2)!}{2(2\pi)^{2n-2k-2}}\zeta(2n-2k-1) \qquad (k = 0, 1, \cdots, n-2), \tag{5.3}$$

which is used in the finite sum in (5.2).

Case ii. $\nu=-n$, $n=0,1,2,\cdots$. Rewrite the expansion (4.7) as

$$
\begin{aligned}
g_\nu(s)=&\frac{2\pi}{\Gamma(\frac12-\nu)}\,\frac{s}{\sin(\pi s)}\,\frac{1}{\sin(\nu\pi)}\\
&\cdot\Bigg[\sum_{k=0}^{n}\frac{\Gamma(2k+\nu-\frac12)}{\Gamma(2k+2\nu)}\zeta(2k)s^{2k+2\nu-2}\\
&\quad+\sum_{k=0}^{\infty}\Bigg\{\frac{\Gamma(2k+2n+\nu+\frac32)}{\Gamma(2k+2n+2\nu+2)}\zeta(2k+2n+2)s^{2k+2n+2\nu}\\
&\qquad-\frac{\Gamma(2k-\nu+\frac32)}{(2k+1)!}\zeta(2k-2\nu+2)s^{2k}\Bigg\}\Bigg]
\end{aligned}
\tag{5.4}
$$

and take limits as $\nu\to-n$. Then, as in Case i, we are led to the expansion

$$
\begin{aligned}
g_{-n}(s)=&\frac{4(-1)^n}{\Gamma(n+\frac12)}\,\frac{s}{\sin(\pi s)}\Bigg[\sum_{k=0}^{n}\Gamma\left(2k-n-\frac12\right)(2n-2k)!\,\zeta(2k)s^{2k-2n-2}\\
&+\sum_{k=0}^{\infty}\frac{\Gamma(2k+n+\frac32)}{(2k+1)!}\zeta(2k+2n+2)s^{2k}\\
&\cdot\left\{\log s+\psi\left(2k+n+\frac32\right)-\psi(2k+2)+\frac{\zeta'(2k+2n+2)}{\zeta(2k+2n+2)}\right\}\Bigg],
\end{aligned}
\tag{5.5}
$$

valid for $n=0,1,2,\cdots$. For $n=0$, the expansion (5.5) agrees with [6, Formula (33)]. Note that the expansion of $g_\nu(s)$ contains logarithmic terms in the case where ν is integral.

Case iii. $\nu=n+\frac12$, $n=0,1,2,\cdots$. When taking limits in (4.7) as $\nu\to n+\frac12$, proper care should be taken because some of the Γ- and ζ-functions become singular. First consider the case $\nu=\frac12$; then we find

$$
\begin{aligned}
g_{1/2}(s)=&\frac{2\pi s}{\sin(\pi s)}\lim_{\nu\to1/2}\frac{1}{\Gamma(\frac12-\nu)}\left[\frac{\Gamma(\nu-\frac12)}{\Gamma(2\nu)}\zeta(0)s^{2\nu-2}-\Gamma\left(\frac32-\nu\right)\zeta(2-2\nu)\right]\\
=&\frac{\pi s}{\sin(\pi s)}\left(\frac1s-1\right),
\end{aligned}
\tag{5.6}
$$

in accordance with (3.14). Generally, for $\nu=n+\frac12$, $n\geqq1$, the expansion (4.7) passes into

$$
\begin{aligned}
g_{n+1/2}(s)=&(-1)^n\frac{2\pi s}{\sin(\pi s)}\\
&\cdot\lim_{\nu\to n+1/2}\frac{1}{\Gamma(\frac12-\nu)}\Bigg[-\sum_{k=0}^{[(n-1)/2]}\frac{\Gamma(2k-\nu+\frac32)}{(2k+1)!}\zeta(2k-2\nu+2)s^{2k}\\
&\qquad-\frac{\Gamma(2n-\nu+\frac32)}{(2n+1)!}\zeta(2n-2\nu+2)s^{2n}\Bigg]\\
=&\frac{2\pi s}{\sin(\pi s)}\left[(-1)^n\sum_{k=0}^{[(n-1)/2]}\binom{n}{2k+1}\zeta(2k-2n+1)s^{2k}-\frac{(n!)^2}{2(2n+1)!}s^{2n}\right],
\end{aligned}
\tag{5.7}
$$

valid for $n=1,2,3,\cdots$. In the final line of (5.7) we may set, by [3, Formula 1.12(22)],

$$
\zeta(2k-2n+1)=-\frac{B_{2n-2k}}{2(n-k)}\qquad(k=0,1,\cdots,n-1)
\tag{5.8}
$$

where B_{2n-2k} is the Bernoulli number. It can be shown that the expansion (5.7) agrees with (3.16).

Case iv. $\nu = -n - \frac{1}{2}$, $n = 0, 1, 2, \cdots$. In this case the expansion (4.7) remains valid, provided that the ratio $\Gamma(2k+\nu-\frac{1}{2})/\Gamma(2k+2\nu)$ is handled with proper care. Thus by means of

$$\lim_{\nu \to -n-1/2} \frac{\Gamma(2k+\nu-\frac{1}{2})}{\Gamma(2k+2\nu)} = \begin{cases} 2(2k-2n-1)_n, & k \leqq n, \\ (2k-2n-1)_n, & k \geqq n+1 \end{cases} \tag{5.9}$$

we obtain the expansion

$$g_{-n-1/2}(s) = \frac{(-1)^n}{n!} \frac{2\pi s}{\sin(\pi s)} \left[-2 \sum_{k=0}^{n} (2k-2n-1)_n \zeta(2k) s^{2k-2n-3} + \sum_{k=1}^{\infty} (-1)^k (k)_n \zeta(k+2n+1) s^{k-2} \right], \tag{5.10}$$

valid for $n = 0, 1, 2, \cdots$. In the special case $n = 0$, $\nu = -\frac{1}{2}$, the expansion (5.10) reduces to

$$\begin{aligned} g_{-1/2}(s) &= \frac{2\pi s}{\sin(\pi s)} \left[s^{-3} + \sum_{k=1}^{\infty} (-1)^k \zeta(k+1) s^{k-2} \right] \\ &= -\frac{2\pi s^{-1}}{\sin(\pi s)} [\psi(s) + \gamma] \end{aligned} \tag{5.11}$$

by use of [3, Formula 1.17(5)]. The same result can also be found by a direct evaluation of the integral (3.3) with $\nu = -\frac{1}{2}$.

Finally, it is pointed out that the infinite series expansions of $g_\nu(s)$, as presented in (5.2), (5.5), and (5.10), are convergent for $0 < |s| < 1$. In Case iii the infinite series reduces to a finite sum; see (5.6) and (5.7).

6. Asymptotic expansion of $G^{(m)}_{\mu,p}(\eta)$. The asymptotic expansion of $G^{(m)}_{\mu,p}(\eta)$ as $\eta \to \infty$ is determined through a term-by-term conversion, based on theorems of Abelian asymptotics [2, Kap. 7], of the series expansion of $s^{-\mu} g_\nu(s)$ around $s = 0$. The conversion is most easily carried out by use of the "dictionary" in Table 1. The left column of the table shows a specific term of the expansion around $s = 0$; the right column shows the corresponding term of the asymptotic expansion as $\eta \to \infty$.

In the expansions of $g_\nu(s)$ as determined in §§ 4 and 5, replace $\pi s/\sin(\pi s)$ by

$$\frac{\pi s}{\sin(\pi s)} = 2 \sum_{k=0}^{\infty} (1 - 2^{1-2k}) \zeta(2k) s^{2k}, \qquad |s| < 1 \tag{6.1}$$

TABLE 1
Inverse Laplace transforms.

$f(s)$	$(1/2\pi i) \int_{c-i\infty}^{c+i\infty} f(s)\, e^{\eta s}\, ds$	
s^λ	$[1/\Gamma(-\lambda)]\eta^{-\lambda-1}$,	$\lambda \neq 0, 1, 2, \cdots$
	0,	$\lambda = 0, 1, 2, \cdots$
$s^\lambda \log s$	$-[1/\Gamma(-\lambda)]\eta^{-\lambda-1}[\log \eta - \psi(-\lambda)]$,	$\lambda \neq 0, 1, 2, \cdots$
	$(-1)^{\lambda+1}\lambda!\,\eta^{-\lambda-1}$,	$\lambda = 0, 1, 2, \cdots$

and multiply the series involved. Then for $2\nu \notin \mathbb{Z}$, the expansion (4.7) of $g_\nu(s)$ takes the form

$$g_\nu(s) = \sum_{k=0}^{\infty} A_k s^{2k+2\nu-2} + \sum_{k=0}^{\infty} B_k s^{2k}, \qquad 0<|s|<1 \tag{6.2}$$

with coefficients

$$\begin{aligned} A_k &= \frac{4}{\Gamma(\frac{1}{2}-\nu)\sin(\nu\pi)} \sum_{l=0}^{k} \frac{\Gamma(2l+\nu-\frac{1}{2})}{\Gamma(2l+2\nu)} \zeta(2l)(1-2^{1-2k+2l})\zeta(2k-2l), \\ B_k &= \frac{-4}{\Gamma(\frac{1}{2}-\nu)\sin(\nu\pi)} \sum_{l=0}^{k} \frac{\Gamma(2l-\nu+\frac{3}{2})}{(2l+1)!} \zeta(2l-2\nu+2)(1-2^{1-2k+2l})\zeta(2k-2l). \end{aligned} \tag{6.3}$$

Similar expansions hold in the case where $2\nu \in \mathbb{Z}$. From (5.2) and (5.5) it follows that the expansion of $g_\nu(s)$ contains logarithmic terms if ν is integral.

Starting from (6.2) multiplied by $s^{-\mu}$, we find by use of Table 1 the desired asymptotic expansion

$$\begin{aligned} G_{\mu,p}^{(m)}(\eta) \sim & \sum_{k=0}^{\infty} \frac{A_k}{\Gamma(\mu-2\nu-2k+2)} \eta^{\mu-2\nu-2k+1} \\ & + \sum_{k=0}^{\infty} \frac{B_k}{\Gamma(\mu-2k)} \eta^{\mu-2k-1} \qquad (\eta \to \infty), \end{aligned} \tag{6.4}$$

valid if $2\nu \notin \mathbb{Z}$. It is pointed out that the first (second) asymptotic series in (6.4) terminates to a finite sum if $\mu-2\nu$ (μ) is an integer. Similar asymptotic expansions hold in the case where $2\nu \in \mathbb{Z}$. If ν is an integer, it is found from (5.2) and (5.5) that the asymptotic expansion of $G_{\mu,p}^{(m)}(\eta)$ contains logarithmic terms $[1/\Gamma(\mu-2k)] \times \eta^{\mu-2k-1} \log \eta$, with $k \geqq \max(\nu-1, 0)$.

As an example, we determine the asymptotic expansion of the integral [5]

$$G_{2,-1/2}^{(2)}(\eta) = \int_{-\infty}^{\eta} (\eta - t)[F''_{-1/2}(t)]^2 \, dt, \tag{6.5}$$

for which $\nu = 2$. In the expansion (5.2) with $n=2$, replace $\pi s/\sin(\pi s)$ by (6.1), and multiply the series involved. Then the expansion of $g_2(s)$ takes the form

$$g_2(s) = \sum_{k=1}^{\infty} c_k s^{2k} \log s + \sum_{k=0}^{\infty} d_k s^{2k}, \qquad |s|<1 \tag{6.6}$$

with coefficients

$$\begin{aligned} c_k &= 6\pi^{-3/2} \sum_{l=1}^{k} \frac{\Gamma(2l-\frac{1}{2})}{(2l+1)!} \zeta(2l-2)(1-2^{1-2k+2l})\zeta(2k-2l), \\ d_k &= 6\pi^{-3/2} \sum_{l=0}^{k} \frac{\Gamma(2l-\frac{1}{2})}{(2l+1)!} \zeta(2l-2)\left[\psi\left(2l-\frac{1}{2}\right) - \psi(2l+2) + \frac{\zeta'(2l-2)}{\zeta(2l-2)}\right] \\ &\quad \cdot (1-2^{1-2k+2l})\zeta(2k-2l). \end{aligned} \tag{6.7}$$

Next, by use of Table 1 in a term-by-term conversion of the expansion of $s^{-2}g_2(s)$, we are led to the asymptotic expansion

$$G_{2,-1/2}^{(2)}(\eta) \sim d_0\eta - \sum_{k=1}^{\infty} (2k-2)! c_k \eta^{-2k+1} \qquad (\eta \to \infty). \tag{6.8}$$

By evaluating the leading terms in (6.8), we find

$$G^{(2)}_{2,-1/2}(\eta)=\frac{3\zeta(3)}{2\pi^3}\eta+\frac{1}{8\pi}\eta^{-1}+\frac{5\pi}{192}\eta^{-3}$$
$$+\frac{43\pi^3}{1920}\eta^{-5}+\frac{323\pi^5}{7168}\eta^{-7}+O(\eta^{-9}) \qquad (\eta\to\infty). \tag{6.9}$$

The asymptotic expansion (6.8) can also be derived in a more elementary manner. To that end we start from the two-sided Laplace transform

$$\int_{-\infty}^{\infty} e^{-st}F''_{-1/2}(t)\,dt=\frac{\pi s^{3/2}}{\sin(\pi s)}, \tag{6.10}$$

obtainable from (2.12) and (2.13). Using (6.1) and Table 1, we expand (6.10) in a power series around $s=0$, whereupon a term-by-term conversion yields the asymptotic expansion

$$F''_{-1/2}(t)\sim-\frac{2}{\pi}\sum_{k=0}^{\infty}(1-2^{1-2k})\Gamma\left(2k+\frac{3}{2}\right)\zeta(2k)t^{-2k-3/2} \qquad (t\to\infty). \tag{6.11}$$

By squaring (6.11) we find

$$[F''_{-1/2}(t)]^2\sim\sum_{k=0}^{\infty}b_k t^{-2k-3} \quad (t\to\infty) \tag{6.12}$$

with coefficients

$$b_k=\frac{4}{\pi^2}\sum_{l=0}^{k}(1-2^{1-2l})\Gamma\left(2l+\frac{3}{2}\right)\zeta(2l)$$
$$\cdot(1-2^{1-2k+2l})\Gamma\left(2k-2l+\frac{3}{2}\right)\zeta(2k-2l). \tag{6.13}$$

Next it is observed from (6.5) that $G^{(2)}_{2,-1/2}(\eta)$ is the repeated integral of order 2, of $[F''_{-1/2}(t)]^2$. As it has been shown in the Appendix of [6], the asymptotic expansion of $G^{(2)}_{2,-1/2}(\eta)$ can now be derived by a twice repeated termwise integration of the expansion (6.12). Thus we find

$$G^{(2)}_{2,-1/2}(\eta)\sim C_1\eta+C_0+\sum_{k=0}^{\infty}\frac{b_k}{(2k+1)(2k+2)}\eta^{-2k-1} \qquad (\eta\to\infty) \tag{6.14}$$

where the constants C_0 and C_1 are yet to be determined. By dividing (6.14) by η and taking limits as $\eta\to\infty$, it readily follows that

$$C_1=\int_{-\infty}^{\infty}[F''_{-1/2}(t)]^2\,dt=g_2(0)=\frac{3\zeta(3)}{2\pi^3} \tag{6.15}$$

where $g_2(0)$ has been evaluated by means of (2.16) and [4, Formula 6.6(4)]. In a similar manner it is found that

$$C_0=-\int_{-\infty}^{\infty}t[F''_{-1/2}(t)]^2\,dt=g'_2(0)=0. \tag{6.16}$$

The asymptotic expansion (6.14) does agree with (6.8) provided that

$$-(2k)!\,c_{k+1}=\frac{b_k}{(2k+1)(2k+2)}, \qquad k=0,1,2,\cdots. \tag{6.17}$$

The latter identity can be proved by a generating function technique.

REFERENCES

[1] R. B. DINGLE, *The Fermi-Dirac integrals* $F_p(\eta) = (p!)^{-1} \int_0^\infty \varepsilon^p (e^{\varepsilon-\eta}+1)^{-1}\, d\varepsilon$, Appl. Sci. Res., B6 (1957), pp. 225-239.

[2] G. DOETSCH, *Handbuch der Laplace-Transformation*, Band II, Birkhäuser-Verlag, Basel, 1955.

[3] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions*, Vols. I, II, McGraw-Hill, New York, 1953.

[4] ———, *Tables of Integral Transforms*, Vol. I, McGraw-Hill, New York, 1954.

[5] D. J. W. GELDART AND M. L. GLASSER, *Finite temperature approach to the gradient expansion for exchange*, in preparation.

[6] M. L. GLASSER AND J. BOERSMA, *Exchange energy of an electron gas of arbitrary dimensionality*, SIAM J. Appl. Math., 43 (1983), pp. 535-545.

[7] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Second edition, Cambridge University Press, Cambridge, 1958.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 821-846, May 1991

017

CONTIGUITY RELATIONS OF AOMOTO-GEL'FAND HYPERGEOMETRIC FUNCTIONS AND APPLICATIONS TO APPELL'S SYSTEM F_3 AND GOURSAT'S SYSTEM ${}_3F_2$ *

TAKESHI SASAKI†

This paper is dedicated to the memory of Yutaka Se-ashi.

Abstract. Aomoto-Gel'fand hypergeometric functions $\Phi(Z, \alpha)$ [K. Aomoto, *Sci. Papers*, College of Arts and Sciences, University of Tokyo, 27 (1977), pp. 49-61], [I. M. Gel'fand, *Soviet Math. Dokl.*, 33 (1986), pp. 573-577] are functions of z defined on the Grassmannian $G_{k,n}$, the set of k-dimensional subspaces of an n-dimensional linear space, and with complex parameters (α). Such a class of functions contains certain classical hypergeometric functions (HGF), such as the HGF of Gauss, the generalized HGF ${}_{p+1}F_p$, and Appell's HGF's F_1, F_2, F_3. On the other hand, W. Miller [*J. Math. Phys.*, 13 (1972), pp. 1393-1399; *SIAM J. Appl. Math.*, 25 (1973), pp. 226-235; *SIAM J. Math. Anal.*, 3 (1972), pp. 31-44] has given contiguity relations for several HGF's, including the HGF's mentioned above, and has shown the Lie-algebraic structure of the equations satisfied by these functions. This paper first presents a principle of obtaining contiguity relations for Aomoto-Gel'fand HGF's and clarifies the Lie-algebraic structure among them. The contiguity relations for Lauricella's HGF F_D are known easily from this principle. The second part of this paper is an application to get complete tables of contiguity relations for F_3 and ${}_3F_2$, which complement the tables for these functions given by Miller in the papers cited above.

Key words. contiguity relations, Aomoto-Gel'fand hypergeometric functions, Appell-Lauricella hypergeometric function F_D, Appell's hypergeometric function F_3, Goursat's hypergeometric function ${}_3F_2$

AMS(MOS) subject classifications. 33A30, 33A75, 33A70

Introduction. This paper is aimed at defining the contiguity relations of the Aomoto-Gel'fand hypergeometric functions and providing, as applications, a complete list of contiguity relations of Appell's hypergeometric function F_3 and that of Goursat's generalized hypergeometric function ${}_3F_2$. We use a short term HGF for "hypergeometric function."

Let us recall the contiguity relations of the Gauss HGF:

$$F(\alpha, \beta, \gamma; x) = \sum_{n=0}^{\infty} \frac{(\alpha, n)(\beta, n)}{(\gamma, n)n!} x^n,$$

where (α, n) denotes the factorial function

$$(\alpha, n) = \begin{cases} 1, & n = 0, \\ \alpha(\alpha+1)(\alpha+2)\cdots(\alpha+n-1), & n = 1, 2, 3, \cdots. \end{cases}$$

Functions such as $F(\alpha \pm 1, \beta, \gamma; x)$, $F(\alpha, \beta \pm 1, \gamma; x), \cdots$ are called *contiguous* to $F(\alpha, \beta, \gamma; x)$, and linear relations among contiguous functions and their derivatives are called *contiguity relations.* The following are typical examples:

$$\alpha F(\alpha+1, \beta, \gamma; x) = \alpha F(\alpha, \beta, \gamma; x) + x\frac{d}{dx} F(\alpha, \beta, \gamma; x),$$

$$(\alpha-\gamma+1)F(\alpha, \beta, \gamma; x) = (\gamma-\alpha-1-\beta x)F(\alpha+1, \beta, \gamma; x) + x(1-x)\frac{d}{dx} F(\alpha+1, \beta, \gamma; x). \tag{0.1}$$

* Received by the editors June 7, 1989; accepted for publication (in revised form) December 11, 1989.
† Department of Mathematics, Faculty of Sciences, Kobe University, Rokko, Kobe, 657 Japan.

Define differential operators H and B which appear in the right-hand sides by

$$H = x\frac{d}{dx} + \alpha \quad \text{and} \quad B = x(1-x)\frac{d}{dx} + (\gamma - \alpha - 1 - \beta x), \tag{0.2}$$

and denote by $S(\alpha, \beta, \gamma)$ the space of solutions of the Gauss hypergeometric differential equation (HGDE):

$$x(1-x)\frac{d^2u}{dx^2} + \{\gamma - (\alpha + \beta + 1)x\}\frac{du}{dx} - \alpha\beta u = 0. \tag{0.3}$$

Then these operators define linear homomorphisms:

$$\begin{aligned} &H : S(\alpha, \beta, \gamma) \to S(\alpha + 1, \beta, \gamma), \\ &B : S(\alpha + 1, \beta, \gamma) \to S(\alpha, \beta, \gamma), \end{aligned} \tag{0.4}$$

which are isomorphisms when $\alpha(\alpha + \gamma - 1) \neq 0$. The differential operator

$$\frac{d}{dx} : S(\alpha, \beta, \gamma) \to S(\alpha + 1, \beta + 1, \gamma + 1),$$

which has been basic in the study of the Gauss HGF in [GA], also reflects a contiguity relation

$$\frac{d}{dx} F(\alpha, \beta, \gamma, x) = \frac{\alpha\beta}{\gamma} F(\alpha + 1, \beta + 1, \gamma + 1; x).$$

As we have seen in [SY3], such relations form a part of the symmetry of the Gauss HGDE (0.3), and symmetry plays a fundamental role in the study of the HGF. It is thus important to understand how the contiguity relations arise for several HGF's. Refer also to [IKSY].

In this paper we first present a general principle of getting such relations and differential operators for a wider class of HGF's, i.e., a class of HGF's defined by Aomoto [A1] and Gel'fand [GE], which we call the Aomoto-Gel'fand HGF's (see § 1).

We next apply this principle to classical HGF's. Contiguity relations for the Appell-Lauricella HGF, denoted by F_D, are completely known by Miller [M1] in relation with the Lie-algebraic structure of the system of differential equations satisfied by these functions. Contiguity relations for Appell's HGF F_3 of two variables and those for Goursat's HGF ${}_pF_q$ are partly known by Miller [M2], [M3]. On the other hand, these functions F_D, F_3, and ${}_pF_q$ ($p = q + 1$) belong to the class of the Aomoto-Gel'fand HGF's. Therefore, as applications, we can reproduce contiguity relations of F_D by a simpler principle and complement those relations of F_3 and ${}_3F_2$ by Miller to yield a complete set of contiguity relations. The Lie-algebraic structure of the associated differential operators is also clarified by the same principle.

Here we give some remarks. Miller et al. in [M1]-[M4], [KMM1]-[KMM3] have determined the symmetry of Horn's two-variable systems. To each system they associated a constant coefficient system, called the canonical system, on a manifold that has fibering over the original two-manifold. In this instance they used (some) contiguity relations. Our present context is, put simply, in the opposite direction; we start with systems which are full of symmetry and then restrict our consideration to base manifolds to get contiguity relations of classical HGF's. This concept itself is not new. Hrabowski [H] introduced the canonical systems, which are generalizations of HGDE's, associated with certain simple Lie groups and classified the Lie-algebraic content of such systems. He also gave a principle, essentially the same as ours, for getting contiguity relations

(recurrence relations, in his terminology) for such systems. Recently, Gel'fand, Zelevinskii, and Kapranov [GZK] gave a precise description of algebraic structure of canonical systems, generalizing that by Hrabowski. So, by a method similar to that exhibited here we may find contiguity relations explicitly for several other HGF's. Appell's F_4, for example, is one such HGF since the function F_C that is a higher-dimensional generalization of F_4 is a special case of this general setting, although the actual computation needs some elaborate working. The author does not know, however, whether F_4 has a representation as one of Aomoto-Gel'fand HGF's. Refer to [T], where Takayama has given another method of obtaining contiguity relations for F_4.

It should be noted that Sato [SA] has already given a different approach to get a generalization of HGF's by referring to prehomogeneous vector spaces and b-functions.

In § 1 we recall the definition of the Aomoto-Gel'fand HGF and the system $E(k, n)$ of differential equations satisfied by this function. In § 2 contiguity relations for this system will be defined. From these relations we obtain *generalized Gauss relations*, namely, hypergeometric difference equations with respect to parameters. We treat in § 3 the system $E(2, n+3)$, which is associated with the Appell-Lauricella HGF F_D, and reproduce contiguity relations due to Miller (see Table 1).

In §§ 4–10 we deal with the system $E(3, 6)$, whose explicit expression is given in § 4. It is noteworthy that, while the system $E(3, 6)$ is defined on $\mathbf{CP}^4$ and of rank 6, it includes, as subsystems, Appell's systems F_1, F_2, and F_3, Gousat's system ${}_3F_2$, and the Gauss HGDE. This will be recalled in § 6. Moreover, this fact enables us to derive contiguity relations for these systems from the contiguity relations of $E(3, 6)$. Refer to [MSY2] for the algebraic-geometric implications of this fact. In § 5 and § 8 we give two representations of contiguity relations of $E(3, 6)$ (see Tables 2, 4). In § 7 we complete enumerating contiguity relations for Appell's F_3 (see Tables 3.1, 3.2). The list of those relations for ${}_3F_2$ will be given in § 9 (Table 5). Section 10 gives proofs of two technical lemmas.

1. The Aomoto–Gel'fand hypergeometric function. We recall the definition of the Aomoto-Gel'fand HGF following [GE] and [GG]. Fix integers n and k so that $k < n$. Let $Z_{k,n}$ denote the space of $k \times n$ complex matrices of rank k such that any column vector is nonzero; $G_{k,n}$ the Grassmannian manifold of k-subspaces in $\mathbf{C}^n$. Since each point in $Z_{k,n}$ defines an imbedding of $\mathbf{C}^k$ into $\mathbf{C}^n$, there is a natural projection $Z_{k,n} \to G_{k,n}$. We let (t^i) be coordinates of $\mathbf{C}^k$ and define a $(k-1)$-form ω by

$$\omega = \sum_{i=1}^{k} (-1)^{i+1} t^i \, dt^1 \wedge \cdots \wedge dt^{i-1} \wedge dt^{i+1} \wedge \cdots \wedge dt^k.$$

For a set of complex numbers $\alpha = (\alpha_1, \cdots, \alpha_n)$ with the property $\sum \alpha_j = n - k$ and for a point $z = (z_{ij})$ in $Z_{k,n}$ we put $\tau = \prod_{j=1}^{n} (\sum_{i=1}^{k} z_{ij} t^i)^{\alpha_j - 1} \omega$, which can be seen as a $(k-1)$-form on the projective space $\mathbf{CP}^{k-1}$. Then we take a suitable $(k-1)$-cycle C in $\mathbf{CP}^{k-1} - S$, where S is the union of n hyperplanes $\sum_{i=1}^{k} z_{ij} t^i = 0, j = 1, \cdots, n$, and define a function by the integral

$$\Phi(z, \alpha) = \int_C \tau, \tag{1.1}$$

which will be called the *Aomoto-Gel'fand hypergeometric function.*

This function is invariant under two kinds of group action. Since $Z_{k,n}$ is a principal GL_k-bundle over $G_{k,n}$, $g \in GL_k$ acts on $Z_{k,n}$ on the left. Under this action the integral

Φ changes as

$$\Phi(gz, \alpha) = (\det g)^{-1}\Phi(z, \alpha). \tag{1.2}$$

On the other hand, the Cartan subgroup H_n, consisting of diagonal matrices, acts on $Z_{k,n}$ on the right. Under this action, Φ transforms as

$$\Phi(zh, \alpha) = \prod_j (h_j)^{\alpha_j - 1}\Phi(z, \alpha), \tag{1.3}$$

where $h = \operatorname{diag}(h_1, \cdots, h_n)$.

Two points ζ and ζ' in $G_{k,n}$ are said to be *equivalent* if $\dim(P \cap \zeta) = \dim(P \cap \zeta')$ for any coordinate planes $P = \{(y^i) \in \mathbf{C}^n;\ y^{i_1} = y^{i_2} = \cdots = y^{i_j} = 0\}$ $(1 \leqq j \leqq n)$. Sets of equivalent points define a *stratification* of $G_{k,n}$. Note that each stratum is invariant under the action of the group H_n. It is of particular importance to study the restriction of the integral to each stratum; combinatorial information on the configuration of strata plays an essential role.

It is easy to see that the function Φ, viewed as a function on $Z_{k,n}$, satisfies the following system $E(k, n) = E(k, n; \alpha)$ of differential equations:

$$\sum_{i=1}^{k} z_{ij}\frac{\partial}{\partial z_{ij}}\Phi - (\alpha_j - 1)\Phi = 0, \tag{1.4}$$

$$\sum_{j=1}^{n} z_{ij}\frac{\partial}{\partial z_{lj}}\Phi + \delta_{il}\Phi = 0, \tag{1.5}$$

$$\frac{\partial^2}{\partial z_{ip}\,\partial z_{jq}}\Phi - \frac{\partial^2}{\partial z_{iq}\,\partial z_{jp}}\Phi = 0. \tag{1.6}$$

Equations (1.4) and (1.5) reflect the invariances (1.2) and (1.3), respectively. It is known [KN] that the rank of $E(k, n; \alpha)$ is generally not greater than $\binom{n-2}{k-1}$ and is equal to this number when z lies in the unique open stratum and when any α_j takes no integral value.

For the precise presentation of the above materials, refer to [A1], [A3], [GE], [GG], [KN].

2. Contiguity relations and Gauss relations. On the space $Z_{k,n}$ acts the general linear group GL_n on the right. We are interested in the associated infinitesimal transformation, which induces the vector fields L_{jl} on $Z_{k,n}$ given by

$$L_{jl} = \sum_{i=1}^{k} z_{ij}\frac{\partial}{\partial z_{il}} \quad \text{for } j, l = 1, \cdots, n. \tag{2.1}$$

The next lemma is a key to understanding what follows.

LEMMA 2.1. *The Lie algebra generated by vector fields L_{jl} is isomorphic to the Lie algebra $\mathfrak{gl}_n$ of general linear matrices.*

Proof. We have only to associate to L_{jl} the $n \times n$ matrix with 1 in the (jl)th component and zero in the others. □

Applying these differential operators on the integral (1.1), we easily see the following proposition.

PROPOSITION 2.2. *The integral Φ satisfies relations*

$$L_{jl}\Phi(z, \alpha) = (\alpha_l - 1)\Phi(z, \alpha + 1_j - 1_l), \tag{2.2}$$

where 1_j denotes the vector with 1 in the jth component and zero in the others.

While these identities are infinitesimal expressions of the right action, they can be seen as differential-difference relations among a solution of the system $E(k, n; \alpha)$

and a solution of another system $E(k, n; \alpha+1_j-1_l)$. So we introduce the following definition.

DEFINITION. We call the relations (2.2) the *contiguity relations* of the integral Φ or of the system $E(k, n)$.

We now look closer at these relations by the invariance (1.2). Take a point $z=(z_1, z_2)$ in $Z_{k,n}$, where z_1 is a nonsingular $k\times k$ matrix and z_2 is a $k\times(n-k)$ matrix. Put

$$u=(u_{ap})=z_1^{-1}z_2, \tag{2.3}$$

which can be seen as a point in $G_{k,n}$. The ranges of indices are supposed to be

$$1\leqq i, j, \cdots, \leqq n, \quad 1\leqq a, b, \cdots, \leqq k, \quad k+1\leqq p, q, \cdots, \leqq n.$$

We define

$$\varphi(u, \alpha)=\Phi((I_k, u), \alpha). \tag{2.4}$$

The invariance (1.2) implies

$$\Phi(z, \alpha)=(\det z_1)^{-1}\varphi(u, \alpha). \tag{2.5}$$

Hence the action of L_{jl} on Φ induces a differential operator $X_{jl}=(\det z_1)L_{jl}(\det z_1)^{-1}$ acting on φ; φ satisfies contiguity relations

$$X_{jl}\varphi(u, \alpha)=(\alpha_l-1)\varphi(u, \alpha+1_j-1_l). \tag{2.6}$$

LEMMA 2.3. *The operators X_{jl} are given by the formulae*

$$X_{jl}=\begin{cases} -\delta_{ab}-\sum_{p=k+1}^{n} u_{bp}\dfrac{\partial}{\partial u_{ap}}, & j=a, \quad l=b, \\ \dfrac{\partial}{\partial u_{ap}}, & j=a, \quad l=p, \\ -u_{ap}-\sum_{q=k+1}^{n} u_{aq}U_{pq}, & j=p, \quad l=a, \\ U_{pq}, & j=p, \quad l=q, \end{cases} \tag{2.7}$$

where

$$U_{pq}=\sum_{a=1}^{k} u_{ap}\frac{\partial}{\partial u_{aq}}.$$

Proof. Denote by (V_{ab}) the inverse matrix of z_1. Assume $1\leqq j, l\leqq k$. Then the derivation of (2.5) by L_{jl} gives

$$\begin{aligned} L_{jl}\Phi &= \sum z_{aj}\frac{\partial}{\partial z_{al}}((\det z_1)^{-1}\varphi(u, \alpha)) \\ &= \sum z_{aj}\frac{\partial}{\partial z_{al}}(\det z_1)^{-1}\cdot\varphi+\sum z_{aj}(\det z_1)^{-1}\frac{\partial u_{bp}}{\partial z_{al}}\frac{\partial\varphi}{\partial u_{bp}} \\ &= -\sum z_{aj}(\det z_1)^{-2}(\text{the } (al)\text{-cofactor})\varphi+\sum z_{aj}(\det z_1)^{-1}(-V_{ba}u_{lp})\frac{\partial\varphi}{\partial u_{bp}} \\ &= -\sum(\det z_1)^{-1}\delta_{jl}\varphi-\sum(\det z_1)^{-1}u_{lp}\frac{\partial\varphi}{\partial u_{jp}}. \end{aligned} \tag{2.8}$$

Hence we get the first formula. Other formulae are similarly shown. □

It is easy to see that these operators X_{jl} are linearly independent and form a Lie algebra isomorphic to $\mathfrak{gl}_n$, which we denote by $\mathfrak{g}$. The subspace $\mathfrak{g}_0 = \mathbf{C}\{U_{pq}\}$ is a subalgebra isomorphic to $\mathfrak{gl}_{n-k}$. The algebra $\mathfrak{g}$ is decomposed into the sum of $\mathfrak{g}_0$ and three subspaces.

$$\mathfrak{g}_1 = \mathbf{C}\{\partial/\partial u_{ap}\}, \quad \mathfrak{g}_{-1} = \mathbf{C}\{u_{ap} + \sum u_{aq} U_{pq}\}, \quad \text{and} \quad \mathfrak{g}_0' = \mathbf{C}\{\delta_{ab} + \sum u_{bp}\, \partial/\partial u_{ap}\}.$$

The triplet $\{\mathfrak{g}_{-1}, \mathfrak{g}_0 \oplus \mathfrak{g}_0', \mathfrak{g}_1\}$ gives a gradation of $\mathfrak{g}$.

The identity (2.6) for each X_{jl} gives

PROPOSITION 2.3. *The contiguity relations with respect to coordinates* (u_{ap}) *of* $G_{k,n}$ *are*

$$-\delta_{ab}\varphi(\alpha) - \sum_{p=k+1}^{n} u_{bp} \frac{\partial \varphi}{\partial u_{ap}}(\alpha) = (\alpha_b - 1)\varphi(\alpha + 1_a - 1_b), \tag{2.9}$$

$$\frac{\partial \varphi}{\partial u_{ap}}(\alpha) = (\alpha_p - 1)\varphi(\alpha + 1_a - 1_p), \tag{2.10}$$

$$-u_{ap}\varphi(\alpha) - \sum_{q=k+1}^{n} u_{aq} \left(\sum_{b=1}^{k} u_{bp} \frac{\partial \varphi}{\partial u_{bq}}(\alpha) \right) = (\alpha_a - 1)\varphi(\alpha + 1_p - 1_b), \tag{2.11}$$

$$\sum_{a=1}^{k} u_{ap} \frac{\partial \varphi}{\partial u_{aq}}(\alpha) = (\alpha_q - 1)\varphi(\alpha + 1_p - 1_q), \tag{2.12}$$

for $1 \leqq a, b \leqq k$, *and* $k+1 \leqq p, q \leqq n$.

In [A2] and [A3] Aomoto has given a general theory on the system of linear difference equations satisfied by the integral $\Phi(z, \alpha)$ viewed as a function of α. In our case the difference equations are obtained from the above formulae: it is enough to get rid of differential terms by using the second identity (2.10). Namely, we have the following proposition.

PROPOSITION 2.4. *The Aomoto-Gel'fand hypergeometric function* $\varphi(u, \alpha)$ *satisfies the system of difference equations with respect to* α:

$$\varphi(u, \alpha) = \sum_{a=1}^{k} u_{ap} \varphi(u, \alpha + 1_a - 1_p), \tag{2.13}$$

$$\alpha_a \varphi(u, \alpha) + \sum_{p=k+1}^{n} (\alpha_p - 1) u_{ap} \varphi(u, \alpha + 1_a - 1_p) = 0, \tag{2.14}$$

$$\begin{aligned} \alpha_p u_{ap} \varphi(u, \alpha) + \sum_{q=k+1, q \neq p}^{n} (\alpha_q - 1) u_{aq} \varphi(u, \alpha + 1_p - 1_q) \\ + (\alpha_a - 1)\varphi(u, \alpha + 1_p - 1_a) = 0, \end{aligned} \tag{2.15}$$

where $1 \leqq a \leqq k$ *and* $k+1 \leqq p \leqq n$.

Proof. By inserting (2.10) into (2.11), (2.9), and (2.12) we have (2.15) and

$$\delta_{ab}\varphi(u, \alpha) + \sum (\alpha_p - 1) u_{bp} \varphi(u, \alpha + 1_a - 1_p) + (\alpha_b - 1)\varphi(u, \alpha + 1_a - 1_b) = 0, \tag{2.16}$$

$$\varphi(u, \alpha + 1_p - 1_q) - \sum u_{ap} \varphi(u, \alpha + 1_a - 1_q) = 0, \tag{2.17}$$

respectively. Replacing $\alpha + 1_p - 1_q$ by α in (2.17) we get (2.13). Replacing $\alpha + 1_a - 1_b$ by α in (2.16) we get (2.14). □

Remark. Equations (2.13) and (2.14) have been given in [GZ].

DEFINITION. We call three systems of difference equations (2.13), (2.14), and (2.15) *generalized Gauss relations* (cf. [GA], [GZ]).

Example. The Goursat HGF ${}_kF_{k-1}$ is given by an integral associated with the one-dimensional stratum in $G_{k,2k}/H_{2k}$ consisting of points

$$\begin{pmatrix} 1 & & & & & 1 & & & & & 1 \\ & 1 & & & & -1 & 1 & & & & \\ & & 1 & & & & -1 & 1 & & & \\ & & & \ddots & & & & & \ddots & & \\ & & & & 1 & & & & & 1 & \\ & & & & & & & & & -1 & -x \end{pmatrix}.$$

In fact it is by definition

$$\begin{aligned} &{}_kF_{k-1}(\alpha_1, \cdots, \alpha_k; \beta_1, \cdots, \beta_{k-1}; x) \\ &\quad = \prod_{j=1}^{k-1} \frac{\Gamma(\beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j - \alpha_j)} \int t_1^{\alpha_1 - \beta_2} \times \prod_{j=2}^{k-2} t_j^{\alpha_j - \beta_{j+1}} t_{k-1}^{\alpha_{k-1} - 1} (1 - t_1)^{\beta_1 - \alpha_1 - 1} \\ &\qquad \cdot \prod_{j=2}^{k-1} (t_{j-1} - t_j)^{\beta_j - \alpha_j - 1} (1 - x t_{k-1})^{-\alpha_k} \, dt_1 \wedge \cdots \wedge dt_{k-1} \end{aligned} \tag{2.18}$$

integrated over the cycle $1 \geqq t_1 \geqq \cdots \geqq t_{k-1} \geqq 0$. The Gauss relations in Proposition 2.4 now consist of $k^2 + 2k$ difference relations, each of which contains three terms. Note that Rainville [R] has given a list of such relations; his list consists of $2k-2$ relations, each connecting three contiguous functions, and $k+1$ relations, each containing $k+1$ contiguous functions.

In the case $k = 2$, where the integral is the Gauss HGF F, we have eight relations; four of them are

$$\begin{aligned} &\gamma F - \alpha F(\alpha^+ \gamma^+) - (\gamma - \alpha) F(\gamma^+) = 0, \\ &\gamma F - \beta F(\beta^+ \gamma^+) - (\gamma - \beta) F(\gamma^+) = 0, \\ &\gamma F - \gamma F(\beta^+) + \alpha x F(\alpha^+ \beta^+ \gamma^+) = 0, \\ &\gamma F - \gamma F(\alpha^+) + \beta x F(\alpha^+ \beta^+ \gamma^+) = 0, \end{aligned}$$

where $F = F(\alpha, \beta, \gamma; x) = {}_2F_1(\alpha, \beta; \gamma; x)$ and $(\alpha^+) = (\alpha + 1, \beta, \gamma)$, $(\alpha^+ \gamma^+) = (\alpha + 1, \beta, \gamma + 1), \cdots$. The remaining four relations follow from these with certain translations of parameters. From the last two relations we have a well-known formula:

$$(\beta - \alpha) F - \beta F(\beta^+) + \alpha F(\alpha^+) = 0.$$

Another coordinate system on the stratum gives different relations; namely, because of the symmetry (1.2), we get relations transformed by so-called Kummer's identities. For example, consider a stratum of points

$$\begin{pmatrix} 1 & & 1 & 1 - x \\ & 1 & 1 & 1 \end{pmatrix}$$

in $G_{2,4}$. It is easy to see that the integral also defines the Gauss HGF with respect to the coordinate x. Then, in addition to the above relations, we have

$$\begin{aligned} &\gamma F - (\gamma - \alpha) F(\beta^+ \gamma^+) + \alpha(x - 1) F(\alpha^+ \beta^+ \gamma^+) = 0, \\ &\gamma F - (\gamma - \beta) F(\alpha^+ \gamma^+) + \beta(x - 1) F(\alpha^+ \beta^+ \gamma^+) = 0. \end{aligned}$$

In this way we can find 15 relations by Gauss [GA, p. 130].

The structure of the difference system for $k=3$ has been studied extensively by Aomoto in [A4]. Refer also to the definition of a generalized beta-function introduced in [GGR3]. The study for general k is interesting.

3. The Appell–Lauricella hypergeometric function F_D. The integral Φ can be determined by its value on the quotient space Γ/H_n of each stratum Γ by the invariance (1.2) and (1.3). We describe in this section contiguity relations (2.6) for the system $E(2, n+3)$ on the quotient of the open stratum.

We introduce coordinates on the open stratum by

$$(I_2, u)=\begin{pmatrix}1 & 0 & u_{13} & u_{14} & \cdots & u_{1,n+3}\\ 0 & 1 & u_{23} & u_{24} & \cdots & u_{2,n+3}\end{pmatrix},$$

where no 2×2 subdeterminants are zeros; to each point $u=(u_{ap})$ we associate a point in $\mathbf{C}^n$ with coordinates $(x^4, \cdots, x^{n+3})$ by

$$x^j=\frac{u_{13}u_{2j}}{u_{23}u_{1j}} \quad \text{for } 4\leqq j\leqq n+3.$$

The integral $\varphi(u, \alpha)=\Phi((I_2, u), \alpha)$ is by definition

$$\int t_1^{\alpha_1-1}t_2^{\alpha_2-1}\prod_{j=3}^{n+3}(u_{1j}t_1+u_{2j}t_2)^{\alpha_j-1}(-t_2\,dt_1+t_1\,dt_2).$$

Define ϕ by

$$\phi(u, \alpha)=u_{13}^{\alpha_2+\alpha_3-1}(-u_{23})^{-\alpha_2}\prod_{j=4}^{n+3}u_{1j}^{\alpha_j-1}.$$

Then the integrand is

$$\phi(u, \alpha)t^{\alpha_2-1}(1-t)^{\alpha_3-1}\prod_{j=4}^{n+3}(1-x^jt)^{\alpha_j-1}\,dt$$

for $t=-u_{23}t_2/u_{13}t_1$. Here recall the Appell–Lauricella HGF in n variables $(y^1, \cdots, y^n)$:

$$\begin{aligned}&F_D(\alpha, \beta_1, \cdots, \beta_n, \gamma; y^1, \cdots, y^n)\\ &\quad=\frac{\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\gamma-\alpha)}\int_0^1 t^{\alpha-1}(1-t)^{\gamma-\alpha-1}\prod_{j=1}^{n}(1-y^jt)^{-\beta_j}\,dt.\end{aligned}\tag{3.1}$$

Then, from the above consideration, if we take the interval $(0, 1)$ as a cycle in t-space, we have

$$\varphi(u, \alpha)=\phi(u, \alpha)w(x, \alpha),\tag{3.2}$$

where

$$w(x, \alpha)=\frac{\Gamma(\alpha)\Gamma(\gamma-\alpha)}{\Gamma(\gamma)}F_D(\alpha, \beta_4, \cdots, \beta_{n+3}, \gamma; x^4, \cdots, x^{n+3}),\tag{3.3}$$

$$\alpha=\alpha_2, \quad \beta_j=1-\alpha_j \quad (4\leqq j\leqq n+3), \quad \text{and} \quad \gamma=\alpha_2+\alpha_3.$$

Now it is easy to compute the action of X_{jl}. In the following we use the notation

$$\partial_j=\frac{\partial}{\partial x^j}, \quad \delta_j=x^j\partial_j, \quad \text{and} \quad \delta=\delta_4+\cdots+\delta_{n+3}.\tag{3.4}$$

Define Y_{jl} by

$$Y_{jl}w=\phi^{-1}X_{jl}\varphi\big|_{u_{23}=u_{13}=u_{14}=\cdots=u_{1,n+3}=1}.\tag{3.5}$$

We can see

$$\phi^{-1}X_{jl}\varphi = M_{jl}(u)\,Y_{jl}w, \tag{3.6}$$

where $M_{jl}(u)=\phi(u,\alpha+1_j-1_l)/\phi(u,\alpha)$, that is, equal to 1 when $u_{23}=u_{13}=\cdots=u_{1,n+3}=1$. The computation of Y_{jl} will be straightforward: we treat, for example, the case $(jl)=(13)$. Since $X_{13}=\partial/\partial u_{13}$,

$$X_{13}\varphi(u,\alpha)=\frac{\phi}{u_{13}}((\alpha_2+\alpha_3-1)w+\delta w).$$

The definition of ϕ implies

$$(\phi w)(\alpha+1_j-1_l)=\frac{1}{u_{13}}\phi(\alpha)w(\alpha+1_j-1_l).$$

Hence

$$Y_{13}w=(\alpha_2+\alpha_3-1)w+\delta w=(\alpha_3-1)w(\alpha+1_j-1_l)\quad\text{and}\quad M_{13}=1/u_{13}.$$

Thus we have the following proposition.

PROPOSITION 3.1. *The Appell-Lauricella* HGF $w(x,\alpha)$ *satisfies contiguity relations in Table* 1. *The diagonal operators* Y_{jj}, $1\leqq j\leqq n+3$, *induce scalar multiplications by* α_j-1.

TABLE 1

Index	Y	a_Y	α^Y	M
[13]	$\gamma-1+\delta$	$\gamma-\alpha-1$	γ^-	$1/u_{13}$
[23]	$-\alpha-\delta$	$-(\gamma-\alpha-1)$	α^+	$1/u_{23}$
[1j]	$-\beta_j-\delta_j$	$-\beta_j$	β_j^+	$1/u_{1j}$
[2j]	δ_j	β_j	$\alpha^+\beta_j^+\gamma^+$	$u_{13}/u_{23}u_{1j}$
[12]	$1-\gamma+\sum_j\beta_jx^j+\sum_j(x^j-1)\delta_j$	$-(\alpha-1)$	$\alpha^-\gamma^-$	u_{23}/u_{13}
[21]	$\alpha+\sum_j(x^j-1)\partial_j$	$-(\sum\beta_j-\gamma)$	$\alpha^+\gamma^+$	u_{13}/u_{23}
[3j]	$-\beta_j+(1-x^j)\partial_j$	$-\beta_j$	$\beta_j^+\gamma^+$	u_{13}/u_{1j}
[j3]	$\gamma-1+(1-x^j)\delta-\alpha x^j$	$\gamma-\alpha-1$	$\beta_j^-\gamma^-$	u_{1j}/u_{13}
[jl]	$-\beta_l+(x^j-x^l)\partial_l$	$-\beta_l$	$\beta_j^-\beta_l^+$	u_{1j}/u_{1l}
[31]	$\sum\beta_j-\gamma+\alpha+\sum_j(x^j-1)\partial_j$	$\sum\beta_j-\gamma$	γ^+	u_{13}
[32]	$\alpha-\gamma+\sum_j\beta_jx^j+\sum_j(x^j-1)\delta_j$	$-(\alpha-1)$	α^-	u_{23}
[j1]	$\sum\beta_j-\gamma+\alpha x^j+x^j\sum_l(x^l-1)\partial_l$	$\sum\beta_j-\gamma$	β_j^-	u_{1j}
[j2]	$(\alpha-1)x^j+\sum_l\beta_lx^l+1-\gamma+\sum(x^l-1)\delta_l$	$-(\alpha-1)$	$\alpha^-\beta_j^-\gamma^-$	$u_{23}u_{1j}/u_{13}$

Each line reads as $Yw(x,\alpha)=a_Yw(x,\alpha^Y)$. The first column denotes indices of Y_{jl}. We use parameters (α,β_j,γ) instead of (α_j) to make the formulae easy to refer to in classical notation. The notation (α^+) in the last column implies that the parameter α is increased by 1 and other parameters are unchanged. The range of indices in Table 1 for $j, l, \cdots$ is from 4 to $n+3$.

Remark. When $n=1$, F_D is the Gauss HGF. The operators H and B defined in (0.4) are $-Y_{23}$ and $-Y_{32}$, respectively.

Remark. The Lie-algebraic structure associated with the above operators given by Miller [M1] is the same as that of $\mathbf{C}\{X_{jl}\}$ (cf. the proof of Proposition 7.2).

Remark. Kametaka and Okamoto [O] introduced the notion of ladder structure and formulated a connection of this structure with the Toda equation. In the present case, their theory can be applied to subalgebras $\mathbf{C}\{X_{jl}, X_{lj}, X_{jj}-X_{ll}\}$ for $j\neq l$, each of

which is isomorphic to $\mathfrak{sl}_2$. For example, when $n=1$, the algebra for $j=2$ and $l=3$ associated with H and B as remarked above is used to give solutions of

$$\left((1-x)\frac{d}{dx}\right)^2 \log \chi_m = \chi_{m-1}\chi_{m+1}/\chi_m^2, \qquad m \in Z.$$

This remark works for each pair (k, n).

4. The system $E(3, 6)$. In this section we derive an explicit form of the system $E(3,6)$ on the quotient of the open stratum of $G_{3,6}$. We fix coordinates as follows. Consider the set of points

$$(I_3, u) = \begin{pmatrix} 1 & & & u_{14} & u_{15} & u_{16} \\ & 1 & & u_{24} & u_{25} & u_{26} \\ & & 1 & u_{34} & u_{35} & u_{36} \end{pmatrix} \tag{4.1}$$

in $Z_{3,6}$. Assume u_{14}, u_{15}, u_{16}, u_{24}, and u_{34} are not equal to zeros and define

$$x^1 = \frac{u_{14}u_{25}}{u_{15}u_{24}}, \quad x^2 = \frac{u_{14}u_{26}}{u_{16}u_{24}}, \quad x^3 = \frac{u_{14}u_{35}}{u_{15}u_{34}}, \quad x^4 = \frac{u_{14}u_{36}}{u_{16}u_{34}}. \tag{4.2}$$

Then $x=(x^i)$ is a system of coordinates around the single H_6-orbit $\{u_{25}=u_{26}=u_{35}=u_{36}=0\}$. It defines a local section of the projection $Z_{3,6} \to GL_3\backslash Z_{3,6}/H_6$ by associating a matrix

$$\begin{pmatrix} 1 & & & 1 & 1 & 1 \\ & 1 & & 1 & x^1 & x^2 \\ & & 1 & 1 & x^3 & x^4 \end{pmatrix}. \tag{4.3}$$

Remark here that this choice of coordinates is by no means the unique one. In § 8 we introduce another choice in order to study Goursat's HGF ${}_3F_2$, although the above choice is useful in the study of Appell's HGF F_3.

Put

$$\phi(u) = u_{14}^{\alpha_{234}-1} u_{24}^{-\alpha_2} u_{34}^{-\alpha_3} u_{15}^{\alpha_5-1} u_{16}^{\alpha_6-1}, \tag{4.4}$$

where $\alpha_{234} = \alpha_2+\alpha_3+\alpha_4$, and define a function w of (x^i) around the origin by

$$w(x, \alpha) = c(\alpha)F(x, \alpha),$$

where

$$c(\alpha) = \Gamma(1-\alpha_{234})\Gamma(\alpha_2)\Gamma(\alpha_3)/\Gamma(1-\alpha_4)$$

and

$$F(x, \alpha) = \sum \frac{(1-\alpha_5, k+l)(1-\alpha_6, m+n)(\alpha_2, k+m)(\alpha_3, l+n)}{(\alpha_{234}, k+l+m+n)(1, k)(1, l)(1, m)(1, n)} \cdot (x^1)^k (x^2)^l (x^3)^m (x^4)^n. \tag{4.5}$$

Summation is taken over nonnegative integers and $(a, k) = a(a+1)\cdots(a+k-1)$. Since the integral Φ is

$$\int \phi t_1^{\alpha_1-1} t_2^{\alpha_2-1} t_3^{\alpha_3-1} (t_1+t_2+t_3)^{\alpha_4-1} (t_1+x^1t_2+x^3t_3)^{\alpha_5-1} (t_1+x^2t_2+x^4t_3)^{\alpha_6-1}\, dt,$$

where $dt = t_1\, dt_2 \wedge dt_3 - t_2\, dt_1 \wedge dt_3 + t_3\, dt_1 \wedge dt_2$, the function $w = \phi^{-1}\Phi$ is a function only of x. Take $\{t_1+t_2+t_3=1,\ t_i \geqq 0\} \subset \mathbf{R}^3$ as the range of integration. Assume that the

real parts of α_2, α_3, α_5, α_6, and $1-\alpha_{234}$ are positive and that α_{234} and α_4 are not integers. Then, by use of the binomial expansion $(1-y)^{\alpha}=\sum((-\alpha, m)/(1, m))y^m$, we get the identity

$$\Phi(u, \alpha)=\phi(u, \alpha)w(x, \alpha). \tag{4.6}$$

We next rewrite the differential equations (1.4)-(1.6) with respect to w. Since equations (1.4) and (1.5) are trivial for w, thanks to the invariance (1.2) and (1.3), we deal with (1.6). Put

$$\theta_{ap}=u_{ap}\frac{\partial}{\partial u_{ap}} \quad \text{for } 1\leqq a\leqq 3 \quad \text{and} \quad 4\leqq p\leqq 6,$$

$$\delta_i=x^i\frac{\partial}{\partial x^i} \quad \text{for } 1\leqq i\leqq 4 \quad \text{and} \quad \delta=\delta_1+\delta_2+\delta_3+\delta_4.$$

LEMMA 4.1. *The vector fields* θ_{ap} $(1\leqq a\leqq 3, 4\leqq p\leqq 6)$ *acts on* $w=\phi^{-1}\Phi$ *as differential operators listed below*:

$$\begin{aligned}
&\theta_{14}=\alpha_{234}-1+\delta, && \theta_{15}=\alpha_5-1-\delta_1-\delta_3, && \theta_{16}=\alpha_6-1-\delta_2-\delta_4,\\
&\theta_{24}=-\alpha_2-\delta_1-\delta_2, && \theta_{25}=\delta_1, && \theta_{26}=\delta_2,\\
&\theta_{34}=-\alpha_3-\delta_3-\delta_4, && \theta_{35}=\delta_3, && \theta_{36}=\delta_4.
\end{aligned}$$

Proof. See (4.1)-(4.3). □

Assume $1\leqq a, b\leqq 3$, and $4\leqq p, q\leqq 6$. Then the identity (2.8) shows that the equations (1.6) become

$$\frac{\partial^2}{\partial u_{ap}\,\partial u_{bq}}-\frac{\partial^2}{\partial u_{bp}\,\partial u_{aq}}=0, \qquad 1\leqq a, b\leqq 3, \quad 4\leqq p, q\leqq 6. \tag{4.7}$$

It is not hard to see that the equations (1.6) for another set of indices reduce to (4.7). Notice that this presentation of the system is called the canonical system in [H]. In terms of θ_{ap}, (4.7) is written as

$$u_{aq}u_{bp}\theta_{ap}\theta_{bq}-u_{ap}u_{bq}\theta_{aq}\theta_{bp}=0.$$

So we have the following proposition.

PROPOSITION 4.2. *The system of equations for* w *consists of the following nine differential equations*:

$$\begin{aligned}
&(\alpha_{234}-1+\delta)\delta_1-x^1(\alpha_5-1-\delta_1-\delta_3)(-\alpha_2-\delta_1-\delta_2)=0,\\
&(\alpha_{234}-1+\delta)\delta_2-x^2(\alpha_6-1-\delta_2-\delta_4)(-\alpha_2-\delta_1-\delta_2)=0,\\
&(\alpha_{234}-1+\delta)\delta_3-x^3(\alpha_5-1-\delta_1-\delta_3)(-\alpha_3-\delta_3-\delta_4)=0,\\
&(\alpha_{234}-1+\delta)\delta_4-x^4(\alpha_6-1-\delta_2-\delta_4)(-\alpha_3-\delta_3-\delta_4)=0,\\
&x^1(\alpha_5-1-\delta_1-\delta_3)\delta_2-x^2(\alpha_6-1-\delta_2-\delta_4)\delta_1=0,\\
&x^3(\alpha_5-1-\delta_1-\delta_3)\delta_4-x^4(\alpha_6-1-\delta_2-\delta_4)\delta_3=0,\\
&x^1(\alpha_2+\delta_1+\delta_2)\delta_3-x^3(\alpha_3+\delta_3+\delta_4)\delta_1=0,\\
&x^2(\alpha_2+\delta_1+\delta_2)\delta_4-x^4(\alpha_3+\delta_3+\delta_4)\delta_2=0,\\
&x^2x^3\delta_1\delta_4-x^1x^4\delta_2\delta_3=0.
\end{aligned}$$

Since this system is of rank 6, by the general theory [SY2] for a system whose rank is greater than the number of independent variables by two, a *conformal structure* is attached to this system. The associated conformal tensor has an expression such as $g_{ij}\,dx^i\,dx^j$ when the equations are written in the form

$$\frac{\partial^2 w}{\partial x^i\,\partial x^j}=g_{ij}\frac{\partial^2 w}{\partial x^1\,\partial x^4}+\text{terms of lower degree of differentiation.}$$

It is known that the conformal class does not depend on the choice of variables. By an easy calculation we have the following proposition.

PROPOSITION 4.3. *The conformal tensor $g_{ij}\,dx^i\,dx^j$ associated with the above system is given by*

$$g_{11}=-\frac{(x^4-x^2)x^3}{(x^1-x^3)x^4}-\frac{(x^4-x^3)x^2}{(x^1-x^2)x^4}+\frac{x^2x^3-x^4}{(1-x^1)x^4},$$

$$g_{22}=-\frac{(x^4-x^3)x^2}{(x^1-x^2)x^4}-\frac{(x^1-x^3)x^2}{(x^4-x^2)x^1}+\frac{(x^1x^4-x^3)x^2}{(1-x^2)x^4x^1},$$

$$g_{33}=-\frac{(x^4-x^2)x^3}{(x^1-x^3)x^4}-\frac{(x^1-x^2)x^3}{(x^4-x^3)x^1}+\frac{(x^1x^4-x^2)x^3}{(1-x^3)x^4x^1},$$

$$g_{44}=-\frac{(x^1-x^3)x^2}{(x^4-x^2)x^1}-\frac{(x^1-x^2)x^3}{(x^4-x^3)x^1}+\frac{x^2x^3-x^1}{(1-x^4)x^1},$$

$$g_{12}=g_{21}=\frac{(x^4-x^3)x^2}{(x^1-x^2)x^4},\quad g_{13}=g_{31}=\frac{(x^4-x^2)x^3}{(x^1-x^3)x^4},\quad g_{14}=g_{41}=1,$$

$$g_{23}=g_{32}=\frac{x^2x^3}{x^1x^4},\quad g_{24}=g_{42}=\frac{(x^1-x^3)x^2}{(x^4-x^2)x^1},\quad g_{34}=g_{43}=\frac{(x^1-x^2)x^3}{(x^4-x^3)x^1}.$$

Remark. This tensor is conformally flat because of the following argument. Consider the mapping defined by six independent solutions of the system. When $\alpha_i=\frac{1}{2}$ for $1\leqq i\leqq 6$, this mapping is, as we have seen in [MSY1], the period map of a four-dimensional family of $K3$ surfaces. Since the period map satisfies Riemann's equality, the image is contained in a quadratic hypersurface in $\mathbf{CP}^5$. On the other hand, it was shown in [SY2] that the associated tensor is conformally flat if the image is contained in a quadratic hypersurface. Hence, the above tensor, which is independent of values of α_i, is conformally flat.

The system of differential equations for the integrals on the stratum $\{x^2=0\}$ of codimension 1 is computed similarly.

PROPOSITION 4.4. *The system of differential equations on the stratum $\{x^2=0\}$ is of rank 5 and consists of five equations*:

$$(\alpha_{234}-1+\delta')\delta_1-x^1(\alpha_5-1-\delta_1-\delta_3)(-\alpha_2-\delta_1)=0,$$

$$(\alpha_{234}-1+\delta')\delta_3-x^3(\alpha_5-1-\delta_1-\delta_3)(-\alpha_3-\delta_3-\delta_4)=0,$$

$$(\alpha_{234}-1+\delta')\delta_4-x^4(\alpha_6-1-\delta_4)(-\alpha_3-\delta_3-\delta_4)=0,$$

$$x^3(\alpha_5-1-\delta_1-\delta_3)\delta_4-x^4(\alpha_6-1-\delta_4)\delta_3=0,$$

$$x^1(\alpha_2+\delta_1)\delta_3-x^3(\alpha_3+\delta_3+\delta_4)\delta_1=0,$$

where $\delta'=\delta_1+\delta_3+\delta_4$.

5. Contiguity relations of $E(3,6)$. We will compute the contiguity relations of $E(3,6)$ with respect to the function $w(x,\alpha)$. For simplicity we use the notation $(\alpha_j^+\alpha_l^-)$ to denote the parameter $\alpha+1_j-1_l$ as in § 3.

Our task is to express (2.6) in terms of the operators θ_{ap}. Note that X_{ii} is a scalar multiplication and that this yields a trivial relation. Define Y_{jl} by

$$Y_{jl}w=\phi^{-1}X_{jl}(\phi w)|_{u_{14}=u_{24}=u_{34}=u_{15}=u_{16}=1}. \tag{5.1}$$

We can see that

$$\phi^{-1}X_{jl}(\phi w)=M_{jl}(u)\,Y_{jl}w, \tag{5.2}$$

where $M_{jl}(u)=\phi(u,\alpha+1_j-1_l)/\phi(u,\alpha)$. For the subspace $\mathfrak{g}_1$ generated by $\partial/\partial u_{ap}$, one of the contiguity relations is

$$\frac{\partial}{\partial u_{14}}(\phi w)(u,\alpha)=(\alpha_4-1)(\phi w)(\alpha_1^+\alpha_4^-)=\frac{1}{u_{14}}(\alpha_4-1)\phi(u,\alpha)w(\alpha_1^+\alpha_4^-).$$

By Lemma 4.1, this yields

$$Y_{14}w=(\alpha_{234}-1+\delta)w(x,\alpha)=(\alpha_4-1)w(x,\alpha_1^+\alpha_4^-) \quad \text{and} \quad M_{14}=1/u_{14}.$$

Similarly, for X_{pq} in $\mathfrak{g}_0$, the identity

$$X_{pq}=\frac{u_{1p}}{u_{1q}}\theta_{1q}+\frac{u_{2p}}{u_{2q}}\theta_{2q}+\frac{u_{3p}}{u_{3q}}\theta_{3q}$$

in Lemma 2.3 shows, for example,

$$X_{45}(\phi w)=\frac{u_{14}}{u_{15}}(\alpha_5-1)\phi w+\phi\left\{\frac{u_{24}}{u_{25}}\delta_1+\frac{u_{34}}{u_{35}}\delta_3-\frac{u_{14}}{u_{15}}(\delta_1+\delta_2)\right\}w.$$

On the other hand, this is equal to $(\alpha_5-1)(\phi w)(\alpha_4^+\alpha_5^-)$. Hence we have

$$Y_{45}w=(\alpha_5-1)w+\{(1-x^1)\partial_1+(1-x^3)\partial_3\}w=(\alpha_5-1)w(\alpha_4^+\alpha_5^-),$$
$$M_{45}=u_{14}/u_{15}.$$

As for the operator X_{ab} in $\mathfrak{g}_0'$, the identity

$$-X_{ab}=\delta_{ab}+\frac{u_{b4}}{u_{a4}}\theta_{a4}+\frac{u_{b5}}{u_{a5}}\theta_{a5}+\frac{u_{b6}}{u_{a6}}\theta_{a6}$$

gives, for example,

$$-X_{12}=\frac{u_{24}}{u_{14}}(\alpha_{234}-1+\delta)+\frac{u_{25}}{u_{15}}(\alpha_5-1-\delta_1-\delta_3)+\frac{u_{26}}{u_{16}}(\alpha_6-1-\delta_2-\delta_4),$$

which yields

$$Y_{12}w=-\{\alpha_{234}-1+\delta+x^1(\alpha_5-1-\delta_1-\delta_3)+x^2(\alpha_6-1-\delta_2-\delta_4)\}w=(\alpha_2-1)w(\alpha_1^+\alpha_2^-),$$
$$M_{12}=u_{24}/u_{14}.$$

We can similarly carry out the computation for the elements in $\mathfrak{g}_{-1}$. Thus we have the following proposition.

PROPOSITION 5.1. *The Aomoto-Gel'fand* HGF w *for the system* $E(3,6)$ *satisfies contiguity relations listed in Table* 2. *The diagonal operators* Y_{jj}, $1\leqq j\leqq 6$, *are scalar multiplications by* α_j-1.

Each line reads as $Y_{jl}w(x,\alpha)=(\alpha_j-1)w(x,\alpha_j^+\alpha_l^-)$. The first column denotes indices of Y_{jl}. The third column denotes the factor $M(u)$ that appears in equation (5.2). ∂_i denotes $\partial/\partial x^i$.

6. Strata in $G_{3,6}$. In the study of HGF's associated with $G_{3,6}$, Gel'fand and Graev [GGR1] gave explicit expressions of the integrals (1.1) for each stratum of $G_{3,6}$. They have shown that the space $G_{3,6}$ has 15 types of strata: one is the open stratum discussed in previous sections and the others are denoted by A, $B1$, $B2$, $B3$, $B4$, $C1$, $C2$, $C3$, $C4$, I, II, III, IV, and V. We do not cite their definitions here (refer to [GGR1, p. 299]), but we give a few remarks.

The dimension of a stratum belonging to respective types is 3 for A, 2 for B's, 1 for C's, and 0 for I-V. A stratum belonging to A is given by the equation $u_{14}=0$ in the coordinates (4.1) and the corresponding system of differential equations has been given in Proposition 4.4. Appell's system F_1, which is the system F_D in two variables,

TABLE 2

Index	Y	M
[12]	$-(\alpha_{234}-1+\delta)-x^1(\alpha_5-1-\delta_1-\delta_3)-x^2(\alpha_6-1-\delta_2-\delta_4)$	u_{24}/u_{14}
[21]	$\alpha_2+(x^1-1)\partial_1+(x^2-1)\partial_2$	u_{14}/u_{24}
[13]	$-(\alpha_{234}-1+\delta)-x^3(\alpha_5-1-\delta_1-\delta_3)-x^4(\alpha_6-1-\delta_2-\delta_4)$	u_{34}/u_{14}
[31]	$\alpha_3+(x^3-1)\partial_3+(x^4-1)\partial_4$	u_{14}/u_{34}
[23]	$\alpha_2+(x^1-x^3)\partial_1+(x^2-x^4)\partial_2$	u_{34}/u_{24}
[32]	$\alpha_3+(x^3-x^1)\partial_3+(x^4-x^2)\partial_4$	u_{24}/u_{34}
[14]	$\alpha_{234}-1+\delta$	$1/u_{14}$
[15]	$\alpha_5-1-\delta_1-\delta_3$	$1/u_{15}$
[16]	$\alpha_6-1-\delta_2-\delta_4$	$1/u_{16}$
[24]	$-\alpha_2-\delta_1-\delta_2$	$1/u_{24}$
[25]	∂_1	$u_{14}/u_{15}u_{24}$
[26]	∂_2	$u_{14}/u_{16}u_{24}$
[34]	$-\alpha_3-\delta_3-\delta_4$	$1/u_{34}$
[35]	∂_3	$u_{14}/u_{15}u_{34}$
[36]	∂_4	$u_{14}/u_{16}u_{34}$
[45]	$\alpha_5-1+(1-x^1)\partial_1+(1-x^3)\partial_3=: Y_{45}$	u_{14}/u_{15}
[54]	$\alpha_{234}-1+\delta-x^1(\alpha_2+\delta_1+\delta_2)-x^3(\alpha_3+\delta_3+\delta_4)=: Y_{54}$	u_{15}/u_{14}
[46]	$\alpha_6-1+(1-x^2)\partial_2+(1-x^4)\partial_4=: Y_{46}$	u_{14}/u_{16}
[64]	$\alpha_{234}-1+\delta-x^2(\alpha_2+\delta_1+\delta_2)-x^4(\alpha_3+\delta_3+\delta_4)=: Y_{64}$	u_{16}/u_{14}
[56]	$\alpha_6-1+(x^1-x^2)\partial_2+(x^3-x^4)\partial_4=: Y_{56}$	u_{15}/u_{16}
[65]	$\alpha_5-1+(x^2-x^1)\partial_1+(x^4-x^3)\partial_3=: Y_{65}$	u_{16}/u_{15}
[41]	$\alpha_{234}-1+\sum_{i=1}^4(x^i-1)\partial_i$	u_{14}
[42]	$-\alpha_4-x^1Y_{45}-x^2Y_{46}$	u_{24}
[43]	$-\alpha_4-x^3Y_{45}-x^4Y_{46}$	u_{34}
[51]	$-\alpha_5-Y_{54}-Y_{56}$	u_{15}
[52]	$-\alpha_5x^1-Y_{54}-x^2Y_{56}$	$u_{15}u_{24}/u_{14}$
[53]	$-\alpha_5x^3-Y_{54}-x^4Y_{56}$	$u_{15}u_{34}/u_{14}$
[61]	$-\alpha_6-Y_{64}-Y_{65}$	u_{16}
[62]	$-\alpha_6x^2-Y_{64}-x^1Y_{65}$	$u_{16}u_{24}/u_{14}$
[63]	$-\alpha_6x^4-Y_{64}-x^3Y_{65}$	$u_{16}u_{34}/u_{14}$

and Horn's system G_2 (cf. [E, § 5.7]) appear both for strata of types B1 and B2. Appell's systems F_2 and F_3 and Horn's system H_2 [E] appear for strata of type B3. The definition of F_3 will be recalled later. The system associated with a stratum of type B4 is a tensor product of two Gauss HGDE's with different parameters. Systems for $C1$, $C2$, or $C3$ are known to be the Gauss HGDE's. The final system associated with a stratum of type $C4$ is Goursat's (so-called generalized) HGF ${}_3F_2$ (cf. § 9). Refer to [GGR3] for these matters and to [MSY2] for another aspect of $G_{3,6}$.

The fact that HGF's F_2, F_3, and H_2 belong to the same type is remarkable. In view of the invariance (1.2) and (1.3), it says that the associated systems are related by certain transformations of coordinates and unknown functions. See [AK, p. 51] for the relation between functions F_2 and F_3 and [SY1, § 5.3] for the relations among systems F_2, F_3, and H_2. Moreover, one and the same function may appear on different strata of the same or different type and may have different expressions depending on the choice of coordinates; this fact explains certain transformation rules for this function such as Kummer's identities for the Gauss HGF. However, because we are concerned only with functions or systems and not with relations among their different expressions, we can restrict our consideration to the systems associated with strata chosen arbitrarily from each type, i.e., to the systems F_1, F_3, ${}_3F_2$, and the Gauss HGDE for types B and C. Since we have treated F_1 and the Gauss HGDE in § 3, we will consider F_3 and ${}_3F_2$ in the following sections.

Remark. A relation between the systems F_1 and G_2 is given as follows. Let $F_1(\alpha, \beta, \beta', \gamma)$ and $G_2(\alpha_1, \alpha_1', \beta_1, \beta_1')$ denote the systems of differential equations just given in [E, eq. (9), p. 233, and eq. (14), p. 234]. Then, for any solution $w(x, y)$ of G_2, the function $z(x, y) = y^{\alpha_1'} w(-x, -1/y)$ is a solution of F_1 and vice versa with the correspondence of parameters given by $\alpha = \alpha_1' + \beta_1'$, $\beta = \alpha_1$, $\beta' = \alpha_1'$, and $\gamma = \alpha_1' - \beta_1 + 1$.

7. Contiguity relations of Appell's F_3. Appel's HGF F_3 is given by the series

$$F_3(\alpha, \alpha', \beta, \beta', \gamma; x, y) = \sum_{m,n=0}^{\infty} \frac{(\alpha, m)(\alpha', n)(\beta, m)(\beta', n)}{(\gamma, m+n)(1, m)(1, n)} x^m y^n.$$

The definition of $F(x, \alpha)$ in (4.5) shows

$$F(x^1, 0, 0, x^4; \alpha_1, \cdots, \alpha_6) = F_3(\alpha, \alpha', \beta, \beta', \gamma; x^1, x^4)$$

and

$$c((\alpha_1, \cdots, \alpha_6)) = \Gamma(1-\gamma)\Gamma(\alpha)\Gamma(\alpha')/\Gamma(\alpha + \alpha' - \gamma + 1),$$

where

$$\alpha = \alpha_2, \quad \alpha' = \alpha_3, \quad \beta = 1 - \alpha_5, \quad \beta' = 1 - \alpha_6, \quad \gamma = \alpha_{234}. \tag{7.1}$$

We have seen in § 5 that the contiguity relation with respect to $w = c(\alpha)F(x, \alpha)$ is written in the form

$$Y(c(\alpha)F(x, \alpha)) = a_Y c(\alpha^Y) F(x, \alpha^Y) \tag{7.2}$$

for an operator Y; a_Y is a linear form on α and $\alpha^Y = \alpha + 1_j - 1_l = (\alpha_j^+ \alpha_l^-)$ for some j and l. If we can restrict both sides of this identity to the stratum $\{x^2 = x^3 = 0\}$, then this readily gives a part of contiguity relations of $F(x^1, 0, 0, x^4; \alpha)$ and thus of F_3. In reference to Table 2, the indices of such operators are

$$12, 13, 14, 15, 16, 24, 25, 34, 36, 52, 54, 63, 64. \tag{7.3}$$

Other operators contain terms like $\partial_2 F$ or $\partial_3 F$, whose restrictions will be examined later.

TABLE 3.1

Index	Y	a_Y	α^Y
[14]	$\gamma - 1 + \delta_x + \delta_y$	$\gamma - \alpha - \alpha' - 1$	γ^-
[15]	$-\beta - \delta_x$	$-\beta$	β^+
[16]	$-\beta' - \delta_y$	$-\beta'$	β'^+
[24]	$-\alpha - \delta_x$	$\gamma - \alpha - \alpha' - 1$	α^+
[25]	∂_x	$-\beta$	$\alpha^+\beta^+\gamma^+$
[34]	$-\alpha' - \delta_y$	$\gamma - \alpha - \alpha' - 1$	α'^+
[36]	∂_y	$-\beta'$	$\alpha'^+\beta'^+\gamma^+$
[52]	$1 - \gamma + (\alpha + \beta - 1)x + (x-1)\delta_x - \delta_y$	$\alpha - 1$	$\alpha^-\beta^-\gamma^-$
[12]	$1 - \gamma + \beta x + (x-1)\delta_x - \delta_y$	$\alpha - 1$	$\alpha^-\gamma^-$
[54]	$\gamma - 1 - \alpha x - (x-1)\delta_x + \delta_y$	$\gamma - \alpha - \alpha' - 1$	$\beta^-\gamma^-$
[64]	$\gamma - 1 - \alpha' y - (y-1)\delta_y + \delta_x$	$\gamma - \alpha - \alpha' - 1$	$\beta'^-\gamma^-$
[13]	$1 - \gamma + \beta' y + (y-1)\delta_y - \delta_x$	$\alpha' - 1$	$\alpha'^-\gamma^-$
[63]	$1 - \gamma + (\alpha' + \beta' - 1)y + (y-1)\delta_y - \delta_x$	$\alpha' - 1$	$\alpha'^-\beta'^-\gamma^-$
[11]	$\beta + \beta' - \gamma$	$\beta + \beta' - \gamma$	—
[22]	$\alpha - 1$	$\alpha - 1$	—
[33]	$\alpha' - 1$	$\alpha' - 1$	—
[44]	$\gamma - \alpha - \alpha' - 1$	$\gamma - \alpha - \alpha' - 1$	—
[55]	$-\beta$	$-\beta$	—
[66]	$-\beta'$	$-\beta'$	—

PROPOSITION 7.1. *The operators with indices listed in* (7.3) *yield the following contiguity relations in Table* 3.1 *with respect to the function* $\{\Gamma(\alpha)\Gamma(\alpha')\Gamma(1-\gamma)/\Gamma(\alpha+\alpha'-\gamma+1)\}F_3(\alpha, \alpha', \beta, \beta', \gamma; x, y)$. *Trivial relations for the operators* X_{ii} *are also listed.*

We use the notation $\partial_x = \partial/\partial x$, $\partial_y = \partial/\partial y$, $\delta_x = x\partial_x$, and $\delta_y = y\partial_y$.

Proof. Transformation from Table 2 to Table 3.1 is done by taking care of the identity (7.1) and the relation $\sum \alpha_i = 3$. The operator X_{14}, for example, increases (α_i) at $i=1$ and decreases at $i=4$. This implies the decrease of γ by 1, which is denoted by γ^-. The multiplier $\alpha_4 - 1$ is equal to $\gamma - \alpha - \alpha' - 1$. This shows the first line. □

The Lie-algebraic structure of the set of operators in this table will be recovered in the following way. Introduce supplementary variables v_{14}, v_{15}, v_{16}, v_{24}, v_{34} and define operators $\tau_{ij} = v_{ij}(\partial/\partial v_{ij})$. Put

$$
\begin{aligned}
&E_{11} = -\tau_{15} - \tau_{16} - \tau_{14} - 1, \quad E_{22} = -\tau_{24} - 1, \quad E_{33} = -\tau_{34} - 1,\\
&E_{44} = \tau_{14} + \tau_{24} + \tau_{34}, \quad E_{55} = \tau_{15}, \quad E_{66} = \tau_{16},\\
&E_{12} = \frac{v_{24}}{v_{14}}\{-\tau_{14} - \delta_x - \delta_y + x(-\tau_{15} + \delta_x)\},\\
&E_{13} = \frac{v_{34}}{v_{14}}\{-\tau_{14} - \delta_x - \delta_y + y(-\tau_{16} + \delta_y)\},\\
&E_{14} = \frac{1}{v_{14}}\{\tau_{14} + \delta_x + \delta_y\}, \quad E_{15} = \frac{1}{v_{15}}\{\tau_{15} - \delta_x\}, \quad E_{16} = \frac{1}{v_{16}}\{\tau_{16} - \delta_y\},\\
&E_{24} = \frac{1}{v_{24}}\{\tau_{24} - \delta_x\}, \qquad E_{25} = \frac{v_{14}}{v_{15}v_{24}}\partial_x,\\
&E_{34} = \frac{1}{v_{34}}\{\tau_{34} - \delta_y\}, \qquad E_{36} = \frac{v_{14}}{v_{16}v_{34}}\partial_y,\\
&E_{52} = \frac{v_{15}v_{24}}{v_{14}}\{x(1 + \tau_{15} - \tau_{24} + \delta_x) - \tau_{14} - \delta_x - \delta_y\},\\
&E_{54} = \frac{v_{15}}{v_{14}}\{\tau_{14} + \delta_x + \delta_y - x(-\tau_{24} + \delta_x)\},\\
&E_{63} = \frac{v_{16}v_{34}}{v_{14}}\{y(1 + \tau_{16} - \tau_{34} + \delta_y) - \tau_{14} - \delta_x - \delta_y\},\\
&E_{64} = \frac{v_{16}}{v_{34}}\{\tau_{14} + \delta_x + \delta_y - y(-\tau_{34} + \delta_y)\}.
\end{aligned}
\tag{7.4}
$$

Then we have the following proposition.

PROPOSITION 7.2. *Differential operators* E_{ij} *listed in* (7.4) *generate a Lie algebra of dimension* 19 *with relations*

$$[E_{ij}, E_{i'j'}] = \delta_{ji'}E_{ij'} - \delta_{j'i}E_{i'j}.$$

Proof. Let X be one of X_{jl} with indices in Table 3.1 and Y be the corresponding operator in Table 2. Then (5.2) says $\phi^{-1}X\phi \cdot w = M_X Y \cdot w$ for the corresponding factor M_X. The set of operators $\phi^{-1}X\phi$ for such X certainly forms a Lie algebra of dimension 19. Hence the proof will be completed by expressing $\phi^{-1}X\phi$ in terms of variables $u_{14}, \cdots, u_{34}$ and x^i and by restricting them onto $\{x^2 = x^3 = 0\}$. While $\phi^{-1}X\phi = X(\log\phi) + X$ as an operator, the operator $M_X^{-1}X$ acts on functions of x as a vector field $\underline{X}$ with respect to variables x and $M_X^{-1}X(\log\phi)$ is seen to be written as $\theta_{\underline{X}}(\log\phi)$,

where θ_X is a linear combination of τ_{jl} defined above. Here we use v for u to avoid confusion. Namely, we have $\phi^{-1}X\phi = M_X\theta_X(\log\phi) + M_X\underline{X} = \phi^{-1}\cdot M_X(\theta_X+\underline{X})\cdot\phi$. Explicit expressions of $E_X = M_X(\theta_X+\underline{X})$ are easily computed to get identities in (7.4). □

Remark. Miller in [M2] has already found nine contiguity relations in Table 3.1: 14, 15, 16, 24, 25, 34, 36, 52, 63. Relations with indices 12, 13, 54, and 64 follow from the Lie algebra structure; for example, $[E_{15}, E_{52}] = E_{12}$.

We next deal with the operators including ∂_x and/or ∂_y. Define two operators by

$$U_{26} = \alpha\beta' - \beta'(1-x)\partial_x - \alpha(1-y)\partial_y + \Delta\partial_x\partial_y, \tag{7.5}$$

$$U_{35} = \alpha'\beta - \alpha(1-x)\partial_x - \beta(1-y)\partial_y + \Delta\partial_x\partial_y, \tag{7.6}$$

where $\Delta = xy - x - y$. Then we see the following lemma.

LEMMA 7.3. *These operators give contiguity relations of second order*:

$$U_{26}\{cF_3\} = \beta'(\alpha' - \gamma + \beta)\{cF_3\}(\alpha^+\beta'^+\gamma^+),$$

$$U_{35}\{cF_3\} = \alpha'(\beta' - \gamma + \alpha)\{cF_3\}(\alpha'^+\beta^+\gamma^+).$$

A proof will be given in the last section. Now assume

$$(\gamma - \alpha' - \beta)(\gamma - \alpha - \beta') \neq 0 \tag{7.7}$$

and define operators by

$$Z_{26} = \frac{1}{\gamma - \alpha' - \beta}\,U_{26} \quad \text{and} \quad Z_{35} = \frac{1}{\gamma - \alpha - \beta'}\,U_{35}. \tag{7.8}$$

Then we have the remaining contiguity relations by the use of Table 2:

PROPOSITION 7.4. *Appell's* HGF $\{\Gamma(\alpha)\Gamma(\alpha')\Gamma(1-\gamma)/\Gamma(\alpha+\alpha'-\gamma+1)\}\times F_3(\alpha, \alpha', \beta, \beta', \gamma; x, y)$ *satisfies the contiguity relations listed in Table* 3.1 *and Table* 3.2.

Each line reads as $Z(cF_3) = a_Z\cdot(cF)(\alpha^Z)$. The first column in Table 3.2 gives the correspondence with Table 2.

TABLE 3.2

Index	Z	a_Z	α^Z
[21]	$\alpha + (x-1)\partial_x - Z_{26}$	$\beta+\beta'-\gamma$	$\alpha^+\gamma^+$
[31]	$\alpha' + (y-1)\partial_y - Z_{35}$	$\beta+\beta'-\gamma$	$\alpha'^+\gamma^+$
[23]	$\alpha + x\partial_x - yZ_{26}$	$\alpha'-1$	$\alpha^+\alpha'^-$
[32]	$\alpha' + y\partial_y - xZ_{35}$	$\alpha-1$	$\alpha^-\alpha'^+$
[45]	$-\beta + (1-x)\partial_x + Z_{35}$	$-\beta$	$\beta^+\gamma^+$
[46]	$-\beta' + (1-y)\partial_y + Z_{26}$	$-\beta'$	$\beta'^+\gamma^+$
[56]	$-\beta' - y\partial_y + xZ_{26}$	$-\beta'$	$\beta^-\beta'^+$
[65]	$-\beta - x\partial_x + yZ_{35}$	$-\beta$	$\beta^+\beta'^-$
[26]	$\frac{1}{\gamma-\alpha'-\beta}(\alpha\beta' - \beta'(1-x)\partial_x - \alpha(1-y)\partial_y + \Delta\partial_{xy}^2) =: Z_{26}$	$-\beta'$	$\alpha^+\beta'^+\gamma^+$
[35]	$\frac{1}{\gamma-\beta'-\alpha}(\beta\alpha' - \alpha'(1-x)\partial_x - \beta(1-y)\partial_y + \Delta\partial_{xy}^2) =: Z_{35}$	$-\beta$	$\beta^+\alpha'^+\gamma^+$
[41]	$\alpha+\alpha'+\beta+\beta'-\gamma+(x-1)\partial_x+(y-1)\partial_y - Z_{26} - Z_{35}$	$\beta+\beta'-\gamma$	γ^+
[42]	$\alpha+\alpha'-\gamma+\beta x + x(x-1)\partial_x - xZ_{35}$	$\alpha-1$	α^-
[43]	$\alpha+\alpha'-\gamma+\beta' y + y(y-1)\partial_y - yZ_{26}$	$\alpha'-1$	α'^-
[51]	$\beta+\beta'-\gamma+\alpha x + x(x-1)\partial_x - xZ_{26}$	$\beta+\beta'-\gamma$	β^-
[61]	$\beta+\beta'-\gamma+\alpha' y + y(y-1)\partial_y - yZ_{35}$	$\beta+\beta'-\gamma$	β'^-
[53]	$1-\gamma+\alpha x+\beta' y + x(x-1)\partial_x + y(y-1)\partial_y - xyZ_{26}$	$\alpha'-1$	$\alpha'^-\beta^-\gamma^-$
[62]	$1-\gamma+\beta x+\alpha' y + x(x-1)\partial_x + y(y-1)\partial_y - xyZ_{35}$	$\alpha-1$	$\beta'^-\alpha^-\gamma^-$

Proof. From [26] in Table 2 we see

$$\frac{\partial}{\partial x^2}(cF)(\alpha)=(\alpha_6-1)(cF)(\alpha_2^+\alpha_6^-).$$

Restriction to $\{x^2=x^3=0\}$ then shows

$$\frac{\partial}{\partial x^2}(cF)(\alpha)\Big|_{x^2=x^3=0}=(\alpha_6-1)c(\alpha_2^+\alpha_6^-)F_3(\alpha_2^+\alpha_6^-;x,y).$$

The change of parameters $(\alpha_2^+\alpha_6^-)$ is equivalent to $(\alpha^+\beta'^+\gamma^+)$ by (7.1). Hence, by Lemma 7.3, the restriction of the function $(\partial/\partial x^2)(cF)$ onto the stratum $\{x^2=x^3=0\}$ can be replaced by the function $Z_{26}(cF)$. The same applies for [35]. Then, with reference to Table 2, we obtain whole relations. □

Remark. The relations in Tables 3.1 and 3.2 form a complete set of contiguity relations in the sense that every contiguous function of F_3 is representable as a linear combination of the original F_3 and its derivatives up to order 2 by a certain combination of differential operators appearing in these tables. See the last columns of these tables.

8. Different expression of contiguity relations of $E(3,6)$. The coordinate system (x^i) introduced in § 4 is not appropriate for the study of the system ${}_3F_2$. In this section we define new coordinate system (y^i) on the quotient of the open stratum in $G_{3,6}$.

Consider the set of points (I_3, u) defined in (4.1). Assume in this section that u_{14}, u_{16}, u_{24}, u_{25}, and u_{35} are not equal to zeros and define

$$y^1=\frac{u_{25}u_{34}}{u_{24}u_{35}},\quad y^2=\frac{u_{14}u_{26}}{u_{16}u_{24}},\quad y^3=\frac{u_{15}u_{24}}{u_{14}u_{25}},\quad y^4=\frac{u_{14}u_{25}u_{36}}{u_{16}u_{24}u_{35}}. \tag{8.1}$$

Then $y=(y^i)$ is a system of coordinates around the single H_6-orbit stratum $\{u_{15}=u_{26}=u_{34}=u_{36}=0\}$. It defines a local section of $Z_{3,6}\to GL_3\backslash Z_{3,6}/H_6$ by associating a matrix

$$\begin{pmatrix} 1 & & & 1 & y^3 & 1 \\ & 1 & & 1 & 1 & y^2 \\ & & 1 & y^1 & 1 & y^4 \end{pmatrix}. \tag{8.2}$$

The connection of y with x is seen to be

$$y^1=\frac{x^1}{x^3},\quad y^2=x^2,\quad y^3=\frac{1}{x^3},\quad y^4=\frac{x^1x^4}{x^3}. \tag{8.3}$$

We use the notation

$$\delta_i=y^i\frac{\partial}{\partial y^i}\quad\text{and}\quad \partial_i=\frac{\partial}{\partial y^i},\qquad 1\leqq i\leqq 4,$$

which should not be confused with that defined in § 4. The integral $\varphi(u,\alpha)=\Phi((I_3,u),\alpha)$ has the expression

$$\varphi(u,\alpha)=\phi(u,\alpha)w(y,\alpha), \tag{8.4}$$

where

$$\phi(u,\alpha)=u_{14}^{\alpha_{2345}-2}u_{24}^{1-\alpha_{235}}u_{25}^{\alpha_{35}-1}u_{35}^{-\alpha_3}u_{16}^{\alpha_6-1} \tag{8.5}$$

and

$$\begin{aligned} w(y,\alpha)=\int\prod_{i=1}^{3}t_i^{\alpha_i-1}(t_1+t_2+y^1t_3)^{\alpha_4-1}(y^3t_1+t_2+t_3)^{\alpha_5-1} \\ \cdot(t_1+y^2t_2+y^4t_3)^{\alpha_6-1}(t_1\,dt_2\wedge dt_3-t_2\,dt_1\wedge dt_3+t_3\,dt_1\wedge dt_2). \end{aligned} \tag{8.6}$$

Here $\alpha_{2345}=\alpha_2+\alpha_3+\alpha_4+\alpha_5$ and so on.

Recall operators $\theta_{ap} = u_{ap}(\partial/\partial u_{ap})$ defined in § 4. We denote the operator $\phi^{-1}\theta_{ap}\phi$ acting on w by the same letter θ_{ap}. Then we see the following lemma, similar to Lemma 4.1.

LEMMA 8.1. *The operators θ_{ap} are given as follows*:

$$\theta_{14} = \alpha_{2345} - 2 + \delta_2 - \delta_3 + \delta_4, \quad \theta_{15} = \delta_3, \quad \theta_{16} = \alpha_6 - 1 - \delta_2 - \delta_4,$$

$$\theta_{24} = 1 - \alpha_{235} - \delta_1 - \delta_2 + \delta_3 - \delta_4, \quad \theta_{25} = \alpha_{35} - 1 + \delta_1 - \delta_3 + \delta_4, \quad \theta_{26} = \delta_2,$$

$$\theta_{34} = \delta_1, \quad \theta_{35} = -\alpha_3 - \delta_1 - \delta_4, \quad \theta_{36} = \delta_4.$$

We next define operators Y_{jl} by

$$\phi^{-1}X_{jl}(\phi w) = M_{jl}(u)\, Y_{jl} w, \tag{8.7}$$

where $M_{jl}(u) = \phi(u, \alpha + 1_j - 1_l)/\phi(u, \alpha)$. Then, similar to Proposition 5.1, we get the following.

PROPOSITION 8.2. *The Aomoto-Gel'fand* HGF w *for the system* $E(3, 6)$ *relative to the coordinate* y *satisfies the contiguity relations listed in Table* 4. *The diagonal operators* Y_{jj}, $1 \leqq j \leqq 6$, *are scalar multiplications by* $\alpha_j - 1$.

The table is read in the same manner as Table 2.

TABLE 4

Index	Y	M
[14]	$\alpha_{2345} - 2 + \delta_2 - \delta_3 + \delta_4$	$1/u_{14}$
[15]	∂_3	$u_{24}/u_{14}u_{25}$
[16]	$\alpha_6 - 1 - \delta_2 - \delta_4$	$1/u_{16}$
[24]	$1 - \alpha_{235} - \delta_1 - \delta_2 + \delta_3 - \delta_4$	$1/u_{24}$
[25]	$\alpha_{35} - 1 + \delta_1 - \delta_3 + \delta_4$	$1/u_{25}$
[26]	∂_2	$u_{14}/u_{16}u_{24}$
[34]	∂_1	$u_{25}/u_{24}u_{35}$
[35]	$-\alpha_3 - \delta_1 - \delta_4$	$1/u_{35}$
[36]	∂_4	$u_{14}u_{25}/u_{16}u_{24}u_{35}$
[12]	$2 - \alpha_{2345} + (1 - \alpha_6)y^2 + (y^2 - 1)(\delta_2 + \delta_4) + (y^3 - 1)\partial_3$	u_{24}/u_{14}
[21]	$\alpha_{235} - 1 + (1 - \alpha_{35})y^3 + (1 - y^3)(\delta_1 + \delta_4) + (y^2 - 1)\partial_2 + (y^3 - 1)\delta_3$	u_{14}/u_{24}
[13]	$(2 - \alpha_{2345})y^1 + (1 - \alpha_6)y^4 + (y^4 - y^1)(\delta_2 + \delta_4) + (y^1y^3 - 1)\partial_3$	$u_{24}u_{35}/u_{14}u_{25}$
[31]	$\alpha_3 y^3 + (y^1y^3 - 1)\partial_1 + (y^3y^4 - 1)\partial_4$	$u_{14}u_{25}/u_{24}u_{35}$
[23]	$1 - \alpha_{35} + (\alpha_{235} - 1)y^1 + (y^1 - 1)(\delta_1 - \delta_3 + \delta_4) + (y^1y^2 - y^4)\partial_2$	u_{35}/u_{25}
[32]	$\alpha_3 + (y^1 - 1)\partial_1 + (y^4 - y^2)\partial_4$	u_{25}/u_{35}
[45]	$\alpha_{35} - 1 - \alpha_3 y^1 + (1 - y^1)(\delta_1 + \delta_4) + (1 - y^3)\partial_3$	u_{24}/u_{25}
[54]	$1 - \alpha_{235} + (\alpha_{2345} - 2)y^3 + (y^3 - 1)(\delta_2 + \delta_4) + (1 - y^1)\partial_1 + (1 - y^3)\delta_3$	u_{25}/u_{24}
[46]	$\alpha_6 - 1 + (1 - y^2)\partial_2 + (y^1 - y^4)\partial_4$	u_{14}/u_{16}
[64]	$\alpha_{2345} - 2 + (1 - \alpha_{235})y^2 + (y^4 - y^1y^2)\partial_1 + (1 - y^2)(\delta_2 - \delta_3 + \delta_4)$	u_{16}/u_{14}
[56]	$(\alpha_6 - 1)y^3 + (1 - y^2y^3)\partial_2 + (1 - y^3y^4)\partial_4$	$u_{14}u_{25}/u_{16}u_{24}$
[65]	$(\alpha_{35} - 1)y^2 - \alpha_3 y^4 + (y^2 - y^4)(\delta_1 + \delta_4) + (1 - y^2y^3)\partial_3$	$u_{16}u_{24}/u_{14}u_{25}$
[41]	$-\alpha_4 - y^3 Y_{45} - Y_{46}$	u_{14}
[42]	$-\alpha_4 - Y_{45} - y^2 Y_{46}$	u_{24}
[43]	$-\alpha_4 y^1 - Y_{45} - y^4 Y_{46}$	$u_{24}u_{35}/u_{25}$
[51]	$-\alpha_5 y^3 - Y_{54} - Y_{56}$	$u_{14}u_{25}/u_{24}$
[52]	$-\alpha_5 - Y_{54} - y^2 Y_{56}$	u_{25}
[53]	$-\alpha_5 - y^1 Y_{54} - y^4 Y_{56}$	u_{35}
[61]	$-\alpha_6 - Y_{64} - y^3 Y_{65}$	u_{16}
[62]	$-\alpha_6 y^2 - Y_{64} - Y_{65}$	$u_{16}u_{24}/u_{14}$
[63]	$-\alpha_6 y^4 - y^1 Y_{64} - Y_{65}$	$u_{16}u_{24}u_{35}/u_{14}u_{25}$

9. Contiguity relations of Goursat's ${}_3F_2$. Goursat's HGF ${}_3F_2$ is given by the series

$$ {}_3F_2(a_1, a_2, a_3; b_1, b_2; y) = \sum_{n=0}^{\infty} \frac{(a_1, n)(a_2, n)(a_3, n)}{(b_1, n)(b_2, n)n!} y^n. $$

We use the abbreviation $(a) = (a_1, a_2, a_3; b_1, b_2)$. This function has an Euler integral representation (2.18) with a trivial change of notation. Hence the restriction of an integral of the form (8.2) to the one-dimensional stratum $\Gamma := \{y^1 = y^2 = y^3 = 0\}$ gives ${}_3F_2$:

$$ \begin{aligned} w(\alpha; 0, 0, 0, y^4) = & \int_{1+t_2 \geqq 0, t_2+t_3 \geqq 0, t_2 \leqq 0} t_2^{\alpha_2 - 1} t_3^{\alpha_3 - 1} \\ & \cdot (1+t_2)^{\alpha_4 - 1} (t_2+t_3)^{\alpha_5 - 1} (1+y^4 t_3)^{\alpha_6 - 1} \, dt_2 \wedge dt_3 \\ = & \gamma(\alpha) \cdot {}_3F_2(a; y), \end{aligned} \tag{9.1} $$

where

$$ \begin{aligned} & \gamma(\alpha) = (-1)^{\alpha_{25} - 1} \frac{\Gamma(\alpha_{235} - 1)\Gamma(\alpha_3)\Gamma(\alpha_4)\Gamma(\alpha_5)}{\Gamma(\alpha_{2345} - 1)\Gamma(\alpha_{35})}, \\ & y = -y^4, \\ & a_1 = \alpha_{235} - 1, \quad a_2 = \alpha_3, \quad a_3 = 1 - \alpha_6, \quad b_1 = \alpha_{2345} - 1, \quad b_2 = \alpha_{35}. \end{aligned} \tag{9.2} $$

With this notation the procedure of determining contiguity relations of ${}_3F_2$ is the same as that used in § 7. We first note that the operators Y_{jl} with indices

$$ 14,\ 16,\ 24,\ 25,\ 35,\ 36 \tag{9.3} $$

can be restricted to the stratum Γ. Namely we have the following proposition.

PROPOSITION 9.1. *The restriction of the differential operators Y_{jl} with indices in* (9.3) *to the stratum Γ yield the six contiguity relations of the function ${}_3F_2$:*

$$ \begin{aligned} & [14] \quad (\theta + b_1 - 1)_3F_2 = (b_1 - 1)_3F_2(b_1^-), \\ & [16] \quad (\theta + a_3)_3F_2 = a_3 \cdot {}_3F_2(a_3^+), \\ & [24] \quad (\theta + a_1)_3F_2 = a_1 \cdot {}_3F_2(a_1^+), \\ & [25] \quad (\theta + b_2 - 1)_3F_2 = (b_2 - 1)_3F_2(b_2^-), \\ & [35] \quad (\theta + a_2)_3F_2 = a_2 \cdot {}_3F_2(a_2^+), \\ & [36] \quad \frac{\partial}{\partial y} {}_3F_2 = \frac{a_1 a_2 a_3}{b_1 b_2} {}_3F_2(a_1^+ a_2^+ a_3^+; b_1^+ b_2^+). \end{aligned} \tag{9.4} $$

We next deal with the remaining operators. The key operators are Y_{15}, Y_{26}, and Y_{34}; namely, ∂_3, ∂_2, and ∂_1. We denote $y(\partial/\partial y)$ by θ and define three operators by

$$ Z_{15} = \frac{1}{1 - a_1} \{(1-y)\theta^2 + (b_{12} - 2 - a_{23}y)\theta - a_2 a_3 y + (b_1 - 1)(b_2 - 1)\}, \tag{9.5} $$

$$ Z_{26} = \frac{1}{(a_2 - b_1)y} \{(1-y)\theta^2 + (b_2 - 1 - a_{13}y)\theta - a_1 a_3 y\}, \tag{9.6} $$

$$ Z_{34} = \frac{1}{(a_3 - b_2)y} \{(1-y)\theta^2 + (b_1 - 1 - a_{12}y)\theta - a_1 a_2 y\}, \tag{9.7} $$

with the assumption

$$ (a_1 - 1)(a_2 - b_1)(a_3 - b_2) \neq 0, \tag{9.8} $$

and with abbreviations such as $b_{12} = b_1 + b_2$, $a_{23} = a_2 + a_3$. Then we can see the following lemma.

TABLE 5

Index	Z	c_Z	a^Z
[14]	$\theta+b_1-1$	b_1-1	b_1^-
[15]	Z_{15}	$-(b_1-1)(b_2-1)/(a_1-1)$	$a_1^-b_1^-b_2^-$
[16]	$-\theta-a_3$	$-a_3$	a_3^+
[24]	$-\theta-a_1$	$-a_1$	a_1^+
[25]	$\theta+b_2-1$	b_2-1	b_2^-
[26]	Z_{26}	a_1a_3/b_1	$a_1^+a_3^+b_1^+$
[34]	Z_{34}	a_1a_2/b_2	$a_1^+a_2^+b_2^+$
[35]	$-\theta-a_2$	$-a_2$	a_2^+
[36]	$-\partial$	$-a_1a_2a_3/b_1b_2$	$a_1^+a_2^+a_3^+b_1^+b_2^+$
[12]	$1-b_1-\theta-Z_{15}$	$(a_1-b_2)(1-b_1)/(a_1-1)$	$a_1^-b_1^-$
[21]	$a_1+\theta-Z_{26}$	$-a_1(a_3-b_1)/b_1$	$a_1^+b_1^+$
[13]	$-y(a_3+\theta)-Z_{15}$	$(b_1-1)(b_2-1)/(a_1-1)$	$a_1^-a_2^-b_1^-b_2^-$
[31]	$\partial-Z_{34}$	$a_1a_2(a_3-b_1)/b_1b_2$	$a_1^+a_2^+b_1^+b_2^+$
[23]	$1-b_2-\theta-yZ_{26}$	$1-b_2$	$a_2^-b_2^-$
[32]	$a_2+\theta-Z_{34}$	$-a_2(a_1-b_2)/b_2$	$a_2^+b_2^+$
[45]	$b_2-1+\theta+Z_{15}$	$(a_1-b_1)(b_2-1)/(a_1-1)$	$a_1^-b_2^-$
[54]	$-a_1-\theta+Z_{34}$	$a_1(a_2-b_2)/b_2$	$a_1^+b_2^+$
[46]	$-a_3-\theta+Z_{26}$	$a_3(a_1-b_1)/b_1$	$a_3^+b_1^+$
[64]	$b_1-1+\theta-yZ_{34}$	b_1-1	$a_3^-b_1^-$
[56]	$-\partial+Z_{26}$	$-a_1a_3(a_2-b_2)/b_1b_2$	$a_1^+a_3^+b_1^+b_2^+$
[65]	$y(a_2+\theta)+Z_{15}$	$-(b_1-1)(b_2-1)/(a_1-1)$	$a_1^-a_3^-b_1^-b_2^-$
[41]	$a_1+a_3-b_1+\theta-Z_{26}$	$-(a_3-b_1)(a_1-b_1)/b_1$	b_1^+
[42]	$1+a_1-b_1-b_2-\theta-Z_{15}$	$(a_1-b_1)(a_1-b_2)/(a_1-1)$	a_1^-
[43]	$1-b_2-\theta-y(a_3+\theta)-Z_{15}+yZ_{26}$	$(a_1-b_1)(1-b_2)/(a_1-1)$	$a_1^-a_2^-b_2^-$
[51]	$a_1+(1+y)\partial-Z_{26}-Z_{34}$	$a_1(a_2-b_2)(a_3-b_1)/b_1b_2$	$a_1^+b_1^+b_2^+$
[52]	$a_1+a_2-b_2+\theta-Z_{34}$	$-(a_1-b_2)(a_2-b_2)/b_2$	b_2^+
[53]	$a_2-b_2-\theta+yZ_{26}$	a_2-b_2	a_2^-
[61]	$a_3-b_1-\theta+yZ_{34}$	a_3-b_1	a_3^-
[62]	$1-b_1-\theta-y(a_2+\theta)-Z_{15}+yZ_{34}$	$(a_1-b_2)(1-b_1)/(a_1-1)$	$a_1^-a_3^-b_1^-$
[63]	$(1-a_3)y-y(a_2+\theta)-Z_{15}$	$(b_1-1)(b_2-1)/(a_1-1)$	$a_1^-a_2^-a_3^-b_1^-b_2^-$

LEMMA 9.2. *These operators yield the contiguity relations of second order for* $\underline{w} := \gamma_3 F_2$:

$$Z_{15}\underline{w}=(\alpha_5-1)\underline{w}(\alpha_1^+\alpha_5^-),\quad Z_{26}\underline{w}=(\alpha_6-1)\underline{w}(\alpha_2^+\alpha_6^-),\quad Z_{34}\underline{w}=(\alpha_4-1)\underline{w}(\alpha_3^+\alpha_4^-).$$

Proof will be given in § 10. Then, similar to Proposition 7.4, we have the following.

PROPOSITION 9.3. *Goursat's* HGF ${}_3F_2(a_1, a_2, a_3; b_1, b_2; y)$ *satisfies the contiguity relations listed in Table* 5.

Each line reads as $Z({}_3F_2(a; y)) = c_Z \cdot ({}_3F_2)(a^Z; y)$. We use parameters (a) in this table. Recall $\theta = y(\partial/\partial y)$ and $\partial = (\partial/\partial y)$. This set of relations is complete in the sense explained in the last remark of § 7.

10. Proof of Lemmas 7.3 and 9.2. We first prove Lemma 7.3. Restricting the system in Proposition 4.4 to the stratum $\{x^3=0\}$ we see that Appell's HGF $F_3(\alpha, \alpha', \beta, \beta', \gamma; x, y)$ satisfies the system

$$\begin{aligned}&(\delta_x+\delta_y+\gamma-1)\delta_x u - x(\delta_x+\alpha)(\delta_x+\beta)u=0,\\ &(\delta_x+\delta_y+\gamma-1)\delta_y u - y(\delta_x+\alpha')(\delta_x+\beta')u=0,\end{aligned}\tag{10.1}$$

where u is an unknown function and $\delta_x = x(\partial/\partial x)$, $\delta_y = y(\partial/\partial y)$. This system can be written in a Pfaffian form: put

$$e_0=u,\quad e_1=\delta_x u,\quad e_2=\delta_y u,\quad e_3=\delta_x\delta_y u$$

and define a column vector e by

$$e = {}^t(e_0, e_1, e_2, e_3).$$

Then the above system is equivalent to the equation for e:

$$de = \omega e, \tag{10.2}$$

where

$$\omega = \begin{pmatrix} 0 & \frac{dx}{x} & \frac{dy}{y} & 0 \\ -\frac{\alpha\beta}{x-1}dx & \left\{\frac{1-\gamma}{x} - \frac{\delta}{x-1}\right\}dx & 0 & \left\{\frac{1}{x-1} - \frac{1}{x}\right\}dx + \frac{1}{y}dy \\ -\frac{\alpha'\beta'}{y-1}dy & 0 & \left\{\frac{1-\gamma}{y} - \frac{\delta'}{y-1}\right\}dy & \left\{\frac{1}{y-1} - \frac{1}{y}\right\}dy + \frac{1}{x}dx \\ 0 & \alpha'\beta'\left\{\frac{dx}{x} - \frac{d\Delta}{\Delta}\right\} & \alpha\beta\left\{\frac{dy}{y} - \frac{d\Delta}{\Delta}\right\} & \delta'\frac{dx}{x} + \delta\frac{dy}{y} + \varepsilon\frac{d\Delta}{\Delta} \end{pmatrix},$$

$$\delta = \alpha + \beta + 1 - \gamma, \quad \delta' = \alpha' + \beta' + 1 - \gamma, \quad \varepsilon = \gamma - \alpha - \beta - \alpha' - \beta' - 1.$$

To denote the dependence of the vector e on the parameters we use the notation $e(\alpha, \alpha', \beta, \beta', \gamma)$; in the sequel we use abbreviations such as (α^+) for $(\alpha+1, \alpha', \beta, \beta', \gamma)$, $(\alpha^+\beta'^+\gamma^+)$ for $(\alpha+1, \alpha', \beta, \beta'+1, \gamma+1)$, and so on. Hence $e(\alpha^+)$ means $e(\alpha+1, \alpha', \beta, \beta', \gamma)$. The notation e always stands for $e(\alpha, \alpha', \beta, \beta', \gamma)$. We need the following lemma.

LEMMA 10.1. *Each solution u of* (10.1) *satisfies the equations*

$$x\Delta u_{xxy} = (\varepsilon y - (\alpha+\beta+1)\Delta)u_{xy} + \alpha\beta(1-y)u_y - \alpha'\beta' u_x,$$

$$y\Delta u_{xyy} = (\varepsilon x - (\alpha'+\beta'+1)\Delta)u_{xy} + \alpha'\beta'(1-y)u_y - \alpha\beta u_x.$$

The proof is straightforward and omitted.

The operator $L(\alpha^+\beta'^+\gamma^+)$ will be found by the following process. The relations [15], [16], and [25] in Table 3.1 provide us with relations expressing $e(\beta^+)$, $e(\beta'^+)$, and $e(\alpha^+\beta^+\gamma^+)$ by $e = e(\alpha, \alpha', \beta, \beta', \gamma)$, respectively. From the first relation we get a relation for (β^-), i.e., a relation expressing $e(\beta^-)$ by e. Then the composition

$$e(\alpha, \beta, \beta', \gamma) \xrightarrow{\beta'^+} e(\alpha, \beta, \beta'+1, \gamma)$$

$$\xrightarrow{\beta^-} e(\alpha, \beta-1, \beta'+1, \gamma)$$

$$\xrightarrow{\alpha^+\beta^+\gamma^+} e(\alpha+1, \beta, \beta'+1, \gamma+1)$$

gives a required operator.

Now let us follow this process. Put

$$\begin{aligned} e_0 &= c(\alpha, \alpha', \gamma)F_3(\alpha, \alpha', \beta, \beta', \gamma; x, y), \\ c(\alpha, \alpha', \gamma) &= \Gamma(1-\gamma)\Gamma(\alpha)\Gamma(\alpha')/\Gamma(\alpha+\alpha'-\gamma+1). \end{aligned} \tag{10.3}$$

The relation [15] in Table 3.1 reads

$$e_0(\beta^+) = e_0(\beta) + \frac{1}{\beta}e_1(\beta).$$

Hence (10.2) shows that we have

$$e(\beta^+) = Ae(\beta),$$

where

$$A = \begin{pmatrix} 1 & \frac{1}{\beta} & 0 & 0 \\ -\frac{\alpha x}{x-1} & \frac{1}{\beta}\left(1+\beta-\gamma-\frac{\delta x}{x-1}\right) & 0 & \frac{1}{\beta(x-1)} \\ 0 & 0 & 1 & \frac{1}{\beta} \\ 0 & -\frac{\alpha'\beta' y}{\beta\Delta} & -\frac{\alpha x(y-1)}{\Delta} & 1+\frac{\delta'}{\beta}+\frac{\varepsilon x(y-1)}{\beta\Delta} \end{pmatrix}.$$

The determinant of this matrix A is

$$\frac{(\alpha'+\beta+1-\gamma)(\beta+\beta'+1-\gamma)y}{\beta^2(x-1)\Delta}.$$

Hence, taking the inverse of this matrix and decreasing the value of β by 1, we get

$$e_0(\beta^-) = \left\{1+\frac{\alpha(\beta+\delta'-1)}{\lambda}x\right\}e_0+\frac{\beta+\delta'-1}{\lambda}(x-1)e_1+\frac{\alpha x(y-1)}{\lambda y}e_2+\frac{\Delta}{\lambda y}e_3,$$

provided that

$$\lambda := (\gamma-\alpha-\beta')(\gamma-\alpha'-\beta) \neq 0.$$

Namely, if we define a differential operator U_{51} by

$$U_{51} = \lambda+\alpha(\beta+\delta'-1)x+(\beta+\delta'-1)x(x-1)\partial_x+\alpha x(y-1)\partial_y+\Delta x\partial_x\partial_y, \tag{10.4}$$

then

$$U_{51}e_0 = \lambda e_0(\beta^-). \tag{10.5}$$

We next define an operator U_{21} by

$$U_{21} = -\frac{1}{\beta}U_{51}(\alpha^+\beta^+\gamma^+)\cdot Y_{25}, \tag{10.6}$$

where $Y_{25} = \partial_x$ in Table 3.1 and $U_{51}(\alpha^+\beta^+\gamma^+)$ denotes the operator U_{51} with parameters α, β, γ increased by 1. By using Lemma 10.1, we see

$$U_{21} = \alpha(\alpha'+\beta+\beta'-\gamma)-(\alpha'+\beta+\beta'-\gamma)(1-x)\partial_x-\alpha(1-y)\partial_y-\Delta\partial_x\partial_y,$$

and we have

$$U_{21}e_0 = \lambda e_0(\alpha^+\gamma^+). \tag{10.7}$$

We finally compute the composition U_{26} of U_{21} and Y_{16} in Table 3.1. A calculation shows

$$U_{26} = \frac{1}{\gamma-\beta-\beta'-1}U_{21}(\beta'^+)\cdot Y_{16},$$

and we have

$$U_{26}e_0 = \beta'(\alpha'+\beta-\gamma)e_0(\alpha^+\beta'^+\gamma^+),$$

which is the first assertion of Lemma 7.3. Note that we can forget the condition $\lambda \neq 0$ in the final step. The second assertion in Lemma 7.3 is obtained from the first one by the symmetry $\alpha \leftrightarrow \alpha'$, $\beta \leftrightarrow \beta'$, and $x \leftrightarrow y$. □

We next prove Lemma 9.2 by a similar method. The function $z = {}_3F_2(a; y)$ satisfies

$$\theta(\theta+b_1-1)(\theta+b_2-1)z - y(\theta+a_1)(\theta+a_2)(\theta+a_3)z = 0. \tag{10.8}$$

Put $e_0 = z$, $e_1 = \theta z$, and $e_2 = \theta^2 z$. Then $e = {}^t(e_0, e_1, e_2)$ is a solution of a Pfaffian equation

$$de = \omega e,$$

where

$$\omega = \begin{pmatrix} 0 & \frac{1}{y} & 0 \\ 0 & 0 & \frac{1}{y} \\ \frac{a_1a_2a_3y}{1-y} & \frac{Ay-B}{1-y} & \frac{a_{123}y-b_{12}+2}{1-y} \end{pmatrix} dy,$$

$$A = a_1a_2+a_2a_3+a_3a_1, \qquad B = (b_1-1)(b_2-1).$$

Recall the convention $a_{123} = a_1+a_2+a_3$ and $b_{12} = b_1+b_2$.

We will derive three operators (9.5)-(9.7) by the use of (9.4). Let us denote differential operators in (9.4) by Y_{14}, Y_{16}, and so on. If we find the inverse U_{42} or Y_{24}, then the composition of Y_{14}, U_{42}, and Y_{25}, in this order, yields Z_{15}. Similarly, denoting by U_{53} and U_{61} the inverses of Y_{35} and Y_{16}, we have Z_{26} and Z_{34} by certain composition of operators.

We follow the first case: By [24] of (9.4), we see

$$e(a_1^+) = \begin{pmatrix} 1 & \frac{1}{a_1} & 0 \\ 0 & 1 & \frac{1}{a_1} \\ \frac{a_2a_3y}{1-y} & \frac{Ay-B}{a_1(1-y)} & 1+\frac{a_{123}y-b_{12}+2}{a_1(1-y)} \end{pmatrix} e.$$

Then, inverting this identity, we see

$$e_0 = \frac{1}{(a_1-b_1+1)(a_1-b_2+1)}[\{(a_1-b_1+1)(a_1-b_2+1)-a_2a_3y\}e_0(a_1^+) \\ -\{a_{23}y+a_1-b_{12}+2\}e_1(a_1^+)+(1-y)e_2(a_1^+)].$$

Hence, by putting

$$U_{42} = \frac{1}{(a_1-b_1)(a_1-b_2)}[\{(a_1-b_1)(a_1-b_2)-a_2a_3y\}-\{a_{23}y+a_1-b_{12}+1\}\theta+(1-y)\theta^2],$$

we have

$$U_{42}e_0 = e_0(a_1^-).$$

Now define $U_{15} = Y_{25}(a_1^-b_1^-)\cdot U_{42}(b_1^-)\cdot Y_{14}$. Then a simple calculation shows

$$U_{15} = (1-y)\theta^2+(-a_{23}y+b_{12}-2)\theta-a_2a_3y+(b_1-1)(b_2-1)$$

and

$$U_{15}e_0 = (b_1-1)(b_2-1)e_0(a_1^-b_1^-b_2^-).$$

By multiplying a suitable constant, we get the operator Z_{15}, which satisfies the first relation in Lemma 9.2 in view of the relation of parameters (9.2).

For the second case look at the symmetry $a_1 \leftrightarrow a_2$ of (10.7). Then for the operator

$$U_{53} = \frac{1}{(a_2-b_1)(a_2-b_2)}[\{(a_2-b_1)(a_2-b_2)-a_1a_3y\}-(a_{13}y+a_2-b_{12}+1)\theta+(1-y)\theta^2],$$

the composition $U_{36} := Y_{36}(a_2^-b_2^-)\cdot U_{53}(b_2^-)\cdot Y_{25}$ gives

$$U_{36} = \frac{a_2-1}{(a_2-b_1)y}\{(y-1)\theta^2+(a_{13}y-b_2+1)\theta+a_1a_3y\}$$

and

$$U_{36}e_0 = -\frac{a_1a_3(a_2-1)}{b_1}e_0(a_1^+a_3^+b_1^+).$$

This is equivalent to the second identity of Lemma 9.2.

The third case can be treated similarly and we complete the proof. □

Acknowledgments. We express hearty thanks to Masaaki Yoshida for having several discussions on the subjects in this paper, and to the referee for showing the author papers [H] and [KMM1]-[KMM3] on canonical systems.

REFERENCES

[A1] K. AOMOTO, *On the structure of integrals of power products of linear functions*, Sci. Papers, College of Arts and Sciences, University of Tokyo, 27 (1977), pp. 49-61.

[A2] ———, *Les équations aux différences linéaires et intégrales des fonctions multiformes*, J. Fac. Sci. Univ. Tokyo Sec. IA Math., 22 (1975), pp. 271-297; *Une correction et un complément à l'article "Les équations aux différences linéaires et les intégrales des fonctions multiformes,"* ibid., 26 (1979), pp. 519-523.

[A3] ———, *Configurations and invariant Gauss-Manin connections of integrals* I, Tokyo J. Math., 5 (1982), pp. 249-287; II, ibid., 6 (1983), pp. 1-24.

[A4] ———, *Special values of the hypergeometric function ${}_3F_2$ and connection formulae among asymptotic expansions*, J. Indian Math. Soc. (N.S.), 51 (1987), pp. 161-221.

[AK] P. APPELL AND J. KAMPÉ DE FERIET, *Fonctions hypergéométriques et hypersphériques—polynomes d'Hermite*, Gauthier-Villars, Paris, 1926.

[E] A. ERDELYI, *Higher Transcendental Functions* I, McGraw-Hill, New York, 1953.

[GA] C. F. GAUSS, *Disquisitiones generales circa seriem infinita*

$$1+\frac{\alpha\beta}{1\cdot\gamma}x+\frac{\alpha(\alpha+1)\beta(\beta+1)}{1\cdot2\cdot\gamma(\gamma+1)}xx+\frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1\cdot2\cdot3\cdot\gamma(\gamma+1)(\gamma+2)}x^3+\text{etc.},$$

in Werke von C. F. Gauss III, Georg Olms Verlag, 1973, Hildesheim, New York, pp. 123-162, 207-229.

[GE] I. M. GEL'FAND, *General theory of hypergeometric functions*, Soviet Math. Dokl., 33 (1986), pp. 573-577.

[GG] I. M. GEL'FAND AND M. I. GEL'FAND, *Generalized hypergeometric equations*, Soviet Math. Dokl., 33 (1986), pp. 643-646.

[GGR1] I. M. GEL'FAND AND M. I. GRAEV, *Hypergeometric functions associated with the Grassmannian $G_{3,6}$*, Soviet Math. Dokl., 35 (1987), pp. 298-303.

[GGR2] ———, *Generalized hypergeometric functions on the Grassmannian $G_{3,6}$*, Preprint 123, Keldysh Institue of Applied Mathematics, 1987. (In Russian.)

[GGR3] ———, *Strata in $G_{3,6}$ and the associated hypergeometric functions*, Preprint 127, Keldysh Institute of Applied Mathematics, 1987. (In Russian.)

[GZ] I. M. GEL'FAND AND A. V. ZELEVINSKII, *Algebraic and combinatorial aspects of the general theory of hypergeometric functions*, Funct. Anal. Appl., 20 (1987), pp. 183-197.

[GZK] I. M. GEL'FAND, A. V. ZELEVINSKII, AND M. M. KAPRANOV, *Hypergeometric functions and toric varieties*, Funkt. Anal. i Prilozhen., 23 (1989), pp. 12-26.

[H] J. HRABOWSKI, *Multiple hypergeometric functions and simple Lie groups SL and S_p*, SIAM J. Math. Anal., 16 (1985), pp. 876-886.

[IKSY] K. IWASAKI, H. KIMURA, S. SHIMOMURA, AND M. YOSHIDA, *From Gauss to Painlevé—a Modern Theory of Special Functions*, Vieweg-Verlag, Wiesbaden, 1991.

[KMM1] E. G. KALNIS, H. L. MANOCHA, AND W. MILLER, *The Lie theory of two-variable hypergeometric functions*, Stud. Appl. Math., 62 (1980), pp. 143-173.

[KMM2] ———, *Transformation and reduction formulas for two-variable hypergeometric function on the sphere S_2*, Stud. Appl. Math., 63 (1980), pp. 155-167.

[KMM3] ———, *Harmonic analysis and expansion formulas for two-variable hypergeometric functions*, Stud. Appl. Math., 66 (1982), pp. 69-89.

[K] T. KIMURA, *Hypergeometric functions of two variables*, Lecture notes, University of Tokyo, Tokyo, 1973.

[KN] M. KITA AND M. NOUMI, *On the structure of cohomology group attached to the integral of certain many-valued analytic functions*, Japan. J. Math. (N.S.), 9 (1983), pp. 113-157.

[M1] W. MILLER, *Lie theory and the Lauricella functions F_D*, J. Math. Phys., 13 (1972), pp. 1393-1399.

[M2] ———, *Lie theory and generalizations of the hypergeometric functions*, SIAM J. Appl. Math., 25 (1973), pp. 226-235.

[M3] ———, *Lie theory and generalized hypergeometric functions*, SIAM J. Math. Anal., 3 (1972), pp. 31-44.

[M4] ———, *Lie algebras and generalizations of hypergeometric functions*, Proc. Sympos. Pure Math., 26 (1973), pp. 355-356.

[MSY1] K. MATSUMOTO, T. SASAKI, AND M. YOSHIDA, *The period map of a 4-parameter family of $K3$ surfaces and the Aomoto-Gel'fand hypergeometric function of type* (3, 6), Proc. Japan Acad. Ser. A Math. Sci., 64 (1988), pp. 307-310.

[MSY2] ———, *The monodromy of the period map of a 4-parameter family of $K3$ surfaces and the Aomoto-Gel'fand hypergeometric function of type* (3, 6), preprint, 1989.

[O] K. OKAMOTO, *Sur les échelles associées aux fonctions spéciales et l'équation de Toda*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 34 (1987), pp. 709-740.

[R] E. D. RAINVILLE, *The contiguous function relations for ${}_pF_q$ with applications to Bateman's $J_n^{u,v}$ and Rice's $H_n(\zeta, p, v)$*, Bull. Amer. Math. Soc., 51 (1945), pp. 714-723.

[SA] M. SATO, *Singular orbits of a prehomogeneous vector space and hypergeometric functions*, Notes (taken by K. Aomoto) of lectures at University of Tokyo, Tokyo, 1971, pp. 1-17. (In Japanese.)

[SE] Y. SE-ASHI, *On differential invariants of integrable finite type linear differential equations*, Hokkaido Math. J., 17 (1988), pp. 151-195.

[SY1] T. SASAKI AND M. YOSHIDA, *Linear differential equations in two variables* I, Math. Ann., 281 (1988), pp. 69-93.

[SY2] ———, *Linear differential equations modeled after hyperquadrics*, Tohoku Math. J., 41 (1989), pp. 321-348.

[SY3] ———, *Tensor products of linear differential equations* II—*new formulae for the hypergeometric functions*, Funkcial. Ekvac., 33 (1990), pp. 527-549.

[T] N. TAKAYAMA, *Gröbner basis and the problem of contiguity relations*, Japan J. Appl. Math., 6 (1989), pp. 147-160.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 847–855, May 1991

018

MEAN CONVERGENCE OF EXPANSIONS IN FREUD-TYPE ORTHOGONAL POLYNOMIALS*

H. N. MHASKAR† AND Y. XU‡

Abstract. Let $\{p_n\}$ be the system of polynomials orthonormal with respect to a weight function of the form $\exp(-2Q(x))$. Under some technical conditions on Q we prove certain norm inequalities for the partial sums of the orthogonal expansion of a function in a suitably weighted L^p- norm. These results are valid, in particular, when $Q(x) = x^m$, where m is an even positive integer.

Key words. orthogonal polynomials, mean convergence, Freud polynomials, Fourier orthogonal expansions

AMS(MOS) subject classifications. 41A25, 42C15, 33A65

1. Introduction.

A classical result of Riesz (cf. [21]) states that if f is a 2π-periodic function, $1 < p < \infty$, and $S_n(f, \bullet)$ is the nth partial sum of its trigonometric Fourier series, then

$$\int_{-\pi}^{\pi} |S_n(f,t)|^p dt \le c \int_{-\pi}^{\pi} |f(t)|^p dt \tag{1.1}$$

where c is a constant depending only on p. In the aperiodic case, the situation is quite different. Thus, for example, if $\int_{-1}^{1} |f(t)|^p dt < \infty$ and we consider the nth partial sum of its Legendre expansion instead of $S_n(f, \bullet)$, then an inequality of the form (1.1) holds only when $4/3 < p < 4$ [16]. Furthermore, if $\int_{-\infty}^{\infty} |f(x)|^p \exp(-x^2)dx < \infty$, and $s_n(f, \bullet)$ denotes the nth partial sum of the expansion of f in Hermite polynomials, then

$$\int_{-\infty}^{\infty} |s_n(f,x)|^p \exp(-x^2)dx \le c_2 \int_{-\infty}^{\infty} |f(x)|^p \exp(-x^2)dx \tag{1.2}$$

(with c_2 depending only on p) can hold only when $p = 2$ [16], [17]. Nevertheless, when $4/3 < p < 4$, then

$$\int_{-\infty}^{\infty} |s_n(f,x)\exp(-x^2/2)|^p dx \le c_3 \int_{-\infty}^{\infty} |f(x)\exp(-x^2/2)|^p dx \tag{1.3}$$

where c_3 depends only on p [1]. In 1970, Muckenhoupt [14] showed that an inequality of the form (1.1) can be proved for the Hermite expansions for all p in the range $1 < p < \infty$ if we take different weights on the two sides of the inequality. More precisely, he proved the following theorem.

THEOREM 1.1 [14]. *Let* $1 < p < \infty$,

$$U(x) := \exp(-x^2/2)(1 + |x|)^{\tilde{b}}$$

*Received by the editors October 22, 1989; accepted for publication (in revised form) May 3, 1990.

†Department of Mathematics, California State University, Los Angeles, California 90032.

‡Department of Mathematics, University of Texas, Austin, Texas 78712.

and

$$V(x) := \exp(-x^2/2)(1+|x|)^{\tilde{B}}(1+\log^+|x|)^{\beta}$$

where $\beta = 1$ *if* $\tilde{b} = \tilde{B}$ *and* $p = 4/3$ *or* 4, *and* $\beta = 0$ *otherwise. Assume that*

$$\begin{aligned} &\tilde{b} < 1 - 1/p, && 1 < p \le 4, \\ &\le \tfrac{2}{3} + 1/(3p), && 4 < p < \infty, \end{aligned} \tag{1.4a}$$

$$\begin{aligned} &\tilde{B} \ge -1 + 1/(3p), && 1 < p < 4/3, \\ &> -1/p, && 4/3 \le p < \infty, \end{aligned} \tag{1.4b}$$

$$\begin{aligned} &\tilde{b} \le \tilde{B} + 1 - 4/(3p), && 1 < p < 4/3, \\ &\le \tilde{B}, && 4/3 \le p \le 4, \\ &\le \tilde{B} - \tfrac{1}{3} + 4/(3p), && 4 < p < \infty, \end{aligned} \tag{1.4c}$$

and if equality occurs in (1.4c) *then equality does not occur in* (1.4a) *or* (1.4b). *Then there exists a constant* c_4 *independent of* f *and* n *such that*

$$\int_{-\infty}^{\infty} |s_n(f,x)U(x)|^p dx \le c_4 \int_{-\infty}^{\infty} |f(x)V(x)|^p dx \tag{1.5}$$

where $s_n(f,\bullet)$ *denotes the* n*th partial sum of the orthogonal expansion of* f *in terms of the Hermite polynomials.*

The proofs of these results concerning Hermite expansions utilize a very detailed knowledge about the asymptotic behavior of the Hermite polynomials.

In this paper, we obtain an analogue of Theorem 1.1 for expansions in Freud polynomials, i.e., polynomials orthogonal on the whole real line with respect to a weight function of the form $\exp(-2Q(x))$ where Q is a suitably chosen function. While several results concerning the asymptotic behavior of such polynomials and related quantities have been proved recently [6]–[9], [15], [18], [20], our knowledge concerning Freud polynomials with a general weight function is still limited to a few relatively imprecise estimations. Our interest, then, is not only in proving an analogue of Theorem 1.1, but also in exploring the extent to which various polynomial inequalities can be used in the study of orthogonal polynomial expansions.

We discuss our main results in § 2 and prove them in § 3.

2. Main results. Let Q be an even and convex function on $\mathbb{R}$, differentiable on $(0,\infty)$, and let $xQ'(x) \to \infty$ as $|x| \to \infty$. We consider a weight function of the form $w_Q^2(x) := \exp(-2Q(x))$ and the sequence of polynomials $\{p_n\}$ orthonormal on $\mathbb{R}$ with respect to w_Q^2. Thus, denoting the class of all polynomials of degree at most n by Π_n, we have

$$p_n(x) := p_n(w_Q^2,x) = \gamma_n x^n + \cdots \in \Pi_n, \quad \gamma_n := \gamma_n(w_Q^2) > 0, \quad n = 0,1,2,\cdots, \tag{2.1a}$$

$$\int_{-\infty}^{\infty} p_n p_m w_Q^2 dx = \delta_{nm}, \qquad n,m = 0,1,\cdots. \tag{2.1b}$$

If f is Lebesgue measurable function on $\mathbb{R}$, we define, when possible,

$$a_k(w_Q,f) := \int_{-\infty}^{\infty} f(t)p_k(t)w_Q^2(t)dt, \qquad k = 0,1,\cdots, \tag{2.2a}$$

$$s_n(w_Q, f, x) := \sum_{k=0}^{n-1} a_k(w_Q, f)p_k(x), \qquad n = 1, 2, \cdots. \tag{2.2b}$$

As usual, we also write

$$\|f\|_p := \begin{cases} \left(\int_{-\infty}^{\infty} |f(x)|^p dx\right)^{1/p}, & 1 \le p < \infty, \\ \operatorname{ess\,sup}_{x\in\mathbb{R}} |f(x)|, & p = \infty \end{cases} \tag{2.3}$$

and the space $L^p(\mathbb{R})$ then denotes the space of all Lebesgue measurable functions f for which $\|f\|_p < \infty$, two functions being considered equal when they are equal almost everywhere.

Our main theorem can now be formulated as follows.

THEOREM 2.1. *Let Q be an even, convex function on $\mathbb{R}$, differentiable on $(0, \infty)$ and $xQ'(x) \uparrow \infty$ as $|x| \to \infty$ and with suitably chosen constants c_5, c_6,*

$$1 < c_5 \le Q'(2x)/Q'(x) \le c_6 < \infty, \qquad x > 0.$$

For every integer $n > 0$, we let q_n be the least positive number satisfying the equation

$$q_n Q'(q_n) = n. \tag{2.4}$$

Suppose that the orthogonal polynomials $\{p_n\}$ satisfy each of the following inequalities where $K(\ge -1/2), A^*, c_7, c_8, c_9$ are suitably chosen constants depending on Q alone:

$$|p_n(x)w_Q(x)| \le c_7 q_n^K, \qquad x \in \mathbb{R}, \quad n = 1, 2, \cdots, \tag{2.5a}$$

$$|p_n(x)w_Q(x)| \le c_8 q_n^{-1/2}, \qquad |x| \le A^* q_n, \quad n = 1, 2, \cdots, \tag{2.5b}$$

$$|p_{n+1}(x) - p_{n-1}(x)|w_Q(x) \le c_9 q_n^{-1/2}, \qquad x \in \mathbb{R}, \quad n = 1, 2, \cdots. \tag{2.5c}$$

Let $1 < p < \infty$ and b, B be constants satisfying

$$b < \min\{1 - 1/p, -K + 1/2 - 1/p\}, \tag{2.6a}$$

$$B > \max\{-1/p, K + 1/2 - 1/p\}, \tag{2.6b}$$

$$B - b \ge 2K + 1. \tag{2.6c}$$

Then, for any Lebesgue measurable function f such that $(1 + |x|)^B w_Q f \in L^p(\mathbb{R})$, we have

$$\|(1 + |x|)^b w_Q(x) s_n(w_Q, f, x)\|_p \le c_{10} \|(1 + |x|)^B w_Q(x) f(x)\|^p \tag{2.7}$$

where c_{10} is a constant depending only on Q, p, b, B.

An estimate of the form (2.5a) can be proved easily using Nikolskii-type inequalities under very mild conditions on Q [11]. However, the constant K obtained in this way is usually not sharp. A sharp estimate of this form is known, at the time of this writing, only when $Q(x) = x^m$ where m is an even integer. In this case, $K = (m-3)/6$ [3]. In particular, in contrast to the Hermite polynomials ($m = 2$) Freud polynomials

for such weights are not uniformly bounded if $m \geq 4$. This fact is partly responsible for our inability to generalize Muckenhoupt's result completely.

Condition (2.5b) has been a subject of great interest in recent years [2], [3], [5], [7], [10]. In [7], it has been proved for the case when $Q(x) = |x|^\alpha$ when $\alpha > 3$. Lopez and Rahmanov have claimed to have proved it under very general conditions including the case of these weights when $\alpha > 0$.

Condition (2.5c) is currently known only in the case where $Q(x) = x^m$, m even positive integer [5]. The known proof requires an asymptotic expansion of quantities γ_{n-1}/γ_n. Thus, our theorem is currently valid only for the case when $Q(x) = x^m$, m even positive integer. Using the results in [7] and [8], it seems possible to obtain stronger results in this case. We hope to return to this in the near future, concentrating in this note on the general principles involved. It is conjectured that the inequalities (2.5) are true more generally, including the case where $Q(x) = |x|^\alpha$, $\alpha > 0$.

The conditions on Q imply that

$$\int_1^\infty \frac{Q(x)}{1+x^2}dx = \infty.$$

It is known [19] that expressions of the form $(1 + |x|)^B w_Q(x)P(x)$, where P is a polynomial, are dense in $L^p(\mathbb{R})$. Therefore, (2.7) easily implies that

$$||(1 + |x|)^b w_Q(x)\Big[s_n(w_Q, f, x) - f(x)\Big]||_p \to 0, \qquad n \to \infty.$$

An inequality similar to (2.7) with an L^1-norm on the left-hand side and an $L\log^+ L$ norm on the right-hand side can also be obtained using our techniques, but we do not intend to pursue this further since the proof does not add any deeper insights to the study of Freud polynomials.

3. Proof. In the sequel, we adopt the following conventions concerning constants. The letters $c, c_1, \cdots$ will denote constants depending only on Q and other fixed parameters involved, but their values may be different at different occurrences, even within the same formula. The notation adopted in § 2 will be continued except for the constants $c, c_1, \cdots$, etc. We also adopt the notation

$$\mathcal{P}_n(x) := p_n(x)w_Q(x), \qquad x \in \mathbb{R}, \quad n = 0, 1, \cdots. \tag{3.1}$$

To begin with, we recall certain facts concerning Freud polynomials and the Hilbert transform.

LEMMA 3.1. (a) [4] *We have*

$$xp_{n-1}(x) = \rho_n p_n(x) + \rho_{n-1}p_{n-2}(x), \quad \rho_n := \gamma_{n-1}/\gamma_n, \quad n = 1, 2, \cdots, \tag{3.2}$$

$$K_n(x, y) := \sum_{k=0}^{n-1} p_k(x)p_k(y) = \rho_n \frac{p_n(x)p_{n-1}(y) - p_n(y)p_{n-1}(x)}{x - y}. \tag{3.3}$$

(b) [10] *For the Freud polynomials, we have*

$$\tfrac{1}{4}q_n \leq \rho_n \leq 4q_n \tag{3.4}$$

where q_n is defined in (2.4).

(c) [12] *Let $0 < p, r < \infty$. Then there exists a constant $L := L(Q, p, r)$ with the following property. For every integer $n = 1, 2, \cdots$ and $P \in \Pi_n$,*

$$|w_Q(x)P(x)|^p \le c\exp(-c_1 n)\int_{|t|\le Lq_n} |w_Q(t)P(t)|^p dt, \qquad |x| \ge Lq_n, \tag{3.5a}$$

$$\left\{\int_{|x|\ge Lq_n} |w_Q(x)P(x)|^r dx\right\}^{p/r} \le c\exp(-c_1 n)\int_{|t|\le Lq_n} |w_Q(t)P(t)|^p dt. \tag{3.5b}$$

(d) [14] *If $1 < p < \infty$, $r < 1 - 1/p$, $R > -1/p$ and $r \le R$, and ϕ is Lebesgue measurable, then*

$$\int_{-\infty}^{\infty}\left|\int_{-\infty}^{\infty}\frac{\phi(y)}{x-y}dy\right|^p (1+|x|)^{rp}dx \le c\int_{-\infty}^{\infty}|\phi(x)(1+|x|)^R|^p dx, \tag{3.6}$$

where the singular integral is taken in the principal value sense.

We would like to warn the reader that the notation in Lemma 3.1(d) is different from the one used in Lemma 8 of [14]. As in [14], the starting point of our proof is the following lemma.

LEMMA 3.2. *We have, for $n = 1, 2, \cdots$,*

$$K_n(x,y)w_Q(x)w_Q(y) = h_1(x,y) + h_2(x,y) + h_3(x,y) \tag{3.7}$$

where

$$h_1(x,y) := \frac{\rho_n}{\rho_n + \rho_{n-1}}\mathcal{P}_{n-1}(x)\mathcal{P}_{n-1}(y), \tag{3.8a}$$

$$h_2(x,y) := \frac{\rho_{n-1}\rho_n}{\rho_n + \rho_{n-1}}\frac{\mathcal{P}_{n-1}(y)\Big[\mathcal{P}_n(x) - \mathcal{P}_{n-2}(x)\Big]}{x-y}, \tag{3.8b}$$

$$h_3(x,y) := h_2(y,x). \tag{3.8c}$$

Proof. Writing

$$D_n(x,y) := K_n(x,y)w_Q(x)w_Q(y), \tag{3.9}$$

we see from (3.3) that

$$\begin{aligned} 2D_n(x,y) &= D_n(x,y) + D_{n-1}(x,y) + \mathcal{P}_{n-1}(x)\mathcal{P}_{n-1}(y) \\ &= \rho_{n-1}\left[\frac{D_n(x,y)}{\rho_n} + \frac{D_{n-1}(x,y)}{\rho_{n-1}}\right] \\ &\quad + \frac{\rho_n - \rho_{n-1}}{\rho_n}D_n(x,y) + \mathcal{P}_{n-1}(x)\mathcal{P}_{n-1}(y). \end{aligned} \tag{3.10}$$

Transferring the middle term to the left-hand side and simplifying, we get

$$D_n(x,y) = \frac{\rho_n\rho_{n-1}}{\rho_n+\rho_{n-1}}\left[\frac{D_n(x,y)}{\rho_n}+\frac{D_{n-1}(x,y)}{\rho_{n-1}}\right]+\frac{\rho_n}{\rho_n+\rho_{n-1}}\mathcal{P}_{n-1}(x)\mathcal{P}_{n-1}(y). \tag{3.11}$$

Using (3.3) and the notation in formulas (3.8), we get (3.7). □

We write

$$u_r(x) := (1+|x|)^r, \qquad g(x) := w_Q(x)f(x). \tag{3.12}$$

Then (cf. [14]) with D_n as in (3.9),

$$u_b(x)w_Q(x)s_n(w_Q,f,x) = u_b(x)\int_{-\infty}^{\infty} g(y)D_n(x,y)dy.$$

Therefore, in order to prove Theorem 2.1, we need to show that

$$\int_{-\infty}^{\infty}\left|\int_{-\infty}^{\infty} g(y)D_n(x,y)u_b(x)dy\right|^p dx \le c\int_{-\infty}^{\infty}|g(y)u_B(y)|^p dy. \tag{3.13}$$

In view of Lemma 3.2, we see that it is enough to show that

$$\int_{-\infty}^{\infty}\left|\int_{-\infty}^{\infty} g(y)h_k(x,y)u_b(x)dy\right|^p dx \le c\int_{-\infty}^{\infty}|g(y)u_B(y)|^p dy, \qquad k=1,2,3. \tag{3.14}$$

The following lemma gives certain estimates on $\{\mathcal{P}_n\}$ which will be needed in the sequel. We note that our assumptions on Q imply that $\{q_{2n}/q_n\}$ is bounded from above and below by positive constants, and also that $K \ge -1/2$.

LEMMA 3.3. *Let $\sigma \in \mathbb{R}$, $0 < r \le \infty$. There exists a constant A depending only on σ, Q such that*

$$|u_\sigma(x)\mathcal{P}_n(x)| \le c\exp(-c_1 n), \qquad |x| \ge Aq_n. \tag{3.15}$$

Moreover,

$$||u_\sigma\mathcal{P}_n||_r \le \begin{cases} q_n^{-1/2}(\log n)^{1/r} & \text{if } \sigma = -1/r, \quad K=-1/2, \\ q_n^T & \text{otherwise} \end{cases} \tag{3.16a}$$

where

$$T := \max\{K+\sigma+1/r, -1/2\}. \tag{3.16b}$$

Proof. Let l be the least positive even integer greater than σ. Then, applying Lemma 3.1(c) to the polynomial $(1+x^l)p_n(x) \in \Pi_{n+l}$ with $p=2$, we get, for $|x| \ge Lq_{n+l}$,

$$\begin{aligned} |u_\sigma(x)\mathcal{P}_n(x)|^2 &\le |(1+x^l)p_n(x)w_Q(x)|^2 \\ &\le c\exp(-c_1 n)\int_{|t|\le Lq_{n+l}}(1+t^l)^2p_n^2(t)w_Q^2(t)dt \\ &\le cq_{n+l}^{2l}\exp(-c_1 n)\int_{-\infty}^{\infty}p_n^2(t)w_Q^2(t)dt \\ &\le c\exp(-c_1 n). \end{aligned} \tag{3.17}$$

Since $q_{n+l} \le cq_n$, (3.15) is now proved. Applying (3.15) with σ replaced by a larger number, we deduce that

$$\int_{|x|\ge A_1 q_n} |u_\sigma(x)\mathcal{P}_n(x)|^r dx \le c\exp(-c_1 n) \tag{3.18a}$$

for a suitably chosen A_1. Using (2.5a), we find that

$$\int_{A^* q_n \le |x| \le A_1 q_n} |u_\sigma(x)\mathcal{P}_n(x)|^r dx \le cq_n^{r\sigma+Kr+1}. \tag{3.18b}$$

Using (2.5b), we get

$$\begin{aligned}
&\int_{|x|\le A^* q_n} |u_\sigma(x)\mathcal{P}_n(x)|^r dx \\
&\qquad \le cq_n^{-r/2} \int_{|x|\le A^* q_n} u_\sigma^r(x)dx \\
&\qquad \le cq_n^{-r/2} \begin{cases} \log q_n & \text{if } \sigma = -1/r, \\ 1 & \text{if } \sigma < -1/r, \\ q_n^{r\sigma+1} & \text{if } \sigma > -1/r. \end{cases}
\end{aligned} \tag{3.18c}$$

The estimate (3.16) can be deduced from (3.18a), (3.18b), and (3.18c). □

Proof of Theorem 2.1. As we observed earlier, it is enough to prove (3.14). Using Hölder's inequality,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} g(y)\mathcal{P}_{n-1}(x)\mathcal{P}_{n-1}(y)u_b(x)dy \right|^p dx \\
&\qquad = \left(\int_{-\infty}^{\infty} \left| u_b(x)\mathcal{P}_{n-1}(x) \right|^p dx \right) \left(\int_{-\infty}^{\infty} \left| g(y)\mathcal{P}_{n-1}(y) \right| dy \right)^p \\
&\qquad \le ||u_b\mathcal{P}_{n-1}||_p^p ||u_B^{-1}\mathcal{P}_{n-1}||_q^p ||gu_B||_p^p
\end{aligned} \tag{3.19}$$

where

$$q := p/(p-1). \tag{3.20}$$

Next, we observe that $u_B^{-1} = u_{(-B)}$ and $q_{n-1} \le q_n$. If $K > -1/2$ then we may use (3.16) and the assumptions (2.6c) to conclude that

$$\int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} g(y)\mathcal{P}_{n-1}(x)\mathcal{P}_{n-1}(y)u_b(x)dy \right|^p dx \le c||gu_B||_p^p. \tag{3.21}$$

If $K = -1/2$, then the estimation can be done by considering four simple cases. We omit these details. Thus, (3.14) is proved for $k = 1$.

Next, we let

$$R := B - 1/2 - K. \tag{3.22}$$

Then, in view of (2.6) and Lemma 3.1(d), for any ϕ with $u_R\phi \in L^p(\mathbb{R})$,

$$\int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} \frac{\phi(y)}{x-y} dy \right|^p u_b(x)^p dx \le c||u_R\phi||_p^p. \tag{3.23}$$

So, using (2.5c) and (3.23), we get

(3.24)
$$\int_{-\infty}^{\infty}\left|\int_{-\infty}^{\infty} g(y)h_2(x,y)u_b(x)dy\right|^p dx$$
$$\le cq_n^p \int_{-\infty}^{\infty}\left|\int_{-\infty}^{\infty}\frac{g(y)\mathcal{P}_n(y)}{x-y}dy\right|^p \left|\mathcal{P}_n(x)-\mathcal{P}_{n-2}(x)\right|^p u_b(x)^p dx$$
$$\le cq_n^{p/2}\int_{-\infty}^{\infty}|g(y)\mathcal{P}_n(y)|^p u_R(y)^p dy.$$

Next, we apply (3.16) with $r=\infty$ and $\sigma = R-B$ and then use (3.22) to see that for $y\in(0,\infty)$,

(3.25)
$$|\mathcal{P}_n(y)u_R(y)| = |\mathcal{P}_n(y)u_{R-B}(y)||u_B(y)|$$
$$\le cq_n^{-1/2}|u_B(y)|.$$

Substituting from (3.25) into (3.24), we get

(3.26)
$$\int_{-\infty}^{\infty}\left|\int_{-\infty}^{\infty} g(y)h_2(x,y)u_b(x)dx\right|^p dy \le c||gu_B||_p^p.$$

We prove the estimate for h_3 in exactly the same way or by using a duality argument. □

Note added in proof. The authors have recently obtained more precise results similar to Theorem 1.1 in the case where $Q(x)=x^m$ (m an even integer).

Acknowledgments. The authors thank Professor Dr. Doron Lubinsky for pointing out a mistake in the original version of the paper and making various useful suggestions towards the improvement of the presentation in the paper.

REFERENCES

[1] R. Askey and S. Wainger, *Mean convergence of expansions in Laguerre and Hermite series,* Amer. J. Math., 87 (1965), pp. 695–708.
[2] W. C. Bauldry, *Estimates of asymmetric Freud polynomials,* J. Approx. Theory, to appear.
[3] S. Bonan and D. S. Clark, *Estimates of the Hermite and Freud polynomials,* J. Approx. Theory, to appear.
[4] G. Freud, *Orthogonal Polynomials,* Pergamon Press, Oxford, 1971.
[5] D. S. Lubinsky,*On Nevai's bounds for orthogonal polynomials associated with exponential weights,* J. Approx. Theory, 44 (1985), pp. 86–91.
[6] ———, *A survey of general orthogonal orthogonal polynomials for weights on the finite and infinite intervals,* Acta Appl. Math., 10 (1987), pp. 237–296.
[7] ———, *Strong asymptotics for extremal errors and polynomials associated with Erdös-type weights,* Pitman Res. Notes, 202, Longman Scientific and Technical, Harlow, Essex and John Wiley, New York, 1989.
[8] D. S. Lubinsky and E. B. Saff, *Strong asymptotics for extremal errors and extremal polynomials associated with weights on* $(-\infty,\infty)$, Lecture Notes, 1305, Springer-Verlag, New York, 1988.
[9] A. Mate, P. Nevai, and V. Totik, *Asymptotics for the zeros of orthogonal polynomials associated with infinite intervals,* J. London Math. Soc., 33 (1986), pp. 303–310.
[10] H. N. Mhaskar, *Bounds for certain Freud-type orthogonal polynomials,* J. Approx. Theory, to appear.

[11] ———, *Weighted analogues of Nikolskii-type inequalities and their applications,* in Conference on Harmonic Analysis in honor of A. Zygmund, Vol. II, W. Beckner, A. P. Calderón, R. Fefferman, and P. W. Jones eds., Wadsworth, Belmont, CA, 1983, pp. 783–801.

[12] H. N. MHASKAR AND E. B. SAFF, *Where does the L^p-norm of a weighted polynomial live?,* Trans. Amer. Math. Soc., 303 (1987), pp. 109–124. (Errata, same journal, 308 (1988), p. 431.)

[13] B. MUCKENHOUPT, *Mean convergence of Hermite and Laguerre series,* I, Trans. Amer. Math. Soc., 147 (1970), pp. 419–431.

[14] ———, *Mean convergence of Hermite and Laguerre series,* II, Trans. Amer. Math. Soc., 147 (1970), pp. 433–460.

[15] P. NEVAI, *Asymptotics for orthogonal polynomials associated with $exp(-x^4)$,* SIAM J. Math. Anal., 15 (1984), pp. 1177–1187.

[16] H. POLLARD, *The mean convergence of orthogonal series,* I, Trans. Amer. Math. Soc., 62 (1947), pp. 387–403.

[17] ———, *The mean convergence of orthogonal series,* II, Trans. Amer. Math. Soc., 63 (1948), pp. 355–367.

[18] E. A. RAHMANOV, *On asymptotic properties of polynomials orthogonal on the real axis,* Math. USSR-Sb., 47 (1984), pp. 155–193.

[19] A. F. TIMAN, *Theory of Approximation of Functions of a Real Variable,* Macmillan, New York, 1963.

[20] W. VAN ASSCHE, *Asymptotics for orthogonal polynomials,* Lecture Notes, 1265, Springer-Verlag, New York, 1987.

[21] A. ZYGMUND, *Trigonometric Series,* Cambridge University Press, Cambridge, 1977.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, pp. 856-876, May 1991

019

REPRODUCING FORMULAS AND DOUBLE ORTHOGONALITY IN BARGMANN AND BERGMAN SPACES*

KRISTIAN SEIP†

Abstract. It is observed that the reproducing kernels of the Bargmann spaces in $\mathbf{C}^n$ act reproducingly over any polyball (apart, of course, from positive constants depending on the radii of the polyball). It is noticed likewise that the reproducing kernels of the Bergman spaces over the unit ball in $\mathbf{C}^n$ act reproducingly over any ball (that is, ball in the Bergman metric). From these observations the eigenvalues and eigenfunctions of certain concentration operators are found. These eigenfunctions can be viewed as analogues to the prolate spheroidal wave functions in the Paley-Wiener space, but they are simpler and have nice properties which the prolate spheroidal wave functions do not have. Such expansions are exploited to yield analogues to results on sampling of bandlimited signals: necessary density conditions for sampling and interpolation, and jittered sampling. These results can be interpreted as results on irregular discrete representations of particular short-time Fourier and wavelet transformations.

Key words. reproducing formula, double orthogonality, concentration operator, Bergman space, Bargmann space, wavelets, sampling, interpolation, jittered sampling

AMS(MOS) subject classifications. 94A05, 41A05, 44A15, 30C40

1. Introduction. Let $\mathcal{H}$ be a reproducing kernel Hilbert space being a subspace of $L^2(X, dm(x))$ where X is a measure space with positive measure m. We denote the reproducing kernel of $\mathcal{H}$ by $K(t, x)$. This is the unique Hermitian function defined on $X \times X$ with the following properties. $K(\cdot, x)$ belongs to $\mathcal{H}$ for each fixed x, and for each $f \in L^2(X, dm(x))$ we have

$$(Pf)(t) = \int_X K(t, x) f(x)\, dm(x),$$

where P denotes projection onto $\mathcal{H}$.

Let next q be some nonnegatively valued and essentially bounded function on X. We denote the inner product of $L^2(X, dm(x))$ by $(\cdot, \cdot)$ and that of $L^2(X, q(x)\, dm(x))$ by $(\cdot, \cdot)_q$. A sequence of elements in $\mathcal{H}$ will be said to be *doubly orthogonal* if it is orthogonal both with respect to $(\cdot, \cdot)$ and to $(\cdot, \cdot)_q$. Along with q we define the operator Q on $L^2(X, dm(x))$ by

$$(Qf)(x) = q(x) f(x).$$

We prove the following easy proposition.

PROPOSITION 1. *Let $\{f_k\}$ be an orthonormal basis of $\mathcal{H}$. Then $\{f_k\}$ is also orthogonal with respect to $(\cdot, \cdot)_q$ if and only if the functions f_k are eigenfunctions of PQ.*

Proof. Assume first double orthogonality. Then for any k,

$$0 = (f_i, f_k)_q = (f_i, PQf_k)$$

holds for all $i \neq k$, which by the assumed completeness of $\{f_k\}$ means that $PQf_k = \lambda_k f_k$.

Assume next that f_k are eigenfunctions of PQ. Then

$$(f_i, f_k)_q = (f_i, PQf_k) = (f_i, \lambda_k f_k),$$

which equals zero whenever $i \neq k$. □

* Received by the editors January 25, 1989; accepted for publication (in revised form) June 11, 1990.
† Division of Mathematical Sciences, University of Trondheim, N-7034 Trondheim-NTH, Norway.

A necessary and sufficient condition that $\mathscr{H}$ possess a doubly orthogonal basis is thus that the positive operator PQ have a purely discrete spectrum. This is definitely the case if PQ is compact. A sufficient condition that PQ be compact is

$$\int_X K(x, x)q(x)\, dm(x) < \infty. \tag{1}$$

Indeed, by the reproducing property of $K(t, x)$ we have

$$\int_X \int_X |K(t, x)q(x)|^2\, dm(t)\, dm(x) = \int_X K(x, x)q(x)^2\, dm(x),$$

which shows that (1) ensures square integrability of $K(t, x)q(x)$. Of course, (1) also implies that PQ is trace class.

If q is the characteristic function of some subset Y of X, PQ may be called a *concentration operator.* In the case that PQ is compact, eigenvalues and eigenfunctions can then be found by successively maximizing the *concentration*

$$\lambda = \frac{\int_Y |f(x)|^2\, dm(x)}{\int_X |f(x)|^2\, dm(x)}$$

over the orthogonal complement in $\mathscr{H}$ of the linear span of the previously determined eigenfunctions.

The first to consider such doubly orthogonal bases seems to have been Bergman in his study of the spaces that now commonly bear his name [4, pp. 14–18]. A later example are the celebrated prolate spheroidal wave functions of Landau, Slepian, and Pollak [40a–c]. By the work of Daubechies and Paul [11] we recognize a rather close connection between these two examples, both giving eigenfunctions of time-frequency-limiting operators.

Daubechies and Paul's problem of finding the eigenvalues and eigenfunctions of certain concentration operators is essentially the same as that of Bergman, but they formulate it in other spaces of functions than Bergman did. Being unaware of the connection to Bergman, they solve the problem in the same way as the prolate spheroidal wave functions were found, that is, by finding a differential operator commuting with the concentration operator.

We shall see below that in Bergman's formulation such problems are solved very easily in a direct manner. This will lead us to expansions with some rather nice properties. In addition to the double orthogonality and the concentration properties, the eigenfunctions will have certain reproducing properties, and the expansions are simply a kind of Taylor series. In the last part of this paper we present applications of such expansions yielding analogues of certain results on irregular sampling of bandlimited signals. These analogues are relevant for the theory of short-time Fourier and wavelet transformations.

We start our discussion of Bargmann and Bergman spaces in $\mathbf{C}^n$. We have chosen to treat the Bergman spaces over the unit ball. We could just as well have chosen the unit polydisk or more generally any unit polyball. To what class of symmetric domains in $\mathbf{C}^n$ our results can be extended we do not know. Our primary interest in these problems has been the above-mentioned applications, and for that purpose considering the ball has been more than sufficient.

It should be mentioned that what we have chosen to call concentration operators constitute a special class of Toeplitz operators, which in recent years have been studied quite extensively (see [6], [36] and the references therein).

Before turning to the main discussion, let us remark that some of the results of this paper are reported in [39].

2. Reproducing formulas and concentration operators. Below are listed some basic definitions and notational conventions that will be needed. For readers not familiar with the Bargmann and Bergman spaces [17] (Bargmann case) and [26] (Bergman case) might be good background references.

Notation. B will denote the unit ball in $\mathbf{C}^n$, $B(R)$ the ball of radius R centered at zero, $D(R_1, \cdots, R_n)$ the polydisk of radii $R_1, \cdots, R_n$ also centered at zero and S the unit sphere in $\mathbf{C}^n$, n a fixed positive integer. μ will denote Lebesgue measure on $\mathbf{C}^n$, normalized so that $\mu(B) = 1$, and σ will mean the rotation-invariant positive Borel measure on S for which $\sigma(S) = 1$. We use standard multi-index notation. If β is any multi-index then D_z^β is the corresponding differential operator, with the index indicating with respect to which variable one is differentiating.

We next introduce some symbols that will have different meaning depending on which case we have under consideration.

We start with the Bargmann case. We introduce the Bargmann kernel of $\mathbf{C}^n$ which is the function

$$K(z, \zeta) = e^{\langle z, \zeta \rangle}$$

defined on $\mathbf{C}^n \times \mathbf{C}^n$. We will use the weighted measure

$$d\omega_\alpha(z) = K(z, z)^{-\alpha}\, d\mu(z),$$

where $\alpha > 0$ is some fixed real number. We let $H(\mathbf{C}^n)$ be the class of all functions holomorphic in $\mathbf{C}^n$ and define the Bargmann spaces

$$F^\alpha = H(\mathbf{C}^n) \cap L^1(\mathbf{C}^n, d\omega_\alpha(z))$$

and

$$F^{\alpha,2}(\mathbf{C}^n) = H(\mathbf{C}^n) \cap L^2(\mathbf{C}^n, d\omega_\alpha(z)).$$

We define

$$T_z(\zeta) = z - \zeta.$$

We next turn to the Bergman case. The Bergman kernel of B is the function

$$K(z, \zeta) = (1 - \langle z, \zeta \rangle)^{-n-1}$$

defined on $B \times B$. We introduce the weighted measure

$$d\omega_\alpha(z) = K(z, z)^{-\alpha}\, d\mu(z),$$

where $\alpha > -1/(n+1)$ is some fixed real number. We let $H(B)$ be the class of all functions holomorphic in B and define the Bergman spaces

$$A^\alpha(B) = H(B) \cap L^1(B, d\omega_\alpha(z))$$

and

$$A^{\alpha,2}(B) = H(B) \cap L^2(B, d\omega_\alpha(z)).$$

The restriction on α is put to make these spaces nontrivial.

We shall denote the automorphism of B that interchanges z and zero (see [38, p. 25]) by T_z.

The Bargmann case. It is easy to show that the following holds.

THEOREM 2.1. *For any $f \in H(\mathbf{C}^n)$, any multi-index β, any $z \in \mathbf{C}^n$, and any $R > 0$ we have the following formula*:

$$\{D_\zeta^\beta [f(T_z\zeta)K(\zeta, z)^\alpha]\}_{\zeta=0} = C \int_{T_z B(R)} f(w)\overline{(T_z w)^\beta K(w, z)^\alpha}\, dw_\alpha(w) \tag{2}$$

with

$$C = C(R, |\beta|, \alpha) = \frac{(n+|\beta|-1)!}{n!}\left(\int_0^{R^2} r^{n+|\beta|-1} e^{-\alpha r}\, dr\right)^{-1}.$$

If $f \in F^\alpha$, the formula is also valid for $R = \infty$.

Proof. For arbitrary $f \in H(\mathbf{C}^n)$, we use the Cauchy formula [38, p. 39] for the function $f(rz)$ yielding

$$r^{|\beta|}\{[D^\beta f](\zeta)\}_{\zeta=0} = \frac{(n+|\beta|-1)!}{(n-1)!}\int_S f(rw)\bar{w}^\beta\, d\sigma(w).$$

We then multiply each side by $2nr^{2n-1}r^{|\beta|}e^{-\alpha r^2}\, dr$ and integrate over r from zero to R. Using the formula for integration in polar coordinates gives

$$\{[D^\beta f](\zeta)\}_{\zeta=0} = C\int_{B(R)} f(w)\overline{w^\beta}\, d\omega_\alpha(w). \tag{3}$$

We apply this formula to $f(T_z\zeta)K(\zeta, z)^\alpha$, then make a change of variables in the integral, and the result follows. □

THEOREM 2.2. *For any $f \in H(\mathbf{C}^n)$, any multi-index β, any $z \in \mathbf{C}^n$, and any $R_1, \cdots, R_n > 0$ we have the following formula*:

$$\{D_\zeta^\beta [f(T_z\zeta)K(\zeta, z)^\alpha]\}_{\zeta=0} = C \int_{T_z D(R_1,\cdots,R_n)} f(w)\overline{(T_z w)^\beta K(w, z)^\alpha}\, d\omega_\alpha(w) \tag{4}$$

with

$$C = C(R_1, \cdots, R_n, |\beta|, \alpha) = \prod_{k=1}^n \beta_k !\left(\int_0^{R_k^2} r^{\beta_k} e^{-\alpha r_k}\, dr_k\right)^{-1}.$$

If $f \in F^\alpha$, the formula is also valid when at least one R_k is infinite.

Proof. The proof is the easy and obvious variant of the proof of Theorem 2.1 where we use the Cauchy formula for polydisks. □

Remark 1. Note that when $\beta = 0$ the above theorems reduce to the following statement. The reproducing kernel of $F^{\alpha,2}$ acts reproducingly over any ball and over any polydisk, apart from positive factors depending only on the radii.

Remark 2. There is of course an obvious generalization of the statements above to "any polyball," which we have not made explicitly.

We shall now relate the discussion of the Introduction to the reproducing formulas above. This amounts merely to a simple extension of an observation made by Bergman [4, pp. 14].

Here I_Ω shall mean the operator of multiplication by the characteristic function of the set Ω.

THEOREM 2.3. *For $R < \infty$ the operator $P_{F^{\alpha,2}} I_{T_z B(R)}$ has eigenvalues*

$$\lambda_{|\beta|}(R) = \frac{C(R, |\beta|, \alpha)}{C(\infty, |\beta|, \alpha)} = \frac{1}{(n+|\beta|-1)!}\int_0^{\alpha R^2} r^{n+|\beta|-1} e^{-r}\, dr.$$

For $R_1, \cdots, R_n < \infty$ the operator $P_{F^{\alpha,2}(B)} I_{T_z D(R_1,\cdots,R_n)}$ has eigenvalues

$$\lambda_\beta(R_1, \cdots, R_n) = \frac{C(R_1, \cdots, R_n, \beta, \alpha)}{C(\infty, \cdots, \infty, \beta, \alpha)} = \prod_{k=1}^{n} \frac{1}{\beta_k!} \int_0^{\alpha R_k^2} r^{k+|\beta|-1} e^{-r}\, dr.$$

In either case the corresponding eigenfunctions are

$$f_\beta^z(\zeta) = \rho_\beta^z K(\zeta, z)^\alpha (T_z\zeta)^\beta,$$

the normalizing factor ρ_β^z being

$$\rho_\beta^z = \left(\frac{\beta! K(z, z)^\alpha}{C(\infty, |\beta|, \alpha)} \right)^{-1/2}. \tag{5}$$

Proof. The operator in question is clearly compact and trace class since, with the obvious definitions, (1) is satisfied.

For any $f \in F^{\alpha,2}(B)$, any $z \in B$, and any multi-index β, we have by Theorem 2.1

$$0 = \frac{\lambda_{|\beta|}(R)\rho_\beta^z}{C(\infty, |\beta|, \alpha)} \{\{D_\zeta^\beta [f(T_z\zeta) K(\zeta, z)^\alpha]\}_{\zeta=0} - \{D_\zeta^\beta [f(T_z\zeta) K(\zeta, z)^\alpha]\}_{\zeta=0}\}$$

$$= \int_{C^n} f(\zeta) \{\lambda_{|\beta|}(R) \overline{f_\beta^z(\zeta)} - \overline{(P_{F^{\alpha,2}} I_{T_z B(R)} f_\beta^z)(\zeta)}\}\, d\omega_\alpha(\zeta),$$

and consequently

$$\lambda_\beta(R) f_\beta^z(\zeta) = (P_{F^{\alpha,2}} I_{T_z B(R)} f_\beta^z)(\zeta).$$

In the case of polydisks, we use Theorem 2.2 in the same way.

The proof is completed by recalling that the set $\{f_\beta^z(\zeta)\}$ constitutes an orthonormal basis for $F^{\alpha,2}$ (the normalizing factor ρ_β^z is easily found using (2)). □

Remark. It may be observed that we have obtained the solutions to the eigenvalue problems considered in [11, I] avoiding the use of commuting second-order differential operators. Actually, we have gained a bit since [11, I] only covers the case of polydisks and not that of balls.

The Bergman case. Here we must work slightly harder. We find it convenient first to collect a few auxiliary facts.

Regarding the group of automorphisms of B and their relation to the Bergman kernel we note the following.

LEMMA 2.4. *For any $z, \zeta, w \in B$ and for any automorphism Φ of B we have*

$$T_z 0 = z, \qquad T_z z = 0, \tag{6}$$

$$T_z^{-1} = T_z, \tag{7}$$

$$K(\Phi z, \Phi\zeta) \det \Phi' z\, \overline{\det \Phi' \zeta} = K(z, \zeta), \tag{8}$$

$$K(T_z w, T_z \zeta) = \frac{K(z, z) K(w, \zeta)}{K(w, z) K(z, \zeta)}, \tag{9}$$

$$\det \Phi'(\Phi^{-1} z) = (\det (\Phi^{-1})'(z))^{-1}. \tag{10}$$

Proof. Equation (10) follows by the chain rule

$$I = (\Phi\Phi^{-1})'(z) = \Phi'(\Phi^{-1} z) \cdot (\Phi^{-1})'(z).$$

The rest are standard results which can be found in Chapter 2 of [38]. □

We next note the following.

LEMMA 2.5. *For any multi-index* β, *any* $f \in H(B)$, *and any* $0 < R < 1$, *we have*

$$\{[D^\beta f](\zeta)\}_{\zeta=0} = C(R, |\beta|, \alpha) \int_{B(R)} f(w)\bar{w}^\beta \, d\omega_\alpha(w) \tag{11}$$

with

$$C(R, |\beta|, \alpha) = \frac{(n+|\beta|-1)!}{n!} \int_0^{R^2} r^{n+|\beta|-1}(1-r)^\alpha \, dr. \tag{12}$$

Proof. This is proved in exactly the same way as formula (3). □

We are now ready to prove the following.

THEOREM 2.6. *For any* $f \in H(B)$, *any multi-index* β, *any* $z \in B$, *and any* $0 < R < 1$, *we have the following formula*:

$$\{D_\zeta^\beta[f(T_z\zeta)K(\zeta, z)^{1+\alpha}]\}_{\zeta=0} = C \int_{T_zB(R)} f(w)\overline{(T_zw)^\beta K(w, z)^{1+\alpha}} \, d\omega_\alpha(w) \tag{13}$$

with $C = C(R, |\beta|, \alpha)$ *as in Lemma* 2.5. *If* $f \in A^\alpha(B)$, *the formula is also valid for* $R = 1$.

Proof. Given $f \in H(B)$, $z \in B$, we form the function

$$g(\zeta) = f(T_z\zeta)(\det T_z'\zeta)^{1+\alpha}.$$

An application of Lemma 2.5 to g yields

$$\{D_\zeta^\beta[f(T_z\zeta)(\det T_z'\zeta)^{1+\alpha}]\}_{\zeta=0} = C(R, |\beta|, \alpha) \int_{B(R)} g(u)\overline{u^\beta K(u, 0)^{1+\alpha}} \, d\omega_\alpha(u). \tag{14}$$

We make the change of variables $u = T_zw$. By (8) you see that

$$dw_\alpha(u) = |\det T_z'w|^2 K(T_zw, T_zw)^{-\alpha} \, d\mu(w) = |\det T_z'w|^{2+2\alpha} \, d\omega_\alpha(w).$$

So we have

$$\begin{aligned} &\int_{B(R)} g(u)\overline{u^\beta K(u, 0)^{1+\alpha}} \, d\omega_\alpha(u) \\ &= \int_{T_zB(R)} f(w)\overline{(T_zw)^\beta K(T_zw, 0)^{1+\alpha}(\det T_z'w)^{1+\alpha}} \, d\omega_\alpha(w) \qquad (15) \\ &= \frac{1}{(\det T_z'z)^{1+\alpha}} \int_{T_zB(R)} f(w)\overline{(T_zw)^\beta K(w, z)^{1+\alpha}} \, d\omega_\alpha(w), \end{aligned}$$

where we have used (6), (7), (8), and (10). By (7), (10), and (8), we have

$$\det T_z'z = (\det T_z'0)^{-1} = K(z, z)\overline{\det T_z'0}. \tag{16}$$

By putting (16) into (15) and then (15) into (14), we get

$$\begin{aligned} &\{D_\zeta^\beta[f(T_z\zeta)K(z, z)^{1+\alpha}(\det T_z'\zeta)^{1+\alpha}\overline{(\det T_z'0)^{1+\alpha}}]\}_{\zeta=0} \\ &= C(R, |\beta|, \alpha) \int_{T_zB(R)} f(w)\overline{(T_zw)^\beta K(w, z)^{1+\alpha}} \, d\omega_\alpha(w). \end{aligned}$$

We finally use (8) and (9) to conclude that we have obtained the desired result. If $f \in A^\alpha(B)$, the validity of the formula for $R = 1$ is obvious. □

Remark 1. The above proof is based on a technique originally used by Bers [5], [29].

Remark 2. Note that when $\beta = 0$ the theorem reduces to the following statement. The reproducing kernel of $A^{\alpha,2}(B)$ acts reproducingly over any ball (ball is now to be understood in the Bergman metric), apart from a positive factor depending only on the radius of the ball.

As in the Bargmann case we shall relate the discussion of the Introduction to the reproducing formulas that we have found.

THEOREM 2.7. *For $R < 1$ the operator $P_{A^{\alpha,2}(B)} I_{T_z B(R)}$ has eigenvalues*

$$\lambda_{|\beta|}(R) = \frac{C(R, |\beta|, \alpha)}{C(1, |\beta|, \alpha)} = \frac{(\alpha+1)\cdots(\alpha+n+|\beta|)}{(n+|\beta|-1)!} \int_0^{R^2} r^{n+|\beta|-1}(1-r)^{\alpha}\, dr$$

with corresponding eigenfunctions

$$f_\beta^z(\zeta) = \rho_\beta^z K(\zeta, z)^{\alpha+1}(T_z\zeta)^\beta,$$

the normalizing factor ρ_β^z being

$$\rho_\beta^z = \left(\frac{\beta!\, K(z,z)^{1+\alpha}}{C(1, |\beta|, \alpha)}\right)^{-1/2}. \tag{17}$$

Proof. The proof is exactly like the proof of Theorem 2.3. □

Remark 1. It may be observed that we have obtained the solutions to the eigenvalue problems considered in [11, II], avoiding the use of commuting second-order differential operators.

Remark 2. To get a feeling for the behaviour of the eigenvalues it is instructive to consider the case $\alpha = 0$ in which they are particlarly simple: $\lambda_{|\beta|}(R) = R^{2(n+|\beta|)}$. For a general discussion on the asymptotics the reader should consult [11, II].

Both in the Bargmann and in the Bergman case the functions f_β^z provide us with a useful tool for a local analysis. We obtain "Taylor expansions," and we have the double orthogonality, the concentration properties, and the fact that the first function f_0^z acts reproducingly over any ball centered at z. In the next sections we shall present two examples of how these features may be exploited.

3. Preparation for two applications. For the sake of simplicity we restrict ourselves to the case where $n = 1$ for the rest of this paper.

We start by introducing some notation which, for much of the discussion that follows, will enable us not to distinguish between the Bargmann and the Bergman cases. From now on, unless otherwise specified, all statements should thus be read with respect to either of the two cases.

We should warn the reader that in doing this we will make some minor changes from the notational conventions of the previous section. This will be very convenient and should not cause any trouble if care is taken to note the differences.

We consider the Bargmann spaces $F^{\alpha,2}$ and the Bergman spaces $A^{\alpha,2}(\Delta)$ (we denote the unit disk in $\mathbf{C}$ by Δ). The reproducing kernels of these spaces will all be denoted by $R(z, \zeta)$, that is, for $F^{\alpha,2}$ we let

$$R(z, \zeta) = \alpha\, e^{\alpha z\bar{\zeta}}$$

and if we consider $A^{\alpha,2}(\Delta)$, we put

$$R(z, \zeta) = (2\alpha+1)(1-z\bar{\zeta})^{-2\alpha-2}.$$

We put $d\omega(\zeta) = d\omega_\alpha(\zeta)$, $(\cdot, \cdot)$ will mean the corresponding inner product and $\|\cdot\|$ the corresponding norm. Let distance $d(\cdot, \cdot)$ mean the Euclidean distance in the Bargmann

case and the hyperbolic distance in the Bergman case. Define

$$c_n = \begin{cases} \alpha^n/\pi & \text{Bargmann case,} \\ (2\alpha+1)(2\alpha+2)\cdots(2\alpha+n+1)/\pi & \text{Bergman case,} \end{cases}$$

$$\lambda_n(R) = \begin{cases} \dfrac{1}{n!}\displaystyle\int_0^{\alpha R^2} t^n e^{-t}\,dt & \text{Bargmann case,} \\ \dfrac{c_n\pi}{n!}\displaystyle\int_0^{\tanh^2 R/2} t^n(1-t)^{2\alpha}\,dt & \text{Bergman case,} \end{cases}$$

$$\Delta_z(R) = \{\zeta\colon d(z,\zeta) < R\}.$$

Note that in the Bergman case the argument R in $\lambda_n(R)$ is now a hyperbolic distance.

With this notation (2) (or (4)) and (13) can be written compactly as

$$\left[\frac{d^n}{dw^n}(f(T_z w)K(z,w))\right]_{w=0} = \frac{c_n}{\lambda_n(R)}\iint_{\Delta_z(R)} f(\zeta)\overline{(T_z\zeta)^n K(\zeta,z)}\,d\omega(\zeta). \tag{18}$$

Note in particular that this means

$$f(z) = \frac{1}{\lambda_0(R)}\iint_{\Delta_z(R)} f(\zeta)\overline{R(\zeta,z)}\,d\omega(\zeta). \tag{19}$$

The eigenfunctions of the concentration operators will now naturally be denoted

$$f_n^z(\zeta) = \rho_n^z R(z,\zeta)(T_z\zeta)^n$$

so that ρ_n^z is modified by a factor of $1/\alpha$ or $1/(2\alpha+1)$ from the definitions (5) or (17).

Let us now indicate the motivation for the applications that will be discussed in the last two sections of this paper. Our Bargmann spaces can be associated with the range of the short-time Fourier transformation when applying a Gaussian window function, as first proposed in [18]. For a discussion of this fact and for a short historical review of the importance of the Bargmann space in signal analysis and in different contexts in mathematical physics, see [12]. Another source of much information is [25].

By making the standard transformation from the upper halfplane to the unit disk, our Bergman spaces can be associated with the range of the wavelet transformation when choosing as wavelet the function $g(t)$ with Fourier transform

$$\hat{g}(\xi) = \xi^{\alpha+1/2}\,e^{-\xi}$$

[22]. During the last few years there has been a great deal of interest in such transformations, and interesting applications have been found in signal analysis, quantum mechanics, applied mathematics, and harmonic analysis [19], [21], [22], [32], [8]. This particular choice of wavelet has been studied extensively by Paul from a quantum mechanical point of view [33], [34]. It leads to coherent state representations related to the radial harmonic oscillator and the Coulomb potential problem for the hydrogen atom.

Discrete representations (or discrete transformations) are, of course, of great importance for various practical and theoretical reasons. In our formulation a discrete representation will be associated with some discrete set of points either in $\mathbf{C}$ or in Δ. For regular lattices we have now reached an almost complete understanding [1], [13], [12], [9]. We urge the reader to consult [9], which should give the necessary background for the present discussion.

A related story should be mentioned here. Different orthonormal bases of wavelets have been discovered recently [41], [32], [2], [10], [30], which have quite amazing

mathematical properties and which have turned out to be of interest for many applications. There is also a recent discovery of a class of orthonormal bases that is related to the Bargmann case [14].

The kinds of lattices that are chosen for discrete representations for short-time Fourier and wavelet transformations may be compared to the usual regular sampling of bandlimited signals. The question has been raised to which extent analogues to results on irregular sampling of bandlimited signals can be found. It seems natural to start such an investigation in a situation which is similar to that of bandlimited signals in the sense that we have available spaces of holomorphic functions.

Returning to our Bargmann and Bergman spaces, let us make precise what we mean by "discrete representations." We say that $\{z_n\}$ is a *set of uniqueness* if

$$f(z_n)=0 \quad \text{for all } z_n \Rightarrow f\equiv 0.$$

This is equivalent to saying that the set $\{R(z_n, z)/R(z_n, z_n)^{1/2}\}$ is complete in the space at hand. $\{z_n\}$ is said to be a *set of interpolation* if to any square-integrable sequence $\{c_n\}$ there exists an f such that $f(z_n)/R(z_n, z_n)^{1/2}=c_n$ for all n.

If the sequences $\{f(z_n)\}$ are to represent the vectors f, we should obviously require $\{z_n\}$ to be a set of uniqueness. But we should also require the inversion problem to be well posed. This means that we need the operator $f\mapsto\{f(z_n)/R(z_n, z_n)^{1/2}\}$ into l^2 to be bounded and to have a bounded inverse. We state this by requiring that for all f in the space in question we should have

$$A\|f\|^2\leqq\sum\frac{|f(z_n)|^2}{R(z_n, z_n)}\leqq B\|f\|^2 \tag{20}$$

with $0<A\leqq B<\infty$. Following Landau [27] we shall say that $\{z_n\}$ is a *set of sampling* if this is satisfied, which is seen to be equivalent to saying that the set $\{R(z_n, z)/R(z_n, z_n)^{1/2}\}$ is a *frame* [16], [42]. The numbers A and B will be denoted the *frame bounds* of $\{z_n\}$. We observe that a set of sampling is in particular a set of uniquenesses.

There exists a kind of duality between sets of sampling and sets of interpolation, which was clearly formulated by Landau [27], [28] in connection with bandlimited signals. This will also be made precise by Theorem 4.4 of the present work. We observe that if $\{z_n\}$ is both a set of sampling and a set of interpolation, then the sequence $\{R(z_n, z)/R(z_n, z_n)^{1/2}\}$ constitutes a Riesz basis in the space at hand [42, p. 188].

4. Necessary density conditions for sampling and interpolation. For a discrete set of points Γ to be one of sampling, must the "density" of points in any part of the domain in question exceed some lower bound? Such a lower bound would correspond to the Nyquist rate of information theory. From what we know about sampling of bandlimited signals [27], [28] we might expect this to be the case. Were it not, we could, at least theoretically, find discrete representations that were more economic than the ones used now. Similarly, for Γ to be a set of interpolation, we would expect that the "density" of points in any part of the domain under consideration could not exceed some upper bound.

Guided by the work of Landau [27], [28], we shall search for such necessary conditions. But first we should know if the word "density" can be given a reasonable meaning.

Following Beurling [28, p. 47] we shall only be concerned with *uniformly discrete sets*, those sets which are separated by some least positive distance (in the Bergman case that is hyperbolic distance). In the Bargmann case we can easily borrow the

concept of density used by Landau [28, p. 47]. We thus fix some "nice" two-dimensional set I of measure 1, like the unit square or the unit disk/π. To any uniformly discrete set Γ we define $n^+(r)$ and $n^-(r)$ to be the largest and smallest number of points of Γ to be found in a translate of rI. We define upper and lower (uniform) densities of Γ,

$$D^+(\Gamma) = \limsup_{r\to\infty} \frac{n^+(rI)}{r^2} \quad \text{and} \quad D^-(\Gamma) = \liminf_{r\to\infty} \frac{n^-(rI)}{r^2}.$$

Landau shows that if the boundary of I has measure zero this definition is independent of I. If the upper and lower densities are equal, we define Γ to have (uniform) density $D(\Gamma) = D^+(\Gamma) = D^-(\Gamma)$.

We shall prove an analogue of Landau's result for bandlimited L^2 functions [27]: If Γ is a set of sampling then $D^-(\Gamma) \geqq m(S)/2\pi$, and if Γ is a set of interpolation then $D^+(\Gamma) \leqq m(S)/2\pi$ (here S denotes the frequency band which can consist of any finite number of intervals and $m(S)$ denotes the measure of S).

In the Bergman case the above density definitions are not well suited. If we chose the disk of unit area as the basic set I, and defined upper and lower densities as above, we would find that the sets constructed below would not have densities, that is, $D^+(\Gamma) > D^-(\Gamma)$. But our intuition tells us that, if any, these "regular" sets should have densities.

Similar problems have been pointed out in slightly different contexts in [9, p. 69] and in [11, I]. It seems to be an interesting challenge to find a suitable density concept in the hyperbolic case.

In both the Bargmann and the Bergman cases we shall exploit the similarities with the Paley-Wiener case to use the ideas contained in [27]. It will be shown that for the Bergman case we can at least find analogies to two of Landau's important auxiliary results, and we will point out the difficulty in finding a "complete" analogy.

Before stating the results, let us quite informally explain the idea of connecting concentration operators to our sampling/interpolation problem. For some large compact set Ω the number of eigenvalues of the corresponding concentration operator significantly different from zero gives us the "dimension" of the space of functions "concentrated" on Ω. Thus if the number of points in Ω from some discrete set Γ was much smaller than this "dimension," we could find a function which was zero at all the points in Ω, and which was essentially zero outside Ω. If this were the case for arbitrarily large Ω, we would expect that the left inequality in (20) could not be satisfied. A similar argument can be made for interpolation.

The following lemmas give estimates reflecting this rather loose statement.

LEMMA 4.1. *Suppose $\{z_k\}_{k=1}^{\infty}$ is a set of sampling. Let ζ be any point (in* **C** *or in* Δ*) and let $\{z_k^\zeta\}_{k=1}^{\infty}$ denote a reordering of this set such that*

$$0 \leqq d(z_1^\zeta, \zeta) \leqq d(z_2^\zeta, \zeta) \leqq \cdots.$$

Furthermore, let r be any positive number and let $n = n(r, \zeta)$ be the number such that

$$d(z_n^\zeta, \zeta) < r + \frac{\delta_0}{2} \leqq d(z_{n+1}^\zeta, \zeta)$$

(if such a number does not exist, define $n = n(r, \zeta) = 0$). Then there exists a constant $\gamma < 1$, independent of r and ζ, such that

$$\frac{|\tilde{\alpha}_n^\zeta|^2 \lambda_n(r) + \cdots + |\tilde{\alpha}_0^\zeta|^2 \lambda_0(r)}{|\tilde{\alpha}_n^\zeta|^2 + \cdots + |\tilde{\alpha}_0^\zeta|^2} \leqq \gamma, \tag{21}$$

where $\alpha_n^\zeta z^n + \cdots + \alpha_1^\zeta z + \alpha_0^\zeta = \prod_{k=1}^n (z - T_\zeta(z_k^\zeta))$ *and*

$$\tilde{\alpha}_k^\zeta = \begin{cases} \left(\dfrac{k!}{\alpha^{k-1}}\right)^{1/2} \alpha_k^\zeta & \textit{Bargmann case,} \\ \left(\dfrac{k!}{2(\alpha+1)\cdot(2(\alpha+1)+1)\cdots(2(\alpha+1)+k-1)}\right)^{1/2} \alpha_k^\zeta & \textit{Bergman case.} \end{cases}$$

Proof. Let $\phi_n^\zeta(z) = R(z,\zeta)\prod_{k=1}^n (T_\zeta(z) - T_\zeta(z_k^\zeta))$. Then

$$A\|\phi_n^\zeta\|^2 \leqq \sum_{k=1}^\infty \frac{|\phi_n^\zeta(z_k^\zeta)|^2}{R(z_k^\zeta, z_k^\zeta)} = \sum_{k=n+1}^\infty \frac{|\phi_n^\zeta(z_k^\zeta)|^2}{R(z_k^\zeta, z_k^\zeta)}.$$

By (19) and the Schwarz inequality we thus get

$$A\|\phi_n^\zeta\|^2 \leqq \frac{1}{\lambda_0(\delta_0/2)}\left(\|\phi_n^\zeta\|^2 - \int\int_{d(z,\zeta)<r} |\phi_n^\zeta(z)|^2\, d\omega(z)\right). \tag{22}$$

From the definitions of ϕ_n^ζ and of the eigenfunctions we may write

$$\phi_n^\zeta(z) = \frac{\|\phi_n^\zeta\|^2}{\sum_{k=0}^n |\tilde{\alpha}_k^\zeta|^2} \sum_{k=0}^n \tilde{\alpha}_k^\zeta f_n^\zeta(z).$$

We put this into the integral of (22) and invoke the double orthogonality and the concentration properties of the eigenfunctions to obtain (21) with $\gamma = 1 - A\lambda_0(\delta_0/2)$. □

Due to the simplicity of the eigenfunctions we have here obtained a sharpening of Lemma 1 of [27], that is, we observe that the above lemma implies the statement analogous to Landau's result.

LEMMA 4.2. *Let the conditions and definitions be as in Lemma* 4.1. *Then we have*

$$\lambda_n(r) \leqq \gamma \tag{23}$$

with γ *as in Lemma* 4.1.

To prove a sharpening of Landau's corresponding result on interpolation seems harder, and we confine ourselves to prove an analogue of Lemma 2 in [27].

LEMMA 4.3. *Suppose* $\{z_k\}_{k=1}^\infty$ *is a set of interpolation. Let* ζ *be any point (in* $\mathbf{C}$ *or in* Δ*) and let* $\{z_k^\zeta\}_{k=1}^\infty$ *denote a reordering of this set such that*

$$0 \leqq d(z_1^\zeta, \zeta) \leqq d(z_2^\zeta, \zeta) \leqq \cdots.$$

Furthermore, let r be any positive number and let $n = n(r,\zeta)$ *be the number such that*

$$d(z_n^\zeta, \zeta) < r - \frac{\delta_0}{2} \leqq d(z_{n+1}^\zeta, \zeta)$$

(if such a number does not exist, define $n = n(r,\zeta) = 0$ *and* $\lambda_{-1}(r) = 1$*). Then there exists a constant* $\delta > 0$*, independent of r and* ζ*, such that*

$$\lambda_{n-1}(r) \geqq \delta. \tag{24}$$

Proof. First we note that it is possible to perform the interpolation in a stable way. This happens if we interpolate in $E_0^\perp$, the orthogonal complement to the space E_0 of functions vanishing on $\{z_k\}$. Hence there exists a constant $A < \infty$ such that to any square-summable sequence $\{a_k\}$ there is a function $\phi \in E_0^\perp$ with $\phi(z_k)/R(z_k, z_k) = a_k$ and

$$A\|\phi\|^2 \leqq \sum |a_k|^2 = \sum \frac{|\phi(z_k)|^2}{R(z_k, z_k)}. \tag{25}$$

For a proof of this fact, see the proof of Proposition 1 in [27].

Second, we need the following form of the Weyl-Courant lemma

$$\lambda_{n-1}(r) \geqq \inf_{\phi \in C_n} \frac{\iint_{d(z,\zeta)<r} |\phi(z)|^2 \, d\omega(z)}{\iint |\phi(z)|^2 \, d\omega(z)}, \tag{26}$$

with C_n any subspace of dimension n. A proof can again be found in [27].

Next, let $\phi_k^\zeta \in E_0^\perp$ be the function whose value is 1 at z_k^ζ and zero at every z_j^ζ, $j \neq k$, and let $\tilde{C}_n = \operatorname{span}\{\phi_k^\zeta\}_{k=1}^n$. For any $\phi \in \tilde{C}_n$ we thus have

$$A\|\phi\|^2 \leqq \sum_{k=1}^{\infty} \frac{|\phi(z_k^\zeta)|^2}{R(z_k^\zeta, z_k^\zeta)} \leqq \sum_{k=1}^{n} \frac{|\phi(z_k^\zeta)|^2}{R(z_k^\zeta, z_k^\zeta)}.$$

By (19) and the Schwarz inequality we find

$$A\|\phi\|^2 \leqq \frac{1}{\lambda_0(\delta_0/2)} \iint_{d(x,y)<r} |\phi(z)|^2 \, d\omega(z).$$

Now the result follows from (26) with $\delta = A\lambda_0(\delta_0/2)$. □

We are now prepared to prove the following general result for the Bargmann case, stating that the Nyquist rate is a critical density both for sampling and for interpolation.

THEOREM 4.4. *In the Bargmann case we have the following. If the uniformly discrete set $\{z_n\}$ is a set of sampling, then $D^-(\{z_n\}) \geqq \alpha/\pi$. If the uniformly discrete set $\{z_n\}$ is a set of interpolation, then $D^+(\{z_n\}) \leqq \alpha/\pi$.*

Proof. We make use of the asymptotic behaviour of the eigenvalues ($0 < \delta < \gamma < 1$)

$$\frac{\#\{\lambda_k(R): \delta \leqq \lambda_k(R) < \gamma\}}{\pi R^2} = O(R^{-1}),$$

$$\frac{\#\{\lambda_k(R): \lambda_k(R) \geqq \tfrac{1}{2}\}}{\pi R^2} = \frac{\alpha}{\pi} + O(R^{-1})$$

(see [11, I]). Thus

$$\frac{\#\{\lambda_k(R): \lambda_k(R) \geqq \gamma\}}{\pi R^2} = \frac{\alpha}{\pi} + O(R^{-1}),$$

and so for $n(R, \zeta)$ of Lemma 4.2 we have

$$\frac{n(R,\zeta)}{\pi R^2} \geqq \frac{\alpha}{\pi} + O(R^{-1}),$$

proving the first part of the theorem. The second part follows in the same way by applying Lemma 4.3. □

Remark 1. We may ask if the inequalities above can be substituted by strict inequalities.

Remark 2. The above result can be viewed as a generalization of what is known for regular lattices [9].

The situation is more delicate in the Bergman case. From the fact that

$$\frac{\#\{\lambda_k(R): \delta \leqq \lambda_k(R) < \gamma\}}{|\Delta_0(R)|} = O(1) \tag{27}$$

(see [11, II], $|\cdot|$ here meaning hyperbolic area) we realize that an attempt to copy the proof of Theorem 4.4 will fail. In view of (27) it is clear that Lemma 4.2 is too weak in this case, and we may ask what can be deduced from Lemma 4.1. At least to indicate the complexity of this problem we construct a special family of point sets below, to which Lemma 4.1 applies.

Note, however, that by Lemma 4.2 and (33) we can conclude that the number of points in any disk D must be $O(|D|)$.

A family of point sets in the Bergman case. In this paragraph we will construct a point set depending on a parameter M which is to represent a characteristic distance between the points. The purpose is to show that this set cannot be one of sampling for M larger than a certain critical bound.

The basic idea of the construction is to place the points regularly in distance M at concentric circles of radii M, $2M$, $3M, \cdots$. We choose $\zeta = 0$ as the center and number the circles accordingly $1, 2, 3, \cdots$ (the choice of the origin as the center is for convenience only, the construction works equally well with any point $\zeta \in \Delta$ playing this role). In Fig. 1 we have drawn the points at the four first circles for $M = 1.1$ with the outer circle indicating the unit circle. How to determine the number of points at each circle is explained below.

Of course, the distance between the points cannot in general be equal to M. We show how to approximate the number of points at circle k. For this purpose, we recall the cosine rule of hyperbolic geometry [3, p. 148] (see Fig. 2)

$$\cosh c = \cosh a \cosh b - \sinh a \sinh b \cos \alpha.$$

In our case we get

$$\cos \alpha_k = \frac{\cosh^2 kM - \cosh M}{\sinh^2 kM} = 1 - \frac{\cosh M - 1}{\sinh^2 kM},$$

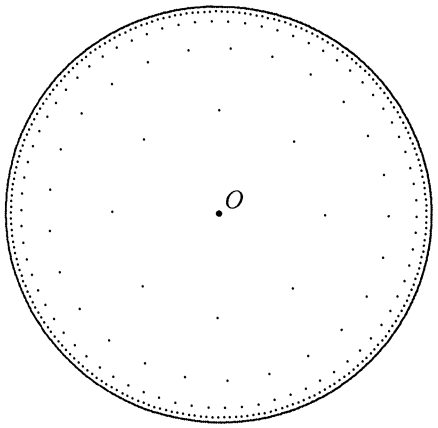


FIG. 1. *An example of a point set in the Bergman case.*

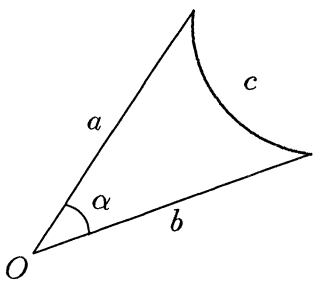


FIG. 2. *A hyperbolic triangle.*

α_k denoting the angle between the lines passing through two adjacent points at circle k and the origin. Since $\alpha_k \to 0$ as $k \to \infty$ we put

$$\alpha_k = \frac{\sqrt{2(\cosh M - 1)}}{\sinh kM}.$$

The number of points at circle k should therefore be approximately $2\pi/\alpha_k$. We define

$$n(k) = [C(M)\, e^{kM}], \tag{28}$$

where

$$C(M) = \frac{\pi}{\sqrt{2(\cosh M - 1)}},$$

and where we have defined $[x]$ to be the integer such that $x = [x] + \delta$ with $-\frac{1}{2} < \delta \leqq \frac{1}{2}$.

For computational ease we shall deviate a bit from the construction above. Let circle k have radius $(k+\varepsilon_k)M$, where ε_k is the solution of the equation

$$\left(\frac{1 - e^{-(k+\varepsilon_k)M}}{1 + e^{-(k+\varepsilon_k)M}}\right)^{[C(M)e^{kM}]} = e^{-2C(M)}.$$

It is easily seen that $\varepsilon_k \to 0$ as $k \to \infty$.

The introduction of the sequence $\{\varepsilon_k\}$ makes the polynomials associated with Lemma 4.1 comfortably simple. For a sufficiently large m, the polynomial $p_m(z)$ corresponding to $\zeta = 0$ and $r = mM$ takes the form

$$p_m(z) = \prod_{k=1}^{m} (z^{n(k)} - e^{-2C(M) + i\beta_k})$$

with $0 \leqq \beta_k < 2\pi$, depending on where we have chosen to arrange the points on each circle.

We next perform an ordering of the terms in this polynomial determined by the order of the exponents. We have

$$\sum_{j=1}^{l} n(j) = C(M)\, e^{M} \frac{e^{lM} - 1}{e^{M} - 1} + O(l) = O(n(l)). \tag{29}$$

Hence we see that the largest $n(k_i)$ appearing in the exponent of a term of the polynomial decides the order of this exponent. We add the coefficients (or rather the squares of the absolute values of the coefficients) of the terms of which $n(k)$ appears as the largest number in the exponent

$$\sum_{n(k)\,\text{largest}} |\alpha_l|^2 = \sum_{j=0}^{k-1} \binom{k-1}{j} e^{-4C(M)(m-j-1)} = e^{-4C(M)(m-1)}(1 + e^{4C(M)})^{k-1}. \tag{30}$$

For future use we note that the total sum of the coefficients is

$$\sum |\alpha_i|^2 = \sum_{k=1}^{m} \left(\sum_{n(k)\,\text{largest}} |\alpha_l|^2 \right) = e^{-4C(M)m}((1 + e^{4C(M)})^m - 1). \tag{31}$$

The scaling factor (cf. Lemma 4.1) of a term with exponent n can be approximated in the following way ($\theta = 2\alpha + 1$):

$$\begin{aligned} \frac{1 \cdot 2 \cdots n}{(\theta+1)\cdot(\theta+2)\cdots(\theta+n)} &= \frac{1 \cdot 2 \cdots n \cdot n^{\theta}}{(\theta+1)\cdot(\theta+2)\cdots(\theta+n)} \cdot n^{-\theta} \\ &= (\Gamma(\theta+1) + e_n) n^{-\theta}, \end{aligned} \tag{32}$$

with $e_n \geqq 0$ and $e_n = O(n^{-1})$. Thus we have for terms with exponent of order $n(k)$

$$\sum_{n(k)\,\text{largest}} |\tilde{\alpha}_l|^2 = O(\exp(-4C(M)m + k(\ln(1+e^{4C(M)}) - M\theta))).$$

This indicates that if our point set is one of sampling, we have

$$M < \theta^{-1} \ln(1+e^{4C(M)})$$

(otherwise the "larger" eigenvalues would dominate the left-hand side of (21) which then could not be expected to be bounded away from 1). We proceed to show that this is indeed the case.

We have

$$\lambda_n(r) = \frac{c_n \pi}{n!} \int_0^{\tanh^2 r/2} t^n (1-t)^{\theta-1}\, dt = 1 - \frac{c_n \pi}{n!} \int_{\tanh^2 r/2}^1 t^n (1-t)^{\theta-1}\, dt \tag{33}$$

$$\geqq 1 - \frac{c_n \pi}{\theta n!} \left(1 - \tanh^2 \frac{1}{2} r\right)^{\theta}.$$

From this we find

$$\frac{\sum |\tilde{\alpha}_i|^2 \lambda_i}{\sum |\tilde{\alpha}_i|^2} \geqq 1 - \cosh^{-2\theta}\left(\frac{1}{2} r\right) \frac{\sum |\alpha_i|^2}{\sum |\tilde{\alpha}_i|^2}. \tag{34}$$

From (30) and (32) we find

$$\sum_{n(k)\,\text{largest}} |\tilde{\alpha}_l|^2 \geqq A(M, \theta) \exp(-4C(M)m + (k-1)(\ln(1+e^{4C(M)}) - M\theta)), \tag{35}$$

where we can choose

$$A(M, \theta) = \Gamma(\theta+1) \left(\frac{C(M)\, e^M}{e^M - 1} + \frac{1}{2M}\right)^{-\theta} e^{4C(M) - M\theta}.$$

Here we have used the following estimate for the exponents l in (35):

$$l \leqq \sum_{j=1}^{k} n(j) \leqq C(M)\, e^M \frac{e^{kM} - 1}{e^M - 1} + \frac{1}{2} k \leqq \left(\frac{C(M)\, e^M}{e^M - 1} + \frac{1}{2M}\right) e^{kM}.$$

Thus we have

$$\sum |\tilde{\alpha}_i|^2 \geqq A(M, \theta)\, e^{-4C(M)m} \sum_{k=0}^{m-1} \exp(k(\ln(1+e^{4C(M)}) - M\theta)),$$

in other words,

$$\sum |\tilde{\alpha}_i|^2 \geqq \begin{cases} A'(M, \theta)\, e^{-4C(M)m}(1 - \exp(m(\ln(1+e^{4C(M)}) - M\theta))), & \\ & \ln(1+e^{4C(M)}) \neq M\theta, \\ A(M, \theta)\, e^{-4C(M)m} m, & \ln(1+e^{4C(M)}) = M\theta, \end{cases} \tag{36}$$

where

$$A'(M, \theta) = A(M, \theta)(1 - \exp(\ln(1+e^{4C(M)}) - M\theta))^{-1}.$$

The radius corresponding to $p_m(z)$ is mM, and consequently

$$\cosh^{-2\theta}(\tfrac{1}{2}r) \leqq 4^{\theta}\, e^{-mM\theta}.$$

We are now ready to draw the desired conclusion. For the case $\ln(1+e^{4C(M)}) < M\theta$ we have from (34) and (36) (with $B(M,\theta) = 4^{\theta}(A'(M,\theta))^{-1}$)

$$\frac{\sum |\tilde{\alpha}_i|^2 \lambda_i}{\sum |\tilde{\alpha}_i|^2} \geqq 1 - B(M,\theta) \frac{\exp(m(\ln(1+e^{4C(M)}) - M\theta)) - e^{-mM\theta}}{1 - \exp(m(\ln(1+e^{4C(M)}) - M\theta))} \to 1$$

as $m \to \infty$, contradicting Lemma 4.1.

For the case $M\theta = \ln(1+e^{4C(M)})$ we find from (34) and (36) (with $B(M,\theta) = 4^{\theta}(A(M,\theta))^{-1}$)

$$\frac{\sum |\tilde{\alpha}_i|^2 \lambda_i}{\sum |\tilde{\alpha}_i|^2} \geqq 1 - B(M,\theta) \frac{1 - e^{-mM\theta}}{m} \to 1$$

as $m \to \infty$, again contradicting Lemma 4.1.

We conclude that if a point set as constructed above is a set of sampling then we have $M/\ln(1+e^{4C(M)}) < (2\alpha+1)^{-1}$.

We remark that in a similar manner it can be proved that inequality (21) with $\zeta = 0$ is indeed satisfied when $M\theta < \ln(1+e^{4C(M)})$. This indicates that what has been found above is the sharpest result that can be deduced from Lemma 4.1.

We now prepare for an interesting observation on the asymptotics of the above inequality. First, we have

$$C(M) = \frac{2\pi}{\sinh(M/2)},$$

yielding for small M

$$M/\ln(1+e^{4C(M)}) = \frac{M^2}{4\pi} + O(M^3). \tag{37}$$

Next, as pointed out in the introduction of [23] the Bargmann case can formally be considered as the limiting case $\alpha \to \infty$. To see this we look at the scaled disk $R\Delta$, put $2\alpha = \tilde{\alpha}R^2$, let $R \to \infty$ and make use of the fact that $(1-|z|^2/R^2)^{\tilde{\alpha}R^2} \to e^{-\tilde{\alpha}|z|^2}$. Finally, observe that close to zero we have, loosely speaking, a Euclidean distance E corresponding to M with $M = 2E$.

We bring these remarks together, rewrite the necessary condition on M as

$$\frac{1}{(ER)^2} > \frac{\tilde{\alpha}}{\pi} + O(E),$$

and observe that formally we recapture the first part of Theorem 4.4.

The interest in the above calculation is not the fact that M cannot be too large since this is well known (for large M there are nonzero functions vanishing on these sets, see [20], [35]). What is interesting is the exact bound, the significance of which is indicated by our formal passage to the Nyquist density.

5. Jittered sampling. The purpose of this section is to show that the family of sets of sampling is in a sense an open set. This we will do by deducing estimates for the frame bounds for jitters of actual sets of sampling.

We should remark that there are available existence results of the following kind (incompletely quoted) [7], [37], [23], [31]. There exists a number $L_0 > 0$ such that if $\{z_n\}$ is any L-lattice (to be defined below) and $L < L_0$, then $\{z_n\}$ is a set of sampling. But we would like to know what is the value of L_0 and how we could estimate the frame bounds.

We shall borrow an idea from the theory of nonharmonic Fourier series [15]. It was used in the proof of one of the forerunners to Kadec's $\frac{1}{4}$-theorem [24]. In [15] it is shown that $\{e^{it_n\xi}\}$ is a Riesz basis for $L^2[-\pi, \pi]$ if $\sup_n |t_n - n| < \log 2/\pi \approx 0.22$ (by [24] the best possible constant is $\frac{1}{4}$). We combine this idea with the local analysis provided by our eigenfunction expansions.

We consider again uniformly discrete sets $\{z_n\}$ of complex numbers. Let us measure the distance between two such sets by

$$\rho(\{z_n\}, \{\zeta_n\}) = \sup_n d(z_n, \zeta_n).$$

We say that $\{\zeta_n\}$ is an *M-jitter* of $\{z_n\}$ if $\rho(\{z_n\}, \{\zeta_n\}) = M < \infty$.

Suppose $\{z_n\}$ has separating distance $2\delta_0 > 0$ and let $\{\zeta_n\}$ be an M-jitter of $\{z_n\}$ with $M < \delta_0$. In the Bargmann case define

$$D(M, \delta_0, A) = \frac{e^{\alpha\delta_0^2}}{(\alpha\delta_0)^2}(1 - A\lambda_0(\delta_0))\left(\left(1 - \frac{M^2}{\delta_0^2}\right)^{-2} - 1\right) \tag{38}$$

and in the Bergman case

$$D(M, \delta_0, A) = \frac{\min\{1, \cosh^{-4\alpha}(\delta_0/2)\}}{(2\alpha+1)\tanh^2(\delta_0/2)}(1 - A\lambda_0(\delta_0))\left(\left(1 - \frac{\tanh^2(M/2)}{\tanh^2(\delta_0/2)}\right)^{-2} - 1\right). \tag{39}$$

With this notation we have the following theorem.

THEOREM 5.1. *Suppose that $\{z_n\}$ with separating distance $2\delta_0$ is a set of sampling with frame bounds A and B. Let M_0 $(< \delta_0)$ be the positive number such that $D(M_0, \delta_0, A) = A$. Then any M-jitter $\{\zeta_n\}$ of $\{z_n\}$ with $M < M_0$ is a set of sampling, and for its frame bounds $\tilde{A}$ and $\tilde{B}$ we have*

$$\tilde{A} \geqq \begin{cases} e^{-\alpha M^2}(\sqrt{A} - \sqrt{D(M, \delta_0, A)})^2 & \text{Bargmann case,} \\ \cosh^{-4\alpha+4}(\tfrac{1}{2}M)(\sqrt{A} - \sqrt{D(M, \delta_0, A)})^2 & \text{Bergman case,} \end{cases} \tag{40}$$

$$\tilde{B} \leqq (\sqrt{B} + \sqrt{D(M, \delta_0, A)})^2. \tag{41}$$

Proof. Let $\{\zeta_n\}$ be an M-jitter of $\{z_n\}$ and f any function in the space at hand. The key to the proof is the following estimate:

$$\sum \left| \frac{R(z_n, z_n)^{1/2} f(\zeta_n)}{R(\zeta_n, z_n)} - \frac{f(z_n)}{R(z_n, z_n)^{1/2}} \right|^2 \leqq D(M, \delta_0, A)\|f\|^2. \tag{42}$$

Let us first show that (42) proves the theorem.

In the Bargmann case we have the identity

$$\frac{R(z, z)^{1/2}}{|R(\zeta, z)} = \frac{\exp(\frac{1}{2}\alpha|z - \zeta|^2)}{R(\zeta, \zeta)^{1/2}},$$

yielding

$$\frac{1}{R(\zeta_n, \zeta_n)^{1/2}} \leqq \frac{R(z_n, z_n)^{1/2}}{|R(\zeta_n, z_n)|} \leqq \frac{\exp(\frac{1}{2}\alpha M^2)}{R(\zeta_n, \zeta_n)^{1/2}}. \tag{43}$$

In the Bergman case we use the identity (see [3, p. 132])

$$\frac{R(z, z)^{1/2}}{|R(\zeta, z)|} = \frac{\cosh^{2\alpha+2}\frac{1}{2}d(z, \zeta)}{R(\zeta, \zeta)^{1/2}},$$

which leads to

$$\frac{1}{R(\zeta_n, \zeta_n)^{1/2}} \leqq \frac{R(z_n, z_n)^{1/2}}{|R(\zeta_n, z_n)|} \leqq \frac{\cosh^{2\alpha+2}\frac{1}{2}M}{R(\zeta_n, \zeta_n)^{1/2}}. \tag{44}$$

By (43) or (44) and Minkowski's inqeuality, (42) implies

$$\left(\sum \frac{|f(\zeta_n)|^2}{R(\zeta_n, \zeta_n)}\right)^{1/2} \leqq \sqrt{D(M, \delta_0, A)}\|f\| + \left(\sum \frac{|f(z_n)|^2}{R(z_n, z_n)}\right)^{1/2}.$$

By the definition of B we arrive at (41). Similarly, by (43) or (44) and Minkowski's inequality we see that (42) implies (40).

We next turn to the proof of (42). To this end we expand $f(z)$ into the eigenfunctions associated with z_n,

$$f(z) = \sum_{k=0}^{\infty} (f, f_k^{z_n}) f_k^{z_n}(z). \tag{45}$$

By the double orthogonality and the fact that $f_0^{z_n}$ acts reproducingly this gives

$$\int_{\Delta_{z_n}(\delta_0)} |f(z)|^2 \, d\omega(z) = \lambda_0(\delta_0) \frac{|f(z_n)|^2}{R(z_n, z_n)} + \sum_{k=1}^{\infty} \lambda_k(\delta_0) |(f, f_k^{z_n})|^2.$$

On summing over n and using the definition of A, we get

$$\sum_n \sum_{k=1}^{\infty} \lambda_k(\delta_0) |(f, f_k^{z_n})|^2 \leqq (1 - A\lambda_0(\delta_0)) \|f\|^2. \tag{46}$$

Let us now put $z = \zeta_n$ in (45). This yields

$$f(\zeta_n) = \frac{R(\zeta_n, z_n)}{R(z_n, z_n)} f(z_n) + R(\zeta_n, z_n) \sum_{k=1}^{\infty} \rho_k^{z_n} (f, f_k^{z_n}) (T_{z_n} \zeta_n)^k. \tag{47}$$

At this stage we split the discussion and treat the two cases separately.

In the Bargmann case (47) implies

$$\frac{R(z_n, z_n)^{1/2} f(\zeta_n)}{R(\zeta_n, z_n)} - \frac{f(z_n)}{R(z_n, z_n)^{1/2}} = \sum_{k=1}^{\infty} \sqrt{\frac{\alpha^{k-1}}{k!}} (f, f_k^{z_n}) (T_{z_n} \zeta_n)^k.$$

To estimate this sum we multiply and divide the kth term by $\delta_0^k / \sqrt{k+1}$ $(k = 1, 2, \cdots)$. By Hölder's inequality we then get

$$\left| \frac{R(z_n, z_n)^{1/2} f(\zeta_n)}{R(\zeta_n, z_n)} - \frac{f(z_n)}{R(z_n, z_n)^{1/2}} \right|^2 \leqq \sum_{k=1}^{\infty} \frac{\alpha^{k-1} \delta_0^{2k}}{k!(k+1)} |(f, f_k^{z_n})|^2 \sum_{j=1}^{\infty} (j+1) \frac{|z_n - \zeta_n|^{2j}}{\delta_0^{2j}}. \tag{48}$$

By the identity

$$\sum_{j=1}^{\infty} (j+1) x^j = \frac{1}{1-x} - 1$$

the last sum in (48) is found to be bounded by $(1 - M^2/\delta_0^2)^{-2} - 1$.

We next sum (48) over n. To apply (46) we use the estimate

$$\lambda_k(\delta_0) = \frac{1}{k!} \int_0^{\alpha\delta_0^2} t^k e^{-t} \, dt \geqq \frac{e^{-\alpha\delta_0^2} (\alpha\delta_0^2)^{k+1}}{k!(k+1)},$$

and we find we have arrived at (42).

The argument is exactly the same in the Bergman case. Here (47) implies

$$\frac{R(z_n, z_n)^{1/2} f(\zeta_n)}{R(\zeta_n, z_n)} - \frac{f(z_n)}{R(z_n, z_n)^{1/2}} = \sum_{k=1}^{\infty} \sqrt{\frac{c_k \pi}{k!(2\alpha+1)}} (f, f_k^{z_n}) (T_{z_n} \zeta_n)^k.$$

We use the estimate

$$\frac{k!(2\alpha+1)\lambda_k(\delta_0)}{c_k\pi} = (2\alpha+1)\int_0^{\tanh^2(\delta_0/2)} t^k(1-t)^2\,dt$$

$$\geqq (2\alpha+1)\min\left\{1, \cosh^{-4\alpha}\frac{\delta_0}{2}\right\}\frac{\tanh^{2(k+1)}(\delta_0/2)}{k+1}$$

and the identity

$$|T_z\zeta| = \tanh\left(\tfrac{1}{2}d(z,\zeta)\right)$$

(see [3, p. 132]) and obtain (42) in the same way as above. □

A slight modificaiton of the argument above shows that the frame bounds depend continuously on the jitters.

Let us now say that a uniformly discrete set $\{\zeta_n\}$ is an *L-lattice* if to any point z (in $\mathbf{C}$ or in Δ) there exists an n such that $d(z,\zeta_n)\leqq L$. With this notation we note the following, which gives us a numerical estimate for the number L_0 mentioned in the introduction to this section.

COROLLARY 5.2. *Let*

$$L_0 = \sup_{\{z_n\}}\{L\colon L<\delta_0(\{z_n\}) \text{ and } A(\{z_n\}) = D(L,\delta_0(\{z_n\}), A(\{z_n\}))\},$$

where the supremum is taken over all uniformly discrete sets $\{z_n\}$ *(each with separating distance* $2\delta_0(\{z_n\})$*) and where*

$$A(\{z_n\}) = \inf_{\|f\|\neq 0}\frac{\sum(|f(z_n)|^2/R(z_n,z_n))}{\|f\|^2}.$$

Then any L-lattice with $L<L_0$ *is a set of sampling.*

Proof. Let $\{\zeta_n\}$ be any L-lattice with $L<L_0$. By definition $\{\zeta_n\}$ is uniformly discrete, say with separating distance $2\tilde{\delta}_0$. So by (19) and the Schwarz inequality we have

$$\sum\frac{|f(z_n)|^2}{R(z_n,z_n))}\leqq\frac{1}{\lambda_0(\tilde{\delta}_0)}\|f\|^2.$$

Since $L<L_0$ we can find a set $\{z_k\}$ such that

$$A = A(\{z_k\}) > D(L,\delta_0,A) \tag{49}$$

and $L<\delta_0$ where we have assumed $\{z_k\}$ to have separating distance $2\delta_0$. Since $\{\zeta_n\}$ is an L-lattice we can pick a subsequence $\{\zeta_{n(k)}\}$ such that

$$\rho(\{\zeta_{n(k)}\},\{z_k\})\leqq L.$$

We have

$$\sum_n\frac{|f(\zeta_n)|^2}{R(\zeta_n,\zeta_n)}\geqq\sum_k\frac{|f(\zeta_{n(k)})|^2}{R(\zeta_{n(k)},\zeta_{n(k)})}$$

and since $\{\zeta_{n(k)}\}$ is an M-jitter of $\{z_k\}$ with $M\leqq L$, the existence of a lower frame bound follows from (49) and Theorem 5.1. □

We should remark that the existence of $L_0>0$ and the possibility of estimating it are ensured by the results in [9].

Acknowledgments. The author takes pleasure in thanking Henrik H. Martens for stimulating discussions throughout this work. Among the various valuable comments

of the referees the author is especially grateful for the one that led him to the formal passage from the Bergman to the Bargmann case at the end of § 4.

Note added in proof (Remark to Theorem 4.4.). Recently R. Wallstén and the author have proved that $D^- > \alpha/\pi$ and $D^+ < \alpha/\pi$ are also sufficient in the respective cases, a result that will appear in a forthcoming publication.

REFERENCES

[1] M. BASTIAANS, *A sampling theorem for the complex spectrogramm and Gabor's expansion of a signal in Gaussian elementary signals*, Optical Engrg., 20 (1981), pp. 594–598.

[2] G. BATTLE, *A block spin construction of ondelettes. Part* I: *Lemarié functions*, Comm. Math. Phys., 110 (1987), pp. 601–615.

[3] A. F. BEARDON, *The Geometry of Discrete Groups*, Springer-Verlag, New York, 1983.

[4] S. BERGMAN, *The Kernel Function and Conformal Mapping*, American Mathematical Society, Math. Surveys V, American Mathematical Society, New York, 1950.

[5] L. BERS, *Automorphic forms and Poincaré series for infinitely generated Fuchsian groups*, Amer. J. Math., 87 (1965), pp. 196–214.

[6] L. A. COBURN, *Toeplitz operators, quantum mechanics and mean oscillation in the Bergman metric*, in AMS Proc. Symposium on Pure Math., from AMS SRI on Operator Theory/Operator Albebras and Applications, Durham, NH, 1988.

[7] R. R. COIFMAN AND R. ROCHBERG, *Representation theorems for holomorphic and harmonic functions in L^p*, Astérisque, 77 (1980), pp. 11–66.

[8] J. M. COMBES, A. GROSSMANN, AND PH. TCHAMITCHIAN, EDS., *Wavelets*, in Proc. Colloque Ondelettes, méthodes temps-fréquence et éspace des phases, Marseille, France, December 1987, Springer-Verlag, Berlin, New York, 1989.

[9] I. DAUBECHIES, *The wavelet transform, time-frequency localization and signal analysis*, IEEE Trans. Inform. Theory, 36 (1990), pp. 961–1005.

[10] ———, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[11a] ———, *Time-frequency localization operators—a geometric phase space approach*, I, IEEE Trans. Inform. Theory, 34 (1988), pp. 605–612.

[11b] I. DAUBECHIES AND T. PAUL, *Time-frequency localization operators—a geometric phase space approach*, II. *Inverse problems*, IEEE Trans. Inform. Theory, 4 (1988), pp. 661–680.

[12] I. DAUBECHIES AND A. GROSSMANN, *Frames in the Bargmann space of entire functions*, Comm. Pure Appl. Math., 41 (1988), pp. 151–164.

[13] I. DAUBECHIES, A. GROSSMANN, AND Y. MEYER, *Painless non-orthogonal expansions*, J. Math. Phys., 27 (1986), pp. 1271–1283.

[14] I. DAUBECHIES, S. JAFFARD, AND J. L. JOURNÉ, *A simple Wilson orthonormal basis with exponential decay*, AT&T Bell Laboratories, Holmdel, NJ, 1989, preprint; SIAM J. Math. Anal., 22 (1991), pp. 549–568.

[15] R. J. DUFFIN AND J. J. EACHUS, *Some notes on an expansion theorem of Paley and Wiener*, Bull. Amer. Math. Soc., 48 (1942), pp. 850–855.

[16] R. J. DUFFIN AND A. C. SCHAEFFER, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc., 72 (1952), pp. 341–366.

[17] G. B. FOLLAND, *Harmonic Analysis in Phase Space*, Princeton University Press, Princeton, NJ, 1989.

[18] D. GABOR, *Theory of communication*, J. Inst. Electr. Engrg. (London), 93 (1946), pp. 429–457.

[19] P. GOUPILLAUD, A. GROSSMANN, AND J. MORLET, *Cyclo-octave and related transforms in seismic signal analysis*, Geoexploration, 23 (1984), pp. 85–102.

[20] C. HOROWITZ, *Zeros of functions in the Bergman spaces*, Duke Math. J., 41 (1977), pp. 693–710.

[21] A. GROSSMANN AND J. MORLET, *Decomposition of Hardy functions into square integrable wavelets of constant shape*, SIAM J. Math. Anal., 15 (1984), pp. 723–736.

[22a] A. GROSSMANN, J. MORLET, AND T. PAUL, *Transforms associated to square integrable group representations.* I, J. Math. Phys., 26 (1985), pp. 2473–2479.

[22b] ———, *Transforms associated to square integrable group representations.* II, Ann. Inst. H. Poincaré Phys. Théor., 45 (1986), pp. 293–309.

[23] S. JANSON, J. PEETRE, AND R. ROCHBERG, *Hankel forms and the Fock space*, Rev. Mat. Iberoamer., 3 (1987), pp. 61–138.

[24] M. I. KADEC, *The exact value of the Paley-Wiener constant*, Dokl. Akad. Nauk. SSSR, 155 (1964), pp. 1253-1254.
[25] J. R. KLAUDER AND B. S. SKAGERSTAM, *Coherent States*, World Scientific, Singapore, 1985.
[26] I. KRA, *Automorphic Forms and Kleinian Groups*, W. A. Benjamin, Reading, 1972.
[27] H. J. LANDAU, *Sampling, data transmission, and the Nyquist rate*, Proc. IEEE, (1967), pp. 1701-1706.
[28] H. J. LANDAU, *Necessary density conditions for sampling and interpolation of certain entire function spaces*, Acta Math., 117 (1967), pp. 37-52.
[29] J. LEHNER, *Discrete Groups and Automorphic Functions*, W. J. Harvey, ed., Academic Press, New York, 1977, pp. 73-120.
[30] P. G. LEMARIÉ, *Une nouvelle construction d'ondelettes de* $L^2(\mathbf{R}^n)$, J. Math. Pures Appl., to appear.
[31] D. LUECKING, *Closed range restriction operators on weighted Bergman spaces*, Pacific J. Math., 110 (1984), pp. 145-160.
[32] Y. MEYER, *Principe d'incertitude, bases hilbertienne et algèbres d'opérateurs*, Séminaire Bourbaki 38 ième année, 1985-1986, n° 662.
[33] T. PAUL, *Functions analytic in a half-plane as quantum mechanical states*, J. Math. Phys., 25 (1984), pp. 3252-3263.
[34] ———, *Affine coherent states and the radial Schrödinger equation*, preprint, Centre de Physique Théorique, Luminy, Marseille, France, 1984.
[35] R. ROCHBERG, *Interpolation by functions in Bergman spaces*, Mich. Math. J., 29 (1982), pp. 229-236.
[36] ———, *Toeplitz and Hankel operators, wavelets, NWO sequences, and almost diagonalization of operators*, in AMS Proc. Symposium on Pure Math., from AMS SRI on Operator Theory/Operator Algebras and Applications, Durham, NH, 1988.
[37] ———, *Decomposition theorems for Bergman spaces and their applications*, in Operators and Function Theory, S. C. Power, ed., NATO ASI Series, D. Reidel, Dordrecht, 1985, pp. 225-277.
[38] W. RUDIN, *Function Theory in the Unit Ball of* $\mathbf{C}^n$, Springer-Verlag, New York, 1980.
[39] K. SEIP, *Mean value theorems and concentration operators in Bargmann and Bergman spaces*, in Wavelets, J. M. Combes, A. Grossmann, and Ph. Tchamitchian, eds., Springer-Verlag, Berlin, New York, 1989, pp. 209-215.
[40a] D. SLEPIAN AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*, I, Bell Systems Tech. J., 40 (1961), pp. 43-64.
[40b] H. J. LANDAU AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*, II, Bell Systems Tech. J., 40 (1961), pp. 65-84.
[40c] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*, III, Bell Systems Tech. J., 40 (1962), pp. 1295-1336.
[41] J. O. STRÖMBERG, *A modified Franklin system and higher-order spline systems on* $\mathbf{R}^n$ *as unconditional bases for Hardy spaces*, in Conference on Harmonic Analysis in Honor of Antoni Zygmund II, W. Beckner, A. P. Calderón, R. Fefferman and P. W. Jones, eds., Wadsworth Mathematics Series, Wadsworth, Belmont, CA, 1981, pp. 475-493.
[42] R. M. YOUNG, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, p. 877, May 1991

020

ERRATA:
Sur une classe de fonctionnelles non convexes et applications*

RABAH TAHRAOUI†

Page 45, line 15: Il faut lire *rdr* à la place de *dr.*

Page 50: Il manque une partie de l'hypothèse (2), i.e., rajouter

$$\left|\frac{\partial h_3}{\partial \eta_i}(x, \eta)\right| \leqq c|\eta|^{p-1} + d.$$

Page 51, line 18: Il faut lire $p_i \in W^{2m,p'}_{\text{loc}}(\Omega)$ au lieu de $p_i \in W^{2m,p}_{(\Omega)}$, p' étant le conjugué de p.

REFERENCE

[1] R. TAHRAOUI, *Sur une classe de fonctionnelles non convexes et applications*, SIAM J. Math. Anal., 21 (1990), pp. 37-52.

* Received by the editors and accepted for publication November 8, 1990.

† Université de Paris-Sud et Centre, Bâtiment 425, Nationale de Recherche Scientifique, Laboratoire d'Analyse Numérique d'Orsay, Orsay 91405, France, and Université de Picardie, U.F.R. Cedex Mathematique et Informatique, Amiens, France.

SIAM J. MATH. ANAL.
Vol. 22, No. 3, p. 878, May 1991

021

ERRATUM:
A Simple Wilson Orthonormal Basis with Exponential Decay*

INGRID DAUBECHIES†, STÉPHANE JAFFARD‡, AND JEAN-LIN JOURNÉ§

The following equation was printed incorrectly in the March issue. The correct equation should read

$$\begin{aligned}
&\sum_{m=1}^{\infty} \hat{f}_m(\xi)\overline{\hat{f}_m(\xi+k)} \\
&\quad = \phi(\xi)\phi(\xi+k)+\frac{1}{2}\sum_{\ell=1}^{\infty}\sum_{\kappa=0}^{1}[\phi(\xi-\ell)+(-1)^{\ell+\kappa}\phi(\xi+\ell)] \\
&\quad\quad \cdot[\phi(\xi-\ell+k)+(-1)^{\ell+\kappa}\phi(\xi+\ell+k)]\, e^{-i\pi\kappa k} \qquad (2.4)\\
&\quad = \phi(\xi)\phi(\xi+k)+\sum_{\ell\in\mathbb{Z},\ell\neq 0}\phi(\xi+\ell)\phi(\xi+\ell+k)\frac{1}{2}(1+(-1)^k) \\
&\quad\quad +\sum_{\ell\in\mathbb{Z},\ell\neq 0}(-1)^{\ell}\phi(\xi+\ell)\phi(\xi-\ell+k)\frac{1}{2}(1-(-1)^k).
\end{aligned}$$

SIAM regrets the error.

REFERENCE

I. DAUBECHIES, S. JAFFARD, AND J.-L. JOURNÉ, *A simple Wilson orthonormal basis with exponential decay*, SIAM J. Math. Anal., 22 (1991), pp. 554–573.

* Received by the editors March 20, 1991; accepted for publication March 20, 1991.

† AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974. This author is "Bevoegdverklaand Navonsen" at the Belgian National Science Foundation (on leave); also on leave from the Department of Theoretical Physics, Vrije Universiteit, Brussels, Belgium.

‡ Centre d'Etude et de Recherche en Mathématique Appliquée, Ecole des Ponts et Chaussées, La Courtine, 93167 Noisy-le-Grand, France.

§ Princeton University, Princeton, New Jersey 08544. This author's work was supported by the National Science Foundation.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 879-895, July 1991

LONGTIME BEHAVIOUR OF STRONGLY DAMPED WAVE EQUATIONS, GLOBAL ATTRACTORS AND THEIR DIMENSION*

J. M. GHIDAGLIA† AND A. MARZOCCHI‡

Abstract. This paper contains some results on asymptotic behaviour of solutions to strongly damped abstract nonlinear wave equations. After reviewing sufficient hypotheses for existence and uniqueness, uniform time estimates are given and a global attractor for the trajectories of the associated dynamical system is constructed. Finally, applications are made to nonlinear wave equations such as Sine-Gordon equation, proving the finite dimensionality of the corresponding attractors.

Key words. nonlinear wave equations, attractors, asymptotic behaviour

AMS(MOS) subject classifications. 35L70, 35B40

Introduction. In this paper we study the longtime behaviour of strongly damped wave equations. We address the case where an external excitation drives the solutions and where nontrivial attractors occur. One of our main results concerns the dimension of these sets that we show to be finite. Let us begin by mentioning two applications that have motivated this work. Firstly, the perturbed Sine-Gordon equation

$$\frac{\partial^2 u}{\partial t^2} - \Delta u + \sin u = -\beta \frac{\partial u}{\partial t} + \alpha \frac{\partial(\Delta u)}{\partial t} + f, \tag{0.1}$$

where $u(x, t)$ is the current in a Josephson junction (see, e.g., [1]), x is the space variable and (0.1) is posed in a bounded domain Ω in $\mathbb{R}^n$ (with appropriate boundary conditions). The parameters α and β are nonnegative and correspond to loss effects. We are concerned in this work with the case where $\alpha > 0$ (refer to [2], [3] for the case $\alpha = 0$). The function $f, f(x, t)$, is time-periodic and figures the external current that drives the device. Another example reads

$$\frac{\partial^2 u}{\partial t^2} - \Delta u + |u|^q u = -\beta \left|\frac{\partial u}{\partial t}\right|^p \frac{\partial u}{\partial t} + \alpha \frac{\partial(\Delta u)}{\partial t} + f, \tag{0.2}$$

which is also a perturbed wave equation occurring in quantum mechanics.

These partial differential equations can be seen as infinite-dimensional dynamical systems, and the questions we address in this paper are related to determining whether or not these systems depend on a finite number of degrees of freedom after a transient period. A mathematical approach to this type of problem has been introduced for dissipative parabolic equations, motivated by the study of turbulence in fluids. It was proved that a global attractor exists and captures all the solutions as time goes to infinity, and that this set was finite dimensional. A bound on this dimension provides an estimate on the number of degrees of freedom in the longtime behaviour.

This point of view has been developed for many other types of partial differential equations such as (weakly) damped wave equations, coupled systems, etc. (see Temam [5] for a recent review), and more recently for nonlinear dispersive equations ([6], [7]).

* Received by the editors March 30, 1990; accepted for publication August 28, 1990.

† Centre de Mathématiques et Leurs Applications, Ecole Normale Supèrieure, Cachan, 94235 Cachan Cedex, France.

‡ Dipartimento di Matematica, Università Cattolica del Sacro Cuore, Via Trieste 17, 25121 Brescia, Italia.

In this article, we shall consider an abstract evolution equation which can be written as

$$\frac{d^2u}{dt^2}+\alpha A\frac{du}{dt}+Au+g(u)+h(u_t)=f, \tag{0.3}$$

where A is a linear positive operator (see below) and show, under several hypotheses, that its longtime behaviour is also described by a global attractor. Results in this direction were derived by Massatt [8] and Webb [9] under quite restrictive hypotheses on the nonlinear terms (g and h). Our techniques are completely different from those in [8] and [9] and we are able, for example, to handle nongradient systems or periodically driven equations. Then, in § 4 we give sufficient conditions on the nonlinear terms g and h that insure the finite dimension of this attractor.

1. Review and complements on the Cauchy problem.

1.1. Functional setting and the linear semigroup. Let H be a real Hilbert space and A a positive linear self-adjoint unbounded operator with domain $D(A)=\{v\in H\,|\,Av\in H\}$ which is dense in H. The space $D(A)$ is a Banach space when endowed with the graph norm $v\to\|v\|_H^2+\|Av\|_H^2$. We will suppose furthermore that $A: D(A)\to H$ is an isomorphism and that A^{-1} is a compact operator. Under these hypotheses it is possible to define the power A^s of $A(s\in\mathbb{R})$ and to see that the spaces

$$V_{2s}=D(A^s)\qquad(s\in\mathbb{R})$$

are Hilbert spaces with scalar products and norms:

$$(u,v)_{2s}=(A^su,A^sv)_H,$$
$$\|u\|_{V_{2s}}=(u,u)_{2s}^{1/2},$$

respectively.

For the sake of notational simplicity we will write throughout the paper

$$(\cdot,\cdot)\quad\text{for }(\cdot,\cdot)_H,$$
$$|\cdot|_{2s}\quad\text{for }\|\cdot\|_{V_{2s}},$$
$$|\cdot|\quad\text{for }|\cdot|_0,$$
$$\|\cdot\|\quad\text{for }|\cdot|_1.$$

Furthermore, we identify H with its dual, which leads to identify V_{-s} with the dual of V_s, and denote also by $(\cdot,\cdot)$ the duality pairing between these spaces.

Let $I\subset\mathbb{R}$ be an interval and X a Banach space with norm $\|\cdot\|_X$. Let $p\in[1,\infty]$ and

$$L^p(I;X)=\{f\,|\,f: I\to X\text{ such that }\|f\|_X(t)\in L^p(I)\}$$

be the space of all X-value functions and L^p-integrable on I. These spaces are Banach spaces when endowed with the norms

$$\|f\|_{L^p(I;X)}=\left[\int_I(\|f\|_X(t))^p\,dt\right]^{1/p}\quad\text{if }1\leqq p<\infty,$$
$$\|f\|_{L^\infty(I;X)}=\operatorname*{ess\,sup}_{t\in I}\{\|f\|_X(t)\}\quad\text{if }p=\infty.$$

Moreover, let $C(I;X)$ be the space of all continuous functions from I into X and $C_b(I;X)=C(I;X)\cap L^\infty(I;X)$ the subspace of the bounded ones.

When inserted in the abstract framework introduced above, the strongly damped wave equation we will consider generalizes as (u_t denote the time derivative)

$$\begin{aligned} &u_{tt}+\alpha Au_t+Au+g(u)+h(u_t)=f(t),\\ &u(0)=u_0, \qquad u_t(0)=u_1, \end{aligned} \tag{1.1}$$

where $\alpha\in\mathbb{R}^+$, f is a given function $I\to H$, and g, h represent given nonlinear mappings whose properties are to be specified later.

As it is well known (see, e.g., [5]) in the linear case, i.e., the case where f, g, and h vanish, (1.1) defines a linear semigroup on $V_1\times H$ which we denote by $\Sigma(t)$:

$$\Sigma(t):(u_0,u_1)\to(u(t),u_t(t)), \qquad t\geqq 0.$$

When $\alpha=0$ we recover the usual wave propagator (which is a unitary group) while when $\alpha>0$ (the dissipative case), the $\Sigma(t)$ are no longer invertible and the trajectory tends exponentially to zero as $t\to+\infty$, as we will see in Proposition 1.1.

1.2. The nonlinear semigroup. In this section we give sufficient conditions that ensure the existence and uniqueness of solutions to (1.1). Although it seems well known that the hypotheses below lead to the desired results, we have not been able to locate them in the literature (see, e.g., [8], [9] for related conditions). The results rely on the classical energy method, i.e., on some a priori estimates on the solutions of (1.1). We consider a smooth solution of (1.1), and assume that the nonlinear interaction term can be split as

$$g(v)=G'(v)+p(v), \tag{1.2}$$

where $G\in C^1(V,\mathbb{R})$ satisfies $G(0)=0$ and $p\in C^0(V, V_{-1})$. The scalar product of (1.1) with u_t in H leads to the "energy relation"

$$\frac{1}{2}\frac{d}{dt}(|u_t|^2+\|u\|^2+2G(u))+\alpha\|u_t\|^2+(h(u_t),u_t)+(p(u),u_t)=(f,u_t). \tag{1.3}$$

In order to deduce from this relation some estimates on the solutions we make the following hypotheses:

$$\text{(H1)} \qquad \liminf_{\|v\|\to\infty}\frac{G(v)}{\|v\|^2}\geqq 0;$$

$$\text{(H2)} \qquad \exists C_1>0,\quad \sigma_1>0 \quad \text{such that} \quad |p(v)|_{-1}^2\leqq C_1(1+|G(v)|)^{1-\sigma_1} \quad \forall v\in V;$$

$$\text{(H3)} \qquad \exists C_2>0,\quad \sigma_2>0 \quad \text{such that} \quad (h(v),v)\geqq -C_2(1+\|v\|^2)^{1-\sigma_2} \quad \forall v\in V.$$

Assuming that the external force term f satisfies $f\in L^2(0,T;V_{-1})$ for all $T>0$, it is straightforward to derive from the above hypotheses an a priori bound on the smooth solutions of (1.1). Existence of solutions can be then obtained via the usual Faedo-Galerkin method. However, supplementary hypotheses are needed in order to pass to the limit in the nonlinear terms. Hence we will assume furthermore that g and h are locally Lipschitzian in the following sense.

There exists $\delta\in]0,1]$ such that for every $R>0$, there exists $C_3=C_3(R)$ and $C_4=C_4(R)$ satisfying

$$\text{(H4)} \qquad |g(v)-g(w)|_{-1}\leqq C_3|v-w|_{1-\delta} \quad \forall v,w\in V \quad \|v\|,\|w\|\leqq R;$$

$$\text{(H5)} \qquad |h(v)-h(w)|_{-1}\leqq C_4(\|v\|+\|w\|)|v-w|_{1-\delta} \quad \forall v,w\in V \quad \|v\|,\|w\|\leqq R.$$

Under these hypotheses (i.e., (H1)-(H5)), for every $f \in L^2(0, T; V_{-1})$ for all $T > 0$ and $u_0 \in V$, $u_1 \in H$ there exists an unique function $u(t)$ such that

$$u \in C(\mathbb{R}^+; V), \qquad u_t \in C(\mathbb{R}^+, H) \cap L^2(0, T; V) \quad \forall T > 0$$

that satisfies (1.1).

Moreover, if f is time independent, i.e., $f(t) \equiv f \in V_{-1}$, the mappings

$$S(t) = \{u_0, u_1\} \to \{u(t), u_t(t)\}$$

form a semigroup on $V \times H$: that is, for fixed $t > 0$, $S(t)$ is continuous on $V \times H$ and

$$S(0) = I, \qquad S(t_1 + t_2) = S(t_1) \circ S(t_2), \quad t_1, t_2 > 0.$$

2. Bounded absorbing sets and attractors.

2.1. Time-uniform estimates in $V_1 \times H$.

2.1.1. The linear case. In this section we show the decay to zero of the linear operator $\Sigma(t)$ as $t \to +\infty$, as we have announced before.

PROPOSITION 2.1. *Let λ_1 be the first eigenvalue of A and set $\varepsilon_0 = \min(1/\alpha, \alpha\lambda_1/(2(1+\lambda_1)))$. Then, for every $\varepsilon \in]0, \varepsilon_0[$,*

$$|\{u, v\}|_\varepsilon^2 \equiv (1 - \varepsilon\alpha)\|u\|^2 + \varepsilon^2|u|^2 + |v + \varepsilon u|^2$$

induces a norm on $V_1 \times H$, equivalent to the usual one, and for every $\{u_0, u_1\} \in V_1 \times H$, we have

$$|\Sigma(t)\{u_0, u_1\}|_\varepsilon \leqq e^{-\varepsilon t}|\{u_0, u_1\}|_\varepsilon \quad \forall t \geqq 0. \tag{2.1}$$

Proof. Actually we are going to show a slightly stronger result than (2.1). We take $g = h = 0$ and $f \in V_{-1}$ and denote by $\{u(t), u_t(t)\}$ the solution of (1.1). Then

$$|\{u, v\}|_\varepsilon^2 \leqq |\{u_0, v_0\}|^2 \, e^{-2\varepsilon t} + \frac{(1 - e^{-2\varepsilon t})}{2\alpha\varepsilon}|f|_{-1}^2. \tag{2.2}$$

In order to prove (2.2) we set $v = u_t + \varepsilon u$, where $\varepsilon \leqq \varepsilon_0$. Thus

$$v_t = u_{tt} + \varepsilon u_t = u_{tt} + \varepsilon(v - \varepsilon u), \tag{2.3}$$

and making use of (1.1) we find

$$\begin{aligned} &v_t + (1 - \varepsilon\alpha)Au + \varepsilon^2 u + (\alpha A - \varepsilon)v = f, \\ &u_t = v - \varepsilon u. \end{aligned} \tag{2.4}$$

Taking the scalar product of $(2.4)_1$ with v we have

$$\frac{1}{2}\frac{d}{dt}|v|^2 + (1 - \varepsilon\alpha)(Au, v) + \varepsilon^2(u, v) + ((\alpha A - \varepsilon)v, u) = (f, v), \tag{2.5}$$

and using $(2.4)_2$, we can deduce that

$$(Au, v) = \frac{1}{2}\frac{d}{dt}\|u\|^2 + \varepsilon\|u\|^2,$$

$$(u, v) = \frac{1}{2}\frac{d}{dt}|u|^2 + \varepsilon|u|^2.$$

Substituting in (2.2) we get

$$\frac{1}{2}\frac{d}{dt}[(1 - \varepsilon\alpha)\|u\|^2 + |v|^2 + \varepsilon^2|u|^2] + \varepsilon(1 - \varepsilon\alpha)\|u\|^2 + ((\alpha A - \varepsilon)v, u) + \varepsilon^3|u|^2 = (f, v). \tag{2.6}$$

Now, we have that

$$\lambda_1|v|^2 \leqq \|v\|^2.$$

Thus, thanks to the hypotheses made on ε, it is easy to see that

$$((\alpha A-\varepsilon)u, v) \geqq \frac{\alpha}{2}\|v\|^2+\varepsilon|v|^2, \tag{2.7}$$

so that by (2.6) we get, setting $y=|\{u, v\}|_\varepsilon^2$,

$$\frac{1}{2}\frac{dy}{dt}+\varepsilon y+\frac{\alpha}{2}\|v\|^2 \leqq |f|_{L^\infty(\mathbb{R}^+, V_{-1})}\|v\| \leqq \frac{\alpha}{2}\|v\|^2+\frac{1}{2\alpha}|f|^2_{L^\infty(\mathbb{R}^+, H)}. \tag{2.8}$$

By this and Gronwall's lemma it is straightforward to derive (2.2). □

By (2.2) the following result also follows:

$$\|\Sigma_\varepsilon(t)\{u_0, u_1+\varepsilon u_0\}\|_{V_1\times H} \leqq e^{-2\varepsilon t}\|\{u_0, u_1+\varepsilon u_0\}\|_{V_1\times H} \quad \forall t \geqq 0, \quad \forall \varepsilon \in]0, \varepsilon_0]. \tag{2.9}$$

2.1.2. The nonlinear case. In order to obtain estimates similar to those of the preceding paragraph we need two more assumptions on the nonlinear terms. Hence we assume that there exists $C_5>0$ such that

$$\liminf_{\|v\|\to+\infty}\frac{(v, g(v))-C_5G(v)}{\|v\|^2} \geqq 0 \tag{H6}$$

and there exist $C_6>0$ and $\sigma_3>0$ such that

$$|h(v)|_{-1} \leqq C_4(1+\|v\|)^{1-\sigma_3} \quad \forall v \in V. \tag{H7}$$

This last hypothesis is somewhat restrictive but we shall relax it below; see Remark 2.2. We are now in position to prove the following.

PROPOSITION 2.2. *Let $f \in H$ and g, h satisfy hypotheses* (H1), (H2), (H3), (H6), *and* (H7). *Then there exists a bounded absorbing set B_0 in the space $V_1\times H$ for the dynamical system represented by* (1.1).

Proof. We are going to prove the following result: For every $\varepsilon \in]0, \varepsilon_1[$, where $\varepsilon_1=\min(1/\alpha, \alpha/3\lambda_1)$ there exist positive constants $k=k(\varepsilon)$ and ρ (independent of ε) such that the solution $u(t)$ of (1.1) satisfies

$$\begin{aligned}(1-\varepsilon\alpha)\|u\|^2+\varepsilon^2|u|^2+|u_t+\varepsilon u|^2+2G(u)\\ \leqq [(1-\varepsilon\alpha)\|u_0\|^2+\varepsilon^2|u_0|^2+|u_1+\varepsilon u_0|^2+2G(u_0)]\,e^{-\varepsilon\rho t}\\ +\frac{1}{\varepsilon\rho}(1-e^{-\varepsilon\rho t})\left(k+\frac{1}{\alpha}|f|^2\right).\end{aligned} \tag{2.10}$$

Proceeding similarly as in Proposition 2.1, it is easy to get

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}[(1-\varepsilon\alpha)\|u\|^2+\varepsilon^2|u|^2+|v|^2]+\varepsilon(1-\varepsilon\alpha)\|u\|^2+\varepsilon^2|u|^2\\ +\alpha\|v\|^2-\varepsilon|v|^2=(f, v)-(g(u), v)-(h(u), v).\end{aligned} \tag{2.11}$$

By $(2.4)_2$ and the decomposition (1.2), we have

$$\begin{aligned}(g(u), v)=\frac{d}{dt}G(u)+(p(u), v_t)+\varepsilon(g(u), u),\\ (h(u_t), v)=(h(u_t), u_t)+\varepsilon(h(u_t), u).\end{aligned} \tag{2.12}$$

Therefore equality (2.11) can be rewritten as

$$\begin{aligned}&\frac{d}{dt}[(1-\varepsilon\alpha)\|u\|^2+\varepsilon^2|u|^2+|v|^2+2G(u)]+\varepsilon[(1-\varepsilon\alpha)\|u\|^2+\varepsilon^2|u|^2]\\&\quad+\varepsilon[(1-\varepsilon\alpha)\|u\|^2+2(g(u),u)]+2\alpha\|v\|^2-2\varepsilon|v|^2+2(h(u_t),u_t)+\varepsilon^3|u|^2\\&\quad=2(p(u),u_t)-2\varepsilon(h(u_t),u).\end{aligned}\tag{2.13}$$

Now hypotheses (H2), (H3), (H6), and (H7) imply that for every choice of $\delta_i>0$ $(i=1,\cdots,4)$ there exist corresponding constants $k_i>0$ $(i=1,\cdots,4)$ such that

$$(g(u),u)\geqq C_5G(u)-\frac{\varepsilon^2\delta_1}{2}\|u\|^2+k_1,\tag{2.14}$$

$$(p(u),u_t)\leqq\|p(u)\|_*\|u_t\|\leqq[\delta_2(1+|G(u)|)^{1/2}+k_2]\|u_t\|,\tag{2.15}$$

$$(h(u_t),u_t)\geqq-\frac{\delta_3}{2}\left(1+\frac{1}{2}\|u_t\|^2\right)-k_3,\tag{2.16}$$

$$-(h(u_t),u)\leqq|h(u_t)|_{-1}\|u\|\leqq C_4(1+\|u_t\|)\|u\|.\tag{2.17}$$

Making use of the Cauch-Schwarz inequality and the fact that $\|u_t\|^2\leqq2\|v\|^2+2\varepsilon^2\|u\|^2$, it is easy to deduce from (2.15)-(2.17) that

$$2(h(u_t),u_t)\geqq-\delta_3(1+\|v\|^2+\varepsilon^2\|u\|^2),\tag{2.18}$$

$$|2\varepsilon(h(u_t),u)|\leqq k_5+\varepsilon^2(\delta_5\|u\|^2+\delta_4\|v\|^2),\tag{2.19}$$

where k_5 and δ_5 depend on $\varepsilon, k_4, \delta_3$,

$$-2(p(u),u_t)\leqq k_6+\varepsilon^2\delta_8\|u\|^2+2\varepsilon\delta_2G(u)+\varepsilon\delta_6\|v\|^2,\tag{2.20}$$

where k_6 and δ_6 depend on $\varepsilon, \delta_2, k_2, \delta_4$.

Inserting all these inequalities in (2.13) and setting $z=(1-\varepsilon\alpha)\|u\|^2+\varepsilon^2|u|^2+|v|^2+2G(u)$ we have

$$\begin{aligned}&\frac{dz}{dt}+\varepsilon[(1-\varepsilon\alpha)\|u\|^2+\varepsilon^2|u|^2+2(C_5-\delta_2)G(u)]\\&\quad+\varepsilon\{[1-\varepsilon(\alpha+\delta_1+\delta_3+\delta_4+\delta_6)]\|u\|^2+\varepsilon^2|u|^2\}\\&\quad+(2\alpha-\delta_2-\varepsilon^2\delta_3-\delta_6)\|v\|^2-2\varepsilon|v|^2\leqq2(f,v)+k_7,\end{aligned}\tag{2.21}$$

where $k_7=k_1+\delta_3+k_5+k_4$.

Now we note that if $\varepsilon<\min(1/\alpha,\alpha/3\lambda_1)$, then we can find $\delta_1, \delta_3, \delta_4, \delta_5, \delta_7$ such that

$$1-\varepsilon(\alpha+\delta_1+\delta_3+\delta_4+\delta_6)\geqq0,$$

$$\frac{1}{\lambda_1}\geqq\frac{3\varepsilon}{\alpha-\delta_3-\varepsilon\delta_4-\delta_6},$$

so that we can forget the term in $\{\}$ and that

$$(2\alpha-\delta_2-\varepsilon^2\delta_3-\delta_6)\|v\|^2\geqq\alpha\|v\|^2+\varepsilon|v|^2.$$

By this inequality and (2.21) we can deduce

$$\frac{dz}{dt}+\varepsilon\rho z\leqq k_7+\frac{|f|^2}{\alpha},\qquad0<\rho<\min\{1,C_3\},\tag{2.22}$$

and using Gronwall's lemma, (2.10) follows. Now hypothesis (H1) implies that for all $\delta_7 > 0$ there exists $k_8 > 0$ such that

$$2G(u) + \delta_7 \|u\|^2 \geqq -k_8,$$

so that from (2.10) we deduce that if $\varepsilon < \varepsilon_2 = \min(1/2\alpha, \alpha/3\lambda_1)$

$$\tfrac{1}{2}\|u\|^2 + \varepsilon^2 |u|^2 + |v|^2 \leqq k_9 + z(0)\, e^{-\varepsilon\rho t}, \tag{2.23}$$

where

$$k_9 = -k_8 + \frac{k_7}{\varepsilon\rho} + \frac{1}{\varepsilon\rho\alpha}|f|^2.$$

Now from (2.22) it readily follows that for every choice of u_0, u_1, $\varepsilon < \varepsilon_2$, and G (of course restricted by (H1) and (H6)), there exists $t_0 = t_0(\varepsilon, \|u_0\|, |u_1|, G)$ such that for $t \geqq t_0$

$$(u(t), u_t(t) + \varepsilon u(t)) \in \{(x, y) \in V \times H : \|(x, y)\|^2_{V\times H} \leqq 2k_9 + 1\} \equiv B_0, \tag{2.24}$$

i.e., B_0 is absorbing for the dynamical system (1.1). □

Remark 2.1. It is possible to assume $f \in C_b(\mathbb{R}^+; H)$ and easily modify the above proof so that $|f|$ is replaced by $|f|_{L^\infty(\mathbb{R}^+;H)}$.

Remark 2.2. The hypothesis (H7) is somewhat restrictive with respect to the applications we have in view. We relax this assumption by writing $h = h_1 + h_2$, where h_1 satisfies (H7) and h_2 is such that there exist $\delta_8 \in]0, \tfrac{1}{2}]$ and C_5 such that

$$|h_2(v)|_{-1} \leqq C_5 (h_2(v), v)^{1-\delta_8}. \tag{2.25}$$

Now Proposition 1.2 can be extended to this case.

PROPOSITION 2.3. *The conclusions of Proposition* 1.2 *hold true if we replace the fact that h satisfies* (H7) *by the assumptions* $h = h_1 + h_2$ *where* h_1 *satisfies* (H7) *and* h_2 *satisfies* (2.25).

Proof. From (2.24) it follows that

$$\varepsilon |(h_2(u_t), u)| \leqq \varepsilon |h_2(u_t)|_{-1} \|u\| \leqq \varepsilon C_5 (h_2(u_t), u_t) + \varepsilon^{2q} C_6 \|u\|^{2q},$$

with $2q = 1/\delta_8$, $q \geqq 1$. Then (2.22) reads as

$$\frac{dz}{dt} + \varepsilon\rho z \leqq k_7 + \frac{|f|^2}{\alpha} + C_6 \varepsilon^{2q} \|u\|^{2q}.$$

But, since for $\varepsilon \leqq 1/2\alpha$ we have $z \geqq \tfrac{1}{2}\|u\|^2$, it follows that

$$\frac{dz}{dt} + \varepsilon\rho z \leqq k_7 + \frac{|f|^2}{\alpha} + C_7 \varepsilon^{2q} z^{2q}.$$

Now, if we take ε sufficiently small, we can suppose

$$\varepsilon C_7^{1/q} \left(k_7 + \frac{2|f|^2}{\alpha} \right) \leqq \left(\frac{|f|^2}{2\alpha} \right)^{1/q}.$$

We choose z_0 and ε such that

$$C_7 \varepsilon^{2q} z_0 \leqq \frac{1}{2} \frac{|f|^2}{\alpha}, \qquad 0 \leqq t \leqq T, \tag{2.26}$$

so that

$$C_7 \varepsilon^{2q} z(t)^q \leqq \frac{|f|^2}{\alpha}, \qquad 0 \leqq t \leqq T, \tag{2.27}$$

and it can be that either

$$T<\infty \quad \text{and} \quad C_7\varepsilon^{2q}z^q(T)=\frac{|f|^2}{\alpha} \tag{2.28}$$

or

$$T=\infty. \tag{2.29}$$

In the interval $0\leqq t\leqq T$, by the previous assumption on ε, we have

$$\frac{dz}{dt}+\varepsilon\rho z\leqq k_7+\frac{2|f|^2}{\alpha},$$

and therefore

$$z(t)\leqq z_0\, e^{-\varepsilon\rho t}+\frac{1}{\varepsilon\rho}\left(k_7+\frac{2|f|^2}{\alpha}\right)(1-e^{-\varepsilon\rho t}), \qquad 0\leqq t\leqq T.$$

Now, since $x\to x^q$ is convex for $q\geqq 1$, we obtain

$$z(t)^q\leqq z_0^q\, e^{-\varepsilon\rho t}+\frac{1}{\varepsilon^q\rho^q}\left(k_7+\frac{2|f|^2}{\alpha}\right)^q(1-e^{-\varepsilon\rho t}),$$

so that

$$C_7\varepsilon^{2q}z(t)^q\leqq C_7\varepsilon^{2q}z_0^q\, e^{-\varepsilon\rho t}+\frac{C_7\varepsilon^q}{\rho^q}\left(k_7+\frac{2|f|^2}{\alpha}\right)(1-e^{-\varepsilon\rho t}), \qquad 0\leqq t\leqq T.$$

But from the assumptions (2.26) and (2.28) on z_0 and ε, it follows that

$$C_7\varepsilon^{2q}z^q(t)\leqq\frac{1}{2}\frac{|f|^2}{\alpha}, \qquad 0\leqq t\leqq T.$$

From this we deduce that (2.28) is impossible and $T=+\infty$, giving us

$$z(t)\leqq z_0\, e^{-\varepsilon\rho t}+\frac{1}{\varepsilon\rho}\left(k_7+\frac{|f|^2}{\alpha}\right)(1-e^{-\varepsilon\rho t}), \qquad 0\leqq t\leqq\infty, \tag{2.30}$$

and

$$C_7\varepsilon^{2q}z_0^q\leqq\frac{1}{2}\frac{|f|^2}{\alpha} \quad \text{for } \varepsilon\leqq\varepsilon_0.$$

Now we still need to prove that there exists an absorbing set in $V_1\times H$, since in the previous argument ε depends on z_0.

Then, if $|z_0|\geqq k_1$ with $R_1\geqq(1/\varepsilon_0^2)((1/2C_7)(|f|^2/\alpha))^{2/q}$, it follows that

$$z(t)\leqq z_0\, e^{-\varepsilon\rho t}+C_8\overline{z_0}(1-e^{-\varepsilon\rho t}) \quad \forall t\geqq 0, \quad \varepsilon\leqq\varepsilon_0.$$

Now, if $R_2=16C_8$, and $R\geqq R_1+R_2$, it is easy to see that

$$z(t)\leqq z_0\, e^{-\varepsilon\rho t}+\frac{R}{4}(1-e^{-\varepsilon\rho t}),$$

so that for $t\geqq t_1(R)$, $z(t)\leqq R/2$. At this point, if $R/2\geqq R_2$, we can continue the argument until $R/2\leqq R_2$. Thus $\{|z|\leqq R_2\}$ is absorbing. □

2.2. Construction of the global attractor. We will assume that the nonlinear mappings g and h, introduced in §§ 1.1–1.2, satisfy the following hypothesis:

(H8) $\exists \delta \in]0, \frac{1}{2}]$ such that $\forall (\xi, \eta) \in V \times V$ and $\forall \rho \geqq 0$,
$\|\xi\| \leqq \rho$ and $|\eta| \leqq \rho \Rightarrow \exists C_9(\rho)$ such that
$|g(\xi)|^2_{2\delta-1} + |h(\eta)|^2_{2\delta-1} \leqq C_9(\rho)(1 + \|\eta\|^2)$.

This section will be devoted to the proof of the following theorem.

THEOREM 2.1. *The ω-limit set of B_0,*

$$\mathscr{A} = \omega(B_0) = \bigcap_{s>0} \operatorname{cl}\left(\bigcup_{t \geqq s} S(t) B_0 \right),$$

where cl *stands for the closure with respect to the topology of $V_1 \times H$, which is the global attractor for $S(t)$ in that space, i.e.,*

(i) *$\mathscr{A}$ is a compact nonempty connected set in $V_1 \times H$;*

(ii) *$\mathscr{A}$ is invariant under $S(t)$; $S(t)\mathscr{A} = \mathscr{A}$ $\forall t \geqq 0$;*

(iii) *$\mathscr{A}$ is globally attracting: for every bounded set $\mathscr{B}$ in $V_1 \times H$, $\operatorname{dist}(S(t)B, \mathscr{A})$ tends to zero as $t \to \infty$.*

Remarks 2.3. (i) A similar result was obtained by Massatt [8] under more restrictive hypotheses on the nonlinear mappings g and h. We also notice that our techniques are completely different from those in [8].

(ii) We have denoted $\operatorname{dist}(X, Y) = \sup_{x \in X} \inf_{y \in Y} d(x, y)$.

(iii) The proof below will also show that $\mathscr{A}$ is bounded in the norm of $V_{1+\delta} \times V_\delta$, where δ is given in (H8).

Proof. The proof will be an easy consequence of the following abstract result [2], [4].

PROPOSITION 2.4. *If a semigroup $S(t)$ on a metric space $\mathscr{E}$ possesses a bounded absorbing set B_α and for every bounded set B in $\mathscr{E}$, there exists a compact set K in $\mathscr{E}$ such that $\lim_{t \to +\infty} \operatorname{dist}(S(t)B, K) = 0$ then the ω-limit set $\omega(B_\alpha)$ is the global attractor of $S(t)$.*

Therefore let $R \geqq 0$ be given with $u_0 \in V$, $u_1 \in H$ such that $\|u_0\| \leqq R$ and $|u_1| \leqq R$. We know from Proposition 1.2 that there exist C_0, $T_0(R)$, and $K_0(R)$ such that

$$\|u(t)\|^2 + |u_t(t)|^2 \leqq C_0 \quad \forall t \geqq T_0(R), \tag{2.31}$$

$$\|u(t)\|^2 + |u_t(t)|^2 \leqq K_0(R) \quad \forall t \geqq 0. \tag{2.32}$$

Let us write now $u = v + w$, where

$$v_{tt} + \alpha A v_t + A v = \varphi(t), \tag{2.33}$$

$$v(0) = 0, \qquad v_t(0) = 0, \tag{2.34}$$

$$\varphi(t) = f - g(u(t)) - h(u_t(t)), \tag{2.35}$$

and

$$w_{tt} + \alpha A w_t + A w = 0, \tag{2.36}$$

$$w(0) = u_0, \qquad w_t(0) = u_1. \tag{2.37}$$

Proposition 2.1 shows that there exist $C_\varepsilon > 0$ such that

$$\|w(t)\|^2 + |w_t(t)|^2 \leqq C_\varepsilon e^{-\varepsilon t} R^2 \quad \forall t \geqq 0. \tag{2.38}$$

Now hypothesis (H8) and (2.31) imply that

$$|g(u(t)) + h(u_t(t))|^2_{2\delta-1} \leqq C_9(R) \|u_t\|^2, \tag{2.39}$$

and integrating (2.21) between t and $t+1$ it is easy to deduce that

$$\int_t^{t+1} \|u_t(t)\|^2 \, d\tau \leqq C_{10}(R) \quad \forall t \geqq 0, \tag{2.40}$$

so that, by virtue of (2.35), we have

$$\begin{aligned} \int_t^{t+1} |\varphi(\tau)|_{2\delta-1}^2 \, d\tau &\leqq 2\left(\int_t^{t+1} |f|_{2\delta-1}^2 \, d\tau + \int_t^{t+1} |g(u(\tau)) + h(u(\tau))|_{2\delta-1}^2 \, d\tau \right) \\ &\leqq C_{11}(R). \end{aligned} \tag{2.41}$$

Suppose now that there exists $C_{12}(R)$ such that

$$|v(t)|_{2\delta+1} + |v_t(t)|_{2\delta} \leqq C_{12}(R); \tag{2.42}$$

then $\bigcup_{t\geqq 0, \|u_0\| \leqq R} \{v(t), v_t(t)\}$ is bounded in $V_{2\delta+1} \times V_\delta$ and being the injection $V_{2\delta+1} \times V_\delta \to V \times H$ compact, we have

$$\overline{\bigcup_{\substack{t\geqq 0 \\ \|u_0\| \leqq R}} \{v(t), v_t(t)\}} = K \quad \text{is compact in } V \times H.$$

But from (2.38) it follows that $\bigcup_{t\geqq 0, \|u_0\| \leqq R} \{u(t) - v(t), u_t(t) - v_t(t)\}$ is contained in the ball $B_{V\times H}(0, C e^{-\varepsilon t} R^2)$ which tends strongly to zero, so that $\overline{\bigcup_{t\geqq 0} S(t) B}^{V\times H}$ is compact in $V \times H$ and the theorem is proved.

It remains to prove (2.42). Setting $\xi = A^\delta v$, by (2.26) we have

$$\xi_{tt} + \alpha A \xi_t + A\xi = A^\delta \varphi, \tag{2.43}$$

from which, setting $\eta = \xi_t + \varepsilon \xi$ and $y = |\{\xi, \eta\}|_\varepsilon$, it is easy to derive

$$\frac{1}{2} \frac{dy}{dt} + \varepsilon y + \frac{\alpha}{2} \|\eta\|^2 = (A^\delta \varphi, \eta) = (A^{\delta - (1/2)} \varphi, A^{(1/2)} \eta) \leqq \|\eta\| \, |\varphi|_{2\delta-1}, \tag{2.44}$$

so that

$$\frac{dy}{dt} + 2\varepsilon y \leqq C_{13} |\varphi|_{2\delta-1}^2 = \zeta \tag{2.45}$$

with

$$\int_t^{t+1} \zeta(\tau) \, d\tau \leqq C_{14}(R). \tag{2.46}$$

Integration of (2.45) now yields

$$y(t) \, e^{2\varepsilon t} \leqq y(0) + \int_0^t \zeta(\tau) \, e^{2\varepsilon \tau} \, d\tau, \qquad n \leqq t \leqq n+1, \quad n \in \mathbb{N},$$

by which

$$\begin{aligned} y(t) \, e^{2\varepsilon t} &\leqq y(0) + \sum_{k=0}^{n} \int_k^{k+1} \zeta(\tau) \, e^{2\varepsilon \tau} \, d\tau \\ &\leqq y(0) + \sum_{k=0}^{n} e^{2\varepsilon(k+1)} \int_k^{k+1} \zeta(\tau) \, d\tau \\ &\leqq y(0) + C_{14}(R) \frac{e^{2\varepsilon(n+2)} - 1}{e^{2\varepsilon} - 1} \\ &\leqq y(0) + C_{14}(R) \frac{e^{2\varepsilon(t+2)}}{e^{2\varepsilon} - 1}, \end{aligned}$$

and, remembering that $y(0)=0$ by (2.35),

$$y(t) \leqq C_{14}(R)\frac{e^{4\varepsilon}}{e^{2\varepsilon}-1} \quad \forall t \geqq 0,$$

which rewrites as

$$\|\xi(t)\|^2+|\xi_t(t)+\varepsilon\eta(t)|^2 \leqq C_{15}(R); \tag{2.47}$$

but $v=A^{-\delta}\xi$, so (2.47) implies (2.42) and the proof is complete. □

3. Applications to wave equations.

3.1. Model problems. We want now to restrict ourselves to a class of model problems which fits in the abstract framework introduced in § 1 and to state sufficient conditions on the nonlinearities to guarantee that hypothesis (H8) and its subsequents be satisfied. Namely, we are going to set $A=-\Delta$ with Dirichlet boundary conditions and consider the mappings g, h as *real functions* acting on $u(x,t)$ and $u_t(x,t)$, respectively. Therefore let Ω be an open bounded connected domain in $\mathbb{R}^n$ with Lipschitz boundary. The model problems that will be considered are of the form

$$\begin{aligned} &\frac{\partial^2 u}{\partial t^2}(x,t)+\alpha\Delta\frac{\partial u}{\partial t}(x,t)-\Delta u(x,t)+g(u(x,t))+h\left(\frac{\partial u}{\partial t}(x,t)\right)=f(x,t), \\ &u(x,t)|_{\partial\Omega}=0, \\ &u(x,0)=u_0(x), \qquad \frac{\partial u}{\partial t}(x,0)=u_1(x). \end{aligned} \tag{3.1}$$

In this case $H=L^2(\Omega)$, $V_1=D(A^{1/2})=H_0^1(\Omega)$, $D(A)=H^2(\Omega)\cap H_0^1(\Omega)$ and $|\cdot|$, $\|\cdot\|$ will denote the $L^2(\Omega)$- and $H_0^1(\Omega)$-norms.

We assume that the functions g and h are such that (H1)-(H6) and (2.25) hold true. This will be the case if, e.g., $g(s)=\lambda|s|^{\beta-1}s$ and $h(s)=\mu|s|^{\gamma-1}s$ with $\lambda\geqq 0$, $\mu\geqq 0$, $\beta>0$, $\gamma>0$, and (for $n\geqq 3$), γ, $\beta<1+(n^2/n-2)$. Of course much more general functions g and h could be considered but we have restricted the exposition to homogeneous ones for the sake of simplicity.

It is merely worthwhile to note that the following results can be obtained in the general framework; we have left that framework only for a better readability of the proofs and theorems.

3.2. Further smoothness properties of the limit sets. The aim of this section is to prove that if the data are more regular, then so is the solution of (3.1), and that there exists an absorbing set for the system (3.1) in the space $V_1\times V_2\equiv H_0^1(\Omega)\times(H^2(\Omega)\cap H_0^1(\Omega))$.

To prove what follows let us make two supplementary hypotheses:

$$\begin{aligned} &\forall R\geqq 0\,\exists\sigma_4\geqq 0,\ \exists C_{16}=C_{16}(R) \text{ such that} \\ &\|v\|\leqq R\Rightarrow|g(v)|\leqq C_{16}(1+|\Delta v|)^{1-\sigma_4/2} \quad \forall v\in V; \end{aligned} \tag{H9}$$

$$\begin{aligned} &\forall R\geqq 0\,\exists\sigma_5>0,\ \exists C_{17}=C_{17}(R) \text{ such that} \\ &\|v\|\leqq R\Rightarrow|h(v)|\leqq C_{17}(1+|\Delta v|)^{1-\sigma_5} \quad \forall v\in V. \end{aligned} \tag{H10}$$

We then have the following proposition.

PROPOSITION 3.1. *Let $f\in H$ and g, h satisfy* (H9) *and* (H10). *Then there exists an absorbing set B_1 in the space $V_1\times V_2$ for the dynamical system* (1.1).

Proof. We begin to show that if $u_0 \in V_2$, $u_1 \in V_1$, then for every $R \geqq 0$ there exists $K_{10} = K_{10}(R)$ such that, for $0 < \varepsilon < \varepsilon_1$,

$$(1-\varepsilon\alpha)|\Delta u|^2 + \varepsilon\|u\|^2 + \|u_t + \varepsilon u\|^2 \leqq [(1-\varepsilon\alpha)|\Delta u_0|^2 + \|u_0\|^2 + \|u_1 + \varepsilon u_0\|^2]\, e^{-\varepsilon t} + \frac{1}{\varepsilon}(1-e^{-\varepsilon t})[K_{10}(R) + |f|]. \tag{3.2}$$

Taking the scalar product of (3.1) with $v = u_t + \varepsilon u$ in V_1, we easily get

$$\frac{1}{2}\frac{d}{dt}((1-\varepsilon\alpha)|\Delta u|^2 + \varepsilon^2\|u\|^2 + \|v\|^2) + \varepsilon(1-\varepsilon\alpha)|\Delta u|^2 + \varepsilon^3\|u\|^2 + ((-\alpha\Delta - \varepsilon)v, Av) = (f, Av) + (g(u), Av) + (h(u_t), \Delta v). \tag{3.3}$$

We choose now $R \geqq 0$ such that $|\Delta u_0|^2 + \varepsilon^2\|u\|^2 \leqq \lambda R^2$ so that, by Poincaré's inequality, we have

$$\|u_0\|^2 + |u_1|^2 \leqq R^2, \tag{3.4}$$

and then, according to the existence results we know that there exists $M = M(R)$ such that

$$\|u(t)\|^2 + |u_t(t)|^2 \leqq M(R) \quad \forall t \geqq 0.$$

From (3.4) and the hypotheses made above it follows that

$$\begin{aligned}(g(u), \Delta v) &\leqq |g(u)|\,|\Delta v| \leqq C_{16}(M(R))|\Delta v|(1+|\Delta u|)^{1-\sigma_4/2} \\ &\leqq \frac{\alpha}{4}|\Delta v|^2 + \frac{C_7(M(R))^2}{\alpha}(1+|\Delta u|)^{2-\sigma_2} \\ &\leqq \frac{\alpha}{4}|\Delta v|^2 + \frac{C_2(M(R))^2}{\alpha}\delta_{10}|\Delta u|^2 + K_{11}(\delta_{10}, R)\end{aligned} \tag{3.5}$$

for every $\delta_{10} > 0$, and

$$\begin{aligned}-(h(u_t), \Delta v) &\leqq |h(u_t)|\,|\Delta v| \leqq C_{17}(M(R))(1+|\Delta v|)^{1-\sigma_5/2}|\Delta v| \\ &\leqq \frac{\alpha}{4}|\Delta v|^2 + \frac{C_8(M(R))^2}{\alpha}(1+|\Delta u_t|)^{2-\sigma_1} \\ &\leqq \frac{\alpha}{4}|\Delta v|^2 + \frac{C_{17}(M(R))^2}{\alpha}\delta_{11}|\Delta u_t|^2 + K_{12}(\delta_{11}, R)\end{aligned} \tag{3.6}$$

for every $\delta_{11} > 0$.

Inserting these inequalities in (3.3) and using the fact that $|\Delta u_t|^2 \leqq |\Delta v|^2 + 2\varepsilon^2|\Delta u|^2$, we get

$$\begin{aligned}&\frac{d}{dt}[(1-\varepsilon\alpha)|\Delta u|^2 + \varepsilon^2\|u\|^2 + \|v\|^2] + \varepsilon[(1-\varepsilon\alpha)|\Delta u|^2 + \varepsilon^2\|u\|^2] \\ &\quad + \varepsilon\{[1-\varepsilon(\alpha+\delta_9)]|\Delta u|^2 + (2\alpha - \delta_{10})|\Delta v|^2 - 2\varepsilon\|v\|^2\} \\ &\quad \leqq \frac{\alpha}{2}|\Delta v|^2 + (f, \Delta v) + K_{13}(R) \leqq \alpha|\Delta v|^2 + \frac{|f|}{2\alpha} + K_{13}(R),\end{aligned} \tag{3.7}$$

where δ_9, δ_{10} depend on ε, C_{16}, C_{17} but can be otherwise chosen arbitrarily small. It is evident now that proceeding exactly as in Proposition 1.2 with the same bound on

ε, one can find (3.2). The proof of the existence of an absorbing set on $V_2 \times V_1$ of the form

$$B_1 = \{(x, y) \in V_2 \times V_1 : \|(x, y)\|_{V_2 \times V_1} \leqq 2K_{10} + 1\} \tag{3.8}$$

is exactly similar to that of the preceding proposition. □

Remark 3.1. If we assume $f, f_t \in C_b(\mathbb{R}^+; H)$ it is easy to prove the same result using the equality

$$-(f, \Delta v) = -\frac{d}{dt}(f, \Delta u) + (f_t, \Delta u) - \varepsilon(f, \Delta u). \tag{3.9}$$

3.3. Sufficient conditions on the nonlinearities. Let us now investigate the meaning of the abstract hypotheses (H8)-(H10) in the case of wave equations of the form (3.1).

We will suppose that the real function $g(\xi)$ satisfies the following conditions: There exists $C_{18} > 0$ such that

$$|g(\xi)| \leqq C_{18}(1 + |\xi|^2)^{\beta/2}, \tag{H11}$$

with

$$0 \leqq \beta < \infty \quad \text{if } n = 2,$$

$$0 \leqq \beta < 5 \quad \text{if } n = 3,$$

$$0 \leqq \beta < \frac{n+2}{n-2} \quad \text{if } n \geqq 4,$$

no such assumption is needed if $n = 1$.

Concerning $h(\xi)$, we will assume that there exists $C_{19} > 0$ such that

$$|h(\eta)| \leqq C_{19}(1 + |\eta|^2)^{\gamma/2}, \tag{H12a}$$

with

$$0 \leqq \gamma < \infty \quad \text{if } n = 2,$$

$$0 \leqq \gamma < \frac{7}{3} \quad \text{if } n = 3,$$

$$0 \leqq \gamma \leqq \frac{n+4}{n} \quad \text{if } n \geqq 2.$$

Now we want to show that (H11), (H12) imply (H9), (H10). We will distinguish four cases, always supposing that $\|u\| \leqq R$ and $|u_t| \leqq R$.

(i) $n = 1$. Since $H^1 \subset L^\infty$ in this case, we have

$$\|u\| \leqq R \Rightarrow |u|_{L^\infty} \leqq C' \Rightarrow |g(u)| \leqq C'' = \sup_{|w|_{L^\infty} \leqq C} |g(w)|,$$

and similarly

$$|u_t| \leqq R \Rightarrow |h(u_t)| \leqq C'''.$$

(ii) $n = 2$. By the Sobolev imbedding theorem, we have $H^1 \hookrightarrow L^{2\beta}$, for every $\beta > 0$. Then

$$|g(u)|^2 \leqq C_{18}(1 + |u|^{2\beta}) \leqq C_9(1 + C_\Omega \|u\|) \leqq C_9(1 + C_\Omega R)$$

and

$$|h(u_1)|^2 \leqq C_{19}(1+|u_t|^{2\beta}) \leqq C_{10}(1+R^{2\gamma}).$$

(iii) $n=3$. By the Sobolev imbedding theorem, we now have that if $\beta<5$,

$$|u_{2\beta}| \leqq C_1\|u\|_{H^{6/5}} \leqq M_2\|u\|^{4/5}|u|_2^{1/5} \leqq M_3R^{4/5}|\Delta u|^{1/5} \qquad (M_i>0),$$

so that

$$|g(u)|^2 \leqq C_9(1+M_3R^{4/5}|\Delta u|^{2\beta/5}) \leqq C_8(R)(1+|\Delta u|)^{2-\sigma_4/2}, \qquad \frac{2\beta}{5}<2.$$

As for h, we have similarly that if $\gamma<7/8$,

$$|u_t|_{2\gamma} \leqq M_4\|u_t\|_{H^{6/7}} \leqq M_5|u_t|^{4/7}|\Delta u|^{3/7},$$

so that

$$|h(u_t)|^2 \leqq C_{19}(1+M_5R^{4/7}|\Delta u_t|^{3\gamma/7}) \leqq C_9(R)(1+|\Delta u_t|)^{1-\sigma_5}, \qquad \frac{3\gamma}{7}<1.$$

(iv) $n \geqq 4$. We omit the calculations since they are exactly similar to the case where $n=3$.

As for hypothesis (H8), we will show that a stronger requirement on h is needed, namely,

(H12b) $$|h(\eta)| \leqq C(1+|\xi|^2)^{\gamma/2} \quad \forall \xi \in \mathbb{R}$$

with

$$\gamma \leqq 6 \quad \text{if } n=1,$$

$$\gamma<1+6/n \quad \text{if } n \geqq 2.$$

More precisely, (H11) and (H12b) imply (H8). To show this, the reasoning is similar to that used previously as concerns the part on $g(\xi)$, having noted that (H11) implies $|g(\xi)|<C(R)$ (that is, $\delta=\frac{1}{2}$) for $n=1,2$ and that, since $H^{-1}\subset L^{6/5}$, $\beta<5$ implies

$$|g(\xi)|_{-1+2\delta} \leqq C_{20}(R) \quad \text{for some } \delta>0.$$

The result for $n \geqq 4$ follows readily. As for $h(\eta)$ we will distinguish two cases: First, $n=1$: We have

$$|h(\eta)|_{-1} \leqq \sup_{\|\xi\| \leqq 1} \int_\Omega h(\eta)\xi\, dx \leqq C|h(\eta)|_{L^1}$$

$$\leqq \int (1+|\eta|)^\gamma\, dx \leqq |(1+|\eta|)|_{L^\infty}^{\gamma-2} \int (1+|\eta|)^2\, dx.$$

Since $(n=1)$ $|\xi|_{L^\infty} \leqq C|\xi|^{1/2}\|\xi\|^{1/2}$ the right-hand side is bounded by $C|\eta|^{(\gamma/2)+1}\|\eta\|^{(\gamma/2)-1}$. Hence (H13) implies that $|h(\eta)|_{-1} \leqq |(1+\|\eta\|^2)$.

Second, $n \geqq 2$: For every $\varepsilon>0$, there exists C_ε such that

$$|h(\eta)|_{-1} \leqq C_\varepsilon|h(\eta)|_{L^{(n+\varepsilon)/(n+2)}} \leqq C_\varepsilon(1+|\eta|_{L^{(2n+\varepsilon)\gamma/(n+2)}}^\gamma) \leqq C_{L^{(n+\varepsilon)/(n+2)}}(1+|\eta|_{H^s}^\gamma),$$

with $n+2/(2n+\varepsilon)\gamma=\frac{1}{2}-(s/n)$. Now, since $\gamma<1+(6/n)$, one can choose ε so that $s\gamma=2$. Therefore

$$|h(\eta)|_{-1} \leqq C_\varepsilon(1+|\eta|_{H^s}^\gamma \leqq C_\varepsilon(1+|\eta|_{L^2}^{\gamma(1-s)}|\eta|_{H^1}^{\gamma s})$$

and (H8) follows.

4. Finite dimensionality of the attractors. We consider a subset $X \subset V_1 \times H$, which satisfies the two following properties:

(4.1) X is included and bounded in $V_{1+\delta} \times V_\delta$ for some $\delta > 0$,

(4.2) X is invariant under (1.1), i.e., $S(t)X = X \quad \forall t \geqq 0$.

We shall supplement the previous hypotheses on g and h with the following ones.

For every $R > 0$, and $(u, \eta) \in V_{1+\delta} \times V_\delta$ with $|u|_{1+\delta} + |\eta|_\delta \leqq R$ there exist $\sigma_1 \in [0, 1[$ and $C \geqq 0$ such that

(H13) $$|\langle g'(u)v, \varphi\rangle| \leqq C|v|_{\sigma_1}\|\varphi\| \quad \forall v \in V, \quad \forall \varphi \in V;$$

there exist $\sigma_2 \in [0, 1[$ and $C \geqq 0$ such that

(H14) $$|\langle h'(\eta)\xi, \xi\rangle| \leqq C\|\xi\|^2_{\sigma_2} \quad \forall \xi \in V.$$

With these notations and hypotheses we can state the following theorem.

THEOREM 4.1. *Let X be a subset in $V_1 \times H$ which satisfies* (4.1) *and* (4.2). *Then X has finite Hausdorff and fractal dimensions.*

COROLLARY 4.1. *The global attractor constructed in Theorem* 2.1 *has finite Hausdorff and fractal dimensions.*

This corollary is obvious since the global attractor satisfies (4.2) by construction and we have noticed in Remark 2.3(ii) that it is bounded in $V_{1+\delta} \times V_\delta$ where $\delta > 0$ is given in (H5). Concerning the proof of Theorem 4.1, we will rely on an abstract result of Ghidaglia [7], which extends that of Constantin, Foias, and Temam [4]. First, we consider the linearized flow

$$v_{tt} + \alpha A v_t + Av + g'(u)v + h'(u_t)v_t = 0,$$

and introduce the quadratic forms $(0 < \varepsilon < 1/2\alpha)$

(4.3) $$q_\varepsilon\{v, v_t\} \equiv (1 - \alpha\varepsilon)\|v\|^2 + \varepsilon^2|v|^2 + |w + \varepsilon v|^2.$$

We immediately see that

(4.4) $$\frac{1}{2}\frac{d}{dt} q_\varepsilon\{v, v_t\} + \varepsilon(1 - \varepsilon\alpha)\|v\|^2 + \varepsilon^3|v|^2 + \alpha\|v_t + \varepsilon v\|^2$$
$$= -(g'(u)v + h'(u_t)v_t - \varepsilon^2 v, v_t + \varepsilon v).$$

Second, we use (H13) and (H14) in order to bound the right-hand side of (4.4) by

(4.5) $$\varepsilon^2|v|\,|v_t + \varepsilon v| + C\|v_t + \varepsilon v\|\,|v|_{\sigma_1} + C|v_t + \varepsilon v|^2_{\sigma_2} + \varepsilon C\|v\|\,\|v_t + \varepsilon v\|.$$

We have written $(h'(u_t)v_t, v_t + \varepsilon v) = (h'(u_t)(v_t + \varepsilon v), v_t + \varepsilon v) - \varepsilon(h'(u_t)v, v_t + \varepsilon v)$ and make use of (H14) in order to bound the last term.

Then combining (4.4) and (4.5) we deduce thanks to the Cauchy-Schwarz inequality that there exists $\varepsilon_0 < 1/2\alpha$ such that for $0 < \varepsilon \leqq \varepsilon$,

(4.6) $$\frac{d}{dt} q_\varepsilon\{v, v_t\} + \varepsilon(1 - \alpha\varepsilon)\|v\|^2 + 2\varepsilon^3|v|^2 + \alpha\|v_t + \varepsilon v\|^2 \leqq C(|v|^2_\sigma + |v_t|^2_\sigma),$$

where $\sigma = \max(\sigma_1, \sigma_2) < 1$.

Third, using the interpolation inequality

$$|\phi|_\sigma \leqq |\phi|^{(1-\sigma)/(2-\sigma)}_{\sigma-1} |\phi|^{1/(2-\sigma)}_1,$$

we deduce that (from now on $\varepsilon = \varepsilon_0$ is fixed)

$$C|v_t|_\sigma^2 \leqq \frac{\alpha}{2}\|v_t + \varepsilon v\|^2 + C|v_t|_{\sigma-1}^2,$$

so that (4.6) reads

$$\frac{d}{dt} q_{\varepsilon_0} + \gamma q_{\varepsilon_0} \leqq C[|v|_\sigma^2 + |v_t|_{\sigma-1}^2], \tag{4.7}$$

where $\gamma = \min(1/2\alpha, \alpha\lambda/2)$ and we have used the fact that

$$\lambda_1|w|^2 \leqq \|w\|^2.$$

Introducing the linear operator on $V_1 \times H$

$$K\{v, w\} = \{A^{\sigma-1}v, A^{\sigma-1}w\},$$

we can rewrite (4.7) as

$$\frac{dq_{\varepsilon_0}}{dt} + \gamma q_{\varepsilon_0} \leqq C(K\{v, v_t\}, \{v, v_t\})_{V_1 \times H}.$$

Finally, since the operator K is compact, we deduce from this inequality [7, Appendix] that X is finite dimensional. □

5. Applications, continued.

5.1. Time-periodic perturbed Sine–Gordon equation. In [1] the following equation, modelling the current in a Josephson junction, is introduced:

$$\frac{\partial^2 u}{\partial t^2} - \Delta u + \sin u = -\beta \frac{\partial u}{\partial t} + \alpha \frac{\partial(\Delta u)}{\partial t} + f(x, t), \tag{5.1}$$

$$\alpha > 0, \quad \beta > 0, \quad u(x, t) \in \mathbb{R}, \quad x \in \mathbb{R}^n.$$

Since $\sin u$ is a bounded function, we can apply the results of the preceding paragraphs to obtain the following theorem.

THEOREM 5.1. *Equation* (5.1), *together with initial and boundary conditions, defines a dynamical system which has a global attractor $\mathcal{A}$ of finite dimension.*

5.2. The power nonlinearity case. The following perturbed wave equation also occurs in quantum mechanics:

$$\frac{\partial^2 u}{\partial t^2} - \Delta u + |u|^q u = -\beta \left|\frac{\partial u}{\partial t}\right|^p \frac{\partial u}{\partial t} + \alpha \frac{\partial(\Delta u)}{\partial t} + f. \tag{5.2}$$

For this equation, bounds on p and q are required by the number of independent variables involved. Thus we have Theorem 5.2.

THEOREM 5.2. *Equation* (5.2), *together with initial and boundary condition, defines a dynamical system which has a finite-dimensional global attractor in the following cases*:

$$n = 1, \quad p \leqq 5, \quad q \text{ arbitrary},$$

$$n = 2, \quad p < 3, \quad q \text{ arbitrary},$$

$$n = 3, \quad p < 4/3, \quad q < 4,$$

$$n \geqq 4, \quad p < 4/n, \quad q < 4/(n-2).$$

REFERENCES

[1] P. S. LANDAHL, O. H. SOERENSEN, AND P. L. CHRISTIANSEN, *Soliton excitations in Josephson tunnel junctions*, Phys. Rev. B, 25 (1982), pp. 5337-5348.

[2] J. M. GHIDAGLIA AND R. TEMAM, *Attractors for damped nonlinear hyperbolic equations*, J. Math. Pures Appl., 66 (1987), pp. 273-319.

[3] ———, *Dimension of the universal attractor describing the periodically driven Sine-Gordon equations*, Transport Theory Statist. Phys., 16 (1987), pp. 253-265.

[4] P. CONSTANTIN, C. FOIAS, AND R. TEMAM, *Attractors representing turbulent flows*, Mem. Amer. Math. Soc., 53 (1985), p. 314.

[5] R. TEMAM, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, New York, Berlin, 1988.

[6] J. M. GHIDAGLIA, *Finite dimensional behaviour for weakly damped driven Schrödinger equations*, Annales de l'I.H.P., Anal. Non Linéaire, 5 (1988), pp. 365-405 C. R. Acad. Sci. Paris Sér. I, 305 (1987), pp. 291-294.

[7] ———, *Weakly damped forced Korteweg-deVries equation behave as a finite dimensional dynamical system in the long time*, J. Differential Equations, 74 (1988), pp. 369-390.

[8] P. MASSATT, *Limiting behaviour for strongly damped nonlinear wave equations*, J. Differential Equations, 48 (1983), pp. 334-349.

[9] G. F. WEBB, *Existence and asymptotic behaviour for a strongly damped nonlinear wave equation*, Canad. J. Math., 32 (1980), pp. 631-643.

SIAM J. MATH. ANAL.
Vol 22, No. 4, pp. 896-904, July 1991

002

A STRONGLY COUPLED SINGULARLY PERTURBED QUASILINEAR SECOND-ORDER SYSTEM*

JOHN S. JEFFRIES†

Abstract. A constructive existence proof is given for solutions of boundary layer type for the singularly perturbed quasilinear second-order system $\varepsilon(d^2x/dt^2) = F(t, x)(dx/dt) + g(t, x)$ subject to Dirichlet boundary conditions. The required assumptions involve only natural conditions that are induced by the O'Malley construction. In particular, restrictive conditions on the structure of $F(t, x)$ which seek to decouple the components of the system are avoided.

Key words. singular perturbations, boundary layer, strongly coupled

AMS(MOS) subject classifications. 34, 54

1. Introduction. We consider solutions exhibiting boundary layer behavior at one endpoint for the following vector differential equation:

$$\varepsilon \frac{d^2x}{dt^2} = F(t, x)\frac{dx}{dt} + g(t, x) \quad \text{for } 0 < t < 1 \tag{1.1}$$

for small values of ε $(\varepsilon \to 0^+)$ subject to the Dirichlet boundary conditions

$$x(0, \varepsilon) = \alpha, \qquad x(1, \varepsilon) = \beta, \tag{1.2}$$

where x and g are real n-dimensional vector-valued functions, α and β are real n-dimensional vectors, and F is a real $n \times n$ matrix-valued function.

The vector Dirichlet problem (1.1)-(1.2) has been considered in Hadlock [9], Chang [1]-[3], Habets [8], Freedman and Kaplan [7], Flaherty and O'Malley [6], O'Donnell [16], O'Malley [17], Chang and Howes [4], Kirschvink [15], Kelley [14], Smith [19], and Jeffries and Smith [11]. However, these works have either made rather strong restrictions on the size of the boundary layer jump or they have placed restrictions on the structure of $F(t, x)$ that effectively require the components to be only weakly coupled. In the present study we avoid these restrictions and thus allow the components of the system to be strongly coupled. It should be noted that the approach we use can be easily modified to handle the case in which $F(t, x)$, $g(t, x)$, α, and β are also functions of ε, i.e., $F = F(t, x, \varepsilon)$, $g = g(t, x, \varepsilon)$, $\alpha = \alpha(\varepsilon)$, and $\beta = \beta(\varepsilon)$. However, to ease the notational burden, we have assumed that F, g, α, and β are independent of ε. (See Smith [19] or Jeffries and Smith [11] for the required modifications if F, g, α, and β are not independent of ε.)

To prove the existence of a boundary layer solution to the problem (1.1)-(1.2) we use, with the aid of a Riccati transformation, the O'Malley construction to obtain an approximate solution. An additional Riccati transformation leads to an explicit construction of a suitable fundamental solution for the linearization of the problem about the proposed approximate solution. A resulting integral representation for the linearization provides directly the existence of a locally unique exact solution for the original problem along with error estimates of the difference between the exact solution and the approximate solution, thereby yielding precise information on the exact solution throughout the interval $0 \leqq t \leqq 1$ as $\varepsilon \to 0^+$.

* Received by the editors March 28, 1990; accepted for publication August 28, 1990.

† Department of Computer Science and Mathematics, New Mexico Highlands University, Las Vegas, New Mexico 87701.

Section 2 contains a discussion of our assumptions, and § 3 discusses the approximate solution provided by the O'Malley construction. A Riccati transformation is used in § 4 to obtain a fundamental solution for the linearization of the problem about the given approximate solution, resulting in an existence and local uniqueness theorem for the original problem along with error estimates.

2. Assumptions.

ASSUMPTION 1. There exists a continuous solution $X_0(t)$ to the reduced equation

$$F(t, X_0(t))\frac{dX_0}{dt}+g(t, X_0(t))=0 \tag{2.1}$$

such that the real parts of the eigenvalues of $F(t, X_0(t))$ are negative.

ASSUMPTION 2. There exists a decaying solution $\hat{X}_0$ to the boundary layer equation

$$\frac{d^2\hat{X}_0}{d\tau^2}=F(0, X_0(0)+\hat{X}_0(\tau))\frac{d\hat{X}_0}{d\tau}, \qquad \hat{X}_0(0)=\alpha-X_0(0). \tag{2.2}$$

It can be shown (see the discussion following formula (3.7)) that if $\hat{X}_0$ decays then it must decay exponentially. Before stating our next assumption, we must first consider the following $2n$-dimensional linear system:

$$\frac{d\hat{Z}}{d\tau}=\begin{pmatrix} 0 & I \\ \hat{A}(\tau) & \hat{B}(\tau)\end{pmatrix}\hat{Z}, \tag{2.3}$$

where

$$\begin{aligned}\hat{A}_{i,j}(\tau)&=\sum_{k=1}^{n}\frac{\partial F_{i,k}}{\partial x_j}(0, X_0(0)+\hat{X}_0(\tau))\frac{d\hat{X}_{0,k}}{d\tau},\\ \hat{B}(\tau)&=F(0, X_0(0)+\hat{X}_0(\tau)).\end{aligned} \tag{2.4}$$

It is shown in Lemma 1 that there exists a fundamental solution $\hat{Z}$ to the above linear system which satisfies the following dichotomy:

$$\begin{aligned}&\|\hat{Z}(\tau)P\hat{Z}^{-1}(u)\|\leqq K e^{-\nu(\tau-u)}, && \tau\geqq u,\\ &\|\hat{Z}(\tau)(I-P)\hat{Z}^{-1}(u)\|\leqq K, && u\geqq\tau,\end{aligned} \tag{2.5}$$

where K and ν are positive constants and $P=\left(\begin{smallmatrix} I_n & 0\\ 0 & 0\end{smallmatrix}\right)$. Essentially, this says that there are n linearly independent solutions to the system (2.3) that decay exponentially and n linearly independent solutions that are bounded in norm away from zero. Defining $P_1=\hat{Z}(0)P\hat{Z}^{-1}(0)$ we make the following assumption.

ASSUMPTION 3. The columns of $(I_n \quad 0)P_1$ span R^n.

The above assumption implies that the solution $\hat{X}_0(\tau)$ to the boundary value problem (2.2) is stable with respect to perturbations in the initial conditions. Assumption 3 is independent of the fundamental solution chosen provided the fundamental solution satisfies the dichotomy (2.5). (This follows from using arguments analogous to those given in Lemma 6.4 of [11] and letting $\varepsilon\to 0^+$.) Furthermore, if there exists a real n-dimensional vector-valued function f such that $F(t, x)=\nabla_x f(t, x)$ then Assumption 3 is always satisfied (see the discussion following formula (3.20)).

ASSUMPTION 4. The given data functions F and g are of class C^{N+1}, where $N\geqq 2$.

3. The approximate solution. In this section we construct an approximate solution to the problem (1.1)–(1.2) using the O'Malley construction. We write the approximate solution $X^N(t, \varepsilon)$ as the sum of an outer solution and a boundary layer correction function of the form

$$X^N(t, \varepsilon)=X(t, \varepsilon)+\hat{X}(\tau, \varepsilon), \qquad \tau:=t/\varepsilon, \tag{3.1}$$

where $X(t, \varepsilon)$ and $\hat{X}(\tau, \varepsilon)$ possess expansions in ε of the form

$$\begin{pmatrix} X(t, \varepsilon) \\ \hat{X}(\tau, \varepsilon) \end{pmatrix} = \sum_{k=0}^{N} \begin{pmatrix} X_k(t) \\ \hat{X}_k(\tau) \end{pmatrix} \varepsilon^k. \tag{3.2}$$

3.1. The outer solution. The outer solution coefficient functions $X_k(t)$, $k = 1, \cdots, N$, are determined by requiring that the outer solution satisfy the differential equation (1.1) and the right boundary conditions to Order (ε^N), i.e.,

$$\varepsilon \frac{d^2X}{dt^2} = F(t, X)\frac{dX}{dt} + g(t, X) + \bar{\rho}(t, \varepsilon), \qquad X(1, \varepsilon) = \beta - \bar{\psi}_N(\varepsilon), \tag{3.3}$$

where $\bar{\rho}(t, \varepsilon)$ and $\bar{\psi}_N(\varepsilon)$ are of Order (ε^{N+1}). A straightforward calculation shows that the outer coefficients functions must satisfy

$$\begin{aligned} &F(t, X_0(t))\frac{dX_k}{dt} = -[X_k(t) \cdot \nabla_x F(t, X_0(t))]\frac{dX_0}{dt} - g_x(t, X_0(t))X_k(t) + P_{k-1}(t), \\ &X_k(1) = 0, \end{aligned} \tag{3.4}$$

where $P_{k-1}(t)$ is known successively in terms of the preceding coefficient functions. Since $F(t, X_0(t))$ is nonsingular (see Assumption 1) there exists a unique solution to each of the above linear terminal value problems.

3.2. The boundary layer correction functions. The boundary layer correction functions are determined by requiring that the approximate solution $X^N(t, \varepsilon)$ satisfies the boundary value problem to Order (ε^N), i.e.,

$$\varepsilon \frac{d^2X^N}{dt^2} = F(t, X^N(t, \varepsilon))\frac{dX^N}{dt} + g(t, X^N(t, \varepsilon)) + \rho(t, \varepsilon), \tag{3.5}$$

$$X^N(0, \varepsilon) = \alpha - \phi_N(\varepsilon), \qquad X^N(1, \varepsilon) = \beta - \psi_N(\varepsilon), \tag{3.6}$$

where $\rho(t, \varepsilon)$ is a continuous function of t and satisfies $\int_0^1 |\rho(s, \varepsilon)|\, ds \leqq \text{Const. } \varepsilon^{N+1}$ and $|\phi_N(\varepsilon)|, |\psi_N(\varepsilon)| \leqq \text{Const. } \varepsilon^{N+1}$.

Using the results of § 3.1, changing variables from t to τ, and expanding about $\varepsilon = 0$, we find that the leading boundary layer correction function $\hat{X}_0$ must satisfy

$$\frac{d^2\hat{X}_0}{d\tau^2} = F(0, X_0(0) + \hat{X}_0(\tau))\frac{d\hat{X}_0}{d\tau}, \qquad \hat{X}_0(0) = \alpha - X_0(0). \tag{3.7}$$

In Assumption 2 we assumed that there exists a decaying solution $\hat{X}_0$ to the above problem. It follows, since $d\hat{X}_0/d\tau$ is a solution to the linear system

$$\frac{d}{d\tau}\left(\frac{d\hat{X}_0}{d\tau}\right) = D(\tau)\left(\frac{d\hat{X}_0}{d\tau}\right), \tag{3.8}$$

where $D(\tau) := F(0, X_0(0) + \hat{X}_0(\tau))$ and the real parts of the eigenvalues of $F(0, X_0(0))$ are negative, that $d\hat{X}_0/d\tau$ must be exponentially decaying. This, in turn, implies that $\hat{X}_0$ is also exponentially decaying, i.e., there exists a positive constant ν_0 such that

$$\left\| \frac{d\hat{X}_0}{d\tau} \right\|, \|\hat{X}_0(\tau)\| \leqq \text{Const. } e^{-\nu_0 \tau}. \tag{3.9}$$

The higher-order boundary layer correction terms must satisfy the following linear differential equations:

$$\begin{aligned}\frac{d^2\hat{X}_k}{d\tau^2} &= F(0, X_0(0)+\hat{X}_0(\tau))\frac{d\hat{X}_k}{d\tau}\\ &\quad +[\hat{X}_k(\tau)\cdot\nabla_x F(0, X_0(0)+\hat{X}_0(\tau))]\frac{d\hat{X}_0}{d\tau}+\hat{g}_{k-1}(\tau),\\ \hat{X}_k(0) &= -X_k(0),\end{aligned} \tag{3.10}$$

where $\hat{g}_{k-1}(\tau)$ is known in terms of the previous boundary layer correction terms and is exponentially decaying providing the previous terms are exponentially decaying.

Defining $\hat{Y}_k := d\hat{X}_k/d\tau$, and using $\hat{A}$ and $\hat{B}$ as defined in (2.4), we can rewrite the differential equation (3.10) as the following inhomogeneous linear system:

$$\frac{d}{d\tau}\begin{pmatrix}\hat{X}_k\\ \hat{Y}_k\end{pmatrix}=\begin{pmatrix}0 & I_n\\ \hat{A}(\tau) & \hat{B}(\tau)\end{pmatrix}\begin{pmatrix}\hat{X}_k\\ \hat{Y}_k\end{pmatrix}+\begin{pmatrix}0\\ \hat{g}_{k-1}(\tau)\end{pmatrix}. \tag{3.11}$$

Using the fundamental solution $\hat{Z}(\tau)$ (see Assumption 3) an exponentially decaying solution is given by

$$\begin{aligned}\begin{pmatrix}\hat{X}_k(\tau)\\ \hat{Y}_k(\tau)\end{pmatrix} &= \hat{Z}(\tau)P\hat{Z}^{-1}(0)d_k+\int_0^\tau \hat{Z}(\tau)P\hat{Z}^{-1}(u)\begin{pmatrix}0\\ \hat{g}_{k-1}(u)\end{pmatrix}du\\ &\quad -\int_\tau^\infty \hat{Z}(\tau)(I-P)\hat{Z}^{-1}(u)\begin{pmatrix}0\\ \hat{g}_{k-1}(u)\end{pmatrix}du,\end{aligned} \tag{3.12}$$

where d_k is a $2n$-dimensional vector. It follows from Assumption 3 that there exists a vector d_k such that $\hat{X}_k(0)=-X_k(0)$. We now turn to the proof of the existence of a fundamental solution $\hat{Z}(\tau)$, satisfying the dichotomy (2.5).

LEMMA 1. *There exists a fundamental solution $\hat{Z}$ to the linear system* (2.3) *satisfying the dichotomy* (2.5).

Proof. The strategy we use is to construct a fundamental solution on an interval of the form $[\tau_0, \infty)$, τ_0 a positive constant, that satisfies the dichotomy (2.5). We then extend this fundamental solution to the interval $[0, \infty)$. It follows from Coppel [5, p. 13] that the extended fundamental solution satisfies the dichotomy on the interval $[0, \infty)$.

Defining

$$\hat{Z}(\tau) := \begin{pmatrix}I_n & \hat{S}(\tau)\\ \hat{T}(\tau) & I_n+\hat{T}(\tau)\hat{S}(\tau)\end{pmatrix}\begin{pmatrix}\hat{\eta}(\tau) & 0\\ 0 & \hat{\xi}(\tau)\end{pmatrix}, \tag{3.13}$$

$$\hat{Z}^{-1}(\tau)=\begin{pmatrix}\hat{\eta}^{-1}(\tau) & 0\\ 0 & \hat{\xi}^{-1}(\tau)\end{pmatrix}\begin{pmatrix}I_n+\hat{S}(\tau)\hat{T}(\tau) & -\hat{S}(\tau)\\ -\hat{T}(\tau) & I_n\end{pmatrix} \tag{3.14}$$

for $\tau \geqq \tau_0$, the linear system decouples

$$\frac{d}{d\tau}\begin{pmatrix}\hat{\eta}\\ \hat{\xi}\end{pmatrix}=\begin{pmatrix}\hat{T} & 0\\ 0 & \hat{B}-\hat{T}\end{pmatrix}\begin{pmatrix}\hat{\eta}\\ \hat{\xi}\end{pmatrix}, \tag{3.15}$$

provided $\hat{T}$ and $\hat{S}$ satisfy the following differential equations for $\tau \geqq \tau_0$:

$$\frac{d\hat{T}}{d\tau}=\hat{B}\hat{T}-\hat{T}^2+\hat{A}, \tag{3.16}$$

$$\frac{d\hat{S}}{d\tau}=\hat{T}\hat{S}+\hat{S}[\hat{T}-\hat{B}]+I_n. \tag{3.17}$$

PROPOSITION 1. *There exists, for τ_0 sufficiently large, a bounded solution $\hat{T}$ to the differential equation* (3.16). *Furthermore, there exist positive constants K_0 and γ_0 such that*

$$\|\hat{T}(\tau)-\hat{B}(\tau)\| \leqq \gamma_0 e^{-\nu_0(\tau-\tau_0)}, \qquad \tau \geqq \tau_0, \tag{3.18}$$

$$\|\gamma_0\| \leqq K_0 e^{-\nu_0\tau_0}. \tag{3.19}$$

Proof. Letting $\hat{T}(\tau)=\hat{B}(\tau)+\hat{E}(\tau)$, we find that $\hat{E}(\tau)$ must satisfy

$$\frac{d\hat{E}}{d\tau}=-\hat{E}\hat{B}-\hat{E}^2+\hat{A}-\frac{d\hat{B}}{d\tau}. \tag{3.20}$$

(Note that if there exists an n-dimensional vector-valued function f such that $F=\nabla_x f$, then $\hat{A}-(d\hat{B}/d\tau)=0$ and so $T=\hat{B}$ is a solution to (3.16), and we may set $\tau_0=0$.) From Assumptions 1 and 2 it follows that there exists a fundamental solution $\hat{\eta}_1$ to the linear system $d\hat{\eta}_1/d\tau=\hat{B}(\tau)\hat{\eta}_1$ such that $\|\hat{\eta}_1(u)\hat{\eta}_1^{-1}(\tau)\| \leqq K_1 e^{-\nu_1(u-\tau)}$ for $u \geqq \tau$, where K_1 and ν_1 are positive constants. Using the fundamental solution $\hat{\eta}_1$ and imposing the terminal condition $\hat{E}(\infty)=0$, we find that $\hat{E}$ satisfies the following integral equation:

$$\hat{E}(\tau)=\int_\tau^\infty \left(\frac{d\hat{B}}{du}-\hat{A}\right)\hat{\eta}_1(u)\hat{\eta}_1^{-1}(\tau)\,du+\int_\tau^\infty \hat{E}^2(u)\hat{\eta}_1(u)\hat{\eta}_1^{-1}(\tau)\,du. \tag{3.21}$$

Letting $\gamma_0<\nu_1/2K_1$ and choosing τ_0 large enough so that $\|\hat{A}(\tau)-(d\hat{B}/d\tau)(\tau)\| \leqq \gamma_0\nu_1/2K_1$ for $u \geqq \tau_0$. (This is possible since $\|\hat{A}(\tau)-(d\hat{B}/d\tau)(\tau)\| \leqq \text{Const. } e^{-\nu_0\tau}$ for $\tau \geqq 0$.) By the Banach–Picard fixed point theorem there exists a solution $\hat{E}$ such that $\|\hat{E}(\tau)\| \leqq \gamma_0$ for $\tau \geqq \tau_0$. Furthermore, using repeated substitution starting with $\hat{E}_0(\tau) := \int_\tau^\infty ((d\hat{B}/du)-\hat{A})\hat{\eta}_1(u)\hat{\eta}_1^{-1}(\tau)\,du$, it can be shown that $\|\hat{E}(\tau)\| \leqq \gamma_0 \exp(-\nu_0(\tau-\tau_0))$ for $\tau \geqq \tau_0$, and since $\|\hat{A}-(d\hat{B}/d\tau)\| \leqq \text{Const. } e^{-\nu_0\tau}$ for $\tau \geqq 0$, γ_0 can be chosen so as to satisfy (3.19).

PROPOSITION 2. *The fundamental solution $\hat{\xi}$ satisfies $\|\hat{\xi}(u)\hat{\xi}^{-1}(u)\| \leqq \text{Const.}$ for $\tau, u \geqq \tau_0$.*

Proof. This follows from the boundedness of $\int_{\tau_0}^\infty \|\hat{B}-\hat{T}\|\,du$.

PROPOSITION 3. *The fundamental solution $\hat{\eta}(\tau)$ satisfies $|\hat{\eta}(\tau)\hat{\eta}^{-1}(u)| \leqq K_2 e^{-\nu_2(\tau-u)}$ for $\tau \geqq u \geqq \tau_0$, for some positive constants K_2 and ν_2.*

Proof. This follows from Assumption 1, the estimates (3.18)–(3.19), and Proposition 1 of [5].

PROPOSITION 4. *There exists a bounded solution $\hat{S}(\tau)$ to the differential equation* (3.17).

Proof. Imposing the initial condition $\hat{S}(\tau_0)=0$ and using the fundamental solutions $\hat{\eta}$ and $\hat{\xi}$ we find that

$$\hat{S}(\tau)=\int_{\tau_0}^\tau \hat{\eta}(\tau)\hat{\eta}^{-1}(u)\hat{\xi}(u)\hat{\xi}^{-1}(\tau)\,du. \tag{3.22}$$

It follows from Propositions 2 and 3 that $\hat{S}(\tau)$ is bounded for $\tau \geqq \tau_0$. Having shown that $\hat{S}$ is bounded we may express $\hat{S}$ as $\hat{S}(\tau)=\int_{\tau_0}^\tau \hat{\eta}(\tau)\hat{\eta}^{-1}(u)[I+\hat{S}(u)\hat{E}(u)]\,du$. This observation will be helpful later in the proof of the nonsingularity of $S(1/\varepsilon, \varepsilon)$. The proof of Lemma 1 now follows from Propositions 1–4.

4. Existence and local uniqueness. In this section we use a Riccati transformation to construct a fundamental solution for the linearization of the problem about the approximate solution $X^N(t, \varepsilon)$. An integral representation for the linearization is then used to prove the existence of an exact solution for (1.1)–(1.2) along with error estimates and local uniqueness.

THEOREM. *Assume there exists a continuous solution $X_0(t)$ to the reduced equation such that the eigenvalues of $F(t, X_0(t))$ have negative real parts* (*Assumption* 1). *If there exists a decaying solution $\hat{X}_0$ to the boundary layer problem* (*Assumption* 2), *the columns*

of $(I_n \;\; 0)P_1$ *span* R^n (*Assumption* 3), *and the given data are sufficiently smooth* (*Assumption* 4), *then there exist constants* ε_0 *and* C_N *such that* (1.1)-(1.2) *has an exact solution* $x(t, \varepsilon)$, *satisfying the estimates*

$$|x(t,\varepsilon)-X^N(t,\varepsilon)| \leqq C_N\varepsilon^N, \qquad \left|\frac{dx}{dt}(t,\varepsilon)-\frac{d}{dt}X^N(t,\varepsilon)\right| \leqq C_N\varepsilon^{N-1} \tag{4.1}$$

uniformly on the region $0 \leqq t \leqq 1, 0 < \varepsilon \leqq \varepsilon_0$. *Moreover,* $x(t, \varepsilon)$ *is unique subject to* (4.1).

Proof. Defining

$$\tilde{x} := x(t,\varepsilon)-X^N(t,\varepsilon), \qquad \tilde{y} := \varepsilon\frac{d\tilde{x}}{dt}, \tag{4.2}$$

a straightforward calculation (see Smith [18]) shows that $\tilde{x}$ and $\tilde{y}$ must satisfy

$$\frac{d}{dt}\begin{pmatrix}\tilde{x}\\ \tilde{y}\end{pmatrix} = \frac{1}{\varepsilon}\begin{pmatrix}0 & I_n\\ \varepsilon A(t,\varepsilon) & B(t,\varepsilon)\end{pmatrix}\begin{pmatrix}\tilde{x}\\ \tilde{y}\end{pmatrix} + \begin{pmatrix}0\\ H(t,\tilde{x},(1/\varepsilon)\tilde{y},\varepsilon)+\rho(t,\varepsilon)\end{pmatrix}, \tag{4.3}$$

subject to the boundary conditions

$$L\begin{pmatrix}\tilde{x}(0,\varepsilon)\\ \tilde{y}(0,\varepsilon)\end{pmatrix} + R\begin{pmatrix}\tilde{x}(1,\varepsilon)\\ \tilde{y}(1,\varepsilon)\end{pmatrix} = \begin{pmatrix}\phi_N(\varepsilon)\\ \psi_N(\varepsilon)\end{pmatrix}, \tag{4.4}$$

where

$$L := \begin{pmatrix}I_n & 0\\ 0 & 0\end{pmatrix}, \qquad R := \begin{pmatrix}0 & 0\\ I_n & 0\end{pmatrix}, \tag{4.5}$$

$$A_{i,j}(t,\varepsilon) := \sum_{k=1}^{n}\frac{\partial F_{i,k}}{\partial x_j}(t, X^N(t,\varepsilon))\frac{dX_k^N}{dt} + \frac{\partial g_i}{\partial x_j}(t, X^N(t,\varepsilon)), \tag{4.6}$$

$$B(t,\varepsilon) := F(t, X^N(t,\varepsilon)), \tag{4.7}$$

$$\begin{aligned} H(t,u,v,\varepsilon) := \int_0^1 \Bigg\{ &\left[\frac{d}{ds}F(t, X^N(t,\varepsilon)+su)\right]v + (1+s)\frac{d^2}{ds^2}\\ &\cdot\left[g(t, X^N(t,\varepsilon)+su) + F(t, X^N(t,\varepsilon)+su)\frac{d}{dt}X^N(t,\varepsilon)\right]\Bigg\}\, ds. \end{aligned} \tag{4.8}$$

Note that $H(t, u, v, \varepsilon)$ satisfies the inequality

$$|H(t,u,v,\varepsilon)| \leqq \text{Const.}\,(\varepsilon^{-1}|u|^2 + \varepsilon|v|^2) \tag{4.9}$$

uniformly as $\varepsilon \to 0^+$, for all $v \in R^n$ and for all u in a fixed compact subset of R^n.

We now construct a fundamental solution to the homogeneous portion of (4.3). Changing variables from t to τ we therefore consider the following linear system:

$$\frac{d}{d\tau}Z(\tau,\varepsilon) = \begin{pmatrix}0 & I_n\\ \varepsilon A(\varepsilon\tau,\varepsilon) & B(\varepsilon\tau,\varepsilon)\end{pmatrix}Z(\tau,\varepsilon). \tag{4.10}$$

As in Lemma 1 we employ a Riccati transformation so as to decouple the exponentially decaying solutions from the solutions bounded away from zero:

$$Z(\tau,\varepsilon) := \begin{pmatrix}I_n & S(\tau,\varepsilon)\\ T(\tau,\varepsilon) & I+T(\tau,\varepsilon)S(\tau,\varepsilon)\end{pmatrix}\begin{pmatrix}\eta(\tau,\varepsilon) & 0\\ 0 & \xi(\tau,\varepsilon)\end{pmatrix}, \tag{4.11}$$

$$Z^{-1}(\tau,\varepsilon) = \begin{pmatrix}\eta^{-1}(\tau,\varepsilon) & 0\\ 0 & \xi^{-1}(\tau,\varepsilon)\end{pmatrix}\begin{pmatrix}I+S(\tau,\varepsilon)T(\tau,\varepsilon) & -S(\tau,\varepsilon)\\ -T(\tau,\varepsilon) & I\end{pmatrix}. \tag{4.12}$$

It follows, using the method of analysis used in Lemma 1, that $Z(\tau, \varepsilon)$ satisfies the following dichotomy

$$\begin{aligned} &|Z(\tau, \varepsilon)PZ^{-1}(u, \varepsilon)| \leqq K e^{-\nu(\tau-u)}, \qquad \tau \geqq u, \\ &|Z(\tau, \varepsilon)(I-P)Z^{-1}(u, \varepsilon)| \leqq K, \qquad u \geqq \tau \end{aligned} \tag{4.13}$$

for positive constants K and ν. We now show that $S(1/\varepsilon, \varepsilon)$ is nonsingular for sufficiently small ε. Using the fundamental solution $\eta(\tau, \varepsilon)$ it follows that

$$S(\tau, \varepsilon) = \int_{\tau_0}^{\tau} \eta(\tau, \varepsilon)\eta^{-1}(u, \varepsilon)[I + S(u)E(u, \varepsilon)]\, du, \tag{4.14}$$

where $E(u, \varepsilon) = T(u, \varepsilon) - B(u, \varepsilon)$. Since $S(\tau, \varepsilon)$ is bounded, $\|E(\tau, \varepsilon)\| \leqq$ Const. $(e^{-\nu_0(\tau-\tau_0)} + \varepsilon)$, and $|\eta(\tau, \varepsilon)P\eta^{-1}(u, \varepsilon)| \leqq K e^{-\nu(\tau-u)}$ for $\tau \geqq u$, we find that

$$S\left(\frac{1}{\varepsilon}, \varepsilon\right) = \int_{1/\varepsilon - 1/\nu \ln 1/\varepsilon}^{1/\varepsilon} \eta\left(\frac{1}{\varepsilon}, \varepsilon\right)\eta^{-1}(u, \varepsilon)\, du + O(\varepsilon). \tag{4.15}$$

Since $(d/du)\eta^{-1} = -\eta^{-1}(u, \varepsilon)B(u, \varepsilon)$ and $B(u, \varepsilon) = F(1, X_0(1)) + O(\varepsilon \ln 1/\varepsilon)$ for $1/\varepsilon - 1/\nu \ln 1/\varepsilon \leqq u \leqq 1/\varepsilon$, it follows that $S(1/\varepsilon, \varepsilon) = -F(1, X_0(1)) + O(\varepsilon(\ln 1/\varepsilon)^2)$. Assumption 1 implies that $F(1, X_0(1))$ is nonsingular and hence, for all sufficiently small ε, $S(1/\varepsilon, \varepsilon)$ is also nonsingular.

We may now use $Z(\tau, \varepsilon)$ to construct a fundamental solution $\hat{Z}_1(\tau, \varepsilon)$ to the linear system

$$\frac{d\hat{Z}_1}{d\tau} = \begin{pmatrix} 0 & I_n \\ \hat{A}(\tau) & \hat{B}(\tau) \end{pmatrix} \hat{Z}_1, \qquad 0 \leqq \tau \leqq \kappa \ln \frac{1}{\varepsilon}, \tag{4.16}$$

κ a sufficiently large positive constant, satisfying the dichotomy

$$\begin{aligned} &\|\hat{Z}_1(\tau, \varepsilon)P\hat{Z}_1^{-1}(u, \varepsilon)\| \leqq K_4 e^{-\nu_4(\tau-u)}, \qquad u \leqq \tau \leqq \kappa \ln 1/\varepsilon, \\ &\|\hat{Z}_1(\tau, \varepsilon)(I-P)\hat{Z}_1^{-1}(u, \varepsilon)\| \leqq K_4, \qquad \tau \leqq u \leqq \kappa \ln 1/\varepsilon, \end{aligned} \tag{4.17}$$

where K_4 and ν_4 are positive constants, and the estimate

$$\hat{Z}_1(0, \varepsilon)P\hat{Z}^{-1}(0, \varepsilon) = Z(0, \varepsilon)PZ^{-1}(0, \varepsilon) + O(\varepsilon(\ln 1/\varepsilon)^2). \tag{4.18}$$

The proof follows from the arguments used in Lemma 6.3 of [11], Proposition 2 of [5], and the estimate $\|\varepsilon A(\varepsilon\tau, \varepsilon) - \hat{A}(\tau)\|$, $\|\hat{B}(\tau) - \hat{B}(\tau)\| \leqq$ Const. $(\varepsilon \ln 1/\varepsilon)$ for $0 \leqq \tau \leqq \kappa \ln 1/\varepsilon$.

We may now apply the arguments used in formulas (6.42)-(6.54) of Lemma 6.4 of [11] to conclude that there exists a bounded matrix $J(\varepsilon)$, with bounded inverse, such that

$$Z(0, \varepsilon)PZ^{-1}(0, \varepsilon) = P_1 J(\varepsilon) + O(\varepsilon(\ln 1/\varepsilon)^2). \tag{4.19}$$

Using the fundamental solution $\tilde{Z}(t, \varepsilon) := Z(t/\varepsilon, \varepsilon)$ we may write the problem (4.3)-(4.4) as the following integral equation:

$$\begin{aligned} \begin{pmatrix} \tilde{x}(t, \varepsilon) \\ \tilde{y}(t, \varepsilon) \end{pmatrix} &= \tilde{Z}(t, \varepsilon)PZ^{-1}(0, \varepsilon)C_L(\tilde{x}, \tilde{y}, \varepsilon) + \tilde{Z}(t, \varepsilon)(I-P)\tilde{Z}^{-1}(1, \varepsilon)C_R(\tilde{x}, \tilde{y}, \varepsilon) \\ &+ \int_0^t \tilde{Z}(t, \varepsilon)P\tilde{Z}^{-1}(s, \varepsilon)\begin{pmatrix} 0 \\ H(s, \tilde{x}(s, \varepsilon), (1/\varepsilon)\tilde{y}(s, \varepsilon), \varepsilon), \varepsilon) + \rho(s, \varepsilon) \end{pmatrix} ds \\ &- \int_t^1 \tilde{Z}(t, \varepsilon)(I-P)\tilde{Z}^{-1}(s, \varepsilon)\begin{pmatrix} 0 \\ H(s, \tilde{x}(s, \varepsilon), (1/\varepsilon)\tilde{y}(s, \varepsilon), \varepsilon) + \rho(x, \varepsilon) \end{pmatrix} ds, \end{aligned} \tag{4.20}$$

where $C_L(\tilde{x}, \tilde{y}, \varepsilon)$ and $C_R(\tilde{x}, \tilde{y}, \varepsilon)$ are determined by the boundary conditions. Imposing the boundary conditions we find that C_L and C_R must satisfy the following linear system:

$$M_L(\varepsilon)C_L(\tilde{x}, \tilde{y}, \varepsilon)+M_R(\varepsilon)C_R(\tilde{x}, \tilde{y}, \varepsilon)=\gamma(\tilde{x}, \tilde{y}, \varepsilon), \tag{4.21}$$

where

$$\begin{aligned} M_L(\varepsilon) &:= L\tilde{Z}(0, \varepsilon)P\tilde{Z}^{-1}(0, \varepsilon)+R\tilde{Z}(1, \varepsilon)P\tilde{Z}^{-1}(0, \varepsilon), \\ M_R(\varepsilon) &:= L\tilde{Z}(0, \varepsilon)(I-P)\tilde{Z}^{-1}(1, \varepsilon)+R\tilde{Z}(1, \varepsilon)(I-P)\tilde{Z}^{-1}(1, \varepsilon), \end{aligned} \tag{4.22}$$

and

$$\begin{aligned} \gamma(\tilde{x}, \tilde{y}, \varepsilon) := &\begin{pmatrix} \phi(\varepsilon) \\ \psi(\varepsilon) \end{pmatrix} + L\int_0^1 \tilde{Z}(0, \varepsilon)(I-P)\tilde{Z}^{-1}(s, \varepsilon) \\ &\cdot \begin{pmatrix} 0 \\ H(s, \tilde{x}(s, \varepsilon), (1/\varepsilon)\tilde{y}(s, \varepsilon), \varepsilon)+\rho(s, \varepsilon) \end{pmatrix} ds \\ &-R\int_0^1 \tilde{Z}(1, \varepsilon)P\tilde{Z}^{-1}(s, \varepsilon)\begin{pmatrix} 0 \\ H(s, \tilde{x}(s, \varepsilon), (1/\varepsilon)\tilde{y}(s, \varepsilon), \varepsilon)+\rho(s, \varepsilon) \end{pmatrix} ds. \end{aligned} \tag{4.23}$$

Since $|R\tilde{Z}(1, \varepsilon)P\tilde{Z}^{-1}(0, \varepsilon)| \leqq \text{Const. } e^{(\nu/\varepsilon)}$, the columns of $(I_n \quad 0)\tilde{Z}(0, \varepsilon)P\tilde{Z}^{-1}(0, \varepsilon) = (I_n \quad 0)P_1J(\varepsilon)+O(\varepsilon(\ln(1/\varepsilon)^2)$ span R^n (see Assumption 3 and (4.19)), and

$$\begin{aligned} &R\tilde{Z}(1, \varepsilon)(I-P)\tilde{Z}^{-1}(1, \varepsilon) \\ &= \begin{pmatrix} 0 & 0 \\ -F(1, X_0(1))T(1/\varepsilon, \varepsilon)T(1/\varepsilon, \varepsilon) & F(1, X_0(1)) \end{pmatrix} + O\left(\varepsilon\left(\ln\frac{1}{\varepsilon}\right)^2\right), \end{aligned} \tag{4.24}$$

it follows that for sufficiently small ε there exists a solution C_L and C_R to the above linear system. We may now apply the Banach-Picard fixed point theorem to conclude that there exists a fixed point to the integral equation and hence a solution to the problem (1.1)-(1.2) satisfying the estimates of (4.1). For the details of such a proof, including a discussion of local uniqueness, see [11, pp. 26-30].

REFERENCES

[1] K. W. CHANG, *Diagonalization method for a vector boundary value problem of singular perturbation type*, J. Math. Anal. Appl., 48 (1974), pp. 652-665.

[2] ———, *Diagonalization methods in singular perturbations*, in International Conference on Differential Equations, H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 164-184.

[3] ———, *Singular perturbations of a boundary value problem for a vector second-order differential equation*, SIAM J. Appl. Math., 30 (1976), pp. 42-54.

[4] K. W. CHANG AND F. A. HOWES, *Nonlinear Singular Perturbation Phenomena: Theory and Applications* Springer-Verlag, Berlin, New York, 1984.

[5] W. A. COPPEL, *Dichotomies in Stability Theory*, Lecture Notes in Math. 629, Springer-Verlag, Berlin, New York, 1978.

[6] J. E. FLAHERTY AND R. E. O'MALLEY, JR., *On the numerical integration of two-point boundary value problems for stiff systems of ordinary equations*, in Boundary and Interior Layers: Computational and Asymptotic Methods, J. J. H. Miller, ed., Boole Press, Dublin, 1980, pp. 93-102.

[7] M. I. FREEDMAN AND J. L. KAPLAN, *Singular perturbations of two-point boundary value problems arising in optimal control theory*, SIAM J. Control Optim., 14 (1976), pp. 189-215.

[8] P. HABETS, *Singular perturbations of a vector boundary value problem*, Lecture Notes in Math. 415, Springer-Verlag, Berlin, New York, 1974, pp. 149-154.

[9] C. R. HADLOCK, *Existence and dependence on a parameter of solutions of a nonlinear two point boundary value problem*, J. Differential Equations, 14 (1973), pp. 498-517.

[10] F. A. HOWES AND R. E. O'MALLEY, JR., *Singular perturbations of semilinear second-order systems*, Lecture Notes in Math., 827, Springer-Verlag, Berlin, New York (1980), pp. 130–150.

[11] J. S. JEFFRIES AND D. R. SMITH, *A Green function approach for a singularly perturbed vector boundary-value problem*, Adv. in Appl. Math., 10 (1989), pp. 1–50.

[12] ———, *The Dirichlet problem for a quasilinear singularly perturbed second-order system*, J. Math. Anal. Appl., to appear.

[13] J. S. JEFFRIES, *Boundary Layer and Shock Layer Solutions to Singularly Perturbed Boundary Value Problems*, Ph.D. thesis, University of California, San Diego, CA, 1987.

[14] W. G. KELLEY, *Existence and uniqueness of solutions for vector problems containing small parameters*, J. Math. Anal. Appl., 131 (1988), pp. 295–312.

[15] S. K. KIRSCHVINK, *Differential Inequalities and Singularly Perturbed Boundary Value Problems*, Ph.D. thesis, University of California, San Diego, CA, 1987.

[16] M. A. O'DONNELL, *Boundary and Interior Layer Behavior in Singularly Perturbed Systems of Boundary Value Problems*, Ph.D. thesis, University of California, Davis, CA, 1983.

[17] R. E. O'MALLEY, JR., *Shock and transition layers for singularly perturbed second-order vector systems*, SIAM J. Appl. Math., 43 (1983), pp. 935–943.

[18] D. R. SMITH, *Singular Perturbation Theory*, Cambridge University Press, London, 1985.

[19] ———, *Single-layer solutions for the Dirichlet problem for a quasilinear singularly perturbed second-order system*, Rocky Mountain J. Math., 18 (1987), pp. 67–103.

SIAM J. MATH ANAL.
Vol. 22, No. 4, pp. 905–933, July 1991

003

EXISTENCE, UNIQUENESS, AND CONTINUOUS DEPENDENCE FOR A SYSTEM OF HYPERBOLIC CONSERVATION LAWS MODELING POLYMER FLOODING*

ASLAK TVEITO† AND RAGNAR WINTHER†

Abstract. The problem of well-posedness for a system of nonstrictly hyperbolic conservation laws is studied. A finite difference scheme is used to prove the existence of an entropy solution with bounded variation. It is proved that the entropy solution of the system is unique, and that the solution depends continuously on its initial data in a proper topology. The analysis is based on a smoothness property of one of the Riemann invariants of the system.

Key words. conservation laws, existence, uniqueness, continuous dependence

AMS(MOS) subject classifications. 65M10, 35L65

1. Introduction.

The purpose of this paper is to study the Cauchy problem for the hyperbolic system of conservation laws

$$\begin{aligned} s_t + f(s,c)_x &= 0, \\ (sc)_t + (c(f(s,c))_x &= 0. \end{aligned} \tag{1.1}$$

This model arises in enhanced oil recovery when oil is displaced in a porous rock by water containing dissolved polymer. The variable s denotes the saturation of the aqueous phase, consisting of water and polymer, while c denotes the concentration of polymer in the aqueous phase. The function $f = f(s,c)$ is usually referred to as the fractional flow function. For a discussion of the application of (1.1) as a model for polymer flooding, we refer to Pope [21], Isaacson [5], and Johansen and Winther [7].

The characteristic speeds of the system (1.1) are given by f_s and f/s. The system will be analyzed in regimes where the ordering of the characteristic speeds depends on the location in the state space. The hyperbolic problem (1.1) is therefore not strictly hyperbolic. Furthermore, the characteristic field associated with the speed f/s is linearly degenerate.

Riemann problems for the nonstrictly hyperbolic system (1.1) have been analyzed by Keyfitz and Kranzer [8] and Isaacson [5]. Temple [25] used these results to establish existence of a solution for the Cauchy problem by the random choice method. A parabolic regularization of the model (1.1) is studied in [29]. In [28], existence of a solution to the Cauchy problem was proved by applying finite difference approximations. However, the analysis given in [28] was limited by rather restrictive assumptions on the fractional flow function f. In particular these assumptions implied that the system was strictly hyperbolic, and they excluded physically relevant fractional flow functions. Generalizations of the problem (1.1) were studied by Serre [23].

It was observed in [28] that if the variable c is smooth initially, then c remains smooth for all time. This property is due to the fact that one of the characteristic fields is noninteracting, i.e., it does not generate discontinuities. This observation should be compared with the result of Keyfitz and Kranzer [9] that global smooth solutions of (1.1) can occur only if the variable f/s is constant for the initial data.

* Received by the editors February 5, 1990; accepted for publication (in revised form) August 9, 1990.

† Department of Informatics, P. O. Box 1080 Blindern, University of Oslo, Norway. This research has been supported by VISTA, a research cooperation between the Norwegian Academy of Science and Letters and Den norske stats oljeselskap a.s. (Statoil).

Hence, in general, discontinuities will develop, but in such a way that the variable c remains smooth.

In the present paper we significantly generalize the theory developed in [28]. The results derived in [28] is based on a simple form of the fractional flow function. In fact, for the model studied in [28], we could formulate the finite difference scheme equivalently in conservative variables and in the Riemann invariants. This simplified the derivation of the bounds on the approximate solutions. In the present paper we generalize the results to physically relevant fractional flow functions, for which we cannot formulate the finite difference scheme in Riemann invariants. In fact, since the system is nonstrictly hyperbolic, the Riemann invariants do not constitute a global coordinate system in the state-space. We also generalize the analysis presented in [28] by including uniqueness and continuous dependence results.

As in [28] the smoothness property of c is utilized. Existence of an entropy solution is derived under rather general assumptions on the fractional flow functions f. The existence argument uses finite difference approximations derived from the nonconservative form

$$\begin{aligned} s_t + f(s,c)_x &= 0, \\ c_t + g(s,c)c_x &= 0 \end{aligned} \tag{1.2}$$

of the system (1.1). Here $g = g(s,c)$ denotes the function $f(s,c)/s$. In order to prove convergence in L^1_{loc} of a subsequence the family (s_Δ, c_Δ) of approximate solutions generated by the finite difference scheme, it is sufficient to establish three estimates: (1) a uniform L^∞ bound, (2) a uniform total variation bound, and (3) L^1-continuity in time of the approximate solution. From these estimates we obtain convergence of a subsequence of the family (s_Δ, c_Δ) in L^1_{loc} (cf. Oleinik [19] or Smoller [24]). Glimm [4] applied this strategy in his famous existence proof for strictly hyperbolic systems of conservation laws with "small" data. He established the proper estimates for his random choice method. This scheme has also been applied by other authors in order to prove existence of weak solutions for special systems. In [15] LeVeque and Temple used this strategy to prove convergence of the Godunov scheme for strictly hyperbolic systems of conservation laws with "line fields." Serre [22] also studied such systems and proved convergence of the Godunov scheme, the random choice method, and the Lax–Friedrichs scheme.

For the present model, the estimates (1) and (3) are straightforward, whereas (2) depends strongly on the smoothness property of c_Δ derived in [28]. Having established the convergence, proving existence of an entropy solution is a matter of verifying that the limit satisfies the proper entropy condition. We remark that, to the author's knowledge, this is the first convergence result for a classical finite difference scheme applied to a nonstrictly hyperbolic system of conservation laws. Convergence of the random choice method for such system has been proved by Temple [25] and Liu and Wang [18].

However, the main contribution of this paper is a proof of uniqueness and continuous dependence results for the system (1.1). We show that, if the initial saturation s is a function of bounded variation and if the initial concentration c is sufficiently smooth, the entropy solution is unique. Furthermore, the solution depends continuously on the initial data in the proper topologies. These results should be compared with an observation done by Isaacson and Temple [6]. They observed, by combining two Riemann problem solutions, that if both the variables s and c are allowed to be discontinuous initially, then the solution would not depend continuously on the initial

data in L^1. Hence, a smoothness condition on c seems to be necessary in order to obtain a well-posed Cauchy problem.

We remark that, as a consequence of the uniqueness result, we obtain convergence of the entire family of approximate solutions, not only a subsequence of it. We hope to be able to investigate this convergence, with an eye to error estimates, in the future. We also remark that the method of proving the smoothness property of the c variable used in this paper also formally applies to more general systems of the form (1.2). The function g may depend on x, t, and other quantities in a model.

Our uniqueness and continuous dependence results are derived by introducing a "Kruzkov-form" for the saturation equation of (1.2). The results obtained from this form are combined with results obtained with the method of characteristics for the concentration equation. We recall that the "Kruzkov-form" for scalar conservation laws was introduced by Kruzkov [10] and used in [10] and Kuznetsov [11] in order to establish uniqueness and continuous dependence results for such problems. For systems of hyperbolic conservation laws there are, however, very few results addressing the question of uniqueness and continuous dependence. Some uniqueness results for rather general systems are derived by DiPerna [2] and Liu [16], and a uniqueness result for the p-system in gas dynamics is established by Oleinik (cf. Smoller [24]). Stability of the constant state for general systems is studied by Temple [26], [27].

The precise assumptions on the model and the main results are stated in §2. In §3 we derive the desired properties of the finite difference approximations, while the convergence arguments are given in §4. The uniqueness and continuous dependence results are proved in §5.

2. Preliminaries and statement of the main result. We begin by introducing some notation. If $\mathcal{K}$ is a domain in $\mathbb{R}$, then $L^p(\mathcal{K})$, $1 \le p \le \infty$, will denote the classical L^p spaces of real valued functions on $\mathcal{K}$. Instead of $L^p(\mathbb{R})$, we will simply write L^p, and the norms on L^p are denoted by $\|\cdot\|_p$. The localized versions of L^p, consisting of functions on $\mathbb{R}$ which are in $L^p(\mathcal{K})$ for any compact subset $\mathcal{K}$ of $\mathbb{R}$, are denoted by L^p_{loc}.

Furthermore, $BV = BV(\mathbb{R})$ denotes the subspace of L^1_{loc} consisting of functions with bounded variation; i.e.,

$$BV = \{v \in L^1_{\text{loc}} : TV(v) < \infty\},$$

where

$$TV(v) = \sup_{h \neq 0} \int_{\mathbb{R}} \frac{|v(x+h) - v(x)|}{|h|}\, dx.$$

The class of Lipschitz continuous functions on a domain $\mathcal{K} \subset \mathbb{R}$ is denoted by $\text{Lip}(\mathcal{K})$. More precisely

$$\text{Lip}(\mathcal{K}) = \{v \in L^\infty(\mathcal{K}) : \|v\|_{\text{Lip}(\mathcal{K})} < \infty\},$$

where

$$\|v\|_{\text{Lip}(\mathcal{K})} = \sup_{x \neq y} \frac{|v(x) - v(y)|}{|x - y|}.$$

As above, we will simply write Lip instead of $\text{Lip}(\mathbb{R})$.

If u is a function on a grid $\Delta = \{j\Delta x\}_{j\in\mathbb{Z}}$, with values u_j at the grid points $j\Delta x$, we define an associated piecewise constant function u_Δ on $\mathbb{R}$ by

$$u_\Delta(x) = u_j \quad \text{for } x \in [(j-\tfrac{1}{2})\Delta x, (j+\tfrac{1}{2})\Delta x).$$

The "discrete" L^1- and L^∞-norms of u are equal to the corresponding "continuous" norms; i.e.,

$$\|u\|_1 \equiv \Delta x \sum_{j\in\mathbb{Z}} |u_j| = \|u_\Delta\|_1,$$
$$\|u\|_\infty \equiv \sup_{j\in\mathbb{Z}} |u_j| = \|u_\Delta\|_\infty.$$

For a grid function u, we also let $TV(u)$ denote the associated "discrete" total variation given by

$$TV(u) \equiv \sum_{j\in\mathbb{Z}} |u_j - u_{j-1}|.$$

It can easily be seen (cf. [1]) that $TV(u)$ is finite if and only if $u_\Delta \in BV$, and that

$$TV(u) = TV(u_\Delta).$$

Consider now the flux-function $f = f(s,c)$ which defines our model (1.2). It will be assumed throughout the paper that there exists an S_0, $0 < S_0 < 1$, such that

$$f(S_0, c) \equiv 0 \quad \forall c \in [0,1]. \tag{2.1}$$

Furthermore we assume that

$$f(1, c) \equiv 1 \quad \forall c \in [0,1] \tag{2.2}$$

(cf. Fig. 1). The associated state-space $\mathcal{S}$ of the model is given by

$$\mathcal{S} = [S_0, 1] \times [0,1].$$

In addition to (2.1) and (2.2), we will also assume that $f(s,c)$ is an increasing and smooth function of s for all $c \in [0,1]$. In order to simplify our notation we introduce a positive constant M_f which bounds all the partial derivatives of f of order 1 or 2; i.e., we assume that $M_f > 0$ satisfies

$$0 \le f_s, |f_c|, |f_{ss}|, |f_{sc}|, |f_{cc}| \le M_f \tag{2.3}$$

for all $(s,c) \in \mathcal{S}$.

We emphasize that no assumptions are made concerning the sign of f_c or f_{ss}. Hence, the conditions here are more general than what has been assumed in some earlier papers on the model (1.2) (cf. [25], [5]).

The initial functions (s^0, c^0) of (1.2) are assumed to be given such that $(s^0(x), c^0(x)) \in \mathcal{S}$ for all $x \in \mathbb{R}$. In particular, this implies that $s^0(x) \ge S_0 > 0$ for all $x \in \mathbb{R}$. In addition we will assume that c^0 is a smooth function in the sense that $c^0 \in \text{Lip} \cap BV$. We recall that this assumption implies that c^0_x exists almost everywhere and, since c^0 is continuous, $c^0_x \in L^1$. Hence, for any bounded interval $\mathcal{K}$ of $\mathbb{R}$, the restriction of c^0 is an element of the Sobolev space $W^{1,1}(\mathcal{K})$ (cf. [30]).

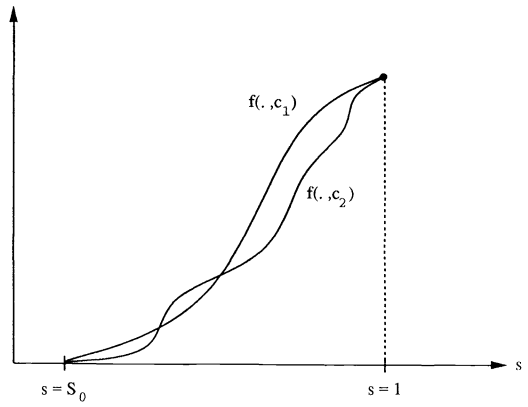


FIG. 1. *A fractional flow function f as a function of s for two values of c.*

Finally, it will be assumed that the functions s^0 and c_x^0 are elements of BV.

To be able to state the assumptions on the initial functions in a compact manner we introduce a subset $\mathcal{B}^0$ of $(L^\infty)^2$. A pair of functions (u, v) is said to be in the class $\mathcal{B}^0$ if

$$\begin{aligned} &\text{(a)} \quad (u(x), v(x)) \in \mathcal{S} \quad \forall x \in \mathbb{R}, \\ &\text{(b)} \quad v \in \mathrm{Lip} \cap BV, \\ &\text{(c)} \quad u, v_x \in BV. \end{aligned} \tag{2.4}$$

With this notation our assumptions on the functions s^0, c^0 can simply be written $(s^0, c^0) \in \mathcal{B}^0$. We also introduce a class $\mathcal{B}$ consisting of functions of two variables x and t; a pair of functions (u, v) is said to be in the class $\mathcal{B}$ if

$$\begin{aligned} &\text{(a)} \quad (u(\cdot, t), v(\cdot, t)) \in \mathcal{B}^0 \text{ for } t \in [0, T_0], \\ &\text{(b)} \quad \|u(\cdot, t) - u(\cdot, \tau)\|_1 + \|v(\cdot, t) - v(\cdot, \tau)\|_1 \le K|t - \tau| \\ &\qquad \text{for } 0 \le t, \tau \le T_0 < \infty, \\ &\text{(c)} \quad v(x, \cdot) \in \mathrm{Lip}[0, T_0] \text{ for } x \in \mathbb{R}. \end{aligned} \tag{2.5}$$

Here K is a finite constant. Throughout the paper T_0 will denote a fixed finite time. The main reason for introducing this notation is that we will show that

$$(s^0, c^0) \in \mathcal{B}^0 \Longrightarrow (s, c) \in \mathcal{B}.$$

Establishing this fundamental property of the solution of (1.2) is an essential part of our existence argument presented below.

Let us now turn to the precise formulation of the initial value problem for the model (1.2). Since we are working with a nonlinear system of hyperbolic conservation

laws, we expect that discontinuities in the solutions will occur. As indicated above, this is only the case for the saturation function s. The concentration function c will remain Lipschitz continuous for all time. Since c is continuous, the shocks in s can be considered, locally in space and time, as a shock of a scalar conservation law. We will therefore require these shocks to satisfy the proper generalization of a scalar entropy condition. This condition is formulated weakly by a modification of the Kruzkov form (cf. [10]).

DEFINITION 1. Let $(s^0, c^0) \in \mathcal{B}^0$ be given. A pair of functions (s, c) is called an entropy solution of (1.2) if they satisfy the following requirements:

1. $(s, c) \in \mathcal{B}$.

2. For all nonnegative C^∞-functions ϕ with compact support in $\mathbb{R} \times [0, T_0]$, all $q \in [S_0, 1]$ and all $T \in [0, T_0]$,

$$\int_0^T \int_{\mathbb{R}} \{|s-q|\phi_t + \operatorname{sign}(s-q)(f(s,c) - f(q,c))\phi_x - \operatorname{sign}(s-q) f(q,c)_x \phi\}\, dx\, dt$$
$$+ \int_{\mathbb{R}} |s^0(x) - q|\phi(x,0)\, dx - \int_{\mathbb{R}} |s(x,T) - q|\phi(x,T)\, dx \geq 0.$$

3. For almost all $(x, t) \in \mathbb{R} \times [0, T_0]$, $(s(x,t), c(x,t))$ satisfies

$$c_t + g(s,c)c_x = 0.$$

Furthermore,

$$\lim_{t \to 0^+} \|c(\cdot, t) - c^0\|_1 = 0.$$

This solution concept is a combination of a weak and a classical formulation. The inequality 2 is a generalization of the Kruzkov form (cf. [10]) for a scalar conservation law

$$s_t + f(s,c)_x = 0,$$

when c is a *given* function of x and t, while 3 is a classical formulation of the second equation of (1.2).

By using a finite difference approximation, we shall be able to prove the following existence result for the model (1.2).

THEOREM 2.1. *For any pair of initial functions $(s^0, c^0) \in \mathcal{B}^0$, there exists an entropy solution of* (1.2).

By assuming some extra regularity on the initial functions, we shall also prove that the entropy solution is unique. Recall that for $(s, c) \in \mathcal{S}$ we have $0 < S_0 \leq s \leq 1$; hence the function $k = c_x/s$ is well defined whenever c_x is. We also define $k^0 = c_x^0/s^0$. Assume that the initial data (s^0, c^0) satisfies the following requirements:

$$\text{(2.6)} \qquad \begin{aligned} &\text{(a)} \quad (s^0, c^0) \in \mathcal{B}^0, \\ &\text{(b)} \quad (s^0(x), c^0(x)) = \begin{cases} (s^L, c^L) & x \leq L, \\ (s^R, c^R) & x \geq R, \end{cases} \\ &\text{(c)} \quad k_x^0 \in L^\infty, \end{aligned}$$

where s^L, s^R, c^L, s^R, L, and R are finite constants. Let $(\bar{s}^0, \bar{c}^0)$ be another pair of initial functions satisfying (2.6) with the same constants; then we have the following uniqueness and continuous dependence result.

THEOREM 2.2. *Let (s^0, c^0) and $(\bar{s}^0, \bar{c}^0)$ be two pairs of initial data as described above, and let (s, c) and $(\bar{s}, \bar{c})$ be the corresponding entropy solutions of (1.2), coinciding for $|x|$ sufficiently large. Then there is a finite constant M such that*

$$\|s(\cdot,t) - \bar{s}(\cdot,t)\|_1 + \|c(\cdot,t) - \bar{c}(\cdot,t)\|_1 + \|c_x(\cdot,t) - \bar{c}_x(\cdot,t)\|_1 \leq M(\|s^0 - \bar{s}^0\|_1 + \|c^0 - \bar{c}^0\|_1 + \|c_x^0 - \bar{c}_x^0\|_1)$$

for $t \in [0, T_0]$.

This theorem shows that the system is well posed in the proper topology. We remark that, in view of the example given by Isaacson and Temple referred to above, the system is not well posed in the L^1-norm applied to the variables (s, c).

Theorem 2.1 will be proved in §4, while Theorem 2.2 will be proved in §5.

3. Properties of the approximate solutions. In order to prove the existence of a solution to the system (1.2), we generate a family (s_Δ, c_Δ) of approximate solutions to the system and prove that a subsequence of this family converges in L^1_{loc} and that the limit is a solution. The family of approximate solutions is generated by a nonconservative finite difference scheme (cf. [28]). The purpose of this section is to establish some bounds on these approximate solutions. First we prove that the approximate solutions remain in the state-space $\mathcal{S}$, and that the total variation of c_Δ is nonincreasing in time. Then, following [28], we show that c_Δ has a certain smoothness property. All these properties of the approximate solutions are used to prove that the total variation of s_Δ remains bounded for all finite time. As a corollary to the TV-estimates we establish the usual L^1-continuity in time for the approximate solutions. Finally we establish a discrete entropy condition.

Let Δx and Δt be the meshsize in space and time, respectively, and let (s_j^n, c_j^n) denote an approximation to $(s(j\Delta x, n\Delta t), c(j\Delta x, n\Delta t))$ for all $(j, n) \in \mathbb{Z} \times \mathbb{Z}^+$. The approximations are generated by the following finite difference scheme:

$$\begin{aligned} s_j^{n+1} &= s_j^n - \mu(f_j^n - f_{j-1}^n), \\ c_j^{n+1} &= c_j^n - \mu g_j^n (c_j^n - c_{j-1}^n), \end{aligned} \tag{3.1}$$

where $\mu = \Delta t/\Delta x$, $f_j^n = f(s_j^n, c_j^n)$, and $g_j^n = g(s_j^n, c_j^n)$. Recall that $g(s, c) = f(s, c)/s$. We assume that the mesh parameters satisfies the following CFL-condition:

$$\frac{\Delta t}{\Delta x} \sup_{(s,c)\in\mathcal{S}} \left(\frac{\partial f}{\partial s}, g \right) \leq 1. \tag{3.2}$$

The iteration is started by putting

$$s_j^0 = \frac{1}{\Delta x} \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} s^0(x)\, dx \quad \text{and} \quad c_j^0 = \frac{1}{\Delta x} \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} c^0(x)\, dx.$$

The family of approximate solutions (s_Δ, c_Δ) is defined by extending the finite difference solutions $\{(s_j^n, c_j^n)\}$ to a function on $\mathbb{R} \times [0, T_0]$,

$$\begin{aligned} &(s_\Delta(x,t), c_\Delta(x,t)) = (s_j^n, c_j^n) \\ &\qquad \text{for } (x,t) \in [(j - \tfrac{1}{2})\Delta x, (j + \tfrac{1}{2})\Delta x) \times [n\Delta t, (n+1)\Delta t). \end{aligned} \tag{3.3}$$

We start by showing that the finite difference solutions remain in the state-space, and that the total variation of c is nonincreasing as a function of time.

LEMMA 3.1. *Suppose that* $(s_j^0, c_j^0) \in \mathcal{S} \quad \forall j \in \mathbb{Z}$ *and that* $TV(c^0) < \infty$ *then*

(a) $(s_j^n, c_j^n) \in \mathcal{S} \quad \forall (j,n) \in \mathbb{Z} \times \mathbb{Z}^+$,

(b) $\min_i c_i^0 \leq c_j^n \leq \max_i c_i^0 \quad \forall (j,n) \in \mathbb{Z} \times \mathbb{Z}^+$,

(c) $TV(c^n) \leq TV(c^0) \quad \forall n \in \mathbb{Z}^+$.

Proof. Assume, for a fixed n, that $(s_j^n, c_j^n) \in \mathcal{S}$ for all $j \in \mathbb{Z}$. Since

$$c_j^{n+1} = (1 - \mu g_j^n) c_j^n + \mu g_j^n c_{j-1}^n,$$

it follows by the CFL-condition (3.2) that c_j^{n+1} is a convex combination of c_j^n and c_{j-1}^n. Hence

$$\min_i c_i^n \leq c_j^{n+1} \leq \max_i c_i^n \quad \forall j \in \mathbb{Z},$$

and

$$TV(c^{n+1}) \leq TV(c^n).$$

It remains to prove that

$$s_j^{n+1} \in [S_0, 1] \quad \forall j \in \mathbb{Z}. \tag{3.4}$$

Define the function $\sigma = \sigma(s^L, s^R, c^L, c^R)$ by

$$\sigma(s^L, s^R, c^L, c^R) = s^R - \mu(f(s^R, c^R) - f(s^L, c^L))$$

for $(s^L, c^L), (s^R, c^R) \in \mathcal{S}$. Then, by the CFL-condition (3.2) and the properties of f, we obtain

$$\frac{\partial \sigma}{\partial s^L} = \mu f_s(s^L, c^L) \geq 0,$$
$$\frac{\partial \sigma}{\partial s^R} = 1 - \mu f_s(s^R, c^R) \geq 0.$$

Consequently,

$$\sigma(s^L, s^R, c^L, c^R) \leq \sigma(1, 1, c^L, c^R) = 1,$$
$$\sigma(s^L, s^R, c^L, c^R) \geq \sigma(S_0, S_0, c^L, c^R) = S_0.$$

Since $s_j^{n+1} = \sigma(s_{j-1}^n, s_j^n, c_{j-1}^n, c_j^n)$, we have established (3.4), and the proof of the lemma is completed by induction on n. □

In [28] we established a smoothness property for the variable corresponding to our c-variable. In the present paper this property is of fundamental importance both for the existence and the uniqueness of a solution to the Cauchy problem. We will therefore, for completeness, briefly review the discussion of smoothness here.

Let us start by deriving this regularity in the case of classical solutions. We assume that (s, c) is a smooth solution of the system

$$\begin{aligned} s_t + (gs)_x &= 0, \\ c_t + gc_x &= 0, \end{aligned} \tag{3.5}$$

and that $(s(x,t), c(x,t)) \in \mathcal{S}$ for all $(x,t) \in \mathbb{R} \times [0, T_0]$. By differentiating the second equation of (3.5) with respect to x, we obtain the equation

$$(c_x)_t + (g\, c_x)_x = 0. \tag{3.6}$$

Now, let

$$k = c_x/s,$$

and recall that for $(s,c) \in \mathcal{S}$, we have $0 < S_0 \le s \le 1$. Then, by using (3.6), we get

$$0 = (ks)_t + (gks)_x = k(s_t + (gs)_x) + s(k_t + g\, k_x),$$

and then, by the first equation of (3.5), we obtain

$$k_t + g\, k_x = 0. \tag{3.7}$$

This equation is easily seen to satisfy the following maximum principle:

$$\|k(\cdot,t)\|_\infty \le \|k^0\|_\infty,$$

and consequently,

$$\|c_x(\cdot,t)\|_\infty \le \|c_x^0\|_\infty / S_0. \tag{3.8}$$

Thus we have established a uniform bound on the spatial derivative of the c-variable under the assumption of classical solutions. However, this is merely formalism since the solution of the system in general only exists in a weak sense, but these calculations clearly motivate a similar result for the approximate solutions. Using the regularity of the approximate solutions we prove, rigorously, that the function c obtained as the limit of c_Δ is Lipschitz continuous.

We remark that the method of deriving a uniform bound on the spatial derivative of c for smooth solutions outlined above is applicable for any systems that can be written on the form (3.5). It is not necessary for the function g to be a function of s and c only; it might also depend explicitly on x, t, and other quantities in a model.

To prove the regularity property for the approximate solutions, we introduce an auxiliary sequence

$$k_j^n = \frac{c_{j+1}^n - c_j^n}{\Delta x\, s_{j+1}^n} \quad \forall (j,n) \in \mathbb{Z} \times \mathbb{Z}^+. \tag{3.9}$$

Again we recall that $0 < S_0 \le s_j^n \le 1$ whenever $(s_j^n, c_j^n) \in \mathcal{S}$. We have the following result.

LEMMA 3.2. *Suppose that $(s_j^0, c_j^0) \in \mathcal{S}$ for all $j \in \mathbb{Z}$; then*

$$\left\| \frac{c_{j+1}^n - c_j^n}{\Delta x} \right\|_\infty \le \left\| \frac{c_{j+1}^0 - c_j^0}{\Delta x\, s_{j+1}^0} \right\|_\infty \le \frac{1}{S_0} \left\| \frac{c_{j+1}^0 - c_j^0}{\Delta x} \right\|_\infty \quad \forall n \in \mathbb{Z}^+.$$

Proof. We first observe that $(s_j^n, c_j^n) \in \mathcal{S}$ for all $(j,n) \in \mathbb{Z} \times \mathbb{Z}^+$ (cf. Lemma 3.1). By using the finite difference scheme (3.1) for (s_j^{n+1}, c_j^{n+1}), we obtain the following scheme for the auxiliary sequence $\{k_j^n\}$:

$$k_j^{n+1} = (1 - \mu g_{j+1}^n) \frac{s_{j+1}^n}{s_{j+1}^{n+1}} k_j^n + \mu g_j^n \frac{s_j^n}{s_{j+1}^{n+1}} k_{j-1}^n \tag{3.10}$$

(we refer to [28] for the details). By the CFL-condition (3.2), the coefficients of the scheme (3.10) are positive, and it follows from (3.1) that their sum is 1. Hence k_j^{n+1} is a convex combination of k_{j-1}^n and k_j^n, and then, by induction, we have

$$\|k^n\|_\infty \leq \|k^0\|_\infty.$$

By (3.9), this proves the lemma. □

From the proof above we collect the following results concerning the properties of the auxiliary sequence $\{k_j^n\}$.

COROLLARY 3.3. *Suppose that* $(s_j^0, c_j^0) \in \mathcal{S}\ \forall j \in \mathbb{Z}$, *and that* $TV(k^0) < \infty$, *then*

$$\text{(a)}\quad \|k^n\|_\infty \leq \|k^0\|_\infty \qquad \forall n \in \mathbb{Z}^+,$$
$$\text{(b)}\quad TV(k^n) \leq TV(k^0) \qquad \forall n \in \mathbb{Z}^+.$$

Both of these properties follow from the fact that k_j^{n+1} is a convex combination of k_{j-1}^n and k_j^n.

Next we prove that the total variation of s_Δ is bounded. We remark that this result depends strongly on the regularity of c_Δ.

LEMMA 3.4. *Suppose that* $(s_j^0, c_j^0) \in \mathcal{S}$ *for all* $j \in \mathbb{Z}$, *and that there is a finite constant* M_0 *such that*

$$TV(s^0), TV(k^0), TV(c^0), \|k^0\|_\infty \leq M_0.$$

Then

$$TV(s^n) \leq (2M_0 + \tfrac{1}{2})e^{2M_f M_0 n\Delta t},$$

where M_f *is defined by* (2.3).

Proof. From the finite difference scheme (3.1), we obtain

$$\begin{aligned} s_{j+1}^{n+1} - s_j^{n+1} &= s_{j+1}^n - s_j^n - \mu(f(s_{j+1}^n, c_{j+1}^n) - f(s_j^n, c_{j+1}^n) + f(s_j^n, c_{j+1}^n) - f(s_j^n, c_j^n)) \\ &\quad +\mu(f(s_j^n, c_j^n) - f(s_{j-1}^n, c_j^n) + f(s_{j-1}^n, c_j^n) - f(s_{j-1}^n, c_{j-1}^n)) \\ &= s_{j+1}^n - s_j^n - \mu f_s(s_{j+1/2}^n, c_{j+1}^n)(s_{j+1}^n - s_j^n) - \mu(f(s_j^n, c_{j+1}^n) - f(s_j^n, c_j^n)) \\ &\quad +\mu f_s(s_{j-1/2}^n, c_j^n)(s_j^n - s_{j-1}^n) + \mu(f(s_{j-1}^n, c_j^n) - f(s_{j-1}^n, c_{j-1}^n)), \end{aligned}$$

where $s_{j+1/2}^n \in \text{int}[s_j^n, s_{j+1}^n]$. Here $\text{int}[a, b]$ denotes the interval $[\min(a, b), \max(a, b)]$. By using a Taylor series expansion in the c-variable we get

$$\begin{aligned} TV(s^{n+1}) &= \sum_j |s_{j+1}^{n+1} - s_j^{n+1}| \\ &= \sum_j |(1 - \mu f_s(s_{j+1/2}^n, c_{j+1}^n))(s_{j+1}^n - s_j^n) + \mu f_s(s_{j-1/2}^n, c_j^n)(s_j^n - s_{j-1}^n) \\ &\quad -\mu f_c(s_j^n, c_j^n)(c_{j+1}^n - c_j^n) + \mu f_c(s_{j-1}^n, c_{j-1}^n)(c_j^n - c_{j-1}^n) \\ &\quad -\frac{\mu}{2} f_{cc}(s_j^n, c_{j+1/2}^n)(c_{j+1}^n - c_j^n)^2 + \frac{\mu}{2} f_{cc}(s_{j-1}^n, c_{j-1/2}^n)(c_j^n - c_{j-1}^n)^2|, \end{aligned}$$

where $c_{j+1/2}^n \in \text{int}[c_j^n, c_{j+1}^n]$. From the CFL-condition and the properties of f, we have

$$\begin{aligned} TV(s^{n+1}) &\leq TV(s^n) + \mu \sum_j |f_c(s_j^n, c_j^n)(c_{j+1}^n - c_j^n) - f_c(s_{j-1}^n, c_{j-1}^n)(c_j^n - c_{j-1}^n)| \\ &\quad +\mu M_f \|c_{j+1}^n - c_j^n\|_\infty \sum_j |c_{j+1}^n - c_j^n| = TV(s^n) + I + II. \end{aligned}$$

Here we bound I by using the properties of the auxiliary sequence $\{k_j^n\}$.

$$\begin{aligned} I &\equiv \frac{\Delta t}{\Delta x}\sum_j |f_c(s_j^n, c_j^n)(c_{j+1}^n - c_j^n) - f_c(s_{j-1}^n, c_{j-1}^n)(c_j^n - c_{j-1}^n)| \\ &= \Delta t \sum_j |f_c(s_j^n, c_j^n)k_j^n s_{j+1}^n - f_c(s_{j-1}^n, c_{j-1}^n)k_{j-1}^n s_j^n| \\ &= \Delta t \sum_j |k_j^n[f_c(s_j^n, c_j^n)(s_{j+1}^n - s_j^n) + s_j^n(f_c(s_j^n, c_j^n) - f_c(s_{j-1}^n, c_{j-1}^n))] \\ &\qquad + f_c(s_{j-1}^n, c_{j-1}^n)s_j^n(k_j^n - k_{j-1}^n)| \\ &\le \Delta t[\|k^n\|_\infty(M_f TV(s^n) + M_f TV(s^n) + M_f TV(c^n)) + M_f TV(k^n)]. \end{aligned}$$

Consequently, we have

$$I \le M_0 M_f(M_0 + 1)\Delta t + 2M_f M_0 \Delta t TV(s^n).$$

The term II is bounded by applying Lemmas 3.1 and 3.2:

$$II \equiv \Delta t M_f \|\frac{c_{j+1}^n - c_j^n}{\Delta x}\|_\infty TV(c^n) \le \Delta t M_f \|k^0\|_\infty TV(c^0) \le M_f M_0^2 \Delta t.$$

By using the bounds for I and II, we get

$$TV(s^{n+1}) \le (1 + 2M_f M_0 \Delta t)TV(s^n) + M_f M_0(2M_0 + 1)\Delta t,$$

and consequently we obtain, using induction, that

$$TV(s^n) \le (TV(s^0) + M_0 + \tfrac{1}{2})(1 + 2M_f M_0 \Delta t)^n - (M_0 + \tfrac{1}{2}).$$

Now the proof is completed by observing that

$$TV(s^n) \le (2M_0 + \tfrac{1}{2})e^{2M_f M_0 n \Delta t}. \qquad \square$$

An immediate consequence of Lemmas 3.1 and 3.4 and Corollary 3.3 is that the approximate solutions and the auxiliary sequence $\{k_j^n\}$ are L^1-Lipschitz continuous in time. For a proof of this fact we refer to [28].

LEMMA 3.5. *Suppose that $(s_j^0, c_j^0) \in \mathcal{S}$ for all $j \in \mathbb{Z}$, and that there is a finite constant M_0 such that*

$$TV(s^0), TV(k^0), TV(c^0), \|k^0\|_\infty \le M_0.$$

Then there exists a finite constant K depending only on M_0 and M_f such that

$$\|s^m - s^n\|_1, \|c^m - c^n\|_1, \|k^m - k^n\|_1 \le K|m - n|\Delta t$$

for $m, n \ge 0$ satisfying $m\Delta t, n\Delta t \le T_0 < \infty$.

We will conclude this section by proving that the approximate solutions satisfy a discrete entropy inequality. By using this discrete entropy inequality we will, in the next section, prove that the pair of functions (s, c), constructed as the limit of the family of approximate solutions, satisfies the entropy inequality formulated in Definition 1. In the discrete entropy condition, we will need a smooth approximation to the signum function. Let $\sigma_\epsilon = \sigma_\epsilon(s)$ be a family of nondecreasing C^∞-function satisfying $\sigma_\epsilon(s) = \text{sign}(s)$ for $|s| > \epsilon$.

LEMMA 3.6. *Suppose that* $(s_j^0, c_j^0) \in \mathcal{S}$ *for all* $j \in \mathbb{Z}$. *Let* ϕ *be a nonnegative* C^∞*-function with compact support on* $\mathbb{R} \times [0, T_0]$, *and let* N *be a positive integer such that* $N\Delta t \le T \le T_0$. *Then*

$$\begin{aligned}
\Delta t \sum_{n=0}^{N-1} \Delta x \sum_{j\in\mathbb{Z}} \Big\{ & \frac{\phi(j\Delta x,(n+1)\Delta t) - \phi(j\Delta x, n\Delta t)}{\Delta t} |s_j^{n+1} - q| \\
& + \frac{\phi((j+1)\Delta x, n\Delta t) - \phi(j\Delta x, n\Delta t)}{\Delta x} (f(s_j^n, c_j^n) - f(q, c_j^n))\operatorname{sign}(s_j^n - q) \\
& - \frac{f(q, c_j^n) - f(q, c_{j-1}^n)}{\Delta x} \sigma_\epsilon\left(s_j^{n+1} - q\right) \phi(j\Delta x, n\Delta t) \Big\} \\
+ \Delta x \sum_{j\in\mathbb{Z}} |s_j^0 - q| \phi(j\Delta x, 0) & - \Delta x \sum_{j\in\mathbb{Z}} |s_j^N - q| \phi(j\Delta x, N\Delta t) \\
& \ge \Delta t \sum_{n=0}^{N-1} \Delta x \sum_{j\in\mathbb{Z}} \Theta_\epsilon(s_j^{n+1} - q)\phi(j\Delta x, n\Delta t),
\end{aligned}$$

where

$$\Theta_\epsilon(s) = \sigma_\epsilon(s)s - |s|.$$

Proof. From the finite difference scheme (3.1), we have the following identity:

$$\begin{aligned}
s_j^{n+1} - q = s_j^n - q & - \mu(f(s_j^n, c_j^n) - f(q, c_j^n)) \\
& + \mu(f(s_{j-1}^n, c_{j-1}^n) - f(q, c_{j-1}^n)) \\
& - \mu(f(q, c_j^n) - f(q, c_{j-1}^n))
\end{aligned} \tag{3.11}$$

for any $q \in [S_0, 1]$. By multiplying this identity by $\sigma_\epsilon(s_j^{n+1} - q)$, we get $\sigma_\epsilon(s_j^{n+1} - q)(s_j^{n+1} - q)$ on the left-hand side of (3.11). By the definition of Θ_ϵ, this equals $|s_j^{n+1} - q| + \Theta_\epsilon(s_j^{n+1} - q)$. To estimate the resulting first term of the right-hand side, we introduce the function

$$\alpha(s) = s - q - \mu(f(s, c_j^n) - f(q, c_j^n)).$$

By the CFL-condition (3.2), we have

$$\alpha'(s) = 1 - \mu f_s(s, c_j^n) \ge 0,$$

hence $\operatorname{sign}(\alpha(s)) = \operatorname{sign}(s - q)$. By using this property of α and the fact that $\sigma_\epsilon(s) \le 1$, we obtain

$$\alpha(s)\sigma_\epsilon(s_j^{n+1} - q) \le |\alpha(s)| = \alpha(s)\operatorname{sign}(s - q).$$

Consequently, we have

$$\begin{aligned}
& (s_j^n - q - \mu(f(s_j^n, c_j^n) - f(q, c_j^n)))\sigma_\epsilon(s_j^{n+1} - q) \\
& \quad \le |s_j^n - q| - \mu(f(s_j^n, c_j^n) - f(q, c_j^n))\operatorname{sign}(s_j^n - q).
\end{aligned}$$

Since f is a nondecreasing function of s, the resulting second term of the right-hand side of (3.11) can be bounded in a similar manner:

$$\begin{aligned}
& \mu(f(s_{j-1}^n, c_{j-1}^n) - f(q, c_{j-1}^n))\sigma_\epsilon(s_j^{n+1} - q) \\
& \quad \le \mu |f(s_{j-1}^n, c_{j-1}^n) - f(q, c_{j-1}^n)| \\
& \quad = \mu(f(s_{j-1}^n, c_{j-1}^n) - f(q, c_{j-1}^n))\operatorname{sign}(s_{j-1}^n - q).
\end{aligned}$$

To summarize, we have derived from (3.11) the following pointwise discrete entropy inequality:

(3.12)
$$\begin{aligned}|s_j^{n+1}-q|-|s_j^n-q|+\Theta_\epsilon(s_j^{n+1}-q)\le&-\mu(f(s_j^n,c_j^n)-f(q,c_j^n))\operatorname{sign}(s_j^n-q)\\&+\mu(f(s_{j-1}^n,c_{j-1}^n)-f(q,c_{j-1}^n))\operatorname{sign}(s_{j-1}^n-q)\\&-\mu(f(q,c_j^n)-f(q,c_{j-1}^n))\sigma_\epsilon(s_j^{n+1}-q).\end{aligned}$$

By multiplying the inequality (3.12) by $\phi(j\Delta x,n\Delta t)$, we obtain, after summation in space and time, that

$$\begin{aligned}&\Delta t\sum_{n=0}^{N-1}\Delta x\sum_{j\in\mathbb{Z}}\Big\{\frac{|s_j^{n+1}-q|-|s_j^n-q|}{\Delta t}\phi(j\Delta x,n\Delta t)\\&+\frac{(f(s_j^n,c_j^n)-f(q,c_j^n))\operatorname{sign}(s_j^n-q)-(f(s_{j-1}^n,c_{j-1}^n)-f(q,c_{j-1}^n))\operatorname{sign}(s_{j-1}^n-q)}{\Delta x}\\&\qquad\cdot\phi(j\Delta x,n\Delta t)+\frac{f(q,c_j^n)-f(q,c_{j-1}^n)}{\Delta x}\sigma_\epsilon(s_j^{n+1}-q)\phi(j\Delta x,n\Delta t)\Big\}\\&\le-\Delta t\sum_{n=0}^{N-1}\Delta x\sum_{j\in\mathbb{Z}}\Theta_\epsilon(s_j^{n+1}-q)\phi(j\Delta x,n\Delta t).\end{aligned}$$

By applying summation by parts in time and space for the first and the second term, respectively, we get

$$\begin{aligned}&\Delta x\sum_{j\in\mathbb{Z}}|s_j^N-q|\phi(j\Delta x,N\Delta t)-\Delta x\sum_{j\in\mathbb{Z}}|s_j^0-q|\phi(j\Delta x,0)\\&-\Delta t\sum_{n=0}^{N-1}\Delta x\sum_{j\in\mathbb{Z}}\frac{\phi(j\Delta x,(n+1)\Delta t)-\phi(j\Delta x,n\Delta t)}{\Delta t}|s_j^{n+1}-q|\\&-\Delta t\sum_{n=0}^{N-1}\Delta x\sum_{j\in\mathbb{Z}}\frac{\phi((j+1)\Delta x,n\Delta t)-\phi(j\Delta x,n\Delta t)}{\Delta x}(f(s_j^n,c_j^n)-f(q,c_j^n))\operatorname{sign}(s_j^n-q)\\&+\Delta t\sum_{n=0}^{N-1}\Delta x\sum_{j\in\mathbb{Z}}\frac{f(q,c_j^n)-f(q,c_{j-1}^n)}{\Delta x}\sigma_\epsilon(s_j^{n+1}-q)\phi(j\Delta x,n\Delta t)\\&\le-\Delta t\sum_{n=0}^{N-1}\Delta x\sum_{j\in\mathbb{Z}}\Theta_\epsilon(s_j^{n+1}-q)\phi(j\Delta x,n\Delta t).\end{aligned}$$

which concludes the proof. □

4. Existence of an entropy solution. In the previous section we established some properties of the family of approximate solutions. In this section we will use these properties to prove that there exists a solution of the problem (1.2) in the sense of Definition 1. We start by showing that a subsequence of the family of approximate solutions (s_Δ,c_Δ) converges in L^1_{loc} to a pair of functions (s,c), and that this limit inherits the properties of the approximate solutions. Then we prove that in conservative variables, i.e., (s,sc), the limit is a weak solution of the Cauchy problem. Using this fact, and the regularity of c, we prove that the pair (s,c) is a solution according to Definition 1.

We start by stating the convergence result.

LEMMA 4.1. *Suppose that* (s^0, c^0) *is in the class* $\mathcal{B}^0$ (*cf.* (2.4)). *Then, as the meshsize tends to zero, there is a subsequence of* (s_Δ, c_Δ), *the family of approximate solutions generated by the finite difference scheme* (3.1), *converging in* L^1_{loc} *to a pair of functions* $(s, c) \in \mathcal{B}$.

Proof. In the previous section, we established the following properties of the approximate solutions:

$$\begin{aligned}
&(1)\quad (s_\Delta(x,t), c_\Delta(x,t)) \in \mathcal{S} \quad \forall (x,t) \in \mathbb{R} \times [0, T_0],\\
&(2)\quad TV(s_\Delta(\cdot,t)), TV(c_\Delta(\cdot,t)) \le K,\\
&(3)\quad \|s_\Delta(\cdot,t) - s_\Delta(\cdot,\tau)\|_1 + \|c_\Delta(\cdot,t) - c_\Delta(\cdot,\tau)\|_1 \le K(|t-\tau| + \Delta t)\\
&\text{for}\quad 0 \le t, \tau \le T_0 < \infty,
\end{aligned}$$

where K is a finite constant independent of Δ. Here (1) follows from Lemma 3.1, (2) follows from Lemmas 3.1 and 3.4, and (3) follows from Lemma 3.5. From these properties, it follows by an argument presented by Oleinik [19] (see also Smoller [24, Chap. 16]) that a subsequence, also denoted by (s_Δ, c_Δ), of the family of approximate solutions converges in L^1_{loc} for all $t \in [0, T_0]$ towards a pair of functions (s, c) as the meshsize tends to zero.

It remains to prove that the limit (s, c) is in the class $\mathcal{B}$ of functions (cf. (2.5)). Since convergence in L^1_{loc} implies pointwise convergence almost everywhere of a subsequence, we can deduce the properties of the limit (s, c) by appealing to the analogous discrete results. It might, however, be necessary to redefine (s, c) on a set of measure zero.

The fact that (s, c) remains in the state-space (cf. (a) of (2.4)), i.e.,

$$(s(x,t), c(x,t)) \in \mathcal{S} \quad \forall (x,t) \in \mathbb{R} \times [0, T_0],$$

follows from Lemma 3.1. The bound on the total variation and the Lipschitz continuity of c (cf. (b) of (2.4)), i.e.,

$$c(\cdot,t) \in \text{Lip} \cap BV \quad \text{for } t \in [0, T_0] \tag{4.1}$$

is a consequence of Lemmas 3.1 and 3.2. A detailed proof of (4.1) based on Lemma 3.2 is given in [28]. In a similar manner we deduce that

$$c(x,\cdot) \in \text{Lip}[0, \mathrm{T}_0] \quad \text{for } x \in \mathbb{R}$$

(cf. (c) of (2.5)), where in addition we have used the scheme (3.1) to derive an obvious bound on

$$\left\| \frac{c_j^{n+1} - c_j^n}{\Delta t} \right\|_\infty .$$

The L^1-Lipschitz continuity in time,

$$\|s(\cdot,t) - s(\cdot,\tau)\|_1 + \|c(\cdot,t) - c(\cdot,\tau)\|_1 \le K|t-\tau| \quad \text{for } 0 \le t, \tau \le T_0 < \infty$$

(cf. (b) of (2.5)) is a consequence of Lemma 3.5. From Lemma 3.4, it follows that

$$s(\cdot,t), \in BV \quad \text{for } t \in [0, T_0]$$

(cf. (c) of (2.4)).

To show that (c) of (2.4) are satisfied for almost all t, we define the family $\{k_\Delta\}$ of functions by

$$k_\Delta(x,t) = k_j^n \quad \text{for } (x,t) \in [(j-\tfrac{1}{2})\Delta x, (j+\tfrac{1}{2})\Delta x) \times [n\Delta t, (n+1)\Delta t).$$

Then, by the properties of k_Δ (cf. Corollary 3.3 and Lemma 3.5), there is a further subsequence, still denoted by k_Δ, converging in L^1_{loc} for all $t \in [0,T_0]$ to a function k satisfying $k(\cdot,t) \in BV \cap L^\infty$ for $t \in [0,T_0]$. Let ϕ be a smooth test function with compact support, and observe that

$$\begin{aligned}\Delta t \sum_{n=0}^{N} \Delta x \sum_{j\in\mathbb{Z}} k_j^n s_{j+1}^n \phi(j\Delta x, n\Delta t) &= \Delta t \sum_{n=0}^{N} \Delta x \sum_{j\in\mathbb{Z}} \frac{c_{j+1}^n - c_j^n}{\Delta x} \phi(j\Delta x, n\Delta t) \\ &= -\Delta t \sum_{n=0}^{N} \Delta x \sum_{j\in\mathbb{Z}} c_{j+1}^n \frac{\phi((j+1)\Delta x, n\Delta t) - \phi(j\Delta x, n\Delta t)}{\Delta x},\end{aligned}$$

where $N = \lceil T/\Delta t \rceil$, $T \le T_0$. Here $\lceil\cdot\rceil$ denotes the truncation operator. By Lebesgue dominated convergence theorem, we get

$$\int_0^T \int_{\mathbb{R}} ks\phi \, dx \, dt = -\int_0^T \int_{\mathbb{R}} c\phi_x \, dx \, dt.$$

Since c_x exists and is well defined almost everywhere, we have

$$\int_0^T \int_{\mathbb{R}} ks\phi \, dx \, dt = \int_0^T \int_{\mathbb{R}} c_x\phi \, dx \, dt.$$

Consequently,

$$ks = c_x,$$

where we, if necessary, have redefined k on a set of measure zero. This proves that (s,c) satisfies (c) of (2.4).

Thus we have verified that (s,c) is in the class $\mathcal{B}$. □

We have seen that a subsequence of the approximate solutions converges towards a pair of functions (s,c), and we will show that this limit is in fact an entropy solution. First, we will show that the limit is a weak solution of the system. Then, by using the smoothness of c, we prove that the second equation of (1.2) is satisfied in a classical sense almost everywhere. Finally, we prove that the limit is an entropy solution of the system.

We first establish that the limit is a weak solution of the system (1.2), i.e., a weak solution in the conservative variables (s, sc). The main difficulty in proving this is that our finite difference scheme is in nonconservative form. Hence, we cannot appeal to the classical theorem of Lax and Wendroff [18], stating that if a family of finite difference approximations generated by a conservative finite difference scheme converges boundedly almost everywhere, then the limit is a weak solution of the system. For our nonconservative scheme we proved a similar result in [28], taking advantage of the smoothness of the variable corresponding to c_Δ. In fact, the following proposition is a direct consequence of Lemma 2 in [28].

PROPOSITION 4.2. *Suppose that (s^0, c^0) is in the class $\mathcal{B}^0$ (cf. (2.4)). Then the pair of functions (s, sc) generated as the limit of the proper subsequence of the family*

of approximate solutions (cf. the previous lemma) is a weak solution of the system (1.1), *in the sense that*

$$\int_0^T \int_{\mathbb{R}} (s\phi_t + f(s,c)\phi_x)\,dx\,dt + \int_{\mathbb{R}} s^0(x)\phi(x,0)\,dx - \int_{\mathbb{R}} s(x,T)\phi(x,T)\,dx = 0,$$
$$\int_0^T \int_{\mathbb{R}} (sc\phi_t + cf(s,c)\phi_x)\,dx\,dt + \int_{\mathbb{R}} s^0(x)c^0(x)\phi(x,0)\,dx$$
$$- \int_{\mathbb{R}} s(x,T)c(x,T)\phi(x,T)\,dx = 0$$

for any smooth test function ϕ of compact support and any $0 \le T \le T_0$.

We next prove that (s,c) satisfies the second equation of (1.2) in a classical sense almost everywhere.

LEMMA 4.3. *Under the assumptions of Lemma* 4.1, *the pair of functions $(s,c) \in \mathcal{B}$ generated as the limit of the proper subsequence of (s_Δ, c_Δ) satisfies*

$$c_t + g(s,c)c_x = 0 \quad a.e.,$$

and

$$\lim_{t \to 0^+} \|c(\cdot,t) - c^0\|_1 = 0.$$

Proof. Let ϕ be a C^∞-function satisfying $\phi(x,t) \equiv 0$ for $|x| \ge R, t = 0, t \ge T$, where R is a finite real number. Then it follows from Proposition 4.2 that

$$\int_0^T \int_{\mathbb{R}} (s\phi_t + f(s,c)\phi_x)\,dx\,dt = 0 \tag{4.2}$$

and

$$\int_0^T \int_{\mathbb{R}} (sc\phi_t + cf(s,c)\phi_x)\,dx\,dt = 0. \tag{4.3}$$

Assume, for the moment, that

$$\int_0^T \int_{\mathbb{R}} (s(c\phi)_t + f(s,c)(c\phi)_x)\,dx\,dt = 0. \tag{4.4}$$

Then, since $(s,c) \in \mathcal{B}$, c_t and c_x exist almost everywhere,

$$\int_0^T \int_{\mathbb{R}} (s\phi c_t + f(s,c)\phi c_x)\,dx\,dt + \int_0^T \int_{\mathbb{R}} (sc\phi_t + cf(s,c)\phi_x)\,dx\,dt = 0,$$

which, by (4.3), implies

$$\int_0^T \int_{\mathbb{R}} s\phi(c_t + g(s,c)c_x)\,dx\,dt = 0.$$

Since $s(x,t) \ge S_0 > 0$, and ϕ is arbitrary, we have

$$c_t + g(s,c)c_x = 0 \quad \text{a.e.}$$

It remains to prove (4.4). Let ψ_n be a sequence of C^∞-functions with $\operatorname{supp}(\psi_n) \subset \operatorname{supp}(\phi)$ converging towards the function $c\phi$ in $W^{1,1}([-R,R]\times[0,T])$; such a sequence exists since C^∞ is dense in $W^{1,1}$ (cf. [30]) and $c\phi$ is easily seen to be in $W^{1,1}$. Consequently, by (4.2), we have

$$\begin{aligned}
&\left|\int_0^T\int_{\mathbb{R}}(s(c\phi)_t + f(s,c)(c\phi)_x)\,dx\,dt\right| \\
&= \left|\int_0^T\int_{\mathbb{R}}(s(c\phi)_t + f(s,c)(c\phi)_x)\,dx\,dt\right. \\
&\qquad \left. -\int_0^T\int_{\mathbb{R}}(s(\psi_n)_t + f(s,c)(\psi_n)_x)\,dx\,dt\right| \\
&\le M_f\int_0^T\int_{\mathbb{R}}(|(c\phi-\psi_n)_t| + |(c\phi-\psi_n)_x)|\,dx\,dt \longrightarrow 0 \quad \text{as } n\longrightarrow\infty.
\end{aligned}$$

We remark that $W^{1,1}$ denotes the Sobolev space with one derivative in L^1.

The second part of the lemma follows from the L^1-Lipschitz continuity in time of c (cf. property (b) of (2.5)), and the fact that $\|c_\Delta(\cdot,0)-c^0\|_1 \to 0$ as $\Delta\to 0$, by the construction. □

We finally prove that (s,c) satisfies the entropy inequality.

LEMMA 4.4. *Under the assumptions of Lemma* 4.1, *the pair of functions* $(s,c)\in\mathcal{B}$ *generated as the limit of the proper subsequence of* (s_Δ, c_Δ) *satisfies*

$$\begin{aligned}
&\int_0^T\int_{\mathbb{R}}\{|s-q|\phi_t + \operatorname{sign}(s-q)(f(s,c)-f(q,c))\phi_x - \operatorname{sign}(s-q)f(q,c)_x\phi\}\,dx\,dt \\
&+\int_{\mathbb{R}}|s^0(x)-q|\phi(x,0)\,dx - \int_{\mathbb{R}}|s(x,T)-q|\phi(x,T)\,dx \ge 0
\end{aligned}$$

for all nonnegative C^∞*-functions* ϕ *with compact support in* $\mathbb{R}\times[0,T_0]$, *all* $q\in[S_0,1]$, *and all* $T\in[0,T_0]$.

Proof. Let ϕ be a nonnegative C^∞-function with compact support on $\mathbb{R}\times[0,T_0]$ and let $N=\lceil T/\Delta t\rceil$. Then, by Lemma 3.6, we have

$$\begin{aligned}
&\Delta t\sum_{n=0}^{N-1}\Delta x\sum_{j\in\mathbb{Z}}\left\{\frac{\phi(j\Delta x,(n+1)\Delta t)-\phi(j\Delta x,n\Delta t)}{\Delta t}|s_j^{n+1}-q|\right. \\
&\quad +\frac{\phi((j+1)\Delta x,n\Delta t)-\phi(j\Delta x,n\Delta t)}{\Delta x}(f(s_j^n,c_j^n)-f(q,c_j^n))\operatorname{sign}(s_j^n-q) \\
&\qquad \left. -f_c(q,\tilde{c_j^n})k_{j-1}^n s_j^n\sigma_\epsilon\left(s_j^{n+1}-q\right)\phi(j\Delta x,n\Delta t)\right\} \\
&+\Delta x\sum_{j\in\mathbb{Z}}|s_j^0-q|\phi(j\Delta x,0) - \Delta x\sum_{j\in\mathbb{Z}}|s_j^N-q|\phi(j\Delta x,N\Delta t) \\
&\quad \ge \Delta t\sum_{n=0}^{N-1}\Delta x\sum_{j\in\mathbb{Z}}\Theta_\epsilon(s_j^{n+1}-q)\phi(j\Delta x,n\Delta t)
\end{aligned}$$

where we have used (3.9), and where $\tilde{c_j^n}\in\operatorname{int}[c_{j-1}^n,c_j^n]$.

Since $|\Theta_\epsilon(s)| \leq \epsilon$ for all s, the right-hand side is greater than

$$-\epsilon\Delta t \sum_{n=0}^{N-1} \Delta x \sum_{j\in\mathbb{Z}} \phi(j\Delta x, n\Delta t).$$

By the proof of Lemma 4.1, a subsequence of the family $(s_\Delta, c_\Delta, k_\Delta)$ converges pointwise almost everywhere towards (s, c, k), where $ks = c_x$; thus by passing to the limit in $(\Delta x, \Delta t)$ we obtain, using the Lebesque dominated convergence theorem, that

$$\begin{aligned}
&\int_0^T \int_{\mathbb{R}} \{|s-q|\phi_t + \operatorname{sign}(s-q)(f(s,c)-f(q,c))\phi_x - \sigma_\epsilon(s-q)f(q,c)_x\phi\}\,dx\,dt \\
&\quad + \int_{\mathbb{R}} |s^0(x)-q|\phi(x,0)\,dx - \int_{\mathbb{R}} |s(x,T)-q|\phi(x,T)\,dx \\
&\qquad \geq -\epsilon \int_0^T \int_{\mathbb{R}} \phi(x,t)\,dx\,dt.
\end{aligned}$$

Now the lemma follows by passing to the limit in ϵ and again using the Lebesgue dominated convergence theorem. □

Proof of Theorem 2.1. We can now summarize the proof of Theorem 2.1. We have constructed a pair of functions (s, c), which by Lemma 4.1 satisfies $(s, c) \in \mathcal{B}$. By Lemma 4.3 the second equation of (1.2) is satisfied in a classical sense almost everywhere, and the solution converges in L^1 towards the initial condition as the time tends to zero. Finally, we proved in Lemma 4.4 that (s, c) satisfies our entropy condition. Consequently, (s, c) is a solution according to Definition 1. □

5. Uniqueness and continuous dependence. The purpose of this section is to prove uniqueness of the entropy solution of (1.2). This result will be established under an extra smoothness assumption on the initial data (s^0, c^0). The technique we shall apply uses characteristics for the concentration equation of (1.2) together with a modification of the analysis given by Kruzkov [10] and Kuznetsov [11] applied to the saturation equation. As a consequence of our uniqueness argument, we will also obtain a continuous dependence result for the solution with respect to the initial data (s^0, c^0). Another consequence of the uniqueness is that the entire family of approximate solutions converges, not only a subsequence of it.

Before we consider the question of uniqueness, we will establish some elementary properties of an entropy solution of (1.2). These properties are based on the simple observation that the characteristic speed g of c is also the particle velocity of the aqueous phase.

Let (s, c) denote an entropy solution of (1.2) with initial data (s^0, c^0). Throughout this section it will be assumed that there exist two constant states (s^L, c^L) and (s^R, c^R) such that for $0 \leq t \leq T$, we have

$$(s(x,t), c(x,t)) = \begin{cases} (s^L, c^L) & x \leq x^L(t), \\ (s^R, c^R) & x \geq x^R(t), \end{cases} \tag{5.1}$$

where x^L and x^R are linear functions of t of the form

$$x^L(t) = L + g(s^L, c^L)t, \qquad x^R(t) = R + g(s^R, c^R)t$$

(cf. Fig. 2). In particular, we observe that the solution constructed above satisfies the condition (5.1) if the initial data is constant for $|x|$ sufficiently large.

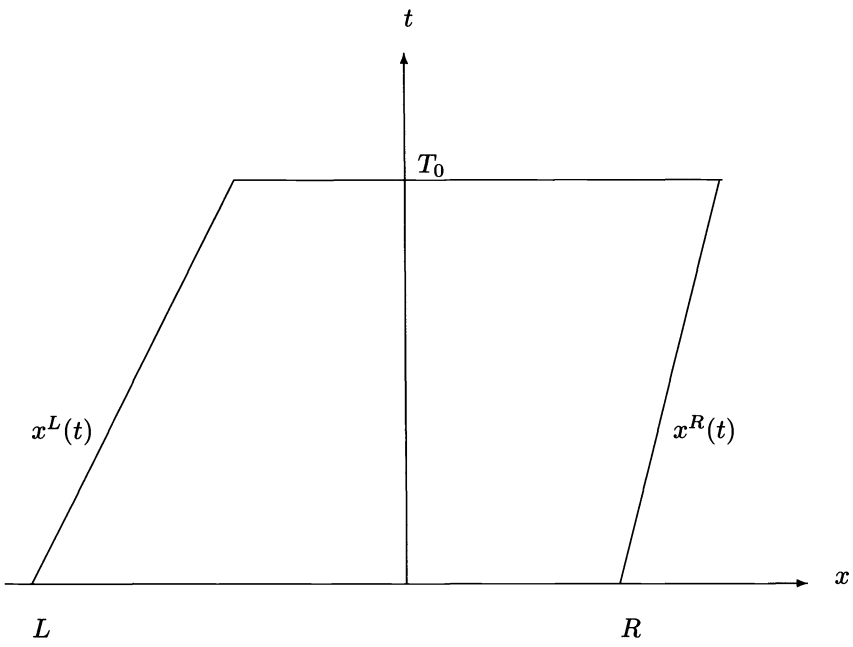


FIG. 2

Let $x(t)$, $0 \le t \le T$, be a smooth curve such that $x^L(t) \le x(t) \le x^R(t)$. From the entropy condition, it follows in particular that (s, c) satisfies the first equation of (1.2) weakly; hence (cf. Oleinik [20]) we have the following conservation relation:

$$\int_{x^L(t)}^{x(t)} s(\xi, t)\, d\xi - \int_L^{x(0)} s^0(\xi)\, d\xi = \int_0^t [s(x(\tau), \tau)x_\tau(\tau) - f(s(x(\tau), \tau), c(x(\tau), \tau))n(\tau)]\, d\tau, \tag{5.2}$$

where

$$n(\tau) = \frac{1}{\sqrt{1 + (x_\tau(\tau))^2}}.$$

For any $t \in [0, T_0]$ and $x \in [x^L(t), x^R(t)]$, define

$$m(x, t) = \int_{x^L(t)}^{x} s(\xi, t)\, d\xi.$$

For each $t \in [0, T_0]$, $m(\cdot, t)$ is a strictly increasing function of x with

$$m_x(x, t) = s(x, t) \ge S_0 > 0.$$

Furthermore, (5.2) implies that

$$m(x^R(t), t) = m(R, 0). \tag{5.3}$$

Also, for $t \in [0, T_0]$, we define a strictly increasing function $y(\cdot, t)$ from $[L, R]$ onto $[x^L(t), x^R(t)]$ by

$$m(y(x,t),t) = m(x,0). \tag{5.4}$$

Since m_x exists, y_x exists and

$$y_x(x,t) = \frac{m_x(x,0)}{m_y(y,t)} = \frac{s^0(x)}{s(y,t)} \geq S_0 > 0.$$

The identity (5.2) implies in particular that

$$m(x,t_1) - m(x,t_2) = -\int_{t_1}^{t_2} f(s(x,t),c(x,t))\,dt.$$

Hence, since the function $f(s(x,t),c(x,t))$ is bounded and measurable as a function of t for almost all x, it follows that $m_t(x,t)$ exists for almost all x and t, and

$$m_t(x,t) = -f(s(x,t),c(x,t)).$$

From (5.4) we therefore obtain that y_t exists for almost all (x,t), and

$$y_t(x,t) = -\frac{m_t(y,t)}{m_y(y,t)} = g(s(y,t),c(y,t)).$$

Note also that $y(x,0) = x$. Hence, the well-defined function $y(x,t)$, which is differentiable almost everywhere, corresponds to the desired characteristic for the concentration equation. In particular,

$$c(y(x,t),t) = c^0(x). \tag{5.5}$$

By differentiating this identity with respect to x, we obtain that

$$k(y(x,t),t) = k^0(x), \tag{5.6}$$

where $k = c_x/s$ and $k^0 = c_x^0/s^0$.

Assume now that $(\bar{s}, \bar{c})$ is another entropy solution of (1.2) with initial data $(\bar{s}^0, \bar{c}^0)$. Furthermore, we shall assume that $(\bar{s}, \bar{c})$ satisfies the condition (5.1), with the same boundary states (s^L, c^L) and (s^R, c^R) as the (s,c), and with the same curves $x^L(t)$ and $x^R(t)$. The difference between c and $\bar{c}$ can now be estimated by using the associated characteristics.

LEMMA 5.1. *Let (s,c) and $(\bar{s},\bar{c})$ be two entropy solutions of* (1.2), *with initial data $(s^0,c^0) \in \mathcal{B}^0$ and $(\bar{s}^0,\bar{c}^0) \in \mathcal{B}^0$, respectively, both satisfying the condition* (5.1). *Then there is a finite constant M such that*

$$\|(c-\bar{c})(\cdot,t)\|_1 \leq M\left\{\|c^0-\bar{c}^0\|_1 + \int_0^t (\|(s-\bar{s})(\cdot,\tau)\|_1 + \|(c-\bar{c})(\cdot,\tau)\|_1)\,d\tau\right\}$$

for $t \in [0, T_0]$.

Proof. Let S and C denote the differences $s - \bar{s}$ and $c - \bar{c}$, respectively. Then C clearly satisfies the following equation:

$$C_t + g(\bar{s},\bar{c})C_x = F(x,t),$$

where $F = c_x(g(\bar{s},\bar{c}) - g(s,c))$. Hence, if $\bar{y} = \bar{y}(x,t)$ represents the characteristic associated the solution $(\bar{s},\bar{c})$, then

$$C(\bar{y}(x,t),t) = C(x,0) + \int_0^t F(\bar{y}(x,\tau),\tau)\,d\tau. \tag{5.7}$$

Since

$$\int_{\mathbb{R}} |C(\bar{y}(x,t),t)|\,dx \geq S_0 \|C(\cdot,t)\|_1,$$

and

$$\int_0^t \int_{\mathbb{R}} |F(\bar{y}(x,\tau),\tau)|\,dx\,d\tau \leq \frac{M_g}{S_0}\|c_x\|_{L^\infty(\mathbb{R}\times[0,T_0])} \int_0^t (\|S(\cdot,\tau)\|_1 + \|C(\cdot,\tau)\|_1)\,d\tau,$$

the desired result follows. □

Recall that $k = c_x/s$, and let $\bar{k} = \bar{c}_x/\bar{s}$. Similarly, $k^0 = c_x^0/s^0$ and $\bar{k}^0 = \bar{c}_x^0/\bar{s}^0$. By assuming some extra regularity on the initial data, we obtain the following bound for the difference $c_x - \bar{c}_x$.

LEMMA 5.2. *Let (s,c) and $(\bar{s},\bar{c})$ be two entropy solutions of (1.2) as above, and assume in addition that $k_x^0 \in L^\infty$. Then there is a finite constant M such that*

$$\|(c_x - \bar{c}_x)(\cdot,t)\|_1 \leq M\Big\{\|s^0 - \bar{s}^0\|_1 + \|c_x^0 - \bar{c}_x^0\|_1 + \|s(\cdot,t) - \bar{s}(\cdot,t)\|_1 + \int_0^t (\|(s-\bar{s})(\cdot,\tau)\|_1 + \|(c-\bar{c})(\cdot,\tau)\|_1)\,d\tau\Big\}$$

for $t \in [0,T_0]$.

Proof. From (5.6), it follows that

$$k_t + g(s,c)k_x = 0.$$

A similar relation holds for $\bar{k}$. Since k_x is bounded by (5.6), an argument analogous to the argument given in the proof of Lemma 5.1 leads to the estimate

$$\|(k-\bar{k})(\cdot,t)\|_1 \leq M\left\{\|k^0 - \bar{k}^0\|_1 + \int_0^t (\|(s-\bar{s})(\cdot,\tau)\|_1 + \|(c-\bar{c})(\cdot,\tau)\|_1)\,d\tau\right\}.$$

Therefore, since

$$c_x - \bar{c}_x = s(k-\bar{k}) + \bar{k}(s-\bar{s}),$$

the estimate follows by a proper modification of the constant M. □

Finally, we estimate the difference $s - \bar{s}$ in L^1.

LEMMA 5.3. *Let (s,c) and $(\bar{s},\bar{c})$ be two entropy solutions of (1.2) as above. Then there is a finite constant K such that*

$$\|s(\cdot,t) - \bar{s}(\cdot,t)\|_1 \leq \|s^0 - \bar{s}^0\|_1 + K\int_0^t \{\|c_x(\cdot,\tau) - \bar{c}_x(\cdot,\tau)\|_1 + \|c(\cdot,\tau) - \bar{c}(\cdot,\tau)\|_1 + \|s(\cdot,\tau) - \bar{s}(\cdot,\tau)\|_1\}\,d\tau$$

for $t \in [0,T_0]$.

The proof of this lemma is given in the end of this section.

Proof of Theorem 2.2. We are now in position to prove Theorem 2.2. Define

$$\Phi(t) = \|c_x(\cdot,t) - \bar{c}_x(\cdot,t)\|_1 + \|c(\cdot,t) - \bar{c}(\cdot,t)\|_1 + \|s(\cdot,t) - \bar{s}(\cdot,t)\|_1.$$

By Lemmas 5.1, 5.2, and 5.3 we obtain

$$\Phi(t) \leq M\left(\Phi(0) + \int_0^t \Phi(\tau)\,d\tau\right)$$

for a finite constant M. Hence, by a proper modification of M, we get

$$\Phi(t) \leq M\Phi(0)$$

for $t \in [0, T_0]$, which is the stability result stated in Theorem 2.2. In particular, this implies the uniqueness result. □

As a consequence of the uniqueness of the entropy solution, we obtain convergence of the entire family (s_Δ, c_Δ) of approximate solutions.

5.1. Proof of Lemma 5.3. We will now prove Lemma 5.3, i.e., the stability estimate for the saturation variable s. The proof is motivated by the work of Kruzkov [10] and Kuznetsov [11] for scalar equations.

We introduce a nonnegative function $\omega \in C^\infty$ satisfying $\omega(\sigma) = \omega(-\sigma)$, $\omega(\sigma) \equiv 0$ for $|\sigma| \geq 1$, and $\int_{\mathbb{R}} \omega(\sigma)d\sigma = 1$. For $\epsilon > 0$, let

$$\omega_\epsilon(\sigma) = \frac{1}{\epsilon}\omega(\sigma/\epsilon);$$

then

$$\begin{aligned}
&\text{(a)} \quad \omega_\epsilon \in C^\infty, \quad \omega_\epsilon(\sigma) \geq 0 \quad \forall \sigma \in \mathbb{R},\\
&\text{(b)} \quad \int_{\mathbb{R}} \omega_\epsilon(\sigma)d\sigma = 1,\\
&\text{(c)} \quad \omega_\epsilon(\sigma) \equiv 0 \quad \text{for} \quad |\sigma| \geq \epsilon,\\
&\text{(d)} \quad \omega_\epsilon(\sigma) \leq M_\omega/\epsilon, \quad |\omega_\epsilon{}'(\sigma)| \leq M_\omega/\epsilon^2,
\end{aligned}$$

where M_ω is a finite constant independent ϵ.

We will need the following auxilary result concerning bounded and measurable functions of compact support.

LEMMA 5.4. (i) *Suppose $v = v(x)$ is a bounded and measurable function of compact support on $\mathbb{R}$; then*

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}\int_{\mathbb{R}}\int_{x-\epsilon}^{x+\epsilon} |v(x) - v(y)|dy\,dx = 0.$$

(ii) *Similarly, suppose $v = v(x,t)$ is a bounded and measurable function of compact support on $\mathbb{R}\times\mathbb{R}$; then*

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon^2}\int_{\mathbb{R}}\int_{\mathbb{R}^+}\int_{x-\epsilon}^{x+\epsilon}\int_{t-\epsilon}^{t+\epsilon} |v(x,t) - v(y,\tau)|\,d\tau\,dy\,dt\,dx = 0.$$

We refer to Kruzkov [10, Lemma 2, p. 222] for a proof.

As above, (s,c) and $(\bar{s},\bar{c})$ are two pairs of solutions, in the sense of Definition 1, of the problem (1.2). Recall that

$$s(x,t)=\bar{s}(x,t) \quad \text{and} \quad c(x,t)=\bar{c}(x,t) \quad \text{for } x\notin[x^L(t),x^R(t)], \quad t\in[0,T_0].$$

For $q\in[S_0,1]$, we define

$$F_1(s,c,q)=\operatorname{sign}(s-q)(f(s,c)-f(q,c))$$

and

$$F_2(s,c,q)=\operatorname{sign}(s-q)f(q,c)_x.$$

Then the entropy condition for (s,c) reads

$$\int_0^T\int_{\mathbb{R}}\{|s-q|\phi_t+F_1(s,c,q)\phi_x-F_2(s,c,q)\phi\}\,dx\,dt + \int_{\mathbb{R}}|s^0(x)-q|\phi(x,0)\,dx-\int_{\mathbb{R}}|s(x,T)-q|\phi(x,T)\,dx\geq 0$$

for all $q\in[S_0,1]$, and all nonnegative C^∞-functions ϕ of compact support. We introduce the test functions

$$\psi_\epsilon(x,t,y,\tau)=\omega_\epsilon(t-\tau)\omega_\epsilon(x-y)$$

for $(x,t,y,\tau)\in(\mathbb{R}\times[0,T_0])^2$, and the form

$$\rho_\epsilon(w,z)=\int_{\mathbb{R}}\int_{\mathbb{R}}\omega_\epsilon(x-y)|w(x)-z(y)|\,dx\,dy.$$

By putting $q=\bar{s}(y,\tau)$ in the entropy condition for (s,c) and integrating with respect to y and τ, we get

$$\begin{aligned}(5.8)\quad 0\leq{}&\int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}}\{|s(x,t)-\bar{s}(y,\tau)|(\psi_\epsilon)_t+F_1(s(x,t),c(x,t),\bar{s}(y,\tau))(\psi_\epsilon)_x\\&-F_2(s(x,t),c(x,t),\bar{s}(y,\tau))\psi_\epsilon\}\,dx\,dt\,dy\,d\tau\\&+\int_0^T\rho_\epsilon(s^0,\bar{s}(\cdot,\tau))\omega_\epsilon(\tau)\,d\tau-\int_0^T\rho_\epsilon(s(\cdot,T),\bar{s}(\cdot,\tau))\omega_\epsilon(T-\tau)\,d\tau\end{aligned}$$

for any $T\in[0,T_0]$. Similarly, by the entropy condition for $(\bar{s},\bar{c})$, we obtain

$$\begin{aligned}(5.9)\quad 0\leq{}&\int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}}\{|s(x,t)-\bar{s}(y,\tau)|(\psi_\epsilon)_\tau+F_1(\bar{s}(y,\tau),\bar{c}(y,\tau),s(x,t))(\psi_\epsilon)_y\\&-F_2(\bar{s}(y,\tau),\bar{c}(y,\tau),s(x,t))\psi_\epsilon\}\,dx\,dt\,dy\,d\tau\\&+\int_0^T\rho_\epsilon(\bar{s}^0,s(\cdot,t))\omega_\epsilon(t)\,dt-\int_0^T\rho_\epsilon(\bar{s}(\cdot,T),s(\cdot,t))\omega_\epsilon(T-t)\,dt.\end{aligned}$$

We observe that

$$(\psi_\epsilon)_t=-(\psi_\epsilon)_\tau \quad \text{and} \quad (\psi_\epsilon)_x=-(\psi_\epsilon)_y.$$

Taking this into account, we obtain, by adding (5.8) and (5.9), that

$$\begin{aligned}
&\int_0^T \rho_\epsilon(s(\cdot,T),\bar{s}(\cdot,\tau))\omega_\epsilon(T-\tau)\,d\tau + \int_0^T \rho_\epsilon(\bar{s}(\cdot,T),s(\cdot,t)\omega_\epsilon(T-t))\,dt \\
&\le \int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}} (F_1(s(x,t),c(x,t),\bar{s}(y,\tau)) \\
&\qquad\qquad -F_1(\bar{s}(y,\tau),\bar{c}(y,\tau),s(x,t)))(\psi_\epsilon)_x\,dx\,dt\,dy\,d\tau \\
&\quad -\int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}} (F_2(s(x,t),c(x,t),\bar{s}(y,\tau)) \\
&\qquad\qquad +F_2(\bar{s}(y,\tau),\bar{c}(y,\tau),s(x,t)))\psi_\epsilon\,dx\,dt\,dy\,d\tau \\
&\quad +\int_0^T \rho_\epsilon(s^0,\bar{s}(\cdot,\tau))\omega_\epsilon(\tau)\,d\tau + \int_0^T \rho_\epsilon(\bar{s}^0,s(\cdot,t))\omega_\epsilon(t)\,dt.
\end{aligned}$$

For notational convenience we write this inequality as

$$L(\epsilon) \le R(\epsilon), \tag{5.10}$$

where

$$L(\epsilon) \equiv L_1(\epsilon) + L_2(\epsilon) \quad \text{and} \quad R(\epsilon) \equiv R_1(\epsilon) + R_2(\epsilon) + R_3(\epsilon) + R_4(\epsilon).$$

Observe that

$$\begin{aligned}
&\left| L_1(\epsilon) - \frac{1}{2}\int_{\mathbb{R}} \left| s(x,T) - \bar{s}(x,T) \right| dx \ \right| \\
&\quad = \left| \int_0^T\int_{\mathbb{R}}\int_{\mathbb{R}} |s(x,T) - \bar{s}(y,\tau)|\omega_\epsilon(T-\tau)\omega_\epsilon(x-y)\,dx\,dy\,d\tau \right. \\
&\qquad \left. - \int_0^T\int_{\mathbb{R}}\int_{\mathbb{R}} |s(x,T) - \bar{s}(x,T)|\omega_\epsilon(T-\tau)\omega_\epsilon(x-y)\,dx\,dy\,d\tau \right| \\
&\quad \le \int_0^T\int_{\mathbb{R}}\int_{\mathbb{R}} |\bar{s}(y,\tau) - \bar{s}(x,T)|\omega_\epsilon(T-\tau)\omega_\epsilon(x-y)\,dx\,dy\,d\tau \\
&\quad \le \int_0^T\int_{\mathbb{R}} |\bar{s}(y,\tau) - \bar{s}(y,T)|\omega_\epsilon(T-\tau)\,dy\,d\tau \\
&\qquad + \frac{1}{2}\int_{\mathbb{R}}\int_{\mathbb{R}} |\bar{s}(y,T) - \bar{s}(x,T)|\omega_\epsilon(x-y)\,dx\,dy \\
&\quad \le \frac{M_\omega}{\epsilon}\int_{T-\epsilon}^T\int_{\mathbb{R}} |\bar{s}(y,\tau) - \bar{s}(y,T)|\,dy\,d\tau + \frac{M_\omega}{2\epsilon}\int_{\mathbb{R}}\int_{x-\epsilon}^{x+\epsilon} |\bar{s}(y,T) - \bar{s}(x,T)|\,dy\,dx.
\end{aligned}$$

Hence, by Lemma 5.4, we have

$$\lim_{\epsilon\to 0} L_1(\epsilon) = \frac{1}{2}\int_{\mathbb{R}} |s(x,T) - \bar{s}(x,T)|\,dx.$$

By applying a similar argument for L_2, R_3, and R_4, we obtain

$$\lim_{\epsilon\to 0}(L_1(\epsilon) + L_2(\epsilon)) = \|s(\cdot,T) - \bar{s}(\cdot,T)\|_1 \tag{5.11}$$

and

$$\lim_{\epsilon\to 0}(R_3(\epsilon) + R_4(\epsilon)) = \|s^0 - \bar{s}^0\|_1. \tag{5.12}$$

It remains to estimate R_1 and R_2, and we start with the latter,

$$\begin{aligned}
|R_2(\epsilon)| &= \Big| \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} (F_2(s(x,t), c(x,t), \bar{s}(y,\tau)) \\
&\qquad + F_2(\bar{s}(y,\tau), \bar{c}(y,\tau), s(x,t))) \psi_\epsilon \, dx \, dt \, dy \, d\tau \Big| \\
&\le \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} |f(\bar{s}(y,\tau), c(x,t))_x - f(s(x,t), \bar{c}(y,\tau))_y| \psi_\epsilon \, dx \, dt \, dy \, d\tau \\
&\le \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} \{ |f_c(\bar{s}(y,\tau), c(x,t))(c_x(x,t) - \bar{c}_y(y,\tau))| \psi_\epsilon + |f_c(\bar{s}(y,\tau), c(x,t)) \\
&\qquad - f_c(s(x,t), \bar{c}(y,\tau))| |\bar{c}_y(y,\tau)| \psi_\epsilon \} \, dx \, dt \, dy \, d\tau \\
&\le K \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} (|c_x(x,t) - \bar{c}_y(y,\tau)| \\
&\qquad + |\bar{s}(y,\tau) - s(x,t)| + |c(x,t) - \bar{c}(y,\tau)|) \psi_\epsilon \, dx \, dt \, dy \, d\tau,
\end{aligned}$$

where K is a finite constant, depending on $\|c_y\|_\infty$, but independent of ϵ. Observe that

$$\begin{aligned}
&\int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} |c(x,t) - \bar{c}(y,\tau)| \psi_\epsilon \, dx \, dt \, dy \, d\tau \\
&\quad \le \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} |c(x,t) - \bar{c}(x,t)| \omega_\epsilon(x-y) \omega_\epsilon(t-\tau) \, dx \, dt \, dy \, d\tau \\
&\qquad + \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} |\bar{c}(x,t) - \bar{c}(y,\tau)| \omega_\epsilon(x-y) \omega_\epsilon(t-\tau) \, dx \, dt \, dy \, d\tau \\
&\quad \le \int_0^T \int_{\mathbb{R}} |c(x,t) - \bar{c}(x,t)| \, dx \, dt \\
&\qquad + \frac{M_\omega^2}{\epsilon^2} \int_0^T \int_{\mathbb{R}} \int_{\tau-\epsilon}^{\tau+\epsilon} \int_{y-\epsilon}^{y+\epsilon} |\bar{c}(x,t) - \bar{c}(y,\tau)| \, dx \, dt \, dy \, d\tau \\
&\quad \xrightarrow{\epsilon \to 0} \int_0^T \int_{\mathbb{R}} |c(x,t) - \bar{c}(x,t)| \, dx \, dt
\end{aligned}$$

by Lemma 5.4. (For simplicity, we have defined $\bar{c}(x,t) = \bar{c}^0(x)$ for $t < 0$.) In a similar manner, we obtain

$$\limsup_{\epsilon \to 0} \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} |c_x(x,t) - \bar{c}_y(y,\tau)| \psi_\epsilon \, dx \, dt \, dy \, d\tau \le \int_0^T \int_{\mathbb{R}} |c_x(x,t) - \bar{c}_x(x,t)| \, dx \, dt$$

and

$$\limsup_{\epsilon \to 0} \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} |\bar{s}(y,\tau) - s(x,t)| \psi_\epsilon \, dx \, dt \, dy \, d\tau \le \int_0^T \int_{\mathbb{R}} |s(x,t) - \bar{s}(x,t)| \, dx \, dt;$$

consequently

$$\begin{aligned}
\limsup_{\epsilon \to 0} |R_2(\epsilon)| \le K \int_0^T \{ &\|c_x(\cdot,t) - \bar{c}_x(\cdot,t)\|_1 \\
&+ \|c(\cdot,t) - \bar{c}(\cdot,t)\|_1 + \|s(\cdot,t) - \bar{s}(\cdot,t)\|_1 \} \, dt.
\end{aligned} \tag{5.13}$$

Next, we estimate R_1. We have

$$\begin{aligned} R_1(\epsilon) &= \int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}} [F_1(s(x,t),c(x,t),\bar{s}(y,\tau)) \\ &\qquad -F_1(\bar{s}(y,\tau),\bar{c}(y,\tau),s(x,t))](\psi_\epsilon)_x\,dx\,dt\,dy\,d\tau \\ &\equiv \int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}} Q(x,t,y,\tau)(\psi_\epsilon)_x\,dx\,dt\,dy\,d\tau, \end{aligned}$$

where

$$\begin{aligned} Q(x,t,y,\tau) = \operatorname{sign}(s(x,t)-\bar{s}(y,\tau))[&(f(s(x,t),c(x,t))-f(s(x,t),\bar{c}(y,\tau))) \\ &-(f(\bar{s}(y,\tau),c(x,t))-f(\bar{s}(y,\tau),\bar{c}(y,\tau)))]. \end{aligned}$$

Observe that

$$\int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}} Q(y,\tau,y,\tau)(\psi_\epsilon)_x\,dx\,dt\,dy\,d\tau = 0$$

for all $\epsilon > 0$. Define the function

$$H(s,c,\bar{s},\bar{c}) = \operatorname{sign}(s-\bar{s})[(f(s,c)-f(s,\bar{c}))-(f(\bar{s},c)-f(\bar{s},\bar{c}))].$$

Then

$$\begin{aligned} &|Q(x,t,y,\tau)-Q(y,\tau,y,\tau)| \\ &\quad = |H(s(x,t),c(x,t),\bar{s}(y,\tau),\bar{c}(y,\tau))-H(s(y,\tau),c(y,\tau),\bar{s}(y,\tau),\bar{c}(y,\tau))| \\ &\quad \le |H(s(x,t),c(x,t),\bar{s}(y,\tau),\bar{c}(y,\tau))-H(s(y,\tau),c(x,t),\bar{s}(y,\tau),\bar{c}(y,\tau))| \\ &\qquad +|H(s(y,\tau),c(x,t),\bar{s}(y,\tau),\bar{c}(y,\tau))-H(s(y,\tau),c(y,\tau),\bar{s}(y,\tau),\bar{c}(y,\tau))| \\ &\quad \le |H_s(\tilde{s},c(x,t),\bar{s}(y,\tau),\bar{c}(y,\tau))||s(x,t)-s(y,\tau)| \\ &\qquad +|H_c(s(y,\tau),\tilde{c},\bar{s}(y,\tau),\bar{c}(y,\tau))||c(x,t)-c(y,\tau)|, \end{aligned}$$

where $\tilde{s} \in [S_0,1]$ and $\tilde{c} \in [0,1]$. Since

$$H_s(s,c,\bar{s},\bar{c}) = \operatorname{sign}(s-\bar{s})(f_s(s,c)-f_s(s,\bar{c}))$$

for almost all s, and

$$H_c(s,c,\bar{s},\bar{c}) = \operatorname{sign}(s-\bar{s})(f_c(s,c)-f_c(\bar{s},c)),$$

we obtain

$$\begin{aligned} &|Q(x,t,y,\tau)-Q(y,\tau,y,\tau)| \\ &\quad \le M_f\{|c(x,t)-\bar{c}(y,\tau)||s(x,t)-s(y,\tau)|+|s(y,\tau)-\bar{s}(y,\tau)||c(x,t)-c(y,\tau)|\}. \end{aligned}$$

Hence,

$$\begin{aligned} |R_1(\epsilon)| &= \left|\int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}} (Q(x,t,y,\tau)-Q(y,\tau,y,\tau))(\psi_\epsilon)_x\,dx\,dt\,dy\,d\tau\right| \\ &\le M_f\int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}} |c(x,t)-\bar{c}(y,\tau)||s(x,t)-s(y,\tau)| \\ &\qquad \cdot|\omega_\epsilon{}'(x-y)|\omega_\epsilon(t-\tau)\,dx\,dt\,dy\,d\tau \\ &\quad +M_f\int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}} |s(y,\tau)-\bar{s}(y,\tau)||c(x,t)-c(y,\tau)| \\ &\qquad \cdot|\omega_\epsilon{}'(x-y)|\omega_\epsilon(t-\tau)\,dx\,dt\,dy\,d\tau \\ &\equiv I(\epsilon)+II(\epsilon). \end{aligned}$$

By using the Lipschitz continuity of $\bar{c}$, we get

$$\begin{aligned}
I(\epsilon) &= M_f \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} (|c(x,t) - \bar{c}(y,\tau)||s(x,t) - s(y,\tau)|) \\
&\quad \cdot |\omega_\epsilon{}'(x-y)|\omega_\epsilon(t-\tau)\,dx\,dt\,dy\,d\tau \\
&\le M_f \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} (|c(x,t) - \bar{c}(x,t)| + |\bar{c}(x,t) - \bar{c}(y,\tau)|) \\
&\quad \cdot |s(x,t) - s(y,\tau)||\omega_\epsilon{}'(x-y)|\omega_\epsilon(t-\tau)\,dx\,dt\,dy\,d\tau \\
&\le M_f \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} |c(x,t) - \bar{c}(x,t)|(|s(x,t) - s(y,t)| \\
&\qquad\qquad + |s(y,t) - s(y,\tau)|)|\omega_\epsilon{}'(x-y)|\omega_\epsilon(t-\tau)\,dx\,dt\,dy\,d\tau \\
&\quad + \frac{M_f M_\omega^2}{\epsilon^2} \int_0^T \int_{\mathbb{R}} \int_{x-\epsilon}^{x-\epsilon} \int_{t-\epsilon}^{t-\epsilon} |s(x,t) - s(y,\tau)|\,dx\,dt\,dy\,d\tau.
\end{aligned}$$

(Recall that we defined $s(x,t) = s^0(x)$ for $t < 0$.) Here, by Lemma 5.4, the second term tends to zero as ϵ tends to zero. The first term is bounded by using the bound on the total variation and the L^1-Lipschitz continuity in time of s. Observe that

$$\begin{aligned}
&\int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} |c(x,t) - \bar{c}(x,t)|(|s(x,t) - s(y,t)| \\
&\qquad\qquad + |s(y,t) - s(y,\tau)|)|\omega_\epsilon{}'(x-y)|\omega_\epsilon(t-\tau)\,dx\,dt\,dy\,d\tau \\
&\le \int_0^T \int_0^T \omega_\epsilon(t-\tau)\|c(\cdot,t) - \bar{c}(\cdot,t)\|_\infty \left\{ \iint_{\mathbb{R}} |s(x,t) - s(y,t)||\omega_\epsilon{}'(x-y)|dx\,dy \right. \\
&\qquad\qquad \left. + \int_{\mathbb{R}} |s(y,t) - s(y,\tau)| \left[\int_{y-\epsilon}^{y+\epsilon} |\omega_\epsilon{}'(x-y)|dx \right] dy \right\} dt\,d\tau \\
&\le \int_0^T \int_0^T \omega_\epsilon(t-\tau)\|c(\cdot,t) - \bar{c}(\cdot,t)\|_\infty \left\{ \int_{\mathbb{R}} |\omega_\epsilon{}'(z)| \left(\int_{\mathbb{R}} |s(x,t) - s(x-z,t)|\,dx \right) dz \right. \\
&\qquad\qquad \left. + \frac{2M_\omega}{\epsilon}\|s(\cdot,t) - s(\cdot,\tau)\|_1 \right\} dt\,d\tau \\
&\le \int_0^T \|c(\cdot,t) - \bar{c}(\cdot,t)\|_\infty \left[\int_{\mathbb{R}} |\omega_\epsilon{}'(z)|\,TV(s(\cdot,t))\,|z|\,dz \right] dt \\
&\quad + \int_0^T \|c(\cdot,t) - \bar{c}(\cdot,t)\|_\infty \frac{\bar{K}}{\epsilon} \int_0^T \omega_\epsilon(t-\tau)|t-\tau|\,d\tau\,dt \\
&\le K \int_0^T \|c(\cdot,t) - \bar{c}(\cdot,t)\|_\infty\,dt,
\end{aligned}$$

where K and $\bar{K}$ are finite constants independent of ϵ. Since

$$\|c(\cdot,t) - \bar{c}(\cdot,t)\|_\infty \le \|c_x(\cdot,t) - \bar{c}_x(\cdot,t)\|_1,$$

we obtain, by a proper modification of K, that

$$\limsup_{\epsilon \to 0} I(\epsilon) \le K \int_0^T \|c_x(\cdot,t) - \bar{c}_x(\cdot,t)\|_1\,dt.$$

Finally we estimate $II(\epsilon)$ by using the Lipschitz continuity of c,

$$II(\epsilon) = M_f \int_0^T \int_{\mathbb{R}} \int_0^T \int_{\mathbb{R}} |s(y,\tau) - \bar{s}(y,\tau)||c(x,t) - c(y,\tau)|$$

$$
\begin{aligned}
&\cdot |\omega_\epsilon{}'(x-y)|\omega_\epsilon(t-\tau)\,dx\,dt\,dy\,d\tau \\
&\leq \bar{K}\int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}} |s(y,\tau)-\bar{s}(y,\tau)|(|x-y|+|t-\tau|) \\
&\qquad \cdot |\omega_\epsilon{}'(x-y)|\omega_\epsilon(t-\tau)\,dx\,dt\,dy\,d\tau \\
&= \bar{K}\int_0^T\int_{\mathbb{R}} |s(y,\tau)-\bar{s}(y,\tau)|\Big\{ \int_{\mathbb{R}} |x-y||\omega_\epsilon{}'(x-y)|\,dx \int_0^T \omega_\epsilon(t-\tau)\,dt \\
&\qquad\qquad + \int_{\mathbb{R}} |\omega_\epsilon{}'(x-y)|\,dx \int_0^T |t-\tau|\omega_\epsilon(t-\tau)\,dt\Big\}\,dy\,d\tau \\
&\leq K\int_0^T \|s(\cdot,t)-\bar{s}(\cdot,t)\|_1\,dt.
\end{aligned}
$$

Consequently,

$$
\limsup_{\epsilon\to 0} R_1(\epsilon) \leq K\int_0^T (\|c_x(\cdot,t)-\bar{c}_x(\cdot,t)\|_1 + \|s(\cdot,t)-\bar{s}(\cdot,t)\|_1)\,dt. \tag{5.14}
$$

By applying (5.10)–(5.14) we obtain

$$
\begin{aligned}
\|s(\cdot,T)-\bar{s}(\cdot,T)\|_1 \leq \|s^0-\bar{s}^0\|_1 + K\int_0^T \{&\|c_x(\cdot,t)-\bar{c}_x(\cdot,t)\|_1 \\
&+\|c(\cdot,t)-\bar{c}(\cdot,t)\|_1 + \|s(\cdot,t)-\bar{s}(\cdot,t)\|_1\}\,dt,
\end{aligned}
$$

which concludes the proof of Lemma 5.3.

REFERENCES

[1] M. G. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws,* Math. Comp., 34 (1980), pp. 1–21.

[2] R. J. DiPerna, *Finite difference schemes for conservation laws,* Comm. Pure Appl. Math., 35 (1982), pp. 379–450.

[3] ———, *Convergence of approximate solutions to conservation laws,* Arch. Rational Mech. Anal., 82 (1982), pp. 27–70.

[4] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems,* Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[5] E. Isaacson, *Global solution of a Riemann problem for a non-strictly hyperbolic system of conservation laws arising in enhanced oil recovery,* Rockefeller University preprints.

[6] E. Isaacson and B. Temple, *The structure of asymptotic states in a singular system of conservation laws,* Adv. Appl. Math., 11 (1990), pp. 205–219.

[7] T. Johansen and R. Winther, *The solution of the Riemann problem for a hyperbolic system of conservation laws modeling polymer flooding,* SIAM J. Math. Anal., 19 (1988), pp. 541–566.

[8] B. Keyfitz and H. Kranzer, *A system of non-strictly hyperbolic conservation laws arising in elasticity theory,* Arch. Rational Mech. Anal., 72 (1980), pp. 219–241.

[9] ———, *Non-strictly hyperbolic systems of conservation laws: Formation of singularities,* Contemporary Math., J. Smoller, ed., 17 (1983), pp. 77–90.

[10] S. N. Kruzkov, *First order quasilinear equations with several space variables,* Math. USSR Sb., 10 (1970), pp. 217–243.

[11] N. N. Kuznetsov, *Accuracy of some approximate methods for computing the weak solutions of a first order quasilinear equation,* USSR Comput. Math. and Math. Phys., 16 (1976), pp. 105–119.

[12] P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves,* Society for Industrial and Applied Mathematics, Philadelphia, PA, 1973.

[13] ———, *Development of singularities of solution of nonlinear hyperbolic partial differential equations,* J. Math. Phys., 5 (1964), pp. 611–613.

[14] P. D. LAX AND B. WENDROFF, *Systems of conservation laws,* Comm. Pure Appl. Math., 13 (1960), pp. 217–237.
[15] R. J. LEVEQUE AND B. TEMPLE, *Stability of Godunovs Method for a class of* 2×2 *systems of conservation laws,* Trans. Amer. Math. Soc., 288 (1985), pp 115–123.
[16] T. P. LIU, *Uniqueness of weak solutions of the Cauchy problem for general* 2×2 *conservation laws,* J. Differential Equations, 20 (1976), pp. 369–388.
[17] ———, *Development of singularities in the nonlinear waves for quasi-linear hyperbolic partial differential equations,* J. Differential Equations, 33 (1979), pp 92–111.
[18] T. P. LIU AND C. H. WANG, *On a non-strictly hyperbolic system of conservation laws,* J. Differential Equations, 57 (1985), pp. 1–14.
[19] O. OLEINIK, *Discontinuous solutions of nonlinear differential equations,* Amer. Math. Soc. Transl. Ser. 2, 26 (1963), pp. 95–172.
[20] ———, *Uniqueness and stability of the generalized solution of the Cauchy problem for a quasi-linear equation,* Amer. Math. Soc. Transl. Ser. 2, 33 (1964), pp. 285–290.
[21] G. A. POPE, *The application of fractional flow theory to enhanced oil recovery,* Soc. Pet. Engrg. J., 20 (1980), pp. 191–205.
[22] D. SERRE, *Solutions à variations bornées pour certains systèms hyperboliques de lois de conservation,* J. Differential Equations, 68 (1987), pp. 137–168.
[23] ———, *Les ondes planes en electromagnetisme non lineaire,* Phys. D, 31 (1988), pp. 227–251.
[24] J. SMOLLER, *Shock waves and reaction–diffusion equations,* Springer-Verlag, New York, 1982.
[25] B. TEMPLE, *Global solution of the Cauchy problem for a class of* 2×2 *non-strictly hyperbolic conservation laws,* Adv. in Appl. Math., 3 (1982), pp. 335–375.
[26] ———, *Stability and decay in systems of conservation laws,* in Nonlinear Hyperbolic Problems, Proceedings, Carasso, Raviart, and Serre, eds., Springer-Verlag, Berlin, New York, 1986.
[27] ———, *Decay with a rate for noncompactly supported solutions of conservation laws,* Trans. Amer. Math. Soc., 98 (1986), pp. 43–82.
[28] A. TVEITO AND R. WINTHER, *Convergence of a non-conservative finite difference scheme for a system of hyperbolic conservation laws,* J. Differential Intgr. Equations, 3 (1990), pp. 979–1000.
[29] A. TVEITO, *Convergence and stability of the Lax–Friedrichs scheme for a nonlinear parabolic polymer flooding problem,* Adv. in Appl. Math., 11 (1990), pp. 220–246.
[30] W. P. ZIEMER, *Weakly Differentiable Functions,* Springer-Verlag, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 934–956, July 1991

004

DIFFUSION OF PENETRANT IN A POLYMER: A FREE BOUNDARY PROBLEM*

BEI HU†

Abstract. A free boundary problem arising in modeling diffusion of a penetrant in a polymer is studied. The asymptotic behavior of the solution for short time and long time, for small and large physical parameter ϵ, and for small and large driving law exponent n are proved. Some explicit error estimates are also given.

Key words. free boundary, variational inequality, asymptotic behavior

AMS(MOS) subject classifications. 35B40, 35R35

1. Introduction.

A model describing the diffusion of a penetrant in a glassy polymer is given by

$$\epsilon u_t = u_{xx} \quad \text{for } 0 < x < s(t), \quad t > 0, \tag{1.1}$$

$$u(0,t) = 1, \tag{1.2}$$

$$[1 + \epsilon u(s(t),t)] \cdot s'(t) = -u_x(s(t),t), \tag{1.3}$$

$$s'(t) = u^n(s(t),t), \tag{1.4}$$

$$s(0) = 0, \tag{1.5}$$

where u is the penetrant concentration over its equilibrium value, s is the penetrant front driven by u. The driving law (1.4) expresses the kinetics of swelling and is assumed to be n-order type. $1/\epsilon$ is the diffusivity in the swollen polymer (see [1], [3]). In special examples in [3], n takes values varying from 10^{-2} to 10^2, ϵ need not be small, and it is an interesting problem to study the dependence of the solution on the parameter ϵ.

The problem with $\epsilon = 1$, (1.3) replaced by $[q + u(s(t),t)] \cdot s'(t) = -u_x(s(t),t)$ $(q \geq 0)$ and (1.4) replaced by $s'(t) = f(u(s(t),t))$ was studied in [4] by Fasano, Meyer, and Primicerio. Their results imply, in particular, the following theorem.

THEOREM 1.1. *Problem* (1.1)–(1.5) *has a unique classical solution* (u,s) *such that* $s \in C^2[0,\infty) \cap C^\infty(0,\infty)$, $u \in C^\infty(D_\infty) \cap C^{2,1}(\overline{D_\infty})$, *where* $D_\infty = \{(x,t) : 0 < x < s(t),\ 0 < t < \infty\}$; *furthermore,*

$$0 < u(x,t) < 1 \quad \textit{for } (x,t) \in D_\infty, \tag{1.6}$$

$$-1 - \epsilon < u_x(x,t) < 0 \quad \textit{for } (x,t) \in D_\infty, \tag{1.7}$$

$$u_t(x,t) > 0 \quad \textit{for } (x,t) \in D_\infty, \tag{1.8}$$

$$s''(t) < 0 \quad \textit{for } 0 \leq t < \infty. \tag{1.9}$$

They also studied the long time behavior of $s(t)$ $(t \to +\infty)$ and proved (for $\epsilon = 1$) that

$$\sqrt{\frac{3}{2}}(1 - o(t)) \leq \frac{s(t)}{\sqrt{t}} \leq \sqrt{2} \qquad \text{for } t \to \infty.$$

* Received by the editors November 20, 1989; accepted for publication (in revised form) September 5, 1990. This work was partially supported by University of Minnesota Graduate School dissertation fellowship.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. Present address, Department of Mathematics, University of Notre Dame, Notre Dame, Indiana 46556.

More recently, Cohen and Erneux [3] obtained more precise asymptotic behavior

$$\lim_{t\to\infty} s(t)s'(t) = \frac{M^2}{2},$$

$$\lim_{t\to\infty} u(ys(t),t) = 1 - \int_0^{My} \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi \Big/ \int_0^{M} \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi, \tag{1.10}$$

where M is given by

$$\exp\left(-\frac{\epsilon}{4}M^2\right) = \frac{1}{2}M\int_0^M \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi;$$

note that (1.10) implies that

$$\lim_{t\to\infty} \frac{s(t)}{\sqrt{t}} = M. \tag{1.11}$$

They also obtained similar formulas for the asymptotic behavior of $u(x,t)$ and $s(t)$ as $t \to 0$ and as $\epsilon \to 0$:

$$u_0(x,t) - C_1\epsilon \le u_\epsilon(x,t) \le u_0(x,t) + C_2\epsilon \quad \text{for } 0 < x < s_\epsilon(t), \quad 0 < t < \infty, \tag{1.12}$$

$$1 - C_3\epsilon < \frac{s_0(t)}{s_\epsilon(t)} < 1 + C_4\epsilon \quad \text{for } 0 < t < \infty, \tag{1.13}$$

where (u_ϵ, s_ϵ) denotes the solution of (1.1)–(1.5); u_0, s_0 are indepedent of ϵ. However, all these results were obtained by formal power series expansions.

In §§1–4, we shall give rigorous justification to these facts. In §2 we consider the short time behavior for (u,s) and give explicit error estimates. In §3 we study the long time behavior and prove (1.10) and (1.11). In §4 we establish (1.12) and (1.13) with $C_1 = 5\max(n,1)$, $C_2 = 0$, $C_3 = 0$, and $C_4 = 4\max(n,1)$. And in §5 we shall prove that $\lim_{\epsilon\to\infty} s_\epsilon(t) = 0$.

Sections 6 and 7 are devoted to studying the effects of the driving law exponent n in (1.4). Denote by (u_n, s_n) the solution of (1.1)–(1.5). In §6 we shall show that $s_n(t) \to 0$ when $n \to \infty$. In §7, we investigate the case $n \to 0$. We shall prove that there exists a critical value $T^* > 0$ such that

$$\lim_{n\to 0}(u_n(x,t), s_n(t)) = (z(x,t),t) \quad \text{for } t \in [0,T^*] \tag{1.14}$$

and

$$\lim_{n\to 0}(u_n(x,t), s_n(t)) = (u_0(x,t), s_0(t)) \quad \text{for } t \in [T^*,\infty), \tag{1.15}$$

where (u_0, s_0) is the solution of an appropriate Stefan problem, and the formulas for $z(x,t)$ and T^* will also be given.

Consider next the case where condition (1.2) for maintaining constant concentration at $x = 0$ is replaced by flux control, namely,

$$u_x(0,t) = g(t). \tag{1.16}$$

THEOREM 1.2 (Anderucci and Ricci [2]). *If*

$$g \in C^1[0,\infty), \quad g(t) \le 0 \quad \text{for } t > 0, \quad g(0) < 0, \tag{1.17}$$

then the problem (1.1), (1.16), (1.3)–(1.5) *has a unique classical solution* (u,s) *with* $s \in C^1[0,\infty)$, $u \in C^{2,1}(D_\infty) \cap C^{1,0}(\overline{D_\infty})$, *where* $D_\infty = \{(x,t) : 0 < x < s(t),\ 0 < t < \infty\}$; *furthermore,*

$$u(x,t) > 0 \quad \text{for } (x,t) \in \overline{D_\infty}, \tag{1.18}$$

$$u_x(x,t) < 0 \quad \text{for } (x,t) \in D_\infty. \tag{1.19}$$

If in addition to (1.17), $g'(t) \ge 0$ *for* $t \ge 0$, *then*

$$s''(t) \le 0 \quad \text{for } 0 \le t < \infty. \tag{1.20}$$

It was also proved in [2] that if $\int_0^\infty g(t)dt > -\infty$, then $\lim_{t\to\infty} s(t)$ exists; the limit was also computed.

In §8 of this paper, we shall study the long time behavior of $s(t)$ when the condition $\int_0^\infty g(t)dt > -\infty$ is dropped. We shall prove that $x = s(t)$ approaches the free boundary of a Stefan problem; in particular, it follows that $\lim_{t\to\infty} s(t) = \infty$ in the case $\int_0^\infty g(t)dt = -\infty$, in contrast to the case $\int_0^\infty g(t)dt > -\infty$, where $\lim_{t\to\infty} s(t)$ is finite.

2. Short time behavior.

THEOREM 2.1. *For the solution of* (1.1)–(1.5), *we have*

$$0 \le u(x,t) - [1-(1+\epsilon)x] \le Ct^2 \quad \text{for } 0 \le x \le s(t), \quad 0 \le t \le \frac{1}{2(1+\epsilon)}, \tag{2.1}$$

$$-C_1(1+\epsilon)^2t^3 \le s(t) - \left(t - \frac{n}{2}(1+\epsilon)t^2\right) \le C_2(1+\epsilon)^2t^3 \quad \text{for } 0 \le t \le \frac{1}{2(1+\epsilon)}, \tag{2.2}$$

where $C = (1+\epsilon)(2n\epsilon+\epsilon+2n)$, $C_1 = 0$ *if* $n \ge 1$ *and* $C_1 = 1/6$ *if* $0 < n < 1$, *and* C_2 *is a constant depending only on* n.

Proof. Let $w(x,t) = 1-(1+\epsilon)x$; then obviously $\epsilon w_t - w_{xx} = 0$, $w(0,t) = u(0,t)$ and

$$w_x(s(t),t) + [1+\epsilon w(s(t),t)]w^n(s(t),t) \le -(1+\epsilon) + (1+\epsilon) = 0. \tag{2.3}$$

Hence, by maximum principle, $w(x,t) \le u(x,t)$ for $0 \le x \le s(t)$, $t \ge 0$.

Next, let $\widetilde{w}(x,t) = 1-(1+\epsilon)x + Cxt$; then $\epsilon\widetilde{w}_t - \widetilde{w}_{xx} = C\epsilon x \ge 0$ for $x \ge 0$ and $\widetilde{w}(0,t) = u(0,t)$. Since $s'(t) = u^n(s(t),t) \in (0,1)$, we have $0 < s(t) < t$, and hence

$$\begin{aligned}
\{\widetilde{w}_x + (1+\epsilon\widetilde{w})\widetilde{w}^n\}\Big|_{x=s(t)} &\ge -1-\epsilon+Ct+[1+\epsilon-\epsilon(1+\epsilon)s(t)][1-(1+\epsilon)s(t)]^n \\
&\ge -1-\epsilon+Ct+[1+\epsilon-\epsilon(1+\epsilon)t][1-(1+\epsilon)t]^n \\
&\equiv a(t) \quad \text{for } 0 \le x \le s(t), \quad 0 \le t \le \frac{1}{2(1+\epsilon)}.
\end{aligned}$$

For $0 \le t \le 1/[2(1+\epsilon)]$,

$$\begin{aligned}
a'(t) &= C - \epsilon(1+\epsilon)[1-(1+\epsilon)t]^n - n(1+\epsilon)[1+\epsilon-\epsilon(1+\epsilon)t][1-(1+\epsilon)t]^{n-1} \\
&\ge C - \epsilon(1+\epsilon) - n(1+\epsilon)^2 \cdot 2 \\
&= 0,
\end{aligned}$$

and thus $a(t) \geq a(0) = 0$ for $0 \leq t \leq 1/[2(1+\epsilon)]$. Now (2.1) follows by maximum principle.

By (2.1),

$$\begin{aligned} s'(t) &\geq [1-(1+\epsilon)s(t)]^n \\ &\geq 1-n(1+\epsilon)s(t)-3C_1[(1+\epsilon)s(t)]^2 \\ &\geq 1-n(1+\epsilon)t-3C_1[(1+\epsilon)t]^2, \end{aligned}$$

where we can take $C_1 = 0$ if $n \geq 1$ and $C_1 = 1/6$ if $0 < n < 1$. Hence

$$s(t) \geq t - \frac{n}{2}(1+\epsilon)t^2 - C_1(1+\epsilon)^2 t^3.$$

Next, using (2.1) again, we get

$$\begin{aligned} s'(t) &\leq [1-(1+\epsilon)s(t)+C[(1+\epsilon)t]^2]^n \\ &\leq 1-n(1+\epsilon)s(t)+C[(1+\epsilon)t]^2, \end{aligned}$$

where the constant C depends only on n. Solving the above inequality we obtain

$$\begin{aligned} s(t) &\leq \frac{1-\exp[-n(1+\epsilon)t]}{n(1+\epsilon)} + C(1+\epsilon)^2 \int_0^t \exp[-n(1+\epsilon)(t-\tau)]\tau^2 d\tau \\ &\leq t - \frac{n}{2}(1+\epsilon)t^2 + C(1+\epsilon)^2 t^3. \end{aligned}$$

□

3. Long time behavior. Define $\phi(x,t)$ to be the corresponding solution of the Stefan problem, i.e.,

$$\begin{aligned} &\epsilon\phi_t = \phi_{xx} \quad \text{for } 0 < x < h(t), \quad t > 0, \\ &\phi(0,t) = 1, \\ &h'(t) = -\phi_x(h(t),t), \\ &\phi(h(t),t) = 0, \\ &h(0) = 0. \end{aligned}$$

It is well known that

$$\phi(x,t) = 1 - \int_0^{x/\sqrt{t}} \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi \Big/ \int_0^M \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi,$$

(3.1)

$$h(t) = M\sqrt{t},$$

where $M = M(\epsilon)$ is such that

$$\exp\left(-\frac{\epsilon}{4}M^2\right) = \frac{1}{2}M \int_0^M \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi. \tag{3.2}$$

The main result of this section is Theorem 3.1.

THEOREM 3.1. *Suppose* (u,s) *is a solution of* (1.1)–(1.5). *Then*

$$\lim_{t\to\infty} s(t)s'(t) = \frac{M^2}{2}, \tag{3.3}$$

which implies that

$$\lim_{t\to\infty} \frac{s(t)}{\sqrt{t}} = M. \tag{3.4}$$

Furthermore,

$$\lim_{t\to\infty} \sup_{0\le x\le s(t)} |u(x,t) - \phi(x,t)| = 0, \tag{3.5}$$

where ϕ is given by (3.1). □

In order to prove (3.3), we need to prove (3.4) first. We shall compare the solution (u, s) of (1.1)–(1.5) to the solution ϕ of the Stefan problem. Let us first transform our problem into a variational inequality. Set

$$\begin{aligned} v(x,t) &= \int_{s^{-1}(x)}^{t} \left[u(x,\tau) - (s'(\tau))^{1/n}\right] d\tau \quad \text{for } x < s(t) \\ &\equiv 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{for } x \ge s(t), \end{aligned} \tag{3.6}$$

where $s^{-1}(x)$ is the inverse function of $s(t)$; it is C^1 since $s'(t) > 0$. A calculation shows that v satisfies the following variational inequality:

$$\epsilon v_t - v_{xx} = -1 - \epsilon(s'(t))^{1/n} \quad \text{for } 0 < x < s(t), \quad t > 0, \tag{3.7}$$

$$v = v_x = 0 \quad \text{on } x = s(t), \quad t > 0, \tag{3.8}$$

$$v = \int_0^t \left[1 - (s'(\tau))^{1/n}\right] d\tau \quad \text{on } x = 0, \quad t > 0, \tag{3.9}$$

$$v = 0 > -1 - \epsilon(s'(t))^{1/n} \quad \text{for } x > s(t), \quad t > 0, \tag{3.10}$$

$$v > 0 \quad \text{for } x < s(t), \quad t > 0 \quad \text{(by using } v_x < 0 \text{ and (3.8))}. \tag{3.11}$$

LEMMA 3.2.

$$s(t) \le M\sqrt{t} \quad \textit{for all } t > 0. \tag{3.12}$$

Proof. Let

$$\begin{aligned} w(x,t) &= \int_{x^2/M^2}^{t} \phi(x,\tau)d\tau \quad \text{for } x < M\sqrt{t} \\ &\equiv 0 \qquad\qquad\qquad\quad \text{for } x \ge M\sqrt{t}. \end{aligned}$$

Then w satisfies the variational inequality:

$$\epsilon w_t - w_{xx} = -1 \quad \text{for } 0 < x < M\sqrt{t}, \quad t > 0, \tag{3.13}$$

$$w = w_x = 0 \quad \text{on } x = M\sqrt{t}, \quad t > 0, \tag{3.14}$$

$$w = t \quad \text{on } x = 0, \quad t > 0, \tag{3.15}$$

$$w = 0 > -1 \quad \text{for } x > M\sqrt{t}, \quad t > 0, \tag{3.16}$$

$$w > 0 \quad \text{for } x < M\sqrt{t}, \quad t > 0. \tag{3.17}$$

Therefore, by comparison principle for variational inequalities (see [7], for example) we get $v(x,t) \le w(x,t)$ for $0 < x < \infty, t > 0$, and hence $s(t) \le M\sqrt{t}$. □

Next, we prove Lemma 3.3.

LEMMA 3.3.

$$\lim_{t\to\infty} \frac{s(t)}{\sqrt{t}} = M. \tag{3.18}$$

Proof. By Theorem 5.2 of [4], we have

$$\lim_{t\to\infty} s'(t) = 0; \tag{3.19}$$

therefore for any $\eta > 0$, there exist $T > 0$ such that

$$0 < \epsilon(s'(t))^{1/n} \le \eta \quad \text{for } t \ge T. \tag{3.20}$$

Let N_η be the solution of

$$(1-\eta)\exp\left(-\frac{\epsilon}{4}N^2\right) = (1+\eta)\frac{1}{2}N\int_0^N \exp\left(-\frac{\epsilon}{4}\xi^2\right)d\xi; \tag{3.21}$$

it is then clear that $N_\eta < M$ and $N_\eta \to M$ as $\eta \to 0$. Next, set

$$\psi(x,t) = (1-\eta)\left(1 - \frac{\displaystyle\int_0^{x/\sqrt{t-T}} \exp\left(-\frac{\epsilon}{4}\xi^2\right)d\xi}{\displaystyle\int_0^{N_\eta} \exp\left(-\frac{\epsilon}{4}\xi^2\right)d\xi}\right) \quad \text{for } t \ge T. \tag{3.22}$$

Then

$$\begin{aligned}
&\epsilon\psi_t = \psi_{xx} \quad \text{for } 0 < x < N_\eta\sqrt{t-T}, \quad t > T,\\
&\psi(0,t) = 1-\eta \quad \text{for } t \ge T,\\
&\psi = 0, \quad -\psi_x = (1+\eta)\frac{d}{dt}\left(N_\eta\sqrt{t-T}\right) \quad \text{on } x = N_\eta\sqrt{t-T}.
\end{aligned}$$

Repeating the proof of Lemma 3.2 we find that

$$s(t) \ge N_\eta\sqrt{t-T} \quad \text{for } t \ge T; \tag{3.23}$$

it follows that

$$\liminf_{t\to\infty} \frac{s(t)}{\sqrt{t}} \ge N_\eta, \tag{3.24}$$

and we conclude the lemma by letting $\eta \to 0$. □

LEMMA 3.4.

$$\lim_{t\to\infty} \sup_{0\le x\le s(t)} |u(x,t) - \phi(x,t)| = 0, \tag{3.25}$$

where ϕ is given by (3.1).

Proof. Set

$$k(x,t) = \frac{1}{t^\alpha}\left(2 - \frac{x^2}{s^2(t)}\right),$$

where $\alpha > 0$ is to be determined. Then

$$\begin{aligned}
\epsilon k_t - k_{xx} &= -\frac{\epsilon\alpha}{t^{\alpha+1}}\left(2 - \frac{x^2}{s^2(t)}\right) + \frac{\epsilon}{t^\alpha}\frac{2x^2 s'(t)}{s^3(t)} + \frac{1}{t^\alpha}\frac{2}{s^2(t)} \\
&\geq \frac{2}{t^{\alpha+1}}\left(-\epsilon\alpha + \frac{t}{s^2(t)}\right) \\
&\geq \frac{2}{t^{\alpha+1}}\left(-\epsilon\alpha + \frac{1}{M^2}\right) \quad \text{(by Lemma 3.2)} \\
&> 0 \quad \text{for } t > 0,\ 0 < x < s(t)
\end{aligned}$$

if α is small enough. For $T > 0$, we set

$$w(x,t) = \phi(x,t) + \sup_{\tau \geq T}(s'(\tau))^{1/n} + T^\alpha k(x,t).$$

Then $\epsilon w_t - w_{xx} > 0$ for $0 < x < s(t), t > T$; $w(0,t) \geq 1 = u(0,t)$ for $t \geq T$; $w(s(t),t) \geq \sup_{\tau \geq T}(s'(\tau))^{1/n} \geq u(s(t),t)$ for $t \geq T$ and $w(x,T) \geq 1 \geq u(x,T)$ for $0 \leq x \leq s(t)$. Therefore by maximum principle,

$$w(x,t) \geq u(x,t) \quad \text{for } 0 \leq x \leq s(t), \quad t \geq T.$$

Hence

$$\limsup_{t\to\infty} \sup_{0\leq x\leq s(t)} [u(x,t) - \phi(x,t)] \leq \sup_{\tau \geq T}(s'(\tau))^{1/n}.$$

Letting $T \to \infty$, we obtain

$$\limsup_{t\to\infty} \sup_{0\leq x\leq s(t)} [u(x,t) - \phi(x,t)] \leq 0. \tag{3.26}$$

Next, by Lemma 3.3,

$$\lim_{t\to\infty} \phi(s(t),t) = 0.$$

Thus by using the subsolution

$$\widetilde{w}(x,t) = \phi(x,t) - \sup_{\tau \geq T} \phi(s(\tau),\tau) - T^\alpha k(x,t),$$

to estimate u from below and letting $T \to \infty$, we get the complement of (3.26), which completes the proof. □

LEMMA 3.5.

$$\limsup_{t\to\infty} s(t)s'(t) \leq \frac{M^2}{2}. \tag{3.27}$$

Proof. By (1.8), $u_{xx} = \epsilon u_t \geq 0$ for $0 \leq x \leq s(t)$, $t > 0$. Hence (as in [3, Prop. 5.1])

$$\begin{aligned}
u(x,t) &\geq u(s(t),t) + u_x(s(t),t)(x - s(t)) \\
&\geq -(1 + \epsilon u(s(t),t))s'(t)(x - s(t)).
\end{aligned}$$

Letting $x = \gamma s(t)$ $(0 < \gamma < 1)$ in the above inequality and letting $t \to \infty$, we obtain

$$\begin{aligned}(1-\gamma)\limsup_{t\to\infty} s(t)s'(t) &\le \limsup_{t\to\infty} \frac{u(\gamma s(t),t)}{1+\epsilon u(s(t),t)} \\ &\le \limsup_{t\to\infty} \phi(\gamma s(t),t) \quad \text{(by Lemma 3.4)} \\ &= 1 - \int_0^{\gamma M} \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi \Big/ \int_0^{M} \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi \\ &\qquad \text{(by Lemma 3.3).}\end{aligned}$$

Dividing by $1-\gamma$ and then letting $\gamma \to 1^-$, we immediately obtain (upon recalling (3.2)) the estimate (3.27). □

LEMMA 3.6. *There exist positive constants C and T such that*

$$u_t(x,t) \le \frac{C}{t} \quad \text{for } 0 \le x \le s(t), \quad T \le t < \infty. \tag{3.28}$$

Proof. It is clear that $\epsilon u_{tt} - u_{txx} = 0$ for $0 < x < s(t)$, $t > 0$, and $u_t(0,t) = 0$ for $t > 0$. Differentiating (1.4) and using $s''(t) < 0$ we get that $u_t(s(t),t) + u_x(s(t),t)s'(t) < 0$. Hence

$$\begin{aligned}u_t(s(t),t) &\le -u_x(s(t),t)s'(t) \\ &= (1+\epsilon u(s(t),t))(s'(t))^2 \quad \text{(by (1.3))} \\ &\le \frac{C^*}{t} \quad \text{for } 0 < t < \infty \quad \text{(by Lemmas 3.3 and 3.5).}\end{aligned}$$

Next, let

$$\theta(x,t) = -\int_0^{(x+1)/\sqrt{t}} \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi.$$

Then $\epsilon\theta_t = \theta_{xx}$ for $t > 0$; and by Lemma 3.3, there exist $T, c_0 > 0$ such that

$$\begin{aligned}\theta_t(s(t),t) &= \exp\left(-\frac{\epsilon}{4}\frac{(s(t)+1)^2}{t}\right)\frac{s(t)+1}{2t\sqrt{t}} \\ &\ge \frac{c_0}{t} \quad \text{for } t \ge T.\end{aligned}$$

It is obvious that $\theta_t(0,t) > 0$ and

$$\inf_{0\le x\le s(T)} \theta_t(x,T) = \inf_{0\le x\le s(T)} \exp\left(-\frac{\epsilon}{4}\frac{(x+1)^2}{T}\right)\frac{x+1}{2T\sqrt{T}} \equiv c_1 > 0.$$

Therefore by maximum principle,

$$u_t(x,t) \le C\theta_t(x,t) \quad \text{for } 0 \le x \le s(t), \quad t \ge T$$

if C is large enough so that $Cc_0 \ge C^*$ and $Cc_1 \ge \max_{0\le x\le s(T)} u_t(x,T)$. It follows that (3.28) holds. □

Proof of Theorem 3.1. Since we have already proved (3.4), (3.5), and (3.27), it remains only to show that

$$\liminf_{t\to\infty} s(t)s'(t) \ge \frac{M^2}{2}. \tag{3.29}$$

By Lemma 3.6,

$$\begin{aligned} u(x,t) &= u(s(t),t) + u_x(s(t),t)(x-s(t)) + \int_x^{s(t)} (\xi - x) u_{xx}(\xi,t)\, d\xi \\ &\le u(s(t),t) + u_x(s(t),t)(x-s(t)) + \frac{C}{t}\frac{(s(t)-x)^2}{2} \quad \text{for } T \le t < \infty. \end{aligned}$$

Similarly to the proof of Lemma 3.5, we can now obtain, for $0 < \gamma < 1$ (note that $\lim_{t\to\infty} u(s(t),t) = 0$)

$$\begin{aligned} &(1-\gamma) \liminf_{t\to\infty} \left(s(t)s'(t) + \frac{C}{2}(1-\gamma)\frac{s^2(t)}{t} \right) \\ &\quad \ge 1 - \int_0^{\gamma M} \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi \Big/ \int_0^{M} \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi. \end{aligned}$$

Since $s^2(t)/t$ is bounded from above, the above inequality implies, after dividing by $1-\gamma$ and letting $\gamma \to 1^-$, that (3.29) holds. □

4. Small ϵ. As in [3], by formally letting $\epsilon \to 0$ in equations (1.1)–(1.5), we obtain equations for the limit functions (u_0, s_0):

$$u_0(x,t) = 1 - B(t)x, \tag{4.1}$$

$$s_0'(t) = (1 - B(t)s_0(t))^n = B(t), \tag{4.2}$$

$$B(0) = 1, \qquad s_0(0) = 0. \tag{4.3}$$

From this, we can uniquely determine $B(t)$ and $s_0(t)$; in particular $B(t)$ satisfies

$$t + \frac{1}{2} + \frac{n-1}{1-2n} = \frac{1}{2}B^{-2} + \frac{n-1}{1-2n}B^{(1/n)-2} \quad \text{if } n \ne \frac{1}{2}, \tag{4.4}$$

$$t + \frac{1}{2} = \frac{1}{2}B^{-2} + \log\frac{1}{B} \quad \text{if } n = \frac{1}{2}. \tag{4.5}$$

It easily follows that

$$0 < B(t) < 1, \qquad 0 < s_0(t)B(t) < 1, \tag{4.6}$$

$$\frac{dB}{dt} = -\frac{B^3}{1 + [(1-n)/n]B^{1/n}} < 0. \tag{4.7}$$

It is also clear that $\lim_{t\to\infty} s_0(t)/\sqrt{t} = \sqrt{2}$. Now we shall prove that the solution (u_ϵ, s_ϵ) of (1.1)–(1.5) converges to (u_0, s_0).

THEOREM 4.1. *We have, for $\epsilon > 0$,*

$$u_\epsilon(x,t) \le 1 - B(t)x \quad \textit{for } 0 \le x \le s_\epsilon(t), \quad 0 \le t < \infty, \tag{4.8}$$

$$s_\epsilon(t) < s_0(t) \quad \textit{for } 0 \le t < \infty. \tag{4.9}$$

Proof. (i) First we show that if (4.9) holds true for $0 < t < t_0$, then (4.8) must be true for $0 \le t \le t_0$. In fact, the function $w = 1 - B(t)x$ satisfies $w_t - \epsilon^{-1}w_{xx} =$

$-B'(t)x \geq 0$ for $0 < x < s_\epsilon(t)$, $t > 0$, and $w(0,t) = u_\epsilon(0,t)$ for $t \geq 0$. Since $1 - B(t)s_\epsilon(t) > 1 - B(t)s_0(t) > 0$ for $0 < t < t_0$,

$$\begin{aligned} \{w_x + (1+\epsilon w)w^n\}\Big|_{x=s_\epsilon(t)} &\geq -B(t) + (1 - B(t)s_\epsilon(t))^n \\ &\geq -B(t) + (1 - B(t)s_0(t))^n \\ &= 0 \quad \text{for } 0 \leq t \leq t_0. \end{aligned} \tag{4.10}$$

Therefore, by maximum principle, $u(x,t) \leq w(x,t)$ for $0 \leq t \leq t_0$, i.e., (4.8) holds for $0 \leq t \leq t_0$.

(ii) Next, we prove that (4.9) holds for small t. In fact, from (4.1)–(4.6) it follows that

$$\begin{aligned} s_0'(t) &\geq 1 - nB(t)s_0(t) - C[B(t)s_0(t)]^2 \\ &\geq 1 - nt - Ct^2; \end{aligned}$$

therefore $s_0(t) \geq t - (n/2)t^2 - Ct^3$. Hence, by (2.2), (4.9) holds for small $t > 0$.

(iii) It now follows that if the theorem is not true, then there exists a $T > 0$ such that

$$s_\epsilon(t) < s_0(t) \quad \text{for } 0 < t < T, \tag{4.11}$$

$$s_\epsilon(T) = s_0(T). \tag{4.12}$$

From (i), we obtain

$$s_\epsilon'(t) = u_\epsilon^n(s_\epsilon(t), t) \leq [1 - B(t)s_\epsilon(t)]^n \quad \text{for } 0 \leq t \leq T. \tag{4.13}$$

Note that $1 - B(t)s_\epsilon(t) \geq 1 - B(t)s_0(t) > c_0 > 0$ for $0 \leq t \leq T$, and the function $f(u) = u^n$ is Lipschitz continuous for $u \geq c_0$. Therefore, by using (4.2), (4.13), and (ii), we can apply the comparison principle of ODE and get

$$s_\epsilon(t) < s_0(t) \quad \text{for } 0 < t \leq T, \tag{4.14}$$

which contradicts (4.12) at $t = T$. □

THEOREM 4.2. *There exists a $\epsilon_0 > 0$ such that, for all $0 < \epsilon < \epsilon_0$,*

$$u_\epsilon(x,t) \geq 1 - B(t)x - 5\max(n,1)\epsilon \quad \textit{for } 0 \leq x \leq s_\epsilon(t), \quad 0 \leq t < \infty, \tag{4.15}$$

$$[1 + 4\max(n,1)\epsilon]s_\epsilon(t) > s_0(t) \quad \textit{for } 0 \leq t < \infty. \tag{4.16}$$

Proof. (i) Set

$$w = 1 - (1 + C^*\epsilon)B(t)x - \frac{C^*\epsilon}{4}B(t)\left(s_\epsilon(t) - \frac{x^2}{s_\epsilon(t)}\right), \tag{4.17}$$

where $C^* = 4\max(n,1)$, then

$$\begin{aligned} &w_t - \epsilon^{-1}w_{xx} \\ &\quad = -(1 + C^*\epsilon)B'(t)x - \frac{C^*\epsilon}{4}B'(t)\left(s_\epsilon(t) - \frac{x^2}{s_\epsilon(t)}\right) \end{aligned}$$

$$
\begin{aligned}
&-\frac{C^*\epsilon}{4}B(t)\left(s'_\epsilon(t)+\frac{x^2}{s^2_\epsilon(t)}s'_\epsilon(t)\right)-\epsilon^{-1}\frac{C^*\epsilon}{2}\frac{B(t)}{s_\epsilon(t)}\\
\le\ & -(1+C^*\epsilon)B'(t)s_\epsilon(t)-\frac{C^*\epsilon}{4}B'(t)s_\epsilon(t)-\frac{C^*}{2}\frac{B(t)}{s_\epsilon(t)}\\
=\ & \frac{1}{s_\epsilon(t)}\left[-\left(1+C^*\epsilon+\frac{C^*\epsilon}{4}\right)B'(t)s^2_\epsilon(t)-\frac{C^*}{2}B(t)\right]\\
\le\ & \frac{1}{s_\epsilon(t)}\left[\left(1+C^*\epsilon+\frac{C^*\epsilon}{4}\right)\frac{B^3(t)}{1+[(1-n)/n]B^{1/n}}s_0^2(t)-\frac{C^*}{2}B(t)\right]\\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by (4.7), (4.9))}\\
\le\ & \frac{1}{s_\epsilon(t)}\left[\left(1+C^*\epsilon+\frac{C^*\epsilon}{4}\right)\max(n,1)B(t)-\frac{C^*}{2}B(t)\right]\quad \text{(by (4.6))}\\
<\ & 0
\end{aligned}
$$

if ϵ is small enough so that $C^*\epsilon + C^*\epsilon/4 < 1$.

Now we show that if (4.16) holds for $0 < t < t_0$, then

$$u_\epsilon(x,t) \ge w(x,t) \quad \text{for } 0 \le x \le s_\epsilon(t),\quad 0 \le t \le t_0, \tag{4.18}$$

which implies that (4.15) is valid for $0 \le t \le t_0$. Clearly, $w(0,t) = u(0,t)$ and

$$
\begin{aligned}
&\{w_x^+ + (1+\epsilon w)(w^+)^n\}\Big|_{x=s_\epsilon(t)}\\
&\quad\le\ -(1+C^*\epsilon)B(t)+\frac{C^*\epsilon}{2}B(t)+(1+\epsilon)\{[1-(1+C^*\epsilon)B(t)s_\epsilon(t)]^+\}^n\\
&\quad\le\ -\left(1+\frac{C^*\epsilon}{2}\right)B(t)+(1+\epsilon)[1-B(t)s_0(t)]^n\quad \text{(by (4.16))}\\
&\quad=\ -\left(1+\frac{C^*\epsilon}{2}\right)B(t)+(1+\epsilon)B(t)\\
&\quad<\ 0\qquad \text{for } 0\le t\le t_0.
\end{aligned}
$$

It follows by maximum principle that (4.18) holds.

(ii) It is clear that $(1+C^*\epsilon)s_\epsilon(t) = (1+C^*\epsilon)t + O(t^2)$ as $t \to 0$, and therefore (4.16) holds for small t. As in the proof of Theorem 4.1, if the theorem is not true, then there exists a $T > 0$ such that

$$(1+C^*\epsilon)s_\epsilon(t) > s_0(t) \quad \text{for } 0 < t < T, \tag{4.19}$$

$$(1+C^*\epsilon)s_\epsilon(T) = s_0(T). \tag{4.20}$$

From (4.18), we obtain

$$(1+C^*\epsilon)s'_\epsilon(t) > s'_\epsilon(t) = u^n_\epsilon(s_\epsilon(t),t) \ge \left\{[1-(1+C^*\epsilon)B(t)s_\epsilon(t)]^+\right\}^n. \tag{4.21}$$

Clearly,

$$\max_{0\le t\le T} B(t)s_0(t) \equiv \lambda < 1;$$

therefore, we can take $\lambda < \overline{\lambda} < 1$. Set

$$G = \{t \in [0,T] :\ (1+C^*\epsilon)B(t)s_\epsilon(t) \ge \overline{\lambda}\}. \tag{4.22}$$

Then for $t \in G$, we have

$$(1 + C^*\epsilon)s_\epsilon(t) \geq \frac{\overline{\lambda}}{B(t)} > s_0(t),$$

whereas for $t \in [0,T] \setminus G$ we have

$$1 - (1 + C^*\epsilon)B(t)s_\epsilon(t) \geq 1 - \overline{\lambda} > 0. \tag{4.23}$$

The set $[0,T] \setminus G$ is open and hence consists of open intervals with endpoints in G. The function $f(u) = u^n$ is Lipschitz continuous for $u \geq 1 - \overline{\lambda}$. Therefore by (4.2) and (4.21), we can apply the comparison principle of ODE and get that (4.19) holds also for $t \in [0,T] \setminus G$, which is a contradiction to (4.20) at $t = T$. □

Remark. Similar result of this section is obtained in [11].

5. Large ϵ. The diffusivity is small if ϵ is large. Therefore the penetrant front $s_\epsilon(t)$ should move very slowly when ϵ is large. This property is described in the following theorem.

THEOREM 5.1. *For the solution* (u_ϵ, s_ϵ) *of* (1.1)–(1.5), *we have*

$$\frac{s_\epsilon(t)}{\sqrt{t}} \to 0 \quad \text{as } \epsilon \to +\infty \tag{5.1}$$

uniformly for $t \in (0,\infty)$.

Proof. By Lemma 3.2,

$$\frac{s_\epsilon(t)}{\sqrt{t}} \leq M(\epsilon) \quad \text{for } 0 < t < \infty, \tag{5.2}$$

where $M(\epsilon)$ satisfies

$$\exp\left(-\frac{\epsilon}{4}M^2(\epsilon)\right) = \frac{1}{2}M(\epsilon)\int_0^{M(\epsilon)} \exp\left(-\frac{\epsilon}{4}\xi^2\right) d\xi. \tag{5.3}$$

If we set $K_\epsilon = \sqrt{\epsilon}M(\epsilon)$, then

$$\exp\left(-\frac{1}{4}K_\epsilon^2\right) = \frac{1}{2\epsilon}K_\epsilon\int_0^{K_\epsilon} \exp\left(-\frac{1}{4}\xi^2\right) d\xi. \tag{5.4}$$

It follows that

$$K_\epsilon \leq 2\sqrt{\log(2\epsilon)} \quad \text{for } \epsilon \geq \epsilon^*, \tag{5.5}$$

where $\epsilon^* \geq 1$ satisfies

$$\int_0^{2\sqrt{\log(2\epsilon^*)}} \exp\left(-\frac{1}{4}\xi^2\right) d\xi \geq 1. \tag{5.6}$$

We can easily check that $\epsilon^* = 1$ would satisfy (5.6). Therefore

$$M(\epsilon) \leq \frac{K_\epsilon}{\sqrt{\epsilon}} \leq \frac{2\sqrt{\log(2\epsilon)}}{\sqrt{\epsilon}} \quad \text{for } \epsilon \geq 1, \tag{5.7}$$

and (5.1) follows. □

6. The case $n \to \infty$. In the next two sections, we investigate the effects of the driving law exponent n. We shall fix ϵ and denote the solution of (1.1)–(1.5) by (u_n, s_n). The main result of this section is Theorem 6.1.

THEOREM 6.1. *For any $t^* > 0$,*

$$\lim_{n\to\infty} s_n(t) = 0 \tag{6.1}$$

uniformly for $t \in [0, t^*]$. □

We first list the available estimates for the solution (u_n, s_n) in the following lemma. Set

$$D_n = \{(x,t) : 0 < x < s_n(t),\ 0 < t < \infty\}.$$

LEMMA 6.2.

$$0 < u_n(x,t) < 1 \quad \textit{for } (x,t) \in D_n, \tag{6.2}$$

$$-1 - \epsilon < \frac{\partial u_n}{\partial x}(x,t) < 0 \quad \textit{for } (x,t) \in D_n, \tag{6.3}$$

$$0 < \frac{\partial u_n}{\partial t}(x,t) < 1 + \epsilon \quad \textit{for } (x,t) \in D_n, \tag{6.4}$$

$$0 < \frac{\partial^2 u_n}{\partial x^2}(x,t) < \frac{1}{\epsilon} + 1 \quad \textit{for } (x,t) \in D_n, \tag{6.5}$$

$$0 < s_n(t) < t \quad \textit{for } t > 0, \tag{6.6}$$

$$0 < s_n'(t) < 1 \quad \textit{for } t > 0, \tag{6.7}$$

$$s_n''(t) < 0 \quad \textit{for } t > 0. \tag{6.8}$$

Proof. (6.2), (6.3), and (6.6)–(6.8) are copied directly from Theorem 1.1. Formula (6.5) is equivalent to (6.4) by using the equation. We also know that $(u_n)_t > 0$. To prove (6.4), it suffices to note that $s_n''(t) < 0$ implies

$$\frac{\partial u_n}{\partial t}(s_n(t),t) + \frac{\partial u_n}{\partial x}(s_n(t),t)s_n'(t) < 0 \quad \text{for } t > 0, \tag{6.9}$$

and therefore by (6.3) and (6.7)

$$\frac{\partial u_n}{\partial t}(s_n(t),t) \le -\frac{\partial u_n}{\partial x}(s_n(t),t)s_n'(t) < 1 + \epsilon \quad \text{for } t > 0. \tag{6.10}$$

Now (6.4) follows by maximum principle. □

Next, we shall prove that $s_n(t)$ is monotone decreasing in n.

LEMMA 6.3. *If $n_1 > n_2$, then*

$$s_{n_1}(t) < s_{n_2}(t) \quad \textit{for } t > 0. \tag{6.11}$$

Proof. The transformation (see [4], for example)

$$v(x,t) = \int_x^{s(t)} (\epsilon u(\xi,t) + 1)d\xi \quad \text{for } 0 < x < s(t), \quad t > 0 \tag{6.12}$$

reduces (1.1)–(1.5) to the following problem:

$$\begin{aligned}
&\epsilon v_t = v_{xx} \quad \text{for } 0 < x < s(t), \quad t > 0,\\
&v_x(0,t) = -\epsilon - 1 \quad \text{for } t > 0,\\
&v(s(t),t) = 0 \quad \text{for } t > 0,\\
&v_x(s(t),t) = -\epsilon(s'(t))^{1/n} - 1 \quad \text{for } t > 0,\\
&s(0) = 0.
\end{aligned}$$

We shall denote by $v_{n_i}(x,t)$ $(i = 1,2)$ the functions obtained from $u_{n_i}(x,t)$ by formula (6.12).

It follows from (2.2) that (6.11) holds for small t. Therefore if (6.11) is not true, then there exists $T > 0$ such that

$$s_{n_1}(t) < s_{n_2}(t) \quad \text{for } 0 < t < T, \tag{6.13}$$

$$s_{n_1}(T) = s_{n_2}(T). \tag{6.14}$$

This implies that

$$s'_{n_1}(T) \geq s'_{n_2}(T), \tag{6.15}$$

and thus

$$\begin{aligned}
-\frac{\partial v_{n_1}}{\partial x}(s_{n_1}(T),T) &= \epsilon(s'_{n_1}(T))^{1/n_1} + 1\\
&\geq \epsilon(s'_{n_2}(T))^{1/n_1} + 1\\
&> \epsilon(s'_{n_2}(T))^{1/n_2} + 1 \quad (\text{since } 0 < s'_{n_2}(T) < 1,\ n_1 > n_2)\\
&= -\frac{\partial v_{n_2}}{\partial x}(s_{n_2}(T),T).
\end{aligned} \tag{6.16}$$

On the other hand, by the maximum principle, the maximum of $w \equiv v_{n_1} - v_{n_2}$ in the region $G \equiv \{(x,t) : 0 \leq x \leq s_{n_1}(t), 0 \leq t \leq T\}$ can only be obtained on $x = s_{n_1}(t), 0 \leq t \leq T$. Clearly,

$$w(0,0) = w(s_{n_1}(T),T) = 0, \tag{6.17}$$

and

$$w(s_{n_1}(t),t) = -v_{n_2}(s_{n_1}(t),t) < 0 \quad \text{for } 0 < t < T. \tag{6.18}$$

Hence w takes its maximum in G at $(x,t) = (s_{n_1}(T),T)$, which contradicts (6.16). Therefore (6.11) holds. □

Proof of Theorem 6.1. By Lemma 6.3 and (6.6)–(6.8), we get that, for any $t^* > 0$,

$$\lim_{n\to\infty} s_n(t) = s_\infty(t) \quad \text{uniformly for } t \in [0,t^*], \tag{6.19}$$

where $s_\infty(t)$ is some Lipschitz continuous function with the following properties:

$$0 \leq s'_\infty(t) \leq 1 \quad \text{a.e.,} \tag{6.20}$$

$$s''_\infty(t) \leq 0 \quad \text{in the distribution sense,} \tag{6.21}$$

$$s_\infty(t) < s_n(t) \quad \text{for } t > 0, \quad n > 0. \tag{6.22}$$

We claim that

(6.23) There exists $\delta > 0$ such that $s_\infty(t) \equiv 0$ for $0 \le t \le \delta$.

Note that (6.20), (6.21), and (6.23) imply that $s_\infty(t) \equiv 0$ for $t > 0$, and then the theorem follows.

If the claim (6.23) is not true, then in view of (6.20)

$$s_\infty(t) > 0 \quad \text{for } t > 0. \tag{6.24}$$

Set

$$D_\infty = \{(x,t) : 0 < x < s_\infty(t),\ 0 < t < t^*\}.$$

By Lemma 6.2 and the embedding theorem (see [10]), there exists a subsequence n_k's of n's ($n_k \to \infty$) such that, for any $0 < \alpha < \frac{1}{2}$ and any $\delta > 0$,

$$u_{n_k},\ \frac{\partial u_{n_k}}{\partial x} \to u_\infty,\ \frac{\partial u_\infty}{\partial x} \quad \text{in } C^\alpha(D_\infty \cap \{t > \delta\}) \text{ norm}, \tag{6.25}$$

$$\frac{\partial u_{n_k}}{\partial t},\ \frac{\partial^2 u_{n_k}}{\partial x^2} \rightharpoonup \frac{\partial u_\infty}{\partial t},\ \frac{\partial^2 u_\infty}{\partial x^2} \quad \text{weakly in } L^p(D_\infty \cap \{t > \delta\}) \text{ for any } p > 1, \tag{6.26}$$

where u_∞ is some function in $W^{2,1}_\infty(D_\infty)$.

Passing the limit in the equations in (1.1)–(1.5) as $n = n_k \to \infty$, we obtain

$$\epsilon(u_\infty)_t = (u_\infty)_{xx} \quad \text{in } D_\infty, \tag{6.27}$$

$$u_\infty(0,t) = 1 \quad \text{for } 0 < t < t^*. \tag{6.28}$$

Next, we compute

$$s_n'(t) = -\frac{(u_n)_x(s_\infty(t),t)}{1+\epsilon u_n(s_\infty(t),t)} + \left(\frac{(u_n)_x(s_\infty(t),t)}{1+\epsilon u_n(s_\infty(t),t)} - \frac{(u_n)_x(s_n(t),t)}{1+\epsilon u_n(s_n(t),t)}\right). \tag{6.29}$$

The first term of the above equality converges to $-(u_\infty)_x(s_\infty(t),t)/(1+\epsilon u_\infty(s_\infty(t),t))$ uniformly in any compact set of $(0,t^*]$, while the second term converges to zero uniformly for $t \in (0,t^*]$ since $(u_n)_x$, $(u_n)_{xx}$ are uniformly bounded. This shows that $s_{n_k}'(t)$ converges uniformly in any compact set of $(0,t^*]$ and the limit is continous. Hence we must have

$$s_{n_k}'(t) \to s_\infty'(t) \quad \text{uniformly in any compact set of } (0,t^*] \text{ as } n_k \to \infty; \tag{6.30}$$

$s_\infty'(t)$ is continuous and furthermore,

$$(1+\epsilon u_\infty(s_\infty(t),t))s_\infty'(t) = -(u_\infty)_x(s_\infty(t),t) \quad \text{for } t \in (0,t^*). \tag{6.31}$$

By (6.24) and (6.20),

$$s_\infty'(t_0) > 0 \quad \text{for some } t_0 \in (0,t^*). \tag{6.32}$$

It follows from (6.21) that

$$s_\infty'(t) \ge s_\infty'(t_0) > 0 \quad \text{for } t \in (0,t_0). \tag{6.33}$$

By virtue of (6.30) and concavity of s_{n_k}, there exists $N > 0$ such that

$$s'_{n_k}(t) \geq \tfrac{1}{2} s'_\infty(t_0) \quad \text{for } 0 < t \leq t_0, \quad n_k > N. \tag{6.34}$$

Since $s_\infty(t) < s_{n_k}(t)$,

$$1 \geq u_{n_k}(s_\infty(t), t) \geq u_{n_k}(s_{n_k}(t), t) = (s_{n_k}(t))^{1/n_k} \geq \left(\tfrac{1}{2} s'_\infty(t_0)\right)^{1/n_k} \tag{6.35}$$

for $0 < t \leq t_0$, $n_k > N$. Letting $n_k \to \infty$ in the above inequality, we obtain

$$u_\infty(s_\infty(t), t) \equiv 1 \quad \text{for } 0 < t \leq t_0.$$

Recalling (6.27) and (6.28), we get

$$u_\infty(x, t) \equiv 1 \quad \text{for } 0 \leq x \leq s_\infty(t), \quad 0 < t \leq t_0,$$

which contradicts (6.31) since $s'_\infty(t) > 0$ for $t \in (0, t_0)$. The proof of the theorem is now complete. □

7. The case $n \to 0$. We may guess that $\lim_{n \to 0} s_n(t) = t$ at the first glance of the driving law (1.4). However, this is not true when t is large enough by Lemma 3.2 (note that the constant M there is independent of n). The interesting result is that there exists a critical value $T^* > 0$ such that

$$\lim_{n \to 0} s_n(t) = t \quad \text{uniformly for } t \in [0, T^*] \tag{7.1}$$

and

$$\lim_{n \to 0} s_n(t) = s_0(t) \quad \text{uniformly for } t \in [T^*, t^*] \tag{7.2}$$

for any $t^* > T^*$, where $s_0(t)$ is the free boundary of an appropriate Stefan problem. In order to make the above statement more explicit, we consider the following problem:

$$\epsilon z_t = z_{xx} \quad \text{for } 0 < x < t, \tag{7.3}$$

$$z(0, t) = 1 \quad \text{for } t > 0, \tag{7.4}$$

$$1 + \epsilon z(t, t) = -z_x(t, t) \quad \text{for } t > 0. \tag{7.5}$$

It is clear that this problem has a unique smooth classical solution $z \in C^{2,1}(\overline{\Delta}) \cap C^\infty(\overline{\Delta} \setminus \{0\})$, where $\Delta = \{(x, t) : 0 < x < t < \infty\}$. The next lemma determines the critical value T^*.

LEMMA 7.1. *There exists a unique $T^* \in (0, 1]$ such that*

$$z(t, t) > 0 \quad \textit{for } 0 \leq t < T^*, \tag{7.6}$$

$$z(t, t) < 0 \quad \textit{for } t > T^*. \tag{7.7}$$

Note that by Lemma 7.1 and maximum principle, $z(x, T^*)$ is smooth and $z(x, T^*) > 0$ for $0 < x < T^*$ and $z(T^*, T^*) = 0$. □

We shall postpone the proof of Lemma 7.1 and first consider the solution (u_0, s_0) of the following Stefan problem:

$$\epsilon (u_0)_t = (u_0)_{xx} \quad \text{for } 0 < x < s_0(t), \quad t > T^*, \tag{7.8}$$

$$u_0(0, t) = 1 \quad \text{for } t > T^*, \tag{7.9}$$

$$s'_0(t) = -(u_0)_x(s_0(t), t) \quad \text{for } t > T^*, \tag{7.10}$$

$$u_0(s_0(t), t) = 0 \quad \text{for } t > T^*, \tag{7.11}$$

$$u_0(x, T^*) = z(x, T^*) \quad \text{for } 0 < x < T^*. \tag{7.12}$$

The main result of this section is Theorem 7.2.

THEOREM 7.2. *Denote by* (u_n, s_n) *the solution of* (1.1)–(1.5). *Suppose that* T^* *and* $z(x,t)$, (u_0, s_0) *are determined as above. Then* $T^* > 1/2(1+\epsilon)$ *and for any* $t^* > T^*$, *we have*

$$\lim_{n\to 0} s_n(t) = t \quad \textit{uniformly for } t \in [0, T^*], \tag{7.13}$$

$$\begin{aligned} &\lim_{n\to 0} u_n(x,t) = z(x,t) \\ &\textit{uniformly in any compact set of } \{(x,t):\ 0 \le x < t, 0 < t \le T^*\}; \end{aligned} \tag{7.14}$$

$$\lim_{n\to 0} s_n(t) = s_0(t) \quad \textit{uniformly for } t \in [T^*, t^*], \tag{7.15}$$

$$\begin{aligned} &\lim_{n\to 0} u_n(x,t) = u_0(x,t) \\ &\textit{uniformly in any compact set of } \{(x,t) : 0 \le x < s_0(t), T^* \le t \le t^*\}. \end{aligned} \tag{7.16}$$

□

Proof of Lemma 7.1. The linear function $w(x,t) \equiv 1 - x$ satisfies

$$w_x + \epsilon w \ge -1 \quad \text{for } 0 < x = t < 1. \tag{7.17}$$

Thus by maximum principle

$$z(x,t) \le w \equiv 1 - x \quad \text{for } 0 \le x \le t \le 1, \tag{7.18}$$

which implies that

$$z(1,1) \le 0. \tag{7.19}$$

It is clear that $z(t,t) > 0$ for t small. Hence in view of (7.19), there exists T^* $(0 < T^* \le 1)$ such that (7.6) holds and

$$z(T^*, T^*) = 0. \tag{7.20}$$

By maximum principle $z(x,t) \le 1$ for $0 \le x \le t < \infty$. Therefore

$$z_x(0,t) \le 0. \tag{7.21}$$

Let $\widetilde{w} = z_t + z_x$. Then

$$\epsilon \widetilde{w}_t = \widetilde{w}_{xx} \quad \text{for } 0 < x < t < \infty, \tag{7.22}$$

and, by (7.4) and (7.21),

$$\widetilde{w}(0,t) \le 0. \tag{7.23}$$

Differentiating (7.5) in t, we obtain

$$\widetilde{w}_x + \epsilon \widetilde{w} = 0 \quad \text{for } 0 < x = t < \infty. \tag{7.24}$$

Therefore by maximum principle

$$\widetilde{w} \le 0 \quad \text{for } 0 \le x \le t < \infty, \tag{7.25}$$

which implies that

$$\frac{d}{dt}z(t,t) \le 0 \quad \text{for } 0 < t < \infty. \tag{7.26}$$

Thus if (7.7) is not true, then

$$z(t,t) \equiv 0 \quad \text{for } T^* \le t \le T^* + \mu \tag{7.27}$$

for some $\mu > 0$. Hence

$$\widetilde{w}(t,t) \equiv 0 \quad \text{for } T^* < t < T^* + \mu. \tag{7.28}$$

Using (7.24), we get

$$\widetilde{w}_x(t,t) \equiv 0 \quad \text{for } T^* < t < T^* + \mu. \tag{7.29}$$

By (7.25) and (7.28), $\widetilde{w}$ takes its maximum on $(x,t) = (t,t)$, $T^* < t < T^* + \mu$. Now it follows from strong maximum principle that

$$\widetilde{w}(x,t) \equiv z_x(x,t) + z_t(x,t) \equiv 0 \quad \text{in } \Delta \cap \{t < T^* + \mu\}, \tag{7.30}$$

which implies that $z(t,t) \equiv z(0,0) = 1$ for $0 < t < T^* + \mu$. This is a contradiction to (7.20). □

Proof of Theorem 7.2. By Lemma 6.3 and (6.6)–(6.8),

$$\lim_{n \to 0} s_n(t) = s_\#(t) \quad \text{uniformly for } t \in [0, t^*], \tag{7.31}$$

where $s_\#(t)$ is some Lipschitz continuous function such that

$$0 \le s'_\#(t) \le 1 \quad \text{a.e.}, \tag{7.32}$$

$$s''_\#(t) \le 0 \quad \text{in the distribution sense}, \tag{7.33}$$

$$s_\#(t) > s_n(t) \quad \text{for } t > 0, n > 0. \tag{7.34}$$

By (2.1),

$$\begin{aligned} u_n(s_n(t),t) &\ge 1 - (1+\epsilon)s_n(t) \\ &\ge 1 - (1+\epsilon)t \\ &\ge \frac{1}{2} \quad \text{for } 0 < t < \frac{1}{2(1+\epsilon)}. \end{aligned} \tag{7.35}$$

Therefore

$$\begin{aligned} t \ge s_n(t) &= \int_0^t s'_n(\xi)\,d\xi \\ &= \int_0^t u_n^n(s_n(\xi),\xi)\,d\xi \\ &\ge \left(\frac{1}{2}\right)^n t \quad \text{for } 0 < t < \frac{1}{2(1+\epsilon)}. \end{aligned} \tag{7.36}$$

It follows by letting $n \to 0$ that

$$s_\#(t) \equiv t \quad \text{for } 0 < t < \frac{1}{2(1+\epsilon)}. \tag{7.37}$$

Define

$$T_{\#} = \sup\{b:\ s_{\#}(t) \equiv t \text{ for } 0 < t < b\} \tag{7.38}$$

and

$$D_{\#} = \{(x,t):\ 0 < x < s_{\#}(t), 0 < t < t^*\}. \tag{7.39}$$

Then $T_{\#} \geq 1/2(1+\epsilon)$. By (7.32) (7.33) and the definition for $T_{\#}$,

$$s_{\#}(t) \equiv t \quad \text{for } 0 \leq t < T_{\#}, \tag{7.40}$$

$$s_{\#}(t) < t \quad \text{for } T_{\#} < t < t^*. \tag{7.41}$$

Next, we introduce a scaling in the x direction so that the solutions u_n are defined in the same domain in the new variables. Let

$$v_n(x,t) = u_n\left(\frac{s_n(t)}{s_{\#}(t)}x, t\right) \qquad \text{for } (x,t) \in D_{\#}. \tag{7.42}$$

Then by Lemma 6.2 and (7.32),

$$\|v_n\|_{W^{2,1}_{\infty}(D_{\#}\cap\{t>\delta\})} \leq C_{\delta} \tag{7.43}$$

for any $\delta > 0$. Thus, for an appropriate subsequence n_k's of n's, we have, as $n_k \to 0$,

$$v_{n_k},\ \frac{\partial v_{n_k}}{\partial x} \to v_{\#},\ \frac{\partial v_{\#}}{\partial x} \quad \text{uniformly in } D_{\#} \cap \{t > \delta\} \text{ for any } \delta > 0, \tag{7.44}$$

$$\frac{\partial v_{n_k}}{\partial t},\ \frac{\partial^2 v_{n_k}}{\partial x^2} \rightharpoonup \frac{\partial v_{\#}}{\partial t},\ \frac{\partial^2 v_{\#}}{\partial x^2} \quad \text{weakly in } L^p(D_{\#} \cap \{t > \delta\}) \text{ for any } p > 1,\ \delta > 0, \tag{7.45}$$

for some function $v_{\#} \in W^{2,1}_{\infty}(D_{\#})$.

Note that

$$s_n'(t) = -\frac{(v_n)_x(s_{\#}(t),t)}{1+\epsilon v_n(s_{\#}(t),t)} \cdot \frac{s_{\#}(t)}{s_n(t)}. \tag{7.46}$$

Therefore, by using the same argument used in the proof of Theorem 6.1 we obtain

$$s_{n_k}'(t) \to s_{\#}'(t) \quad \text{uniformly in } t \in [\delta, t^*] \text{ as } n_k \to 0, \text{ for any } \delta > 0; \tag{7.47}$$

furthermore, $s_{\#}'(t)$ is continuous and

$$(1+\epsilon v_{\#}(s_{\#}(t),t))s_{\#}'(t) = -(v_{\#})_x(s_{\#}(t),t)). \tag{7.48}$$

A direct computation shows

$$\epsilon(v_n)_t = \left(\frac{s_{\#}}{s_n}\right)^2 (v_n)_{xx} + \epsilon x\left(\frac{s_n'}{s_n} - \frac{s_{\#}'}{s_{\#}}\right)(v_n)_x \quad \text{for } (x,t) \in D_{\#}. \tag{7.49}$$

By (7.31) and (7.47),

$$\lim_{n=n_k\to 0}\left|\frac{s_n'}{s_n} - \frac{s_{\#}'}{s_{\#}}\right| = 0 \quad \text{uniformly in } t \in [\delta, t^*] \text{ for any } \delta > 0. \tag{7.50}$$

Therefore, by letting $n = n_k \to 0$ in (7.49), using (7.31) and (7.50), we obtain

$$\epsilon(v_\#)_t = (v_\#)_{xx} \quad \text{for } (x,t) \in D_\#, \tag{7.51}$$

$$v_\#(0,t) = 1 \quad \text{for } 0 < t < t^*. \tag{7.52}$$

Since $s'_\#(t) \equiv 1$ for $0 < t < T_\#$, (7.48) implies that

$$1 + \epsilon v_\#(s_\#(t),t) = -(v_\#)_x(s_\#(t),t) \quad \text{for } 0 < t < T_\#. \tag{7.53}$$

Clearly also,

$$v_\#(s_\#(t),t) = \lim_{n\to 0} v_n(s_\#(t),t) \geq 0 \quad \text{for } 0 < t < t^*. \tag{7.54}$$

Recall that z is a solution of (7.3)–(7.5), and by uniqueness we have

$$v_\# = z \quad \text{for } 0 < t < x \leq T_\#. \tag{7.55}$$

Therefore (7.54) and (7.7) imply

$$T_\# \leq T^*. \tag{7.56}$$

Next, we shall prove that

$$v_\#(s_\#(t),t) = 0 \quad \text{for } T_\# < t < t^*. \tag{7.57}$$

Note that (7.57) implies $v_\#(s_\#(T_\#),T_\#) = 0$, by continuity; and hence $T_\# = T^*$ by Lemma 7.1.

Since $s'_\#(t)$ is continuous, (7.33), (7.40), and (7.41) imply that

$$s'_\#(t) < 1 \quad \text{for } T_\# < t < t^*. \tag{7.58}$$

Now we fix $t \in (T_\#, t^*)$ and take β with $s'_\#(t) < \beta < 1$. By virtue of (7.47),

$$s'_{n_k}(t) \leq \beta \quad \text{if } 0 < n_k < n_0 \tag{7.59}$$

for some sufficiently small n_0. It follows that

$$0 \leq v_n(s_\#(t),t) = (s'_n(t))^{1/n} \leq \beta^{1/n}. \tag{7.60}$$

Letting $n_k \to 0+$ in the above inequality, we obtain (7.57). Thus we have proved

$$T_\# = T^*. \tag{7.61}$$

Equations (7.51), (7.52), (7.57), (7.48), and (7.55), (7.61) imply that $(v_\#, s_\#)$ for $T_\# < t < t^*$ is a solution of the Stefan problem (7.8)–(7.12). Therefore by uniqueness

$$v_\# = u_0 \quad \text{for } T_\# < x < s_\#(t),\ T_\# < t < t^*, \tag{7.62}$$

$$s_\# = s_0 \quad \text{for } T_\# < t < t^*. \tag{7.63}$$

Now the limits in (7.44) and (7.45) are unique and independent of the choice of subsequences. And hence the whole sequences themselves converge. This implies that (7.13)–(7.16) hold. The theorem is proved. □

8. Neumann boundary condition. In this section we consider the Neumann boundary condition at $x = 0$. Taking for simplicity $\epsilon = 1$, the system becomes

$$u_t = u_{xx} \quad \text{for } 0 < x < s(t), \quad t > 0, \tag{8.1}$$

$$u_x(0,t) = g(t), \tag{8.2}$$

$$[1 + u(s(t),t)] \cdot s'(t) = -u_x(s(t),t), \tag{8.3}$$

$$s'(t) = u^n(s(t),t), \tag{8.4}$$

$$s(0) = 0. \tag{8.5}$$

It is shown in [2] that if

$$g \in C^2[0,\infty), \quad g(t) \le 0 \quad \text{for } t > 0, \quad g(0) < 0; \quad g'(t) \ge 0 \quad \text{for } t \ge 0, \tag{8.6}$$

then (1.18)–(1.20) hold. The asymptotic behavior of $s(t)$ as $t \to \infty$ is also studied in the case $\int_0^\infty g(t)dt > -\infty$.

We shall study the asymptotic behavior in the general case. Define $\psi_\eta = \psi_\eta(x,t)$, $h_\eta = h_\eta(t)$ to be the solution for the Stefan problem:

$$\psi_t = \psi_{xx} \quad \text{for } 0 < x < h(t), \quad t > 0, \tag{8.7}$$

$$\psi_x(0,t) = g(t), \tag{8.8}$$

$$\psi = 0, \ -\psi_x = (1+\eta)h'(t) \quad \text{on } x = h(t), \tag{8.9}$$

$$h(0) = 0, \tag{8.10}$$

where g satisfies (8.6) and $\eta \ge 0$.

THEOREM 8.1. *Assume in addition to assumption* (8.6) *that*

$$\lim_{t\to\infty} g(t) = 0. \tag{8.11}$$

Then

$$\lim_{t\to\infty} \frac{s(t)}{h_0(t)} = 1, \tag{8.12}$$

where $x = h_0(t)$ corresponds to the free boundary of the Stefan problem (8.7)–(8.10) *with $\eta = 0$.* □

We first establish several lemmas.

LEMMA 8.2. *For ψ_η and h_η defined in* (8.7)–(8.10), *we have*

$$(1+\eta)h_\eta(t) \ge h_0(t) \quad \textit{for } t > 0. \tag{8.13}$$

Proof. Let $w_\eta(x,t)$ be defined as

$$\begin{aligned} w_\eta(x,t) &= \int_{h_\eta^{-1}(x)}^{t} \psi_\eta(x,\tau)d\tau \quad \text{for } x < h_\eta(t) \\ &\equiv 0 \qquad\qquad\qquad\quad \text{for } x \ge h_\eta(t) \end{aligned} \tag{8.14}$$

($h_\eta^{-1}(x)$ exists since $h'(t) > 0$). Then

$$(w_\eta)_t - (w_\eta)_{xx} = -1 - \eta \quad \text{for } 0 < x < h_\eta(t), \quad t > 0, \tag{8.15}$$

$$w_\eta = (w_\eta)_x = 0 \quad \text{on } x = h_\eta(t), \quad t > 0, \tag{8.16}$$

$$(w_\eta)_x = \int_0^t g(\tau)d\tau \quad \text{on } x = 0, \quad t > 0, \tag{8.17}$$

$$w_\eta = 0 > -1 - \eta \quad \text{for } x > h_\eta(t), \quad t > 0. \tag{8.18}$$

Introducing the change of variables $y = (1+\eta)x$ and $\widetilde{w}_\eta(y,t) = (1+\eta)w_\eta(x,t)$, we find that (8.15) is equivalent to

$$(\widetilde{w}_\eta)_t - (\widetilde{w}_\eta)_{yy} = -1 + \left((1+\eta) - \frac{1}{1+\eta}\right)(w_\eta)_t(x,t) \quad \text{for } 0 < y < (1+\eta)h_\eta(t), \quad t > 0. \tag{8.19}$$

Since $(w_\eta)_t(x,t) = \psi_\eta(x,t) \geq 0$, it follows from the comparison principle for variational inequalities that

$$\widetilde{w}_\eta(y,t) \geq w_0(y,t) \quad \text{for all } y > 0, \quad t > 0. \tag{8.20}$$

Hence (8.13) holds. □

LEMMA 8.3. *For any* $T > 0$

$$\lim_{t\to\infty} \frac{h_\eta(t-T)}{h_\eta(t)} = 1. \tag{8.21}$$

Proof. We have $h_\eta(t) \geq h_\eta(t-T)$ since $h'_\eta \geq 0$. Equation (8.21) is obviously valid if $h_\eta(t)$ has a finite limit as $t \to \infty$. If, however, $\lim_{t\to\infty} h_\eta(t) = \infty$, then

$$\left|\frac{h_\eta(t-T)}{h_\eta(t)} - 1\right| \leq \frac{T|h'_\eta(\xi)|}{h_\eta(t)}, \tag{8.22}$$

where $\xi = \xi(t) \in (t-T, t)$. It is clear that $h'_\eta(t)$ is bounded as $t \to \infty$ since $g(t)$ is bounded. Therefore, the right-hand side of inequality (8.22) converges to zero as $t \to \infty$. □

Proof of Theorem 8.1. First we prove that

$$\lim_{t\to\infty} s'(t) = 0. \tag{8.23}$$

Since $s''(t) \leq 0$, $\lim_{t\to\infty} s'(t) = c_0$ exists. If $c_0 > 0$, then

$$\lim_{t\to\infty} \frac{s(t)}{t} = c_0 > 0.$$

However, by the mass balance

$$s(t) = -\int_0^{s(t)} u(x,t)dx - \int_0^t g(\tau)d\tau, \tag{8.24}$$

and hence

$$\frac{s(t)}{t} \leq -\frac{1}{t}\int_0^t g(\tau)d\tau \longrightarrow 0 \quad \text{as } t \to \infty$$

by (8.11), which is a contradiction.

Now, by using (8.23), Lemmas 8.2 and 8.3, we can proceed in the same proof as in Lemma 3.3 to finish the proof of this theorem. □

Having proved Theorem 8.1, we now reduce the asymptotic behavior of $s(t)$ near ∞ to that of the corresponding Stefan problem, which is well known. Therefore, using, for example, [6, Thm. 3, Chap. 8], we immediately get the behavior of $s(t)$ near ∞ if $g(t)$ is like $t^{-\delta}$ $(1/2 \leq \delta < 1)$ near ∞.

Acknowledgments. The author thanks Professor Avner Friedman for his directions and help. The author also thanks the referees for several helpful suggestions and comments.

REFERENCES

[1] G. ASTARITA AND G. C. SARTI, *A class of mathematical models for sorption of swelling solvents in glassy polymers,* Polymer Engrg. Sci., 18 (1978), pp. 388–395.

[2] D. ANDREUCCI AND R. RICCI, *A free boundary problem arising from sorption of solvents in glassy polymers,* Quart. Appl. Math., 44 (1987), pp. 649–657.

[3] D. S. COHEN AND T. ERNEUX, *Free boundary problem in controlled release pharmaceuticals.* I: *Diffusion in glassy polymers,* SIAM. J. Appl. Math., 48 (1988), pp. 1451–1465.

[4] A. FASANO, F. H. MEYER, AND M. PRIMICERIO, *On a problem in the polymer industry: Theoretical and numerical investigation of swelling,* SIAM. J. Math. Anal., 17 (1986), pp. 945–960.

[5] A. FASANO AND M. PRIMICERIO, *Free boundary problems for nonlinear parabolic equations with nonlinear free boundary conditions,* J. Math. Anal. Appl., 72 (1979), pp. 247–273.

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

[7] ———, *Variational Principles and Free Boundary Problems,* John Wiley, New York, 1982.

[8] ———, *Asymptotic behavior for the free boundary of parabolic variational inequalities and applications to sequential analysis,* Illinois J. Math., 26 (1982), pp. 653–697.

[9] ———, *Free boundary problems for parabolic equations* II: *Evaporation or condensation of a liquid drop,* J. Math. Mech., 9 (1960), pp. 19–66.

[10] O. A. LADYŽENSKAJA, V. A. SOLONNIKOV, AND N. N. URALĆEVA, *Linear and Quasi-Linear Equations of Parabolic Type,* Translations of Mathematical Monographs, Vol. 23, 1968.

[11] E. COMPARINI AND R. RICCI, *Convergence to pseudo-steady-state approximation for the unreacted core model,* Appl. Anal., 26 (1987), pp. 305–325.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 957-972, July 1991

005

THE ONE-DIMENSIONAL WIGNER-POISSON PROBLEM AND ITS RELATION TO THE SCHRÖDINGER-POISSON PROBLEM*

H. STEINRÜCK†

Abstract. This paper shows the existence of a solution of the Wigner-Poisson problem by expanding the solution into a series of solutions of the Schrödinger equation, and proves the convergence of the solutions of the Wigner-Poisson problem to a generalized solution of the Vlasov-Poisson problem in the classical limit.

Key words. quantum transport, nonlinear evolution equations

AMS(MOS) subject classifications. 35K22, 81C10, 82A70

1. Introduction. In this paper we consider the one-dimensional Wigner-Poisson problem

$$V(x) = \int_R |x-y| n(y)\, dy, \qquad n(y) = \int_R w(y, v)\, dv, \tag{1.1}$$

$$w_t + v.w_x - \Theta_\varepsilon[V]w = 0, \qquad x \in R_x, \qquad v \in R_v, \tag{1.2}$$

$$w(x, v, t=0) = w_I(x, v). \tag{1.3}$$

Equations (1.1)-(1.3) govern the temporal evolution of the Wigner function w, defined on the (x, v) phase space under the action of a self-consistent Coulomb potential $V(x)$. The particle density $n(x)$ is the mean value of w with respect to the velocity variable v. The pseudodifferential operator $\Theta_\varepsilon[V]$ is defined by

$$\Theta_\varepsilon[V]w = \frac{i}{2\pi\varepsilon} \int_R \int_R \left(V\left(x+\frac{\varepsilon\eta}{2}\right) - V\left(x-\frac{\varepsilon\eta}{2}\right)\right) w(x, v') \cdot e^{i\eta(v-v')}\, dv'\, d\eta. \tag{1.4}$$

Note that (1.1)-(1.3) are already in dimensionless form and that the parameter ε is the scaled (dimensionless) Planck constant. The Wigner function for a pure quantum mechanical state ϕ (a solution of the Schrödinger equation) was first introduced by Wigner in 1932 [14] and is defined by

$$w_\phi(x, v, t) = \frac{1}{\sqrt{2\pi\varepsilon}} \int_R \overline{\phi(x+\varepsilon\eta/2, t)}\phi(x-\varepsilon\eta/2, t)\, e^{i\eta v}\, d\eta.$$

An easy calculation shows that w_ϕ solves the Wigner equation (1.2). Therefore the Schrödinger and the Wigner equation are equivalent with respect to the temporal evolution of a pure state.

* Received by the editors June 5, 1989; accepted for publication (in revised form) July 9, 1990. This work was supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung under grant P6771.

† Institut für Angewandte und Numerische Mathematik, Technische Universität Wien, Wiedner Hauptstrasse 8-10, 1140 Vienna, Austria. Present address, IBM, T. J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.

The advantage of the Wigner function is that it is defined on the (x, v) phase space and is therefore amenable to a comparison with the phase space formulation of classical mechanics. On the other hand, there are qualitative differences between the quantum and the classical case. While in classical mechanics the distribution function of particles in the phase space evolves according to the Liouville equation

$$w_t + v \cdot w_x - V_x \cdot w_v = 0, \tag{1.5}$$

which preserves the positivity of the initial function, the Wigner equation is in general not positivity preserving. That means that the solution w of (1.2) might assume negative values, though the initial function is nonnegative. Therefore the Wigner function cannot be interpreted as a distribution function of particles in the phase space. For a pure quantum state the mean values with respect to v (respectively, x) of w_ϕ are the probability densities to find a particle, which is described by its wave function ϕ at x (respectively, to have the velocity v). Therefore the Wigner function is referred as a "quasi distribution."

Note that for a harmonic oscillator (quadratic potential) the Wigner equation is identical to the classical Liouville equation. Considering the classical limit ($\varepsilon \to 0$) the Wigner equation tends at least formally to the Liouville equation. The pseudodifferential operator $\Theta_\varepsilon[V]$ then becomes the differential operator $V_x \, \partial/\partial v$.

Up to now, we have discussed the Wigner equation only as an equivalent alternative to the one particle Schrödinger equation. But starting with the many-body Schrödinger equation, using the density matrix formulation and an appropriate ansatz to get self-consistent equations we derive the Wigner–Poisson problem (1.1)–(1.3) (see [3]). Taking again the classical limit ($\varepsilon \to 0$), the Wigner–Poisson problem becomes at least formally the Vlasov–Poisson problem, which is a coupled system consisting of the Liouville and the Poisson equation.

The recent interest in the Wigner–Poisson problem is motivated by the miniaturization of semiconductor devices, where quantum effects in potential wells near heterojunctions [8] or tunneling effects [7] must be taken into account.

Mathematically the Wigner equation for a prescribed potential was analyzed in [9] and [10], using semigroup theory. In [2] these results are extended to the case of a particle with spin in a prescribed electromagnetic field. The existence of a solution of the coupled Wigner–Poisson problem has been shown for some special cases: periodic boundary conditions [1], and a bounded Brillouin zone [6], [13].

In this paper we will prove the existence of a unique globally defined solution of the Wigner–Poisson problem in the case of one space dimension. Due to the dimension dependence of the Green function of the Laplace operator the higher-dimensional cases have to be treated separately. An existence proof for the two-dimensional problem in the charge neutral case will be given by Arnold and Nier [4] and for the three-dimensional case by Brezzi and Markowich [5] in subsequent papers.

A second goal of this paper is the classical limit $\varepsilon \to 0$: A sequence of solutions of (1.1)–(1.3) has a subsequence, which converges weakly to a generalized solution of the Vlasov–Poisson problem as $\varepsilon \to 0$.

The paper is organized as follows. In § 2 we prove the existence of a solution of the one-dimensional Schrödinger–Poisson problem. In § 3 we use the relation between the Schrödinger and the Wigner equation to derive a solution representation of the Wigner equation in terms of the solution operator of the Schrödinger equation. The tools developed in §§ 2 and 3 will be used in § 4 to show the existence and uniqueness of a globally defined solution of the Wigner–Poisson problem. Section 5 is devoted to the classical limit.

2. The Schrödinger–Poisson problem. In this section we will prove the global existence of a solution of the Schrödinger-Poisson problem:

$$V(x) = \int_R |x-y|\,\overline{\psi(y)}\psi(y)\,dy, \tag{2.1}$$

$$i\varepsilon\psi_t = -\frac{\varepsilon^2}{2}\psi_{xx} + V\psi, \qquad x \in R, \quad t > 0, \tag{2.2}$$

$$\psi(x, t=0) = \psi_I(x). \tag{2.3}$$

THEOREM 2.1. *Let* $W := \{\psi \in L^2(R)\colon \|\psi\|_W := \|\psi\|_2 + \|\psi_x\|_2 + \|x\psi\|_2 < \infty\}$. *If* $\psi_I \in W$, *then the Schrödinger-Poisson problem has a unique mild solution* $\psi \in C([0,\infty) \to L^2(R))$. *If additionally* $\psi_I \in H_2$, *then the Schrödinger-Poisson problem has a unique classical solution.*

Remark. The term "mild" solution refers to the fact that the Schrödinger equation (2.2) is satisfied in the mild sense (see the definition on [11, p. 106]).

For the proof we first show that the Schrödinger equation (2.2)-(2.3) for a given sufficiently smooth potential, which is bounded by a polynomial of degree one, has a mild solution. Then inserting this solution into (2.1) we obtain a new potential, for which the Schrödinger equation can be solved and we will show that an iteration can be defined that way. Considering a sufficient small time interval this iteration contracts to a unique fixed point, which is a solution of the Schrödinger-Poisson problem. Using a priori estimates this solution can be extended for all positive times $t > 0$.

LEMMA 2.2. *Let* $V(x, t)$ *satisfy*:

$$|V(x, t)| \leqq \alpha(|x| + 1), \tag{2.4}$$

$$|V_x(x, t)| \leqq \beta, \tag{2.5}$$

$$|V_{xx}(x, t)| \leqq \gamma, \tag{2.6}$$

$$\limsup_{s \to t} \sup_{x \in R} \frac{|V(x, t) - V(x, s)|}{|x| + 1} = 0 \tag{2.7}$$

for some constants α, β, γ; *then there exists a strongly continuous family of unitary operators* $U(t, s)\colon L^2(R) \to L^2(R)$ *with*

$$U(t, r)U(r, s) = U(t, s) \quad \text{for } 0 \leqq s \leqq r \leqq t, \tag{2.8}$$

$$\frac{\partial^+}{\partial t} U(t, s)\psi|_{t=s} = i\frac{\varepsilon}{2}\psi_{xx} - \frac{i}{\varepsilon} V(\cdot, s)\psi \quad \text{for } \psi \in Y, \tag{2.9}$$

$$\frac{\partial}{\partial s} U(t, s)\psi = -U(t, s)\left(i\frac{\varepsilon}{2}\psi_{xx} - \frac{i}{\varepsilon} V(\cdot, s)\psi\right) \quad \text{for } \psi \in Y, \tag{2.10}$$

where $Y := \{\psi \in L^2$ *with* $x\psi, \psi_{xx} \in L^2\}$.

Proof. The conditions for the existence of an evolution system $U(t, s)$ satisfying (2.8)-(2.10) are the following [11]:

- (E1) The operators $H_s := (i\varepsilon/2)(\partial^2/\partial x^2) - (i/\varepsilon)V(\cdot, s)$ form a stable family of generators of C_0-semigroups. That is, let $0 \leqq s_1 \leqq s_2 \leqq \cdots \leqq s_n$ be a sequence of real numbers; then the estimate $\|\prod_{j=1}^n (H_{s_n} - \lambda)^{-1}\|_2 \leqq \lambda^{-n}$ for $\lambda > 0$ holds.
- (E2) The subspace $Y \subseteq D(H_s)$ is admissible. In other words, the restriction of H_s to Y generates a C_0-semigroup on Y for every fixed s.
- (E3) The mapping $t \to H_t$ is continuous in the norm of bounded operators from Y into L^2.

The operator iH_s is essentially self-adjoint (see the corollary to Theorem X.38 in [12]) and therefore H_s is the generator of a unitary C_0-group of operators on L^2.

Let $\lambda > 0$ and define ψ as the solution of $(H_s - \lambda)\psi = f$ for $f \in L^2$. Taking the inner product with $\bar{\psi}$ and using that H_s is skew symmetric yields

$$\|\psi\|_2 = \|(H_s - \lambda)^{-1} f\|_2 \leqq 1/\lambda \|f\|_2. \tag{2.11}$$

Since (2.11) holds independently of s, (E1) immediately follows.

To show (E2) we have to prove that H_s generates a *C_0-semigroup* on Y. We define a norm on Y by

$$\|\psi\|_Y := \|\psi\|_2 + \|x\psi\|_2 + \|\psi_x\|_2 + \|\psi_{xx}\|_2. \tag{2.12}$$

Since H_s is a closed operator on L^2, its restriction to Y is closed too, and it remains to estimate the resolvent $\|(H_s - \lambda)^{-1}\|_Y$.

Let $f \in Y$ and let ψ be the solution of

$$(H_s - \lambda)\psi = f, \qquad \lambda > 0;$$

then ψ_x is the solution of

$$(H_s - \lambda)\psi_x = f_x + i/\varepsilon V_x \psi. \tag{2.13}$$

We take the inner product of (2.13) with ψ_x, use that H_s is skew symmetric, and applying the Schwarz inequality we obtain

$$\begin{aligned} \|\psi_x\|_2 &\leqq \frac{1}{\lambda}\left(\|f_x\|_2 + \frac{\|V_x\|_\infty}{\varepsilon}\|\psi\|_2\right) \\ &\leqq \frac{1}{\lambda}\left(\|f_x\|_2 + \frac{\beta}{\lambda\varepsilon}\|f\|_2\right). \end{aligned} \tag{2.14}$$

Analogously we obtain from

$$(H_s - \lambda)\psi_{xx} = f_{xx} + 2i/\varepsilon V_x \psi_x + i/\varepsilon V_{xx}\psi \tag{2.15}$$

and (2.4)-(2.7)

$$\|\psi_{xx}\|_2 \leqq \frac{1}{\lambda}\left(\|f_{xx}\|_2 + \frac{2\beta}{\varepsilon\lambda}\|f_x\|_2 + \left(\frac{2\beta^2}{\varepsilon^2\lambda^2} + \frac{\gamma}{\varepsilon\lambda}\right)\|f\|_2\right). \tag{2.16}$$

In the same way we obtain

$$\|x\psi\|_2 \leqq \frac{1}{\lambda}\left(\|xf\|_2 + \frac{\varepsilon}{\lambda}\|f_x\|_2 + \frac{\beta}{\lambda^2}\|f\|_2\right). \tag{2.17}$$

Summing (2.14)-(2.17) yields

$$\|(H_s - \lambda)^{-1} f\|_Y = \|\psi\|_Y \leqq \frac{1}{\lambda}\left(1 + \frac{c}{\lambda} + \frac{c}{\lambda^2}\right)\|f\|_Y \quad \text{for } \lambda > 0 \tag{2.18}$$

with $c > \max\{2\beta/\varepsilon + \gamma/\varepsilon,\ 2\beta/\varepsilon + \varepsilon,\ 2\beta^2/\varepsilon^2 + \beta\}$, which implies

$$\|(H_s - \lambda)^{-n}\|_Y \leqq \frac{1}{(\lambda - c - 1)^n} \quad \text{for } \lambda > c + 1,$$

and by the Hille-Yoshida theorem [11] H_s generates a C_0-semigroup on Y.

It remains to check condition (E3). Let $\psi \in Y$; then we have

$$\begin{aligned} \lim_{t \to s} \|(H_t - H_s)\psi\|_2 &= \frac{1}{\varepsilon} \lim_{t \to s} \|(V(\cdot, t) - V(\cdot, s))\psi\|_2 \\ &\leqq \lim_{t \to s} \sup_x \frac{1}{\varepsilon} \frac{|V(\cdot, t) - V(\cdot, s)|}{|x| + 1} \|\psi\|_Y = 0. \end{aligned}$$

Therefore the mapping $t \to H_t$, $R^+ \to \mathcal{B}(Y, L^2)$, where $\mathcal{B}(Y, L^2)$ denotes the space of bounded operators from Y into L^2, is continuous. Therefore an evolution system $U(t, s)$ satisfying (2.8)-(2.10) exists.

Finally, we show that $U(t, s)$ is unitary. Let $\psi(t) = U(t, s)\psi_0$ and $\varphi(t) = U(t, s)\varphi_0$ with ψ_0 and $\varphi_0 \in Y$. Using the self-adjointness of iH_t we have

$$\frac{d}{dt}(\psi(t), \varphi(t)) = \left(\frac{d}{dt}\psi, \varphi\right) + \left(\psi, \frac{d}{dt}\varphi\right)$$

$$= (H_t\psi, \varphi) + (\psi, H_t\varphi) = (H_t\psi, \varphi) - (H_t\psi, \varphi) = 0.$$

Since Y is dense in L^2, we therefore have $(U(t, s)\psi, U(t, s)\varphi) = (\psi, \varphi)$ for $\psi, \varphi \in L^2$ and hence $U(t, s)$ is unitary. □

Now we can prove the following estimates.

LEMMA 2.3. *Let $V(x, t)$ satisfy* (2.4)-(2.7), *and let $\psi(t) = U(t, 0)\psi_I$ be the solution of* (2.2); *then the following estimates hold*:

$$\|\psi(t)\|_2 = \|\psi_I\|_2, \tag{2.19}$$

$$\|\psi_x(t)\|_2 \leqq \|\psi_{I,x}\|_2 + \frac{t\beta}{\varepsilon}\|\psi_I\|_2, \tag{2.20}$$

$$\|x\psi(t)\|_2 \leqq \|x\psi_I\|_2 + t\varepsilon\|\psi_{I,x}\|_2 + \frac{\beta t^2}{2}\|\psi_I\|_2. \tag{2.21}$$

Thus the subspace W is invariant under $U(t, s)$.

Proof. To prove (2.20) we differentiate (2.2) with respect to x:

$$\frac{\partial}{\partial t}\psi_x = \frac{i\varepsilon}{2}\frac{\partial^2}{\partial x^2}\psi_x - i/\varepsilon V\psi_x - i/\varepsilon V_x\psi.$$

Using the method of the variation of the constant we obtain

$$\psi_x = U(t, 0)\psi_{I,x} - i/\varepsilon \int_0^t U(t, s)V_x(\cdot, s)\psi(s)\, ds,$$

$$\|\psi_x\|_2 \leqq \|\psi_{I,x}\|_2 + t\beta/\varepsilon\|\psi_I\|_2,$$

which proves (2.20) and (2.21) is obtained similarly. □

In the next lemma we will show that the potential $V(x, t)$ satisfies the assumptions (2.4)-(2.7).

LEMMA 2.4. *Let $V(x)$ be defined in* (2.1) *with* (i) $\psi(t) \in W$ *and* (ii) *the mapping $t \to \psi(x, t)$ is continuous in $[0, \infty) \to L^2$; then the following estimates hold*:

$$|V(x)| \leqq (|x|\|\psi\|_2 + \|y\psi(y)\|_2)\|\psi\|_2, \tag{2.22}$$

$$|V_x(x)| \leqq \|\psi\|_2^2, \tag{2.23}$$

$$|V_{xx}| \leqq 2\|\psi\|_2\|\psi_x\|_2, \tag{2.24}$$

$$\limsup_{s \to t} \sup_{x \in R} \frac{|V(x, t) - V(x, s)|}{|x| + 1} = 0. \tag{2.25}$$

Proof. Statements (2.22) and (2.23) follow immediately from (2.1) and (2.24) is obtained from

$$|V_{xx}| = \left|\frac{\partial}{\partial x}\int_R \operatorname{sign}(y)|\psi(x + y)|^2\, dy\right| \leqq 2\|\psi\|_2\|\psi_x\|_2,$$

and (2.25) follows from

$$\limsup_{s\to t}\sup_{x\in R}\frac{|V(x,t)-V(x,s)|}{|x|+1}$$
$$\leqq \|\psi(t)-\psi(s)\|_2(\|(|y|+1)\psi(t)\|_2+\|(|y|+1)\psi(s)\|_2)=0. \qquad \square$$

Combining the estimates (2.20), (2.21) with (2.22), (2.24) we obtain the following a priori estimates.

LEMMA 2.5. *A solution $\psi(x,t)$, $V(x,t)$ of the Schrödinger–Poisson problem has to satisfy the a priori estimates*

$$\|\psi(t)\|_2=\|\psi_I\|_2, \tag{2.26}$$

$$\|\psi_x(t)\|_2\leqq\|\psi_{I,x}\|_2+\frac{t}{\varepsilon}\|\psi_I\|_2^3, \tag{2.27}$$

$$\|x\psi(t)\|_2\leqq\|x\psi_I\|_2+t\varepsilon\|\psi_{I,x}\|_2+\frac{t^2}{2}\|\psi_I\|_2^3, \qquad |V_{x,t}|\leqq\|\psi_I\|_2. \tag{2.28}$$

To show the existence of a solution of the Schrödinger–Poisson problem on the time interval $[0,T]$, we define the fixed point operator

$$F_T: C([0,T]\to W)\to C([0,T]\to L^2), \qquad F_T\psi(t)=U_{V_\psi}(t,0)\psi_I, \tag{2.29}$$

where $V_\psi:=\int_R|x-y||\psi(y)|^2\,dy$ and $U_V(t,s)$ is the evolution system associated with the family of operators $H_t(V):=(i\varepsilon/2)(d^2/dx^2)-(i/\varepsilon)V(t)$.

We prove the following lemma.

LEMMA 2.6. *The set*

$$S_T:=\{\psi\in C([0,T]\to W):\ \psi(t=0)=\psi_I,\ \psi \text{ satisfies } (2.26)\text{–}(2.28)\}, \tag{2.30}$$

is closed in $C([0,T]\to L^2)$ and invariant under the action of F_T.

Proof. Let $\psi_n\in S_T$ and $\psi_0\in C([0,T]\to L^2)$ with $\lim_{n\to\infty}\psi_n=\psi_0$ in the sense of $C([0,T]\to L^2)$, that is,

$$\lim_{n\to\infty}\sup_t\|\psi_n(t)-\psi_0(t)\|_2=0.$$

We have to show that $\psi_0\in S_T$. Since ψ_n converges to ψ_0 in L^2 for every fixed t, we have $\|\psi_0(t)\|_2=\|\psi_I\|_2$ and $\psi_0(0)=\psi_I$. Now consider t fixed. Since $\psi_n(t)$ converges in $L^2(R)$, $x\psi_n(t)$ converges in $L^2([-X,X])$ to $x\psi_0(t)$ for every $X>0$. Therefore we have

$$\|x\psi_0(t)\|_2^2=\lim_{x\to\infty}\lim_{n\to\infty}\int_{-X}^{X}x^2|\psi_n(x,t)|^2\,dx$$
$$\leqq(1+2\varepsilon t+t^2\|\psi_I\|_2^2)^2\|\psi_I\|_W^2.$$

To show that $\psi_0(t)$ has a derivative with respect to x, which is in L^2, we define $\psi_{0,x}$ in a distributional sense. Let $\phi\in C_0^\infty(R)$; then we define

$$\int_R\psi_{0,x}(x,t)\phi(x)\,dx=-\int_R\psi_0(x,t)\phi_x(x)\,dx$$
$$=-\lim_{n\to\infty}\int_R\psi_n(x,t)\phi_x\,dx$$
$$=\lim_{n\to\infty}\int_R\psi_{n,x}(x,t)\phi\,dx.$$

Using the estimate for $\|\psi_{n,x}\|_2$ we conclude that $\psi_{0,x}$ is a linear functional on L^2 and therefore a L^2-function itself with

$$\|\psi_{0,x}(t)\|_2 \leqq \|\psi_{I,x}\|_2 + \frac{t}{\varepsilon}\|\psi_I\|_2^3.$$

Therefore S_T is closed and applying (2.20), (2.21), (2.19), and (2.23) yields that S_T is invariant under F_T. □

Now we can prove the existence of a solution of the Schrödinger-Poisson problem.

LEMMA 2.7. *For T sufficiently small, the operator F_T has a unique fixed point in S_T.*

Proof. Since S_T is closed, it suffices to show that F_T is a contraction. Let $\phi_1, \phi_2 \in S_T$; then

$$F_T\phi_1 - F_T\phi_2 = \frac{i}{\varepsilon}\int_0^t U_{V_{\phi_1}}(t, s)(V_{\phi_1} - V_{\phi_2})F_T\phi_2\, ds.$$

We estimate

$$\begin{aligned}
&\|F_T\phi_1(t) - F_T\phi_2(t)\|_2 \\
&\quad \leqq \frac{t}{\varepsilon}\|(|x|+1)F_T\phi_2\|_2 \sup_{0\leqq\tau\leqq T, x\in R} \frac{|V_{\phi_1}(x,\tau) - V_{\phi_2}(x,\tau)|}{|x|+1} \\
&\quad \leqq \frac{2t}{\varepsilon}\left(2 + \frac{T}{\varepsilon} + \frac{T^2}{\|\psi_I\|_2^2}\right)^2 \|\psi_I\|_W^2 \sup_{0\leqq\tau\leqq T}\|\phi_1(\tau) - \phi_2(\tau)\|_2.
\end{aligned}$$

To get the last inequality we made use of

$$\begin{aligned}
|V_{\phi_1} - V_{\phi_2}| &= \int_R |x-y|(|\phi_1(y)|^2 - |\phi_2(y)|^2)\, dy \\
&\leqq (|x|+1)(\|(|y|+1)\phi_1\|_2 + \|(|y|+1)\phi_2\|_2)\|\phi_1 - \phi_2\|_2.
\end{aligned}$$

Choosing T sufficiently small, F_T is a contraction on S_T and has therefore a unique fixed point, which is a solution of the Schrödinger-Poisson problem. Using the a priori estimates (2.26)-(2.28) this solution can be extended for all positive times $t > 0$. □

3. The equivalence of the Schrödinger and the Wigner equation. In this section we assume that the potential $V(x, t)$ satisfying (2.4)-(2.7) is prescribed and that $U(t, s)$ is the family of operators defined in Lemma 2.2. We will use the evolution system U of the Schrödinger equation to construct a solution representation for the Wigner equation

$$w_t + v \cdot w_x - \Theta_\varepsilon[V]w = 0, \qquad w(t=0) = w_I. \tag{3.1}$$

By $\mathcal{F}$ we denote the Fourier transform with respect to the velocity variable v:

$$\mathcal{F}w(x, \eta, t) := \frac{1}{\sqrt{2\pi}}\int_R w(x, v, t)\, e^{-i\eta v}\, dv.$$

Then $\mathcal{F}w$ satisfies the Fourier transformed Wigner equation

$$\begin{aligned}
&\mathcal{F}w_t + i\mathcal{F}w_{x\eta} - i/\varepsilon(V(x+\varepsilon\eta/2, t) - V(x-\varepsilon\eta/2, t))\mathcal{F}w = 0, \\
&\mathcal{F}w(t=0) = \mathcal{F}w_I.
\end{aligned} \tag{3.2}$$

By employing the transform

$$p = x + \varepsilon\eta/2, \qquad q = x - \varepsilon\eta/2,$$
$$(Cg)(p, q) = \varepsilon^{-1} g(x, \eta), \qquad C : L^2(R_x \times R_\eta) \to L^2(R_p \times R_q), \tag{3.3}$$

and setting $z = C\mathcal{F}w$, $z_I = C\mathcal{F}w_I$, we obtain an evolution equation for z:

$$z_t := Qz, \qquad z(t=0) = z_I, \tag{3.4}$$

with

$$Qz := \left[i\frac{\varepsilon}{2}\frac{\partial^2}{\partial q^2} - \frac{i}{\varepsilon} V(q, t)) \right] z - \left[i\frac{\varepsilon}{2}\frac{\partial^2}{\partial p^2} - \frac{i}{\varepsilon} V(p, t) \right] z.$$

Using the notation of tensor products of Hilbert spaces [12] we can represent the operator Q as

$$Q = I \otimes H_t - H_t \otimes I. \tag{3.5}$$

Now we can write the solution operator of (3.4) by using the evolution system $U(t, s)$ defined in Lemma 2.2.

LEMMA 3.1. *The family of operators*

$$\{T(t, s) = \overline{U(t, s)} \otimes U(t, s) \; 0 \leqq s \leqq t\} \tag{3.6}$$

is a strongly continuous unitary evolution system with

$$\frac{\partial^+}{\partial t} T(t, s) z|_{t=s} = Qz \quad \textit{for } z \in Y \otimes Y, \tag{3.7}$$

$$\frac{\partial}{\partial s} T(t, s) z = -T(t, s) Qz \quad \textit{for } z \in Y \otimes Y. \tag{3.8}$$

The proof runs along the same lines as in [9]. This yields the following solution representation.

THEOREM 3.2. *Let $w_I \in L^2(R^2)$ be real valued; then there exists a complete orthonormal system $\varphi_{I,k}$, $k \in N$ of $L^2(R)$ and real numbers λ_k, $k \in N$, so that the solution of the Wigner equation is given by*

$$w(t) = \mathcal{F}^{-1} C^{-1} \sum_{k \in N} \lambda_k \overline{(U(t, 0)\varphi_k)(p)} (U(t, 0)\varphi_k)(q). \tag{3.9}$$

We associate with the initial condition $z_I = C\mathcal{F}w_I$ the operator

$$Z_I : L^2(R) \to L^2(R), \qquad (Z_I f)(p) = \int_R z_I(p, q) f(q)\, dq.$$

Since w_I is real valued, $z_I(p, q) = \overline{z_I(q, p)}$ and thus Z_I is self-adjoint. Since $z_I \in L^2(R \times R)$, Z_I is a Hilbert–Schmidt operator and therefore compact. Therefore Z_I has a complete orthogonal set of eigenfunctions $\varphi_{I,k}$, $k \in N$ and real eigenvalues λ_k with

$$z_I(p, q) = \sum_{k \in N} \lambda_k \overline{\varphi_{I,k}(p)}\, \varphi_{I,k}(q) \tag{3.10}$$

with $\sum_{k \in N} \lambda_k^2 < \infty$. Now (3.10) and (3.6) imply the assertion of the theorem. □

In other words, the solution of the Wigner equation (1.2) is equivalent to the solution of a countable series of Schrödinger equations:

$$i\varepsilon \frac{\partial}{\partial t} \psi_k = -\frac{\varepsilon^2}{2} \frac{\partial^2}{\partial x^2} \psi_k + V\psi_k, \quad \psi_k(t=0) = \varphi_{I,k}, \quad k \in N, \tag{3.11}$$

where $(\lambda_k, \varphi_{I,k})$ are eigenvalues and eigenfunctions of the operator Z_I associated with the initial condition $z_I = C\mathcal{F}w_I$.

We define the Hilbert space $\mathcal{H}$ of sequences of L^2-functions:

$$\mathcal{H} := \left\{ \psi = (\psi_1, \psi_2, \cdots),\ \psi_k \in L^2(R),\ \sum_{k=1}^{\infty} \lambda_k \|\psi_k\|_2^2 < \infty \right\},$$

with the inner product

$$(\varphi, \psi)_{\mathcal{H}} := \sum_{k=1}^{\infty} \lambda_k (\varphi_k, \psi_k).$$

We define the sequences $x\psi$, $(\partial/\partial x)\psi$ elementwise and using Lemma 2.3 we obtain for a solution of (3.11) the following lemma.

LEMMA 3.3. *Let* $\|\varphi_I\|_{\mathcal{H}} < \infty$, $\|x\varphi_I\|_{\mathcal{H}} < \infty$, $\|(\partial/\partial x)\varphi_I\|_{\mathcal{H}} < \infty$; *then a solution of* (3.11) *satisfies*

$$\|\psi(t)\|_{\mathcal{H}} = \|\varphi_I\|_{\mathcal{H}}, \tag{3.12}$$

$$\|\psi_x(t)\|_{\mathcal{H}} \leqq \left\| \frac{\partial}{\partial x} \varphi_I \right\|_{\mathcal{H}} + \frac{t\beta}{\varepsilon} \|\varphi_I\|_{\mathcal{H}}, \tag{3.13}$$

$$\|x\psi(t)\|_{\mathcal{H}} \leqq \|\varphi_I\|_{\mathcal{H}} + t\varepsilon \left\| \frac{\partial}{\partial x} \varphi_I \right\|_{\mathcal{H}} + t^2\beta/2 \|\varphi_I\|_{\mathcal{H}} \tag{3.14}$$

with $\beta \geqq \|V_x\|_\infty$.

The next lemma gives a criterion for $x\varphi_I$, $(\partial/\partial x)\varphi_I \in \mathcal{H}$ in terms of the initial condition z_I.

LEMMA 3.4. *Suppose that the operators* Z_I, Z_I^*, Z_I' *with the kernels* z_I, $p \cdot q \cdot z_I$, $(\partial^2/\partial p\, \partial q) z_I \in L^2$ *are self-adjoint, of trace class, and* $\lambda_k \geqq 0$; *then* $\sum_{k=1}^{\infty} \lambda_k \|x\varphi_{I,k}\|_2^2$, $\sum_{k=1}^{\infty} \lambda_k \|(\partial/\partial x)\varphi_{I,k}\|_2^2$ *converge.*

Proof. The trace of Z_I' is given by

$$\begin{aligned}
\operatorname{tr}(Z_I') &= \sum_{k=1}^{\infty} (\varphi_{I,k}, Z_I' \varphi_{I,k}) \\
&= \sum_{k=1}^{\infty} \int_R \int_R \varphi_{I,k}(p) \overline{\varphi_{I,k}(q)} \overline{\frac{\partial^2}{\partial p\, \partial q} z_I(p, q)}\, dp\, dq, \\
&\sum_{k=1}^{\infty} \int_R \int_R \overline{z_I(p, q)} \frac{\partial}{\partial p} \varphi_{I,k}(p) \frac{\partial}{\partial q} \overline{\varphi_{I,k}(q)}\, dp\, dq \\
&= \sum_{k=1}^{\infty} \int_R \int_R \sum_{n=1}^{\infty} \lambda_n \overline{\varphi_{I,n}(p)} \varphi_{I,n}(q) \frac{\partial}{\partial p} \varphi_{I,k}(p) \frac{\partial}{\partial q} \overline{\varphi_{I,k}(q)}\, dp\, dq \\
&= \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \lambda_n \left| \left(\varphi_{I,n}, \frac{\partial}{\partial p} \varphi_{I,k} \right) \right|^2 = \sum_{n=1}^{\infty} \lambda_n \left\| \frac{\partial}{\partial x} \varphi_{I,n} \right\|^2.
\end{aligned}$$

A similar calculation gives

$$\operatorname{tr}(Z_I^*) = \sum_{n=1}^{\infty} \lambda_n \|x\varphi_{I,n}\|_2^2. \qquad \square$$

4. The Wigner–Poisson problem. In the previous section we have considered the Wigner equation for a prescribed potential. The goal of this section is to show the existence of a solution of (1.1)–(1.3). To define the potential by the Poisson equation

in integral form (1.1) we need that the particle density $n(x)=\int_R w(x, v)\,dv$ and $xn(x)$ are L^1-functions. This fact is not guaranteed a priori if we assume that only the initial density n_I and xn_I are well-defined L^1 functions. To exploit the equivalence of the Schrödinger and Wigner equations we will define a generalized particle density. Suppose that $w_I(x, v)$ is sufficiently smooth and decays sufficiently fast; then we have formally

$$\begin{aligned} n(x,t) &= \int_R w(x,v,t)\,dv = \frac{1}{\sqrt{2\pi}}\,\mathcal{F}w(x,\eta=0,t) \\ &= \frac{\varepsilon}{\sqrt{2\pi}}\,z(x,x,t) = \frac{\varepsilon}{\sqrt{2\pi}}\sum_{k=1}^{\infty}\lambda_k|U(t,0)\varphi_{I,k}(x)|^2, \end{aligned} \tag{4.1}$$

where λ_k and $\varphi_{I,k}$ are defined in Theorem 3.2.

We say that an initial function z_I is of trace class (positive definite) if the associated operator Z_I is of trace class (positive definite). Now we define a generalized particle density in the following lemma.

LEMMA 4.1. *Let z_I be of trace class; then we define the generalized particle density by*

$$n_g(\cdot,t) := \frac{\varepsilon}{\sqrt{2\pi}}\sum_{k=1}^{\infty}\lambda_k|U(t,s)\varphi_{I,k}(\cdot)|^2 \in L^1(R).$$

If, additionally, z_I is positive definite, n_g is not negative.

Proof. Since z_I is of trace class, $\sum_{k=1}^{\infty}\lambda_k$ converges absolutely and therefore

$$\begin{aligned} \|n_g(t)\|_1 &= \frac{\varepsilon}{\sqrt{2\pi}}\int_R\left|\sum_{k=1}^{\infty}\lambda_k|(U(t,0)\varphi_{I,k})(x)|^2\right|dx \\ &\leqq \frac{\varepsilon}{\sqrt{2\pi}}\sum_{k=1}^{\infty}|\lambda_k|. \end{aligned}$$

If Z_I is positive definite, then the eigenvalues λ_k of Z_I are positive and therefore n_g is nonnegative. Note that the positive definiteness of z_I is not equivalent with $w_I \geqq 0$. □

LEMMA 4.2. *Let vw and $w_x \in L^2$ and $\sum_{k=1}^{\infty}\lambda_k < \infty$; then*

$$n_g(x) = n(x) = \int_R w(x,v)\,dv \in L^1(R).$$

Proof. From *vw*, $w_x \in L^2$ we conclude that $z \in H^1$, which implies that $n(x,t) = (\varepsilon/\sqrt{2\pi})z(x,x,t) \in L^1_{\text{loc}}$ and therefore coincides with n_g in the L^1-sense, and by Lemma 4.1 we have $n \in L^1(R)$. □

Now we can prove the existence theorem.

THEOREM 4.3. *Let w_I be real valued, so that $z_I = C\mathcal{F}w_I$ is positive definite and that z_I, pqz_I, $(\partial^2/\partial p\,\partial q)z \in L^2$ are of trace class; then the Wigner–Poisson problem*

$$V(x) = \int_R |x-y|n_g(y)\,dy, \tag{4.2}$$

$$w_t + v\cdot w_x - \Theta_\varepsilon[V]w = 0, \qquad x \in R_x, \quad v \in R_v, \tag{4.3}$$

$$w(x, v, t=0) = w_I(x, v), \tag{4.4}$$

where n_g is the generalized particle density, has a unique mild solution.

Remark. The term "mild solution" means that the Wigner equation is satisfied in the mild sense.

Proof. Let $\varphi_{I,k}$, λ_k be the eigenfunctions and eigenvalues of Z_I. Then the Wigner-Poisson problem is equivalent to

$$V(x) = \int_R |x-y| \sum_{k=1}^{\infty} \frac{\varepsilon}{\sqrt{2\pi}} \lambda_k |\psi_k(y)|^2 \, dy,$$

$$i\varepsilon \frac{\partial}{\partial t} \psi_k = -\frac{\varepsilon^2}{2} \frac{\partial^2}{\partial x^2} \psi_k + V\psi_k, \qquad \psi_k(t=0) = \varphi_{I,k}.$$

We define the set $\mathcal{S}_T \subset C([0, T] \to \mathcal{H})$:

$$\mathcal{S}_T := \Big\{ \psi(\cdot, t) \in \mathcal{H} \text{ for } 0 \leqq t \leqq T, \text{ with } \psi_k(t=0) = \varphi_{I,k},\ k \in N, \tag{4.5}$$
$$\cdot \left\| \frac{\partial}{\partial x} \psi(t) \right\|_{\mathcal{H}} \leqq B(t),\ \|x\psi(t)\|_{\mathcal{H}} \leqq C(t),\ \|\psi_k(t)\|_2 = 1 \Big\},$$

where

$$B(t) = \left\| \frac{\partial}{\partial x} \varphi_I \right\|_{\mathcal{H}} + t/\varepsilon \|\varphi_I\|_{\mathcal{H}}^3 \quad \text{and} \quad C(t) = \|\varphi_I\|_{\mathcal{H}} + t\varepsilon \left\| \frac{\partial}{\partial x} \varphi_I \right\|_{\mathcal{H}} + \frac{t^2}{2} \|\varphi_I\|_{\mathcal{H}}^3.$$

Proceeding analogously as in § 2, it follows that $\mathcal{S}_T$ is a closed set invariant under the operator F_T:

$$F_T : \mathcal{S}_T \to C([0, T] \to \mathcal{H}), \quad (F_T\psi)_k(t) = U_{V_\psi}(t, 0)\varphi_{I,k}, \quad k \in N.$$

In the next step we show that F_T is a contraction for T small enough.

From

$$\|(F\psi)_k(t) - (F\varphi)_k(t)\|_2$$
$$\leqq T \sup_{0 \leqq t \leqq T,\, x \in R} \frac{|V_\psi(x, t) - V_\varphi(x, t)|}{|x|+1} (\|x(F_T\psi)_k(t)\|_2 + \|(F_T\varphi)_k\|_2)$$

and

$$\left| \frac{V_\psi(x, t) - V_\varphi(x, t)}{|x|+1} \right| = \frac{\varepsilon}{\sqrt{2\pi}} \left| \int_R \frac{|x-y|}{|x|+1} \sum_{k=1}^{\infty} \lambda_k (|\varphi_k|^2 - |\psi_k|^2) \, dy \right|$$
$$\leqq \|\psi - \varphi\|_{\mathcal{H}} (\|\psi\|_{\mathcal{H}} + \|x\psi\|_{\mathcal{H}} + \|\varphi\|_{\mathcal{H}} + \|x\varphi\|_{\mathcal{H}}),$$

follows the estimate

$$\|F_T\psi - F_T\varphi\|_{\mathcal{H}}$$
$$\leqq \frac{\varepsilon}{\sqrt{2\pi}} T(\|\varphi\|_{\mathcal{H}} + \|x\varphi\|_{\mathcal{H}} + \|\psi\|_{\mathcal{H}} + \|x\psi\|_{\mathcal{H}})(\|F_T\varphi\|_{\mathcal{H}} + \|xF_T\psi\|_{\mathcal{H}}) \|\psi - \varphi\|_{\mathcal{H}}$$
$$\leqq \frac{\varepsilon}{\sqrt{2\pi}} 2C(T)^2 T \|\psi - \varphi\|_{\mathcal{H}}.$$

Now choosing T sufficiently small yields that F_T is a contraction and therefore has a unique fixed point in $\mathcal{S}_T$, which is a mild solution of the Wigner-Poisson problem on the interval $[0, T]$. Due to the a priori estimates (4.5) this solution can be extended to any $t \in R^+$.

5. The classical limit. In this section we will prove that solutions of the Wigner-Poisson problem (1.1)-(1.3) converge (after extracting a subsequence) to a generalized solution of the Vlasov-Poisson problem as $\varepsilon \to 0$. The analysis in this section is guided by the analysis in the corresponding sections in [13] and [1].

THEOREM 5.1. *Let $T>0$ and $(\varepsilon_m)_{m\in N}$ be a monotone decaying sequence of positive real numbers with limit zero (in the following we drop the index m) and let $w^{(\varepsilon)}$ be the solution of the Wigner–Poisson problem:*

$$V^{(\varepsilon)}(x)=\int_R |x-y|n^{(\varepsilon)}\,dy, \qquad n^{(\varepsilon)}(x)=\int_R w^{(\varepsilon)}(x,v)\,dv, \tag{5.1}$$

$$\frac{\partial}{\partial t}w^{(\varepsilon)}+v\cdot\frac{\partial}{\partial x}w^{(\varepsilon)}-\Theta_\varepsilon[V^{(\varepsilon)}]w^{(\varepsilon)}=0, \qquad x\in R,\quad v\in R, \tag{5.2}$$

$$w^{(\varepsilon)}(x,v,t=0)=w_I^{(\varepsilon)}(x,v) \tag{5.3}$$

with $w_I^{(\varepsilon)}$ satisfies the assumption of the existence Theorem 4.3 and the terms

$$\|w_I^{(\varepsilon)}\|_2, \quad \|xw_I^{(\varepsilon)}\|_2, \quad \|vw_I^{(\varepsilon)}\|_2, \tag{5.4}$$

$$\|w_I^{(\varepsilon)}\|_1, \quad \|x^2w_I^{(\varepsilon)}\|_1, \quad \|v^2w_I^{(\varepsilon)}\|_1, \quad \varepsilon^2\left\|\frac{\partial^2}{\partial x^2}w_I^{(\varepsilon)}\right\|_1, \tag{5.5}$$

are uniformly bounded with respect to ε. Then there exists a subsequence ε_{m_j} (which we denote as the original sequence) and functions $w^{(0)}$, $w_I^{(0)}$, $V_x^{(0)}$ with

$$\lim_{\varepsilon\to 0} w^{(\varepsilon)}=w^{(0)} \quad \text{in } L^2([0,T]\to L^2(R^2)) \text{ weakly}, \tag{5.6}$$

$$\lim_{\varepsilon\to 0} vw^{(\varepsilon)}=vw^{(0)} \quad \text{in } L^2([0,T]\to L^2(R^2)) \text{ weakly}, \tag{5.7}$$

$$\lim_{\varepsilon\to 0} w_I^{(\varepsilon)}=w_I^{(0)} \quad \text{in } L^2([0,T]\to L^2(R^2)) \text{ weakly}, \tag{5.8}$$

$$\lim_{\varepsilon\to 0} V_x^{(\varepsilon)}=V_x^{(0)} \quad \text{in } L^\infty(R\times[0,T]) \text{ weak}^*, \tag{5.9}$$

where $w^{(0)}$, $V^{(0)}$ is a generalized solution of the Vlasov–Poisson problem:

$$V^{(0)}(x)=\int_R |x-y|n^{(0)}\,dy, \qquad n^{(0)}(x)=\int_R w^{(0)}(x,v)\,dv, \tag{5.10}$$

$$\frac{\partial}{\partial t}w^{(0)}+v\cdot\frac{\partial}{\partial x}w^{(0)}-V_x^{(0)}\frac{\partial}{\partial v}w^{(0)}=0, \qquad x\in R,\quad v\in R, \tag{5.11}$$

$$w^{(0)}(x,v,t=0)=w_I^{(0)}(x,v). \tag{5.12}$$

In the following two lemmas we will show that certain quantities are bounded uniformly with respect to ε.

LEMMA 5.2. *Let the assumptions of Theorem 5.1 hold; then there exists a constant $D(T)$ independent of ε with*

$$\|vw^{(\varepsilon)}(t)\|_2\leqq D(\|w_I^{(\varepsilon)}\|_2+\|vw_I^{(\varepsilon)}\|_2) \quad \text{for } 0\leqq t\leqq T, \tag{5.13}$$

and

$$\|xw^{(\varepsilon)}(t)\|_2\leqq D(\|w_I^{(\varepsilon)}\|_2+\|xw_I^{(\varepsilon)}\|_2+\|vw_I^{(\varepsilon)}\|_2) \quad \text{for } 0\leqq t\leqq T. \tag{5.14}$$

Proof. The function $vw^{(\varepsilon)}$ solves the equation

$$(vw^{(\varepsilon)})_t+v(vw^{(\varepsilon)})_x-\Theta_\varepsilon[V^{(\varepsilon)}](vw^{(\varepsilon)})$$
$$=-\frac{1}{4\pi}\int_R\int_R\left(V_x^{(\varepsilon)}\left(x-\frac{\varepsilon\eta}{2}\right)+V_x^{(\varepsilon)}\left(x+\frac{\varepsilon\eta}{2}\right)\right)e^{i\eta(v-v')}w^{(\varepsilon)}(x,v',t)\,dv'\,d\eta.$$

Using $\|V_x^{(\varepsilon)}\|_\infty \leqq \|n_I^{(\varepsilon)}\|_1 \leqq \|w_I^{(\varepsilon)}\|_1$ we obtain

$$\|vw^{(\varepsilon)}(t)\|_2 \leqq \|vw_I^{(\varepsilon)}\|_2 + T\|w_I^{(\varepsilon)}\|_1 \|w_I^{(\varepsilon)}\|_2.$$

The second inequality is obtained similarly. □

LEMMA 5.3. *Let the assumptions of Theorem* 5.1 *hold; then there exists a constant* c *depending only on the quantities in* (5.5) *with*

$$\|xn(t)\|_1 \leqq (t^2+1)c. \tag{5.15}$$

Proof.

$$\begin{aligned}
\|xn\|_1 &= \frac{\varepsilon}{\sqrt{2\pi}} \int_R |xz^{(\varepsilon)}(x, x)|\, dx \\
&= \frac{\varepsilon}{\sqrt{2\pi}} \int_R \sum_{k=1}^{\infty} \lambda_k |xU(t,0)\varphi_{I,k}|\, |U(t,0)\varphi_{I,k}|\, dx \\
&\leqq \frac{\varepsilon}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \lambda_k \|xU(t,0)\varphi_{I,k}\|_2 \\
&\leqq \frac{\varepsilon}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \lambda_k \left(\|x\varphi_{I,k}\|_2 + t\varepsilon \left\| \frac{\partial}{\partial x} \varphi_{I,k} \right\|_2 + \frac{t^2\beta}{2} \|\varphi_{I,k}\|_2 \right)
\end{aligned} \tag{5.16}$$

with $\beta = \sup |V_x| = \|w_I\|_1$. Using the definition of the generalized particle density we estimate

$$\begin{aligned}
\frac{\varepsilon}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \lambda_k \|x\varphi_{I,k}\|_2 &\leqq \sqrt{\frac{\varepsilon}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \lambda_k \|x\varphi_{I,k}\|_2^2} \sqrt{\frac{\varepsilon}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \lambda_k} \\
&\leqq \sqrt{\|x^2 w_I\|_1 \|w\|_1}
\end{aligned}$$

and similarly

$$\frac{\varepsilon^3}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \lambda_k \left\| \frac{\partial}{\partial x} \varphi_{I,k} \right\|_2 \leqq \sqrt{\left\| \frac{\varepsilon^2}{4} w_{I,xx} + v^2 w_I \right\|_1 \|w_I\|_1}.$$

Inserting these estimates into (5.16) yields Lemma 5.3. □

Now we are able to prove Theorem 5.1. By (5.13), (5.14), and the assumption of Theorem 5.1 the functions $w^{(\varepsilon)}$, $vw^{(\varepsilon)}$, and $w_I^{(\varepsilon)}$ are uniformly bounded with respect to ε in the L^2-norm. Due to the weak compactness of a bounded set in L^2 there exist subsequences of $w^{(\varepsilon)}$, $vw^{(\varepsilon)}$, $w_I^{(\varepsilon)}$ which converge weakly to the functions $w^{(0)}$, $vw^{(0)}$, $w_I^{(0)}$. Since the L^1 norm of $n^{(\varepsilon)}$ is bounded uniformly, $V_x^{(\varepsilon)}$ is bounded uniformly and therefore there exist a subsequence of $V_x^{(\varepsilon)}$ and a limit function $V_x^{(0)}$ with

$$\lim_{\varepsilon \to 0} V_x^{(\varepsilon)} = V_{I,x}^{(0)} \quad \text{in } L^\infty(R \times [0, T]) \text{ weak*}.$$

It remains to show that $w^{(0)}$, $V_x^{(0)}$ satisfy the Vlasov-Poisson equation.

We multiply the Poisson equation by a test function $\sigma \in C_0^\infty(R \times [0, T])$ and choose a positive constant A:

$$\begin{aligned}
\int_0^T \int_R V_x^{(\varepsilon)} \sigma\, dx\, dt &= \int_0^T \int_R \int_R \operatorname{sign}(x-y) n^{(\varepsilon)}(y, t) \sigma(x, t)\, dy\, dx\, dt \\
&= \int_0^T \int_{-A}^{A} \int_R (1+|v|) w^{(\varepsilon)}(y, v, t) \frac{\int \operatorname{sign}(x-y)\sigma(x, t)\, dx}{1+|v|}\, dv\, dy\, dt \\
&\quad + \int_0^T \int_{R \setminus [-A, A]} \int_R \operatorname{sign}(x-y) n^{(\varepsilon)}(y, v, t) \sigma(x, t)\, dx\, dy\, dt.
\end{aligned}$$

We can estimate the second term by $\mu(\text{supp}(\sigma))\|\sigma\|_\infty\|xn^{(\varepsilon)}\|_1/A$. Note that μ denotes the Lebesgue measure. Using that $vw^{(\varepsilon)}$ converges weakly to $vw^{(0)}$, we obtain

$$\left|\int_0^T\int_R\left(V_x^{(0)}-\int_R \text{sign}(x-y)n^{(0)}(y,t)\,dy\right)\sigma(x,t)\,dx\,dt\right|$$
$$\leqq c(1+t^2)\mu(\text{supp}(\sigma))\|\sigma\|_\infty/A.$$

Taking the limit $A\to\infty$, we conclude that $V^{(0)}$, $n^{(0)}$ satisfy the Poisson equation.

We denote by $\Omega_A := [-A, A]^2$ and $\Omega_{A,T} := [-A, A]^2\times[0, T]$. Then we multiply the Fourier transformed Wigner equation by a test function $\varphi\in C_0^\infty(R^2\times[0,T])$ with $\text{supp}(\varphi)\subset\Omega_{A,T}$. Integration by parts yields

$$-\int_{\Omega_{A,T}}\mathcal{F}w^{(\varepsilon)}\varphi_t\,d\eta\,dx\,dt+i\int_{\Omega_{A,T}}\mathcal{F}w^{(\varepsilon)}\varphi_{x\eta}\,d\eta\,dx\,dt$$
$$-i/\varepsilon\int_{\Omega_{A,T}}\mathcal{F}w^{(\varepsilon)}(V^{(\varepsilon)}(x+\varepsilon\eta/2)-V^{(\varepsilon)}(x-\varepsilon\eta/2))\varphi\,d\eta\,dx\,dt$$
$$=\int_{\Omega_A}\mathcal{F}w_I^{(\varepsilon)}\pi(t=0)\,d\eta\,dx.$$

Now using the weak convergence of $w^{(\varepsilon)}$, $w_I^{(\varepsilon)}$ we obtain

$$\int_{\Omega_{A,T}}\mathcal{F}w^{(\varepsilon)}\varphi_t\,d\eta\,dx\,dt\to\int_{\Omega_{A,T}}\mathcal{F}w^{(0)}\varphi_t\,d\eta\,dx\,dt,$$
$$\int_{\Omega_{A,T}}\mathcal{F}w^{(\varepsilon)}\varphi_{x\eta}\,d\eta\,dx\,dt\to\int_{\Omega_{A,T}}\mathcal{F}w^{(0)}\varphi_{x\eta}\,d\eta\,dx\,dt,$$
$$\int_{\Omega_A}\mathcal{F}w_I^{(\varepsilon)}\varphi(t=0)\,d\eta\,dx\to\int_{\Omega_A}\mathcal{F}w_I^{(0)}\varphi(t=0)\,d\eta\,dx.$$

Only the third term on the left-hand side has to be taken care of. In the following we set $\delta_\varepsilon V=(V(x+\varepsilon\eta/2)-V(x-\varepsilon\eta/2))/\varepsilon$.

$$\int_{\Omega_{A,T}}\delta_\varepsilon V^{(\varepsilon)}\mathcal{F}w^{(\varepsilon)}\varphi\,dx\,d\eta\,dt$$

(5.17)
$$=\int_{\Omega_{A,T}}\eta V_x^{(0)}\mathcal{F}w_\varphi^{(0)}\,dx\,d\eta\,dt$$

(5.18)
$$+\int_{\Omega_{A,T}}(\eta V_x^{(\varepsilon)}-\eta V_x^{(0)})\mathcal{F}w^{(0)}\varphi\,dx\,d\eta\,dt$$

(5.19)
$$+\int_{\Omega_{A,T}}\eta V_x^{(\varepsilon)}(\mathcal{F}w^{(\varepsilon)}-\mathcal{F}w^{(0)})\varphi\,dx\,d\eta\,dt$$

(5.20)
$$+\int_{\Omega_{A,T}}(\delta_\varepsilon V^{(\varepsilon)}-\eta V_x^{(\varepsilon)})\mathcal{F}w^{(\varepsilon)}\varphi\,dx\,d\eta\,dt.$$

Since $V_x^{(\varepsilon)}$ converges in $L^\infty(R\times[0,T])$ weak* and $\eta\mathcal{F}w^{(0)}\in L^1$, the first term (5.18) converges to zero:

$$\int_{\Omega_{A,T}}(\eta V_x^{(\varepsilon)}-\eta V_x^{(0)})\mathcal{F}w^{(0)}\varphi\,dx\,d\eta\,dt\to 0\quad\text{as }\varepsilon\to 0.$$

Since $\mathcal{F}w_\eta^{(\varepsilon)}$ and $\delta_\varepsilon V^{(\varepsilon)}\mathcal{F}w^{(\varepsilon)}$ are uniformly bounded in $L^\infty([0, T]\to L^2(\Omega_A))$ and using

$$\mathcal{F}w_t^{(\varepsilon)} = -i\frac{\partial}{\partial x}\mathcal{F}w_\eta^{(\varepsilon)} + \delta_\varepsilon V^{(\varepsilon)}\mathcal{F}w^{(\varepsilon)},$$

we can interpret $\mathcal{F}w^{(\varepsilon)}(t)$ as a functional on $H^1(\Omega_A)$ with

$$\|\mathcal{F}w^{(\varepsilon)}\|_{W^{1,\infty}([0,T]\to H^{-1}(\Omega_A))} \leqq F,$$

where F is a constant independent of ε. Therefore we can extract a subsequence with

$$\mathcal{F}w^{(\varepsilon)} \to \mathcal{F}w^{(0)} \quad \text{strongly in } L^\infty([0, T]\to H^{-1}(\Omega_A)).$$

Since

$$\|V_{xx}^{(\varepsilon)}\|_2^2 = \|n^{(\varepsilon)}\|_2^2 \leqq \text{const. } \|(1+|v|)w^{(\varepsilon)}\|_2^2,$$

we conclude that $\eta V_x^{(\varepsilon)}\varphi$ is uniformly bounded in $L^\infty([0, T]\to H^1(\Omega_A))$ as $\varepsilon\to 0$ and therefore we have

$$\int_{\Omega_{A,T}} \eta V_x^{(\varepsilon)}(\mathcal{F}w^{(\varepsilon)} - \mathcal{F}w^{(0)})\varphi\, dx\, d\eta\, dt \to 0 \quad \text{as } \varepsilon\to 0.$$

To show the convergence of the fourth term (5.20), we estimate

$$\begin{aligned}
|\delta_\varepsilon V^{(\varepsilon)} - \eta V_x^{(\varepsilon)}| &= \int_R \left(\frac{|x-y+\varepsilon\eta/2| - |x-y-\varepsilon\eta/2|}{\varepsilon} - \eta \operatorname{sign}(x-y)\right) n^{(\varepsilon)}\, dy \\
&\leqq \int_{x-\varepsilon|\eta|/2}^{x+\varepsilon|\eta|/2} (2|x-y|/\varepsilon + |\eta|) n^{(\varepsilon)}(y)\, dy \\
&\leqq \text{const. } \varepsilon^{1/2}\eta^{3/2}\|(1+|v|)w^{(\varepsilon)}\|_2,
\end{aligned}$$

and we obtain

$$\int_{\Omega_{A,T}} (\delta_\varepsilon V^{(\varepsilon)} - \eta V_x^{(\varepsilon)})\mathcal{F}w^{(\varepsilon)}\varphi\, dx\, d\eta\, dt \to 0 \quad \text{as } \varepsilon\to 0,$$

which concludes the proof of Theorem 5.1. □

REFERENCES

[1] A. ARNOLD AND P. MARKOWICH, *The periodic quantum Liouville Poisson problem*, Boll.Un. Mat. Ital. (7)4-B (1990), pp. 449-484.

[2] A. ARNOLD AND H. STEINRÜCK, *The electromagnetic Wigner equation for an electron with spin*, Z. Angew. Math. Phys., 40 (1989), pp. 793-815.

[3] G. BERTSCH, *Heavy ion dynamics of intermediate energy*, manuscript, Cyclotron Laboratory and Physics Department, Michigan State University, East Lansing, MI, 1980.

[4] A. ARNOLD AND F. NIER, *The two-dimensional Wigner-Poisson problem for an electron gas in the charge neutral case*, manuscript, Fachbereich Mathematik, TU, Berlin, Strasse des 17 Juni 136, D-1000 Berlin, FRG, 1990, Math. Methods Appl. Sci., to appear.

[5] F. BREZZI AND P. MARKOWICH, *The* 3D *Wigner Poisson problem, existence, uniqueness and approximations*, manuscript, Fachbereich Mathematik, TU, Berlin, Strasse des 17 Juni 136, D-1000 Berlin, FRG; Comm. Math. Phys. (1990), submitted.

[6] P. DEGOND AND P. MARKOWICH, *A quantum transport model for semi-conductors: the Wigner Poisson problem on a bounded Brillouin zone*, manuscript, Centre de Mathématique Appliquées, Ecole Polytechnique, F-91128 Palaiseau, France, 1989.

[7] N. KLUCKSDAHL, A. KRIMAN, D. FERRY, AND C. RINGHOFER, *Selfconsistent study of the resonant tunneling diode*, Phys. Rev. B, 39 (1989), pp. 7720-7735.

[8] W. LUI AND J. FREY, *A simplified method for quantum size effect analysis in submicron devices*, J. Appl Phys. 64 (1988), pp. 6790–6794.

[9] P. MARKOWICH, *On the equivalence of the Schrödinger and the quantum Liouville equation*, Math. Methods Appl. Sci., 11 (1989), pp. 459–469.

[10] P. MARKOWICH AND C. RINGHOFER, *An analysis of the quantum Liouville equation*, Z. Angew. Math. Mach., 69 (1989), pp. 121–127.

[11] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, Tokyo, 1983.

[12] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics*, I. *Functional Analysis*, II. *Fourier Analysis, Selfadjointness*, Academic Press, New York, San Francisco, London, 1975.

[13] H. STEINRÜCK, *Wigner–Poisson problem in crystal: existence, uniqueness, semiclassical limit in the one-dimensional case*, Z. Angew. Math. Mech. (1990), to appear.

[14] E. WIGNER, *On the quantum correction for thermodynamic equilibrium*, Phys. Rev., 40 (1932), pp. 749–759.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 973-981, July 1991

006

A BOUNDARY INTEGRAL EQUATION FOR THE TWO-DIMENSIONAL FLOATING-BODY PROBLEM*

Y. W. LIU†

Abstract. The time-harmonic two-dimensional finite-depth floating-body problem is reformulated as a boundary integral equation. As a result of choosing a simple kernel function, the integral equation extends over both the wetted portion of the floating body and the free surface. It is also shown that this integral equation suffers no irregular frequencies, that is, it has at most one solution.

Key word. floating-body problem

AMS(MOS) subject classification. 76B15

1. Introduction.

In 1950, Fritz John [4] published a paper analyzing time-harmonic motion of a fluid in which an impenetrable body is partially immersed. This floating-body problem was formulated mathematically as a boundary value problem for Laplace's equation in $\mathcal{R}^3$ with appropriate boundary and radiation conditions. The two-dimensional problem is formulated in a similar way with slight modification. Using the Green's function suggested by John, the boundary value problem can be reduced to an integral equation over the wetted portion of the floating body. Just as John demonstrated the existence of irregular frequencies for the three-dimensional case, the two-dimensional case also suffers irregular frequencies (e.g., Ursell [7] and Liu [6]). By irregular frequencies is meant frequencies for which the integral equation is not uniquely solvable even though the solution of the original boundary value problem is unique.

Another way to treat this problem is to employ a simpler Green's function, which only satisfies the boundary condition at the bottom of the fluid. Hence the corresponding integral equation is defined over the wetted surface of the floating body and the free surface as well. For both three- and two-dimensional problems, such integral equations have been derived and even solved numerically for certain cases, e.g., Yeung [9] and Bai and Yeung [2]. Numerical evidence indicates that these integral equations do not exhibit irregular frequencies but does not constitute a conclusive analytical argument to support this conjecture. Recently, Angell, Hsiao, and Kleinman [1] presented a proof of the conjecture that the integral equation for the three-dimensional problem has no irregular frequencies provided that the original boundary value problem is uniquely solvable.

The present paper provides a proof of the conjecture that an integral equation for a two-dimensional floating-body problem having no irregular frequencies is available. Making use of a proper Green's function, we arrive at an integral equation which has the same form as the one derived by Angell, Hsiao, and Kleinman. However, the proof of the uniqueness theorem is quite different. The unique solvability of the original boundary value problem is also required.

*Received by the editors March 27, 1989; accepted for publication (in revised form) June 11, 1990.

†Mathematics Department, Tennessee Technological University, Cookeville, Tennessee 38505. This work is contained in the author's doctoral dissertation at the University of Delaware.

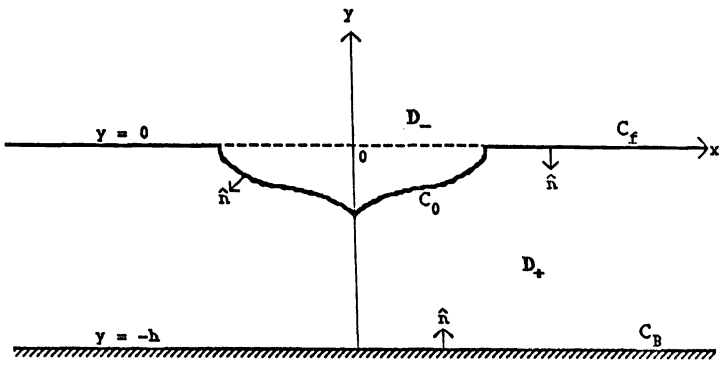


FIG. 1

2. Notation and statement of problem. The geometry of the two-dimensional floating-body problem with finite depth h is described as follows and is illustrated in Fig. 1. We denoted the fluid domain by $\mathbf{D}_+$, whose boundary consists of C_0, the wetted portion of the floating body, the free surface C_f, and the bottom C_B, and we denote by $\mathbf{D}_-$ the domain consisting of the upper half space and the interior of the floating body.

The function ϕ solves the floating-body problem if

$$\begin{aligned} &\nabla^2 \phi = 0 \quad \text{in } \mathbf{D}_+, \quad \frac{\partial \phi}{\partial n} = V \quad \text{on } C_0, \quad \frac{\partial \phi}{\partial n} = 0 \quad \text{on } C_B, \\ &\frac{\partial \phi}{\partial n} + k\phi = 0 \quad \text{on } C_f, \end{aligned} \tag{1}$$

and provided ϕ satisfies a radiation condition. Here $\partial/\partial n$ denotes the normal derivative pointing into $\mathbf{D}_+$ and V is a given function. The radiation condition is specified in the form

$$\frac{\partial \phi}{\partial |x|} - ik_0 = o(1) \quad \text{as } |x| \to \infty, \tag{2}$$

where k_0 is the root with largest real part of the transcendental equation

$$k = k_0 \tan h(k_o h). \tag{3}$$

Condition (2) may be shown to guarantee that

$$\phi(x, y) = e^{ik_0|x|} \cdot a(y) + O\left(e^{-\mu|x|}\right) \quad \text{as } |x| \to \infty, \tag{4}$$

uniformly in y, for some complex-valued function $a(y)$, (x, y) being rectangular coordinates and μ is a positive constant.

We now define the Green's function

$$\gamma(p, q) = \frac{1}{\pi} \left\{ \log \frac{|q|}{|p - q|} + \log \frac{|q_1|}{|p - q_1|} \right\}, \tag{5}$$

where $p = (x_p, y_p)$, $q = (x_q, y_q)$ and $q = (x_q, -2h - y_q)$. With this Green's function, Green's theorem for a solution of Laplace's equation in $\mathbf{D}_+$ which satisfies the radiation condition (2) then takes the form

$$\int_{C_0 \cup C_f \cup C_B} \left\{ \gamma(p, q) \frac{\partial \phi(q)}{\partial n_q} - \phi(q) \frac{\partial \gamma(p, q)}{\partial n_q} \right\} dS_q = \alpha(p)\phi(p), \tag{6}$$

with

$$\alpha(p) := \lim_{\varepsilon \to 0} \int_{[\partial B_\varepsilon(p)] \cap \mathbf{D}_+} \frac{\partial \gamma(p,q)}{\partial n_q} dS_q, \tag{7}$$

where $\partial B_\varepsilon(p)$ denotes the boundary of a disc of radius ε centered at p. If ϕ satisfies all the boundary conditions in (1) we arrive at the boundary integral equation

$$\begin{aligned} \alpha(p)\phi(p) + \int_{C_0} \phi(q) \frac{\partial \gamma}{\partial n_q}(p,q) dS_q + \int_{C_f} \phi(q) \left[\frac{\partial \gamma}{\partial n_q}(p,q) + k\gamma(p,q) \right] dS_q \\ = \int_{C_0} \gamma(p,q) V(q) dS_q, \end{aligned} \tag{8}$$

where p lies either on C_0 or C_f. The integral on C_B vanishes since both γ and ϕ satisfy the homogeneous Neumann condition there and the integrals over large enough artificial bounds $x = R$ and $x = -R$ can be shown to be of order $O(R^{-1})$ (see Liu [6]). This equation has irregular frequencies if there are real values of k for which the corresponding homogeneous equation ($V = 0$) has nontrivial solutions. We shall prove in the following that such irregular frequencies do not exist.

3. Uniqueness. Our result can be summarized as follows.

THEOREM. *If*

(a) $\phi(p) = e^{ik_0|x_p|} \cdot a(y_p) + O(e^{-\mu|x_p|})$ *as* $|p| \to \infty$, *uniformly in* y_p, $\mu > 0$;

(b) $\alpha(p)\phi(p) + \int_{C_0} \phi(q) \frac{\partial \gamma}{\partial n_q}(p,q)\, dS_q + \int_{C_f} \phi(q) [\frac{\partial \gamma}{\partial n_q}(p,q) + k\gamma(p,q)]\, dS_q = 0$ *for all* $p \in C_0 \cup C_f$; *and*

(c) ϕ *is continuous on* $C_0 \cup C_f$,

then $\phi(p) = 0$ *for all* $p \in C_0 \cup C_f$.

Proof. Assume that ϕ is a function satisfying (a), (b), and (c) of the theorem and define the functions u_+ and u_- in $\mathbf{D}_+$ and $\mathbf{D}_-$, respectively, as

$$\begin{aligned} \left.\begin{matrix} u_+ \\ u_- \end{matrix}\right\} := \int_{C_0} \phi(q) \frac{\partial \gamma}{\partial n_q}(p,q) dS_q \\ + \int_{C_f} \phi(q) \left[\frac{\partial \gamma}{\partial n_q}(p,q) + k\gamma(p,q) \right] dS_q, \qquad p \in \left\{ \begin{matrix} \mathbf{D}_+, \\ \mathbf{D}_-. \end{matrix} \right. \end{aligned} \tag{9}$$

It follows that

$$\nabla^2 u_\pm = 0, \qquad p \in \mathbf{D}_\pm, \tag{10}$$

since $p \neq q$ in either case. This property is inherited from the function $\gamma(p,q)$. Using the jump conditions for double layer potentials defined on $C_0 \cup C_f$, we have

$$\lim_{\substack{p \to C_0 \cup C_f \\ p \in \mathbf{D}_\pm}} \int_{C_0 \cup C_f} \phi(q) \frac{\partial \gamma}{\partial n_q}(p,q) dS_q = \left.\begin{matrix} (\alpha(p) - 2) \\ \alpha(p) \end{matrix}\right\} \phi(p) + \int_{C_0 \cup C_f} \phi(q) \frac{\partial \gamma}{\partial n_q}(p,q) dS_q, \tag{11}$$

$$p \in C_0 \cup C_f.$$

Together with the continuity of single layer potenials, this implies that

$$\lim_{\substack{p\to C_0\cup C_f\\ p\in\mathbf{D}_-}} u_-(p) = \alpha(p)\phi(p) + \int_{C_0} \phi(q)\frac{\partial\gamma}{\partial n_q}(p,q)\,dS_q + \int_{C_f} \phi(q)\left[\frac{\partial\gamma}{\partial n_q}(p,q) + k\gamma(p,q)\right] dS_q, \tag{12}$$

and hence

$$\lim_{p\to C_0\cup C_f} u_-(p) = 0, \qquad p\in\mathbf{D}_- \tag{13}$$

in view of (b). As is established in Appendix A, the growth of the function u_- in $\mathbf{D}_-$ is

$$u_-(p) = O(\log r_p) + O(1) + O(r_p^{-1}) \quad \text{as } r_p = |p| \to \infty, \tag{14}$$

for $p\in\mathbf{D}_-$. This together with (13) implies that

$$u_-(p) \equiv 0, \qquad p\in\mathbf{D}_- \tag{15}$$

by the maximum principle (see Appendix B). Consequently, we have

$$\frac{\partial u_-}{\partial n_-} = 0 \quad \text{on } C_0\cup C_f, \tag{16}$$

where $\partial/\partial n_-$ indicates the normal derivative from $\mathbf{D}_-$. From the definition of u_- in (9), it follows that

$$\frac{\partial}{\partial n_p}\int_{C_0\cup C_f} \phi(q)\frac{\partial\gamma}{\partial n_q}(p,q,)\,dS_q + k\int_{C_f}\phi(q)\frac{\partial\gamma}{\partial n_p}(p,q)\,dS_q + \beta(p)\phi(p) = 0, \tag{17}$$

with

$$\beta(p) = \begin{cases} 0, & p\in C_o,\\ -k, & p\in C_f, \end{cases} \tag{18}$$

where the jump condition of simple layer is employed. Note that the existence of the normal derivative of a double layer is not guaranteed for a merely continous density ϕ. But using the fact that $u_-\equiv 0$ in $\mathbf{D}_-$ and hence u_- has a ordinary normal derivative, $\partial u_-/\partial n_-\equiv 0$ on $C_0\cup C_f$, together with the fact that the single layers have ordinary normal derivatives, we conclude from definition (9) that

$$\frac{\partial}{\partial n_p}\int_{C_0\cup C_f}\phi(q)\frac{\partial\gamma}{\partial n_q}(p,q)\,dS_q$$

exists in the ordinary sense.

For properties of u_+, we find that

$$u_+(p) = (\alpha(p)-2)\phi(p) + \int_{C_0}\phi(q)\frac{\partial\gamma}{\partial n_q}(p,q)\,dS_q + \int_{C_f}\phi(q)\left[\frac{\partial\gamma}{\partial n_q}(p,q) + k\gamma(p,q)\right] dS_q, \qquad p\in C_0\cup C_f, \tag{19}$$

from the usual jump conditions. Consequently, using condition (b) we have that

$$u_+(p) = -2\phi(p), \qquad p \in C_0 \cup C_f. \tag{20}$$

Equation (20) means that $u_+(p)$ has the same growth as $\phi(p)$ on C_f, which is specified by (a). Taking the normal derivative yields

$$\begin{aligned}\frac{\partial u_+}{\partial n_+}(p) = \frac{\partial}{\partial n_p} \int_{C_0 \cup C_f} \phi(q) \frac{\partial \gamma}{\partial n_q}(p,q) dS_q \\ + k \int_{C_f} \phi(q) \frac{\partial \gamma}{\partial n_p}(p,q) dS_q - \beta(p)\phi(p).\end{aligned} \tag{21}$$

Since the normal derivatives of the double layer potential with continuous density are the same from either side provided that one of them exists, we obtain

$$\frac{\partial u_+}{\partial n_+} = -2\beta(p)\phi(p) = \beta(p)u_+(p) \tag{22}$$

from (17) and (20), or equivalently

$$\frac{\partial u_+}{\partial n_+} = 0, \qquad p \in C_0, \tag{23}$$

and

$$\frac{\partial u_+}{\partial n_+} = -ku_+(p), \qquad p \in C_f. \tag{24}$$

The function u_+ also satisfies

$$\frac{\partial u_+}{\partial n_+} = 0, \qquad p \in C_B, \tag{25}$$

which is inherited from $\gamma(p,q)$. Thus equations (10), (24), and (25) imply that u_+ satifies Laplace's equation in $\mathbf{D}_+$, together with the homogeneous Neumann and Robin condition on C_B and C_f, respectively. Following Weinstein [8] and Kreisel [5], u_+ has the representation

$$u_+(x,y) = \sum_{n=0}^{\infty} a_n(x) \cosh\left[k_n(y+h)\right], \qquad |x| \geq A, \tag{26}$$

where k_n's are the roots of the transcendental equation (3), and A is any number greater than the diameter of the wetted portion of the floating body. That is, $A > \max |x_p|, p \in C_0$. Hence the Fourier coefficients $a_n(x)$ in (26) are of the form

$$a_n(x) = C_n \int_{-h}^{0} u_+(x,y) \cosh\left[k_n(y+h)\right] dy \tag{27}$$

for some constant C_n, and they all have the same growth as $u_+(x,y)$ when $|x| \geq A$, i.e.,

$$a_n(x) = O(\log|x|) + O(1) + O(|x|^{-1}) \tag{28}$$

(see Appendix A). Recall that the most general form of $a_n(x)$ is

$$a_n(x) = C_{n,1} e^{ik_n|x|} + C_{n,2} e^{-ik_n|x|} \tag{29}$$

since $u_+(x,y)$ satisfies Laplace's equation. The fact that k_n has positive imaginary part for $n \geq 1$ makes the term $\exp(-ik_n|x|)$ grow exponentially. Computing $a_n(x)\exp(ik_n|x|)$ from both (27) and (28), we see that

$$e^{ik_n|x|}\{O(\log|x|) + O(1) + O(|x|^{-1})\} = C_{n,1}e^{i2k_n|x|} + C_{n,2}.$$

Hence $C_{n,2}$, being the only term that survives as $|x| \to \infty$, must be zero for $n \geq 1$. We may then rewrite the expression (26) as

$$u_+(x,y) = \sum_{n=0}^{\infty} C_{n,1}e^{ik_n|x|}\cosh[k_n(y+h)] + C_{n,2}e^{-ik_n|x|}\cosh[k_0(y+h)]. \tag{30}$$

Because $u_+(x,0)$ has the same asymptotic growth as ϕ, i.e., $0(\exp(ik_0|x|))$, we conclude that $C_{0,2} = 0$. Then the representation (30) can be further simplified as

$$u_+(x,y) = \sum_{n=0}^{\infty} C_n e^{ik_n|x|}\cosh[k_n(y+h)] \quad \text{for } |x| \geq A. \tag{31}$$

With this representation, it is obvious that u_+ satisfies the radiation condition (2). That is

$$\frac{\partial u_+}{\partial|x|} - ik_0u_+ = o(1) \quad \text{as} |x| \to \infty, \tag{32}$$

for $-h \leq y \leq 0$.

Now the uniqueness theorem for the two-dimensional floating-body problem (Liu [6]) implies that $u_+ \equiv 0$ in $\mathbf{D}_+$ and hence on $C_0 \cup C_f$, provided that C_0 satisfies the geometric restriction. Equation (24) then ensures that $\phi(p) = 0$ on $C_0 \cup C_f$. It means that the only solution of (a), (b), and (c) is $\phi \equiv 0$ and the integral equation (8) has no irregular frequencies. This completes the proof.

Appendix A. On the growth of $u\pm$. Here we establish the lemma regarding the growth of the functions u_+ and u_- defined in the uniqueness theorem. It is restated as follows.

LEMMA. *If*

(a) $\phi(p) = e^{ik_0|x_p|}.a(y) + O(e^{-\mu|x_p|})$ *as* $|p| \to \infty$ *uniformly in* y_p, $\mu > 0$;

(b) $$\alpha(p)\phi(p) + \int_{C_0} \phi(q)\frac{\partial\gamma}{\partial n_q}(p,q)\,dS_q + \int_{C_f} \phi(q)\left\{\frac{\partial\gamma}{\partial n_q}(p,q) + k\gamma(p,q)\right\}dS_q = 0$$
$$\forall p \in C_0 \cup C_f;$$

(c) ϕ *is continuous on* $C_0 \cup C_f$; *and*

(d) $$u_\pm(p) = \int_{C_0} \phi(q)\frac{\partial\gamma}{\partial n_q}(p,q)\,dS_q + \int_{C_f} \phi(q)\left\{\frac{\partial\gamma}{\partial n_q}(p,q) + k\gamma(p,q)\right\}dS_q,$$
$$p \in \mathbf{D}_\pm,$$

then $u_-(p) = O(\log r_p) + O(1) + O(r_p^{-1})$, *and* $u_+(p) = O(\log|x_p|) + O(1) + O(|x_p|^{-1})$, *as* $r_p \to \infty$, *where* $r_p = |p| = (x_p^2 + y_p^2)^{1/2}$.

Proof. With the Green's function $\gamma(p,q)$ defined in (5), we have

$$\gamma(p,q) = \frac{1}{2\pi}\left\{\log\frac{(x_q^2 + y_q^2)}{[(x_p - x_q)^2 + (y_p - y_q)^2]} + \log\frac{[x_q^2 + (2h + y_q)^2]}{[(y_p - x_q)^2 + (y_p + 2h + y_q)^2]}\right\} \tag{A.1}$$

and

$$\frac{\partial\gamma}{\partial y_q}(p,q) = \frac{1}{\pi}\left\{\frac{y_p - y_q}{(x_p - x_q)^2 + (y_p - y_q)^2} + \frac{y_q}{x_q^2 + y_q^2} - \frac{y_p + 2h + y_q}{(x_p - x_q)^2 + (y_p + 2h + y_q)^2} + \frac{2h + y_q}{x_q^2 + (2h + y_q)^2}\right\}; \tag{A.2}$$

hence

$$\int_{C_0} \phi(q)\frac{\partial\gamma}{\partial n_q}(p,q)dS_q = O(r_p^{-1}), \tag{A.3}$$

$$\int_{C_f \cap B_A} \phi(q)\frac{\partial\gamma}{\partial n_q}(p,q)dS_q = O(r_p^{-1}), \tag{A.4}$$

and

$$\int_{C_f \cap B_A} \phi(q)k\gamma(p,q)dS_q = O(\log r_p) + O(r_p^{-1}), \tag{A.5}$$

where $C_f \cap B_A$ is that portion on the free surface contained in the disc B_A centered at the origin with radius A, $A > \max\{|x_p|, p \in C_0\}$. We also have to establish the growth of

$$\int_{C_f \cap B_A^C} \phi(q)\left[\frac{\partial\gamma}{\partial n_q}(p,q) + k\gamma(p,q)\right]dS_q \tag{A.6}$$

with B_A^C being the complement of B_A. To examine the growth of the first piece in (A.6), it suffices to establish the growth of the integral

$$-\frac{1}{\pi}\int_{A'}^{\infty}\left[a(0)e^{ik_0x_q} + O(e^{-\mu|x_q|})\right]\left\{\frac{y_p}{(x_p - x_q)^2 + y_p^2} + \frac{y_p + 2h}{x_q^2 + (y_p + 2h)^2} - \frac{y_p + 2h}{(x_p - x_q)^2 + (y_p + 2h)^2}\right\}dx_q,$$

where the asymptotic behavior of ϕ has been employed, A' being a fixed number sufficiently large. Integrating by parts shows that its growth is of $O(r_p^{-1})$. Hence we conclude that

$$\int_{C_f \cap B_A^C} \phi(q)\frac{\partial\gamma}{\partial n_q}(p,q)dS_q = 0(r_p^{-1}) \quad \text{as } r_p \to \infty. \tag{A.7}$$

Next we consider the second piece in (A.6). As in the previous estimate, it suffices to examine the growth of

$$\int_{A'}^{\infty}\left[a(0)e^{ik_0x_q} + O(e^{-\mu|x_q|})\right]\log\frac{x_q^2}{(x_p - x_q)^2 + y_p^2}dx_q.$$

Integrating by parts twice, we see that

$$\int_{A'}^{\infty} e^{ik_o x_q} \log \frac{x_q^2}{(x_p - x_q)^2 + y_p^2} dx_q$$

$$= -\frac{e^{ik_o A'}}{ik_o} \log \frac{A'^2}{(x_p - A')^2 + y_p^2} - \frac{2}{ik_o} \int_{A'}^{\infty} e^{ik_o x_q} \left[\frac{1}{x_q} + \frac{x_p - x_q}{(x_p - x_q)^2 + y_p^2} \right] dx_q$$

$$= -\frac{e^{ik_o A'}}{ik_o} \log \frac{A'^2}{(x_p - A')^2 + y_p^2} - \frac{2}{(ik_o)^2} e^{ik_o A'} \left[\frac{1}{A'} + \frac{x_p - A'}{(x_p - A')^2 + y_p^2} \right]$$

$$+ \frac{2}{(ik_o)^2} \int_{A'}^{\infty} e^{ik_o x_q} \left[\frac{-1}{x_q^2} + \frac{(x_p - x_q)^2 + y_p^2}{[(x_p - x_q)^2 + y_p^2]^2} \right] dx_q.$$

Hence

$$\text{(A.8)} \qquad \int_{C_f \cap B_A^C} k\phi(q)\gamma(p,q) dS_q = O(\log r_p) + O(1) + O(r_p^{-1}) \quad \text{as } r_p \to \infty.$$

It follows from (A.3)–(A.5), (A.7), and (A.8) that

$$\text{(A.9)} \qquad u_-(p) = O(\log r_p) + O(1) + O(r_p^{-1}) \quad \text{as } r_p \to \infty, \quad p \in \mathbf{D}_-,$$

which is the required growth for the function $u_-(p)$ in $\mathbf{D}_-$. There are only two places where the estimate of u_+ is different from that of u_-.The first is that for every p in $\mathbf{D}_+$, y_p is always bounded, which results in the fact that $r_p \to \infty$ implies only $|x_p| \to \infty$. The second is that the normal derivatives on $C_0 \cup C_f$ have opposite sign to those of $u_-(p)$, but this will not affect any growth order. Thus we may arrive at

$$\text{(A.3)}' \qquad \int_{C_0} \phi(q) \frac{\partial \gamma}{\partial n_q}(p,q) dS_q = O(|x_p|^{-1}),$$

$$\text{(A.4)}' \qquad \int_{C_f \cap B_A} \phi(q) \frac{\partial \gamma}{\partial n_q}(p,q) dS_q = O(|x_p|^{-1}),$$

$$\text{(A.5)}' \qquad \int_{C_f \cap B_A} \phi(q) k\gamma(p,q) dS_q = O(\log |x_p|) + O(|x_p|^{-1}),$$

and

$$\text{(A.8)}' \qquad \int_{C_f \cap B_A^C} \phi(q) \left[\frac{\partial \gamma}{\partial n_q}(p,q) + k\gamma(p,q) \right] dS_q = O(\log |x_p|) + O(1) + O(|x_p|^{-1}).$$

Then it follows that

$$\text{(A.10)} \qquad u_+(p) = O(\log |x_p|) + O(1) + O(|x_p|^{-1}) \quad \text{as } r_p \to \infty, \quad p \in \mathbf{D}_+,$$

and the proof is complete.

Appendix B. Maximum principle on u_-. We wish to prove that the function $u_-(p)$ defined in (9) is a zero function in $\mathbf{D}_-$.

Proof. From (13) and (14) we know that $u_- = 0$ on $C_0 \cup C_f$ and that $u_-(p) = O(\log r_p)$ as $r_p \to \infty$ in $\mathbf{D}_-$. Choose r_0 large enough such that C_0 lies completely inside the circle $r = r_0$. Consider the potential u_- in the domain $\mathbf{D}_- \cap \{(r,\theta)|r > r_0\}$. Since $u_- = 0$ on C_f it follows that u_- can be continued, by Schwarz's symmetry principle, into the domain $\mathbf{D}_+ \cap \{(r,\theta)|r > r_0\}$ by the relation $u_-(x,y) = -u_-(x,-y)$. Because of the fact that $u_-(x,y)$ is an odd function of y as $r > r_0$, there is an expansion

$$u_-(r\cos\theta, r\sin\theta) = \sum_{m=1}^{\infty} b_m r^m \sin m\theta + \sum_{m=1}^{\infty} b_{-m} r^{-m} \sin m\theta$$

for $r > r_0$, $-\pi \le \theta \le \pi$, where the coefficients $\{b_m\}_{m=1}^{\infty}$ and $\{b_{-m}\}_{m=1}^{\infty}$ satisfy

$$b_m r^m - b_{-m} r^{-m} = \frac{1}{\pi} \int_{-\pi}^{\pi} u_-(r\cos\theta, r\sin\theta) \sin m\theta d\theta.$$

Let $r \to \infty$. We find that (14) requires $b_m = 0$ for $m \ge 1$, and hence

$$u_-(r\cos\theta, r\sin\theta) = \sum_{m=1}^{\infty} b_{-m} r^{-m} \sin m\theta.$$

We then have

$$|u_-(r\cos\theta, r\sin\theta)| < M/r$$

for r sufficiently large, say $r > 2r_0$.

Now we consider u_- in a bounded domain $\mathbf{D}_- \cap \{(r,\theta)|r < R\}$ where $R > 2r_0$. On the boundary we have

$$|u_-(R\cos\theta, R\sin\theta)| < M/R$$

when $-\pi < \theta < \pi$, while $u_- = 0$ on $C_0 \cup C_f$. Thus, for $p \in \mathbf{D}_-$ and $|p| < R$, we have

$$|u_-(p)| < M/R,$$

by the maximum principle. Let $R \to \infty$, keeping p fixed. We obtain $|u_-(p)| \le 0$, i.e., $u_-(p) \equiv 0$ in $\mathbf{D}_-$. This is the required result.

Acknowledgments. The author is indebted to his supervisors Professors G. C. Hsiao and R. E. Kleinman for their guidance, criticism, and suggestions.

REFERENCES

[1] T. S. ANGELL, G. C. HSIAO, AND R. E. KLEINMAN, *An integral equation for the floating-body problem*, J. Fluid Mech., 166 (1986), pp. 161–171.

[2] K. J. BAI AND R. W. YEUNG, *Numerical solutions to free surface flow problems*, 10th Symposium on Naval Hydrodynamics, Massachusetts Institute of Technology, Cambridge, MA, 1974, pp. 609–647.

[3] G. C. HSIAO, R. E. KLEINMAN, AND Y. W. LIU, *A boundary element method for the floating-body problem*, in Boundary Elements, Proc. Internat. Conference. Beijing, China, Pergamon Press, New York, 1986, pp. 199–206.

[4] F. JOHN, *On the motion of floating bodies:* II, Comm. Pure Appl. Math., 3 (1950), pp. 45–101.

[5] G. KREISEL, *Surface waves*, Quart. Appl. Math., 7 (1949), pp. 21–44.

[6] Y.W. LIU, *A boundary integral method for the two-dimensional floating body problem*, Ph.D. dissertation, University of Delaware, Newark, DE, 1987.

[7] F. URSELL, *Irregular frequencies and the motion of floating bodies*, J. Fluid Mech., 105 (1981), pp. 143–156.

[8] A. WEINSTEIN, *Sur un problème aux limites dans une bande indéfinie*, Comptes rendus de l'Academie des Sciences, 184 (1927), p. 497.

[9] R. W. YEUNG, *A hybrid integral equation method for time harmonic free surface flows*, Proc. First Int. Conf. on Numerical Ship Hydrodynamics, Gaithersburg, MD, 1975, pp. 581–607.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 982-990, July 1991

007

A PRESCRIBED MEAN CURVATURE PROBLEM ON DOMAINS WITHOUT RADIAL SYMMETRY*

CHARLES V. COFFMAN† AND WILLIAM K. ZIEMER‡

Abstract. The existence is proved of a nontrivial solution to the problem

$$\operatorname{div}(\nabla u/\sqrt{1+|\nabla u|^2}) - \mu u + \lambda u^q = 0 \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega,$$

on a smooth, bounded but not necessarily radially symmetric domain Ω when λ is sufficiently large.

Key words. prescribed mean curvature problem, elliptic boundary value problem

AMS(MOS) subject classifications. 35J20, 35J25, 35J60

1. Introduction. This note is concerned with the existence of positive solutions to the Dirichlet problem for the so-called "prescribed mean curvature equation"

$$\operatorname{div}(\nabla u/\sqrt{1+|\nabla u|^2}) - \mu u + \lambda u^q = 0 \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{1}$$

on a bounded $C^{1,1}$ domain $\Omega \subseteq R^N$, $N \geqq 2$. Here $\mu \geqq 0$; $1 < q < (N+2)/(N-2)$, and λ is a real parameter whose value will be assigned in advance. We stress that the domain Ω is not assumed to be radially symmetric. Our main result is that when μ, q, and Ω are fixed as above then (1) has a smooth positive solution for all sufficiently large λ.

The problem (1) has been treated in a number of recent works. The papers [1], [4], [5], [9] deal with radially symmetric ground states (positive symmetric solutions on R^N which tend to zero at ∞) in the case $\mu > 0$; it is shown in [4] that positive solutions vanishing at ∞ must be radially symmetric.

Radial symmetry of positive solutions on a ball is proved in [6]; the case when Ω is a ball is also treated in [8] and [11]. When $\mu = 0$ and Ω is a ball, Theorem 3.4 of [8] implies nonexistence of nontrivial solutions to (1) for supercritical q, i.e., for $q \geqq (N+2)/(N-2)$; a similar nonexistence result is given in [10] for $\mu > 0$ and Ω star-shaped. In [11] it is shown that when q is subcritical and Ω is a ball, then, independently of the radius of Ω, there is no nontrivial solution to (1) if

$$\mu > (2(q+1)/(q-1))^{(q-1)/(q+1)} \lambda^{2/(q+1)}.$$

Nonexistence is also proved in [11] for $\lambda = 1$ and $\mu > 0$ when the radius of Ω is too small; an existence result is given for the case where μ is sufficiently small and the radius of Ω sufficiently large.

It is instructive to compare (1) with the analogous semilinear problem in which the mean-curvature operator is replaced by the Laplacian

$$\Delta u - \mu u + \lambda u^q = 0 \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega. \tag{2}$$

The nonexistence results quoted above for the case of supercritical q of course apply also to (2), while the nonexistence results quoted from [11] reflect features special to the quasilinear problem (1). In particular, in the case of (2), unlike that of (1), we can limit attention to the case where $\lambda = 1$ and $\mu = 0$ or 1.

* Received by the editors March 22, 1990; accepted for publication July 6, 1990.

† Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. The research of this author was supported by National Science Foundation grant DMS 870-4530.

‡ Department of Mathematics, California State University at Long Beach, Long Beach, California 90840.

Our approach to (1) makes use of variational methods which generalize those that were introduced by Nehari [7] to treat boundary value problems for nonlinear ordinary differential equations; the application of these methods to (2) is described in [2]; see also [3].

We seek solutions to (1) as minima of the functional

$$H_\lambda(u)=\int_\Omega [2\sqrt{1+|\nabla u|^2}-2+\mu u^2-2\lambda(q+1)^{-1}|u|^{q+1}]\,dx \tag{3}$$

over nonzero functions subject to the constraint

$$N_\lambda(u)=\int_\Omega [|\nabla u|^2/\sqrt{1+|\nabla u|^2}+\mu u^2-\lambda|u|^{q+1}]\,dx=0. \tag{4}$$

We first note that (4) is necessarily satisfied by a solution to (1). Secondly, on a ray in $W_0^{1,2}(\Omega)$ emanating from the origin, $H_\lambda(u)$ achieves its maximum at the unique nonzero u that satisfies (4). Thus minimizing (3) subject to (4) is equivalent to maximizing (3) on rays from the origin and then finding a u that gives the minimum of these maxima.

The feature of (1) that enables one to extend Nehari's method is that the integrand on the right in (3) is concave in $(u^2, |\nabla u|^2)$. As a consequence, if $u\in W_0^{1,2}(\Omega)\backslash\{0\}$ with $N_\lambda(u)=0$ and we solve the linear boundary value problem

$$\operatorname{div}(\nabla v/\sqrt{1+|\nabla u|^2})-\mu v+\alpha u^q=0 \quad \text{in } \Omega, \qquad v=0 \quad \text{on } \partial\Omega, \tag{5}$$

adjusting $\alpha>0$ so that $N_\lambda(v)=0$, then the mapping

$$T_\lambda: u\to v \tag{6}$$

reduces H_λ. That is $H_\lambda(v)\leqq H_\lambda(u)$ and equality holds only if u and v are proportional; if the latter is the case then u satisfies (1) (μ and λ are assumed to be fixed throughout).

Several difficulties are encountered when one attempts to implement the method suggested above. First, the term

$$\int_\Omega (\sqrt{1+|\nabla|^2}-1)\,dx$$

(see (3)) is not coercive on $W_0^{1,p}(\Omega)$ for any $p>1$. Second, in contrast to the case of the semilinear analogue (2) (cf. [2]), the set

$$\{u\in W_0^{1,2}(\Omega): u\neq 0,\ N_\lambda(u)=0\}$$

is not bounded away from zero. Thus, in fact, the problem (3), (4), as formulated above, does not have a nontrivial global minimizer. We seek, therefore, to minimize H_λ under additional constraints. Assuming $\mu\geqq 0$ to be fixed, we construct a set

$$S(\lambda)\subseteq W^{2,r}(\Omega)\cap\{u: u\in W_0^{1,2}(\Omega)\backslash\{0\},\ N_\lambda(u)=0\} \qquad (r>N)$$

that is bounded in $W^{2,r}(\Omega)$, and such that when λ is sufficiently large we have the following: (i) $S(\lambda)$ is nonempty; (ii) H_λ has a positive lower bound on $S(\lambda)$; (iii) for $u\in S(\lambda)$, problem (5) is solvable and (iv) $S(\lambda)$ is invariant under the mapping T_λ or some iterate thereof. When λ is such that these latter four conditions hold, then H_λ can be minimized over $S(\lambda)$ and the resulting minimizer is a solution to (1).

2. Preliminaries. We shall actually consider a more general problem than (1), namely,

$$\operatorname{div}(\nabla u/\sqrt{1+|\nabla u|^2})-\mu u+\lambda f(u)=0 \quad \text{in } \Omega, \qquad u=0 \quad \text{on } \partial\Omega, \tag{7}$$

where the functional $f \in C[0, \infty)$ is assumed to satisfy the conditions

$$1 < s^{-q} f(s) < C \tag{8}$$

for some q such that

$$1 < q < (N+2)/(N-2), \tag{9}$$

and also to satisfy Nehari's condition, which in our notation takes the following form: for some $\varepsilon > 0$,

$$s^{-1-\varepsilon} f(s) < t^{-1-\varepsilon} f(t) \quad \text{for } 0 < s < t. \tag{10}$$

In place of (3) and (4) the definitions of H and N are

$$H_\lambda(u) = \int_\Omega [2(\sqrt{1+|\nabla u|^2} - 1) + \mu u^2 - 2\lambda F(u)]\, dx \tag{11}$$

and

$$N_\lambda(u) = \int_\Omega [|\nabla u|^2/\sqrt{1+|\nabla u|^2} + \mu u^2 - \lambda u f(u)]\, dx; \tag{12}$$

F in (11) is defined by

$$F(u) = \int_0^u f(s)\, ds. \tag{13}$$

3. Inequalities. Below K will be used to denote a generic constant that can depend on N and Ω but not on f, μ, or λ.

We find readily from (10) that F, defined by (13), satisfies

$$(2+\varepsilon) F(u) \leqq u f(u). \tag{14}$$

We will make frequent use below of the inequalities

$$p \geqq 2(\sqrt{1+p} - 1) \geqq p/\sqrt{1+p}. \tag{15}$$

From (11), (12), and (14) we conclude that for $u \in W_0^{1,2}(\Omega) \backslash \{0\}$ we have

$$H_\lambda(u) \geqq \varepsilon(2+\varepsilon)^{-1} \int_\Omega (2[\sqrt{1+|\nabla u|^2} - 1] + \mu u^2)\, dx, \tag{16}$$

when u is subject to the constraints

$$N_\lambda(u) = 0 \quad \text{and} \quad u \neq 0. \tag{17}$$

Next we observe that

$$2\sqrt{1+\|\nabla u\|_\infty^2} \int_\Omega [\sqrt{1+|\nabla u|^2} - 1]\, dx \geqq \int_\Omega |\nabla u|^2\, dx.$$

Combining this with (16) we have, when (17) holds,

$$\sqrt{1+\|\nabla u\|_\infty^2}\, H_\lambda(u) \geqq \varepsilon(2+\varepsilon)^{-1} \int_\Omega (|\nabla u|^2 + \mu u^2)\, dx. \tag{18}$$

Let $u \in W_0^{1,2}(\Omega) \backslash \{0\}$ satisfy (17) and define the formal operator $L = L_u$ by

$$Lw = -\operatorname{div}(\nabla w/\sqrt{1+|\nabla u|^2}) + \mu w. \tag{19}$$

We note that if $\|\nabla u\|_\infty < \infty$ and $h \in L^2(\Omega)$, then by the Lax-Milgram theorem, the boundary value problem

$$Lw = h \quad \text{in } \Omega, \qquad w = 0 \quad \text{on } \partial\Omega$$

has a unique weak solution in $W_0^{1,2}(\Omega)$; for elementary reasons w will be nonnegative if h is.

LEMMA 3.1. *Let λ be fixed and let $u \in W_0^{1,2}(\Omega)\backslash\{0\}$ and satisfy* (17). *Let $v \in W_0^{1,2}(\Omega)$ be the weak solution to*

$$Lv = \alpha\lambda f(u), \tag{20}$$

for any $\alpha > 0$. Then

$$H_\lambda(v) \leqq H_\lambda(u), \tag{21}$$

with equality only if $N_\lambda(v) = 0$ and u is a solution to (7).

Proof. By an elementary concavity inequality we have

$$H_\lambda(v) - H_\lambda(u) \leqq \int_\Omega [vLv - uLu - \lambda(v^2 - u^2)g(u)]\, dx,$$

where

$$g(0) = 0, \quad g(u) = f(u)/u, \qquad u \neq 0.$$

From (17) and (20) we have

$$H_\lambda(v) - H_\lambda(u) \leqq \lambda \int_\Omega [\alpha uv - v^2]g(u)\, dx. \tag{22}$$

Using (20), the Schwarz inequality, and (17) we get

$$\begin{aligned} \alpha\lambda \int_\Omega u^2 g(u)\, dx = \int_\Omega uLv\, dx &\leqq \left(\int_\Omega uLu\, dx\right)^{1/2} \left(\int vLv\right)^{1/2} \\ &\leqq \left(\lambda \int_\Omega u^2 g(u)\, dx\right)^{1/2} \left(\alpha\lambda \int_\Omega vug(u)\, dx\right)^{1/2}. \end{aligned} \tag{23}$$

We multiply through the last inequality by $(\int_\Omega vug(u)\, dx)^{1/2}$ and use the Schwarz inequality to get

$$\alpha \int_\Omega vug(u)\, dx \leqq \int_\Omega v^2 g(u)\, dx,$$

which, when substituted in (22), gives (21).

From the way in which the Schwarz inequality was used it is clear that for equality we must have v proportional to u; since $H_\lambda(v)$ achieves its maximum when $\alpha(>0)$ is chosen so that $N_\lambda(v) = 0$, the assertion follows.

Henceforth we assume that, in (20), $\alpha > 0$ is chosen so that $N_\lambda(v) = 0$. We seek a bound for this α. Using (17) we have from (23)

$$\alpha \left(\int_\Omega uLu\, dx\right)^{1/2} \leqq \left(\int_\Omega vLv\, dx\right)^{1/2}.$$

We also have

$$\sqrt{1 + \|\nabla u\|_\infty^2} \int_\Omega uLu\, dx \geqq H_\lambda(u),$$

and, since $N_\lambda(v)=0$,

$$\int_\Omega vLv\,dx \leqq \varepsilon^{-1}(2+\varepsilon)\sqrt{1+\|\nabla v\|_\infty^2}\,H_\lambda(v),$$

and thus, in view of (21),

$$\alpha \leqq (\varepsilon^{-1}(2+\varepsilon)\sqrt{1+\|\nabla v\|_\infty^2}\sqrt{1+\|\nabla u\|_\infty^2})^{1/2}. \tag{24}$$

If, in addition to the constraint (17), we also require that

$$\|\nabla u\|_\infty \leqq A,$$

then under these conditions $H_\lambda(u)$ will have a positive infimum. We seek upper and lower estimates for this infimum; these will be expressed as functions of μ and λ and in the case of the lower estimate also of A. Let $w\in C_0^\infty(\Omega)\backslash\{0\}$ be given and let $\beta=\beta(\lambda)>0$ be such that $N_\lambda(\beta w)=0$. Using the definition (12), it follows from (17) and (8) that

$$\int_\Omega (|\nabla w|^2+\mu w^2)\,dx \geqq \lambda\beta^{q-1}\int_\Omega |w|^{q+1}\,dx,$$

from which we conclude that

$$\beta \leqq K(1+\mu)^{1/(q-1)}\lambda^{-1/(q-1)}.$$

Thus we have shown the following.

LEMMA 3.2. *There exists a constant K such that for every $\lambda>0$ there is a $u\in C_0^\infty(\Omega)$ satisfying* (17) *and such that*

$$\|u\|_{1,2} \leqq K(1+\mu)^{1/(q-1)}\lambda^{-1/(q-1)} \tag{25}$$

(*here we are taking* $\|u\|_{1,2}=(\int_\Omega |\nabla u|^2\,dx)^{1/2}$).

From (25) there follows

$$H_\lambda(u) \leqq K^2(1+\mu)^{(q+1)/(q-1)}\lambda^{-2/(q-1)}; \tag{26}$$

this estimate then holds for the infimum of H_λ subject to (17).

Finally, if u satisfies (17) then we have

$$\begin{aligned}(\sqrt{1+\|\nabla u\|_\infty^2})^{-1}\int_\Omega (|\nabla u|^2+\mu u^2)\,dx &\leqq C\lambda\int_\Omega |u|^{q+1}\,dx\\ &\leqq KC\lambda\left(\int (|\nabla u|^2+\mu u^2)\,dx\right)^{(q+1)/2},\end{aligned}$$

from which we have

$$(1+\|\nabla u\|_\infty^2)^{1/(q-1)}\int_\Omega (|\nabla u|^2+\mu u^2)\,dx \geqq KC^{-2/(q-1)}\lambda^{-2/(q-1)}.$$

Thus we have the following.

LEMMA 3.3. *If $m_\lambda(A)$ denotes the infimum of H_λ subject to* (17) *and*

$$\|\nabla u\|_\infty \leqq A,$$

then

$$m_\lambda(\infty) \leqq K^2(1+\mu)^{(q+1)/(q-1)}\lambda^{-2/(q-1)}$$

and

$$m_\lambda(A) \geqq K\varepsilon(2+\varepsilon)^{-1}(1+A^2)^{(1-q)/2(1+q)}C^{-2/(q-1)}\lambda^{-2/(q-1)}.$$

LEMMA 3.4. *Let Ω be a region with $C^{1,1}$ boundary, let r: $N<r<\infty$ be given, and let*

$$1<s_0<r.$$

Then there is a constant $\delta>0$ and a constant M such that for

$$s_0 \leqq s \leqq r$$

if $u \in W^{2,r}(\Omega)$ and

$$\|u\|_{2,r}<\delta,$$

then the operator L_u defined by (19) *acts from $W^{2,s}(\Omega)\cap W_0^{1,2}(\Omega)$ to $L^s(\Omega)$ and is invertible with inverse*

$$L_u^{-1}: L^s(\Omega)\to W^{2,s}(\Omega)$$

satisfying the norm estimate

$$\|L_u^{-1}\|<M.$$

4. Determination of $S(\lambda)$. The set $S(\lambda)$ will be taken to be of the form

$$S(\lambda)=\{u:\ u\in W^{2,r}(\Omega)\cap W_0^{1,2}(\Omega),\ N_\lambda(u)=0,\ u\geqq 0,\ \text{a.e. in } \Omega,$$
$$0<\|u\|_{2,r}\leqq m_1\lambda^{-1/(q-1)},\ H_\lambda(u)\leqq m_2^2\lambda^{-2/(q-1)}\},$$

where r and the constant m_1 (which depends on μ, C, N, Ω, and the choice of r) are to be determined; m_2 is chosen in accordance with (26), i.e.,

$$m_2=K(1+\mu)^{(q+1)/2(q-1)}, \tag{27}$$

with K as in Lemma 3.2. Note that with the exception of the inequality on the $(2, r)$-norm all of the conditions in the definition are preserved by the operator T_λ.

We first determine the order of the iterate of the operator T_λ (defined by (5), (6); see also Lemma 3.1) that will leave $S(\lambda)$ invariant. What we actually seek is the number n such that, if we begin with $u\in W_0^{1,2}$, then after n iterations the resulting function will belong to $W^{2,r}(\Omega)$ with $r>N$. If $N=2$ then $r>2$ is chosen arbitrarily and we take $s_0=r$, $n=1$. If $N>2$ then we define inductively a finite sequence $s_0, s_1, \cdots, s_{n-1}$ by putting

$$s_0=p^*/q, \qquad p^*=2N/N-2,$$
$$s_{k+1}=Ns_k/q(N-2s_k), \qquad 0<k<n$$

(note that this sequence is strictly increasing only if (9) holds). We take n to be the least natural number such that

$$N/2<s_{n-2} \quad \text{or} \quad N<s_{n-1}<\infty,$$

we choose a finite $r>N$ arbitrarily in the former case and we take $r=s_{n-1}$ in the latter case. We have $n=1$ only if $N=2$ or $N=3$ and $q<2$.

Let $\lambda>0$ be fixed and let $u=u_0$ be an arbitrary element of $S(\lambda)$. Let the finite sequence

$$u_0, u_1, \cdots, u_n$$

be determined as follows: u_0 is as above and

$$Lu_{k+1}=\alpha_k\lambda f(u_k) \qquad (L=L_v,\ v=u_k),\quad 0<k<n, \tag{28}$$

where α_k is chosen so that $N(u_{k+1})=0$. We want to estimate the norms of the u_k simultaneously in $W^{2,s}(\Omega)$ $(s=s_{k-1})$ and $W^{2,r}(\Omega)$; more specifically, we shall derive estimates of the form

$$\|u_k\|_{2,s} \leqq \tau_k \lambda^{-1/(q-1)} \qquad (s=s_{k-1}), \tag{29}$$

$$\|u_k\|_{2,r} \leqq \tau_k \lambda^{-1/(q-1)}, \tag{30}$$

$k=1, 2, \cdots, n$. *In the course of this derivation we shall assume*

$$\|u_k\|_{2,r} < \delta, \qquad 0 \leqq k \leqq n, \tag{31}$$

where, with s_0 and r chosen as above, δ is as in Lemma 3.4; this assumption will be justified subsequently.

First, since $r > N$, (31) gives a uniform bound on the $\|\nabla u_k\|_\infty$; through (24) this gives a uniform estimate on the α_k,

$$\alpha_k \leqq a_0, \qquad 0 \leqq k < n. \tag{32}$$

We have from $u_0 \in S(\lambda)$,

$$H_\lambda(u_0) \leqq m_2^2 \lambda^{-2/(q-1)};$$

thus, in view of (18),

$$\begin{aligned} \|u_0\|_{1,2}^2 &\leqq a_0 \sqrt{(2+\varepsilon)/\varepsilon}\, H_\lambda(u_0) \\ &\leqq a_0 \sqrt{(2+\varepsilon)/\varepsilon}\, m_2^2 \lambda^{-2/(q-1)}, \end{aligned}$$

where a_0 is as in (32) (here we are taking $\|u\|_{1,2} = (\int_\Omega |\nabla u|^2\, dx)^{1/2}$). From the latter inequality, via (8),

$$\begin{aligned} \|f(u_0)\|_{s_0} &\leqq C \|u_0\|_{p^*}^q \\ &\leqq KCm_2^q (a_0^2(2+\varepsilon)/\varepsilon)^{q/4} \lambda^{-q/(q-1)}, \end{aligned}$$

and hence from (28), (32)

$$\|u_1\|_{2,s_0} \leqq KMCa_0^{1+q/2} m_2^q ((2+\varepsilon)/\varepsilon)^{q/4} \lambda^{-1/(q-1)} \leqq \tau_1 \lambda^{-1/(q-1)} \tag{33}$$

(here K incorporates an imbedding constant and M is as in Lemma 3.4).

Similarly we have

$$\begin{aligned} \|f(u_0)\|_r &\leqq CK \|u_0\|_\infty^q \\ &\leqq KCm_1^q \lambda^{-q/(q-1)}; \end{aligned}$$

hence

$$\|u_1\|_{2,r} \leqq KMCa_0 m_1^q \lambda^{-1/(q-1)} \leqq \bar{\tau}_1 \lambda^{-1/(q-1)}.$$

Proceeding inductively we get, for $1 \leqq k \leqq n$,

$$\|f(u_k)\|_{s_k} \leqq C \|u_k\|_{qs_k}^q \leqq CK^q \|u_k\|_{2,s_{k-1}}^q \leqq CK^q \tau_k^q \lambda^{-q/(q-1)},$$

and hence

$$\|u_{k+1}\|_{2,s_k} \leqq CMK^q a_0 \tau_k^q \lambda^{-1/(q-1)} \leqq \tau_{k+1} \lambda^{-1/(q-1)}.$$

Thus, with τ_1 defined implicitly by (33), the τ_k are determined recursively by

$$\tau_{k+1} = CMK^q a_0 \tau_k^q,$$

where K is an embedding constant that depends on k. The $\bar{\tau}_k$ satisfy a similar recursive relation with different values for K. The τ_k are completely determined and we now choose m_1 so that

$$m_2, \tau_n < m_1. \tag{34}$$

The $\bar{\tau}_k$ now also can be determined and we shall assume, as we can, that they are nondecreasing with $m_1 \leqq \bar{\tau}_1$.

Let $\Lambda > 0$ be such that

$$\bar{\tau}_n \Lambda^{-1/(q-1)} < \delta.$$

We easily see then that the assumption (31) is justified provided that $\lambda \geqq \Lambda$. Thus if $u_0, \cdots, u_n$ are as above then the inequalities (29) and (30) will hold provided $\lambda \geqq \Lambda$.

It follows from (34), (27), Lemma 3.2, and (25) that $S(\lambda) \neq \varnothing$. Let $\lambda \geqq \Lambda$, let $u = u_0 \in S(\lambda)$ be chosen arbitrarily and let u_n be related to u as above, i.e., let $u_n = T^n u$. Then we easily see that $u_n \in S(\lambda)$. Indeed $N_\lambda(u_n) = 0$ by the definition of T_λ, $H_\lambda(u_n) \leqq H_\lambda(u) \leqq m_2^2 \lambda^{2/(q-1)}$ by Lemma 3.1, $u_n \geqq 0$ almost everywhere for elementary reasons, and finally we have

$$\|u_n\|_{2,r} \leqq \tau_n \lambda^{-1/(q-1)} \leqq m_1 \lambda^{-1/(q-1)}$$

by (34). Thus, as claimed, T^n maps $S(\lambda)$ into itself provided $\lambda \geqq \Lambda$.

5. Statement of results. From the preceding section we deduce the following result.

THEOREM 5.1. *Let Ω be a bounded domain in $\mathbb{R}^N$ with $C^{1,1}$ boundary. Let $f \in C[0, \infty)$ satisfy* (8), *where* (9) *holds, and* (10). *Then there is a function $\Lambda = \Lambda(\Omega, \mu, f)$, monotone nondecreasing in μ for Ω, f fixed, such that for all $\lambda \geqq \Lambda(\Omega, \mu, f)$, the problem* (7) *has a nontrivial, nonnegative C^1-solution.*

Remarks. 1. It is clear from the proof that f need only be defined and/or satisfy (8), (10) on some interval $[0, c)$.

2. Explicit dependence of f on x can be allowed provided (8), (9) hold with the constant C independent of x and (10) is satisfied.

REFERENCES

[1] F. V. ATKINSON, L. PELETIER, AND J. SERRIN, *Ground states for the prescribed mean curvature equation: the supercritical case*, in Nonlinear Diffusion Equations and Their Equilibrium States I, Proc. of a Microprogram held August 25–September 12, 1986, W.-M. Ni, L. A. Peletier, and J. Serrin, eds., Springer-Verlag, New York, 1988.

[2] C. V. COFFMAN AND M. M. MARCUS, *Existence theorems for superlinear elliptic Dirichlet problems in exterior domains*, in Nonlinear Functional Analysis and Its Applications, Proc. of Symposia in Pure Mathematics, 45, American Mathematical Society, Providence, RI, 1986.

[3] C. V. COFFMAN, *An existence theorem for a class of non-linear integral equations with applications to a non-linear boundary value problem*, J. Math. and Mech., 18 (1968), pp. 411–422.

[4] B. FRANCHI AND E. LANCONELLI, *Radial symmetry of the ground states for a class of quasi-linear elliptic equations*, in Nonlinear Diffusion Equations and Their Equilibrium States I, Proc. Microprogram, August 25–September 12, 1986, W.-M. Ni, L. A. Peletier, and J. Serrin, eds., Springer-Verlag, New York, 1988.

[5] B. FRANCHI, E. LANCONELLI, AND J. SERRIN, *Existence and uniqueness of ground state solutions of quasi-linear equations*, in Nonlinear Diffusion Equations and Their Equilibrium States I, Proc. Microprogram, August 25–September 12, 1986, W.-M. Ni, L. A. Peletier, and J. Serrin, eds., Springer-Verlag, New York, 1988.

[6] B. GIDAS, W.-M. NI, AND L. NIRENBERG, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[7] Z. NEHARI, *On a class of non-linear second-order differential equations*, Trans. Amer. Math. Soc., 95 (1960), pp. 101–123.

[8] W.-M. NI AND J. SERRIN, *Non-existence theorems for quasi-linear partial differential equations*, Rend. Circ. Mat. Palermo (2), Suppl. 8 (1985), pp. 171–185.

[9] L. A. PELETIER AND J. SERRIN, *Ground states for the prescribed mean curvature equation*, Proc. Amer. Math. Soc., 100 (1987), pp. 694–700.

[10] P. PUCCI AND J. SERRIN, *A general variational identity*, Indiana Univ. Math. J., 35 (1986), pp. 681–703.

[11] J. SERRIN, *Positive solutions of a prescribed mean curvature problem*, in Calculus of Variations and Differential Equations, Lecture Notes in Math. 1340, Springer-Verlag, New York, 1988.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 991-1006, July 1991

008

ON TRANSMISSION PROBLEMS FOR THE SCHRÖDINGER EQUATION*

G. F. ROACH† AND BO ZHANG†

Abstract. In this paper, weighted Sobolev spaces and the method of limiting absorption are used to settle questions of existence and uniqueness of solutions to transmission problems for the Schrödinger equation. The methods used here are suitable for problems in which there are inhomogeneous terms and variable coefficients that vanish asymptotically at infinity. Furthermore, it is shown that transmission solutions which satisfy the radiation condition

$$(1+r)^{-\mu/2}\left(\frac{\partial u_2}{\partial r}-iku_2\right)\in L_2(\Omega_2), \qquad r=|x|, \quad x\in\mathbf{R}^n, \quad 0\leqq\mu\leqq 1$$

also satisfy an integrability condition at infinity of the form

$$(1+r)^{-1/2}(\ln(e+r))^{-1/2-\delta/2}u_2\in L_2(\Omega_2), \qquad \delta>0.$$

Key words. transmission problem, Schrödinger equation, limiting absorption principle, radiation condition, weighted Sobolev space

AMS(MOS) subject classifications. 35J10, 78A45

1. Introduction.

Exterior boundary value problems involving compact boundaries associated with the Helmholtz equation

$$(\Delta+k^2)u=f, \qquad k^2>0 \tag{1.1}$$

have already been investigated by many authors. In these investigations a crucial role is played by the Sommerfeld radiation conditions. These were obtained by A. Sommerfeld from physical consideration and require that solutions should have the behaviour

$$u=O(r^{-(n-1)/2}), \tag{1.2}$$

$$\frac{\partial u}{\partial r}-iku=o(r^{-(n-1)/2}), \tag{1.3}$$

as $r=|x|\to\infty$. Examples of corresponding existence and uniqueness theorems can be found in the book by Kupradze [13]. Vekua [22] has shown that (1.2) is superfluous since (1.3) on its own guarantees uniqueness, while Rellich [19] has shown that (1.3) can be weakened to the following integral form:

$$\lim_{R\to\infty}\int_{s(R)}|\partial u/\partial r-iku|^2\,ds=0, \tag{1.4}$$

where $s(R)$ is a sphere of radius R. The existence of radiating solutions to (1.1) has been settled using, for instance, integral equation methods (see, for example, Jager [5]). Similar problems involving differential equations of more general elliptic type in the limiting case where the losses vanish have been studied by a number of Russian authors [2], [3], [21].

The existence and uniqueness of a radiating solution for boundary value problems involving an infinite boundary for general elliptic differential equations has been

* Received by the editors August 7, 1989; accepted for publication (in revised form) May 3, 1990.
† Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, Scotland.

considered by Vogelsang [23] and Saranen [20]. These authors require that radiating solutions should satisfy the following conditions: a finiteness condition

$$(1+|x|^2)^{-1/4-\delta/2}u \in L_2(\Omega) \tag{1.5}$$

and a radiating condition

$$(1+|x|^2)^{-\delta/2}\nabla(e^{-ikr}u) \in (L_2(\Omega))^n. \tag{1.6}$$

In Neittaanmaki and Roach [15] it is shown that under the assumption $(1+|x|)\times(\ln(e+|x|))^{\delta/2}f \in L_2(\Omega)$ there exists a unique solution u of (1.1) satisfying the radiation condition

$$\frac{\partial u}{\partial r} - iku \in L_2(\Omega). \tag{1.7}$$

In [15] it is also shown that the radiation solution has the following integrability property:

$$(1+|x|)^{-1/2}(\ln(e+|x|))^{-1/2-\delta/2}u \in L_2(\Omega), \tag{1.8}$$

where δ is some fixed number satisfying $0<\delta<\frac{1}{2}$.

Transmission problems for the Helmholtz equation (1.1) have also been studied. In [11] Kress and Roach considered the existence and uniqueness of transmission solutions for (1.1), (1.2), and (1.3), with $f=0$ and established existence and uniqueness theorems using integral equation methods. In this connection we also mention the works of Kittappa and Kleinman [9], Kleinman and Martin [10], Kupradze [13], Ramm [18], and Werner [25], [26]. In particular, we remark that in [25] and [26] Werner used integral equation methods to study the existence and uniqueness of transmission solutions to the reduced wave equation in which the smooth coefficients and inhomogeneous terms are assumed to be constants and zero, respectively, outside a sphere of sufficiently large radius. Jones [6] established a uniqueness theorem for elastodynamics, with spatially varying parameters, by making extensive use of the properties of the spherical harmonics. Kristensson [12] obtained a uniqueness theorem for Helmholtz equation in penetrable media with an infinite interface in the lossless case, and in this connection we also mention the results of Odeh [16]. Beck [1] studied a transmission problem for a second-order uniformly elliptic differential operator of Helmholtz type in the case of a more general geometry compared to [12]. Here he derived a Rellich-type estimate for the transmission solutions and established the uniqueness theorem using this estimate in the case where the field u and its covariant derivative are continuous on the interface.

In the present paper we study the uniqueness and existence of transmission solutions to a non-selfadjoint differential equation of Schrödinger type. In [4] Eidus proved that under certain conditions the non-selfadjoint differential operator in $L_2(\mathbf{R}^n)$ defined by

$$L := -\Delta + B(x)\nabla + q(x) \tag{1.9}$$

has no positive eigenvalues. Using this result of Eidus and a radiation condition of the form

$$(1+r)^{-\mu/2}\left(\frac{\partial u_2}{\partial r} - iku_2\right) \in L_2(\Omega_2), \qquad 0 \leqq \mu \leqq 1, \tag{1.10}$$

we can obtain a uniqueness theorem for transmission problems for the Schrödinger equation

$$-\Delta u + B(x)\nabla u + q(x)u - k^2 u = f, \qquad k^2 > 0. \tag{1.11}$$

This is achieved by imposing suitable conditions on the coefficients and by using the unique continuation principle [14]. Weighted Sobolev spaces are then used to prove the limiting absorption principle, from which an existence theorem for transmission solutions can be derived. Furthermore, it is shown that transmission solutions satisfy the integrability property

$$(1+|x|)^{-1/2}(\ln(e+|x|))^{-1/2-\delta/2}u_2 \in L_2(\Omega_2), \tag{1.12}$$

where δ is some fixed number and $0<\delta\leqq 1$.

2. Statement of the problem and notation. Let $\Omega_1 \subset \mathbf{R}^n$ be a bounded domain and Ω_2 be the exterior. The interface boundary $\Gamma := \partial\Omega_1 = \partial\Omega_2$ is assumed to be of class C^2. We denote by $\nu=(\nu_1, \nu_2, \cdots, \nu_n)$ the unit normal to Γ drawn in the direction from Ω_1 to Ω_2.

We are concerned with the uniqueness and existence of solutions to the following problem.

Transmission problem. Find $u=(u_1, u_2)\in C^2(\Omega_1)\times C^2(\Omega_2)$ such that

$$\begin{gathered} -\Delta u_j + B_j\nabla u_j + q_j u_j - k^2 u_j = f_j \quad \text{in } \Omega_j, \quad j=1,2, \\ \alpha u_1 - u_2 = g_1 \quad \text{on } \Gamma, \\ \beta\,\partial u_1/\partial\nu - \partial u_2/\partial\nu = g_2 \quad \text{on } \Gamma, \\ (1+r)^{-\mu/2}\left(\frac{\partial u_2}{\partial r} - iku_2\right) \in L_2(\Omega_2), \qquad 0\leqq\mu\leqq 1, \end{gathered} \tag{2.1}$$

where B_j, q_j, f_j, and u_j are complex-valued functions on Ω_j, $j=1,2$, g_j, $j=1,2$ are complex-valued functions on Γ, $k^2>0$, $\alpha, \beta\in\mathbf{C}\backslash\{0\}$ are constants. In Ω_j, $j=1,2$, we assume that the following relations hold:

$$|B_j|+|q_j|\leqq C, \qquad j=1,2, \tag{2.2}$$

$$q_1, B_1\in C(\Omega_1), \qquad q_2, B_2\in C(\Omega_2), \tag{2.3}$$

$$|q_2|+|B_2| = O(r^{-2}(\ln r)^{-1-\delta}), \qquad 0<\delta\leqq 1, \quad r\to\infty, \tag{2.4}$$

$$f_j\in L_2(B(R)\cap\Omega_j) \quad \forall R>0, \quad j=1,2, \quad g_1\in H^{1/2}(\Gamma), \quad g_2\in H^{-1/2}(\Gamma).$$

With $D\subset\mathbf{R}^n$ an open set, we introduce complex Sobolev spaces $H^k(D)$, $k=0,1,2,\cdots$, and $H^l(\Gamma)$, $l\in\mathbf{R}$. The inner product on $H^k(D)$ is denoted by

$$(u, v)_{k,D} = \sum_{|\alpha|\leqq k}\int_D \partial^\alpha u\,\partial^\alpha\bar{v}\,dx,$$

where $\partial^\alpha = \partial_1^{\alpha_1}\cdots\partial_n^{\alpha_n}$, $\partial_i = \partial/\partial x_i$, and the corresponding norm by $\|u\|_{k,D}=\sqrt{(u,u)_{k,D}}$. The duality relation between $H^l(\Gamma)$ and $H^{-l}(\Gamma)$ will be denoted by $\langle\circ,\circ\rangle$. By $\mathcal{D}(D)$ we denote the space of infinitely continuously differential functions having compact support in D.

It is convenient to introduce the following weight functions:

$$\begin{aligned} P_\delta(x) &:= (1+|x|)^{-1/2}(\ln(e+|x|))^{-1/2-\delta/2}, \\ q_\delta(x) &:= (\ln(e+|x|))^{-\delta/2}, \\ \rho(x) &:= (1+|x|)(\ln(e+|x|))^{\delta/2}, \end{aligned} \tag{2.5}$$

and for any weight function g the following weighted Sobolev spaces:

$$\begin{aligned} L_g^2(D) &:= \{u\in L_2(D\cap B(R)) \ \forall R>0 \,|\, gu\in L_2(D)\}, \\ H_g^k(D) &:= \{u\in H^k(D\cap B(R)) \ \forall R>0 \,|\, g\partial^\alpha u\in L_2(D), \ |\alpha|\leqq k\}, \end{aligned}$$

with the weighted norm

$$\|u\|_{k,g,D} := \left(\sum_{|\alpha| \leqq k} \int_D g^2 |\partial^\alpha u|^2 \, dx \right)^{1/2},$$

and the Sobolev space

$$L_g^2(\Omega) := L_g^2(\Omega_1) \times L_g^2(\Omega_2),$$

with the norm

$$\|u\|_{0,g,\Omega} := \left(\sum_{j=1}^{2} \|u_j\|_{0,g,\Omega_j}^2 \right)^{1/2},$$

where $\Omega = \Omega_1 \cup \Omega_2$.

Let $r = |x|$, $\hat{x} = x/r$, and $\partial_r u = \partial u / \partial r = \hat{x} \nabla u$, where $\nabla u = (\partial_1 u, \cdots, \partial_n u)$. We introduce the abbreviations

$$E(R) := \{x \in \mathbf{R}^n \,|\, |x| > R\},$$
$$B(R) := \{x \in \mathbf{R}^n \,|\, |x| < R\},$$
$$S(R) := \{x \in \mathbf{R}^n \,|\, |x| = R\},$$
$$D(R) := \{x \in \Omega_2 \,|\, |x| \leqq R\}.$$

For convenience, we suppose that $0 \in \Omega_1$ and that R_0 is a fixed number such that $\bar{\Omega}_1 \subset B(R_0)$. Thus, let

$$\Omega(R) := \{x \in E(R_0) \,|\, |x| < R\},$$
$$\Omega(R_1, R_2) := \{x \in E(R_0) \,|\, R_1 < |x| < R_2\}.$$

To give a definition of a weak formulation of (2.1) we introduce the following bounded, bilinear, and linear functionals:

$$a_j(u_j, v_j) := \int_{\Omega_j} (\nabla u_j \nabla \bar{v}_j + B_j \nabla u_j \bar{v}_j + q_j u_j \bar{v}_j - k^2 u_j \bar{v}_j) \, dx \qquad \forall u_j, v_j \in H^1(\Omega_j), \quad j = 1, 2, \tag{2.6}$$

$$a(u, v) := \bar{\alpha}\beta a_1(u_1, v_1) + a_2(u_2, v_2), \qquad \forall u = (u_1, u_2), \quad v = (v_1, v_2) \in H^1(\Omega_1) \times H^1(\Omega_2), \tag{2.7}$$

$$F(v) := \bar{\alpha}\beta \int_{\Omega_1} f_1 \bar{v}_1 \, dx + \int_{\Omega_2} f_2 \bar{v}_2 \, dx + \int_\Gamma g_2 \bar{v}_2 \, ds \qquad \forall v = (v_1, v_2) \in H^1(\Omega_1) \times H^1(\Omega_2), \tag{2.8}$$

together with the sets

$$H_E := \{v = (v_1, v_2) \,|\, v_j \in H^1(B(R) \cap \Omega_j) \ \forall R > 0, \ j = 1, 2, \ \alpha v_1 = v_2 \text{ on } \Gamma\}$$

and

$$H_E^* := \{v = (v_1, v_2) \,|\, v_j \in H^1(B(R) \cap \Omega_j) \ \forall R > 0, \ j = 1, 2, \ \alpha v_1 - v_2 = g_1 \text{ on } \Gamma\}.$$

With the above notation, a weak formulation of (2.1) is given:

$$\text{Find } u = (u_1, u_2) \in H_E^* \quad \text{such that } a(u, \phi v) = F(\phi v) \quad \forall v \in H_E, \quad \phi \in \mathcal{D}(\mathbf{R}^n),$$
$$(1 + r)^{-\mu/2} (\partial u_2 / \partial r - iku_2) \in L_2(\Omega_2). \tag{2.9}$$

Solutions of (2.9) are called weak solutions of (2.1). In what follows we are concerned with showing that weak solutions of the transmission problem (2.1) exist and, moreover, are unique.

3. On the uniqueness of the transmission solutions. To establish the required uniqueness of the solutions to (2.9), we make use of the following result due to Eidus [4, Thm. 3.2].

LEMMA 3.1 [4]. *Under the assumptions* (2.2)-(2.4), *every solution* $u \in H^2(E(R_1) \cap B(R))$ *for all* $R > R_1$ *of the equation*

$$-\Delta u + B_2 \nabla u + q_2 u = k^2 u, \qquad k^2 > 0 \quad \text{in } E(R_1), \quad R_1 \geqq R_0$$

is equal to zero almost everywhere in $E(R_1)$ *if* $u \in L_2(E(R_1))$.

THEOREM 3.1. *The transmission problem* (2.9) *has at most one solution.*

Proof. Let $u = (u_1, u_2) \in H_E$ be such that

$$a(u, \phi v) = 0 \quad \forall v \in H_E, \quad \phi \in \mathcal{D}(\mathbf{R}^n) \tag{3.1}$$

and

$$(1+r)^{-\mu/2}\left(\frac{\partial u_2}{\partial r} - iku_2\right) \in L_2(\Omega_2). \tag{3.2}$$

Then we have

$$u_j \in H^1(B(R) \cap \Omega_j) \quad \forall R > 0, \qquad a_j(u_j, \phi) = 0 \quad \forall \phi \in \mathcal{D}(\Omega_j), \quad j = 1, 2. \tag{3.3}$$

Furthermore, by the coerciveness estimates [14, Thm. 2.7], we see that

$$u_j \in H^2(B(R) \cap \Omega_j) \quad \forall R > 0, \tag{3.4}$$

and that for any $R > 0$ there is a constant C_R such that

$$|\Delta u_j| \leqq C_R(|u_j| + |\nabla u_j|) \quad \text{a.e. in } B(R) \cap \Omega_j \quad \forall R > 0, \tag{3.5}$$

and

$$-\Delta u_j + B_j \nabla u_j + q_j u_j - k^2 u_j = 0 \quad \text{a.e. in } B(R) \cap \Omega_j \quad \forall R > 0, \quad j = 1, 2. \tag{3.6}$$

Hence we can obtain from the definition of H_E and (3.1) the transmission conditions

$$\alpha u_1 = u_2 \qquad \text{on } \Gamma, \tag{3.7}$$

$$\frac{\beta\, \partial u_1}{\partial n} = \frac{\partial u_2}{\partial \nu} \quad \text{on } \Gamma. \tag{3.8}$$

Bearing in mind Lemma 3.1 and (3.6), it remains to prove that $u_2 \in L_2(\Omega_2)$.

Step 1. We prove that

$$(1+r)^{-1/2}(\ln(e+r))^{-1/2-\delta/2}(|u_2| + |\nabla u_2|) \in L_2(\Omega_2)$$

and

$$(1+r)^{-1/2}|\nabla(e^{-ikr}u_2)| \in L_2(\Omega_2).$$

Let $R > R_1 > R_0 + 1$, $\eta(r) = r^{-1/2}(\ln r)^{-1/2-\delta/2}$ and $g(r) = r^{-1/2}$ for $r > 1$. By direct calculation with $w = e^{-ikr}u_2$, we see from (3.6) that

$$(-\Delta - ik(2\partial_r + (n-1)r^{-1}) + B_2\nabla + q_2 + ikB_2\hat{x})w = 0 \quad \text{a.e. in } \Omega_2. \tag{3.9}$$

On multiplying the equation (3.9) by $(\ln r)^{-\delta}\bar{w}$, integrating over $\Omega(R_1, R)$, and taking the imaginary part of both sides of the equation thus obtained, we get by appropriate partial integration that

$$\begin{aligned}&\delta k\|u_2\|^2_{0,\eta,\Omega(R_1,R)}+k(\ln R)^{-\delta}\|u_2\|^2_{0,s(R)}\\&=\operatorname{Im}\int_{\Omega(R_1,R)}[B_2\nabla w\bar{w}+q_2|w|^2+ikB_2\hat{x}|w|^2-\delta(r\ln r)^{-1}\partial_r w\bar{w}](\ln r)^{-\delta}\,dx\\&\quad-(\ln R)^{-\delta}\operatorname{Im}(\partial_r w, w)_{0,s(R)}+K_{R_1}(w),\end{aligned}\tag{3.10}$$

where $|K_{R_1}(w)|\leqq C\|w\|^2_{1,s(R_1)}<+\infty$. By (2.4) and Cauchy's inequality we obtain from (3.10) that

$$\begin{aligned}&\delta k\|u_2\|^2_{0,\eta,\Omega(R_1,R)}+k(\ln R)^{-\delta}\|u_2\|^2_{0,s(R)}\\&\leqq C(\ln R_1)^{-\delta}\{\|\nabla w\|^2_{0,g,\Omega(R_1,R)}+\|u_2\|^2_{0,\eta,\Omega(R_1,R)}+(\ln R)^{-\delta}\|u_2\|^2_{0,s(R)}\}\\&\quad+\|\partial_r w\|^2_{0,s(R)}+K_{R_1}(w).\end{aligned}\tag{3.11}$$

On the other hand, multiply (3.9) by $2g^2(r\partial_r\bar{w}+(n-1)/2\bar{w})$, integrate over $\Omega(R_1, R)$, and make use of the basic identity

$$\begin{aligned}2\operatorname{Re}[\Delta w r h(r)\partial_r\bar{w}]&=\operatorname{div}[2\operatorname{Re}\nabla w r h(r)\partial_r\bar{w}-h(r)|\nabla w|^2x]+(n-2)h(r)|\nabla w|^2\\&\quad+r\partial_r h(r)(|\nabla w|^2-2|\partial_r w|^2)\quad\text{a.e. in } E(R_0),\end{aligned}\tag{3.12}$$

with $h(r)=g^2=r^{-1}$ to obtain, after integrating by parts, the result

$$\begin{aligned}&2\|\nabla w\|^2_{0,g,\Omega(R_1,R)}-2\|\partial_r w\|^2_{0,g,\Omega(R_1,R)}-(n-1)\operatorname{Re}(\partial_r w, r^{-2}w)_{0,\Omega(R_1,R)}\\&\quad+2\operatorname{Re}((B_2\nabla+q_2+ikB_2\hat{x})w, g^2(r\partial_r+(n-1)/2)w)_{0,\Omega(R_1,R)}\\&\quad-\int_{s(R)}(2|\partial_r w|^2-|\nabla w|^2)\,ds-(n-1)R^{-1}\operatorname{Re}\int_{s(R)}\bar{w}\partial_r w\,ds\\&\quad+\int_{s(R_1)}(2|\partial_r w|^2-|\nabla w|^2)\,ds+(n-1)R_1^{-1}\operatorname{Re}\int_{s(R_1)}\bar{w}\partial_r w\,ds=0.\end{aligned}\tag{3.13}$$

Taking into account (2.4) and the fact that $|\nabla w|^2-|\partial_r w|^2\geqq 0$, we obtain from (3.13) that

$$\begin{aligned}\|\nabla w\|^2_{0,g,\Omega(R_1,R)}&\leqq CR_1^{-1/2}\{\|\nabla w\|^2_{0,g,\Omega(R_1,R)}+\|u_2\|^2_{0,\eta,\Omega(R_1,R)}\\&\quad+(\ln R)^{-\delta}\|u_2\|^2_{0,s(R)}\}+\|\partial_r w\|^2_{0,s(R)}\\&\quad+\|\partial_r w\|^2_{0,g,\Omega(R_1,R)}+K_{R_1}(w).\end{aligned}\tag{3.14}$$

Thus for R_1 sufficiently large we get from (3.11) and (3.14) that

$$\|u_2\|^2_{0,\eta,\Omega(R_1,R)}+\|\nabla w\|^2_{0,g,\Omega(R_1,R)}\leqq C(R_1)\{\|\partial_r w\|^2_{0,s(R)}+\|\partial_r w\|^2_{0,g,\Omega(R_1,R)}\}+K_{R_1}(w).$$

As $(1+r)^{-\mu/2}\partial_r w\in L_2(\Omega_2)$ and $0\leqq\mu\leqq 1$, there is a sequence $R_j\to\infty$ such that $\|\partial_r w\|^2_{0,s(R_j)}$ tends to zero. Thus, we get

$$\|u_2\|^2_{0,\eta,E(R_1)}+\|\nabla w\|^2_{0,g,E(R_1)}\leqq C(R_1)\|\partial_r w\|^2_{0,g,E(R_1)}+K_{R_1}(w)<+\infty.$$

Consequently, $(1+r)^{-1/2}(\ln(e+r))^{-1/2-\delta/2}u_2\in L_2(\Omega_2)$ and $(1+r)^{-1/2}|\nabla(e^{-ikr}u_2)|\in L_2(\Omega_2)$. This, combined with (3.6), implies that $(1+r)^{-1/2}(\ln(e+r))^{-1/2-\delta/2}|\nabla u_2|\in L_2(\Omega_2)$.

Step 2. To show that $|\nabla w|\in L_2(\Omega_2)$, we need the following.

In the equation (3.13), integrating

$$\operatorname{Re}\int_{\Omega(R_1,R)}r^{-2}\bar{w}\partial_r w\,dx$$

by parts and letting $R \to \infty$, we get

$$2\int_{E(R_1)} r^{-1}\{|\nabla w|^2 - |\partial_r w|^2\}\,dx + \int_{s(R_1)} \{2|\partial_r w|^2 - |\nabla w|^2 + (n-1)R_1^{-1}\,\mathrm{Re}\,\bar{w}\partial_r w\}\,ds$$
$$= -(n-1)/2\left\{(n-3)\int_{E(R_1)} r^{-3}|u_2|^2\,dx + \int_{s(R_1)} r^{-2}|u_2|^2\,ds\right\}$$
$$-\mathrm{Re}\,((B_2\nabla + q_2 + ikB_2\hat{x})w, (2\partial_r w + (n-1)r^{-1}w))_{0,E(R_1)}.$$

From the assumption (2.4) and Step 1 we infer that the right-hand side of the above equality is integrable with respect to R_1 from R to ∞ and, therefore, so is the left-hand side. Integrating both sides of the above equality with respect to R_1 from R to ∞ and denoting the left-hand side and the right-hand side of the equality thus obtained by $LH(R)$ and $RH(R)$, respectively, we then have $RH(R) < +\infty$, while

$$LH(R) = \liminf_{R_2\to\infty}\left\{\int_{\Omega(R,R_2)} [|\nabla w|^2 + (n-1)r^{-1}\,\mathrm{Re}\,\bar{w}\partial_r w]\,dx\right.$$
$$+2R_2\int_{E(R_2)} r^{-1}[|\nabla w|^2 - |\partial_r w|^2]\,dx$$
$$\left.-2R\int_{E(R)} r^{-1}[|\nabla w|^2 - |\partial_r w|^2]\,dx\right\}$$
$$\geqq \liminf_{R_2\to\infty}\left\{\frac{1}{2}\int_{\Omega(R,R_2)} |\nabla w|^2\,dx - (n-1)^2/2\int_{\Omega(R,R_2)} r^{-2}|u_2|^2\,dx\right.$$
$$\left.-2R\int_{E(R)} r^{-1}[|\nabla w|^2 - |\partial_r w|^2]\,dx\right\},$$

where we have used Young's inequality and the fact that

$$|\nabla w|^2 - |\partial_r w|^2 \geqq 0.$$

Hence, from the above inequality and Step 1 it follows immediately that $|\nabla w| \in L_2(\Omega_2)$.

Step 3. To show that $\liminf_{R\to\infty}\int_{s(R)}(|u_2|^2 + |\nabla u_2|^2)\,ds = 0$, we need the following.

Let $R > R_0$ and j be an arbitrary integer with $j > R$. Choose $\phi_j(r) \in \mathscr{D}(\mathbf{R}^n)$, satisfying

$$\phi_j(r) = 1, \quad r \leqq R, \quad \phi_j(r) = 0, \quad r \geqq j, \quad 0 \leqq \phi_j \leqq 1 \quad \text{in } \mathbf{R}^n, \tag{3.15}$$

$$|\partial_r \phi_j| \leqq C(j-R)^{-1}, \qquad \partial_r \phi_j \leqq 0, \quad R < r < j, \tag{3.16}$$

where C is a constant independent of r, R, j, and ϕ_j. The existence of suitable ϕ_j is indicated in Remark 3.1 below. On multiplying (3.9) by $\phi_j \bar{w}$ and integrating it over $\Omega(R,j)$, we have

$$\mathrm{Im}\,(\partial_r u_2, u_2)_{0,s(R)} = -\mathrm{Im}\int_{\Omega(R,j)} \{[B_2\nabla w + q_2 w + ikB_2\hat{x}w]\phi_j\bar{w} + \partial_r\phi_j\partial_r w\bar{w}\}\,dx$$
$$-k\int_{\Omega(R,j)} \partial_r\phi_j|u_2|^2\,dx. \tag{3.17}$$

By assumption (2.4), Step 1, Step 2, (3.15), and (3.16), we see that the first term of the right-hand side of (3.17) remains finite as $j \to \infty$ and therefore the second one is finite as well. Define

$$f_j(r) := k|\partial_r\phi_j|\,\|u_2\|^2_{0,s(r)} \quad \text{for all } r \geqq R;$$

then $f_j(r) \to 0$ almost everywhere in $[R, \infty)$ as $j \to \infty$ from (3.15) and (3.16) and

$$I_j(R) := -k \int_{\Omega(R,j)} \partial_r \phi_j |u_2|^2 \, dx = \int_R^\infty f_j(r) \, dr$$

is uniformly bounded for $j \geqq R$ and is convergent as $j \to \infty$ from (3.17). For all $j > R$ let $g_j = \sup\{f_j, f_{j+1}, \cdots\}$. Then g_j is integrable. In fact, $Mf_{jk} := \max\{f_j, f_{j+1}, \cdots, f_{j+k}\}$ is integrable. Moreover, $Mf_{jk} \to g_j$ as $k \to \infty$ and for any $j > R$ there is a constant $A(j) > 0$, independent of k and Mf_{jk}, such that

$$\int_R^\infty Mf_{jk} \, dr \leqq A(j) \quad \text{for all positive integer } k \tag{3.18}$$

for $G_j(x) := \sup_k \int_x^\infty Mf_{jk} \, dr$ with $x \geqq R$ is a nonincreasing positive function of x, and thus for any $j > R$, $G_j(x)$ tends to M_j, with $0 \leqq M_j < \infty$ as $x \to \infty$. Therefore the function g_j is indeed integrable by Levi's theorem (see [24, p. 368]). The sequence $\{g_j\}$ is nonincreasing, the integrals $\int g_j \, dr \geqq 0$ and $g_j \to 0$ almost everywhere in $[R, \infty)$. Hence,

$$\int_R^\infty g_j \, dr \to 0 \tag{3.19}$$

as $j \to \infty$ by Levi's theorem. The inequalities $f_j \leqq g_j$ for all $j > R$ imply that $I_j(R) \to 0$ as $j \to \infty$. This combined with (3.17) yields

$$\operatorname{Im} (\partial_r u_2, u_2)_{0,s(R)} \to 0$$

as $R \to \infty$. From this and the fact that

$$|\nabla w|^2 = |\nabla u_2|^2 + k^2 |u_2|^2 - 2k \operatorname{Im} \partial_r u_2 \bar{u}_2, \qquad |\nabla w| \in L_2(\Omega_2),$$

the required result follows.

Step 4. We prove that $u_2 \in L_2(\Omega_2)$ and $u_j = 0$ almost everywhere in Ω_j, $j = 1, 2$.

Multiply the equation (3.6) by $2\partial_r \bar{u}_2 + (n-1) r^{-1} \bar{u}_2$, integrate it over $E(R)$, and take the real part of both sides of the equality thus obtained to get

$$\begin{aligned} 2(\|\nabla u_2\|^2_{0,g,E(R)} - \|\partial_r u_2\|^2_{0,g,E(R)}) &- (n-1) \operatorname{Re} (\partial_r u_2, r^{-2} u_2)_{0,E(R)} \\ &+ \operatorname{Re} ((B_2 \nabla + q_2) u_2, 2\partial_r u_2 + (n-1) r^{-1} u_2)_{0,E(R)} \\ &+ \int_{s(R)} \{2|\partial_r u_2|^2 + k^2 |u_2|^2 - |\nabla u_2|^2\} \, ds \\ &+ (n-1) \operatorname{Re} \int_{s(R)} r^{-1} \bar{u}_2 \partial_r u_2 \, ds = 0. \end{aligned}$$

From this equality, (2.4), Step 1, and Step 2 we are able to obtain $u_2 \in L_2(\Omega_2)$ by following the same argument used in Step 2. Consequently, we have $u_2 = 0$ almost everywhere in $E(R)$ for some fixed $R > R_0$ by Lemma 3.1. The unique continuation principle [14, p. 65] and the transmission conditions (3.7) and (3.8) imply that $u_1 = 0$ almost everywhere in Ω_1 and $u_2 = 0$ almost everywhere in Ω_2. The theorem is proved.

Remark 3.1. Define $\phi_j \in \mathscr{D}(\mathbf{R}^n)$ as follows:

$$\phi_j(r) = \begin{cases} 1, & r \leqq R, \\ 0, & r \geqq j, \end{cases} \qquad \phi_j(r) = \exp\{1 - (j-R)^2 [(j-R)^2 - (r-R)^2]^{-1}\}, \quad R < r < j.$$

Then ϕ_j satisfies (3.15) and (3.16).

4. On the existence of the transmission solutions. The existence of the transmission solutions will be established by the principle of limiting absorption as given by Eidus [2], [3]. To this end, we introduce the following bilinear functionals:

$$a_{2\varepsilon}(u_2, v_2) = a_2(u_2, v_2) - i\varepsilon(u_2, v_2)_{0,\Omega_2} \quad \forall u_2, v_2 \in H^1(\Omega_2),$$

$$a_\varepsilon(u, v) = \bar{\alpha}\beta a_1(u_1, v_1) + a_{2\varepsilon}(u_2, v_2) \quad \forall u, v \in H^1(\Omega_1) \times H^1(\Omega_2), \quad 0 < \varepsilon \leqq 1,$$

the Hilbert space

$$H := \{v = (v_1, v_2) \mid v_i \in H^1(\Omega_i),\ i = 1, 2,\ \alpha v_1 - v_2 = 0, \text{ on } \Gamma\},$$

with the inner product

$$(u, v)_{1,\Omega} := \sum_{i=1}^{2} (u_i, v_i)_{1,\Omega_i},$$

and the set

$$H^* := \{v = (v_1, v_2) \mid v_i \in H^1(\Omega_i),\ i = 1, 2,\ \alpha v_1 - v_2 = g_1 \text{ on } \Gamma\}.$$

In this section we will consider the following problem:

$$\begin{aligned} &\text{Find } u_\varepsilon = (u_{1\varepsilon}, u_{2\varepsilon}) \in H^* \quad \text{such that} \\ &a_\varepsilon(u_\varepsilon, v) = F(v) \quad \forall v \in H. \end{aligned} \tag{4.1}$$

We will prove that (4.1) has a unique solution $u_\varepsilon \in H^*$ and, with respect to some suitable weighted norm, the limit

$$\lim_{\varepsilon \to 0^+} u_\varepsilon = u \tag{4.2}$$

exists. The limit function u will turn out to be the solution of the transmission problem (2.9) and will be seen to possess certain integrability properties. Finally, we introduce the bounded linear operator $A_\varepsilon : H \to H$ as follows:

$$(A_\varepsilon u, v)_{1,\Omega} = a_\varepsilon(u, v) \quad \forall u, v \in H. \tag{4.3}$$

To settle the existence and uniqueness of solutions to (4.1), we shall need the following lemmas.

LEMMA 4.1. *If either* $\operatorname{Re}(\bar{\alpha}\beta) > 0$ *or* $\operatorname{Im}(\bar{\alpha}\beta) \neq 0$, *then there is a constant* $C > 0$ *such that for every* $u \in H$

$$\|u\|_{1,\Omega} \leqq C(\|A_\varepsilon u\|_{1,\Omega} + (1+\varepsilon)\|u\|_{0,\Omega}), \tag{4.4}$$

where C *is independent of* ε, u, *and* A_ε.

Proof. Let $a = \operatorname{Re}(\bar{\alpha}\beta)$ and $b = \operatorname{Im}(\bar{\alpha}\beta)$. We distinguish the following two cases:

Case 1. $a > 0$. Since

$$\begin{aligned} \bar{\alpha}\beta \|u\|_{1,\Omega}^2 = (A_\varepsilon u, u)_{1,\Omega} &+ \bar{\alpha}\beta \int_{\Omega_1} (1 + k^2 - q_1)|u_1|^2\, dx - \bar{\alpha}\beta \int_{\Omega_1} B_1 \nabla u_1 \bar{u}_1\, dx \\ &+ (\bar{\alpha}\beta - 1) \int_{\Omega_2} |\nabla u_2|^2\, dx + \int_{\Omega_2} (\bar{\alpha}\beta + k^2 + i\varepsilon - q_2)|u_2|^2\, dx \\ &- \int_{\Omega_2} B_2 \nabla u_2 \bar{u}_2\, dx, \end{aligned}$$

by taking the real part in the above equality and using Cauchy's and Young's inequalities, we obtain

$$\|u\|_{1,\Omega}^2 \leqq C(|(A_\varepsilon u, u)_{1,\Omega}| + \|u\|_{0,\Omega}^2). \tag{4.5}$$

The estimate (4.4) follows from (4.5) and Cauchy's inequality.

Case 2. $b \neq 0$. From the equality

$$\begin{aligned}\|u\|_{1,\Omega}^2 = (A_\varepsilon u, u)_{1,\Omega} + (1-\bar{\alpha}\beta)\int_{\Omega_1} |\nabla u_1|^2\,dx + \int_{\Omega_1} (1+\bar{\alpha}\beta k^2 - \bar{\alpha}\beta q_1)|u_1|^2\,dx \\ -\bar{\alpha}\beta\int_{\Omega_1} B_1\nabla u_1 \bar{u}_1\,dx + \int_{\Omega_2}(1+k^2+i\varepsilon-q_2)|u_2|^2\,dx \\ -\int_{\Omega_2} B_2\nabla u_2\bar{u}_2\,dx,\end{aligned} \tag{4.6}$$

we obtain, by multiplying the imaginary part by $b^{-1}a$, adding the result to the real part of (4.6), and using the Young's inequality, that

$$\|u\|_{1,\Omega}^2 \leqq C[|(A_\varepsilon u, u)_{1,\Omega}| + (1+\varepsilon)\|u\|_{0,\Omega}^2], \tag{4.7}$$

which gives the estimate (4.4).

LEMMA 4.2. *If either* $\operatorname{Re}(\bar{\alpha}\beta) > 0$ *or* $\operatorname{Im}(\bar{\alpha}\beta) \neq 0$, *then for any* $R \geqq R_0$ *there exists a constant* $C > 0$ *such that for every* $u \in H$

$$\|u\|_{1,B(R)} \leqq C(\|A_\varepsilon u\|_{1,B(R+1)} + (1+\varepsilon)\|u\|_{0,B(R+1)}), \tag{4.8}$$

where C *is independent of* ε, u, *and* A_ε.

Proof. For any $R \geqq R_0$ let $\phi(r) \in \mathscr{D}(\mathbf{R}^n)$ be such that

$$\phi(r) = \begin{cases} 1, & r \leqq R, \\ 0, & r \geqq R+1, \end{cases} \quad 0 \leqq \phi(r) \leqq 1, \quad |\partial_r \phi| \leqq C\phi^{1/2}, \quad R < r < R+1, \tag{4.9}$$

where C is a constant independent of R, r, and ϕ. From (4.5) and (4.7) we have

$$\|\phi u\|_{1,B(R+1)}^2 \leqq C[|(A_\varepsilon(\phi u), \phi u)_{1,B(R+1)}| + (1+\varepsilon)\|\phi u\|_{0,B(R+1)}^2]. \tag{4.10}$$

By making use of (4.3), (4.10), Cauchy's inequality and a direct calculation, the required inequality (4.8) follows from (4.10).

Remark 4.1. The same arguments used to prove (4.4) and (4.8) can also be used to obtain similar results when A_ε is replaced by $A_\varepsilon + \lambda I$ with $\operatorname{Re}\lambda \geqq 0$ and $\operatorname{Im}\lambda \geqq 0$, where I denotes the identity operator on H.

To deal with nonhomogeneous data we reduce (4.1) to an equivalent problem involving homogeneous data as follows. First, we note that for $g_1 \in H^{1/2}(\Gamma)$ there is a $w_0 \in H^1(\Omega_1)$ such that $w_0 = g_1$ on Γ and $\|w_0\|_{1,\Omega_1} \leqq C\|g_1\|_{1/2,\Gamma}$ [17, p. 141]. We then set

$$u'_{1\varepsilon} = u_{1\varepsilon} - w_0/\alpha$$

and

$$F'(v) = F(v) - (\bar{\alpha}\beta/\alpha)a_1(w_0, v_1) \quad \forall v = (v_1, v_2) \in H.$$

Consequently, (4.1) is equivalent to the following problem:

$$\begin{aligned}&\text{Find } u'_\varepsilon = (u'_{1\varepsilon}, u'_{2\varepsilon}) \in H \quad \text{such that} \\ &a_\varepsilon(u'_\varepsilon, v) = F'(v) \quad \forall v \in H\end{aligned} \tag{4.11}$$

and

$$u_\varepsilon = (u_{1\varepsilon}, u_{2\varepsilon}) = (u'_{1\varepsilon} + w_0/\alpha, u'_{2\varepsilon}). \tag{4.12}$$

If u'_ε are solutions of (4.11) with $f_2 \in L_2(\Omega_2)$, then it follows from the fact that $|F'(v)| \leqq C(\sum_{j=1}^{2} \|f_j\|_{0,\Omega_j} + \|g_1\|_{1/2,\Gamma} + \|g_2\|_{-1/2,\Gamma})\|v\|_{1,\Omega}$ for all $v \in H$, the definition of A_ε, (4.3), and (4.11) that

$$\|A_\varepsilon u'_\varepsilon\|_{1,\Omega} \leqq C\left(\sum_{j=1}^{2} \|f_j\|_{0,\Omega_j} + \|g_1\|_{1/2,\Gamma} + \|g_2\|_{-1/2,\Gamma}\right). \tag{4.13}$$

To prove that (4.1), or equivalently, (4.11), has a unique solution and that (4.2) exists in an appropriate weighted Sobolev space, we need the following a priori estimate for solutions to (4.1).

LEMMA 4.3. *Let $u_\varepsilon \in H^*$, $0 < \varepsilon \leqq 1$, be a solution of* (4.1) *with $f_2 \in L^2_\rho(\Omega_2)$. If either* $\operatorname{Re}(\bar{\alpha}\beta) > 0$ *or* $\operatorname{Im}(\bar{\alpha}\beta) \neq 0$, *then*

$$\begin{aligned} &\|\nabla(e^{-ikr}u_\varepsilon)\|_{0,\Omega} + \|u_\varepsilon\|_{0,P_\delta,\Omega} \\ &\qquad \leqq C\left(\sum_{j=1}^{2} \|f_j\|_{0,\rho,\Omega_j} + \|g_1\|_{1/2,\Gamma} + \|g_2\|_{-1/2,\Gamma} + \|u_\varepsilon\|_{0,B(R_2)}\right) \end{aligned} \tag{4.14}$$

for some fixed $R_2 > R_0$ and $C > 0$ independent of ε, u_ε, f_j, and g_j, $j = 1, 2$.

Proof. We first note that, by coerciveness estimates [14, Thm. 2.7], we have

$$u_2 \in H^2(B(R) \cap \Omega_2) \quad \forall R > 0$$

and

$$(-\Delta + B_2\nabla + q_2 - k^2 - i\varepsilon)u_2 = f_2 \quad \text{a.e. in } \Omega_2.$$

Lemma 4.3 is proved in several steps following [15] and [23]. For $R_1 > R_0$ set $\phi \in C^\infty(\mathbf{R}^n)$ be such that $\phi(x) = 0$, $|x| \leqq R_1$, $\phi(x) = 1$, $|x| \geqq 2R_1$, and $0 \leqq \phi \leqq 1$, $R_1 < |x| < 2R_1$. Let $w = (w_1, w_2) = e^{-ikr}\phi(u_{1\varepsilon}, u_{2\varepsilon})$. Then $w \in H$ and by a direct calculation we find that w_2 satisfies

$$(A_1 + A_2 + A_3 + B_2\nabla)w_2 = \tilde{f}, \tag{4.15}$$

where $A_1 = -\Delta$, $A_2 = -ik(2\partial_r + (n-1)/r)$, $A_3 = q_2 + ikB_2\hat{x} - i\varepsilon$, $\tilde{f} = e^{-ikr}[-\Delta(\phi u_{2\varepsilon}) + B_2\nabla(\phi u_{2\varepsilon}) + q_2\phi u_{2\varepsilon} - (k^2 + i\varepsilon)\phi u_{2\varepsilon}]$.

Step 1. We show that there exists a constant $C_1 > 0$, independent of ε, w_2, $\tilde{f}$, and R, such that for $R > R_1$

$$\begin{aligned} &k\|\nabla w_2\|^2_{0,\Omega(R)} + \varepsilon\|\sqrt{r}\,\nabla w_2\|^2_{0,\Omega(R)} \\ &\quad \leqq C_1\{(\ln R_1)^{\delta/2}\|\tilde{f}\|^2_{0,\rho,\Omega(R)} + (\ln R_1)^{-\delta/2}(\|\nabla w_2\|^2_{0,\Omega(R)} + \|w_2\|^2_{0,P_\delta,\Omega(R)}) \\ &\qquad + \varepsilon(\ln R_1)^{-\delta/2}(\|\sqrt{r}\,\nabla w_2\|^2_{0,\Omega(R)} + \|w_2\|^2_{0,q_\delta,\Omega(R)})\} + K_R(w_2), \end{aligned} \tag{4.16}$$

where $K_R(w_2)$ denotes a functional satisfying

$$|K_R(w_2)| \leqq CR(\|\nabla w_2\|^2_{0,s(R)} + \|w_2\|^2_{0,s(R)}).$$

Step 2. We show that for $R_2 > R_1$, arbitrarily, the following estimate holds:

$$\begin{aligned} &k\|w_2\|^2_{0,P_\delta,\Omega(R_1,R_2)} + (\varepsilon/\delta)\|w_2\|^2_{0,q_\delta,\Omega(R_1,R_2)} \\ &\quad \leqq C_1\{\|\tilde{f}\|^2_{0,\rho,\Omega(R_1,R_2)} + (\ln R_1)^{-\delta/2}(\|\nabla w_2\|^2_{0,\Omega(R_1,R_2)} + \|w_2\|^2_{0,P_\delta,\Omega(R_1,R_2)})\} \\ &\qquad + (\varepsilon/\delta)(\ln R_2)^{-\delta}\|w_2\|^2_{0,\Omega(R_1,R_2)}, \end{aligned} \tag{4.17}$$

where $C_1 > 0$ is independent of ε, w_2, $\tilde{f}$, R_1, and R_2.

The estimates (4.16) and (4.17) can be easily proved by using the same arguments as used in [15] and (2.3), (2.4), and (2.5).

Step 3. We now combine Step 1 and Step 2 to obtain the required estimate (4.14). By (4.16) and (4.17) we know, for any $R_2 > R_1 \geqq R_0$, that

$$\begin{aligned} & k\|\nabla w_2\|^2_{0,\Omega(R_1,R_2)} + \varepsilon\|\sqrt{r}\,\nabla w_2\|^2_{0,\Omega(R_1,R_2)} \\ & \quad + (\varepsilon/\delta)\|w_2\|^2_{0,q_\delta,\Omega(R_1,R_2)} + k\|w_2\|^2_{0,P_\delta,\Omega(R_1,R_2)} \\ & \leqq C\{(\ln R_1)^{\delta/2}\|\tilde{f}\|^2_{0,\rho,\Omega(R_1,R_2)} \\ & \quad + (\ln R_1)^{-\delta/2}(\|\nabla w_2\|^2_{0,\Omega(R_1,R_2)} + \|w_2\|^2_{0,P_\delta,\Omega(R_1,R_2)}) \\ & \quad + \varepsilon(\ln R_1)^{-\delta/2}(\|\sqrt{r}\,\nabla w_2\|^2_{0,\Omega(R_1,R_2)} + \|w_2\|^2_{0,q_\delta,\Omega(R_1,R_2)})\} \\ & \quad + \varepsilon\delta^{-1}(\ln R_2)^{-\delta}\|w_2\|^2_{0,\Omega(R_1,R_2)} + K_{R_2}(w_2), \end{aligned}$$

which implies that, for $R_1 \geqq R_0$ sufficiently large,

$$\begin{aligned} & \|\nabla w_2\|^2_{0,\Omega(R_1,R_2)} + \|w_2\|^2_{0,P_\delta,\Omega(R_1,R_2)} \\ & \quad \leqq C(\|\tilde{f}\|^2_{0,\rho,\Omega(R_1,R_2)} + (\ln R_2)^{-\delta}\|w_2\|^2_{0,\Omega(R_1,R_2)}) + K_{R_2}(w_2). \end{aligned} \tag{4.18}$$

Since $|K_{R_2}(w_2)| \leqq CR_2(\|w_2\|^2_{0,s(R_2)} + \|\nabla w_2\|^2_{0,s(R_2)})$ and $w_2 \in H^1(\Omega_2)$, there exists a sequence $\{R_{2j}\}_{j=1}^{\infty}$ such that $R_{2j} \to \infty$ and $K_{R_{2j}}(w_2) \to 0$ as $j \to \infty$. From this, the definition of ϕ, (4.18), and (4.15), we obtain by taking $R_2 = R_{2j}$ and passing to the limit in (4.18) as $j \to \infty$ that for some fixed $R_1 > 0$

$$\|\nabla(e^{-ikr}\phi u_{2\varepsilon})\|_{0,E(R_1)} + \|\phi u_{2\varepsilon}\|_{0,P_\delta,E(R_1)} \leqq C\|\tilde{f}\|_{0,\rho,E(R_1)}.$$

From this and by making use of the definition of ϕ and $\tilde{f}$, we arrive at the following inequality:

$$\|\nabla(e^{-ikr}u_{2\varepsilon})\|_{0,E(2R_1)} + \|u_{2\varepsilon}\|_{0,P_\delta,E(2R_1)} \leqq C(\|f_2\|_{0,\rho,E(R_1)} + \|u_{2\varepsilon}\|_{1,\Omega(R_1,2R_1)}). \tag{4.19}$$

On the other hand, since in any bounded region the weighted norms are equivalent to similar norms with unit weight, by Lemma 4.2 and (4.12), (4.13) we get

$$\begin{aligned} & \|\nabla(e^{-ikr}u_\varepsilon)\|_{0,B(2R_1)} + \|u_\varepsilon\|_{0,P_\delta,B(2R_1)} + \|u_{2\varepsilon}\|_{1,\Omega(R_1,2R_1)} \\ & \quad \leqq C\|u_\varepsilon\|_{1,B(2R_1)} \\ & \quad \leqq C\left[\sum_{j=1}^{2}\|f_j\|_{0,\rho,\Omega_j} + \|g_1\|_{1/2,\Gamma} + \|g_2\|_{-1/2,\Gamma} + \|u_\varepsilon\|_{0,B(2R_1+1)}\right]. \end{aligned} \tag{4.20}$$

Then (4.19) and (4.20) imply the required estimate (4.14).

With Lemma 4.3 we can now establish the principle of limiting absorption in the following form.

THEOREM 4.1. *Let* $f_j \in L^2_\rho(\Omega_j)$, $j = 1, 2$, *and either* $\operatorname{Re}(\bar{\alpha}\beta) > 0$ *or* $\operatorname{Im}(\bar{\alpha}\beta) \neq 0$. *Let* u_ε, $0 < \varepsilon \leqq 1$, *be a solution to* (4.1). *Then there exists a* $u \in H^*_E \cap L^2_{P_\delta}(\Omega)$ *such that* $\lim_{\varepsilon\to 0^+} u_\varepsilon = u$ *weakly in* $L^2_{P_\delta}(\Omega)$ *and* u *is the unique solution to the transmission problem* (2.9). *Moreover,* u *satisfies the estimate*

$$\|\nabla(e^{-ikr}u)\|_{0,\Omega} + \|u\|_{0,P_\delta,\Omega} \leqq C\left(\sum_{j=1}^{2}\|f_j\|_{0,\rho,\Omega_j} + \|g_1\|_{1/2,\Gamma} + \|g_2\|_{-1/2,\Gamma} + \|u\|_{0,B(R_2)}\right) \tag{4.21}$$

where C *is a positive constant independent of* $u, f_j, g_j, j = 1, 2$ *and* R_2 *is some fixed number chosen as in Lemma* 4.3.

Proof. We distinguish the following two cases.

Case 1. $\operatorname{Sup}_\varepsilon \|u_\varepsilon\|_{0,B(R_2)} < \infty$.

By Lemma 4.3 we find

$$\sup_{\varepsilon} [\|\nabla(e^{-ikr}u_\varepsilon)\|_{0,\Omega} + \|u_\varepsilon\|_{0,P_\delta,\Omega}] < \infty. \tag{4.22}$$

Define the Hilbert space Ξ as follows:

$$\Xi := \{v \in L^2(B(R) \cap \Omega_1) \times L^2(B(R) \cap \Omega_2) \ \forall R > 0 \,|\, \|\nabla(e^{-ikr}v)\|_{0,\Omega} + \|v\|_{0,P_\delta,\Omega} < \infty\},$$

with the norm

$$\|v\|_\Xi := [\|\nabla(e^{-ikr}v)\|^2_{0,\Omega} + \|v\|^2_{0,P_\delta,\Omega}]^{1/2}.$$

Then, by the reflexive property of Hilbert spaces and (4.22) we can find a $u \in \Xi \subset L^2_{P_\delta}(\Omega)$ and a subsequence which we simply denote by $\{u_\varepsilon\}_{0<\varepsilon\leqq 1}$ such that u_ε converges weakly to u in Ξ. This implies that u_ε converges weakly to u in $L^2_{P_\delta}(\Omega)$ and that u_ε converges weakly to u in $L_2(B(R))$ with $R \geqq 0$, arbitrarily. On the other hand, from (4.22) we see $\sup_\varepsilon \|u_\varepsilon\|_{0,B(R)} < \infty$ for arbitrary $R \geqq 0$, which, together with Lemma 4.2, gives $\sup_\varepsilon \|u_\varepsilon\|_{1,B(R)} < \infty$ for arbitrary $R \geqq 0$. Rellich's selection theorem asserts that there is a subsequence $\{u_{\varepsilon_i}\}_{i=1}^\infty$ and a $v \in L_2(B(R))$ such that u_{ε_i} converges strongly to v in $L_2(B(R))$ for arbitrary $R \geqq 0$. Hence $v = u$ in $L_2(B(R))$ for arbitrary $R \geqq 0$. Lemma 4.2 yields that u_{ε_i} also converges strongly to u in $H^1(B(R))$ for arbitrary $R \geqq 0$. Combining this with (4.1), we obtain that $u \in H^*_E \cap L^2_{P_\delta}(\Omega)$ is the unique solution to (2.9). By the weak lower semicontinuity of $\|v\|_\Xi$ on Ξ, we get

$$\|u\|_\Xi \leqq \liminf_{i\to\infty} \|u_{\varepsilon_i}\|_\Xi.$$

Then, from (4.14) we obtain that u satisfies the estimate (4.21). Furthermore, the same argument as above and Theorem 3.1 imply that the original sequence $\{u_\varepsilon\}_{0<\varepsilon\leqq 1}$ weakly converges to u in $L^2_{P_\delta}(\Omega)$.

Case 2. $\operatorname{Sup}_\varepsilon \|u_\varepsilon\|_{0,B(R_2)} = \infty$.

Let $v_\varepsilon = \|u_\varepsilon\|^{-1}_{0,B(R_2)} u_\varepsilon$. Then v_ε satisfies (4.1) with f_j and g_j replaced by $f_{j\varepsilon} = \|u_\varepsilon\|^{-1}_{0,B(R_2)} f_j$ and $g_{j\varepsilon} = \|u_\varepsilon\|^{-1}_{0,B(R_2)} g_j$, respectively, $j = 1, 2$, and $\|v_\varepsilon\|_{0,B(R_2)} = 1$, where $f_{j\varepsilon}$ and $g_{j\varepsilon}$, $j = 1, 2$, tend to zero as $\varepsilon \to 0$. The same argument as in Case 1 can be used to prove that there is a $v \in L^2_{P_\delta}(\Omega)$ such that v_ε converges to v weakly in $L^2_{P_\delta}(\Omega)$, where v is a solution of the homogeneous problem corresponding to (2.9) and $\|v\|_{0,B(R_2)} = 1$. By Theorem 3.1 the solution of this homogeneous problem is unique, and therefore $v = 0$ almost everywhere in Ω. This is a contradiction. The proof is complete.

To prove the existence of solutions to the transmission problem we have to prove that at least for $\varepsilon > 0$ sufficiently small (4.1) has indeed a unique solution. With the help of Theorems 3.1 and 4.1, (4.11), and (4.12), this can be established as follows.

THEOREM 4.2. *Let $f_j \in L_2(\Omega_j)$, $j = 1, 2$ and either* $\operatorname{Re}(\bar{\alpha}\beta) > 0$ *or* $\operatorname{Im}(\bar{\alpha}\beta) \neq 0$. *Then there is an $\varepsilon_0 > 0$ such that for $0 < \varepsilon \leqq \varepsilon_0$ problem* (4.1) *has a unique solution $u_\varepsilon \in H^*$, satisfying*

$$\|u_\varepsilon\|_{1,\Omega} \leqq C(\varepsilon)\left(\sum_{j=1}^{2} \|f_j\|_{0,\Omega_j} + \|g_1\|_{1/2,\Gamma} + \|g_2\|_{-1/2,\Gamma}\right). \tag{4.23}$$

Proof. (a) *Uniqueness.* Consider the following homogeneous problem corresponding to (4.1):

$$\begin{aligned} &\text{Find } v_\varepsilon = (v_{1\varepsilon}, v_{2\varepsilon}) \in H \quad \text{such that} \\ &a_\varepsilon(v_\varepsilon, v) = 0 \quad \forall v \in H. \end{aligned} \tag{4.24}$$

To prove the uniqueness of solutions to (4.1) it is sufficient to prove that there is an $\varepsilon_0 > 0$ such that for $0 < \varepsilon \leqq \varepsilon_0$ problem (4.24) has only a trivial solution. If this is false,

then there will be a subsequence $\{v_{\varepsilon_i}\}$ such that $\varepsilon_i \to 0$ as $i \to \infty$, $v_{\varepsilon_i} \neq 0$ almost everywhere in Ω, and v_{ε_i} is a solution to (4.24) with ε replaced by ε_i. We now see that $\|v_{\varepsilon_i}\|_{0,B(R_2)} \neq 0$, because otherwise we get $\|v_{\varepsilon_i}\|_{0,P_\delta,\Omega} = 0$ from Lemma 4.3, which implies that $v_{\varepsilon_i} = 0$ almost everywhere in Ω. Thus, we can assume that $\|v_{\varepsilon_i}\|_{0,B(R_2)} = 1$. Theorem 3.1 and the same argument as used in Case 1 in the proof of Theorem 4.1 imply that v_{ε_i} converges to zero in $L_2(B(R_2))$. This contradicts the fact that $\|v_{\varepsilon_i}\|_{0,B(R_2)} = 1$, $i = 1, 2, \cdots$. Uniqueness is proved.

(b) *Existence.* Using (4.3) and the uniqueness property we find that the kernel of the operator A_ε satisfies $\operatorname{Ker}(A_\varepsilon) = \{0\}$. On the other hand, since A_ε is a bounded linear operator, $A_\varepsilon + \lambda I$ is invertible for all $\lambda \in C$ with $|\lambda|$ sufficiently large. Therefore, if we can also prove that $A_\varepsilon + \lambda I$ is a semi-Fredholm operator for all $\lambda \in C$ with $\operatorname{Re} \lambda \geqq 0$ and $\operatorname{Im} \lambda \geqq 0$, well-known results from perturbation theory imply that $A_\varepsilon + \lambda I$ is not only semi-Fredholm but also Fredholm with index zero for all $\lambda \in C$ with $\operatorname{Re} \lambda \geqq 0$ and $\operatorname{Im} \lambda \geqq 0$ [8, Thm. 5.17]. In particular, if $\lambda = 0$, we will obtain that A_ε is a Fredholm operator with index zero, which, together with $\operatorname{Ker}(A_\varepsilon) = \{0\}$ implies that A_ε is bijective from H onto H. Therefore, (4.11) has a unique solution. Thus by (4.11) and (4.12), (4.1) also has a unique solution and the estimate (4.23) follows from (4.12) and (4.13). Hence, we only have to prove that $A_\varepsilon + \lambda I$ is a semi-Fredholm operator with $\operatorname{Re} \lambda \geqq 0$ and $\operatorname{Im} \lambda \geqq 0$, that is, that $R(A_\varepsilon + \lambda I) = \overline{R(A_\varepsilon + \lambda I)}$ and the dimension of $\operatorname{Ker}(A_\varepsilon + \lambda I)$ is finite. To this end we first derive a coerciveness estimate. From (4.3) we find for $0 < t \leqq 1$ and $u \in H$ that

$$
\begin{aligned}
\sqrt{2}|((A_\varepsilon + \lambda I)u, u)_{1,\Omega}| &\geqq t|\operatorname{Re}((A_\varepsilon + \lambda I)u, u)_{1,\Omega}| + |\operatorname{Im}((A_\varepsilon + \lambda I)u, u)_{1,\Omega}| \\
&\geqq t\|\nabla u_2\|^2_{0,\Omega_2} - tk^2\|u_2\|^2_{0,\Omega_2} + \varepsilon\|u_2\|^2_{0,\Omega_2} - C\|u_1\|^2_{1,\Omega_1} \\
&\quad - |(B_2 \nabla u_2, u_2)_{0,\Omega_2}| - |(q_2 u_2, u_2)_{0,\Omega_2}|.
\end{aligned} \tag{4.25}
$$

Since by (2.4) we have

$$
\begin{aligned}
&|(B_2 \nabla u_2, u_2)_{0,\Omega_2}| + |(q_2 u_2, u_2)_{0,\Omega_2}| \\
&\quad \leqq CR_1^{-1}(\ln R_1)^{-1-\delta}(\|\nabla u_2\|^2_{0,E(R_1)} + \|u_2\|^2_{0,E(R_1)}) + C\|u_2\|^2_{1,D(R_1)},
\end{aligned}
$$

it follows from (4.25) that there is an $R_1 > 0$ sufficiently large and a $t > 0$ sufficiently small such that

$$
\|u\|_{1,\Omega} \leqq C(\varepsilon, R_1)[\|(A_\varepsilon + \lambda I)u\|_{1,\Omega} + \|u\|_{1,B(R_1)}] \quad \forall u \in H. \tag{4.26}
$$

On the other hand, by Remark 4.1 we have

$$
\|u\|_{1,B(R_1)} \leqq C[\|(A_\varepsilon + \lambda I)u\|_{1,\Omega} + (1+\varepsilon)\|u\|_{0,B(R_1+1)}] \quad \forall u \in H. \tag{4.27}
$$

Now combine (4.26) and (4.27) to obtain the coerciveness estimate

$$
\|u\|_{1,\Omega} \leqq C(\varepsilon)[\|(A_\varepsilon + \lambda I)u\|_{1,\Omega} + \|u\|_{0,B(R_1+1)}] \quad \forall u \in H. \tag{4.28}
$$

From (4.28) and Rellich's selection theorem it follows that $\dim \operatorname{Ker}(A_\varepsilon + \lambda I)$ is finite.

To prove that $R(A_\varepsilon + \lambda I)$ is closed, we consider the operator $T: H/\operatorname{Ker}(A_\varepsilon + \lambda I) \to H$, defined as follows:

$$
T[u] := (A_\varepsilon + \lambda I)u \quad \forall u \in H,
$$

where $H/\operatorname{Ker}(A_\varepsilon + \lambda I)$ denotes the quotient space with the norm $\|[u]\| = \inf\{\|u - v\|_{1,\Omega} | v \in \operatorname{Ker}(A_\varepsilon + \lambda I)\}$ and $[u] \in H/\operatorname{Ker}(A_\varepsilon + \lambda I)$, an equivalence class such that for arbitrary $u', u'' \in [u]$ we have $u' - u'' \in \operatorname{Ker}(A_\varepsilon + \lambda I)$. Evidently, $R(T) = R(A_\varepsilon + \lambda I)$ and T is a bounded linear operator satisfying $\operatorname{Ker}(T) = \{[0]\}$; that is, T

is invertible. Consequently, to prove $R(A_\varepsilon+\lambda I)=\overline{R(A_\varepsilon+\lambda I)}$ it is enough to prove that T^{-1} is continuous, that is,

$$\|[u]\| \leqq C\|T[u]\|_{1,\Omega} \quad \forall u \in H. \tag{4.29}$$

However, by (4.28) we know that

$$\|u-v\|_{1,\Omega} \leqq C(\varepsilon)(\|(A_\varepsilon+\lambda I)u\|_{1,\Omega}+\|u-v\|_{0,B(R_1+1)}) \quad \forall v \in \operatorname{Ker}(A_\varepsilon+\lambda I), \quad u \in H,$$

which, together with Rellich's selection theorem and the fact that $T[u]=(A_\varepsilon+\lambda I)u$ and $\operatorname{Ker}(T)=\{[0]\}$, implies that

$$\|u-v\|_{1,\Omega} \leqq C(\varepsilon)\|(A_\varepsilon+\lambda I)u\|_{1,\Omega} \quad \forall v \in \operatorname{Ker}(A_\varepsilon+\lambda I), \quad u \in H.$$

From this, (4.29) follows. Hence, the proof is complete.

As a direct consequence of Theorems 4.1 and 4.2, we obtain the existence of solutions to the transmission problem (2.9).

THEOREM 4.3. *Let* $f_j \in L^2_\rho(\Omega_j)$, $j=1,2$ *and either* $\operatorname{Re}(\bar{\alpha}\beta)>0$ *or* $\operatorname{Im}(\bar{\alpha}\beta)\neq 0$. *Then* (2.9) *has a unique solution* $u \in H^*_E \cap L^2_{P_\delta}(\Omega)$ *such that*

$$\|\nabla(e^{-ikr}u)\|_{0,\Omega}+\|u\|_{0,P_\delta,\Omega} \leqq C\left(\sum_{j=1}^{2}\|f_j\|_{0,\rho,\Omega_j}+\|g_1\|_{1/2,\Gamma}+\|g_2\|_{-1/2,\Gamma}+\|u\|_{0,B(R_2)}\right). \tag{4.30}$$

Remark 4.2. If $g_1 \in H^{3/2}(\Gamma)$ and $g_2 \in H^{1/2}(\Gamma)$, then, by coerciveness estimates [14, Thm. 2.7], we find that the transmission solution $u=(u_1,u_2)$ has the regularity that $u_j \in H^2(B(R)\cap\Omega_j)$ for all $R>0$, $j=1,2$.

Remark 4.3. For exterior boundary value problems, that is, Dirichlet, Neumann, and Robin problems, our methods are still applicable and similar results hold.

Remark 4.4. If the Laplace operator Δ is replaced by the operator $\sum_{k,j=1}^{n}(\partial/\partial x_k)\times(a_{kj}^{(i)}\partial/\partial x_j)$, $i=1,2$, with $a_{kj}^{(i)}$ satisfying the following conditions:

(1) The $a_{kj}^{(i)}$ are real-valued functions, $a_{kj}^{(i)}=a_{jk}^{(i)}$, $a_{kj}^{(i)} \in C^1(\Omega_i)$, $i=1,2$;

(2) For all $x \in \Omega_i$ and for all $\xi \in \mathbf{R}^n$ the following inequality holds:

$$\sum_{k,j=1}^{n} a_{kj}^{(i)}(x)\xi_k\xi_j \geqq C_0|\xi|^2,$$

where C_0 is a positive constant;

(3) There is a $R_1 \geqq R_0$ such that $a_{kj}^{(i)}(x)=\delta_{kj}$ for $|x| \geqq R_1$, $i=1,2$, $k,j=1,\cdots,n$;

(4) In (2.1) the transmission condition $\beta\partial u_1/\partial\nu-\partial u_2/\partial\nu=g_2$ on Γ is replaced by the condition $\beta a_{kj}^{(1)}\nu_k\partial u_1/\partial x_j-a_{kj}^{(2)}\nu_k\partial u_2/\partial x_j=g_2$ on Γ.

Then Theorems 3.1, 4.1, and 4.3 still hold.

Acknowledgment. The authors thank the referee for his constructive comments on this problem and for his pointing out an error in the proof of Lemma 4.3 in the original version of this paper.

REFERENCES

[1] P. BECK, *Ein Eindeutigkeitssatz für Strahlungslösungen eines Übergangsproblems zur Schwingungsgleichung mit unbeschränkter Trennflache*, Math. Methods Appl. Sci., 7 (1985), pp. 290–308.

[2] D. M. EIDUS, *The principle of limiting absorption*, Amer. Math. Soc. Transl., (2) 47 (1965), pp. 157–191.

[3] ———, *The principle of limiting amplitude*, Russian Math. Surveys, 24 (1969), pp. 97–167.

[4] ———, *On the spectra and eigenfunctions of the Schrödinger and Maxwell operators*, J. Math. Anal. Appl., 106 (1985), pp. 540–568.

[5] W. JAGER, *Zur Theorie der Schwingungsgleichung mit variablen Koeffizienten in Aussengebieten*, Math. Z., 102 (1967), pp. 62–88.

[6] D. S. JONES, *A uniqueness theorem in elastodynamics*, Quart. J. Mech. Appl. Math., 37 (1984), pp. 121-141.

[7] T. KATO, *Growth properties of solutions of the reduced wave equation with a variable coefficient*, Comm. Pure Appl. Math., 12 (1959), pp. 403-425.

[8] ———, *Perturbation Theory of Linear Operators*, Springer-Verlag, Berlin, New York, 1966.

[9] R. KITTAPPA AND R. E. KLEINMAN, *Acoustic scattering by penetrable homogeneous objects*, J. Math. Phys., 16 (1975), pp. 421-432.

[10] R. E. KLEINMAN AND P. A. MARTIN, *On single integral equations for the transmission problem of acoustics*, SIAM J. Appl. Math., 48 (1988), pp. 307-325.

[11] R. KRESS AND G. F. ROACH, *Transmission problems for the Helmholtz equation*, J. Math. Phys., 19 (1978), pp. 1433-1437.

[12] G. KRISTENSSON, *A uniqueness theorem for Helmholtz equation: penetrable media with an infinite interface*, SIAM J. Math. Anal., 11 (1980), pp. 1104-1117.

[13] W. D. KUPRADZE, *Randwertaufgaben der Schwingungstheorie und Integralgleichungen*, Deutscher Verlag der Wissenschaften, Berlin, 1956.

[14] R. LEIS, *Initial Boundary Value Problems in Mathematical Physics*, John Wiley, Chichester, 1986.

[15] P. NEITTAANMAKI AND G. F. ROACH, *Weighted Sobolev spaces and exterior problems for the Helmholtz equation*, Proc. Roy. Soc. London Ser. A, 410 (1987), pp. 373-383.

[16] F. M. ODEH, *Uniqueness theorems for the Helmholtz equation in domains with infinite boundaries*, J. Math. Mech., 12 (1963), pp. 857-868.

[17] J. T. ODEN AND J. N. REDDY, *An Introduction to the Mathematical Theory of Finite Elements*, John Wiley, New York, 1976.

[18] A. G. RAMM, *Scattering by a penetrable body*, J. Math. Phys., 25 (1984), pp. 469-471.

[19] F. RELLICH, *Über das asymptotische Verhalten der Lösungen von* $\Delta u + \lambda u = 0$ *in unendlichen Gebieten*, J. Deut. Math. Ver., 53 (1943), pp. 56-64.

[20] J. SARANEN, *Ausstrahlungsproblem des Laplace-operators für eine Klasse der Gebiete mit unbeschränktem Rand*, Manuscripta Math., 20 (1977), pp. 355-376.

[21] B. R. VAINBERG, *Principle of radiation, limit absorption and limit amplitude in general theory of partial differential equations*, Russian Math. Surveys, 21 (1966), pp. 115-193.

[22] I. N. VEKUA, *New Methods in Solving Elliptic Boundary Value Problems*, North-Holland, Amsterdam, 1967.

[23] V. VOGELSANG, *Das Ausstrahlungsproblem für elliptische Differentialgleichungen in Gebieten mit unbeschränktem Rand*, Math. Z., 144 (1975), pp. 101-124.

[24] J. WEIDMANN, *Linear Operators in Hilbert Spaces*, Springer-Verlag, Berlin, New York, 1980.

[25] P. WERNER, *Zur mathematischen Theorie akustischen Wellenfelder*, Arch. Rational Mech. Anal., 6 (1960), pp. 231-260.

[26] ———, *Beugungsprobleme der mathematischen Akustik*, Arch. Rational Mech. Anal., 12 (1963), pp. 155-184.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1007-1015, July 1991

009

NONTRIVIAL SOLUTIONS TO NONLINEAR VOLTERRA INTEGRAL EQUATIONS*

W. OKRASIŃSKI†

Abstract. The Volterra integral equation

$$u(x)=\int_0^x k(x-s)g(u(s))\,ds \qquad (x\geqq 0)$$

is considered, where $k\geqq 0$ is a monotonic integrable function and g is an increasing, continuous function such that $g(0)=0$. Some necessary and sufficient conditions for the existence of nontrivial solutions to the above equation are presented. Physical problems that motivate these results are also discussed.

Key words. nonlinear Volterra equation, existence of nontrivial solutions

AMS(MOS) subject classifications. 45D05, 45G10

1. Introduction and statement of results. We consider the nonlinear Volterra integral equation

$$u(x)=\int_0^x k(x-s)g(u(s))\,ds \qquad (x\geqq 0), \tag{1.1}$$

where

(k) $k:(0,\delta\rangle\to(0,+\infty)$ $(\delta>0)$ is a monotonic absolutely continuous function such that $\int_0^\delta k(s)\,ds<\infty$,

(g) g is a nondecreasing absolutely continuous function such $g(0)=0$, $g(u)>0$ for $u>0$ and $u/g(u)\to 0$ as $u\to 0+$.

Equation (1.1) has been studied recently in connection with nonlinear diffusion and shock-wave propagation problems (see the Appendix). A typical example of g considered in applications is $g(u)=u^p$ $(p\in(0,1))$. Obviously, $u\equiv 0$ is the trivial solution to (1.1). But the question of physical interest is the existence of nontrivial solutions to (1.1), i.e., continuous functions u such that $u(x)>0$ for $x>0$.

The purpose of this paper is to present some sufficient and necessary conditions for the existence of nontrivial solutions to (1.1) on an interval $\langle 0,\delta_1\rangle$ $(\delta_1>0)$. We formulate the results which will be proved in this paper. Let K^{-1} denote the inverse function to $K(x)\doteq\int_0^x k(s)\,ds$. At first, we present two theorems involving sufficient conditions for the existence of nontrivial solutions to (1.1).

THEOREM 1.1. *Suppose k is an increasing function satisfying* (k). *Let g satisfy* (g). *If*

$$\int_0^{\delta_0}[g'(s)/g(s)]K^{-1}(s/g(s))\,ds<\infty \qquad (\delta_0>0), \tag{1.2}$$

then equation (1.1) *has a nontrivial solution on an interval* $\langle 0,\delta_1\rangle$ $(\delta_1>0)$.

THEOREM 1.2. *Suppose k is a decreasing function satisfying* (k) *such that* $\ln k$ *is convex. Let g satisfy* (g). *If*

$$\int_0^{\delta_0}[g(s)k\circ K^{-1}(s/g(s))]^{-1}\,ds<\infty \qquad (\delta_0>0), \tag{1.3}$$

then equation (1.1) *has a nontrivial solution on an interval* $\langle 0,\delta_1\rangle$ $(\delta_1>0)$.

* Received by the editors January 2, 1990; accepted for publication (in revised form) June 11, 1990.
† Institute of Mathematics, University of Wrocław, Pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland.

We now present two necessary conditions for the existence of nontrivial solutions to (1.1).

THEOREM 1.3. *Suppose k is an increasing function satisfying* (k) *such that* $\ln k$ *is concave. Let g satisfy* (g). *If equation* (1.1) *has a nontrivial solution on an interval* $\langle 0, \delta_1\rangle$ $(\delta_1 > 0)$, *then*

$$\int_0^{\delta_0} [g(s)k \circ K^{-1}(s/g(s))]^{-1}\, ds < \infty \qquad (\delta_0 > 0). \tag{1.4}$$

THEOREM 1.4. *Suppose k is a decreasing function satisfying* (k). *Let g satisfy* (g). *If equation* (1.1) *has a nontrivial solution on an interval* $\langle 0, \delta_1\rangle$ $(\delta_1 > 0)$, *then*

$$\int_0^{\delta_0} [g'(s)/g(s)]K^{-1}(s/g(s))\, ds < \infty \qquad (\delta_0 > 0). \tag{1.5}$$

Proofs of the theorems presented above are given in § 3.

In [5] Gripenberg considered the following special case of (1.1):

$$u(x) = \int_0^x (x-s)^{\alpha-1} g(u(s))\, ds \qquad (\alpha > 0), \tag{1.6}$$

where

(i) $g(u)/u$ is continuous positive and nonincreasing on $(0, a\rangle$ $(a > 0)$,

(ii) for each $q > 0$ the function $u[g(u)/u]^q$ is nondecreasing on $\langle 0, a_q\rangle$ $(a_q > 0)$.

In [5] it is shown that $u \equiv 0$ is the unique solution to (1.6) if and only if

$$\int_0^{\delta_0} [s[g(s)/s]^{1/\alpha}]^{-1}\, ds = \infty \qquad (\delta_0 > 0). \tag{1.7}$$

Let us note that for $\alpha = 1$ from (1.7) we obtain the famous Osgood condition. On the basis of Gripenberg's results it can be concluded that (1.6) has a nontrivial solution if and only if

$$\int_0^{\delta_0} [s[g(s)/s]^{1/\alpha}]^{-1}\, ds < \infty \qquad (\delta_0 > 0). \tag{1.8}$$

It is easy to see that in the case where $\lim_{u\to 0+} u/g(u) > 0$, equation (1.6) only has the trivial solution. We must emphasize that under Gripenberg's assumptions the case where $g(u) = u^p$ $(p \in (0, 1))$ is not allowed. Results similar to Gripenberg's are presented in [2], [3], and [10] under weaker assumptions of g and such g's as mentioned above are admissible. In § 2 of this paper we show that Theorems 1.1–1.4 generalize condition (1.8).

2. Some consequences and comments. On the basis of Theorem 1.1 we can formulate the corollary.

COROLLARY 2.1. *Suppose the assumptions of Theorem* 1.1 *and* (i) *are satisfied. If*

$$\int_0^{\delta_0} K^{-1}(s/g(s))/s\, ds < \infty, \tag{2.1}$$

then equation (1.1) *has a nontrivial solution on an interval* $\langle 0, \delta_1\rangle$ $(\delta_1 > 0)$.

By (g) *and* (i) *we have*

$$g'(s) \leqq g(s)/s \quad \text{a.e.} \tag{2.2}$$

By (2.2) *we obtain*

$$[g'(s)/g(s)]K^{-1}(s/g(s)) \leqq K^{-1}(s/g(s))/s \quad \text{a.e.} \tag{2.3}$$

By (2.1) and (2.3) we infer (1.2) is fulfilled. The corollary is true.

Now we show Corollary 2.2.

COROLLARY 2.2. *Suppose the assumptions of Theorem* 1.4 *and* (i) *are satisfied. Moreover, there exists a number* $q_0 > 1$ *such that* $u[g(u)/u]^{q_0}$ *is nondecreasing. If equation* (1.1) *has a nontrivial solution on an interval* $\langle 0, \delta_1\rangle$ $(\delta_1 > 0)$, *then*

$$\int_0^{\delta_0} K^{-1}(s/g(s))/s \, ds < \infty. \tag{2.4}$$

As in [5] it can be shown that

$$g'(s) \geqq (1 - 1/q_0)g(s)/s \quad \text{a.e.} \tag{2.5}$$

By (1.5) and (2.5) we infer (2.4) is satisfied.

Now we present results concerning the special kernel $k(x) = x^{\alpha-1}$ $(\alpha > 0)$.

Remark 2.1. Let us note that for the kernel $k(x) = x^{\alpha-1}$ $(\alpha > 0)$ conditions (1.3), (1.4), (2.1), and (2.4) are equivalent to Gripenberg's condition (1.8).

We can generalize Gripenberg's results as follows.

COROLLARY 2.3. *Let* $\alpha > 0$. *Suppose g satisfies* (g), (i) *and there exists a number* $q_0 > 1$ *such that* $u[g(u)/u]^{q_0}$ *is nondecreasing. Equation* (1.6) *has a nontrivial solution if and only if* (1.8) *is fulfilled.*

This corollary is a consequence of Theorems 1.2–1.3, Corollaries 2.1–2.2, and Remark 2.1.

Remark 2.2. Function $g(u) = u^p$ $(p \in (0, 1))$ satisfies the assumptions of Corollary 2.3.

In [3] it is shown that the equation

$$u(x) = \int_0^x \exp(-1/(x-s)^{\alpha})[u(s)]^p \, ds \qquad (\alpha \geqq 1,\ p \in (0, 1))$$

has a nontrivial solution. In this case our sufficient conditions do not work. This suggests that we must look for other kinds of conditions.

3. Proofs of theorems. We start with two remarks, which are consequences of theorems presented in [2], [7], and [8].

Remark 3.1. Let (k) and (g) be satisfied. If there exists a nontrivial solution to (1.1), then it is the unique nontrivial solution on an interval $\langle 0, \delta_1\rangle$ $(\delta_1 > 0)$. Moreover, it is a strictly increasing absolutely continuous function. We shall denote this solution by u_0.

Remark 3.2. Let (k) and (g) be satisfied. Consider the equation

$$u(x) = \varepsilon x + \int_0^x k(x-s)g(u(s)) \, ds. \tag{3.1}$$

For every $\varepsilon \in (0, 1)$ there exists a unique strictly increasing absolutely continuous solution u_ε to (3.1) on an interval $\langle 0, \delta_1\rangle$, where $\delta_1 > 0$ is independent of ε. Moreover, $u_{\varepsilon_1} \leqq u_{\varepsilon_2}$ for $\varepsilon_1 \leqq \varepsilon_2$.

Now we can formulate the following lemma.

LEMMA 3.1. *Let* $\varepsilon \in \langle 0, 1)$. *Let k satisfy* (k) *and g satisfy* (g). *If equation* (3.1) *has a nontrivial solution* u_ε *on* $\langle 0, \delta_1\rangle$ $(\delta_1 > 0)$, *then there exists a function* $v_\varepsilon \in L(0, u_\varepsilon(\delta_1))$

such that $v_\varepsilon > 0$ *almost everywhere and*

$$v_\varepsilon(x) = 1 \Big/ \left(\int_0^x k \int_s^x v_\varepsilon(\xi)\, d\xi\, g'(s)\, ds + \varepsilon \right) \quad \text{a.e.} \tag{3.2}$$

Proof. Let u_ε be the nontrivial solution to (3.1) mentioned in Remarks 3.1 and 3.2. It is an increasing absolutely continuous function. Differentiating (3.1) we get

$$u'_\varepsilon(x) = \int_0^x k(x-s) g'(u_\varepsilon(s)) u'_\varepsilon(s)\, ds + \varepsilon \tag{3.3}$$

for almost all $x \in \langle 0, \delta_1 \rangle$. By assumptions and (3.3) we have $u'_\varepsilon > 0$ almost everywhere. We infer u_ε^{-1} is an absolutely continuous function. From (3.3) we get

$$u'_\varepsilon(x) = \varepsilon + \int_0^{u_\varepsilon(x)} k(x - u_\varepsilon^{-1}(s)) g'(s)\, ds \quad \text{a.e.}$$

Substituting $u_\varepsilon^{-1}(x)$ instead of x, we obtain

$$u'_\varepsilon(u_\varepsilon^{-1}(x)) = \varepsilon + \int_0^x k(u_\varepsilon^{-1}(x) - u_\varepsilon^{-1}(s)) g'(s)\, ds \quad \text{a.e.} \tag{3.4}$$

Let $v_\varepsilon \doteq [u_\varepsilon^{-1}]$ almost everywhere. Since $v_\varepsilon = 1/u'_\varepsilon \circ u_\varepsilon^{-1}$ almost everywhere and $u_\varepsilon^{-1}(x) - u_\varepsilon^{-1}(s) = \int_s^x v_\varepsilon(\xi)\, d\xi$, by (3.4) we get (3.2).

COROLLARY 3.1. *If* u_ε *is the nontrivial solution mentioned in Remarks* 3.1 *and* 3.2, *then we put* $v_\varepsilon \doteq [u_\varepsilon^{-1}]'$ *almost everywhere. Since* u_ε^{-1} *is absolutely continuous,* $u_\varepsilon^{-1}(x) = \int_0^x v_\varepsilon(s)\, ds$.

COROLLARY 3.2. *Let* $\varepsilon > 0$. *If the integrable function* v_ε *satisfies* (3.2), *then*

$$\int_0^x K\left(\int_s^x v_\varepsilon(\xi)\, d\xi \right) g'(s)\, ds \leqq x. \tag{3.5}$$

Let $\varepsilon = 0$. *If the integrable function* v_0 *satisfies* (3.2), *then*

$$\int_0^x K\left(\int_s^x v_0(\xi)\, d\xi \right) g'(s)\, ds = x. \tag{3.6}$$

From (3.2) *we have*

$$v_\varepsilon(x) \int_0^x k\left(\int_s^x v_\varepsilon(\xi)\, d\xi \right) g'(s)\, ds + \varepsilon v_\varepsilon(x) = 1 \quad \text{a.e.} \tag{3.7}$$

Integrating (3.7) *we get*

$$\int_0^x K\left(\int_s^x v_\varepsilon(\xi)\, d\xi \right) g'(s)\, ds + \varepsilon \int_0^x v_\varepsilon(s)\, ds = x. \tag{3.8}$$

From (3.8) *we obtain* (3.5) *and* (3.6).

Proof of Theorem 1.1. Consider (3.1). Let $\{u_\varepsilon\}$ $(\varepsilon \in (0, 1))$ denote the family of increasing absolutely continuous solutions to (3.1) on $\langle 0, \delta_1 \rangle$ mentioned in Remark 3.2. By Corollary 3.1 the function $v_\varepsilon \doteq [u_\varepsilon^{-1}]'$ almost everywhere on $\langle 0, u_\varepsilon(\delta_1) \rangle$ satisfies (3.5). Since K is convex then from (3.5) by the Jensen inequality, we get

$$g(x) K\left(\int_0^x \int_s^x v_\varepsilon(\xi)\, d\xi\, g'(s)\, ds / g(x) \right) \leqq x$$

or after simple calculations

$$V_\varepsilon(x) \leqq K^{-1}(x/g(x)), \tag{3.9}$$

where

$$V_\varepsilon(x) \doteq \int_0^x v_\varepsilon(s)g(s)\,ds/g(x).$$

We have

$$V'_\varepsilon(x) = v_\varepsilon(x) - [g'(x)/g(x)]V_\varepsilon(x) \quad \text{a.e.} \tag{3.10}$$

From (3.9) and (3.10) we get

$$V'_\varepsilon(x) \geqq v_\varepsilon(x) - [g'(x)/g(x)]K^{-1}(x/g(x)) \quad \text{a.e.} \tag{3.11}$$

From (3.11) we obtain

$$v_\varepsilon(x) \leqq V'_\varepsilon(x) + [g'(x)/g(x)]K^{-1}(x/g(x)) \quad \text{a.e.} \tag{3.12}$$

Since $u_\varepsilon^{-1}(x) = \int_0^x v_\varepsilon(s)\,ds$ and (1.2) holds, integrating (3.12) we get

$$u_\varepsilon^{-1}(x) \leqq V_\varepsilon(x) + \int_0^x [g'(s)/g(s)]K^{-1}(s/g(s))\,ds.$$

Hence by (3.9) we obtain

$$u_\varepsilon^{-1}(x) \leqq K^{-1}(x/g(x)) + \int_0^x [g'(s)/g(s)]K^{-1}(s/g(s))\,ds. \tag{3.13}$$

Let

$$F^{-1}(x) \doteq K^{-1}\left(\sup_{s\in\langle 0,x\rangle}(s/g(s))\right) + \int_0^x [g'(s)/g(s)]K^{-1}(s/g(s))\,ds + x.$$

It is a strictly increasing continuous function such that $F^{-1}(0) = 0$. By (3.13) we have

$$u_\varepsilon^{-1}(x) \leqq F^{-1}(x) \quad \text{on } \langle 0, u_\varepsilon(\delta_1)\rangle$$

or

$$u_\varepsilon(x) \geqq F(x) \quad \text{on } \langle 0, \delta_1\rangle. \tag{3.14}$$

Let $\varepsilon \searrow 0+$. Since the sequence u_ε is decreasing with respect to ε, we infer that

$$u(x) \doteq \lim_{\varepsilon\searrow 0+} u_\varepsilon(x) \qquad (x \in \langle 0, \delta_1\rangle)$$

is a nondecreasing continuous solution to (1.1). Since (3.14) holds for any $\varepsilon \in (0, 1)$, $u(x) \geqq F(x)$ on $\langle 0, \delta_1\rangle$. It means we have found a nontrivial solution u to (1.1).

Proof of Theorem 1.2. Consider (3.1). Let $\{u_\varepsilon\}$ $(\varepsilon \in (0, 1))$ denote the family of increasing absolutely continuous solutions to (3.1) on $\langle 0, \delta_1\rangle$ mentioned in Remark 3.2. By Corollary 3.1 and (3.2) we infer that the function $v_\varepsilon \doteq [u_\varepsilon^{-1}]'$ almost everywhere on $\langle 0, u_\varepsilon(\delta_1)\rangle$ must satisfy the following inequality:

$$v_\varepsilon(x) \leqq 1 \Big/ \int_0^x k\left(\int_s^x v_\varepsilon(\xi)\,d\xi\right) g'(s)\,ds \quad \text{a.e.} \tag{3.15}$$

We can write (3.15) as

$$v_\varepsilon(x) \leqq 1 \Big/ \int_0^x k \circ K^{-1}\left(K\left(\int_s^x v_\varepsilon(\xi)\,d\xi\right)\right) g'(s)\,ds \quad \text{a.e.} \tag{3.16}$$

Since $\ln k$ is convex it follows that $k \circ K^{-1}$ is convex. By the Jensen inequality from (3.16) we obtain

$$v_\varepsilon(x) \leqq 1/g(x) k \circ K^{-1}\left(\int_0^x K\left(\int_s^x v_\varepsilon(\xi)\, d\xi\right) g'(s)\, ds/g(x)\right) \quad \text{a.e.} \tag{3.17}$$

The function $1/k \circ K^{-1}$ is increasing because k is decreasing and K is increasing. By (3.6) and (3.17) we obtain

$$v_\varepsilon(x) \leqq 1/g(x) k \circ K^{-1}(x/g(x)) \quad \text{a.e. on } \langle 0, u_\varepsilon(\delta_1)\rangle. \tag{3.18}$$

Let

$$G^{-1}(x) \doteq \int_0^x [g(s) k \circ K^{-1}(s/g(s))]^{-1}\, ds. \tag{3.19}$$

By assumption (1.3) the function G^{-1} is well defined. Integrating (3.18), we get

$$u_\varepsilon^{-1}(x) \leqq G^{-1}(x) \quad \text{on } \langle 0, u_\varepsilon(\delta_1)\rangle$$

or

$$u_\varepsilon(x) \geqq G(x) \quad \text{on } \langle 0, \delta_1\rangle. \tag{3.20}$$

Let $\varepsilon \searrow 0+$. Since the sequence u_ε is decreasing with respect to ε, we infer that

$$u(x) \doteq \lim_{\varepsilon \searrow 0+} u_\varepsilon(x) \qquad (x \in \langle 0, \delta_1\rangle)$$

is a nondecreasing continuous solution to (1.1). Since (3.20) holds for any $\varepsilon \in (0, 1)$, $u(x) \geqq G(x)$ on $\langle 0, \delta_1\rangle$. It means that the function u is a nontrivial solution to (1.1).

Proof of Theorem 1.3. Let u_0 be the nontrivial solution to (1.1) mentioned in Remark 3.1. Let $v_0 \doteq [u_0^{-1}]'$ almost everywhere on $\langle 0, u_0(\delta_1)\rangle$. By Corollary 3.1 and (3.2) we have

$$v_0(x) = 1 \Big/ \int_0^x k \circ K^{-1}\left(K\left(\int_s^x v_0(\xi)\, d\xi\right)\right) g'(s)\, ds \quad \text{a.e.} \tag{3.21}$$

Since $\ln k$ is concave it follows that $k \circ K^{-1}$ is concave. From (3.21) by the Jensen inequality we get

$$v_0(x) \geqq 1/g(x) k \circ K^{-1}\left(\int_0^x K\left(\int_s^x v_0(\xi)\, d\xi\right) g'(s)\, ds/g(x)\right) \tag{3.22}$$

and by (3.6)

$$v_0(x) \geqq 1/g(x) k \circ K^{-1}(x/g(x)). \tag{3.23}$$

Since v_0 is integrable, we infer that (1.4) holds.

Proof of Theorem 1.4. Let u_0 be the nontrivial solution to (1.1) mentioned in Remark 3.1. Let $v_0 = [u_0^{-1}]'$ almost everywhere on $\langle 0, u_0(\delta_1)\rangle$. By Corollary 3.1 the function v_0 satisfies (3.6). Since K is convex, then from (3.6) by the Jensen inequality we get

$$x \leqq g(x) K\left(\int_0^x \int_s^x v_0(\xi)\, d\xi\, g'(s)\, ds/g(x)\right),$$

or after simple calculations,

$$V_0(x) \geqq K^{-1}(x/g(x)), \tag{3.24}$$

where

$$V_0(x) \doteq \int_0^x v_0(s)g(s)\,ds/g(x).$$

We have

$$V_0'(x) = v_0(x) - [g'(x)/g(x)]V_0(x) \quad \text{a.e.} \tag{3.25}$$

From (3.24) and (3.25) we get

$$V_0'(x) \leqq v_0(x) - [g'(x)/g(x)]K^{-1}(x/g(x)) \quad \text{a.e.} \tag{3.26}$$

or

$$[g'(x)/g(x)]K^{-1}(x/g(x)) \leqq v_0(x) - V_0'(x) \quad \text{a.e.} \tag{3.27}$$

Since v_0 and V_0' are integrable, then by (3.27) we get (1.5).

Appendix. We present two physical problems leading to nonlinear Volterra integral equations having the form (1.1) and satisfying (k) and (g).

A.1. Travelling wave solutions to shock-wave problems. We consider shock waves in gas-filled shock-wave tubes [6]. We want to find the axial component of the particle velocity behind the shock wave. We assume that the x-axis of the coordinate system is directed along the axis of the tube. Moreover, the shock-wave front passes through the origin of the x-axis at time $t=0$. Let $c_s \geqq c_0$ (c_0 is the sound speed) denote the speed of the shock-wave front. We denote by $v(x, t)$ the axial component of the particle velocity behind the wave front at point x and time t. The function v must satisfy the following equation (see [6]):

$$D_t v + c_s D_x v = -(B_1 v + (c_0 - c_s))D_x v + \frac{1}{2} B_2 \int_0^{t-x/c_s} D_t v(x, t-s)s^{\alpha-1}\,ds, \tag{A1}$$

where B_1, B_2, $\alpha > 0$ are physical parameters, and must also satisfy the condition

$$v(x, x/c_s) = (c_s - c_0)/B_1. \tag{A2}$$

This last condition describes the discontinuity of the axial velocity at the wave front. We look for so-called travelling wave solutions of (A1)-(A2) having the form

$$v(x, t) = v(t - x/c_s). \tag{A3}$$

In the case of such solutions the problem (A1)-(A2) will be reduced to

$$\frac{B_1}{c_s}([v(x) - (c_s - c_0)/B_1]^2)' = B_2 \int_0^x v'(x-s)s^{\alpha-1}\,ds, \tag{A4}$$

where $v = v(x)$ is the unknown function such that

$$v(0) = (c_s - c_0)/B_1. \tag{A5}$$

Substituting into (A4) and (A5) instead of v the function $B_2 c_s[v - (c_s - c_0)/B_1]/B_1$, we get

$$([v(x)]^2)' = \int_0^x v'(x-s)s^{\alpha-1}\,ds \tag{A6}$$

and

$$v(0) = 0. \tag{A7}$$

Integrating (A6) and using (A7), we get

$$[v(x)]^2 = \int_0^x (x-s)^{\alpha-1} v(s)\, ds \tag{A8}$$

or after substituting $v = u^{1/2}$, we obtain

$$u(x) = \int_0^x (x-s)^{\alpha-1} [u(s)]^{1/2}\, ds \tag{A9}$$

for $x \geqq 0$. With respect to physical meaning only nonnegative nontrivial solutions are interesting. Such solutions are considered in [6] and [11].

A.2. Subsolutions of nonlinear diffusion problems. We study the following diffusion problem [1], [9]:

$$D_t h = r^{-1} D_r (r D_r (h^{1/p})) \qquad (p \in (0, 1)) \tag{A10}$$

with conditions

$$h(r, 0) = 0 \quad \text{for } r > 1, \tag{A11}$$

$$h(1, t) = 1 \quad \text{for } t > 0. \tag{A12}$$

Let us note that in the case $p = \frac{1}{2}$ the problem (A10)-(A12) may describe the infiltration of the fluid from a cylindrical reservoir. It may be shown that (A10) with conditions (A11)-(A12) has a unique so-called weak solution $h(r, t)$ in the domain $(1, +\infty) \times (0, +\infty)$ (for details see [4]). It is important for applications that the weak solution is classical at these points (r, t) for which $h(r, t) > 0$. Moreover, it is shown that $\operatorname{supp} h(\cdot, t) = \langle 1, r_0(t) \rangle$ for $t > 0$, where $r_0(t)$ is a continuous increasing function. With respect to applications it is interesting to give even approximate information about the function h. We try to construct an auxiliary function approximating the exact solution h from below. This new function will be a so-called subsolution. We try to construct a subsolution $\underline{h}$ having the form

$$\underline{h}(r, t) = \begin{cases} \dot{A}(t) f(r/[A(t)]^{1/2}, & r \leqq [A(t)]^{1/2}, \\ 0, & r > [A(t)]^{1/2}. \end{cases} \tag{A13}$$

The above function will be a subsolution if there exist both a sufficiently smooth decreasing function f satisfying the problem

$$s^{-1}(s(f^{1/p})')' = -\tfrac{1}{2} s f' \quad \text{for } s \in (0, 1\rangle \tag{A14}$$

with

$$f(1) = 0 \quad \text{and} \quad \lim_{s \to 1-} [f^{1/p}(s)]' = 0 \tag{A15}$$

and the function A satisfying the differential equation

$$\dot{A} f(1/A^{1/2}) = 1 \quad \text{with } A(0) = 1. \tag{A16}$$

To solve (A14)-(A15) we use the substitution

$$f(s) = w(-\log s). \tag{A17}$$

We reduce (A14) and (A15) to

$$(w^{1/p})'' = \tfrac{1}{2} e^{-2x} w' \quad \text{for } x \geqq 0 \tag{A18}$$

with conditions

$$w(0)=0 \quad \text{and} \quad \lim_{x\to 0+}[w^{1/p}(x)]'=0. \tag{A19}$$

Integrating (A18) twice and using (A19), we get

$$[w(x)]^{1/p}=\int_0^x e^{-2s}\left[\frac{1}{2}+x-s\right]w(s)\,ds \qquad (x\geqq 0). \tag{A20}$$

Substituting $e^{-2px/(1-p)}w(x)$ instead of $w(x)$, we get

$$[w(x)]^{1/p}=\int_0^x k(x-s)w(s)\,ds \tag{A21}$$

where

$$k(x)=[\tfrac{1}{2}+x]\,e^{2x/(1-p)}. \tag{A22}$$

We look for continuous solutions w of (A21) such that $w(x)>0$ for $x>0$. We substitute

$$w=u^p \tag{A23}$$

into (A21) to obtain

$$u(x)=\int_0^x k(x-s)[u(s)]^p\,ds \qquad (p\in(0,1)). \tag{A24}$$

Mathematical considerations concerning the existence of nontrivial solutions to (A24), or more exactly to (A20), are presented in [9].

REFERENCES

[1] J. BEAR, *Dynamics of Fluids in Porous Media*, American Elsevier, New York, 1972.

[2] P. J. BUSHELL AND W. OKRASIŃSKI, *Uniqueness of solutions for a class of nonlinear Volterra integral equations with convolution kernel*, Math. Proc. Cambridge Philos. Soc., 106 (1989), pp. 547-552.

[3] ———, *Nonlinear Volterra integral equations with convolution kernel*, J. London Math. Soc., to appear.

[4] J. GONCERZEWICZ, *Porous medium-type equation with irregular boundary data*, in Proc. International Conference, Nancy, March 1988, Pitman Res. Notes in Math. Sci. Series 208, Pitman, Boston, 1989, pp. 59-67.

[5] G. GRIPENBERG, *Unique solutions of some Volterra integral equations*, Math. Scand., 48 (1981), pp. 59-67.

[6] J. J. KELLER, *Propagation of simple nonlinear waves in gas filled tubes with friction*, Z. Angew. Math. Phys., 32 (1981), pp. 170-181.

[7] R. K. MILLER, *Nonlinear Volterra Equations*, W. A. Benjamin, New York, 1971.

[8] W. OKRASIŃSKI, *On a nonlinear Volterra equation*, Math. Methods Appl. Sci., 8 (1986), pp. 345-350.

[9] ———, *On subsolutions of a nonlinear diffusion problem*, Math. Methods Appl. Sci., 11 (1989), pp. 409-416.

[10] ———, *Remarks on nontrivial solutions of a Volterra integral equation*, Math. Methods Appl. Sci., to appear.

[11] W. R. SCHNEIDER, *The general solution of a nonlinear equation of convolution type*, Z. Angew. Math. Phys., 33 (1982), pp. 140-142.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1016-1020, July 1991

010

STABILITY OF TRAVELING WAVEFRONTS FOR THE DISCRETE NAGUMO EQUATION*

B. ZINNER†

Abstract. It has been shown that the discrete Nagumo equation

$$\dot{u}_n = d(u_{n-1} - 2u_n + u_{n+1}) + f(u_n), \qquad n \in \mathbb{Z},$$

has a traveling wavefront solution for sufficiently strong coupling d. In this paper it is shown that such a traveling wavefront is unique (up to a shift in time) and globally stable.

Key words. traveling waves, lower solution technique, myelinated axon, discrete cells

AMS(MOS) subject classifications. 34K20, 35K57

1. Introduction. Consider the infinite system of coupled nonlinear differential equations

$$\dot{u}_n = d(u_{n-1} - 2u_n + u_{n+1}) + f(u_n), \qquad n \in \mathbb{Z} \tag{1}$$

where d is a positive real number.

A typical example for the nonlinearity f is the cubic polynomial $f(x) = x(x-a)(1-x)$, $0 < a < \frac{1}{2}$. Equation (1) is the discrete analogue to the well-known Nagumo equation [5]

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(u). \tag{2}$$

The discrete Nagumo equation is interesting because it has been used to derive (2) [8] and it has also been proposed as a model for conduction in myelinated nerve axons [1]. The continuous Nagumo equation (2) is well studied [4] and it has been proven that there exist globally stable monotone traveling wavefront solutions.

The analytic approach has been less developed for the discrete than for the continuous Nagumo equation. The first results about the discrete Nagumo equation were concerned with threshold properties, that is, conditions forcing nonconvergence to zero of solutions as time approaches infinity, and bounds on the speed of propagation of a "wave of excitation" [1], [2]. The next results were concerned with wave propagation, that is, with solutions of the form

$$u_n(t) = U(n + ct).$$

In particular, failure of propagation for small d and local stability of traveling wavefronts were shown in [6] and [7].

Then traveling wavefronts were analyzed numerically for certain cubic polynomials f [3]. Only recently has the existence of monotone traveling wavefronts for sufficiently large d been proved in [9] and [10]. It is the purpose of this paper to prove that such traveling wavefronts are globally stable.

By a *traveling wavefront* with *velocity* c, $c > 0$, we mean a solution $\{u_n(t)\}_{n=-\infty}^{\infty}$ of (1) for which there exists $U \in C^1(\mathbb{R}, (0,1))$, $U(-\infty) = 0$, $U(\infty) = 1$, such that $u_n(t) = U(n + ct)$ for all $t \in \mathbb{R}$. If in addition $U'(z) > 0$ for all $z \in \mathbb{R}$, then the wavefront is *monotone*. The following theorem allows f to have several zeros in $(0, 1)$ even though

* Received by the editors September 18, 1989; accepted for publication (in revised form) June 21, 1990.
† Division of Mathematics, 120 Mathematics Annex, Auburn University, Auburn, Alabama 36849-5307.

existence of a monotone traveling wavefront has only been shown for the case where f has exactly one zero in $(0, 1)$.

THEOREM 1.1. *Suppose* $f \in C^1([0,1], \mathbb{R})$, *satisfies*

(i) $f(0) = f(1) = 0$, $f'(0) < 0, f'(1) < 0$,

(ii) $f(u) < 0$ *for* $0 < u < \alpha_0$,

(iii) $f(u) > 0$ *for* $\alpha_1 < u < 1$, $0 < \alpha_0 \leqq \alpha_1 < 1$,

and suppose there exists a monotone traveling wavefront $\{v_n(t)\}$, $v_n(t) = U(n + ct)$. *Then for any solution* $\{u_n(t)\}$ *of* (1) *which satisfies*

$$0 \leqq u_n(0) \leqq 1 \quad \textit{for all integers } n, \quad \textit{and}$$

$$\limsup_{n \to -\infty} u_n(0) < \alpha_0 \leqq \alpha_1 < \liminf_{n \to \infty} u_n(0),$$

there exists a constant s *such that*

$$\lim_{t \to \infty} \left(\sup_{n \in \mathbb{Z}} |u_n(t) - U(n + ct - s)| \right) = 0.$$

The following corollary is a direct consequence of Theorem 1.1.

COROLLARY 1.2. *Suppose* $\{u_n(t)\}$ *and* $\{v_n(t)\}$ *are traveling wavefronts of* (1). *Then there exists a constant* t_0 *such that* $\{u_n(t)\} = \{v_n(t - t_0)\}$. *In particular, there is a unique speed* c *for traveling wavefronts.*

2. Proof of the theorem. We will make use of the following lemma [6, Thms. 4.1, 4.2].

LEMMA 2.1. *If the hypotheses of Theorem* 1.1 *hold, then there are constants* z_1, z_2, q_0, μ_0 (*the last two positive*) *such that*

$$U(n + ct - z_1) - q_0 e^{-\mu_0 t} \leqq u_n(t) \leqq U(n + ct - z_2) + q_0 e^{-\mu_0 t}. \tag{3}$$

Furthermore, if there are constants t_0, z_0, *and* ε *for which*

$$\sup_{n \in \mathbb{Z}} |u_n(t_0) - U(n + ct_0 - z_0)| < \varepsilon,$$

then there is a number $\omega(\varepsilon)$ *with* $\lim_{\varepsilon \to 0} \omega(\varepsilon) = 0$ *such that*

$$\sup_{n \in \mathbb{Z}} |u_n(t) - U(n + ct - z_0)| < \omega(\varepsilon) \quad \textit{for all } t \geqq t_0.$$

Lemma 2.1 says that $\{u_n(t)\}$ is "more or less" bounded between two shifted wavefronts and if $\{u_n(t)\}$ is close to a wavefront U at some instant then it will remain close to U. The main idea for the following proof is the attempt to replace the constants z_1, z_2 in (3) by functions $z_1(t)$, $z_2(t)$ with $\lim_{t \to \infty} z_1(t) = \lim_{t \to \infty} z_2(t)$.

Let

$$w_n^s(t) := u_n(t) - U(n + ct - s),$$

$$A := \left\{ s \in [z_2, z_1] : \limsup_{t \to \infty} \left[\sup_{n \in \mathbb{Z}} w_n^s(t) \right] \leqq 0 \right\}, \quad \textit{and}$$

$$B := \left\{ s \in [z_2, z_1] : \limsup_{t \to \infty} \left[\sup_{n \in \mathbb{Z}} -w_n^s(t) \right] \leqq 0 \right\}.$$

Note that $z_2 \in A$, $z_1 \in B$, and Theorem 1.1 will be proved if we show that $A \cap B \neq \varnothing$.

By assumption (i) of Theorem 1.1 there exist positive constants μ and δ_0 such that

$$f(x) - f(y) \leqq -\mu(x - y) \quad \text{for all } x, y \in [0, \delta_0], \quad \text{and} \tag{4a}$$

$$f(x) - f(y) \leqq -\mu(x - y) \quad \text{for all } x, y \in [1 - \delta_0, 1]. \tag{4b}$$

Choose $\delta \in (0, \delta_0)$, arbitrary. Since $U(-\infty)=0$ and $U(\infty)=1$, we may choose $n_0 \in \mathbb{Z}$ and $n_1 \in \mathbb{Z}$ such that

$$U(n_0 - z_2) \leqq \frac{\delta}{2} \quad \text{and} \quad U(n_1 - z_1) \geqq 1 - \frac{\delta}{2}.$$

Now let $I(t) := \{n \in \mathbb{Z}: n_0 - ct \leqq n \leqq n_1 - ct\}$,

$$A_f := \left\{ s \in [z_2, z_1]: \limsup_{t\to\infty} \left[\sup_{n \in I(t)} w_n^s(t) \right] \leqq 0 \right\},$$

$$B_f := \left\{ s \in [z_2, z_1]: \limsup_{t\to\infty} \left[\sup_{n \in I(t)} -w_n^s(t) \right] \leqq 0 \right\}.$$

LEMMA 2.2. *$A_f = A$ and $B_f = B$.*

Proof. From the definition of A_f, A, B_f, and B it is clear that $A \subset A_f$ and $B \subset B_f$. It suffices to show $A_f \subset A$, since the proof of $B_f \subset B$ is identical.

So suppose $s \in A_f$. Then

$$\limsup_{t\to\infty} \left[\sup_{n \in I(t)} w_n^s(t) \right] \leqq 0$$

and we have to show that

$$\limsup_{t\to\infty} \left[\sup_{n \in \mathbb{Z}} w_n^s(t) \right] \leqq 0,$$

i.e., given $\varepsilon > 0$ there exists T such that

$$w_n^s(t) \leqq \varepsilon \quad \text{for all } n \in \mathbb{Z}, \quad t \geqq T. \tag{5}$$

Fix $\varepsilon > 0$. Then there exists T_0 such that

$$w_n^s(t) \leqq \varepsilon \quad \text{for all } n \in I(t), \quad t \geqq T_0$$

and

$$u_n(t) \notin [\delta, 1-\delta] \quad \text{for all } n \notin I(t), \quad t \geqq T_0. \tag{6}$$

Let $q \in (0, 1)$ be such that

$$d\left(\frac{1}{q} + \frac{1}{q} - 2\right) \leqq \frac{1}{2}\mu. \tag{7}$$

We will show that the existence of a number $T_k \geqq T_0$, $k \in \mathbb{N}_0$, such that

$$w_n^s(t) \leqq \max\{q^k\delta, \varepsilon\} \quad \text{for all } n \in \mathbb{Z}, \quad t \geqq T_k, \tag{8}$$

will imply the existence of $T_{k+1} \geqq T_0$ such that

$$w_n^s(t) \leqq \max\{q^{k+1}\delta, \varepsilon\} \quad \text{for all } n \in \mathbb{Z}, \quad t \geqq T_{k+1}. \tag{9}$$

Note that (9) only needs to be shown for $n \notin I(t)$. Suppose (8) holds and for some index n we have

$$\max\{q^{k+1}\delta, \varepsilon\} \leqq w_n^s(t) \leqq \max\{q^k\delta, \varepsilon\}, \qquad t \geqq T_k. \tag{10}$$

Since u_n and U satisfy (1) we obtain

$$\dot{w}_n^s = d(w_{n-1}^s - 2w_n^s + w_{n+1}^s) + f(u_n) - f(U), \tag{11}$$

which can be rewritten to

$$\dot{w}_n^s = d\left(\frac{w_{n-1}^s}{w_n^s}+\frac{w_{n+1}^s}{w_n^s}-2\right)w_n^s+\frac{f(u_n)-f(U)}{w_n^s}\,w_n^s. \tag{12}$$

From (7), (8), and (10), we deduce

$$d\left(\frac{w_{n-1}^s}{w_n^s}+\frac{w_{n+1}^s}{w_n^s}-2\right)\leqq\frac{1}{2}\mu \tag{13}$$

and from (4a), (4b), and (6) we get

$$\frac{f(u_n)-f(U)}{w_n^s}\leqq-\mu. \tag{14}$$

Finally, from (12), (13), and (14) we estimate

$$\dot{w}_n^s\leqq-\tfrac{1}{2}\mu w_n^s. \tag{15}$$

With Gronwall's lemma we deduce (9) where T_{k+1} is determined by

$$\exp\left(-\tfrac{1}{2}\mu(T_{k+1}-T_k)\right)=q.$$

The claim (5) then follows by induction. □

Because of Lemma 2.2 it now suffices to show that $A_f\cap B_f\neq\varnothing$. To reach a contradiction suppose $A_f\cap B_f=\varnothing$.

We consider A_f and B_f as topological subspaces of the interval $[z_2, z_1]$. It is easy to check that B_f is closed and therefore we may pick $s\in\partial B_f\backslash A_f$. Since $s\in\partial B_f$ there exist sequences $\{t_k\}_{k=1}^{\infty}$, $\lim_{k\to\infty}t_k=\infty$, $\{n_k\}_{k=1}^{\infty}$, $n_k\in I(t_k)$, such that

$$\lim_{k\to\infty}w_{n_k}^s(t_k)=0. \tag{16}$$

Since $s\notin A_f\cap B_f=A\cap B$ there exists $\varepsilon>0$ such that

$$\sup_{n\in\mathbb{Z}}|w_n^s(t_k)|\geqq\varepsilon \tag{17}$$

in view of Lemma 2.1.

It follows from the definition of $I(t)$ that there exists T_0 such that

$$|w_n^s(t)|<\delta\quad\text{for all } n\notin I(t),\quad t\geqq T_0. \tag{18}$$

Either

$$\begin{aligned}&\lim_{k\to\infty}w_{m_k}^s(t_k)=0\text{ implies }\lim_{k\to\infty}\left(|w_{m_k-1}^s(t_k)|+|w_{m_k+1}^s(t_k)|\right)=0\\&\text{holds for all sequences }\{m_k\}_{k=1}^{\infty},\ m_k\in I(t_k),\end{aligned} \tag{19}$$

or (19) does not hold.

If (19) holds, then it follows from (16) and by induction that $\lim_{k\to\infty}\sup_{n\in I(t_k)}|w_n^s(t_k)|=0$. Together with (18) this implies that there exists $K\in\mathbb{N}$ such that $\sup_{n\in\mathbb{Z}}|w_n^s(t_k)|<\delta$ for all $k\geqq K$. Since δ may be chosen to be less than ε, this contradicts (17).

Therefore (19) does not hold and hence there exist a subsequence of $\{t_k\}$, which we also denote by $\{t_k\}$, a sequence $\{m_k\}$, $m_k\in I(t_k)$, and $\varepsilon_0>0$ such that

$$\lim_{k\to\infty}w_{m_k}^s(t_k)=0\quad\text{and}\quad|w_{m_k-1}^s(t_k)|+|w_{m_k+1}^s(t_k)|\geqq\varepsilon_0. \tag{20}$$

It follows from (20) and $s \in \partial B_f \subset B$ that there exists $K \in \mathbb{N}$ such that

$$-2w^s_{m_k}(t_k) > -\frac{\varepsilon_0}{4},$$

$$w^s_{m_k-1}(t_k) + w^s_{m_k+1}(t_k) > \frac{\varepsilon_0}{2},$$

and

$$f(u_{m_k}(t_k)) - f(U(m_k - ct_k - s)) > -\frac{d\varepsilon_0}{8}$$

holds for all $k \geqq K$. Therefore (11) implies

$$\dot{w}^s_{m_k}(t_k) > \frac{d\varepsilon_0}{8} \quad \text{for all } k \geqq K. \tag{21}$$

It also follows from (11) that there exists $M > 0$ such that

$$\ddot{w}_n(t) < M \quad \text{for all } n \in \mathbb{N} \text{ and } t \geqq 0,$$

which implies together with (21) that there exists $h > 0$ such that

$$\dot{w}_{m_k}(t) > \frac{d\varepsilon_0}{16} \quad \text{for all } t \in [t_k - h, t_k], \quad k \geqq K.$$

Therefore $-w^s_{m_k}(t_k - h) > (hd\varepsilon_0/16) - w^s_{m_k}(t_k)$ for all $k \geqq K$ which implies together with (20) that

$$\limsup_{t \to \infty} \sup_{m \in \mathbb{N}} -w^s_n(t) \geqq \frac{hd\varepsilon_0}{16} > 0,$$

in contradiction to $s \in B_f$.

REFERENCES

[1] J. BELL, *Some threshold results for models of myelinated nerves*, Math. Biosci., 54 (1981), pp. 181–190.

[2] J. BELL AND C. COSNER, *Threshold behaviour and propagation for nonlinear differential-difference systems motivated by modeling myelinated axons*, Quart. Appl. Math., 42 (1984), pp. 1–14.

[3] H. CHI, J. BELL, AND B. HASSARD, *Numerical solution of a nonlinear advance-delay-differential equation from nerve conduction theory*, J. Math. Biol., 24 (1986), pp. 583–601.

[4] P. C. FIFE AND J. B. MCLEOD, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 333–361.

[5] H. K. MCKEAN, *Nagumo's equation*, Adv. in Math., 4 (1970), pp. 209–223.

[6] J. P. KEENER, *Propagation and its failure in the discrete Nagumo equation*, in Proc. Conference on Ordinary and Partial Differential Equations, Dundee, 1986.

[7] ———, *Propagation and its failure in coupled systems of discrete excitable cells*, SIAM J. Appl. Math., 47 (1987), pp. 556–572.

[8] A. C. SCOTT, *Active and Nonlinear Wave Propagation in Electronics*, Wiley-Interscience, New York, 1970.

[9] B. ZINNER, *Traveling wavefront solutions for the discrete Nagumo equation*, Ph.D. thesis, University of Utah, Salt Lake City, UT, 1988.

[10] ———, *Existence of traveling wavefront solutions for the discrete Nagumo equation*, J. Differential Equations, to appear.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1021-1042, July 1991

011

AN OSCILLATION METHOD FOR FOURTH-ORDER, SELF-ADJOINT, TWO-POINT BOUNDARY VALUE PROBLEMS WITH NONLINEAR EIGENVALUES*

LEON GREENBERG†

Abstract. An oscillation method is presented for finding the eigenvalues of a fourth-order, self-adjoint, two-point boundary value problem. The eigenvalue may occur nonlinearly in the differential equation, and may occur in the boundary conditions. The method can approximate the nth eigenvalue without consideration of other eigenvalues. It provides an a posteriori error estimate for the approximate eigenvalue.

Key words. eigenvalue, eigenfunction, self-adjoint, oscillation, energy inner product, Wronskian, Sturm-Liouville equation

AMS(MOS) classification. 65L15

1. Introduction. Physical problems often lead to linear two-point boundary value problems in which the eigenvalue occurs nonlinearly in the differential equation, and may occur in the boundary conditions. Typically, this type of equation arises when we linearize a nonlinear equation and look for normal modes. Examples of such problems occur in acoustics (Porter and Reiss [10]), mechanics of beams (Roseau [12]), and hydrodynamics (Drazin and Reid [5], Chandrasekhar [2], [3]).

A classical example of this kind of problem is a Sturm-Liouville equation

$$\begin{aligned} &(p(x,\lambda)u')'+q(x,\lambda)u=0 \quad \text{for } 0<x<1,\\ &\alpha_0(\lambda)u(0)+\beta_0(\lambda)u'(0)=0,\\ &\alpha_1(\lambda)u(1)+\beta_1(\lambda)u'(1)=0. \end{aligned} \tag{1.1}$$

In [8], Greenberg and Babuška studied the problem (1.1) by means of a function $N(\lambda)$, which is related to the number of oscillations of a solution of the differential equation. Under suitable conditions, (1.1) has $N(\lambda'')-N(\lambda')$ eigenvalues in the interval $[\lambda', \lambda'')$. This enables us to decide whether or not a given interval $[\lambda', \lambda'')$ contains an eigenvalue. Furthermore, if $N(\lambda')\leqq n-1$ and $N(\lambda'')\geqq n$, then the nth eigenvalue satisfies $\lambda'\leqq\lambda_n<\lambda''$. Thus we can determine if λ_n lies in a given interval. By applying the bisection method to $N(\lambda)$, we can approximate the nth eigenvalue without consideration of other eigenvalues, and with an a posteriori error estimate.

In this paper, the above results will be extended to fourth-order boundary value problems. $N(\lambda)$ will be related to the oscillations of the Wronskian of two solutions of the differential equation. In [8], the analysis for the second-order problem was based on the Sturm comparison and oscillation theorems and on Sturm sequences for tridiagonal matrices. These methods are not applicable to the fourth-order problem. An entirely different analysis is given here, based on variational properties of eigenvalues and inertial properties of the energy inner product. It seems likely that these methods can be applied to higher-order problems.

We shall consider two types of fourth-order problems: A scalar equation

$$\text{(I)} \qquad L(u;\lambda)=(p(x,\lambda)u'')''-(q(x,\lambda)u')'+r(x,\lambda)u=0 \quad \text{for } 0<x<1$$

* Received by the editors December 11, 1989; accepted for publication (in revised form) August 2, 1990.
† Mathematics Department, University of Maryland, College Park, Maryland 20742.

with boundary conditions of the form

$$\mathrm{BC}_0(u;\lambda)=\begin{bmatrix}\alpha_{01}(\lambda) & \alpha_{02}(\lambda) & \alpha_{03}(\lambda) & \alpha_{04}(\lambda)\\ \beta_{01}(\lambda) & \beta_{02}(\lambda) & \beta_{03}(\lambda) & \beta_{04}(\lambda)\end{bmatrix}\begin{bmatrix}u(0)\\ u'(0)\\ u''(0)\\ u'''(0)\end{bmatrix}=0$$

and

$$\mathrm{BC}_1(u;\lambda)=\begin{bmatrix}\alpha_{11}(\lambda) & \alpha_{12}(\lambda) & \alpha_{13}(\lambda) & \alpha_{14}(\lambda)\\ \beta_{11}(\lambda) & \beta_{12}(\lambda) & \beta_{13}(\lambda) & \beta_{14}(\lambda)\end{bmatrix}\begin{bmatrix}u(1)\\ u'(1)\\ u''(1)\\ u'''(1)\end{bmatrix}=0;$$

and a vector equation

$$\text{(II)}\qquad L(u;\lambda)=-(P(x,\lambda)u'+Q(x,\lambda)u)'+Q^T(x,\lambda)u'+R(x,\lambda)u=0 \quad \text{for } 0<x<1$$

with boundary conditions of the form

$$\mathrm{BC}_0(u;\lambda)=\begin{bmatrix}\alpha_{01}(\lambda) & \alpha_{02}(\lambda) & \alpha_{03}(\lambda) & \alpha_{04}(\lambda)\\ \beta_{01}(\lambda) & \beta_{02}(\lambda) & \beta_{03}(\lambda) & \beta_{04}(\lambda)\end{bmatrix}\begin{bmatrix}u_1(0)\\ u_2(0)\\ u_1'(0)\\ u_2'(0)\end{bmatrix}=0$$

and

$$\mathrm{BC}_1(u;\lambda)=\begin{bmatrix}\alpha_{11}(\lambda) & \alpha_{12}(\lambda) & \alpha_{13}(\lambda) & \alpha_{14}(\lambda)\\ \beta_{11}(\lambda) & \beta_{12}(\lambda) & \beta_{13}(\lambda) & \beta_{14}(\lambda)\end{bmatrix}\begin{bmatrix}u_1(1)\\ u_2(1)\\ u_1'(1)\\ u_2'(1)\end{bmatrix}=0.$$

In problem (II) $u(x)=(u_1(x), u_2(x))^T$ and $P(x,\lambda)$, $Q(x,\lambda)$, and $R(x,\lambda)$ are 2×2 matrices. (Note that $Q^T(x,\lambda)$ denotes the transpose of $Q(x,\lambda)$.) The boundary conditions are required to be self-adjoint. This will be discussed in § 3.

We shall need to make certain assumptions about the coefficient functions in problems (I) and (II). As indicated below, these occur in several categories. The standard assumptions will always be implicitly assumed. The other assumptions will be explicitly assumed when needed. (The monotonicity and limit assumptions are similar to assumptions in the Sturm oscillation theorem for second-order problems.) The reader may wish to skip over or scan through these assumptions at first. The monotonicity and limit assumptions will only be used in § 4. In the following, λ will vary in an interval (Λ_1, Λ_2). We do not exclude the possibilities $\Lambda_1=-\infty$ or $\Lambda_2=\infty$.

Standard assumptions for problem (I).

(S1) $p(x,\lambda)$, $(\partial/\partial x)p(x,\lambda)$, $(\partial^2/\partial x^2)p(x,\lambda)$, $q(x,\lambda)$, $(\partial/\partial x)q(x,\lambda)$, and $r(x,\lambda)$ are continuous functions on $[0,1]\times(\Lambda_1,\Lambda_2)$.

(S2) $p(x,\lambda)\geqq k>0$ for $0\leqq x\leqq 1$, $\Lambda_1<\lambda<\Lambda_2$.

Standard assumptions for problem (II).

(S1a) $P(x,\lambda)$, $(\partial/\partial x)P(x,\lambda)$, $Q(x,\lambda)$, $(\partial/\partial x)Q(x,\lambda)$, and $R(x,\lambda)$ are continuous matrix functions on $[0,1]\times(\Lambda_1,\Lambda_2)$.

(S1b) $P(x, \lambda) = P^T(x, \lambda)$ and $R(x, \lambda) = R^T(x, \lambda)$ for $0 \leqq x \leqq 1$, $\Lambda_1 < \lambda < \Lambda_2$.

(S2) There is a number $k > 0$ such that for any vector $\xi = (\xi_1, \xi_2)^T$, $\xi^T P(x, \lambda)\xi \geqq k(\xi_1^2 + \xi_2^2)$ for $0 \leqq x \leqq 1$, $\Lambda_1 < \lambda < \Lambda_2$.

Standard assumptions for problems (I) *and* (II).

(S3) The coefficients in the boundary conditions are continuous functions of λ on (Λ_1, Λ_2).

(S4) At each endpoint $x = 0, 1$, the two given boundary conditions are linearly independent for $\Lambda_1 < \lambda < \Lambda_2$.

(S5) Any essential boundary conditions have constant coefficients (see Remark 2, below).

Remark 1. We do not need the full strength of assumption (S1). It would be enough to assume that these functions are piecewise smooth in x (up to the indicated degree of smoothness), and continuous in λ. We assume (S1) for simplicity.

Remark 2. Regarding (S5), recall that an essential boundary condition for (I) involves u and u', but no higher derivatives. An essential boundary condition for (II) involves u_1 and u_2, but no derivatives. The relevant Hilbert space for problem (I) (respectively, (II)) is a subspace of the Sobolev space $H^2[0, 1]$ (respectively, $H^1[0, 1] \times H^1[0, 1]$) that depends on the essential boundary conditions. We have assumed condition (S5) in order to prevent this space from changing as λ varies. This will enable us to apply well-known monotonicity properties of eigenvalues.

The energy inner product for (I) is of the form

$$B(u, v; \lambda) = -\beta_0(u, v; \lambda) + \beta_1(u, v; \lambda) + \int_0^1 [p(x, \lambda)u''v'' + q(x, \lambda)u'v' + r(x, \lambda)uv]\, dx. \tag{1.2}$$

Here the inner product $\beta_i(u, v; \lambda)$ (for $i = 0, 1$) arises from the natural boundary conditions at the endpoint $x = i$, and has the form

$$\beta_i(u, v; \lambda) = a_i(\lambda)u(i)v(i) + b_i(\lambda)(u(i)v'(i) + u'(i)v(i)) + c_i(\lambda)u'(i)v'(i). \tag{1.3}$$

For problem (II), we shall use the "dot product" notation: $\xi \cdot \eta = \xi_1\eta_1 + \xi_2\eta_2$, where $\xi = (\xi_1, \xi_2)^T$ and $\eta = (\eta_1, \eta_2)^T$. The energy inner product for (II) has the form

$$B(u, v; \lambda) = -\beta_0(u, v; \lambda) + \beta_1(u, v; \lambda) + \int_0^1 [P(x, \lambda)u' \cdot v' + Q(x, \lambda)u \cdot v' + Q(x, \lambda)v \cdot u' + R(x, \lambda)u \cdot v]\, dx, \tag{1.4}$$

where

$$\beta_i(u, v; \lambda) = a_i(\lambda)u_1(i)v_1(i) + b_i(\lambda)(u_1(i)v_2(i) + u_2(i)v_1(i)) + c_i(\lambda)u_2(i)v_2(i). \tag{1.5}$$

(These inner products will be discussed in § 3, where we shall classify the boundary conditions.)

Please note that the inner products in this paper are often indefinite. For a linear eigenvalue problem, the number of negative eigenvalues equals the negative index of inertia of the energy inner product $B(u, v)$.

We shall assume monotonicity conditions on the coefficients of (I) and (II) in order to force $B(u, u; \lambda)$ to be a strictly decreasing function of λ for each $u \neq 0$.

Monotonicity assumptions for (I).

(M1) For each x, $r(x, \lambda)$ is a strictly decreasing function of λ.

(M2) For each x, $p(x, \lambda)$ and $q(x, \lambda)$ are nonincreasing functions of λ.

Monotonicity assumptions for (II).

(M1) For each x and for $\xi = (\xi_1, \xi_2)^T \neq 0$, $R(x, \lambda)\xi \cdot \xi$ is a strictly decreasing function of λ.

(M2) For each x and for any vectors $\xi = (\xi_1, \xi_2)^T$, $\eta = (\eta_1, \eta_2)^T$, $P(x, \lambda)\xi \cdot \xi + 2Q(x, \lambda)\eta \cdot \xi + R(x, \lambda)\eta \cdot \eta$ is a nonincreasing function of λ.

Monotonicity assumptions for (I) *and* (II).

(M3) For each vector $\xi = (\xi_1, \xi_2)^T$, the boundary term $\beta_0 = a_0(\lambda)\xi_1^2 + 2b_0(\lambda)\xi_1\xi_2 + c_0(\lambda)\xi_2^2$ is a nondecreasing function of λ.

(M4) For each vector $\xi = (\xi_1, \xi_2)^T$, the boundary term $\beta_1 = a_1(\lambda)\xi_1^2 + 2b_1(\lambda)\xi_1\xi_2 + c_1(\lambda)\xi_2^2$ is a nonincreasing function of λ.

Remark 3. If $A(\lambda)$ is an $n \times n$ symmetric matrix whose coefficients $a_{ij}(\lambda)$ are differentiable functions of λ, and $\xi = (\xi_1, \xi_2, \cdots, \xi_n)^T \neq 0$, then $\xi^T A(\lambda)\xi$ is a strictly decreasing function of λ if $(d/d\lambda)A(\lambda)$ is negative definite. (If $A(\lambda)$ is not differentiable, an analogous condition is that $A(\lambda'') - A(\lambda')$ be negative definite for $\lambda' < \lambda''$.) Suppose that the coefficients in (I) and (II) are differentiable functions of λ. Then (M1) is satisfied for (II) if $(\partial/\partial\lambda)R(x, \lambda)$ is negative definite. (M2) is satisfied for (II) if $(\partial/\partial\lambda)\left[\begin{smallmatrix} P & Q \\ Q^T & R \end{smallmatrix}\right]$ is negative semidefinite. (M3) is satisfied if $(d/d\lambda)\left[\begin{smallmatrix} a_0 & b_0 \\ b_0 & c_0 \end{smallmatrix}\right]$ is positive semidefinite. (M4) is satisfied if $(d/d\lambda)\left[\begin{smallmatrix} a_1 & b_1 \\ b_1 & c_1 \end{smallmatrix}\right]$ is negative semidefinite.

The main result in this paper is Theorem 4.2, which indicates how $N(\lambda)$ can be used to locate eigenvalues. It is an analogue for fourth-order problems of Theorem 2.1 in [8], which deals with second-order problems. The condition (L1) below will be needed to ensure that $B(u, v; \lambda)$ is positive definite when λ is near Λ_1. In a companion paper [7], an existence theorem, similar to the Sturm oscillation theorem, will be proved. The condition (L2) below will be needed to guarantee the existence of infinitely many eigenvalues. For (I), we can place the burden for this either on $r(x, \lambda)$ or $q(x, \lambda)$, and so we shall give two possible conditions for each of these cases. We shall use the following notation:

$$f^*(\lambda) = \max_{0 \leqq x \leqq 1} f(x, \lambda), \qquad f_*(\lambda) = \min_{0 \leqq x \leqq 1} f(x, \lambda). \tag{1.6}$$

Limit assumptions for (I).

(L1) Either (a) $\lim_{\lambda \to \Lambda_1} r_*(\lambda) = \infty$ or (b) $\lim_{\lambda \to \Lambda_1} q_*(\lambda) = \infty$ and there exists λ_+ in (Λ_1, Λ_2) such that $r_*(\lambda_+) > 0$, $a_0(\lambda_+) \leqq 0$, and $a_1(\lambda_+) \geqq 0$.[1]

(L2) Either (a) $\lim_{\lambda \to \Lambda_2} r^*(\lambda) = -\infty$ or (b) $\lim_{\lambda \to \Lambda_2} q^*(\lambda) = -\infty$.

In problem (II), let $R(x, \lambda)$ have eigenvalues $\mu_R(x, \lambda) \leqq \nu_R(x, \lambda)$, and let $\mu_R(\lambda) = \min_{0 \leqq x \leqq 1} \mu_R(x, \lambda)$, $\nu_R(\lambda) = \max_{0 \leqq x \leqq 1} \nu_R(x, \lambda)$. The matrix $\left[\begin{smallmatrix} 0 & Q(x, \lambda) \\ Q^T(x, \lambda) & 0 \end{smallmatrix}\right]$ has eigenvalues $\pm\mu_Q(x, \lambda)$ and $\pm\nu_Q(x, \lambda)$, where $0 \leqq \mu_Q(x, \lambda) \leqq \nu_Q(x, \lambda)$. Let $\nu_Q(\lambda) = \max_{0 \leqq x \leqq 1} \nu_Q(x, \lambda)$.

[1] The functions $a_0(\lambda)$ and $a_1(\lambda)$ in (L1b) are coefficients in the boundary terms $\beta_0(u, v; \lambda)$ and $\beta_1(u, v; \lambda)$ in (1.2) and (1.3).

Limit assumptions for (II).

(L1) $$\lim_{\lambda \to \Lambda_1} \frac{\mu_R(\lambda)}{1+\nu_Q(\lambda)^2} = \infty.$$

(L2) $$\lim_{\lambda \to \Lambda_2} \frac{\nu_R(\lambda)}{1+\nu_Q(\lambda)} = -\infty.$$

Remark 4. For (II), the denominator in (L1) is $1+\nu_Q(\lambda)^2$ rather than $\nu_Q(\lambda)^2$ in order to counteract the effect of small $\nu_Q(\lambda)$, and similarly for (L2). If $\nu_Q(\lambda)$ is bounded, then (L1) may be replaced by $\lim_{\lambda \to \Lambda_1} \mu_R(\lambda) = \infty$, and (L2) replaced by $\lim_{\lambda \to \Lambda_2} \nu_R(\lambda) = -\infty$.

The construction of the function $N(\lambda)$ begins in § 2, where we consider linear eigenvalue problems of the form $Lu = \lambda u$. Here λ does not appear in the coefficients of L or the boundary conditions. A number $N_0(L)$ is defined, using the Wronskian $w(x)$ of two solutions of the differential equation $Lu = 0$. $N_0(L)$ is related to the zeros of $w(x)$. Theorem 2.2 asserts that $N_0(L)$ equals the number of negative eigenvalues of L if L has Dirichlet boundary conditions at $x = 1$. This theorem was proved (for problem (II)) by Morse [9], in the context of extremal curves in the calculus of variations. It is also one of the main theorems in Edwards [6], where it is proved by topological methods. The proof given in § 2 is simpler and more direct than previous proofs. It uses only standard tools from eigenvalue theory.

In order to deal with operators that do not have Dirichlet boundary conditions, a "correction term" $\sigma(L)$ is defined, using an orthogonal decomposition of the energy inner product. The integer $\sigma(L)$ can have values 0, 1, or 2. Theorem 2.5 asserts that the number $N(L) = N_0(L) + \sigma(L)$ equals the number of negative eigenvalues of L. This theorem is due to Edwards [6]. However, it should be mentioned that Edwards never explicitly addresses the boundary conditions, and so $\sigma(L)$ cannot be related to the coefficients of the boundary value problem. Here we emphasize the detailed calculation of formulas, which enables numerical computation.

In § 3 we classify the separated, self-adjoint boundary conditions for (I) and (II) into four types (for each problem). The inner products $\beta_0(u, v)$ and $\beta_1(u, v)$ are calculated for each type of boundary condition. We also calculate the inner product $B_s(u, v)$, which is used to compute $\sigma(L)$.

By its definition in (2.22), $B_s(u, v)$ is an integral. But an important feature of our method is that $B_s(u, v)$ and $\sigma(L)$ depend only on values (of u, v their derivatives, and the coefficients of L) at $x = 1$. In this respect, $\sigma(L)$ is analogous to the term $\sigma(\lambda)$ in [8], for Sturm–Liouville problems. The main purpose of § 3 is to calculate $B_s(u, v)$ and $\sigma(L)$ for the different types of boundary conditions.

In § 4, we turn to the case where the operator L and the boundary conditions depend on λ. Let L_λ denote the operator L for fixed λ, and let $\mu_k(\lambda)$ be the kth eigenvalue of L_λ. The monotonicity assumptions imply that the $\mu_k(\lambda)$ are strictly decreasing functions of λ. The zeros of these functions are the eigenvalues of L. These are the key facts that allow the theorems of § 2 to carry over to the λ-dependent case. We define $N(\lambda) = N(L_\lambda)$, and Theorem 4.2 shows that $N(\lambda)$ has properties similar to those proved in [8] for Sturm–Liouville problems.

In § 5 we discuss properties of the Wronskians that are used to calculate $N(\lambda)$. Our method requires the calculation of $n(x; \lambda) = \text{nullity } W(x; \lambda)$, where $W(x; \lambda)$ is a certain Wronskian matrix. The integers $n(x; \lambda)$ can be 0, 1, or 2, and $n(x_0; \lambda) > 0$ if and only if x_0 is a zero of the Wronskian determinant $w(x) = w(x; \lambda) = \det W(x; \lambda)$. An important fact established in § 5 is that $w(x)$ changes sign at x_0 if $n(x_0; \lambda) = 1$, but

does not change sign if $n(x_0; \lambda) = 2$. This enables us to distinguish the two cases numerically.

2. Linear eigenvalue problems. In this section, we shall consider differential operators L of types (I) and (II) in the case that the coefficients (in L and in the boundary conditions) do not depend on λ. Our main objective in this section is to prove two formulas (Theorems 2.2 and 2.5) for the number of negative eigenvalues of L.

Let $H^k[a, b]$ be the Sobolev space of functions $u(x)$ that have derivatives $u^{(j)}(x) \in L^2[a, b]$ for $0 \leqq j \leqq k$. The basic Hilbert space for (I) is

$$H = \{u \in H^2[0, 1] \mid u \text{ satisfies the essential boundary conditions at } x = 0, 1\}. \tag{2.1}$$

The basic Hilbert space for (II) is

$$H = \{u \in H^1[0, 1] \times H^1[0, 1] \mid u \text{ satisfies the essential boundary conditions at } x = 0, 1\}. \tag{2.2}$$

The energy inner products in these spaces are given in (1.2) and (1.4). (Note that the above spaces are the relevant Hilbert spaces for (I) and (II) whether or not the coefficients depend on λ. The standard assumption (S5) implies that these spaces do not change when λ varies. In the present section, the operator L and the energy inner product B do not depend on λ.) Let $\|u\|$ denote the norm $(\int_0^y u(x)^2\,dx)^{1/2}$ in $L^2[0, y]$. The following estimates are immediate consequences of the fundamental theorem of calculus and Schwarz's inequality.

LEMMA 2.1. *Let $u(x)$ be a function in $H^1[0, y]$ such that $u(x_0) = 0$ for some $x_0 \in [0, y]$. Then*

(1) $u(x)^2 \leqq y\|u'\|^2$ *for* $0 \leqq x \leqq y$, *and*

(2) $\|u\| \leqq y\|u'\|$.

DEFINITION 2.1. Let L be an operator of type (I) or (II), where we assume that the coefficients in L and in the boundary conditions do not depend on λ. For $0 < y \leqq 1$, let $L|_{[0,y]}$ denote the operator L restricted to the interval $[0, y]$, with the given boundary conditions $BC_0(u) = 0$ at $x = 0$, and Dirichlet boundary conditions at $x = y$. (Recall that Dirichlet conditions for problem (I) are $u(y) = u'(y) = 0$, and for problem (II) are $u_1(y) = u_2(y) = 0$.)

THEOREM 2.1. *Let $\lambda_1(y)$ be the smallest eigenvalue of $L|_{[0,y]}$. Then* $\lim_{y \to 0} \lambda_1(y) = \infty$.

Proof. We shall use the formula

$$\lambda_1(y) = \min_{u \in H, u \neq 0} \frac{B(u, u)}{\|u\|^2}, \tag{2.3}$$

where the Hilbert space H, energy inner product B, and norm $\|u\|$ have all been restricted to the interval $[0, y]$ (and $u \in H$ is required to satisfy Dirichlet conditions at $x = y$). Thus, for example, $\|u\|^2 = \int_0^y u(x)^2\,dx$.

For (I), the energy inner product is

$$B(u, u) = -\beta_0(u, u) + \int_0^y [p(x)u''(x)^2 + q(x)u'(x)^2 + r(x)u(x)^2]\,dx, \tag{2.4}$$

where

$$\beta_0(u) = a_0 u(0)^2 + 2b_0 u(0)u'(0) + c_0 u'(0)^2. \tag{2.5}$$

(The term $\beta_y(u, u)$ is absent because we have Dirichlet conditions at $x = y$.) Using the notation $f_* = \min_{0\leqq x\leqq 1} f(x)$, and recalling that $p(x) \geqq k > 0$, we have

$$B(u, u) \geqq -\beta_0(u, u) + k\|u''\|^2 + q_*\|u'\|^2 + r_*\|u\|^2. \tag{2.6}$$

By Lemma 2.1,

$$\|u'\| \leqq y\|u''\|, \qquad \|u\| \leqq y^2\|u''\|, \tag{2.7}$$

$$|\beta_0(u, u)| \leqq (|a_0|y^3 + 2|b_0|y^2 + |c_0|y)\|u''\|^2. \tag{2.8}$$

Therefore,

$$\frac{B(u, u)}{\|u\|^2} \geqq \frac{(k - \alpha y - \beta y^2 - \gamma y^3 - \delta y^4)\|u''\|^2}{y^4\|u''\|^2}, \tag{2.9}$$

where $\alpha = |c_0|$, $\beta = |q_*| + 2|b_0|$, $\gamma = |a_0|$, and $\delta = |r_*|$. This implies that

$$\lambda_1(y) \geqq \frac{(k - \alpha y - \beta y^2 - \gamma y^3 - \delta y^4)}{y^4}, \tag{2.10}$$

and so $\lim_{y\to 0} \lambda_1(y) = \infty$. This proves the theorem for (I).

For (II), the energy inner product is

$$B(u, u) = -\beta_0(u, u) + \int_0^y [P(x)u' \cdot u' + 2Q(x)u \cdot u' + R(x)u \cdot u]\, dx, \tag{2.11}$$

where

$$\beta_0(u, u) = a_0 u_1(0)^2 + 2b_0 u_1(0)u_2(0) + c_0 u_2(0)^2. \tag{2.12}$$

The matrix $\begin{bmatrix} 0 & Q(x) \\ Q^T(x) & 0 \end{bmatrix}$ has eigenvalues $\pm\mu_Q(x)$ and $\pm\nu_Q(x)$, where $0 \leqq \mu_Q(x) \leqq \nu_Q(x)$. Let $\nu_Q = \max_{0\leqq x\leqq 1} \nu_Q(x)$. Then, for any $\varepsilon > 0$,

$$2|Q(x)u \cdot u'| = 2\left| Q(x)\left(\frac{u}{\varepsilon}\right) \cdot (\varepsilon u') \right| \leqq \nu_Q\left(\frac{u \cdot u}{\varepsilon^2} + \varepsilon^2 u' \cdot u'\right).$$

Let $\mu_R(x)$ be the smallest eigenvalue of $R(x)$, and $\mu_R = \min_{0\leqq x\leqq 1} \mu_R(x)$. Then

$$R(x)u \cdot u \geqq \mu_R u \cdot u.$$

Recall that $P(x)u' \cdot u' \geqq ku' \cdot u'$, where $k > 0$. The above remarks imply that

$$B(u, u) \geqq -\beta_0(u, u) + \int_0^y \left[(k - \varepsilon^2\nu_Q)u' \cdot u' + \left(\mu_R - \frac{\nu_Q}{\varepsilon^2}\right) u \cdot u \right] dx. \tag{2.13}$$

Let $\|u\|^2 = \int_0^y u \cdot u\, dx = \|u_1\|^2 + \|u_2\|^2$, and similarly $\|u'\|^2 = \int_0^y u' \cdot u'\, dx = \|u_1'\|^2 + \|u_2'\|^2$. Lemma 2.1 implies that $\|u\|^2 \leqq y^2\|u'\|^2$. Therefore,

$$B(u, u) \geqq -\beta_0(u, u) + \left[(k - \varepsilon^2\nu_Q) - \left| \mu_R - \frac{\nu_Q}{\varepsilon^2} \right| y^2 \right] \|u'\|^2. \tag{2.14}$$

We can estimate $\beta_0(u, u)$ as in (I) to obtain

$$|\beta_0(u, u)| \leqq (|a_0| + 2|b_0| + |c_0|)y\|u'\|^2. \tag{2.15}$$

Therefore,

$$B(u, u) \geqq (\alpha - \beta y - \gamma y^2)\|u'\|^2, \tag{2.16}$$

where $\alpha = k - \varepsilon^2\nu_Q$, $\beta = |a_0| + 2|b_0| + |c_0|$, and $\gamma = |\mu_R - (\nu_Q/\varepsilon^2)$. Now choose ε small enough so that $k - \varepsilon^2\nu_Q > 0$, and the proof concludes as for (I). □

Remark. Inequalities (2.9) and (2.16) show that the Rayleigh quotient has a lower bound, and so the eigenvalues are bounded from below on any interval $[0, y]$. This is a well-known property of elliptic equations.

DEFINITION 2.2. Let L be a differential operator of type (I) or (II) on the interval $[0, 1]$. (We assume that the coefficients in L and the boundary conditions do not depend on λ.) For a given number λ_0, let u, v be linearly independent solutions of $Lz = \lambda_0 z$ that satisfy the boundary conditions at $x = 0$. Let $W(x) = W(x; L, \lambda_0) = W[u, v](x; L, \lambda_0)$ be the Wronskian matrix, and $w(x) = w(x; L, \lambda_0) = w[u, v](x; L, \lambda_0)$ the Wronskian determinant. Thus, for problem (I), $W(x) = \begin{bmatrix} u & v \\ u' & v' \end{bmatrix}$, and for problem (II), $W(x) = \begin{bmatrix} u_1 & v_1 \\ u_2 & v_2 \end{bmatrix}$. In each case, $w(x) = \det W(x)$.

Please note that these Wronskian matrices and determinants have size 2×2. This is in contrast to the standard practice for a fourth-order problem, where four independent solutions are used, so that the Wronskians have size 4×4.

We now define the following functions:

$$n(x; L, \lambda_0) = \text{nullity } W(x) = 2 - \text{rank } W(x), \tag{2.17}$$

$$N_0(L, \lambda_0) = \sum_{0<x<1} n(x; L, \lambda_0), \tag{2.18}$$

$$N_0(L) = N_0(L, 0). \tag{2.19}$$

Remark. The integer $n(x; L, \lambda_0)$ can be 0, 1, or 2. Moreover, there can be only finitely many points $x \in (0, 1)$ such that $n(x; L, \lambda_0) > 0$. (This will be verified in the proof of the following theorem.) Thus, the right-hand side of (2.18) is a finite sum.

THEOREM 2.2. *Let L be a differential operator of type* (I) *or* (II). (*We assume that the coefficients in L and in the boundary conditions do not depend on λ.*) *Suppose that L has Dirichlet boundary conditions at $x = 1$. Then L has exactly $N_0(L, \lambda_0)$ eigenvalues* (*counting multiplicity*) *that are less than λ_0. In particular, L has $N_0(L)$ negative eigenvalues.*

Proof. For $0 < y \leqq 1$, consider the restricted operator $L|_{[0,y]}$ defined in Definition 2.1. Let u, v be linearly independent solutions of $Lz = \lambda_0 z$ that satisfy the boundary conditions at $x = 0$ (as in Definition 2.2). The given number λ_0 is an eigenvalue of $L|_{[0,y]}$ if and only if there exist constants a, b (not both zero) such that $z = au + bv$ satisfies the Dirichlet conditions at $x = y$. Such constants exist if and only if the Wronskian $w[u, v](y; L, \lambda_0) = 0$. The integer $n(y; L, \lambda_0)$ is the multiplicity of the eigenvalue λ_0 for $L|_{[0,y]}$.

Let $\lambda_k(y)$ be the kth eigenvalue of $L|_{[0,y]}$. Then $n(y; L, \lambda_0) > 0$ if and only if $\lambda_k(y) = \lambda_0$ for some k. Furthermore, $n(y; L, \lambda_0)$ is the number of functions $\lambda_k(y)$ (i.e., the number of indices k) such that $\lambda_k(y) = \lambda_0$.

It is well known that $\lambda_k(y)$ is a continuous, strictly decreasing function of y. (See, for example, Weinberger [13].) Therefore, for each k, there can be at most one y such that $\lambda_k(y) = \lambda_0$. Theorem 2.1 implies that all of the functions $\lambda_k(y)$ decrease from ∞ to $\lambda_k(1)$ as y varies from 0 to 1. Therefore there exists $y \in (0, 1)$ such that $\lambda_k(y) = \lambda_0$ if and only if $\lambda_k(1) < \lambda_0$. But the numbers $\lambda_k(1)$ are the eigenvalues of L. Therefore, there are only finitely many indices k such that $\lambda_k(1) < \lambda_0$. This shows that only finitely many terms $n(x; L, \lambda_0)$ in (2.18) can be nonzero. It also shows that $N_0(L, \lambda_0)$ equals the number of eigenvalues $\lambda_k(1)$ that are less than λ_0. □

The formula given in Definition 2.2 and Theorem 2.2 is the first objective of this section. It deals with the case where L has Dirichlet conditions at $x = 1$. Next we shall find a formula that deals with more general boundary conditions. A "correction term" $\sigma(L)$ will be constructed that depends on the boundary conditions at $x = 1$. It will be

shown that the function $N(L) = N_0(L) + \sigma(L)$ counts the negative eigenvalues of L.

Let H be one of the spaces (2.1) or (2.2) (corresponding to (I) or (II)), and let $B(u, v)$ be the energy inner product in H. The negative eigenvalues of L are, of course, related to the inertial properties of $B(u, v)$. This point of view is based on the following theorem, which is easily proved by using eigenfunction expansions. The proof will be omitted. We state the theorem to set the mood for the subsequent discussion.

THEOREM 2.3. *The following numbers are equal*:

(1) *The maximum dimension of a subspace* $V \subset H$ *on which* $B(u, v)$ *is negative definite*;

(2) *The dimension of a maximal subspace* $V \subset H$ *on which* $B(u, v)$ *is negative definite*;

(3) *The number of negative eigenvalues of* L (*counted with multiplicity*).

DEFINITION 2.3. We shall consider the following subspaces of H:

$$H_0 = \{u \in H \mid u \text{ satisfies Dirichlet conditions at } x = 1\}. \tag{2.20}$$

$$S(L) = \{u \in H \mid Lu = 0 \text{ and } u \text{ satisfies the given boundary conditions at } x = 0\}. \tag{2.21}$$

The restrictions of the energy inner product B to these subspaces will be denoted as follows:

$$B_0 = B|_{H_0},\ B_s = B|_{S(L)}. \tag{2.22}$$

We shall also consider the following operator:

(2.23) L_0 is the differential operator that has the same differential expression as L, with the given boundary conditions at $x = 0$, and Dirichlet boundary conditions at $x = 1$.

Remarks. In the notation of Definition 2.1, $L_0 = L|_{[0,1]}$. H_0 is the relevant Hilbert space for L_0 and B_0 is the energy inner product in H_0. The condition in brackets on the right-hand side of (2.21) should be interpreted in the weak sense: $B(u, v) = 0$ for all $v \in H_0$. This forces u to have enough smoothness so that the condition in (2.21) can be interpreted and is valid in the classical sense. Note that the functions in $S(L)$ are required to satisfy the essential boundary conditions at $x = 1$ (if any), since this is part of the definition of H. Thus dim $S(L)$ can be 0, 1, or 2. If zero is not an eigenvalue of L_0, then dim $S(L) = 2 - \nu$, where ν is the number of independent essential boundary conditions at $x = 1$. In this case dim $S(L) =$ codim H_0.

THEOREM 2.4. (a) $S(L)$ *is the subspace of* H *that is orthogonal to* H_0 *with respect to the energy inner product* B.

(b) *If* 0 *is not an eigenvalue of* L_0, *then* $H = H_0 \oplus S(L)$ (*orthogonal decomposition*).

Proof. (a) The equation

$$B(u, v) = 0 \quad \text{for all } v \in H_0 \tag{2.24}$$

is the weak form of

$$Lu = 0 \tag{2.25}$$

and u satisfies the given boundary conditions at $x = 0$.

(b) Since zero is not an eigenvalue of L_0, dim $S(L) =$ codim H_0 (as was remarked above). Furthermore $H_0 \cap S(L) = \{0\}$, because any nontrivial function $u \in H_0 \cap S(L)$ is an eigenfunction of L_0 with eigenvalue zero. This shows that $H = H_0 \oplus S(L)$. □

DEFINITION 2.4. If $F(u, v)$ is an inner product on a space X, the index of F is

$$\text{Ind } F = \sup \{\dim V \mid V \text{ is a subspace of } X \text{ on which } F \text{ is negative definite}\}. \tag{2.26}$$

We now define

$$\sigma(L) = \text{Ind } B_s \tag{2.27}$$

and

$$N(L) = N_0(L) + \sigma(L), \tag{2.28}$$

if zero is not an eigenvalue of L_0;

$$N(L) = \lim_{\varepsilon \to 0^+} N(L+\varepsilon), \tag{2.29}$$

if zero is an eigenvalue of L_0.

Remarks. Theorem 2.2 asserts that Ind $B_0 = N_0(L)$. Note that $\sigma(L)$ can be 0, 1, or 2. One way to calculate $\sigma(L)$ is to represent B_s by a (symmetric) matrix A, with respect to some basis of $S(L)$. Then $\sigma(L)$ is the number of negative eigenvalues of A.

Note that to calculate $N(L+\varepsilon)$, we would find solutions of $Lz = -\varepsilon z$.

THEOREM 2.5. *Let L be a differential operator of type* (I) *or* (II). (*We assume that the coefficients in L and in the boundary conditions do not depend on λ.*) *Then L has exactly $N(L)$ negative eigenvalues* (*counting multiplicity*).

Proof. First suppose that zero is not an eigenvalue of L_0. Then $N(L) = N_0(L) + \sigma(L)$. By Theorem 2.3, the number of negative eigenvalues of L is Ind B. The orthogonal decomposition $H = H_0 \oplus S(L)$ (in Theorem 2.4) implies that Ind $B =$ Ind $B_0 +$ Ind B_s, which equals $N_0(L) + \sigma(L) = N(L)$.

If zero is an eigenvalue of L_0, then it is not an eigenvalue of $L_0 + \varepsilon$ for small ε. Let $\lambda_1, \lambda_2, \cdots, \lambda_n$ be the negative eigenvalues of L. Then $\lambda_1 + \varepsilon, \cdots, \lambda_n + \varepsilon$ are eigenvalues of $L + \varepsilon$, and for small ε they remain negative. Thus Ind $B = n = N(L+\varepsilon)$, for small ε. □

Remark. In the next section we shall express the inner product B_s explicitly in terms of the boundary conditions.

3. Self-adjoint boundary conditions. The separated, self-adjoint boundary conditions for problems (I) and (II) will now be classified. The inner products $\beta_0(u, v)$, $\beta_1(u, v)$ (that occur in (1.2)-(1.4)) and $B_s(u, v)$ (in (2.22)) will be calculated for each type of boundary condition. As in the previous section, we assume that the operator L and the boundary conditions do not depend on λ.

LEMMA 3.1. *Let V be a two-dimensional subspace of $\mathbb{R}^4$ such that*

$$(x_2y_3 - x_3y_2) + (x_4y_1 - x_1y_4) = 0 \tag{*}$$

for all $x, y \in V$. Then V is the solution space of one of the following systems of equations:

(1) $$\begin{cases} x_1 = 0, \\ x_2 = 0; \end{cases}$$

(2) $$\begin{cases} x_1 = 0, \\ cx_2 + x_3 = 0; \end{cases}$$

(3) $$\begin{cases} ax_1 + bx_3 - x_4 = 0, \\ bx_1 - x_2 = 0; \end{cases}$$

$$\begin{cases} ax_1 + bx_2 - x_4 = 0, \\ bx_1 + cx_2 + x_3 = 0. \end{cases} \tag{4}$$

The above coefficients a, b, c can be arbitrary real numbers.

Proof. Using the Euclidean inner product $\langle x, y\rangle = \sum_{i=1}^{4} x_i y_i$ in $\mathbb{R}^4$, let $V^\perp$ be the orthogonal complement of V. Consider the matrix

$$S = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}. \tag{3.1}$$

Note that

$$\langle x, Sy\rangle = (x_2y_3 - x_3y_2) + (x_4y_1 - x_1y_4). \tag{3.2}$$

The equation (*) says that $\langle x, Sy\rangle = 0$ for all $x, y \in V$. This implies that $S(V) = V^\perp$. Since $S^2 = -I$, $S(V^\perp) = S^2(V) = V$. Thus S interchanges V and $V^\perp$. Therefore if $x, y \in V^\perp$, then $S(y) \in V$ and so $\langle x, S(y)\rangle = 0$. This shows that equation (*) is satisfied for all $x, y \in V^\perp$.

Now let $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$, $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^T$ be a basis for $V^\perp$. Then V is the solution space of the system of equations

$$\begin{aligned} &\alpha_1x_1 + \alpha_2x_2 + \alpha_3x_3 + \alpha_4x_4 = 0, \\ &\beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 = 0. \end{aligned} \tag{3.3}$$

Since α, β satisfy equation (*), we have

$$\alpha_2\beta_3 - \alpha_3\beta_2 = \alpha_1\beta_4 - \alpha_4\beta_1. \tag{3.4}$$

Using (3.4), the lemma can be proved by considering several cases, depending on the rank of the matrix $A = \left[\begin{smallmatrix} \alpha_3 & \alpha_4 \\ \beta_3 & \beta_4 \end{smallmatrix}\right]$. The details are left to the reader. □

Notation 3.1. The boundary conditions for (I) will be given using the following functions: $u_0 = u$, $u_1 = u'$, $u_2 = pu''$, $u_3 = (pu'')' - qu'$.

THEOREM 3.1. *For problem* (I), *self-adjoint* (*separated*) *boundary conditions at* $x = i$ (*where* $i = 0, 1$) *are equivalent to one of the following types.* (*The conditions at* $x = 0$ *and* $x = 1$ *are independent of each other.*)

(1) $$\begin{cases} u_0(i) = 0 \quad (\textit{Dirichlet conditions}), \\ u_1(i) = 0; \end{cases}$$

(2) $$\begin{cases} u_0(i) = 0, \\ cu_1(i) + u_2(i) = 0, \end{cases}$$

(3) $$\begin{cases} au_0(i) + bu_2(i) - u_3(i) = 0, \\ bu_0(i) - u_1(i) = 0; \end{cases}$$

(4) $$\begin{cases} au_0(i) + bu_1(i) - u_3(i) = 0, \\ bu_0(i) + cu_1(i) + u_2(i) = 0. \end{cases}$$

Proof. Let L be the operator of (I). Integration by parts shows that

$$\int_0^1 (Lu)v\,dx = -\gamma_0(u, v) + \gamma_1(u, v) + \int_0^1 u(Lv)\,dx,$$

where (for $i = 0, 1$)

$$\gamma_i(u, v) = (u_1(i)v_2(i) - u_2(i)v_1(i)) + (u_3(i)v_0(i) - u_0(i)v_3(i)). \tag{3.5}$$

The boundary conditions

$$\begin{aligned} &\alpha_{0i}u_0(i) + \alpha_{1i}u_1(i) + \alpha_{2i}u_2(i) + \alpha_{3i}u_3(i) = 0, \\ &\beta_{0i}u_0(i) + \beta_{1i}u_1(i) + \beta_{2i}u_2(i) + \beta_{3i}u_3(i) = 0 \end{aligned} \tag{3.6}$$

at $x = i$ are self-adjoint if and only if $\gamma_i(u, v) = 0$ for all u, v satisfying (3.6). Setting $(x_1, x_2, x_3, x_4) = (u_0(i), u_1(i), u_2(i), u_3(i))$, we see that Lemma 3.1 implies that the equations (3.6) are equivalent to one of the types listed in the theorem. □

Notation 3.2. The boundary conditions for (II) will be given using the functions $u = (u_1, u_2)^T$ and $\dot{u} = (\dot{u}_1, \dot{u}_2)^T = Pu' + Qu$.

The following theorem is proved similarly to the previous one by setting $(x_1, x_2, x_3, x_4) = (u_1(i), u_2(i), \dot{u}_2(i), -\dot{u}_1(i))$ in Lemma 3.1.

THEOREM 3.2. *For* (II), *self-adjoint* (*separated*) *boundary conditions at* $x = i$ (*where* $i = 0, 1$) *are equivalent to one of the following types*: (*The conditions at* $x = 0$ *and* $x = 1$ *are independent of each other.*)

(1) $\begin{cases} u_1(i) = 0 & (\textit{Dirichlet conditions}), \\ u_2(i) = 0; \end{cases}$

(2) $\begin{cases} u_1(i) = 0, \\ cu_2(i) + \dot{u}_2(i) = 0; \end{cases}$

(3) $\begin{cases} au_1(i) + \dot{u}_1(i) + b\dot{u}_2(i) = 0, \\ bu_1(i) - u_2(i) = 0; \end{cases}$

(4) $\begin{cases} au_1(i) + bu_2(i) + \dot{u}_1(i) = 0, \\ bu_1(i) + cu_2(i) + \dot{u}_2(i) = 0. \end{cases}$

The following two theorems can be proved using integration by parts. The details are left for the reader.

THEOREM 3.3. *For* (I), *the following formulas give the inner products* $\beta_i(u, v)$ *and* $B_s(u, v)$, *corresponding to the different types of boundary conditions.*

(1) $\begin{cases} u_0(i) = 0, \\ u_1(i) = 0; \end{cases}$

$\beta_i(u, v) = 0$. *If* $i = 1$, *then* $B_s(u, v) = 0$.

(2) $\begin{cases} u_0(i) = 0, \\ cu_1(i) + u_2(i) = 0; \end{cases}$

$\beta_i(u, v) = cu_1(i)v_1(i)$. *If* $i = 1$, *then* $B_s(u, v) = [cu_1(1) + u_2(1)]v_1(1)$.

(3) $\begin{cases} au_0(i) + bu_2(i) - u_3(i) = 0, \\ bu_0(i) - u_1(i) = 0; \end{cases}$

$\beta_i(u, v) = au_0(i)v_0(i)$.

If $i = 1$, *then* $B_s(u, v) = [au_0(1) + bu_2(1) - u_3(1)]v_0(1)$.

(4) $\begin{cases} au_0(i) + bu_1(i) - u_3(i) = 0, \\ bu_0(i) + cu_1(i) + u_2(i) = 0; \end{cases}$

$\beta_i(u, v) = au_0(i)v_0(i) + b(u_0(i)v_1(i) + u_1(i)v_0(i)) + cu_1(i)v_1(i)$.

If $i = 1$, *then*

$$B_s(u, v) = [au_0(1) + bu_1(1) - u_3(1)]v_0(1) + [bu_0(1) + cu_1(1) + u_2(1)]v_1(1).$$

THEOREM 3.4. *For* (II), *the following formulas give the inner products* $\beta_i(u, v)$ *and* $B_s(u, v)$, *corresponding to the different types of boundary conditions.*

(1) $$\begin{cases} u_1(i) = 0, \\ u_2(i) = 0; \end{cases}$$

$\beta_i(u, v) = 0$. *If* $i = 1$, *then* $B_s(u, v) = 0$.

(2) $$\begin{cases} u_1(i) = 0, \\ cu_2(i) + \dot{u}_2(i) = 0; \end{cases}$$

$\beta_i(u, v) = cu_2(i)v_2(i)$. *If* $i = 1$, *then* $B_s(u, v) = [cu_2(1) + \dot{u}_2(1)]v_2(1)$.

(3) $$\begin{cases} au_1(i) + \dot{u}_1(i) + b\dot{u}_2(i) = 0, \\ bu_1(i) - u_2(i) = 0; \end{cases}$$

$\beta_i(u, v) = au_1(i)v_1(i)$.

If $i = 1$, *then* $B_s(u, v) = [au_1(1) + \dot{u}_1(1) + b\dot{u}_2(1)]v_1(1)$.

(4) $$\begin{cases} au_1(i) + bu_2(i) + \dot{u}_1(i) = 0, \\ bu_1(i) + cu_2(i) + \dot{u}_2(i) = 0; \end{cases}$$

$\beta_i(u, v) = au_1(i)v_1(i) + b(u_1(i)v_2(i) + u_2(i)v_1(i)) + cu_2(i)v_2(i)$.

If $i = 1$, *then*

$$B_s(u, v) = [au_1(1) + bu_2(1) + \dot{u}_1(1)]v_1(1) + [bu_1(1) + cu_2(1) + \dot{u}_2(1)]v_2(1).$$

Remark. The inner products $B_s(u, v)$ do not appear to be symmetric in the above formulas. But this is only an illusion, since $B(u, v)$ is symmetric.

4. Nonlinear eigenvalues. We now turn to the case where the operator L and the boundary conditions depend on λ. The coefficients a, b, c in the boundary conditions (see Theorems 3.1 and 3.2) may now be functions of λ. But in the type (3) boundary conditions (for (I) and (II)) the coefficient b is assumed to be constant, because essential boundary conditions are assumed to have constant coefficients. (See the standard assumption (S5) in § 1.) The type of boundary condition does not change as λ varies, because the type is determined by the essential boundary conditions.

We begin with the following estimates, which will be needed in the proof of Theorem 4.1. As before, we use the notation $\|u\|^2 = \int_0^1 u(x)^2\,dx$.

LEMMA 4.1. *Let* $\varepsilon > 0$.

(1) *If* $u \in H^1[0, 1]$, *then* $u(x)^2 \leqq \varepsilon^2\|u'\|^2 + (1 + 1/\varepsilon^2)\|u\|^2$.

(2) *There exists* $C(\varepsilon) > 0$ *so that if* $u \in H^2[0, 1]$ *then* $\|u'\|^2 \leqq \varepsilon^2\|u''\|^2 + C(\varepsilon)\|u\|^2$.

Proof. (1) Let $\min_{0\leqq x\leqq 1} u(x)^2 = m^2 = u(x_0)^2$. Then $\|u\|^2 = \int_0^1 u(x)^2\,dx \geqq m^2$, so $m \leqq \|u\|$.

$$\begin{aligned} u(x)^2 &= u(x_0)^2 + \int_{x_0}^{x} 2u(x)u'(x)\,dx \\ &\leqq m^2 + 2\left(\frac{\|u\|}{\varepsilon}\right)(\varepsilon\|u'\|) \\ &\leqq m^2 + \left(\varepsilon^2\|u'\|^2 + \frac{1}{\varepsilon^2}\|u\|^2\right) \\ &\leqq \varepsilon^2\|u'\|^2 + \left(1 + \frac{1}{\varepsilon^2}\right)\|u\|^2. \end{aligned}$$

(2) Integrating by parts, we have

$$\|u'\|^2 = \int_0^1 u'(x)^2\,dx = u(1)u'(1) - u(0)u'(0) - \int_0^1 uu''\,dx,$$

and therefore,

$$\|u'\|^2 \leqq |u(1)u'(1)| + |u(0)u'(0)| + \|u\|\,\|u''\|. \tag{4.1}$$

Let $\varepsilon_1, \varepsilon_2, \delta > 0$ be constants to be determined later. For $i = 0, 1$ we have

$$|u(i)u'(i)| \leqq \frac{1}{2}\left(\varepsilon_1^2 u'(i)^2 + \frac{1}{\varepsilon_1^2}u(i)^2\right), \tag{4.2}$$

$$\|u\|\,\|u''\| \leqq \frac{1}{2}\left(\varepsilon_2^2\|u''\|^2 + \frac{1}{\varepsilon_2^2}\|u\|^2\right). \tag{4.3}$$

Part (1) of the lemma implies that

$$u(i)^2 \leqq \delta^2\|u'\|^2 + \left(1+\frac{1}{\delta^2}\right)\|u\|^2, \tag{4.4}$$

$$u'(i)^2 \leqq \|u''\|^2 + 2\|u'\|^2. \tag{4.5}$$

The above inequalities imply that

$$\left(1-2\varepsilon_1^2-\frac{\delta^2}{\varepsilon_1^2}\right)\|u'\|^2 \leqq \left(\varepsilon_1^2+\frac{\varepsilon_2^2}{2}\right)\|u''\|^2 + \left[\frac{1}{\varepsilon_1^2}\left(1+\frac{1}{\delta^2}\right)+\frac{1}{2\varepsilon_2^2}\right]\|u\|^2. \tag{4.6}$$

We may assume that $\varepsilon < 1$ (or else replace ε by $\varepsilon' = \frac{1}{2}$). Now let $\varepsilon_1 = \varepsilon/2$ and $\varepsilon_2 = \varepsilon/\sqrt{2}$, so that $\varepsilon_1^2 + (\varepsilon_2^2/2) = \varepsilon^2/2$. Define δ so that $1 - 2\varepsilon_1^2 - (\delta^2/\varepsilon_1^2) = \frac{1}{2}$ (i.e., $\delta^2 = \varepsilon^2(1-\varepsilon^2)/8$), and let $C(\varepsilon) = 2[(1/\varepsilon_1^2)(1+1/\delta^2)+1/(2\varepsilon_2^2)]$. Now (4.6) implies that $\|u'\|^2 \leqq \varepsilon^2\|u''\|^2 + C(\varepsilon)\|u\|^2$. □

LEMMA 4.2. *Let Q be an $n \times 2n$ matrix and let $|Q|$ be its norm with respect to the Euclidean norm in $\mathbb{R}^n$. Let $\hat{Q}$ be the $2n \times 2n$ matrix $\begin{bmatrix} 0 & Q \\ Q^T & 0 \end{bmatrix}$ and let λ_0 be the maximum eigenvalue of $\hat{Q}$. Then $|Q| = \lambda_0$.*

Proof. Let $x, y \in \mathbb{R}^n$ and $z = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^{2n}$. The equation $\hat{Q}z = \lambda z$ is equivalent to the pair of equations

$$Q^T x = \lambda y, \qquad Qy = \lambda x, \tag{4.7}$$

which imply that

$$QQ^T x = \lambda^2 x, \qquad Q^T Q y = \lambda^2 y. \tag{4.8}$$

Therefore, if λ is an eigenvalue of $\hat{Q}$, then λ^2 is an eigenvalue of Q^TQ. Conversely, if $Q^TQy = \lambda^2 y$, let $Qy = \lambda x$. Then $\lambda Q^T x = Q^T Q y = \lambda^2 y$. If $\lambda \neq 0$, then $Q^T x = \lambda y$. Since $Qy = \lambda x$ and $Q^T x = \lambda y$, it follows that λ is an eigenvalue of $\hat{Q}$. If $\lambda = 0$, then Q and Q^T have a zero eigenvalue, so $\hat{Q}$ does also. Thus λ is an eigenvalue of $\hat{Q}$ if and only if λ^2 is an eigenvalue of Q^TQ. Since λ_0 is the maximum eigenvalue of $\hat{Q}$, λ_0^2 is the maximum eigenvalue of Q^TQ.

Let $\langle x, y\rangle$ denote the Euclidean inner product in $\mathbb{R}^n$, and $|x|^2 = \langle x, x\rangle$. Then

$$|Q|^2 = \max_{x\neq 0}\frac{|Qx|^2}{|x|^2} = \max_{x\neq 0}\frac{\langle Qx, Qx\rangle}{|x|^2} = \max_{x\neq 0}\frac{\langle Q^TQx, x\rangle}{|x|^2} = \lambda_0^2.$$

Therefore, $|Q| = \lambda_0$. □

Remark. The above proof shows that if λ is an eigenvalue of $\hat{Q}$, then $-\lambda$ is also an eigenvalue. In fact, the characteristic polynomial $p(\lambda) = \det(\lambda I - \hat{Q})$ is an even function. This can be seen by multiplying the first n rows and the last n columns of $\det(\lambda I - \hat{Q})$ by -1.

The lemma shows that the norm $|Q|$ of the matrix $Q(x, \lambda)$ (which occurs in (II)) equals the maximum eigenvalue $\nu_Q(x, \lambda)$ of $\hat{Q} = \begin{bmatrix} 0 & Q \\ Q^T & 0 \end{bmatrix}$. This implies that

$$|Q(x, \lambda)u \cdot u'| \leqq \nu_Q(\lambda)|u||u'|, \tag{4.9}$$

where $\nu_Q(\lambda) = \max_{0\leqq x\leqq 1} \nu_Q(x, \lambda)$. This fact will be used in the proof of the following theorem.

THEOREM 4.1. *Let L be a differential operator of type* (I) *or* (II). *Suppose that L satisfies the monotonicity assumptions and the limit assumption* (L1). *Then the energy inner product $B(u, v; \lambda)$ is positive definite for λ near Λ_1.*

Proof. Suppose that L is of type (I). Recall that

$$B(u, u; \lambda) = -\beta_0(u, u; \lambda) + \beta_1(u, u; \lambda) + \int_0^1 (p(x, \lambda)(u'')^2 + q(x, \lambda)(u')^2 + r(x, \lambda)u^2)\, dx, \tag{4.10}$$

where, for $i = 0, 1$,

$$\beta_i(u, u; \lambda) = a_i(\lambda)u(i)^2 + 2b_i(\lambda)u(i)u'(i) + c_i(\lambda)u'(i)^2. \tag{4.11}$$

Suppose that L satisfies the limit assumption (L1a): $\lim_{\lambda\to\Lambda_1} r_*(\lambda) = \infty$. Let $\Lambda_1 < \lambda_0 < \Lambda_2$. The standard assumption (S2) and the monotonicity assumptions imply that for $\Lambda_1 < \lambda < \lambda_0$,

$$B(u, u; \lambda) \geqq -\beta_0(u, u; \lambda_0) + \beta_1(u, u; \lambda_0) + \int_0^1 (k(u'')^2 + q(x, \lambda_0)(u')^2 + r(x, \lambda)u^2)\, dx, \tag{4.12}$$

and therefore,

$$B(u, u; \lambda) \geqq -|\beta_0(u, u; \lambda_0)| - |\beta_1(u, u; \lambda_0)| + k\|u''\|^2 + q_*(\lambda_0)\|u'\|^2 + r_*(\lambda)\|u\|^2. \tag{4.13}$$

Letting $a_i = a_i(\lambda_0)$, $b_i = b_i(\lambda_0)$, and $c_i = c_i(\lambda_0)$, we have

$$|\beta_i(u, u; \lambda_0)| \leqq |a_i|u(i)^2 + 2|b_i||u(i)u'(i)| + |c_i|u'(i)^2, \tag{4.14}$$

$$2|u(i)u'(i)| \leqq u(i)^2 + u'(i)^2. \tag{4.15}$$

Let $\varepsilon > 0$ be a constant to be chosen later. Lemma 4.1 implies that

$$u(i)^2 \leqq \|u'\|^2 + 2\|u\|^2, \tag{4.16}$$

$$u'(i)^2 \leqq \varepsilon^2\|u''\|^2 + \left(1 + \frac{1}{\varepsilon^2}\right)\|u'\|^2. \tag{4.17}$$

These inequalities imply that

$$|\beta_i(u, u; \lambda_0)| \leqq A_i\varepsilon^2\|u''\|^2 + B_i(\varepsilon)\|u'\|^2 + C_i\|u\|^2, \tag{4.18}$$

where $A_i = |b_i| + |c_i|$, $B_i(\varepsilon) = (|a_i| + |b_i|) + (|b_i| + |c_i|)(1 + 1/\varepsilon^2)$ and $C_i = 2(|a_i| + |b_i|)$. Letting $A = A_0 + A_1$, $B(\varepsilon) = B_0(\varepsilon) + B_1(\varepsilon)$, and $C = C_0 + C_1$, the above inequalities imply that for $\Lambda_1 < \lambda < \lambda_0$,

$$B(u, u; \lambda) \geqq (k - A\varepsilon^2)\|u''\|^2 + (q_*(\lambda_0) - B(\varepsilon))\|u'\|^2 + (r_*(\lambda) - C)\|u\|^2. \tag{4.19}$$

Let $\delta > 0$ be a constant to be determined later. Lemma 4.1 implies that there exists $D(\delta) > 0$ so that

$$\|u'\|^2 \leqq \delta^2\|u''\|^2 + D(\delta)\|u\|^2. \tag{4.20}$$

By letting $E(\varepsilon) = |q_*(\lambda_0) - B(\varepsilon)|$, the inequalities (4.19) and (4.20) imply that

$$B(u, u; \lambda) \geqq (k - A\varepsilon^2 - E(\varepsilon)\delta^2)\|u''\|^2 + (r_*(\lambda) - C - D(\delta)E(\varepsilon))\|u\|^2, \tag{4.21}$$

for $\Lambda_1 < \lambda < \lambda_0$. Now choose ε and δ so that $A\varepsilon^2 < k/2$ and $E(\varepsilon)\delta^2 < k/2$. Then $k - A\varepsilon^2 - E(\varepsilon)\delta^2 > 0$ and (L1a) implies that $r_*(\lambda) - C - D(\delta)E(\varepsilon) > 0$ for λ near Λ_1. This concludes the proof under the assumption (L1a).

Next we suppose that L is of type (I) and satisfies (L1b): $\lim_{\lambda \to \Lambda_1} q_*(\lambda) = \infty$ and there exists λ_+ such that $r_*(\lambda_+) > 0$, $a_0(\lambda_+) \leqq 0$, and $a_1(\lambda_+) \geqq 0$. The standard assumption (S2) and the monotonicity assumptions imply that for $\Lambda_1 < \lambda < \lambda_+$,

$$\begin{aligned} B(u, u; \lambda) \geqq & -\beta_0(u, u; \lambda_+) + \beta_1(u, u; \lambda_+) + k\|u''\|^2 \\ & + q_*(\lambda)\|u'\|^2 + r_*(\lambda_+)\|u\|^2. \end{aligned} \tag{4.22}$$

Letting $a_i = a_i(\lambda_+)$, $b_i = b_i(\lambda_+)$, $c_i = c_i(\lambda_+)$ and using

$$2|u(i)u'(i)| \leqq \varepsilon^2 u(i)^2 + \frac{1}{\varepsilon^2} u'(i)^2, \tag{4.23}$$

$$u(i)^2 \leqq \|u'\|^2 + 2\|u\|^2, \tag{4.24}$$

and

$$u'(i)^2 \leqq \delta^2\|u''\|^2 + \left(1 + \frac{1}{\delta^2}\right)\|u'\|^2, \tag{4.25}$$

we obtain

$$\beta_1(u, u; \lambda_+) \geqq a_1 u(1)^2 - A_1(\varepsilon)\delta^2\|u''\|^2 - B_1(\varepsilon, \delta)\|u'\|^2 - C_1\varepsilon^2\|u\|^2, \tag{4.26}$$

$$-\beta_0(u, u; \lambda_+) \geqq -a_0 u(0)^2 - A_0(\varepsilon)\delta^2\|u''\|^2 - B_0(\varepsilon, \delta)\|u'\|^2 - C_0\varepsilon^2\|u\|^2, \tag{4.27}$$

where

$$A_i(\varepsilon) = |c_i| + \frac{|b_i|}{\varepsilon^2}, \quad B_i(\varepsilon, \delta) = |b_i|\varepsilon^2 + \left(|c_i| + \frac{|b_i|}{\varepsilon^2}\right)\left(1 + \frac{1}{\delta^2}\right), \quad \text{and } C_i = 2|b_i|.$$

Letting $A = A_0 + A_1$, $B = B_0 + B_1$, and $C = C_0 + C_1$, (4.22) now implies that

$$\begin{aligned} B(u, u; \lambda) \geqq & -a_0 u(0)^2 + a_1 u(1)^2 + (k - (A(\varepsilon)\delta^2)\|u''\|^2 \\ & + (q_*(\lambda) - B(\varepsilon, \delta))\|u'\|^2 + (r_*(\lambda_+) - C\varepsilon^2)\|u\|^2 \end{aligned} \tag{4.28}$$

for $\Lambda_1 < \lambda < \lambda_+$. Choose ε and δ so that $r_*(\lambda_+) - C\varepsilon^2 > 0$ and $k - A(\varepsilon)\delta^2 > 0$. The assumption (L1b) implies that $q_*(\lambda) - B(\varepsilon, \delta) > 0$ for λ near Λ_1. This concludes the proof under the assumption (L1b).

Now suppose that L is of type (II). Recall that

$$\begin{aligned} B(u, u; \lambda) = & -\beta_0(u, u; \lambda_0) + \beta_1(u, u; \lambda_0) \\ & + \int_0^1 (P(x, \lambda)u' \cdot u' + 2Q(x, \lambda)u \cdot u' + R(x, \lambda)u \cdot u)\, dx, \end{aligned} \tag{4.29}$$

where

$$\beta_i(u, u; \lambda) = a_i(\lambda)u_1(i)^2 + 2b_i(\lambda)u_1(i)u_2(i) + c_i(\lambda)u_2(i)^2. \tag{4.30}$$

The standard assumption (S2) and the monotonicity assumptions imply that for $\Lambda_1 < \lambda < \lambda_0$,

$$\begin{aligned} B(u, u; \lambda) \geqq & -\beta_0(u, u; \lambda_0) + \beta_1(u, u; \lambda_0) \\ & + \int_0^1 (ku' \cdot u' + 2Q(x, \lambda)u \cdot u' + R(x, \lambda)u \cdot u)\, dx. \end{aligned} \tag{4.31}$$

The eigenvalues of $R(x,\lambda)$ are $\mu_R(x,\lambda) \leqq \nu_R(x,\lambda)$, and $\mu_R(\lambda) = \min_{0\leqq x\leqq 1} \mu_R(x,\lambda)$. For $u=(u_1,u_2)^T$, let $|u|^2 = u_1^2+u_2^2$ and $\|u\|^2 = \int_0^1 |u|^2\,dx$. Then

$$\int_0^1 R(x,\lambda)u\cdot u\,dx \geqq \mu_R(\lambda)\|u\|^2. \tag{4.32}$$

The eigenvalues of $\begin{bmatrix} 0 & Q(x,\lambda) \\ Q^T(x,\lambda) & 0 \end{bmatrix}$ are $\pm\mu_Q(x,\lambda)$ and $\pm\nu_Q(x,\lambda)$, where $0 \leqq \mu_Q(x,\lambda) \leqq \nu_Q(x,\lambda)$ and $\nu_Q(\lambda) = \max_{0\leqq x\leqq 1} \nu_Q(x,\lambda)$. It follows from (4.9) that

$$2|Q(x,\lambda)u\cdot u'| \leqq 2\nu_Q(\lambda)|u||u'| \leqq \varepsilon^2|u'|^2 + \frac{\nu_Q(\lambda)^2}{\varepsilon^2}|u|^2,$$

and therefore,

$$\int_0^1 2Q(x,\lambda)u\cdot u'\,dx \geqq -\left(\varepsilon^2\|u'\|^2 + \frac{\nu_Q(\lambda)^2}{\varepsilon^2}\|u\|^2\right). \tag{4.33}$$

The inequalities (4.31), (4.32), and (4.33) imply

$$\begin{aligned} B(u,u;\lambda) \geqq & -\beta_0(u,u;\lambda_0)+\beta_1(u,u;\lambda_0) \\ & +(k-\varepsilon^2)\|u'\|^2 + \left[\mu_R(\lambda) - \frac{\nu_Q(\lambda)^2}{\varepsilon^2}\right]\|u\|^2 \end{aligned} \tag{4.34.}$$

for $\Lambda_1 < \lambda < \lambda_0$. Using the estimates

$$2|u_1(i)u_2(i)| \leqq u_1(i)^2 + u_2(i)^2, \tag{4.35}$$

and

$$u_j(i)^2 \leqq \varepsilon^2\|u_j'\|^2 + \left(1+\frac{1}{\varepsilon^2}\right)\|u_j\|^2, \tag{4.36}$$

we obtain

$$|\beta_i(u,u;\lambda_0)| \leqq A_i\left(\varepsilon^2\|u'\|^2 + \left(1+\frac{1}{\varepsilon^2}\right)\|u\|^2\right), \tag{4.37}$$

where $A_i = |a_i(\lambda_0)| + |b_i(\lambda_0)| + |c_i(\lambda_0)|$. Letting $A = 1 + A_0 + A_1$ and using the fact that $\nu_Q(\lambda)^2/\varepsilon^2 + A(1+1/\varepsilon^2) \leqq A(1+\nu_Q(\lambda)^2)(1+1/\varepsilon^2)$, the inequalities (4.34) and (4.37) imply that

$$\begin{aligned} B(u,u;\lambda) \geqq & (k-A\varepsilon^2)\|u'\|^2 \\ & +(1+\nu_Q(\lambda)^2)\left[\frac{\mu_R(\lambda)}{1+\nu_Q(\lambda)^2} - A\left(1+\frac{1}{\varepsilon^2}\right)\right]\|u\|^2 \end{aligned} \tag{4.38}$$

for $\Lambda_1 < \lambda < \lambda_0$. Now choose $\varepsilon > 0$ so that $k - A\varepsilon^2 > 0$. The limit assumption (L1) implies that

$$\frac{\mu_R(\lambda)}{1+\nu_Q(\lambda)^2} - A\left(1+\frac{1}{\varepsilon^2}\right) > 0 \quad \text{for } \lambda \text{ near } \Lambda_1. \qquad \square$$

Remark. If (I) has a boundary condition of type (1) or (2) at $x=0$ or 1 (see Theorem 3.1 for the types of boundary conditions), then Lemma 2.1 implies that $u(x)^2 \leqq \|u'\|^2$ and $\|u\|^2 \leqq \|u'\|^2$ for $u \in H$. In this case, estimates similar to those in the the previous proof show that $B(u,v;\lambda)$ is positive definite for λ near Λ_1 if $\lim_{\lambda\to\Lambda_1} q_*(\lambda) = \infty$.

DEFINITION 4.1. L_λ will denote the operator L for fixed λ, and

$$N_0(\lambda) = N_0(L_\lambda), \tag{4.39}$$

$$\sigma(\lambda) = \sigma(L_\lambda), \tag{4.40}$$

$$N(\lambda) = N(L_\lambda) = N_0(\lambda) + \sigma(\lambda). \tag{4.41}$$

The following theorem is the main result of this paper. It is analogous to Theorem 2.1 in [8], which applies to second-order problems.

THEOREM 4.2. *Let L be an operator of type* (I) *or* (II).

(1) *Let $\Lambda_1 < \lambda' < \lambda'' < \Lambda_2$. If L satisfies the monotonicity assumptions, then L has exactly $N(\lambda'') - N(\lambda')$ eigenvalues (counting multiplicity) in the interval $[\lambda', \lambda'')$.*

(2) *Let $\Lambda_1 < \lambda < \Lambda_2$. If L also satisfies the limit assumption* (L1), *then L has exactly $N(\lambda)$ eigenvalues (counting multiplicity) in the interval (Λ_1, λ).*

Proof. (1) Let $\mu_1(\lambda) \leqq \mu_2(\lambda) \leqq \cdots \leqq \mu_n(\lambda) \leqq \cdots$ be the eigenvalues of L_λ. The following statements are equivalent:

(a) λ_0 is an eigenvalue of L;

(b) zero is an eigenvalue of L_{λ_0};

(c) $\mu_k(\lambda_0) = 0$, for some k.

Furthermore, λ_0 is a double eigenvalue of L if and only if $\mu_k(\lambda_0) = \mu_{k+1}(\lambda_0) = 0$, for some k.

The monotonicity assumptions imply that for each $u \in H$, $u \neq 0$, $B(u, u; \lambda)$ is a strictly decreasing function of λ. This implies that the $\mu_k(\lambda)$ are strictly decreasing functions. By Theorem 2.5, L_λ has exactly $N(\lambda)$ negative eigenvalues. As λ increases from λ' to λ'', the number of negative eigenvalues increases from $N(\lambda')$ to $N(\lambda'')$. Thus $N(\lambda'') - N(\lambda')$ new eigenvalues $\mu_k(\lambda)$ have become negative, thereby passing through zero. This shows that L has $N(\lambda'') - N(\lambda')$ eigenvalues in the interval $[\lambda', \lambda'')$.

(2) If L satisfies the limit assumption (L1), then Theorem 4.1 shows that $N(\lambda_0) = 0$ for λ_0 near Λ_1. Part (1) of the present theorem implies that there are no eigenvalues in the interval (Λ_1, λ_0), and $N(\lambda) = 0$ for $\Lambda_1 < \lambda \leqq \lambda_0$. If $\lambda_0 < \lambda < \Lambda_2$, then any eigenvalues in (Λ_1, λ) are in the interval $[\lambda_0, \lambda)$, and the number of these eigenvalues is $N(\lambda) - N(\lambda_0) = N(\lambda)$. □

Remark. In [7] it is shown that if L satisfies the limit assumption (L2), then $\lim_{\lambda \to \Lambda_2} N(\lambda) = \infty$ and L has infinitely many eigenvalues.

5. Wronskians. The Wronskians used to calculate $N(\lambda)$ will now be considered in greater detail. We shall define six Wronskians that are related to each other by a system of differential equations. Let the given boundary conditions at $x = 0$ be denoted by $\mathrm{BC}_0(y; \lambda) = 0$. For a given λ_0, let u, v be linearly independent solutions of

$$\begin{aligned} L(y; \lambda_0) &= 0, \\ \mathrm{BC}_0(y; \lambda_0) &= 0. \end{aligned} \tag{5.1}$$

For problem (I) we continue to use the notation $u_0 = u$, $u_1 = u'$, $u_2 = pu''$, and $u_3 = (pu'')' - qu'$. Now let

$$\begin{aligned} w_1 &= u_0 v_1 - u_1 v_0, & w_2 &= u_0 v_2 - u_2 v_0, \\ w_3 &= u_0 v_3 - u_3 v_0, & w_4 &= u_1 v_2 - u_2 v_1, \\ w_5 &= u_1 v_3 - u_3 v_1, & w_6 &= u_2 v_3 - u_3 v_2. \end{aligned} \tag{5.2}$$

For problem (II) let $u = (u_1, u_2)^T$, $\dot{u} = (\dot{u}_1, \dot{u}_2)^T = Pu' + Qu$, and let

$$\begin{aligned} w_1 &= u_1 v_2 - u_2 v_1, & w_2 &= u_1 \dot{v}_1 - \dot{u}_1 v_1, \\ w_3 &= u_1 \dot{v}_2 - \dot{u}_2 v_1, & w_4 &= u_2 \dot{v}_1 - \dot{u}_1 v_2, \\ w_5 &= u_2 \dot{v}_2 - \dot{u}_2 v_2, & w_6 &= \dot{u}_1 \dot{v}_2 - \dot{u}_2 \dot{v}_1. \end{aligned} \tag{5.3}$$

Note that the Wronskian determinant $w(x)$ (that occurs in Definition 2.2) is the same as $w_1(x)$ for both (I) and (II). The self-adjointness of the problems implies the following fact.

THEOREM 5.1. (1) *For* (I), $w_4 = w_3$.

(2) *For* (II), $w_5 = -w_2$.

Proof. (1) Integration by parts shows that

$$\int_0^x (Lu)v\,dt - \int_0^x u(Lv)\,dt = [(u_3v_0 - u_0v_3) + (u_1v_2 - u_2v_1)]_0^x. \tag{5.4}$$

(Here we have abbreviated $L(u;\lambda_0)$ by Lu.) If u and v are solutions of (5.1), then $Lu = 0 = Lv$, and (5.4) shows that

$$(u_3v_0 - u_0v_3) + (u_1v_2 - u_2v_1) = \text{constant}. \tag{5.5}$$

But the boundary conditions at $x = 0$ are self-adjoint precisely when the left-hand side of (5.5) is zero at $x = 0$ (see the proof of Theorem 3.1). Therefore, the constant in (5.5) is zero. Referring to (5.2), this implies that $w_4 = w_3$.

(2) Similarly, for (II), integration by parts shows that

$$\int_0^x (Lu)\cdot v\,dt - \int_0^x u\cdot(Lv)\,dt = [u\cdot\dot{v} - \dot{u}\cdot v]_0^x. \tag{5.6}$$

As before, this leads to the equation

$$u\cdot\dot{v} - \dot{u}\cdot v = \text{constant}. \tag{5.7}$$

Since the boundary conditions at $x = 0$ are self-adjoint, the constant in (5.7) is zero, which implies that $w_5 = -w_2$. □

Theorem 5.1 allows us to reduce the system of six Wronskians to five by omitting w_4 for (I) and w_5 for (II). A simple calculation now shows that for (I), the Wronskians satisfy the following equations:

$$\begin{aligned} w_1' &= (1/p)w_2, \\ w_2' &= qw_1 + 2w_3, \\ w_3' &= w_5, \\ w_5' &= rw_1 + (1/p)w_6, \\ w_6' &= rw_2 + qw_5. \end{aligned} \tag{5.8}$$

In (II), the equation $L(u;\lambda_0) = 0$ can be split into a first-order system:

$$\begin{aligned} u' &= -P^{-1}Qu + P^{-1}\dot{u}, \\ \dot{u}' &= (-Q^TP^{-1}Q + R)u + Q^TP^{-1}\dot{u}. \end{aligned} \tag{5.9}$$

Let

$$\begin{aligned} A &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = -P^{-1}Q, \\ B &= \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = P^{-1}, \\ C &= \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = -Q^TP^{-1}Q + R. \end{aligned} \tag{5.10}$$

Note that $B = B^T$ and $C = C^T$. Equation (5.9) may be rewritten as

$$\begin{aligned} u' &= Au + B\dot{u}, \\ \dot{u}' &= Cu - A^T\dot{u}. \end{aligned} \tag{5.11}$$

Using (5.11), a calculation shows that the Wronskians for (II) satisfy the following equations:

$$\begin{aligned} w_1' &= (a_{11}+a_{22})w_1 + 2b_{12}w_2 + b_{22}w_3 - b_{11}w_4, \\ w_2' &= c_{12}w_1 - a_{21}w_3 + a_{12}w_4 - b_{12}w_6, \\ w_3' &= c_{22}w_1 - 2a_{12}w_2 + (a_{11}-a_{22})w_3 + b_{11}w_6, \\ w_4' &= -c_{11}w_1 + 2a_{21}w_2 + (-a_{11}+a_{22})w_4 - b_{22}w_6, \\ w_6' &= -2c_{12}w_2 + c_{11}w_3 - c_{22}w_4 - (a_{11}+a_{22})w_6. \end{aligned} \tag{5.12}$$

Recall that $n(x; \lambda_0) = \text{nullity } W(x; \lambda_0)$, where $W(x; \lambda_0) = \left[\begin{smallmatrix} u_0 & v_0 \\ u_1 & v_1 \end{smallmatrix}\right]$ for (I) and $W(x; \lambda_0) = \left[\begin{smallmatrix} u_1 & v_1 \\ u_2 & v_2 \end{smallmatrix}\right]$ for (II).

THEOREM 5.2. *For* (I) *and* (II), *the following are equivalent.*

(1) $n(x_0; \lambda_0) = 2$,

(2) $w_1(x_0; \lambda_0) = w_2(x_0; \lambda_0) = w_3(x_0; \lambda_0) = w_4(x_0; \lambda_0) = w_5(x_0; \lambda_0) = 0$.

Proof. Statement (1) is true for (I) if and only if $u_0(x_0) = u_1(x_0) = v_0(x_0) = v_1(x_0) = 0$. It is true for (II) if and only if $u_1(x_0) = u_2(x_0) = v_1(x_0) = v_2(x_0) = 0$. Referring to (5.2) and (5.3), we see that (1) implies (2).

Conversely, suppose that (2) is true for (I). If $(u_0(x_0), v_0(x_0)) \neq (0, 0)$, then, since $w_2(x_0) = w_3(x_0) = 0$, it follows that there are constants α, β such that $(u_2(x_0), v_2(x_0)) = \alpha(u_0(x_0), v_0(x_0))$ and $(u_3(x_0), v_3(x_0)) = \beta(u_0(x_0), v_0(x_0))$. This implies that $w_6(x_0) = 0$. Since all six Wronskians are zero at $x = x_0$, the vectors $(u_0(x_0), u_1(x_0), u_2(x_0), u_3(x_0))$ and $(v_0(x_0), v_1(x_0), v_2(x_0), v_3(x_0))$ are linearly dependent. This implies that $u(x)$ and $v(x)$ are linearly dependent. This contradiction implies that $u_0(x_0) = v_0(x_0) = 0$.

Similarly, if $(u_1(x_0), v_1(x_0)) \neq (0, 0)$, then $w_4(x_0) = w_5(x_0) = 0$ implies that $(u_2(x_0), v_2(x_0)) = \alpha(u_1(x_0), v_1(x_0))$ and $(u_3(x_0), v_3(x_0)) = \beta(u_1(x_0), v_1(x_0))$. This again implies that $w_6(x_0) = 0$, which leads to a contradiction. Therefore $u_1(x_0) = v_1(x_0) = 0$, and $n(x_0; \lambda_0) = 2$.

A similar argument shows that (2) implies (1) for (II). □

THEOREM 5.3. *The following are true for* (I).

(1) *If* $n(x_0; \lambda_0) = 1$, *then* $w(x)$ *has a zero of order* 1 *or* 3 *at* x_0.

(2) *If* $n(x_0; \lambda_0) = 2$, *then* $w(x)$ *has a zero of order* 4 *at* x_0.

Proof. Using equations (5.8), the derivatives of w can be expressed in the form

$$\begin{aligned} w &= w_1, \\ w' &= (1/p)w_2, \\ w'' &= \alpha_{21}w_1 + \alpha_{22}w_2 + (2/p)w_3, \\ w''' &= \alpha_{31}w_1 + \alpha_{32}w_2 + \alpha_{33}w_3 + (2/p)w_5, \\ w^{(iv)} &= \alpha_{41}w_1 + \alpha_{42}w_2 + \alpha_{43}w_3 + \alpha_{45}w_5 + (2/p^2)w_6. \end{aligned} \tag{5.13}$$

By Theorem 5.2, $n(x_0; \lambda_0) = 2$ if and only if $w_1(x_0) = w_2(x_0) = w_3(x_0) = w_4(x_0) = w_5(x_0) = 0$. Recalling that $w_3 = w_4$, (5.13) shows that $n(x_0; \lambda_0) = 2$ if and only if $w(x_0) = w'(x_0) = w''(x_0) = w'''(x_0) = 0$. Moreover, if $w^{(iv)}(x_0) = 0$ also, then (5.13) shows that $w_6(x_0) = 0$, so that all six Wronskians are zero at x_0. This leads to a contradiction (as in the proof of Theorem 5.2). Therefore, $w(x)$ has a zero of order 4 at x_0 if $n(x_0; \lambda_0) = 2$.

Now suppose that $n(x_0; \lambda_0) = 1$. Then $w(x_0) = 0$, and we may suppose that $u_0(x_0) = u_1(x_0) = 0$. This implies that $w_2(x_0) = -u_2(x_0)v_0(x_0)$, $w_3(x_0) = -u_3(x_0)v_0(x_0)$ and $w_4(x_0) = -u_2(x_0)v_1(x_0)$. If $w'(x_0) \neq 0$, then $w(x)$ has a zero of order 1 at x_0. Suppose that $w'(x_0) = 0$. Then (5.13) shows that $w_2(x_0) = 0$, which implies that either $v_0(x_0) = 0$ or $u_2(x_0) = 0$. Therefore, either $w_3(x_0) = 0$ or $w_4(x_0) = 0$. But $w_3 = w_4$, by Theorem 6.1. Thus, $w_3(x_0) = 0$, and (5.13) implies that $w''(x_0) = 0$. If $w'''(x_0) = 0$ also, then $n(x_0; \lambda_0) = 2$, as shown above. Therefore, $w(x)$ has a zero of order 1 or 3 at x_0 if $n(x_0; \lambda_0) = 1$. □

THEOREM 5.4. *For problem* (II), $n(x_0; \lambda_0)$ *equals the order of the zero of* $w(x)$ *at* x_0.

Proof. We know that $n(x_0; \lambda_0) > 0$ if and only if $w(x_0) = 0$. We claim that $n(x_0; \lambda_0) = 2$ if and only if $w(x_0) = w'(x_0) = 0$. If $n(x_0; \lambda_0) = 2$, then $w_1(x_0) = w_2(x_0) = w_3(x_0) = w_4(x_0) = w_5(x_0) = 0$, by Theorem 5.2. The first equation in (5.12), which is

$$w' = (a_{11} + a_{22})w_1 + 2b_{12}w_2 + b_{22}w_3 - b_{11}w_4, \tag{5.14}$$

now implies that $w'(x_0) = 0$.

Conversely, suppose that $w(x_0) = w'(x_0) = 0$. We may assume that $u_1(x_0) = u_2(x_0) = 0$. Therefore, $w_2(x_0) = -\dot{u}_1(x_0)v_1(x_0)$, $w_3(x_0) = -\dot{u}_2(x_0)v_1(x_0)$ and $w_4(x_0) = -\dot{u}_1(x_0)v_2(x_0)$. Equation (5.14) implies that

$$\begin{aligned} w'(x_0) = &-2b_{12}(x_0)\dot{u}_1(x_0)v_1(x_0) - b_{22}(x_0)\dot{u}_2(x_0)v_1(x_0) \\ &+ b_{11}(x_0)\dot{u}_1(x_0)v_2(x_0). \end{aligned} \tag{5.15}$$

By Theorem 6.1, $w_5 = -w_2$. Since $u_1(x_0) = u_2(x_0) = 0$, this implies that

$$\dot{u}_1(x_0)v_1(x_0) + \dot{u}_2(x_0)v_2(x_0) = 0. \tag{5.16}$$

Substituting in (5.15), and setting $w'(x_0) = 0$, we obtain

$$\begin{aligned} &b_{11}(x_0)\dot{u}_1(x_0)v_2(x_0) + b_{12}(x_0)[\dot{u}_2(x_0)v_2(x_0) - \dot{u}_1(x_0)v_1(x_0)] \\ &\quad - b_{22}(x_0)\dot{u}_2(x_0)v_1(x_0) = 0. \end{aligned} \tag{5.17}$$

Let $\langle \xi, \eta \rangle$ denote the Euclidean inner product for vectors $\xi, \eta \in \mathbb{R}^2$, and consider the vectors $\dot{u}(x_0) = (\dot{u}_1(x_0), \dot{u}_2(x_0))^T$ and $v(x_0)^\perp = (v_2(x_0), -v_1(x_0))^T$. Note that $\dot{u}(x_0) \neq (0, 0)$, since $u(x)$ is not the trivial solution of $L(u; \lambda_0) = 0$. Let

$$B_0 = B(x_0) = \begin{bmatrix} b_{11}(x_0) & b_{12}(x_0) \\ b_{21}(x_0) & b_{22}(x_0) \end{bmatrix} = P(x_0)^{-1}. \tag{5.18}$$

Since B_0 is positive definite, $B_0\dot{u}(x_0) \neq 0$. Equation (5.17) is equivalent to

$$\langle B_0\dot{u}(x_0), v(x_0)^\perp \rangle = 0. \tag{5.19}$$

Since $B_0\dot{u}(x_0)$ and $v(x_0) = (v_1(x_0), v_2(x_0))^T$ are both orthogonal to $v(x_0)^\perp$, and $B_0\dot{u}(x_0) \neq 0$, it follows that there is a constant c such that $v(x_0) = cB_0\dot{u}(x_0)$. This implies that

$$P_0 v(x_0) = c\dot{u}(x_0), \tag{5.20}$$

where $P_0 = P(x_0)$. Equation (5.16) says that $\langle \dot{u}(x_0), v(x_0) \rangle = 0$. By (5.20), $\langle P_0 v(x_0), v(x_0) \rangle = c\langle \dot{u}(x_0), v(x_0) \rangle = 0$, and therefore, $v(x_0) = 0$. This shows that $n(x_0; \lambda_0) = 2$ if $w(x_0) = w'(x_0) = 0$.

We now know that $n(x_0; \lambda_0) = 2$ if and only if $w(x_0) = w'(x_0) = 0$. Furthermore, we claim that $w''(x_0) \neq 0$ in this case. To see this, differentiate (5.14) and use (5.12) to obtain

$$w'' = \alpha_1 w_1 + \alpha_2 w_2 + \alpha_3 w_3 + \alpha_4 w_4 + 2(b_{11}b_{22} - b_{12}^2)w_6. \tag{5.21}$$

Since $n(x_0; \lambda_0) = 2$,

$$w_1(x_0) = w_2(x_0) = w_3(x_0) = w_4(x_0) = w_5(x_0) = 0.$$

If $w''(x_0) = 0$, then (5.21) implies that $w_6(x_0) = 0$ also. This leads to a contradiction (as in the proof of Theorem 5.2). Therefore, x_0 is a zero of order 2 (of $w(x)$) if $n(x_0; \lambda_0) = 2$, and it must be a zero of order 1 if $n(x_0; \lambda_0) = 1$. □

Theorems 5.3 and 5.4 imply the important fact that $w(x)$ changes sign at x_0 if $n(x_0; \lambda_0) = 1$, and does not change sign if $n(x_0; \lambda_0) = 2$.

REFERENCES

[1] M. BÔCHER, *Leçons sur les Méthodes de Sturm dans la Théorie des Equations Différéntielles Linéaires*, Gauthier-Villars, Paris, 1917.

[2] S. CHANDRASEKHAR, *Hydrodynamic and Hydromagnetic Stability*, Oxford University Press, London, 1961.

[3] ———, *On characteristic value problems in high order differential equations which arise in studies on hydrodynamic and hydromagnetic stability*, Amer. Math. Monthly, 61 (1954), pp. 32–45.

[4] L. COLLATZ, *Eigenwertprobleme und ihre Numerische Behandlung*, Chelsea reprint, New York, 1948.

[5] P. G. DRAZIN AND W. H. REID, *Hydrodynamic Stability*, Cambridge University Press, Cambridge, 1981.

[6] H. M. EDWARDS, *A generalized Sturm theorem*, Ann. of Math., 80 (1964), pp. 22–57.

[7] L. GREENBERG, *Existence theorems for nonlinear eigenvalues in fourth order, self-adjoint, two-point boundary value problems*, preprint.

[8] L. GREENBERG AND I. BABUŠKA, *A continuous analogue of Sturm sequences in the context of Sturm–Liouville equations*, SIAM J. Numer. Anal., 26 (1989), pp. 920–945.

[9] M. MORSE, *The Calculus of Variations in the Large*, Amer. Math. Soc. Colloq. Publ., 18, Providence, RI, 1934.

[10] M. B. PORTER AND E. L. REISS, *A numerical method for acoustic normal modes for shear flows*, J. Sound Vibration, 100 (1985), pp. 91–105.

[11] W. T. REID, *Sturmian Theory for Ordinary Differential Equations*, Springer-Verlag, Berlin, New York, 1980.

[12] M. ROSEAU, *Vibrations in Mechanical Systems*, Springer-Verlag, Berlin, New York, 1987.

[13] H. WEINBERGER, *Variational Methods for Eigenvalue Approximation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1043-1064, July 1991

012

DIFFUSIVE LOGISTIC EQUATIONS WITH INDEFINITE WEIGHTS: POPULATION MODELS IN DISRUPTED ENVIRONMENTS II*

ROBERT STEPHEN CANTRELL† AND CHRIS COSNER†

Abstract. The dynamics of a population inhabiting a strongly heterogeneous environment are modeled by diffusive logistic equations of the form $u_t = \nabla \cdot (d(x, u)\nabla u) - \mathbf{b}(x) \cdot \nabla u + m(x)u - cu^2$ in $\Omega \times (0, \infty)$, where u represents the population density, $d(x, u)$ the (possibly) density dependent diffusion rate, $\mathbf{b}(x)$ drift, c the limiting effects of crowding, and $m(x)$ the local growth rate of the population. The growth rate $m(x)$ is positive on favorable habitats and negative on unfavorable ones. The environment Ω is bounded and surrounded by uninhabitable regions, so that $u = 0$ on $\partial\Omega \times (0, \infty)$. In a previous paper, the authors considered the special case $d(x, u) = d$, a constant, and $\mathbf{b} = 0$, and were able to make an analysis based on variational methods. The inclusion of density dependent diffusion and/or drift makes for more flexible and realistic models. However, variational methods are mathematically insufficient in these more complicated situations. By employing methods based on monotonicity and positive operator theory, many previous results on the dependence on m of the overall suitability of the environment can be recovered and some new results can be established concerning environmental quality dependence on $\mathbf{b}$. In the process, a bifurcation and stability analysis is made of the model which includes some new estimates on eigenvalues for associated linear problems.

Key words. diffusive logistic equations, heterogeneous environments, population dynamics, monotone flows, bifurcation and stability analysis, eigenvalue problems, indefinite weights

AMS(MOS) subject classifications. 35J65, 35K60, 92A15

1. Introduction.

Reaction-diffusion equations have been widely used as models for populations whose densities vary with location as well as time. If the environment is strongly heterogeneous, the coefficients describing the growth and diffusion of the population may vary with location as well. In an earlier article, we studied a diffusive logistic model in which the diffusion rate of the population was constant but the growth rate was assumed to vary with position, being positive in regions of favorable habitat and negative in unfavorable regions. The present article is devoted to extending our results to models in which the diffusion rate varies with position and population density, and the population may be subject to drift in addition to pure diffusion. These more complicated models incorporate effects which are often present in real situations, and thus can give more complete descriptions of biological phenomena. Since these models are quasilinear rather than semilinear, and in general are not in variational form, the technical aspects of the analysis are somewhat different and more difficult than in the case of simple Fickian diffusion. The technical complexity is inherent in the models we consider, and cannot be avoided if we are to give a rigorous analysis. However, we have introduced into our models only the sorts of effects which theoretical ecologists have suggested to us as being especially important.

Since the implications of our analysis should be of some interest to biological scientists, we give a brief summary of them in the last section, and conclude each section of the paper with a fairly detailed discussion of the main results from a biological viewpoint. Some of the discussion overlaps that given in [6], [7], and [9]. Additional references are given in [6] and [9]. In [7] we use relatively simple mathematics to analyze a number of special cases of the general models considered here and give a

* Received by the editors July 17, 1989; accepted for publication (in revised form) June 11, 1990. This research was supported by National Science Foundation grant DMS88-02346.

† Department of Mathematics and Computer Science, University of Miami, Coral Gables, Florida 33124.

moderately detailed biological discussion of their interpretation. A reader whose primary interest is in the biological aspects of the work may find the present article more accessible after reading [7].

The situation we wish to model is that of a species inhabiting a bounded region of variable habitat, and dispersing throughout that region via a process of diffusion which may be affected by population density and location and which may also involve drift due to wind, current, or environmental gradients. The questions we address are those of deciding how environmental factors and/or the density dependence of the diffusion affect the population. Obviously, if the overall environment includes too much poor habitat the population cannot be expected to persist. However, the *arrangement* of favorable and unfavorable regions turns out to play an important role in determining the overall suitability of an environment. Questions about the effects of the arrangement of favorable and unfavorable regions are especially important in refuge theory; for example, is one large preserve likely to be more or less effective in sustaining a population than several small preserves? Of course, the answers to such questions will depend on the details of the biology but models can suggest answers in some cases and serve to sharpen discussion in others. Island biogeography theory has been widely used in the context of refuge theory. The sort of models we consider provide an alternative approach. We discuss this point in some detail and give a number of references in [6]. One object of the present work is to extend the results of [6] to include models with more complex and realistic sorts of diffusion and drift processes. Another object is to study directly the effects of density dependent diffusion and of drift. We find that density dependence in the diffusion rate may have effects similar to those of depensation in the growth dynamics, as studied in [20]. Specifically, models with density dependent diffusion may admit multiple equilibria even if the corresponding dynamics with constant diffusion yield a unique equilibrium. The effects of drift in the case of a homogeneous environment and constant diffusion are discussed in [22]. We extend some of the results of [22] to models with variable diffusion and growth coefficients, but we have so far been unable to give a complete description of the effects of drift on the population dynamics.

There is a vast literature on traveling waves in reaction-diffusion models. Most of that literature is not directly relevant to our work, because we are concerned exclusively with bounded regions. A general overview of the literature on waves is given in [10] and [28]. Some specific problems related to wavelike propagation in ecological models and to considerations of domain size are studied in [4]. Another class of models which have a more direct relation to our work are patch models, where a population is assumed to inhabit a number of discrete patches rather than a continuous region. Some topics similar to those we consider are discussed from the viewpoint of patch models in [26]. (Threshold results for propagation through an infinite number of patches are derived in [3]; the specific model in [3] arises in neurophysiology, but the same methods would also apply to ecological models.) Our present work is most closely related to the ideas discussed in [20], [22], and [27], and of course [6], [9], where the models are primarily reaction-diffusion equations on bounded spatial domains.

The viewpoint we take in our modeling is essentially that taken in the pioneering work of Skellam [27], who deduced reaction-diffusion models for population growth and dispersal from the random-walk problem, and analyzed some of those models via classical methods. A representative result is that the density u of a population with linear growth law inhabiting a uniform disc of radius r_0 surrounded by a completely inhospitable region can be described by the equation $u_t = d\Delta u + mu$ with homogeneous Dirichlet boundary conditions, and hence will grow rather than decline provided

$m-(dj_1^2/r_0^2)>0$, where j_1 is the first zero of the Bessel function J_0. Another way to state this result is that the population will grow if the first eigenvalue λ_1 for the problem $-d\Delta\phi=\lambda m\phi$ on the disc of radius r_0 with homogeneous Dirichlet conditions satisfies $\lambda_1<1$. Skellam considered several other models, the most complicated being of the general form $u_t=d\Delta u+m(x)u-c(x)u^2$. Of such models, Skellam wrote (in 1951) that "orthodox analytical methods appear inadequate." (See [27, p. 212].) Since that time, there has been much work on reaction-diffusion models for population dynamics, and a number of new analytical methods have been introduced. For general background on the modeling aspects of population dynamics, see [19] or [24]; for mathematical methods and results, see [9], [10], or [28]. In our previous article [6], we study models that include those considered by Skellam, and discuss their biological interpretation.

The models in [6] have the form

$$u_t=d\Delta u+f(x,u)u \quad \text{in } \Omega, \qquad u=0 \quad \text{on } \partial\Omega, \tag{1.1}$$

where f is decreasing in u and Ω is a bounded domain in $\mathbb{R}^n$, with $n\leqq 3$ in applications. The intrinsic local growth rate of the population is given by $f(x,0)$, which is assumed to change sign on Ω, with positive values indicating favorable habitat and negative ones unfavorable habitat. Our results imply that (1.1) has a unique positive steady state which is a global attractor for nonnegative, nontrivial solutions, provided the first positive eigenvalue $\lambda_1(d,f(x,0))$ of the problem

$$-d\Delta\phi=\lambda f(x,0)\phi \quad \text{in } \Omega, \qquad \phi=0 \quad \text{on } \partial\Omega,$$

satisfies $\lambda_1(d,f(x,0))<1$. We also examined the question of how $\lambda_1(d,m(x))$ depends on the arrangement of positive and negative regions for $m(x)$. We showed that for a sequence $\{m_j(x)\}$ of weights, a necessary and sufficient condition for having

$$\lim_{j\to\infty}\lambda_1(d,m_j(x))=\infty \text{ is that } \limsup_{j\to\infty}\int_\Omega \psi m_j\leqq 0$$

for any $\psi\in L^1(\Omega)$ with $\psi\geqq 0$ almost everywhere. (This result is Theorem 3.1 of [6].) One implication of these results is that if the unfavorable regions are greater than or equal to the favorable ones in strength and extent, and the two sorts of regions are too closely interspersed, the population will not persist, *even though it might persist if the favorable habit formed a single larger region.* We also showed that in a certain sense, the most favorable situations will occur if there is a relatively large favorable region located some distance away from the boundary of Ω.

In the present article we consider models of the form

$$u_t=\nabla\cdot d(x,u)\nabla u-\mathbf{b}\cdot\nabla u+m(x)u-cu^2 \quad \text{in } \Omega, \qquad u=0 \quad \text{on } \partial\Omega, \tag{1.2}$$

and attempt to recover some of the results of [6]. Since the analysis of [6] was based largely on variational methods, we have had to substantially modify our techniques. In many cases, we replace ideas and results based on variational principles with others based on monotonicity or positive operator theory. Also, since we assume only $m(x)\in L^\infty(\Omega)$ (for various reasons which are discussed in [6]), we must work with weak solutions, so the standard Hopf maximum principle must generally be replaced by the maximum and comparison principles for weak solutions of elliptic equations discussed in [11, Chaps. 8, 9], or the corresponding results for the parabolic case which follow readily from similar arguments. (We do not always state this explicitly, and in some

cases we will simply cite references where the Hopf maximum principle is used but whose results extend directly to our situation via maximum principles for weak solutions.)

Models such as (1.2) display some different features than those of the form (1.1). In particular, if $d(x, u)$ is not monotone nondecreasing in u, (1.2) may have multiple positive solutions. (We give an example in § 3.) It is known that a similar phenomenon can occur in (1.1) in the presence of depensation (that is, if $f(x, u)$ is allowed to be increasing in u for some values of x and/or u) but not in the case of logistic growth. This situation is not surprising, since the equation $\nabla \cdot d(u)\nabla u + g(u) = 0$ can be converted to the form $\Delta U + G(U) = 0$ by letting $U = D(u)$ where $D(0) = 0$ and $D'(u) = d(u)$, and such a change of variables may destroy monotonicity or concavity properties of $g(u)$.

To analyze (1.2) we observe that the recent work of Hirsch [16] on monotone flows implies that the dynamics of (1.2) are determined by its steady states, we "unfold" the steady-state problem by introducing a parameter λ multiplying the undifferentiated terms, and we then analyze the steady states by using λ as a bifurcation parameter and applying the results of Rabinowitz [25]. Our main results state that under suitable restrictions on d, $\mathbf{b}$, and m, the problem (1.2) has a unique, stable, positive steady state provided $\lambda_1(d(x, 0), \mathbf{b}(x), m(x)) < 1$, where $\lambda_1(d, \mathbf{b}, m)$ is the first positive eigenvalue of

$$-\nabla d\nabla\phi + \mathbf{b}\cdot\nabla\phi = \lambda m\phi \quad \text{in } \Omega, \qquad \phi = 0 \quad \text{on } \partial\Omega, \tag{1.3}$$

and give a partial description of how that eigenvalue depends on d, $\mathbf{b}$, and m. In particular, we show that under a mild coercivity assumption on the left side of (1.3), the necessary and sufficient condition for $\lambda_1(d, 0, m_j(x)) \to \infty$ as $j \to \infty$ given in Theorem 3.1 of [6] extends to the case of $\lambda_1(d(x), \mathbf{b}(x), m_j(x))$. Since environments may vary in ways best described by discontinuous functions (for example, if a field is crossed by a paved road with sharp boundaries) we consider the case of $m \in L^\infty(\Omega)$ with $m > 0$ on a set of positive measure, but with m taking both positive and negative values. In that situation, we have to extend known results somewhat to obtain the existence of a first positive eigenvalue $\lambda_1(d, \mathbf{b}, m)$. The analysis is based on work of Hess and Kato [15] and Hess [14]. Our results on the behavior of $\lambda_1(d, \mathbf{b}, m)$ overlap slightly with those of Murray and Sperb [22], who considered the case of $\lambda_1(1, \mathbf{b}, 1)$. Other results implying bounds for eigenvalues for $\lambda_1(d, \mathbf{b}, m)$ under various hypotheses are given in [12], [13], [15], and [17], but they either do not apply in our situation or do not appear to be sharp enough for our purposes. We have observed that the presence of a drift term can either raise or lower $\lambda_1(d, \mathbf{b}, m)$. Our analysis of the existence problem for positive steady states of (1.2) is fairly complete, but to obtain uniqueness we must make additional structure assumptions (specifically that either $\partial d/\partial u \equiv 0$ or $\mathbf{b} \equiv 0$), and there remain many open questions about the dependence of $\lambda_1(d, \mathbf{b}, m)$ on d, $\mathbf{b}$, and m.

The paper is structured as follows. We derive the basic existence theory for positive equilibria in § 2, and obtain conditions on the uniqueness and stability of equilibria in § 3. Many of the results are somewhat technical, but they have some interesting biological implications. In § 4 we examine how the eigenvalue $\lambda_1(d, \mathbf{b}, m)$, whose size determines whether the model predicts extinction or persistence for the population, depends on the environment, drift, and diffusion. In § 5 we obtain some population estimates, again in terms of $\lambda_1(d, \mathbf{b}, m)$. Since the answers to the questions of greatest biological interest are determined by the size of $\lambda_1(d, \mathbf{b}, m)$, we consider the results of § 4 to have the greatest applied significance because they relate $\lambda_1(d, \mathbf{b}, m)$ to the physical conditions in the model. In § 6 we give a biologically oriented summary of

our conclusions. We also conclude each section with a description of the biological interpretation of the main results of that section.

2. A qualitative overview. In this section, we consider the positive steady-state solutions of the parabolic problem

$$\begin{aligned} &u_t = \nabla \cdot (d(x,u)\nabla u) - \mathbf{b}(x)\cdot\nabla u + \lambda(m(x)u - cu^2) && \text{in } \Omega\times(0,\infty),\\ &u(x,0) = u_0(x) \geqq 0 && \text{for } x\in\Omega,\\ &u(x,t) = 0 && \text{on } \partial\Omega\times(0,\infty). \end{aligned} \tag{2.1}$$

Here, as noted in the Introduction, λ is a real parameter and we wish to observe the structure of said solutions when viewed as a subset of an appropriate function space via global bifurcation theory [25] as well as determine the stability properties of the solution when viewed as solutions to (2.1). We assume that $d\in C^2(\bar{\Omega}\times\mathbb{R})$ such that $d(x,s)\geqq d_1>0$ for all $(x,s)\in\bar{\Omega}\times\mathbb{R}$, $\mathbf{b}\in[C^1(\bar{\Omega})]^n$, and $m\in L^\infty(\bar{\Omega})$ and consider

$$\begin{aligned} &-\nabla\cdot(d(x,u)\nabla u) + \mathbf{b}(x)\cdot\nabla u = \lambda(m(x)u - cu^2) && \text{in } \Omega,\\ &u = 0 && \text{on } \partial\Omega. \end{aligned} \tag{2.2}$$

Observe first of all that (2.2) can be expressed as

$$\begin{aligned} &-\Delta u + \left[\frac{\mathbf{b}(x) - \mathbf{d}_x(x,0)}{d(x,0)}\right]\cdot\nabla u\\ &\quad = \lambda\frac{m(x)}{d(x,0)}u + \left[\frac{d_u(x,u)}{d(x,u)}|\nabla u|^2 + \left(\frac{\mathbf{d}_x(x,u)}{d(x,u)} - \frac{\mathbf{d}_x(x,0)}{d(x,0)}\right)\cdot\nabla u\right.\\ &\quad\quad + \left(\frac{1}{d(x,0)} - \frac{1}{d(x,u)}\right)(\mathbf{b}(x)\cdot\nabla u) + \lambda\left(\frac{1}{d(x,u)} - \frac{1}{d(x,0)}\right)m(x)u - \left.\frac{\lambda cu^2}{d(x,u)}\right] \quad \text{in } \Omega,\\ &u = 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.3}$$

Denote the expression in brackets in (2.3) by $H(\lambda,u)$. Then for a sufficiently large p, $H:\mathbb{R}\times W_0^{1,p}(\Omega)\to L^{p/2}(\Omega)$ is continuous and $\lim_{\|u\|_{1,p}\to 0} H(\lambda,u)/\|u\|_{1,p} = 0$, where $\|\ \|_{1,p}$ denotes the norm in $W_0^{1,p}(\Omega)$ and the limit is uniform for λ contained in compact intervals. (That such is the case relies on the fact $W_0^{1,p}(\Omega)$ embeds into $C_0^\alpha(\bar{\Omega})$ for sufficiently large p.) Consequently, if L denotes the elliptic operator on the left-hand side of (2.3) subject to zero Dirichlet boundary data and M/D denotes multiplication by $m(x)/d(x,0)$, a solution u to (2.3) is equivalent to a solution u of

$$u = \lambda L^{-1}\left(\frac{M}{D}\right)u + L^{-1}H(\lambda,u). \tag{2.4}$$

Since $L^{-1}: L^{p/2}(\Omega)\to W^{2,p/2}(\Omega)\cap W_0^{1,p/2}(\Omega)$ is continuous, $W^{2,p/2}(\Omega)\cap W_0^{1,p/2}(\Omega)$ embeds compactly into $C_0^{1+\alpha}(\bar{\Omega})$, $0<\alpha<1$ for p sufficiently large, and $C_0^{1+\alpha}(\bar{\Omega})$ embeds into $W_0^{1,p}(\Omega)$ for any p, the right-hand side of (2.4) may be viewed as a completely continuous operator on $W_0^{1,p}(\Omega)$ for a sufficiently large p with $\lim_{\|u\|_{1,p}\to 0}\|L^{-1}H(\lambda,u)\|_{1,p}/\|u\|_{1,p} = 0$ uniformly for λ in compact intervals. Consequently, $\mathbb{R}\times W_0^{1,p}(\Omega)$ is an appropriate space in which to apply global bifurcation theory [25].

In order to invoke global bifurcation theory to guarantee the existence of a continuum of positive solutions to (2.2) in $\mathbb{R}\times W_0^{1,p}(\Omega)$, it suffices to establish that there is a unique $\lambda = \lambda_1(m) > 0$ such that

$$v = \lambda L^{-1}\left(\frac{M}{D}\right)v \tag{2.5}$$

has as generalized null space the span of a positive function. Note that (2.5) is equivalent

to

$$\begin{aligned} &-\nabla\cdot(d(x,0)\nabla v)+\mathbf{b}(x)\cdot\nabla v=\lambda m(x)v \quad \text{in } \Omega,\\ &v=0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{on } \partial\Omega. \end{aligned} \tag{2.6}$$

In the special cases $\mathbf{b}\equiv\mathbf{0}$ or $m\in C(\bar{\Omega})$, the result follows from the results of [21] and [15], respectively, provided that $\{x\in\Omega: m(x)>0\}$ has positive measure. In the case that $\mathbf{b}\not\equiv\mathbf{0}$ and $m\in(L^{\infty}(\Omega)-C(\bar{\Omega}))$, to our knowledge, the result does not explicitly appear in the literature. Since such is the case and since the result is of independent interest, we include a brief proof.

THEOREM 2.1. *Suppose that d, $\mathbf{b}$, and m are as described above and that $\{x\in\Omega: m(x)>0\}$ has positive measure. Then there is a unique $\lambda=\lambda_1(m)>0$ such that*

$$\begin{aligned} &-\Delta v+\left[\frac{\mathbf{b}(x)-\mathbf{d}_x(x,0)}{d(x,0)}\right]\cdot\nabla v=\lambda\frac{m(x)}{d(x,0)}v \quad \text{in } \Omega,\\ &v=0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{on } \partial\Omega \end{aligned} \tag{2.7}$$

has a solution $v\in C_0^{1+\alpha}(\bar{\Omega})$ with $v(x)>0$ in Ω and $(\partial v/\partial n)(x)<0$ on $\partial\Omega$. Moreover, $\cup_{r\geqq 1}N\{(I-\lambda L^{-1}(M/D))^r\}=\langle v\rangle$.

Proof. The uniqueness and simplicity assertions of the theorem follow as in [15] once the existence of such a λ has been established. To this end, let $R>0$ be such that $m(x)/d(x,0)+R>0$ on Ω almost everywhere and consider the operator $A_\lambda=\lambda(L+R\lambda)^{-1}(M/D+R)$, which may be viewed as a compact positive operator on $C_0^0(\bar{\Omega})$. A_λ is continuous in λ and consequently so is its spectral radius $r(A_\lambda)$ [23]. Moreover, $\lim_{\lambda\to 0}r(A_\lambda)=0$. Hence, as in [8], the existence of an eigenvalue λ with the required properties follows from the Krein–Rutman theorem and the maximum principle as long as there is a $\lambda>0$ so that $r(A_\lambda)\geqq 1$. The assumption that $\{x\in\Omega: m(x)>0\}$ has positive measure guarantees that (2.7) has infinitely many eigenvalues with positive real part [14, Thm. 2]. For any such eigenvalue λ^* and any associated eigenfunction v, Lemma 3 of [15] implies that

$$|v|\leqq A_{\mathrm{Re}\,\lambda^*}|v|.$$

Hence $r(A_{\mathrm{Re}\,\lambda^*})\geqq 1$, and the result is established.

It is of substantial interest from the biological point of view not only to have the existence of an unbounded continuum of positive solutions to (2.2) but also to know there is a solution (λ,u) on the continuum for all $\lambda>\lambda_1(m)$. Such an observation requires information in addition to that provided by global bifurcation theory. The *a priori* estimates given in the following theorem are sufficient for this purpose.

THEOREM 2.2. *Suppose (λ,u) is a positive solution to (2.2) and that $\lambda\in[a,b]$, where $0\leqq a\leqq b<\infty$. Then there is a constant $K>0$ such that $\|u\|_{1,p}\leqq K$.*

Proof. We know that $u\in W^{2,p}(\Omega)\cap W_0^{1,p}(\Omega)$ and consequently $u\in C_0^{1+\alpha}(\bar{\Omega})$. Hence, as in § 2 of [6], the maximum principle implies that $\|u\|_{\infty}\leqq$ ess $\sup_{x\in\bar{\Omega}}(m^+(x)/c)$. The result then will follow if we can show $\|\nabla u\|_{\infty}$ is bounded uniformly with respect to $\lambda\in[a,b]$.

To this end, we employ results in §§ 4 and 5 of Chapter 4 and in § 2 of Chapter 6 of [18]. Equation (2.2) satisfies the ellipticity and structure conditions there imposed, with ellipiticity (and other constants) bounded for $\lambda\in[a,b]$. Moreover, as noted, $0\leqq u(x)\leqq$ ess $\sup_{x\in\bar{\Omega}}(m^+(x)/c)$. By the proof of Theorem 4.1 of [18, Chap. 4], ess $\sup_{\Omega}|\nabla u|$ can then be bounded in terms of ess $\sup_{\partial\Omega}|\nabla u|$ and integrals which are in essence $\|u\|_{2,2}^2$ and $\|u\|_{1,4}^2$. Theorem 5.1 of [18, Chap. 4] implies that these last integrals are bounded in terms of $\|u\|_{\infty}$ and constants depending on Ω and the coefficients of (2.2), all of which are uniformly bounded for $\lambda\in[a,b]$.

In order to see that ess $\sup_{\partial\Omega} |\nabla u|$ is uniformly bounded with respect to $\lambda \in [a, b]$, we employ Lemma 4.1 of [18, Chap. 4] or Lemma 2.1 of [18, Chap. 6]. (These are two statements of the same result.) The idea behind the lemmas is Bernstein's, namely, to compare u with an appropriate auxiliary function via differential inequalities and the maximum principle to obtain bounds on $\partial u/\partial n$ on subsets of $\partial\Omega$. In [18], the classical maximum principle is used, and u is assumed to have classical second derivatives throughout Ω. However, we may replace the classical maximum principle with the maximum principle for weak solutions as stated in Chapter 8 of [11] and only require that $u \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$. The bound we obtain depends only on Ω, $\|u\|_\infty$, and ellipticity and structure constants which are bounded as long as $\|u\|_\infty$ and λ are uniformly bounded so the result follows.

Consequently, there is an unbounded continuum $\mathscr{C}$ in $\mathbb{R} \times W_0^{1,p}(\Omega)$ of positive solutions to (2.2) with the property that if $(\lambda, u) \in \mathscr{C}$ and $\lambda \in [a, b]$, where $0 \leqq a < b < \infty$, there is a $K > 0$ such that $\|u\|_{1,p} \leqq K$. Furthermore, it is evident that (2.2) has only the trivial solution when $\lambda = 0$. Hence, the projection $\Pi(\mathscr{C})$ into $\mathbb{R}$ of $\mathscr{C}$ must satisfy $(\lambda_1(m), \infty) \subseteq \Pi(\mathscr{C}) \subseteq (0, \infty)$. In particular, there is at least one positive solution of (2.2) for every $\lambda > \lambda_1(m)$.

For any fixed λ, $(\lambda, u) \in \mathscr{C}$ of course implies that u is an equilibrium solution to (2.1). It is sometimes possible to determine that u is globally asymptotically stable with respect to smooth initial data $u_0(x) \geqq 0$. We first observe that if $\lambda > \lambda_1(m)$, then the zero solution of (2.2) is unstable. To this end, observe that the linearization A of $-\nabla \cdot (d(x, u)\nabla u) + \mathbf{b}(x) \cdot \nabla u - \lambda(m(x)u - cu^2)$ with respect to u at $u = 0$ is given by $A(\phi) = -\nabla \cdot (d(x, 0)\nabla\phi) + \mathbf{b}(x) \cdot \nabla\phi - \lambda m(x)\phi$ and that the zero solution is unstable provided that

$$A\phi = \sigma\phi \quad \text{in } \Omega,$$

$$\phi = 0 \quad \text{on } \partial\Omega,$$

and $\phi(x) > 0$ in Ω implies that $\sigma < 0$. If $\sigma \geqq 0$, then $v = \phi$ is a positive solution to the inhomogeneous boundary value problem

$$-\nabla \cdot (d(x, 0)\nabla v) + \mathbf{b}(x) \cdot \nabla v = \lambda m(x)v + h \quad \text{in } \Omega,$$

$$v = 0 \quad \text{on } \partial\Omega,$$

where $h = \sigma\phi \geqq 0$. As $\lambda > \lambda_1(m)$, Proposition 3 of [15] is violated. As a consequence, the zero solution of (2.2) is unstable as an equilibrium to (2.1) if $\lambda > \lambda_1(m)$, and we are able to establish the following theorem.

THEOREM 2.3. *Suppose that for some $\lambda > \lambda_1(m)$, there is a unique positive solution $\bar{u}$ to* (2.2). *Then $\bar{u}$ is a globally asymptotically stable equilibrium for* (2.1) *provided we require the initial data u_0 to lie in an appropriate Sobolev–Slobedickii space $W_0^{\sigma,p}$. This will be the case if $u_0 \in C_0^2(\bar{\Omega})$, for example.*

Proof. The methods of Amann [1], [2] imply that (2.1) generates a monotone flow on a Sobolev–Slobedickii space $W_0^{\sigma,p}(\Omega)$ with $W_0^{\sigma,p}(\Omega) \subseteq W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega) \subseteq C^1(\bar{\Omega})$. (See also [16, Thm. 4.6]. We assume slightly less regularity than Amann since $m(x) \in L^\infty(\Omega)$, but a comparison principle for weak solutions to quasilinear parabolic problems can be readily obtained by modifying the proof of Theorem 9.5 of [11] to treat the parabolic case, so this is not a problem.) Consequently, the results of [16] are applicable in this situation. Since the zero solution of (2.2) is an unstable equilibrium and since $[0, \bar{u}]$ is an order interval containing no other equilibria, Theorem 0.6 of [16] implies that if $0 \leqq \bar{u}_0(x) \leqq \bar{u}(x)$ and $u_0(x) \not\equiv 0$, then the solution $u(x, t)$ of (2.1) corresponding to $u_0(x)$ converges to $\bar{u}(x)$ as $t \to \infty$, uniformly on $\bar{\Omega}$. Moreover, for any sufficiently

large constant K, Theorem 0.7 of [16] implies that the solution $u_K(x,t)$ of (2.1) corresponding to initial condition K converges to $\bar{u}(x)$ as $t\to\infty$, uniformly on $\bar{\Omega}$. The result now follows from monotonicity, since for any smooth initial data $v(x)>0$ and $v\not\equiv 0$, we can find $u_0(x)\in[0,\bar{u}]$ and K sufficiently large so that $u_0(x)\leqq v(x)\leqq K$. We conclude this section with the following result, which is a corollary to Theorems 2.1–2.3.

THEOREM 2.4. *Consider equation* (2.6) *and let* $\lambda_1(d,\mathbf{b},m)$ *be as in Theorem* 2.1. *If* $\lambda_1(d,\mathbf{b},m)<1$, *then the problem* (2.1) *with* $\lambda=1$ *has a positive equilibrium solution* $\bar{u}$. *If* $\bar{u}$ *is unique, then it is globally asymptotically stable with respect to smooth initial data* $u_0(x)$.

2.1. Biological interpretation. The primary result of biological interest in this section is Theorem 2.4. That result asserts that there exists a positive equilibrium density for the population being modeled provided that the eigenvalue $\lambda_1(d,\mathbf{b},m)$ is less than 1. The significance of the result lies in the fact that $\lambda_1(d,\mathbf{b},m)$ depends directly on the terms in the model describing biological properties of the population and the environment. Thus Theorem 2.4 provides a criterion for the possible persistence of a population in terms of diffusion, drift, and growth rates which vary with location. In some simple cases it is possible to compute $\lambda_1(d,\mathbf{b},m)$ as the solution of an equation involving trigonometric and hyperbolic functions (which can be approximated by Newton's method). This is discussed in [7]; an example is given below.

In general, the numerical problem of finding $\lambda_1(d,\mathbf{b},m)$ is fairly difficult but has been studied extensively. Approximation schemes for the case $\mathbf{b}=0$ (no drift) are discussed in detail in [29]. There is a substantial literature on numerical approximation for solutions of eigenvalue problems with or without drift terms; [29] gives a large number of references. It is not surprising that the computation of the eigenvalue λ_1 may be complicated, since if λ_1 gives a reasonable synthesis of the various factors such as the size, shape, and quality of the environment and the effects of winds, currents, temperature or chemical gradients, it must reflect a large number of complex biological factors. In giving an accurate description of a complex phenomenon, a certain amount of mathematical sophistication may be required. Even so, the computational problem of finding $\lambda_1(d,\mathbf{b},m)$ is likely to be simpler than that of evaluating the results of a comparably detailed simulation.

A major advantage of having a criterion for persistence based on $\lambda_1(d,\mathbf{b},m)$ is that it is possible to make a number of *qualitative* statements about the ways in which changes in the environment affect a population. That is the main theme of [6] and [7] and of § 4 of this article. We discuss the topic at some length in § 4.

As an example, suppose that we consider a one-dimensional region $\Omega=(0,\ell)$, with no drift, constant diffusion rate, and a growth rate $m(x)$ which is a positive constant on a subinterval of Ω and a negative constant on the remainder of Ω. For appropriate d, this problem can easily be rescaled into the form

$$\begin{aligned} &u''+m(x)u-cu^2=0 \quad \text{on } (0,1),\\ &u(0)=u(1)=0 \end{aligned} \tag{2.8}$$

with

$$m(x)=\begin{cases} -1, & 0<x<a,\\ k, & a<x<a+T,\\ -1, & a+T<x<1, \end{cases} \tag{2.9}$$

for some $T<1$ describing the relative size of the favorable region, some $a\in[0,1-T]$ describing its location relative to the boundary, and some k describing the relative

quality of the favorable habitat compared with the unfavorable. (The diffusion coefficient is scaled into $m(x)$; the carrying capacity relative to habitat quality is described by c but does not directly enter the computation of λ_1.) We show in [7] that for (2.8), (2.9), we have $\lambda_1 = \alpha^2$ where α is the smallest positive solution of

$$\cot \alpha\sqrt{k}\, T = \frac{k \tanh[\alpha(1-a-T)] \tanh \alpha a - 1}{\sqrt{k}\,[\tanh \alpha a + \tanh \alpha(1-a-T)]}. \tag{2.10}$$

Note that for a uniformly favorable environment we have $\lambda_1 = \alpha^2 = \pi^2/k$ so that we can expect persistence only for $k > \pi^2$. For $k = 16$, $T = 1$, $a = 0$ (indicating a uniformly favorable environment), we have $\lambda_1 \approx .61$. For $k = 16$, $a = .1$, $T = .8$ we find by solving (2.10) that $\lambda_1 \approx .63$. For $k = 16$, $a = .3$, $T = .4$, $\lambda_1 \approx .86$. The results of [7] show how in (2.8) a number of other forms of $m(x)$ can be treated via equations similar to (2.10).

3. Uniqueness. In this section we shall consider the question of uniqueness for positive steady states in our model. Our analysis includes some results on the direction of bifurcation with respect to the unfolding parameter λ, and on the linearized stability of the steady state. We begin with an example that shows that some restrictions are needed if uniqueness is to hold. In the general semilinear problem $\Delta u + f(x, u) = 0$, some conditions must be imposed on f to obtain uniqueness, and the problem $\nabla \cdot d(u)\nabla u + m(x)u - cu^2 = 0$ is equivalent to the semilinear problem $\Delta U + m(x)D^{-1}(U) - c[D^{-1}(U)]^2 = 0$ where $U = D(u)$ with $D'(s) = d(s)$, $D(0) = 0$; so we must expect that some conditions will be needed on $d(s)$ if the corresponding semilinear problem is to have a unique solution. The nature of those conditions is indicated by the problem

$$(d(u)u')' + \lambda(u - u^2) = 0, \qquad u > 0 \quad \text{on } (0, \pi), \qquad u(0) = u(\pi) = 0, \tag{3.1}$$

where $d(s) = 1 - 2d_0 s$ for $0 \leqq s \leqq d_0/4$, $d(s)$ is smooth for $0 \leqq s < \infty$, and $d(s) \geqq d_1 > 0$. By the analysis in § 2, a branch of positive solutions to (3.1) bifurcates from the trivial solution at $\lambda = 1$. If we multiply the equation in (3.1) by u, integrate by parts, and use the fact that

$$\int_0^\pi u'(x)^2\, dx \geqq \int_0^\pi u(x)^2\, dx,$$

then we obtain the relation

$$d_1 \int_0^\pi u^2\, dx \leqq d_1 \int_0^\pi (u')^2\, dx \leqq \int_0^\pi d(u)(u')^2\, dx = \lambda \int_0^\pi u^2\, dx - \lambda \int_0^\pi u^3\, dx.$$

Since $u > 0$ on $(0, \pi)$, it follows that $\lambda \geqq d_1 > 0$. Also, $0 < u < 1$ on $(0, \pi)$ by the maximum principle, so a standard application of the Rabinowitz global bifurcation theorem implies that the branch of positive solutions emanating from the zero solution at $\lambda = 1$ must meet infinity in λ. However, if we multiply (3.1) by $\sin x$ and integrate by parts, then as long as $0 \leqq u \leqq d_0/4$ (which will be true locally near the bifurcation point) we have

$$\begin{aligned}
\lambda \int_0^\pi u \sin x\, dx - \lambda \int_0^\pi u^2 \sin x\, dx &= -\int_0^\pi (d(u)u')' \sin x\, dx \\
&= -\int_0^\pi u'' \sin x\, dx + \int_0^\pi (2d_0 uu')' \sin x\, dx \\
&= \int_0^\pi u \sin x\, dx - \int_0^\pi d_0 u^2 \sin x\, dx.
\end{aligned}$$

Hence, as long as $0 < u \leqq d_0/4$, we have

$$\lambda \int_0^\pi (u-u^2)\sin x\,dx = \int_0^\pi (u-u^2)\sin x\,dx + (1-d_0)\int_0^\pi u^2 \sin x\,dx,$$

so if $d_0 > 1$, we must have $\lambda < 1$. But the branch of solutions must meet infinity in λ, so there must be a solution with $\sup u > d_0/4$ corresponding to $\lambda = 1$. It follows from the fact that the branch of positive solutions is a continuum and the leftward direction of bifurcation that for some $\varepsilon > 0$, the problem (3.1) has at least two solutions, one with $\sup u \leqq d_0/4$ and one with $\sup u > d_0/4$, for $\lambda = 1-\varepsilon$.

To avoid the phenomenon observed in this example, we must ensure that the bifurcation is to the right rather than to the left. If we have $\lambda > \lambda_1(d(x,0), \mathbf{b}, m)$ for all positive solutions and they are all linearly stable, that is enough for uniqueness.

THEOREM 3.1. *Assume that $d(x,u)$ is of class C^1. Suppose that for any solution of*

$$\begin{aligned} &\nabla d(x,u)\nabla u - \mathbf{b}(x)\cdot\nabla u + \lambda(m(x)u - cu^2) = 0 && \text{in } \Omega, \\ &u = 0 && \text{on } \partial\Omega, \\ &\lambda > 0, \quad u > 0 && \text{in } \Omega, \end{aligned} \tag{3.2}$$

we have $\lambda > \lambda_1(d(x,0), \mathbf{b}, m)$, and that the first eigenvalue of the linearized problem

$$\begin{aligned} &-\nabla\cdot d(x,u)\nabla\phi - \nabla\cdot\phi\frac{\partial d(x,u)}{\partial u}\nabla u + \mathbf{b}\cdot\nabla\phi + \lambda(2cu-m)\phi = \sigma\phi \quad \text{in } \Omega, \\ &\phi = 0 \qquad\qquad \text{on } \partial\Omega \end{aligned} \tag{3.3}$$

satisfies $\sigma_1 > 0$ for any positive solution u. Then the positive solution for (3.2) *is unique for any given λ.*

Remarks. Combined with the comparison principle for the corresponding parabolic problem and the results of Hirsch [16], uniqueness implies stability. The condition $\sigma_1 > 0$ already implies linearized (and hence local) stability.

Proof. Choose $p > 1$ large enough that $W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$ embeds in $C_0^{1+\alpha}(\Omega)$ and $W^{1,p/2}(\Omega)$ embeds in $C^\beta(\Omega)$ for some $\alpha, \beta \in (0,1)$. Then the nonlinear function $F(\lambda, u) = \nabla\cdot d(x,u)\nabla u - \mathbf{b}(x)\cdot\nabla u + \lambda(m(x)u - cu^2)$ maps $\mathbb{R}\times(W^{2,p}(\Omega)\cap W_0^{1,p}(\Omega)) \to L^p(\Omega)$. The map is continuously differentiable, and the derivative with respect to the second variable is the negative of the operator on the left side of (3.3). If $\sigma_1 > 0$ in (3.3), then the linearized operator is invertible from L^p to $W^{2,p}\cap W_0^{1,p}$ by standard elliptic theory. Thus, if (λ_0, u_0) satisfies $F(\lambda, u) = 0$ and $u_0 > 0$ in Ω, then there is a bounded neighborhood U of u_0 in $W^{2,p}\cap W_0^{1,p}$, an interval $\Lambda = (\lambda_0 - \delta, \lambda_0 + \delta)$ with $\delta > 0$, and a function $g: \Lambda \to U$ such that for any $\lambda \in \Lambda$, the unique solution in U of (3.2) is $u = g(\lambda)$. Let $\lambda_k \uparrow \lambda_0 + \delta$ as $k \to \infty$. By (3.2), we have for $u_k = g(\lambda_k)$

$$\begin{aligned} u_k &= -\Delta^{-1}[(\nabla d(x,u_k)\cdot\nabla u_k - \mathbf{b}\cdot\nabla u_k + \lambda_k(mu_k - cu_k^2))/d(x,u_k)] \\ &\equiv -\Delta^{-1} w_k. \end{aligned} \tag{3.4}$$

Since U is bounded in $W^{2,p}\cap W_0^{1,p}$, so is $\bar{U}$; thus the right side of (3.4) is of the form $-\Delta^{-1}w_k$, where $\{w_k\}$ is uniformly bounded in $W^{1,p/2}$. (Here the $p/2$ is due to the presence on the right side of (3.4) of terms of the form $(\partial d/\partial u)|\nabla u|^2$; also, we have used the fact that $d \geqq d_1 > 0$, the embedding $W^{2,p}\cap W_0^{1,p} \hookrightarrow C_0^{1+\alpha}$, and the differentiability of d.) Our choice of p is such that $W^{1,p/2} \hookrightarrow C^\beta$, so since C^β embeds compactly in

C^0, we may choose a subsequence and reindex so that $\{w_k\}$ converges in C^0, and hence in L^p. Then (3.4) implies that the sequence $\{u_k\}$ converges in $W^{2,p} \cap W_0^{1,p}$, thus producing a nonnegative solution to (3.2) at $\lambda = \lambda_0 + \delta$. A similar argument applies at $\lambda = \lambda_0 - \delta$. For these values of λ, the solution can be extended further if it is positive. However, if we choose $K > 0$ sufficiently large we have from (3.2) that

$$-\nabla \cdot d(x, u)\nabla u + \mathbf{b}(x) \cdot \nabla u + Ku = [\lambda m(x) - \lambda cu + K]u \geqq 0,$$

so if $u \geqq 0$ and $u = 0$ somewhere in Ω, then $u = 0$ almost everywhere by the strong maximum principle for weak solutions (see [11, Thm. 8.19]). Hence, the only way that continuation in λ can fail is if λ is a bifurcation point from the branch of trivial solutions. The unique point where positive solutions can bifurcate from that branch is $\lambda = \lambda_1(d(x, 0), \mathbf{b}, m)$. It follows that if (3.2) has a positive solution u_0 for some λ_0, then there is a curve $(\lambda, u(\lambda))$ of positive solutions passing through (λ_0, u_0) which can be extended at least until $\lambda = \lambda_1(d(x, 0), \mathbf{b}, m)$. Suppose that for some $\lambda_0 > \lambda_1(d(x, 0), \mathbf{b}, m)$ there are two distinct positive solutions of (3.2). Then each lies on an arc which extends until $\lambda = \lambda_1(d(x, 0), \mathbf{b}, m)$, and the arcs cannot intersect as long as the solutions to (3.2) remain positive. If an arc contains a positive solution to (3.2) at $\lambda = \lambda_1(d(x, 0), \mathbf{b}, m)$, then the preceding argument based on the implicit function theorem implies that there are positive solutions of (3.2) on $\lambda \in (\lambda_1 - \delta, \lambda_1)$, for some $\delta > 0$, contradicting our hypotheses. But both branches cannot connect to the zero solution at the point $\lambda = \lambda_1(d(x, 0), \mathbf{b}, m)$, since the Crandall–Rabinowitz constructive bifurcation theorem for simple eigenvalues implies that there is a unique branch of positive solutions in some neighborhood of the bifurcation point. Hence, assuming the existence of two distinct positive solutions for some λ yields a contradiction, so for any λ the positive solution of (3.2) must be unique.

Remark. A similar argument is used to obtain a uniqueness theorem for a diffusive Lotka–Volterra competition model in [5].

So far, we have been unable to establish that the hypotheses of Theorem 3.1 are satisfied, in general, for equations of the form (3.2). However, we can show that they will be met if the differential operator in (3.2) is either linear or in divergence form. The two cases require different arguments, so we consider them separately.

COROLLARY 3.2. *Suppose that* $\mathbf{b} \equiv 0$ *and* $\partial d(x, u)/\partial u \geqq 0$ *for all* $x \in \Omega$ *and* $u \in [0, \operatorname{ess\,sup}(m^+/c)]$. *Then* (3.2) *has a unique positive solution for* $\lambda > \lambda_1(d(x, 0), 0, m)$.

Proof. Suppose that (3.2) has a positive solution for some $\lambda > 0$. Multiplying by u, integrating by parts, and using the hypothesis that $d(x, u)$ is monotone increasing in u, we have

$$0 < \int_\Omega d(x, 0)|\nabla u|^2 \, dx < \int_\Omega d(x, u)|\nabla u|^2 + \lambda c \int_\Omega u^3 \, dx = \lambda \int_\Omega m(x)u^2 \, dx.$$

Since $\int_\Omega mu^2 \, dx > 0$, it follows from results in [16] that

$$\int_\Omega d(x, 0)|\nabla u|^2 \, dx \geqq \lambda_1(d(x, 0), 0, m) \int_\Omega m(x)u^2 \, dx,$$

so that

$$\lambda_1(d(x, 0), 0, m) \int_\Omega mu^2 \, dx < \lambda \int_\Omega mu^2 \, dx$$

and

$$\lambda_1(d(x, 0), 0, m) < \lambda,$$

as required.

Suppose that $\phi_1 > 0$ is an eigenfunction for (3.3) corresponding to the principa eigenvalue σ_1. Multiplying (3.3) by u and integrating by parts yields

$$\sigma_1 \int_\Omega u\phi_1 \, dx = \int_\Omega \left[-\nabla \cdot d(x, u)\nabla u + \frac{\partial d(x, u)}{\partial u} |\nabla u|^2 + 2\lambda cu^2 - \lambda mu \right] \phi_1 \, dx,$$

so by (3.2) we have

$$\sigma_1 \int_\Omega u\phi_1 \, dx = \int_\Omega \left[\frac{\partial d(x, u)}{\partial u} |\nabla u|^2 + \lambda cu^2 \right] \phi_1 \, dx > 0$$

and hence $\sigma_1 > 0$ as required.

COROLLARY 3.3. *Suppose that* $d(x, u) \equiv d(x, 0)$, *so that* (3.2) *is semilinear. The* (3.2) *has a unique positive solution for* $\lambda > \lambda_1(d(x, 0), \mathbf{b}, m)$.

Proof. Suppose that (3.2) has a positive solution for some $\lambda > 0$. Then we have

$$-\nabla \cdot d(x, 0)\nabla u + \mathbf{b}(x) \cdot \nabla u = \lambda(m(x) - cu)u,$$

where $u > 0$ in Ω and $u = 0$ on $\partial\Omega$, so that

$$\lambda = \lambda_1(d(x, 0), \mathbf{b}, m - cu) > \lambda_1(d(x, 0), \mathbf{b}, m)$$

by the monotonicity of the positive principal eigenvalue with respect to the weigh (see [11]). In this case, (3.2) can be written as

$$-\nabla \cdot d(x, 0)\nabla u + \mathbf{b} \cdot \nabla u + \lambda(2cu - m)u = \lambda cu > 0. \tag{3.5}$$

Since (3.5) admits a positive solution u for the positive inhomogeneous term λcu it follows that the principal eigenvalue for the operator $L\phi \equiv -\nabla \cdot d(x, 0)\nabla\phi - \mathbf{b} \cdot \nabla\phi + \lambda(2cu - m)\phi$ must be positive. Since $\partial d(x, u)/\partial u \equiv 0$, that eigenvalue is σ_1 so $\sigma_1 > 0$.

Remark. It would be of interest to find a natural set of conditions including thos of both corollaries under which the hypotheses of Theorem 3.1 are satisfied. So far w have been unable to find such general conditions. It is well known that results fo quasilinear problems not in divergence form are typically much weaker and/or mor difficult than for either linear or divergence form problems (see the discussion in [7 and [14]). There are numerous open questions about uniqueness even in the case o ordinary differential equations.

3.1. Biological interpretation. The results of this section serve largely to sharpe those of the previous section by giving criteria for the uniqueness of the positive stead state for the population. Uniqueness is important in the context of our models becaus it implies the global stability of the positive steady state and thus the persistence o the population. Perhaps the most interesting observation from a biological viewpoir is that uniqueness may *fail* if the rate of diffusion is allowed to decrease with respec to the population density. Such a phenomenon occurs when the diffusion rate i constant but the logistic growth term is replaced by something of the form $uf(x, u$ with $f(x, u)$ sometimes increasing with u; in other words, in the presence of depensatio in the growth rate. That observation is made in [20] in connection with a model fo the population dynamics of the spruce budworm. As far as we know, it has not bee observed previously that the same sort of effect can be induced by density depender diffusion, which can sometimes produce multiple steady states even with a simpl logistic growth term. In the absence of drift, we show that such an effect can onl occur if the diffusion rate decreases relative to the population density at some densities We have not been able to determine the effects of drift terms on this phenomenon. I

the case of a density independent diffusion rate, we show that a positive steady state is unique and stable if it exists, and similarly for models with no drift and a diffusion rate which increases with population density.

4. Properties of the principal eigenvalue. We have seen that under fairly general hypotheses, the condition $\lambda_1(d(x,0), \mathbf{b}(x), m(x)) < 1$ is sufficient for the existence of a positive steady state for our model, and under somewhat stronger hypotheses the condition is also necessary and the positive steady state is unique and hence stable. Thus, it is natural to ask how $\lambda_1(d(x,0), \mathbf{b}(x), m(x))$ depends on d, $\mathbf{b}$, and m. Some results for the case $d(x,0) \equiv 1$, $m(x) \equiv 1$ are given in [21], and for the case $d(x,0) \equiv 1$ and $\mathbf{b} \equiv 0$ in [6].

Our first result is an extension of Theorem 3.1 of [6].

THEOREM 4.1. *Suppose that* $d(x) = d(x,0) \in C^{1+\alpha}(\bar{\Omega})$, $\mathbf{b}(x) = (b_1, \cdots, b_n)$ *with* $b_i(x) \in C^{\alpha}(\bar{\Omega})$ *for* $i = 1, \cdots, n$, *and* $m_j(x) \in L^{\infty}(\Omega)$ *for* $j = 1, 2, \cdots$. *Suppose that* d *and* $\mathbf{b}$ *are such that for any* $\phi \in W_0^{1,2}(\Omega)$ *we have*

$$\int_{\Omega} d|\nabla\phi|^2 + (\mathbf{b}\cdot\nabla\phi)\phi \geqq d_0 \int_{\Omega} |\nabla\phi|^2 \tag{4.1}$$

for some $d_0 > 0$, *and for each* j,

$$\|m_j\|_{\infty} \leqq m_0 \quad \text{and} \quad m_j > 0 \quad \text{on a set of positive measure.} \tag{4.2}$$

To have $\lim_{j\to\infty} \lambda_1(d, \mathbf{b}, m_j) = \infty$, *it is necessary and sufficient that*

$$\limsup_{j\to\infty} \int_{\Omega} m_j\beta \leqq 0 \tag{4.3}$$

for all $\beta \in L^1(\Omega)$ *with* $\beta \geqq 0$ *almost everywhere.*

Proof. Suppose that (4.1), (4.2), and (4.3) hold but $\lambda_1(d, \mathbf{b}, m_j) \not\to \infty$ as $j \to \infty$. We may then choose a subsequence $\{\lambda_1(d, \mathbf{b}, m_k)\}$ which is bounded. Let ϕ_k be the positive eigenfunction corresponding to $\lambda_1(d, \mathbf{b}, m_k)$ and normalized so that $\int_{\Omega} |\nabla\phi_k|^2 = 1$. Then the sequence $\{\phi_k\}$ is uniformly bounded in $W_0^{1,2}(\Omega)$, and since $W_0^{1,2}(\Omega)$ embeds compactly in $L^2(\Omega)$, we may choose a subsequence $\{\phi_l\}$ which converges in $L^2(\Omega)$ to some function ϕ. We have

$$\begin{aligned} d_0 = d_0 \int_{\Omega} |\nabla\phi_l|^2 &\leqq \int_{\Omega} d(x)|\nabla\phi_l|^2 + \phi_l \mathbf{b}(x)\cdot\nabla\phi_l \\ &= \lambda_1(d, \mathbf{b}, m_l) \int_{\Omega} m_l\phi_l^2 \\ &= \lambda_1(d, \mathbf{b}, m_l)\left(\int_{\Omega} m_l(\phi_l^2 - \phi^2) + \int_{\Omega} m_l\phi^2 \right). \end{aligned} \tag{4.4}$$

Letting $l \to \infty$, the first integral in the last formula goes to zero since $\|m_l\|_{\infty} \leqq m_0$ and $\phi_l \to \phi$ in $L^2(\Omega)$; the second goes to zero by (4.3). Since $\{\lambda_1(d, \mathbf{b}, m_l)\}$ is bounded, this implies $d_0 \leqq 0$, which is a contradiction, so we must have $\lambda_1(d, \mathbf{b}, m_j) \to \infty$.

To show that our hypotheses are necessary as well as sufficient, we use a device due to Holland [17] in a form similar to that used by Hess [13] in the context of periodic-parabolic problems. (In fact, this device provides a method for estimating the size of $\lambda_1(d, \mathbf{b}, m)$ from above for fixed d, $\mathbf{b}$, and m, but we shall not pursue this.) Consider the problem

$$-\nabla d(x)\nabla\psi + \mathbf{b}\cdot\nabla\psi - \lambda m\psi = \mu\psi \quad \text{in } \Omega, \qquad \psi = 0 \quad \text{on } \partial\Omega, \tag{4.5}$$

where $m \in L^\infty(\Omega)$ and $m > 0$ on a set of positive measure. The first eigenvalue $\mu_1(\lambda)$ admits an eigenfunction with $\psi > 0$ on Ω. Let $\theta = -\ln \psi$; then θ is defined on Ω and satisfies

$$d\Delta\theta - d|\nabla\theta|^2 - (\mathbf{b} - \nabla d)\cdot\nabla\theta - \lambda m = \mu_1(\lambda). \tag{4.6}$$

Suppose that $\phi \in C_0^\infty(\Omega)$ satisfies $\int_\Omega m\phi^2 > 0$ and $\int_\Omega \phi^2 = 1$. (We will return to the question of deciding if such functions exist later.) Multiplying (4.6) by ϕ^2 and using the divergence theorem, we have

$$\int_\Omega \nabla\cdot d\phi^2\nabla\theta - \int_\Omega [d\phi^2|\nabla\theta|^2 + 2d\phi\nabla\phi\cdot\nabla\theta + \phi^2\mathbf{b}\cdot\nabla\theta] - \lambda\int_\Omega m\phi^2 = \mu_1(\lambda). \tag{4.7}$$

Another application of the divergence theorem shows that the first term in (4.7) is zero. Adding the quantity

$$\int_\Omega d|\phi\nabla\theta + [\phi\mathbf{b} + 2d\nabla\phi]/2d|^2 \geqq 0$$

to the left side of (4.7) and rearranging terms, we have

$$\int_\Omega [|\phi\mathbf{b} + 2d\nabla\phi|^2/4d] - \lambda\int m\phi^2 \geqq \mu_1(\lambda). \tag{4.8}$$

If we let

$$\lambda = \Lambda(\phi, m) \equiv \int_\Omega [|\phi\mathbf{b} + 2d\nabla\phi|^2/4d] \Big/ \int_\Omega m\phi^2,$$

then (4.8) implies $\mu_1(\lambda) \leqq 0$; but since $\psi > 0$, it then follows from (4.5) and the positivity lemma of [15] that $\lambda_1(d, \mathbf{b}, m) \leqq \Lambda(\phi, m)$. Now suppose that $\limsup_{j\to\infty}\int_\Omega m_j\beta = \varepsilon_0 > 0$ for some $\beta \in L^1(\Omega)$ with $\beta \geqq 0$ almost everywhere. Then we can take a subsequence $\{m_k\}$ so that $\int_\Omega m_k\beta \geqq \varepsilon_0/2$, and we can approximate $\sqrt{\beta}$ as closely as we wish in $L^2(\Omega)$ with a function $\phi \in C_0^\infty(\Omega)$. If we choose ϕ so that $\int_\Omega |\beta - \phi^2| \leqq \varepsilon_0/4m_0$, we obtain $\int m_k\phi^2 \geqq \varepsilon_0/4 > 0$. It follows that for such ϕ the denominator of $\Lambda(\phi, m_k)$ is uniformly bounded away from zero, and since the numerator is independent of m, we have $\lambda_1(d, \mathbf{b}, m_k) \leqq \Lambda_0 < \infty$ for some Λ_0 and all m_k in the subsequence. Hence we cannot have $\lim_{j\to\infty}\lambda_1(d, \mathbf{b}, m_j) = \infty$ if (4.3) does not hold.

Remarks. If we consider a set of weights $\{m_j\}$ with $\int m_j \geqq m_1 > 0$ for all j, then (4.3) fails for $\beta \equiv 1$, so the corresponding set of principal eigenvalues $\lambda_1(d, \mathbf{b}, m_j)$ must be bounded, since otherwise we could find a sequence m_l with $\lambda_1(d, \mathbf{b}, m_l) \to \infty$ as $l \to \infty$. There are various conditions on d and $\mathbf{b}$ under which (4.1) must hold. For example, if we assume $d \geqq d_1 > 0$ and $|\mathbf{b}| \leqq b_0$, then we have for any $\varepsilon > 0$

$$\left|\int_\Omega (\mathbf{b}\cdot\nabla\phi)\phi\right| \leqq \int_\Omega (\varepsilon(\mathbf{b}\cdot\nabla\phi)^2/2) + (\phi^2/2\varepsilon) \leqq [(\varepsilon b_0^2/2) + (1/2\varepsilon\lambda_0)]\int_\Omega |\nabla\phi|^2$$

where $\lambda_0 \equiv \lambda_1(1, 0, 1)$. If we minimize with respect to ε, we obtain $|\int_\Omega (\mathbf{b}\cdot\nabla\phi)\phi| \leqq b_0/\sqrt{\lambda_0}$, so that (4.1) is satisfied if $d_1 > b_0/\sqrt{\lambda_0}$. If we assume $\mathbf{b} \in [C^1(\bar{\Omega})]^n$, we have

$$\int_\Omega (\mathbf{b}\cdot\nabla\phi)\phi = \frac{1}{2}\int_\Omega \mathbf{b}\cdot\nabla(\phi^2) = -\frac{1}{2}\int_\Omega (\nabla\cdot\mathbf{b})\phi^2.$$

If $\nabla\cdot\mathbf{b} \leqq b_1$ and $d \geqq d_1 > 0$, then (4.1) must hold provided $d_1 > b_1/2\lambda_0$.

Next we consider the problem of how $\mathbf{b}$ affects $\lambda_1(d, \mathbf{b}, m)$. For the case where $d \equiv 1$, $m \equiv 1$, and $\mathbf{b} = -\nabla B$, $B \in C^2(\bar{\Omega})$, Murray and Sperb [22] showed that if γ_1, γ_2 are such that $\gamma_1 \leqq \frac{1}{2}\Delta B + \frac{1}{4}|\nabla B|^2 \leqq \gamma_2$, then

$$\lambda_1(1, 0, 1) + \gamma_1 \leqq \lambda_1(1, \mathbf{b}, 1) \leqq \lambda_1(1, 0, 1) + \gamma_2. \tag{4.9}$$

They also showed that if $\Omega \subseteq \mathbb{R}^2$ is convex, the matrix $((\partial b_i/\partial x_j))$ is positive semidefinite in Ω, $\omega = \max_{\partial\Omega} |\mathbf{b}|$, $\alpha(\lambda, \omega) = \omega((\lambda + \omega^2)^{1/2} + \omega)$, and ρ is the radius of the largest disc contained in Ω, then $\lambda_1(\mathbf{1}, \mathbf{b}, 1)$ is greater than or equal to the first positive root λ of $\alpha(\lambda, \omega)/(\lambda + \alpha(\lambda, \omega)) = \cos(\rho\sqrt{\lambda})$.

Inequality (4.9) is obtained via a change of variables. If we have $\mathbf{b}(x)/d(x) = -\nabla B$ for some B, we can make the corresponding change of variables; letting $\psi = e^{B/2}\phi$ converts the problem $-\nabla \cdot d\nabla\phi + \mathbf{b} \cdot \nabla\phi = \lambda m\phi$ to

$$-\nabla \cdot d\nabla\psi + [(\nabla \cdot d\nabla B/2) + (d|\nabla B|^2/4)]\psi = \lambda m\psi, \tag{4.10}$$

while preserving the homogeneous Dirichlet boundary condition. If $\lambda = \lambda_1(d, \mathbf{b}, m)$ then since $\psi = e^{B/2}\phi > 0$, λ is also the first eigenvalue for (4.10), which has the variational characterization (see [21])

$$\lambda = \inf_{\substack{\psi \in W_0^{1,2}(\Omega) \\ \int m\psi^2 > 0}} \frac{\int_\Omega [d|\nabla\psi|^2 + \gamma\psi^2]}{\int_\Omega m\psi^2}, \tag{4.11}$$

where $\gamma = [(\nabla \cdot d\nabla B/2) + (d|\nabla B|^2/4)]$. In the special case $d \equiv 1$, $m \equiv 1$, (4.11) implies the bound (4.9). In general, if $\mathbf{b}$ satisfies $\mathbf{b} = -d\nabla B$, with B such that $\gamma \geqq 0$, then we may conclude $\lambda_1(d, \mathbf{b}, m) \geqq \lambda_1(d, 0, m)$, and if $\gamma \leqq 0$, $\lambda_1(d, \mathbf{b}, m) \leqq \lambda_1(d, 0, m)$. Since m is indefinite, it is not clear how to obtain bounds analogous to (4.9).

In the case where $\mathbf{b}$ is not a gradient, we can still obtain some information if $\mathbf{b} \in [C^1(\Omega)]^n$. If we multiply the equation $-\nabla \cdot d\nabla\phi + \mathbf{b} \cdot \nabla\phi = \lambda_1 m\phi$ by ϕ, integrate by parts, and use the boundary condition, we obtain

$$\int_\Omega [d|\nabla\phi|^2 - (\nabla \cdot \mathbf{b}/2)\phi^2] = \lambda_1 \int_\Omega m\phi^2. \tag{4.12}$$

If condition (4.1) is satisfied (which will clearly be the case if $\nabla \cdot \mathbf{b} \leqq 0$) then we have $\int_\Omega m\phi^2 > 0$, so that we may again use the variational formulation of [21] to see that if $\nabla \cdot \mathbf{b} \leqq 0$, then

$$\begin{aligned} \lambda_1(d, \mathbf{b}, m) &\geqq \inf_{\substack{\psi \in W_0^{1,2}(\Omega) \\ \int_\Omega m\psi^2 > 0}} \frac{\int_\Omega [d|\nabla\psi|^2 - (\nabla \cdot \mathbf{b}/2)\psi^2]}{\int_\Omega m\psi^2} \\ &\geqq \inf_{\substack{\psi \in W_0^{1,2}(\Omega) \\ \int_\Omega m\psi^2 > 0}} \frac{\int_\Omega d|\nabla\psi|^2}{\int_\Omega m\psi^2} = \lambda_1(d, 0, m). \end{aligned}$$

We have thus proved the following result.

THEOREM 4.2. *Suppose that either* $\mathbf{b} \in [C^1(\bar{\Omega})]^n$ *and* $\nabla \cdot \mathbf{b} \leqq 0$, *or that* $\mathbf{b} = -d\nabla B$ *for some* $B \in C^2(\bar{\Omega})$ *such that* $\gamma \equiv [(\nabla \cdot d\nabla B/2) + (d|\nabla B|^2/4)] \geqq 0$. *Then* $\lambda_1(d, \mathbf{b}, m) \geqq \lambda_1(d, 0, m)$.

Theorem 4.2 generalizes a result of [22] which implies that adding a constant drift term to the Laplacian always raises the principal eigenvalue. It can be shown via a perturbation argument that if $\mathbf{b} \in [C^1(\bar{\Omega})]^n$ with $\nabla \cdot \mathbf{b} > 0$, then for $\varepsilon > 0$ sufficiently small, $\lambda_1(1, \varepsilon\mathbf{b}, m) \leqq \lambda_1(1, 0, m)$. The general question of deciding how $\lambda_1(d_1, \mathbf{b}_1, m_1)$ and $\lambda_1(d_2, \mathbf{b}_2, m_2)$ are related is to our knowledge an open problem.

THEOREM 4.3. *Suppose that* $d \geqq d_1 > 0$ *and that* $\mathbf{b} \in [C^1(\bar{\Omega})]^n$ *with* $\nabla \cdot \mathbf{b} \leqq 0$. *Let* $M(x)$ *be any solution to*

$$\nabla \cdot d\nabla M + \mathbf{b} \cdot \nabla M + (\nabla \cdot \mathbf{b})M = m. \tag{4.13}$$

Suppose $M_1 \geqq \sup_\Omega M$ *and* $M_2 \geqq \operatorname{ess\,sup}_\Omega (-Mm)$. *If* $M_2 \leqq 0$, *then* $\lambda_1(d, \mathbf{b}, m) \geqq 1/2M_1$. *If* $M_2 > 0$, *then*

$$\lambda_1(d, \mathbf{b}, m) \geqq \frac{-2M_1 + [4M_1^2 + (8M_2/d_1\lambda_1(1, 0, 1))]^{1/2}}{(4M_2/d_1\lambda_1(1, 0, 1))}.$$

Remarks. Observe that no boundary condition is imposed on M in (4.13). Since $\nabla \cdot \mathbf{b} \leqq 0$, there will exist a solution for any reasonable boundary data.

Proof. Suppose ϕ satisfies

$$\begin{aligned} -\nabla \cdot d\nabla\phi + \mathbf{b} \cdot \nabla\phi &= \lambda_1(d, \mathbf{b}, m)m\phi \quad \text{in } \Omega, \\ \phi &= 0 \qquad\qquad\qquad\quad \text{on } \partial\Omega. \end{aligned} \tag{4.14}$$

Then we have via integration by parts

$$0 = \int_\Omega M(\nabla \cdot d\nabla\phi^2) - \phi^2(\nabla \cdot d\nabla M).$$

Since

$$\begin{aligned} \nabla \cdot d\nabla\phi^2 &= 2\phi\nabla \cdot d\nabla\phi + 2d|\nabla\phi|^2 \\ &= 2\phi\mathbf{b} \cdot \nabla\phi - 2\lambda_1 m\phi^2 + 2d|\nabla\phi|^2 \end{aligned}$$

and

$$\int_\Omega 2M\phi(\mathbf{b} \cdot \nabla\phi) = \int_\Omega M\mathbf{b} \cdot \nabla\phi^2 = -\int_\Omega \phi^2[\nabla M \cdot \mathbf{b} + M\nabla \cdot \mathbf{b}],$$

it follows that

$$0 = 2\int_\Omega Md|\nabla\phi|^2 - \int_\Omega \phi^2[\nabla \cdot d\nabla M + \mathbf{b} \cdot \nabla M + (\nabla \cdot \mathbf{b})M] - 2\lambda_1 \int_\Omega Mm\phi^2. \tag{4.15}$$

From (4.14) it follows that

$$\begin{aligned} \lambda_1 \int_\Omega m\phi^2 &= \int_\Omega [d|\nabla\phi|^2 + (\mathbf{b} \cdot \nabla\phi)\phi] \\ &= \int_\Omega [d|\nabla\phi|^2 - (\nabla \cdot \mathbf{b}/2)\phi^2] > 0, \end{aligned}$$

and we may assume that ϕ is normalized so that $\lambda_1 \int_\Omega m\phi^2 = 1$. By (4.13) we may replace the middle term in (4.15) by $-\int_\Omega m\phi^2$; if we then rearrange terms and multiply by λ_1, we obtain

$$1 = \lambda_1 \int_\Omega m\phi^2 = -2\lambda_1^2 \int_\Omega Mm\phi^2 + 2\lambda_1 \int_\Omega Md|\nabla\phi|^2. \tag{4.16}$$

Since

$$\nabla \cdot \mathbf{b} \leqq 0, \qquad 0 < \int_\Omega d|\nabla \phi|^2 \leqq \lambda_1 \int_\Omega m\phi^2 = 1;$$

also,

$$\int_\Omega \phi^2 \leqq \frac{1}{\lambda_1(1,0,1)d_1} \int_\Omega d|\nabla \phi|^2.$$

If we estimate the two integrals on the right of (4.16), we obtain

$$\int_\Omega d|\nabla \phi|^2 \leqq 1 \leqq 2M_2\lambda_1^2 \int_\Omega \phi^2 + 2M_1\lambda_1 \int_\Omega d|\nabla \phi|^2.$$

If $M_2 \leqq 0$, we have $1 \leqq 2M_1\lambda_1$. If $M_2 > 0$, then we have

$$1 \leqq (2M_2/\lambda_1(1,0,1)d_1)\lambda_1^2 + 2M_1\lambda_1.$$

The bounds on λ_1 follow immediately.

Example. Let $\Omega = (0, \pi)$, $d \equiv 1$, $\mathbf{b} \equiv b_0$, and $M = \sin nx$. Then (4.13) becomes $M'' + bM' = \sin nx$, which has a solution

$$M = -[b_0/n(n^2+b_0^2)] \cos nx - [1/(n^2+b_0^2)] \sin nx.$$

We may use $M_1 = 1/n(n^2+b_0^2)^{1/2}$ and $M_2 = (n+b_0)/n(n^2+b_0^2)$. Theorem 4.3 then yields

$$\lambda_1(1, b_0, \sin nx) \geqq \frac{1}{2}\left[\frac{-(n^2+b_0^2)^{1/2}}{n+b_0} + \left(\frac{n^2+b_0^2}{(n+b_0)^2} + \frac{2n(n^2+b_0^2)}{n+b_0}\right)^{1/2}\right], \tag{4.17}$$

which implies that $\lambda_1(1, b_0, \sin nx) \to \infty$ with order n as $n \to \infty$ and with order $\sqrt{b_0}$ as $b_0 \to \infty$.

4.1. Biological interpretation. While the results of this section are technical in appearance in the sense that they represent extensions of existing results, they are potentially the most relevant for studying the effects of environmental factors on population dynamics. In previous sections we established that the size of the eigenvalue $\lambda_1(d, \mathbf{b}, m)$ gives a criterion for persistence, namely, $\lambda_1(d, \mathbf{b}, m) < 1$, so that $\lambda_1(d, \mathbf{b}, m)$ serves as a reasonable measure of the overall suitability of an environment. In the next section we shall strengthen the case for using λ_1 as such a measure by deriving a population estimate in terms of λ_1. The results of this section give some information on how $\lambda_1(d, \mathbf{b}, m)$ is affected by the aspects of the environment described by the diffusion rate, drift, and local growth rate. Thus, they provide a means of using our models to infer the likely effects of certain environmental changes. The first two major results are qualitative, in that they describe the general behavior of λ_1 when the environment is perturbed in certain ways. The third is quantitative and allows a comparison of the relative impact of different effects, at least in simple cases. Our biological conclusions are somewhat tentative because of the enormous complexity of the problems they address, but they provide a starting point and direction for further work. We undertake a much more detailed analysis of some specific situations in [7].

Theorem 4.1 is a generalization of a result in [6]. Its main significance, we believe, is that it allows us to gain some insight into the effects of habitat fragmentation via reaction-diffusion models. The problem of understanding habitat fragmentation on populations is one of the most important topics in conservation biology. The theory

of island biogeography has been used to a considerable extent in the theoretical work on this problem, and it generally suggests that a few large regions of favorable habitat can be expected to sustain more species than a great many very small regions of the same total area. Theorem 4.1 allows us to consider the question at the species level rather than the community level, but leads to conclusions which are similar in spirit. Specifically, if we consider an environment in which the average habitat quality (as measured by the integral of the growth rate $m(x)$) is zero and vary the spatial distribution of favorable habitat so that it becomes more and more fragmented and more closely interspersed with unfavorable regions, then λ_1 will eventually tend to infinity so that our model predicts extinction.

As a simple example, if we consider a one-dimensional environment with fixed diffusion and drift coefficients and take $m_j(x) = \sin(jx)$, then $\lambda_1(d, b, \sin(jx)) \to \infty$ as $j \to \infty$. Whenever j is large enough that $\lambda_1(d, b, \sin(jx)) > 1$, the population cannot be expected to persist. It is important to keep in mind the asymptotic nature of this result; some of our work in [7] indicates that under certain conditions a few medium-sized favorable regions may provide a more suitable overall environment than a single large one. In some cases, we can obtain more precise quantitative information from Theorem 4.3. We have given an example immediately prior to this discussion. For more details on the connections between our work, island biogeography theory, and conservation biology, along with some references, see [6].

Theorem 4.2 gives some information on the effects of drift on the population. It is well known that (in the presence of a hostile exterior) increasing the diffusion rate tends to cause a more rapid loss of population across the boundary of the environment. Under certain conditions the effects of drift can produce the same results, and the theorem described some of those conditions. The case of constant growth and diffusion rates was treated by Murray and Sperb [22], and our results can be viewed as an extension of theirs to the case of variable diffusion. In realistic models we should expect $\nabla \cdot \mathbf{b}$ to change sign unless $\mathbf{b}$ is constant, since otherwise the drift term itself acts as a source or a sink. The condition $\mathbf{b} = -d\nabla B$ with B satisfying $(\nabla \cdot d\nabla B/2) + (d|\nabla B|^2/4) \geqq 0$ says roughly that the drift acts to augment the effects of diffusion. This condition can be checked via standard techniques from vector calculus. It was shown in [22] that for constant diffusion and growth rates, constant drift always makes the environment less suitable for the population under the assumption of a hostile exterior. Our results show that the same conclusion holds in the case of variable growth and diffusion rates. In both situations, the effect is due essentially to the drift pushing the population toward the hostile exterior region in one direction, while contributing no inward flux from the other since there will be no population in the hostile exterior region.

The qualitative results of Theorems 4.1 and 4.2 are augmented by the quantitative bounds on λ_1 given by Theorem 4.3. The example following the proof of that theorem shows how it can be used to draw conclusions about the persistence of a population from data on the diffusion, drift, and growth coefficients in a specific case. If the lower bound given in (4.17) is larger than 1, our model predicts extinction for the population. Other situations could be treated in a similar way. Of course, more complicated situations will require more effort in the analysis, but the estimate is based on the well-developed theory of linear differential equations. The specific bound (4.17) is already of some interest biologically since it gives an indication of the relative significance of drift and environmental heterogeneity. If we consider a one-dimensional environment, the estimate increases with the same order as the number of fragments of equal size into which the regions of favorable and unfavorable habitat are divided. It increases with the order of the square root of the coefficient describing the drift.

5. Population estimates. In the situations covered by Corollary 3.2 (density dependent diffusion in divergence form) and Corollary 3.3 (density independent diffusion not necessarily in divergence form), we are able to estimate the total size $\int_\Omega u\,dx$ of the positive steady states to (2.1) in a manner analogous to that in Theorem 4.1 of [6]. Since the results for these two cases are different from each other, we include them in this paper for the sake of completeness. We begin with the case of density independent diffusion.

THEOREM 5.1. *Suppose that u is the positive solution to*

$$\begin{aligned} -\nabla\cdot(d(x)\nabla u)+\mathbf{b}(x)\cdot\nabla u &= \lambda[m(x)u-u^2] \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{5.1}$$

where $\lambda>\lambda_1(d,\mathbf{b},m)$. Suppose that the differential operator satisfies the coercivity condition (4.1) *and that $\tilde\lambda>0$ is the principal eigenvalue for*

$$\begin{aligned} -d_0\Delta z &= \mu m z \quad \text{in } \Omega, \\ z &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Then $\tilde\lambda\leqq\lambda_1(d,\mathbf{b},m)$ and

$$\|u\|_1\leqq\left(1-\frac{\tilde\lambda}{\lambda}\right)\|m_+\|_3|\Omega|^{2/3}.$$

Proof. Suppose that $w>0$ on Ω and satisfies

$$\begin{aligned} -\nabla\cdot(d(x)\nabla w)+\mathbf{b}(x)\cdot\nabla w &= \lambda_1(d,\mathbf{b},m)m(x)w \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Then

$$\lambda_1(d,\mathbf{b},m)\int_\Omega m(x)w^2=\int_\Omega w(-\nabla\cdot(d(x)\nabla w)+\mathbf{b}(x)\cdot\nabla w)\geqq d_0\int_\Omega|\nabla w|^2.$$

Consequently, $\int_\Omega m(x)w^2>0$ and $d_0\int_\Omega|\nabla w|^2\geqq\tilde\lambda\int_\Omega m(x)w^2$ by the variational characterization of $\tilde\lambda$ [21]. Hence $\tilde\lambda\leqq\lambda_1(d,\mathbf{b},m)$. Multiplying (5.1) by u and integrating gives

$$\begin{aligned} \int_\Omega u^3 &= \int_\Omega m(x)u^2-\int_\Omega u\left[\frac{-\nabla\cdot(d(x)\nabla u)+\mathbf{b}(x)\cdot\nabla u}{\lambda}\right] \\ &\leqq \int_\Omega m(x)u^2-\frac{d_0}{\lambda}\int_\Omega|\nabla u|^2 \\ &\leqq \left(1-\frac{\tilde\lambda}{\lambda}\right)\int_\Omega m(x)u^2, \end{aligned}$$

since $\int_\Omega m(x)u^2>0$ by (4.1). Since $\int_\Omega mu^2\leqq\int_\Omega m_+u^2\leqq\|m_+\|_3\|u\|_3^2$ and $\|u\|_1\leqq\|u\|_3|\Omega|^{2/3}$, the result follows.

Two comments are in order at this point. The first is that the reader will recall that § 4 contains a discussion of conditions under which the coercivity condition (4.1) obtains. The second is that Theorem 5.1 does not provide an estimate of the rate at which $\|u\|_1$ approaches zero as $\lambda\to\lambda_1(d,\mathbf{b},m)$ which we know must be the case by the results of §§ 2 and 3. This limitation is due to the presence of the drift term. However, Theorem 5.1 does provide the useful global estimate $\|u\|_1\leqq\|m+\|_3|\Omega|^{2/3}$. In the density dependent case in divergence form, we can obtain the same global estimate

as well as estimate the rate at which $\|u\|_1$ tends to zero as $\lambda \to \lambda_1(d(x,0),0,m)$ as the next result shows.

THEOREM 5.2. *Suppose that u is the positive solution to*

$$-\nabla \cdot (d(x,u)\nabla u) = \lambda[m(x)u - u^2] \quad \text{in } \Omega,$$
$$u = 0 \quad \text{on } \partial\Omega,$$

where $\lambda > \lambda_1(d(x,0),0,m)$ and we assume that $\partial d/\partial u \geqq 0$. Then

$$\|u\|_1 \leqq \left(1 - \frac{\lambda_1(d(x,0),0,m)}{\lambda}\right)\|m_+\|_3 |\Omega|^{2/3}.$$

Proof.

$$0 < \int_\Omega \frac{d(x,u)|\nabla u|^2}{\lambda} + \int_\Omega u^3 = \int_\Omega mu^2.$$

So $\int_\Omega mu^2 > 0$, and hence

$$\begin{aligned}\int_\Omega u^3 &= \int_\Omega mu^2 - \int_\Omega \frac{d(x,u)}{\lambda}|\nabla u|^2 \\ &\leqq \int_\Omega mu^2 - \int_\Omega \frac{d(x,0)}{\lambda}|\nabla u|^2 \\ &\leqq \left(1 - \frac{\lambda_1(d(x,0),0,m)}{\lambda}\right)\int_\Omega mu^2\end{aligned}$$

by the positivity of $\int_\Omega mu^2$ and the variational characterization of $\lambda_1(d(x,0),0,m)$. The remainder of the proof follows as in the proof of Theorem 5.1.

Finally, we note that in both these situations, we can obtain estimates on the rate of decay of solutions to (2.1) which are analogous to the result of Theorem 4.7 in [6]. The modifications needed to obtain these results from Theorem 4.7 of [6] are similar to those needed to obtain Theorems 5.1 and 5.2 above from Theorem 4.1 of [6]. Consequently, we omit them from this paper.

5.1. Biological interpretation. The immediate biological interpretation of the results of this section is clear. They yield bounds on the total population which our models predict a given environment can sustain. Theorem 5.1 is less sharp than Theorem 5.2, but for regions with simple geometry $\tilde{\lambda}$ may be easier to compute than λ_1. A deeper interpretation of Theorem 5.2 is that $\lambda_1(d, \mathbf{b}, m)$ is, in fact, an appropriate measure of environmental suitability, for in the original form of our models with $\lambda = 1$, Theorem 5.2 gives a bound on the population in which $1 - \lambda_1(d, \mathbf{b}, m)$ appears as a factor. Thus, if we vary d, $\mathbf{b}$, and m so that $\lambda_1(d, \mathbf{b}, m)$ approaches 1, the bound on the population goes to zero. (We have considered only the case where the carrying capacity is taken to be 1, but that can always be achieved by a rescaling if the carrying capacity is a constant.)

6. Conclusions. Reaction-diffusion models have been widely used to model population dynamics (see [4]–[7], [9], [10], [19], [20], [24], [27], [28]). We consider a class of such models which incorporate environmental variation, drift, and density dependent diffusion. We establish that in many cases the eigenvalue $\lambda_1(d, \mathbf{b}, m)$ for an associated linear problem is a reasonable measure of environmental suitability by showing that the condition $\lambda_1(d, \mathbf{b}, m) < 1$ implies persistence and obtaining upper bounds for the population in which $1 - \lambda(d, \mathbf{b}, m)$ appears as a factor. The significance of this

observation is that $\lambda_1(d, \mathbf{b}, m)$ is a quantity which depends directly on the diffusion, drift, and growth rates for the population and which can be computed by using well-known (although sometimes fairly sophisticated) mathematical techniques. In some cases we can calculate $\lambda_1(d, \mathbf{b}, m)$ fairly easily, but what is perhaps more important is that we can make qualitative inferences about the effects of changing various aspects of the environment on its overall suitability for a population as measured by λ_1. Specifically, our models predict that a high degree of fragmentation of favorable habitat increases λ_1 and thus decreases environmental suitability, and that the presence of drift may either increase or decrease environmental suitability. (It turns out that under the assumption of a hostile exterior region that constant drift always decreases the overall environmental suitability, but a spatially varying drift term may actually increase it.) Similar conclusions have been drawn in other ways, but largely on the basis of either heuristic arguments or different modeling viewpoints. A conclusion that does not rely on properties of λ_1 is that the presence of density dependent diffusion can lead to multiple equilibria. A similar effect has been observed for models with constant diffusion but depensatory growth rate, but the observation that multiple equilibria can occur with a logistic growth term and density dependent diffusion is apparently new.

REFERENCES

[1] H. AMANN, *Existence and regularity for semilinear parabolic evolution equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 11 (1984), pp. 576–593.

[2] ———, *Quasilinear evolution equations and parabolic systems*, Trans. Amer. Math. Soc., 293 (1986), pp. 191–227.

[3] J. BELL AND C. COSNER, *Threshold behavior and propagation for nonlinear differential-difference systems motivated by modelling myelinated nerves*, Quart. Appl. Math., 17 (1984), pp. 1–14.

[4] ———, *Wavelike solutions to reaction-diffusion equations on a cylinder: dependence on cylinder width*, SIAM J. Appl. Math., 47 (1987), pp. 534–543.

[5] R. S. CANTRELL AND C. COSNER, *On the uniqueness and stability of positive solutions in the Lotka–Volterra competition model with diffusion*, Houston J. Math., 15 (1989), pp. 15–34.

[6] ———, *Diffusive logistic equations with indefinite weights: population models in disrupted environments*, Proc. Roy. Soc. Edinburgh Sect. A, 112 (1989), pp. 293–318.

[7] ———, *The effects of spatial heterogeneity in population dynamics*, J. Math. Biology, to appear.

[8] R. S. CANTRELL AND K. SCHMITT, *On the eigenvalue problem for coupled elliptic systems*, SIAM J. Math. Anal., 17 (1986), pp. 850–862.

[9] C. COSNER, *Eigenvalue problems with indefinite weights and reaction-diffusion models in population dynamics*, preprint.

[10] P. C. FIFE, *Mathematical Aspects of Reacting and Diffusing Systems*, Lecture Notes in Biomath. 28, Springer-Verlag, Berlin, 1979.

[11] D. GILBARG AND N. W. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[12] J. P. GOSSEZ AND E. LAMI DOZO, *On the principal eigenvalue of a second order elliptic problem*, Arch. Rational Mech. Anal., 89 (1985), pp. 169–175.

[13] P. HESS, *On positive solutions of semilinear periodic-parabolic problems*, in Infinite-Dimensional Systems, Lecture Notes in Math. 1076, Springer-Verlag, Berlin, New York, 1984, pp. 101–114.

[14] ———, *On the relative completeness of the generalized eigenvectors of elliptic eigenvalue problems with indefinite weight functions*, Math. Ann., 270 (1985), pp. 467–475.

[15] P. HESS AND T. KATO, *On some linear and nonlinear eigenvalue problems with an indefinite weight function*, Comm. Partial Differential Equations, 5 (1980), pp. 999–1030.

[16] M. HIRSCH, *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math., 383 (1988), pp. 1–53.

[17] C. J. HOLLAND, *A minimum principle for the principal eigenvalue for second-order linear elliptic equations with natural boundary conditions*, Comm. Pure Appl. Math., 31 (1978), pp. 509–519.

[18] O. A. LADYZHENSKAYA AND N. N. URAL'TSEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[19] S. LEVIN, *Population models and community structure in heterogeneous environments*, in Mathematical Ecology, T. G. Hallam and S. Levin, eds., Biomathematics 17, Springer-Verlag, Berlin, 1986.

[20] D. LUDWIG, D. G. ARONSON, AND H. F. WEINBERGER, *Spatial patterning of the spruce budworm*, J. Math. Biol., 8 (1979), pp. 217–258.

[21] A. MANES AND A. M. MICHELETTI, *Un'estensione della teoria variazionale classica degli autovalori per operatori ellitici del secondo ordine*, Boll. Un. Mat. Ital. (6), 7 (1973), pp. 285–301.

[22] J. D. MURRAY AND R. P. SPERB, *Minimum domains for spatial patterns in a class of reaction-diffusion equations*, J. Math. Biol., 18 (1983), pp. 169–184.

[23] R. NUSSBAUM, *Periodic solutions of some nonlinear integral equations*, in Proc. Internat. Symposium on Dynamical Systems, A. R. Bednarek and L. Cesari, eds., Gainesville, FL, 1976.

[24] A. OKUBO, *Diffusion and Ecological Problems: Mathematical Models*, Biomathematics 10, Springer-Verlag, Berlin, 1980.

[25] P. H. RABINOWITZ, *Some aspects of nonlinear eigenvalue problems*, Rocky Mountain J. Math., 3 (1973), pp. 161–202.

[26] H. SENO, *Effect of a singular patch on population persistence in a multi-patch system*, Ecological Modelling, 43 (1988), pp. 271–286.

[27] J. G. SKELLAM, *Random dispersal in theoretical populations*, Biometrika, 38 (1951), pp. 196–218.

[28] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, 1983.

[29] H. F. WEINBERGER, *Variational Methods for Eigenvalue Approximation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1065–1080, July 1991

013

GLOBAL BEHAVIOR OF AN AGE-STRUCTURED EPIDEMIC MODEL*

STAVROS N. BUSENBERG†, MIMMO IANNELLI‡, AND HORST R. THIEME§

Abstract. The global behavior of the general $s \to i \to s$ age-structured epidemic model in a population of constant size is obtained. It is shown that there is a sharp threshold which determines the existence and global stability of an endemic state; hence, periodic solutions are ruled out. The threshold is identified as the spectral radius of a positive linear operator. The analysis employs the theory of semigroups and positive operator methods, and is based on the formulation of the problem as an abstract differential equation in a Banach space.

Key words. epidemic model, age structure, semigroup, monotone operator, global behavior

AMS(MOS) subject classifications. 92A15, 35L60, 47D05, 47H07

1. Introduction.

Periodic oscillations in the incidence rates of infectives occur for a number of diseases, and there are several dynamic mechanisms that could be the cause of this oscillatory behavior. For example, there is a pronounced biennial oscillation in the occurence of measles that is reported in the data of several large cities [7], [10], [13], and these have been variously attributed to the incubation delays [10], to the age-structured seasonal interaction rates that are due to school attendance [13], and to more complicated nonlinear transmission dynamics [7]. The simplest epidemiological interactions occur for diseases which do not impart immunity and which can be described by models that include only two epidemiological classes composed of the susceptible and infective parts of the population, that is, they are of the $s \to i \to s$ type. Since it has been suggested that the age dependence of the transmission rate of a disease may cause oscillations that do not occur in the corresponding situation without age structure, a basic problem in epidemic modeling is to rigorously establish either the possibility or the impossibility of the occurence of such oscillations.

The purpose of this and of a subsequent paper, where we will include vertical transmission, is to show that, for the general $s \to i \to s$ age-structured model, periodic oscillations do not occur. In fact, we give the complete global dynamic behavior for such a model when the total population is at its equilibrium distribution. To our knowledge, this is the first general age-structured epidemic model for which the global dynamic behavior is resolved. Global behavior results for this model have recently been obtained by Busenberg, Cooke, and Iannelli [2] when the force of infection term assumes certain particular age-dependent forms. These forms represent the cases where the disease transmission interactions are limited to members of the same age group (the *intracohort* case), and the other extreme possibility where the age of the infectives does not affect their contact rates with other individuals (the *intercohort* case.) In fact, for the intercohort case, periodic oscillations are ruled out in [2], only

*Received by the editors November 20, 1989; accepted for publication (in revised form) June 11, 1990.

†Department of Mathematics, Harvey Mudd College, Claremont, California 91711. The work of this author was partially supported by National Science Foundation grant DMS–8902712.

‡Dipartimento di Matematica, Università di Trento, 38050 Povo (Trento), Italy.

§Department of Mathematics, Arizona State University, Tempe, Arizona 85287. The work of this author was partially supported by a Heisenberg scholarship from "Deutsche Forschungsgemeinschaft."

with additional assumptions on the form of the age-dependent probability that a contact of a susceptible with an infective will lead to the transmission of the disease.

There is a recent surge of interest in analyzing the effects of age-structure on the dynamics of epidemics [1]–[6], [8], [11], [12] because of the recognition that the transmission dynamics of certain diseases could not be correctly described by the traditional epidemic models with no age dependence. Much of the work in this area has been limited to the derivation of threshold conditions for the existence of endemic steady states and, in some cases, to the the study of the local stability of these endemic states. Global stability analyses of such models have been obtained only in special cases [2], [4]. Even though the $s \to i \to s$ model exhibits one of the simplest types of disease transmision interactions, its complete analysis which we provide here settles a basic open question in the modeling of epidemics.

The model that we consider divides the affected population into two epidemiologically distinct classes, composed of the susceptible and the infective individuals. The age-specific densities of the susceptibles at time t and age a is denoted by $s(a,t)$, and that of the infectives by $i(a,t)$. Thus

$$\int_{a_1}^{a_2} s(a,t)\,da \quad \text{and} \quad \int_{a_1}^{a_2} i(a,t)\,da,$$

are the respective numbers of susceptible and infective individuals at time t whose ages fall between a_1 and a_2. Let $K(a,a')$ denote the rate at which an infective individual of age a' comes into a disease transmitting contact with a susceptible individual of age a also let $K_0(a)$ be the infection rate for pure intracohort interaction. Then the rate at which susceptible individuals of age a are moved over into the infective class is given by

$$s(a,t)\lambda[a|u(\cdot,t)] = s(a,t)\left[K_0(a)i(a,t) + \int_0^\infty K(a,a')i(a',t)\,da'\right].$$

The decomposition of the trasmission rate into the two parts K_0 and K allows us to include a pure intracohort term without using kernels which are measures.

Denoting the age-specific mortality, birth rate, and cure rate by $\mu(a)$, $\beta(a)$, and $\gamma(a)$, respectively, and using the notation $\lambda(a,t)$ for $\lambda[a|u(\cdot,t)]$, we obtain the following system of equations that describe the dynamics of this model:

$$\begin{aligned} &\frac{\partial s(a,t)}{\partial t} + \frac{\partial s(a,t)}{\partial a} + \mu(a)s(a,t) = -s(a,t)\lambda(a,t) + \gamma(a)i(a,t), \\ &\frac{\partial i(a,t)}{\partial t} + \frac{\partial i(a,t)}{\partial a} + \mu(a)i(a,t) = s(a,t)\lambda(a,t) - \gamma(a,t)i(a,t) \end{aligned} \tag{1.1}$$

with initial and boundary conditions

$$\begin{aligned} &s(0,t) = \int_0^\infty \beta(a)s(a,t)\,da, \qquad i(0,t) = 0, \\ &s(a,0) = s_0(a), \qquad i(a,0) = i_0(a). \end{aligned} \tag{1.2}$$

We note that in our assumption the disease does not affect the natural mortality and fertility rates.

We shall be dealing with these equations under the hypothesis that the total population, $p(a,t) = s(a,t) + i(a,t)$, has reached a steady-state distribution $p_\infty(a)$. From the classical theory of the linear McKendrick–Von Foerster equation which $p(a,t)$ satisfies, this occurs when the basic reproductive number R of the total population is equal to 1. That is,

$$R = \int_0^\infty \beta(a) \exp\left(-\int_0^a \mu(a')\,da'\right) da = 1. \tag{1.3}$$

It is worth noting that, if the parameters β, μ, γ, and K in the model did not depend on age, the corresponding equations for the total populations $S(t)$ and $I(t)$, can be obtained by integrating (1.1) from $a = 0$ to ∞, and using the conditions (1.2), and $s(\infty,t) = i(\infty,t) = 0$. These last conditions simply mean that no individual in the population can survive to infinite age. The resulting equations are

$$\begin{aligned} \frac{dS(t)}{dt} &= (\beta - \mu)S(t) + \gamma I(t) - KS(t)I(t), \\ \frac{dI(t)}{dt} &= -(\mu + \gamma)I(t) + KS(t)I(t). \end{aligned} \tag{1.4}$$

This is the standard $S \to I \to S$ epidemic model with vital dynamics. The condition (1.3) implies that $\beta = \mu$. Hence, we are in the case where the births and deaths are in balance, and consequently, the total population $S(t) + I(t) = P$ is constant. Writing $S(t) = P - I(t)$, and substituting in the second equation in (1.4), we obtain a first-order equation in the single variable $I(t)$ which can be explicitly integrated by elementary methods. The system (1.4) cannot exhibit time-periodic behavior, and the main result of this paper shows that the introduction of age dependence, with a general form $K(a,a')$ of the contact rate, does not produce a model which has periodic solutions, regardless of the choice of the age-dependent parameters entering in (1.1)–(1.2).

In the next section we transform the problem (1.1)–(1, 2) to one that involves only a single dependent variable $u(a,t) = i(a,t)/p_\infty(a)$. We then proceed to show that the problem can be viewed as a dynamical system for which we prove the existence of an appropriate unique solution, under mild restrictions on the parameters appearing in the equations. In the other sections we exploit the monotonicity and convexity properties of our formulation of the problem to show that all solutions stabilize to steady states, regardless of their initial data. The conditions we impose on the contact rate $K(a,a')$ include all epidemiologically significant cases. Thus, for the general $s \to i \to s$ age structured model, the global behavior of solutions is settled, and oscillatory solutions do not occur. In the final section we discuss the epidemiological implications of our results and of the hypotheses that we need to place on the pertinent parameters in the model.

2. Setting the problem. Let us first state the assumptions on the demographic parameters $\beta(a)$ and $\mu(a)$, namely, we assume that there is a maximum age $a_\dagger$ for the population so that we can restrict ourselves to the age interval $[0, a_\dagger]$, and

(2.1) $\beta(a)$ is nonnegative and belongs to $L^\infty(0, a_\dagger)$,

(2.2) $\mu(a)$ is nonnegative and measurable,

$$R = \int_0^{a_\dagger} \beta(a) \exp\left(-\int_0^a \mu(a')\,da'\right) da = 1. \tag{2.3}$$

This last parameter R is the basic reproductive number of the population and represents the mean number of newborns from an individual, during his whole lifespan. Thus, by (2.3) we are assuming that the demographic parameters are such that a steady asymptotic state exists:

$$p_\infty(a) = b_0 \exp\left(-\int_0^a \mu(\sigma)\,d\sigma\right), \tag{2.4}$$

where

$$b_0 = \int_0^{a_\dagger} \beta(a) p_\infty(a)\,da. \tag{2.5}$$

Now we assume that the total population $p(a,t) = s(a,t) + i(a,t)$ has reached the steady-state distribution (2.4). Thus we have

$$s(a,t) = p_\infty(a) - i(a,t) \tag{2.6}$$

and the system can be reduced to a single equation for $i(a,t)$. In fact, fitting (2.6) into the second equation in (1.1), we have

$$\begin{aligned}
&\frac{\partial i(a,t)}{\partial t} + \frac{\partial i(a,t)}{\partial a} + \mu(a) i(a,t) = \lambda[a \mid i(.,t)][p_\infty(a) - i(a,t)] - \gamma(a) i(a,t),\\
&\lambda[a \mid i(.,t)] = K_0(a) i(a,t) + \int_0^{a_\dagger} K(a,a') i(a',t)\,da',\\
&i(0,t) = 0, \qquad i(a,0) = i_0(a).
\end{aligned}$$

Furthermore, we perform the following change $u(a,t) = i(a,t)/p_\infty(a)$ so that

$$\begin{aligned}
&\frac{\partial u(a,t)}{\partial t} + \frac{\partial u(a,t)}{\partial a} = \lambda[a \mid u(\cdot,t)][1 - u(a,t)] - \gamma(a) u(a,t),\\
&\lambda[a \mid u(\cdot,t)] = K_0(a) p_\infty(a) u(a,t) + \int_0^{a_\dagger} K(a,a') p_\infty(a') u(a',t)\,da',\\
&u(0,t) = 0, \qquad u(a,0) = \frac{i_0(a)}{p_\infty(a)} = u_0(a),
\end{aligned} \tag{2.7}$$

and this is the problem that we are going to consider in the sequel.

Let us now state the assumptions on $\gamma(a)$ and on the contact rates $K_0(a)$ and $K(a,a')$. We assume that

$$\gamma(a), K_0(a) \quad \text{are nonnegative and belong to } L^\infty(0,a_\dagger); \tag{2.8}$$

moreover, $K(a,a')$ is measurable, and there exist a positive constant ϵ and nonnegative functions $K_1(a), K_2(a)$ such that

$$\epsilon\, K_1(a) K_2(a') \le K(a,a') p_\infty(a') \le K_1(a) K_2(a'), \tag{2.9}$$

$$K_1 \in L^1(0,a_\dagger)\,, \qquad K_2 \in L^\infty(0,a_\dagger). \tag{2.10}$$

There are $0 \le a_1, a_2, b_1, b_2 \le a_\dagger$ such that

$$a_1 < b_1, \quad a_2 < b_2, \quad a_1 < b_2, \tag{2.11}$$

$$K_1(a) > 0 \quad \text{if } a_1 < a < b_1, \tag{2.12}$$

$$K_2(a) > 0 \quad \text{if } a_2 < a < b_2. \tag{2.13}$$

The implications of these hypotheses from the viewpoint of the biological model will be discussed in §6.

3. Resetting the problem. We now identify the problem as an abstract semilinear equation. To this purpose we consider the Banach space $E = L^1(0, a_\dagger)$ and define

$$A = \begin{cases} D_A = \{f \in E \mid f \text{ is absolutely continuous; } f(0) = 0\}, \\ (Af)(a) = -f'(a), \end{cases} \tag{3.1}$$

$$F(f)(a) = \lambda[a \mid f(.)][1 - f(a)] - \gamma(a)f(a), \tag{3.2}$$

where

$$\lambda[a \mid f(.)] = K_0(a)p_\infty(a)f(a) + \int_0^{a_\dagger} K(a, a')p_\infty(a')f(a')\,da'.$$

Thus (2.2) can be expressed as the following Cauchy problem in E:

$$\frac{d}{dt}u(t) = Au(t) + F(u(t)), \qquad u(0) = u_0, \tag{3.3}$$

and, due to the meaning of $u(a, t)$, we look for a solution in the closed convex set:

$$C = \{f \in L^1(0, a_\dagger);\ 0 \le f(a) \le 1 \text{ a.e.}\}. \tag{3.4}$$

Concerning the operator A we have that it is the generator of the C_0-semigroup on E:

$$(e^{tA}u_0)(a) = \begin{cases} u_0(a - t) & \text{for} \quad t < a, \\ 0 & \text{for} \quad t > a \end{cases} \tag{3.5}$$

with the following properties:

$$e^{tA}(C) \subset C, \tag{3.6}$$

$$\text{if } u_0 \le v_0, \quad \text{then } e^{tA}u_0 \le e^{tA}v_0, \tag{3.7}$$

where $\le$ denotes the usual ordering in L^1.

We also note that if $\alpha > 0$ the resolvent $(I - \alpha A)^{-1}$ is given by the formula:

$$((I - \alpha A)^{-1}f)(a) = \frac{1}{\alpha}\int_0^a e^{-\frac{1}{\alpha}(a-s)}f(s)\,ds \tag{3.8}$$

and satisfies:

$$(I - \alpha A)^{-1}(C) \subset C \tag{3.9}$$

$$\text{if } u \le v, \quad \text{then } (I - \alpha A)^{-1}u \le (I - \alpha A)^{-1}v. \tag{3.10}$$

Concerning F we have Proposition 3.1.

PROPOSITION 3.1. *The function $F : C \to E$ is Lipschitz continuous and there exists a constant $\alpha \in (0,1)$ such that*

$$\text{if } u \le v, \quad \text{then } u + \alpha F(u) \le v + \alpha F(v), \tag{3.11}$$

$$(I + \alpha F)(C) \subset C. \tag{3.12}$$

Proof. Lipschitz continuity is a simple matter. Moreover, let γ^+ and λ^+ be such that $\gamma(a) < \gamma^+$ and $\lambda[a \mid f(.)] < \lambda^+$ for any $f \in C$. Take

$$\alpha < \frac{1}{\gamma^+ + \lambda^+}, \tag{3.13}$$

then, if $u \le v$,

$$\begin{aligned}
(u + \alpha F(u))(a) &= u(a) + \alpha\lambda[a \mid u(\cdot)][1 - u(a)] - \alpha\gamma(a)u(a) \\
&\le u(a) + \alpha\lambda[a \mid v(\cdot)][1 - u(a)] - \alpha\gamma(a)u(a) \\
&= u(a)[1 - \alpha\lambda[a \mid v(\cdot)] - \alpha\gamma(a)] + \alpha\lambda[a \mid v(\cdot)] \\
&\le v(a)[1 - \alpha\lambda[a \mid v(\cdot)] - \alpha\gamma(a)] + \alpha\lambda[a \mid v(\cdot)] \\
&= (v + \alpha F(v))(a)
\end{aligned}$$

with the last estimate being possible because the square bracket is nonnegative, thanks to the choice of α.

Finally (3.12) is a consequence of (3.11), in fact if $0 \le u(a) \le 1$ then, by (3.2),

$$0 \le \alpha F(0) \le u + \alpha F(u) \le 1 + \alpha F(1) = 1 - \alpha\gamma(a) \le 1. \tag{3.14}$$

□

Now we are ready to prove existence of a mild solution to problem (3.3), namely, we look for a solution of the integral equation (see A. Pazy [12]):

$$u(t) = e^{tA}u_0 + \int_0^t e^{(t-s)A}F(u(s))\,ds. \tag{3.15}$$

We then have Theorem 3.2.

THEOREM 3.2. *Let $u_0 \in C$; then problem* (3.3) *has a unique mild solution in C. This defines a flow $S(t)u_0$ which has the following properties:*

$$S(t)(C) \subset C, \tag{3.16}$$

$$\text{if } u_0 \le v_0, \quad \text{then } S(t)u_0 \le S(t)v_0, \tag{3.17}$$

$$\text{if } 0 \le \xi \le 1, \quad \text{then } \xi S(t)u_0 \le S(t)(\xi u_0). \tag{3.18}$$

Proof. It is easy to prove that problem (3.15) is equivalent to the following one:

$$u(t) = e^{-\frac{1}{\alpha}t}e^{tA}u_0 + \frac{1}{\alpha}\int_0^t e^{-\frac{1}{\alpha}(t-s)}e^{(t-s)A}[u(s) + \alpha F(u(s))]\,ds, \tag{3.19}$$

where α is chosen as in (3.13). We start the standard iterative procedure:

$$
\begin{aligned}
&u^0(t) = u_0,\\
&u^{n+1}(t) = e^{-\frac{1}{\alpha}t}e^{tA}u_0 + \frac{1}{\alpha}\int_0^t e^{-\frac{1}{\alpha}(t-s)}e^{(t-s)A}[u^n(s) + \alpha F(u^n(s))]\,ds.
\end{aligned}
\tag{3.20}
$$

Now, thanks to (3.6) and (3.12), $u^{n+1}(t) \in C$; in fact, the right-hand side of (3.20) is a convex linear combination of $e^{tA}u_0$ and $[u^n(s) + \alpha F(u^n(s))]$ (note that $e^{-\frac{1}{\alpha}t} + \frac{1}{\alpha}\int_0^t e^{-\frac{1}{\alpha}(t-s)}ds = 1$); thus, because of the Lipschitz continuity of F, the sequence $u^n(t)$ converges uniformly to $S(t)u_0 \in C$.

Furthermore, if we start the procedure with $v_0 \geq u_0$, thanks to (3.7) and (3.11), the iterates v^n and u^n satisfy $v^n \geq u^n$ so that, in the limit, we have (3.17).

Finally, let u_ξ^n be the iterates (3.20) with u_0 replaced by ξu_0. Since $0 \leq \xi \leq 1$, we have

$$\xi u + \alpha F(\xi u) \geq \xi(u + \alpha F(u))$$

so that

$$u_\xi^n \geq \xi u^n$$

and, going to the limit, we obtain (3.18). □

4. Existence and uniqueness of an endemic equilibrium. We are now concerned with existence and uniqueness of a nontrivial equilibrium point to problem (3.3); namely, we look for a solution to the equation

$$Au_\infty + F(u_\infty) = 0. \tag{4.1}$$

We use the symbol u_∞ because we shall show below that any such fixed point is a limit of other solutions of (3.3). This is equivalent to looking for a fixed point of the flow:

$$S(t)u_\infty = u_\infty. \tag{4.2}$$

In fact, if u_∞ satisfies (4.2), from equation (3.15) we get

$$u_\infty = e^{tA}u_\infty + \int_0^t e^{(t-s)A}\,ds\; F(u_\infty) \quad \forall t \geq 0, \tag{4.3}$$

that is,

$$e^{tA}u_\infty = u_\infty - \int_0^t e^{sA}\,ds\; F(u_\infty), \tag{4.4}$$

which implies that $e^{tA}u_\infty$ is differentiable, and consequently, that $u_\infty \in D_A$. Differentiating (4.4) at $t = 0$, we obtain (4.1). This argument can be followed backward proving (4.2), starting from (4.1).

We now state a necessary condition which should be satisfied by any nontrivial solution of (4.1).

PROPOSITION 4.1. *Let u_∞ be a nontrivial solution of* (4.1)*; then*

$$\int_0^{a_\dagger} K_2(a)u_\infty(a)da > 0. \tag{4.5}$$

Proof. If (4.5) is not fulfilled, then

$$\int_0^{a_\dagger} K(a,a')p_\infty(a')u_\infty(a')\,da' = 0 \quad \forall a \in [0,a_\dagger] \tag{4.6}$$

and u_∞ satisfies

$$u'_\infty(a) = -\gamma(a)u_\infty(a) + K_0(a)p_\infty(a)u_\infty(a)[1-u_\infty(a)],$$

thus $u_\infty(a)$ is either identically zero or positive on $[0,a_\dagger]$, and this latter case is not compatible with (4.6) and (2.9)–(2.13). □

We next state a preliminary estimate that must be satisfied by any possible nontrivial fixed point. Our estimate involves a comparison function, namely,

$$m(a) = \int_0^a K_1(a')\,da'. \tag{4.7}$$

PROPOSITION 4.2. *Let u_∞ be a nontrivial solution to* (4.1). *Then*

$$\epsilon_1 m(a) \le u_\infty(a) \le \epsilon_2 m(a), \tag{4.8}$$

where ϵ_1, ϵ_2 are positive constants depending on u_∞.

Proof. As u_∞ is a nontrivial solution, (4.5) is satisfied so that

$$c_1K_1(a) \le \lambda[a \mid u_\infty(.)] \le k_0^+u_\infty(a) + c_2K_1(a),$$

where c_1, c_2 are positive and $k_0^+ \ge b_0K_0(a)$. Hence u_∞ satisfies the inequality

$$c_1K_1(a) - (\lambda^+ + \gamma^+)u_\infty(a) \le u'_\infty(a) \le k_0^+u_\infty(a) + c_2K_1(a)$$

and, since $u_\infty(0) = 0$, we can find positive ϵ_1, ϵ_2 such that (4.8) is satisfied. □

To deal with problem (4.1) it is convenient to transform this problem into the following one:

$$u_\infty = (I - \alpha A)^{-1}(I + \alpha F)u_\infty,$$

where α is as in (3.13). Thus we are led to investigate fixed points of the mapping

$$T = (I - \alpha A)^{-1}(I + \alpha F),$$

which by (3.9)–(3.12) has the following properties:

$$T(C) \subset C, \tag{4.9}$$

$$\text{if } u \le v, \quad \text{then } Tu \le Tv. \tag{4.10}$$

We first show the uniqueness of a nontrivial equilibrium by following the line of reasoning in Krasnosel'skii [9]. In order to do this, we need some more estimates.

LEMMA 4.3. *Let $u_\infty \in C$ be a nontrivial fixed point of T. Let ξ be a constant and $v \in C$, such that $0 < \xi < 1$, $v \ge 0$, and*

$$\xi u_\infty + v \in C, \qquad \int_0^{a_\dagger} K_2(a)v(a)\,da > 0; \tag{4.11}$$

then there exists a constant $\epsilon_3 > 0$ *such that*

$$T(\xi u_\infty + v) \geq \xi u_\infty + \epsilon_3(1-\xi)m. \tag{4.12}$$

Proof. Note that (see (3.13))

$$\begin{aligned}
&(\xi u_\infty + v + \alpha F(\xi u_\infty + v))(a)\\
&\quad = \xi u_\infty(a) + v(a) + \alpha[1 - \xi u_\infty(a) - v(a)]\lambda[a \mid \xi u_\infty(\cdot) + v(\cdot)]\\
&\qquad - \alpha\gamma(a)(\xi u_\infty(a) + v(a))\\
&\quad = \xi u_\infty(a) + \alpha F(\xi u_\infty)(a)\\
&\qquad + v(a)(1 - \alpha\lambda[a \mid \xi u_\infty(\cdot) + v(\cdot)] - \alpha\gamma(a)) + \alpha[1 - \xi u_\infty(a)]\lambda[a \mid v(\cdot)]\\
&\quad \geq \xi u_\infty(a) + \alpha\xi F(u_\infty)(a) + \alpha(1-\xi)\lambda[a \mid v(\cdot)]\\
&\quad \geq \xi(u_\infty + \alpha F u_\infty)(a) + \alpha\epsilon(1-\xi)K_1(a)\int_0^{a_\dagger} K_2(a')v(a')\,da'.
\end{aligned}$$

Then by (3.8), (3.10), and (4.7), we obtain (4.12). □

LEMMA 4.4. *Let* $u_\infty \in C$ *be a nontrivial fixed point of* T *and let* $0 < \xi < 1$. *Then there exists a constant* $\epsilon_4 > 0$ *such that*

$$T(\xi u_\infty) \geq \xi u_\infty + \xi(1-\xi)\epsilon_4 m_1, \tag{4.13}$$

where $m_1(a) = m^2(a)$.

Proof. Note that

$$\begin{aligned}
&(\xi u_\infty + \alpha F(\xi u_\infty))(a)\\
&\quad = \xi(u_\infty + \alpha F(u_\infty))(a) + \alpha\xi(1-\xi)u_\infty(a)\lambda[a \mid u_\infty(.)]\\
&\quad \geq \xi(u_\infty + \alpha F(u_\infty))(a) + \alpha\xi(1-\xi)\epsilon u_\infty(a)K_1(a)\int_0^{a_\dagger} K_2(a')u_\infty(a')da'\\
&\quad \geq \xi(u_\infty + \alpha F(u_\infty))(a) + \alpha\xi(1-\xi)\epsilon_4\frac{d}{da}\left[\int_0^a K_1(a')\,da'\right]^2,
\end{aligned}$$

where we have used (4.8). Thus, by (3.8) and (3.10) we get (4.13). □

Note that by (4.13) it follows that

$$\xi u_\infty + \xi(1-\xi)\epsilon_4 m_1 \in C,$$

and moreover, since $a_1 < b_2$, it follows that

$$\int_0^{a_\dagger} K_2(a')m_1(a')\,da' > 0.$$

Finally, we have Theorem 4.5.

THEOREM 4.5. *T has at most one nontrivial fixed point.*

Proof. Let $u_\infty \not\equiv v_\infty$ be two nontrivial fixed points. Without losing generality we can assume that we do not have $u_\infty \leq v_\infty$. By (4.8), $v_\infty \geq (\epsilon_1/\epsilon_2)u_\infty$. Let ξ be the maximum constant such that

$$v_\infty \geq \xi u_\infty. \tag{4.14}$$

It must be that $0 < \xi < 1$. Then, by Lemma 4.4

$$v_\infty = T^2(v_\infty) \geq T^2(\xi u_\infty) \geq T(\xi u_\infty + \xi(1-\xi)\epsilon_4 m_1) \tag{4.15}$$

and by Lemma 4.3 (take $v = \xi(1-\xi)\epsilon_4 m_1$)

$$v_\infty \geq \xi u_\infty + (1-\xi)\epsilon_3 m$$

so that

$$v_\infty \geq \xi u_\infty + (1-\xi)\frac{\epsilon_3}{\epsilon_2} u_\infty,$$

which is impossible because it contradicts the definition of ξ. □

The existence of a nontrivial equilibrium is related to the Fréchet derivative of T, which we denote by $DT[u]$, and depends on the spectral radius ρ of $DT[0]$; actually we have Proposition 4.6.

PROPOSITION 4.6. *Let $\rho \leq 1$; then the mapping T has no nontrivial fixed points in C.*

Proof. Suppose that u_∞ is a nontrivial fixed point of T and let $0 < \xi < 1$. Then, by Lemmas 4.3 and 4.4

$$T^2(\xi u_\infty) \geq T(\xi u_\infty + \xi(1-\xi)\epsilon_4 m_1) \geq \xi u_\infty + (1-\xi)\frac{\epsilon_3}{\epsilon_2} u_\infty = \xi(1+\delta)u_\infty$$

and, by noting that

$$(Tu)(a) = (DT[0]u)(a) - \int_0^a e^{-\frac{1}{\alpha}(a-s)} \lambda[s \mid u(\cdot)]u(s)\, ds, \tag{4.16}$$

which implies

$$T(u) \leq DT[0]u \quad \text{for any } u \in C,$$

we have

$$\xi \left(DT[0]\right)^2 u_\infty \geq T^2(\xi u_\infty) \geq \xi(1+\delta)u_\infty$$

so that

$$\left(DT[0]\right)^2 u_\infty \geq (1+\delta)u_\infty,$$

which implies $\rho > 1$, thus contradicting the assumption. □

On the other hand, we also have Proposition 4.7.

PROPOSITION 4.7. *If $\rho > 1$, T has at least one nontrivial fixed point.*

Proof. $DT[0]$ is linear, completely continuous, and leaves the positive cone $K \equiv \{h \in E \mid h \geq 0\}$ invariant. Then by the Krein–Rutman theorem, there exists an eigenvector $h^* \geq 0$ with eigenvalue ρ. Since h^* satisfies the problem:

$$\begin{aligned} &\alpha\rho \frac{d}{da} h^*(a) = [1 - \alpha\gamma(a) - \rho]h^*(a) + \alpha\lambda[a \mid h^*(\cdot)], \\ &h^*(0) = 0. \end{aligned}$$

Proceeding as in the proof of Proposition 4.2, we can find positive constants c_1, c_2 such that

$$c_1\, m(a) \leq h^*(a) \leq c_2\, m(a).$$

This, in particular, implies that h^* belongs to $L^\infty(0, a_\dagger)$, so that, without loss of generality, we can assume that $h^* \in C$. Moreover, by (4.16), we have, for $0 < \xi < 1$,

$$T(\xi h^*)(a) = \xi \rho h^*(a) - \xi^2 \int_0^a e^{-\frac{1}{\alpha}(a-s)} \lambda[s \mid h^*(\cdot)] h^*(s)\, ds.$$

Since

$$\int_0^a \lambda[s \mid h^*(\cdot)] h^*(s) ds \le \lambda^+ c_2 m(a) \le \lambda^+ \frac{c_2}{c_1} h^*(a),$$

we have

$$T(\xi h^*)(a) \ge \xi h^*(a) + \xi \left[(\rho - 1) - \xi \lambda^+ \frac{c_2}{c_1} \right] h^*(a),$$

and, if ξ is sufficiently small,

$$T(\xi h^*) \ge \xi h^*.$$

Thus the sequence

$$u_n = T^n(\xi h^*) \in C$$

is monotonic nondecreasing and converges to some u_∞ which is a nontrivial fixed point of T. □

5. Convergence to the equilibrium. Our first statement concerns the case when only the trivial equilibrium exists.

THEOREM 5.1. *Assume that no nontrivial equilibrium exists; then*

$$\forall u_0 \in C, \quad S(t)u_0 \longrightarrow 0 \quad \text{as } t \to \infty. \tag{5.1}$$

Proof. Assume that (5.1) is not true for some initial datum u_0, then define

$$\bar{u} = \limsup_{t \to \infty} S(t)u_0.$$

This $\bar{u} \not\equiv 0$ does exist in C because $L^1(0, a_\dagger)$ is a complete lattice. Moreover, since $\sup_{s \ge r} S(s)u_0$ is nonincreasing as a function of r, then

$$\bar{u} = \lim_{r \to \infty} \sup_{s \ge r} S(s)u_0.$$

Now

$$\begin{aligned} S(t)\bar{u} &= \lim_{r \to \infty} S(t) \sup_{s \ge r} S(s)u_0 \ge \lim_{r \to \infty} \sup_{s \ge r} S(t+s)u_0 \\ &= \lim_{r \to \infty} \sup_{s \ge t+r} S(s)u_0 = \lim_{r \to \infty} \sup_{s = r} S(s)u_0 = \bar{u}, \end{aligned}$$

which implies that $S(t)\bar{u}$ is nondecreasing as a function of t; in fact, if $h > 0$,

$$S(t+h)\bar{u} = S(t)S(h)\bar{u} \ge S(t)\bar{u}.$$

Then

$$u_\infty = \lim_{t \to \infty} S(t)\bar{u}$$

exists and is a nontrivial fixed point of $S(t)$, i.e., it is a nontrivial equilibrium. This contradicts the assumptions and proves that (5.1) is true. □

Let us now consider the case in which a (unique) nontrivial equilibrium u_∞ exists. In this case we have to introduce the following condition:

$$\int_t^{a\dagger} K_2(a)u_0(a-t)\,da > 0 \quad \text{for some } t \geq 0. \tag{5.2}$$

If this condition is satisfied, we say that u_0, with support in $[0, a_\dagger]$, is a *nontrivial initial datum.*

Then we have Proposition 5.2.

PROPOSITION 5.2. *Let u_0 be a nontrivial initial datum. Then there exist $t_0 > 0$ and $\xi > 0$ such that*

$$S(t_0)u_0 \geq \xi u_\infty.$$

Proof. We start from equation (3.19) (where $u(t) = S(t)u_0$), which yields

$$u(t) \geq \frac{1}{\alpha}\int_0^t e^{-\frac{1}{\alpha}(t-s)}e^{(t-s)A}[u(s) + \alpha F(u(s))]\,ds. \tag{5.3}$$

Now, since $\alpha < 1/(\lambda^+ + \gamma^+)$ we have

$$u(s) + \alpha F(u(s)) \geq w(s),$$

where

$$(w(s))(a) = \alpha\epsilon K_1(a)\int_0^{a_\dagger} K_2(a')(u(s))(a')\,da'$$

so that

$$\left(e^{(t-s)A}w(s)\right)(a) \geq \begin{cases} \alpha\epsilon K_1(a+s-t)\displaystyle\int_0^{a_\dagger} K_2(a')(u(s))(a')\,da' & \text{for } a > t-s, \\ 0 & \text{for } a < t-s, \end{cases}$$

and, plugging this into (5.3), we obtain

$$(u(t))(a) \geq \epsilon e^{-\frac{1}{\alpha}t}\int_{(t-a)\vee 0}^t K_1(a+s-t)h(s)\,ds \tag{5.4}$$

with

$$h(s) = \int_0^{a_\dagger} K_2(a')(u(s))(a')\,da'.$$

We will prove that

$$\text{There exists } s_\infty \text{ such that } h(s) > 0 \text{ for } s > s_\infty. \tag{5.5}$$

To this aim, we first prove that

$$\text{If } h(s) > 0 \text{ for } s \in [s_1, s_2], \text{ then } h(s) > 0 \text{ for } s \in [s_1 + c_1, s_2 + c_2], \tag{5.6}$$

where $c_1 = (a_1 \vee a_2) - b_1$ and $c_2 = b_2 - a_1$.

Note first that $u(t)(a) > 0$ if

$$a > a_1, \quad t > s_1, \quad s_1 - b_1 < t - a < s_2 - a_1. \tag{5.7}$$

In fact, if these are fulfilled then $(t-a)\vee s_1 < t\wedge s_2$ and, setting $I \equiv [(t-a)\vee s_1, t\wedge s_2]$ we have

$$I \subset [s_1, s_2], \qquad \{a - t + I\} \cap [a_1, b_1] \neq \emptyset \tag{5.8}$$

so that from (5.4)

$$(u(t))(a) \geq \epsilon e^{-\frac{1}{\alpha}t} \int_{(t-a)\vee s_1}^{t\wedge s_2} K_1(a+s-t)h(s)\,ds > 0 \tag{5.9}$$

because by (5.8) the integrand is positive on some interval contained in I.

Finally, fix $t \in [s_1 + c_1, s_2 + c_2]$ and note that

$$h(t) \geq \int_{a_2}^{b_2} K_2(a')(u(t))(a')\,da' > \int_{a_1\vee a_2}^{b_2} K_2(a')(u(t))(a')\,da'.$$

Now, K_2 is positive on $[a_1 \vee a_2, b_2]$ and, by (5.7), $u(t)(a)$ is positive for $a > a_1$ and $a \in J \equiv [a_1 + t - s_2, t + b_1 - s_1]$ but our choice of t yields

$$a_1 + t - s_2 < a_1 + c_2 = b_2,$$

$$t + b_1 - s_1 > b_1 + c_1 = a_1 \vee a_2,$$

so that $J \cap [a_1 \vee a_2, b_2] \neq \emptyset$, and $h(t) > 0$.

Now that (5.6) is proven, note that if u_0 is a nontrivial initial datum, then from (3.19) we see that for some positive constant δ

$$h(s) = \int_0^{a_\dagger} K_2(a)u(s)(a)\,da \geq \delta \int_0^{a_\dagger} K_2(a)(e^{sA}u_0)(a) = \delta \int_s^{a_\dagger} K_2(a)u_0(a-s)\,da > 0$$

for s in some interval $[s_1, s_2]$. Thus, iterating statement (5.6) we have, for any integer n

$$h(s) > 0 \quad \text{in } [s_1 + nc_1, s_2 + nc_2] \tag{5.10}$$

so that, since $c_2 - c_1 > 0$ and $c_2 > 0$ by (2.11), if n is sufficiently large, successive intervals are overlapping, and consequently statement (5.5) is proven.

We now use (5.5) in order to prove that

$$\text{There exists } t_\infty \text{ such that } (u(t))(a) > 0 \text{ for } a > a_1 \text{ and } t > t_\infty. \tag{5.11}$$

In fact, if $t > s_\infty + a_\dagger + b_1$, we have $t - a > s_\infty$ and, by (5.4)

$$\begin{aligned}(u(t))(a) &\geq \epsilon e^{-\frac{1}{\alpha}t} \int_{(t-a)}^{t} K_1(a+s-t)h(s)\,ds \\ &\geq \epsilon e^{-\frac{1}{\alpha}t} \int_{a_1}^{a\wedge b_1} K_1(\sigma)h(\sigma+t-a)\,d\sigma > 0.\end{aligned}$$

Now let $t_0 > s_\infty + a_\dagger$; then $t_0 - a > s_\infty$ and

$$\begin{aligned}(u(t_0))(a) &\geq \epsilon e^{-\frac{1}{\alpha}t_0} \int_{(t_0-a)}^{t_0} K_1(a+s-t_0)h(s)ds \\ &> \epsilon e^{-\frac{1}{\alpha}t_0} h_- \int_{(t_0-a)}^{t_0} K_1(a+s-t_0)ds \\ &= \epsilon e^{-\frac{1}{\alpha}t_0} h_- \int_0^a K_1(s)ds = \epsilon e^{-\frac{1}{\alpha}t_0} h_- m(a),\end{aligned}$$

where $h_- = \inf_{s\in[s_\infty,t_0]} h(s)$. Finally, by (4.8) we have

$$(u(t_0))(a) \geq \xi u_\infty(a), \tag{5.12}$$

where $\xi = (\epsilon/\epsilon_2)e^{-\frac{1}{\alpha}t_0}h_-$, and the proof is done. □

Finally, we come to the following theorem.

THEOREM 5.3. *Let u_∞ be the unique nontrivial equilibrium; then for any nontrivial initial datum u_0*

$$S(t)u_0 \longrightarrow u_\infty \quad \text{as } t \to \infty. \tag{5.13}$$

If u_0 is not nontrivial, then

$$S(t)u_0 = 0 \quad \text{for } t \geq a_\dagger. \tag{5.14}$$

Proof. Let ξ and t_0 be as in the previous proposition; then letting **1** denote the function equal to 1 almost everywhere on $[0, a_\dagger]$,

$$\xi u_\infty \leq S(t_0)u_0 \leq \mathbf{1}$$

and

$$S(t)\xi u_\infty \leq S(t+t_0)u_0 \leq S(t)\mathbf{1}. \tag{5.15}$$

Now

$$\xi u_\infty = \xi S(t)u_\infty \leq S(t)\xi u_\infty \quad \text{and} \quad S(t)\mathbf{1} \leq \mathbf{1}.$$

Consequently, $S(t)\xi u_\infty$ and $S(t)\mathbf{1}$ are monotonic with respect to t (nondecreasing and nonincreasing, respectively) and converge to a nontrivial equilibrium which coincides with u_∞ by the uniqueness Theorem 4.5; then (5.13) follows from (5.15).

Now let

$$\int_t^{a_\dagger} K_2(a)u_0(a-t)da = 0 \quad \text{for all } t \geq 0.$$

Let us show that the iterates $u^n(t)$ defined in (3.20) also satisfy

$$\int_s^{a_\dagger} K_2(a)u^n(t)(a-s)\,da = 0 \quad \text{for all } t \geq 0 \text{ and } s \geq 0. \tag{5.16}$$

Actually, if (5.16) is true, then

$$\begin{aligned}\lambda[a \mid u^n(t)(\cdot)] &\leq K_0(a)p_\infty(a)u^n(t)(a) + K_1(a)\int_0^{a_\dagger} K_2(a)u^n(t)(a)\,da \\ &= K_0(a)p_\infty(a)u^n(t)(a)\end{aligned}$$

and

$$u^n(t) + \alpha F(u^n(t)) \leq (1+\alpha k_0^+)u^n(t).$$

This implies that

$$u^{n+1}(t) \leq e^{tA}u_0 + c\int_0^t e^{(t-\sigma)A}u^n(\sigma)\,d\sigma,$$

where $c > 0$; and, consequently,

$$\int_s^{a_\dagger} K_2(a)u^{n+1}(t)(a-s)\,da \le \int_{s+t}^{a_\dagger} K_2(a)u_0(a-t-s)da + c\int_0^t \int_{s+t-\sigma}^{a_\dagger} K_2(a)u^n(\sigma)(a-t-s+\sigma)\,da\,d\sigma = 0.$$

Thus (5.16) is true by induction.

Now (5.16) implies that

$$\int_s^{a_\dagger} K_2(a)u(t)(a-s)da = 0 \quad \text{for all } t \ge 0 \text{ and } s \ge 0 \tag{5.17}$$

so that if $t > a_\dagger$

$$u(t) \le c\int_0^t e^{(t-s)A}u(s)\,ds, \tag{5.18}$$

which implies (5.14). In fact, fix $t > a_\dagger, a \in [0, a_\dagger]$ and define

$$\phi(x) = u(t-a+x)(x) \quad \text{for } x \in [0, a_\dagger];$$

then by (5.18)

$$\phi(x) \le c\int_0^x \phi(s)\,ds,$$

which implies $\phi(x) \equiv 0$. □

6. Conclusion. As noted in the Introduction, our results provide a complete global stability analysis and show that sustained periodic behavior is not possible in the general $s - i - s$ age-dependent epidemic model with constant total population size. Consequently, the addition of age-structure does not change the gross dynamics of this model, even though it adds considerable demographic detail and realism to the model. We demonstrate the existence of a sharp endemic threshold and our results can be used to obtain the age distribution of the susceptible and the infective subpopulations. The basic endemic threshold for this model is identified as the spectral radius of an explicit linear operator which depends on the pertinent demographic and epidemiological parameters. It generalizes the explicit expressions that were obtained for this threshold by Busenberg, Cooke, and Iannelli [2] for the two special cases of purely intercohort and intracohort forces of infection. The proof that we provide for the existence and uniqueness of the endemic state, when it exists, is constructive. It can be used as the basis of an iterative method for obtaining approximations, within prescribed error bounds, of the subpopulation age distributions at the endemic state.

The generality of our results is limited only by the restrictions (2.9)–(2.13) on the age-dependent force of infection term. The conditions (2.10)–(2.13) do not place any essential restrictions on $K(a, a')$ that go beyond what would be dictated by the biological situation that is being modeled. However, from (2.9) it is seen that if $K(\tilde{a}, \tilde{a}') = 0$ for some fixed $(\tilde{a}, \tilde{a}') \in [0, a_\dagger] \times [0, a_\dagger]$, then either $K_1(\tilde{a}) = 0$, or $K_2(\tilde{a}') = 0$. Hence, we either have $K(\tilde{a}, a) = 0$ or $K(a, \tilde{a}') = 0$ for all $a \in [0, a_\dagger]$. That is, if the disease is never transmitted to any age $\tilde{a}$ individual from every age $\tilde{a}'$ individual, then either it cannot be transmitted from any age group $a \ne \tilde{a}$ to age $\tilde{a}$ individuals, or it

cannot be transmitted from any age $\tilde{a}'$ individual to an age $a \neq \tilde{a}'$ individual for all $a \in [0, a_\dagger]$. Since we are dealing with densities which are defined almost everywhere, this represents a real restriction only when there exist two age-class intervals $I_1, I_2 \subset [0, a_\dagger]$, for which $K(a, a') = 0$ for all $(a, a') \subset I_1 \times I_2$. In the epidemiological situtation that we are modeling, such complete immunity to disease transmission across two separate age classes is extremely unlikely. Hence, condition (2.9) includes the situations that are of interest from the viewpoint of the epidemiological model. In particular, asymmetric forms of K with $K(a, a') \neq K(a', a)$ are included in our result, as well as cases where individuals from two separate age groups have only limited, but nonvanishing, disease transmitting interactions. Nevertheless, it would be interesting to explore the dynamic implications of any substantial weakening of condition (2.9).

We finally note that we have excluded the possibility of vertical transmission in our model by requiring that $i(0, t) = 0$ in (1.2). This restriction is not essential, and was made for technical reasons since the inclusion of vertical transmission adds complications to the the model equations. Similar results to the ones that we have proved here also hold for the model with vertical transmission and they will be taken up in a separate paper.

REFERENCES

[1] R. ANDERSON AND R. MAY, *Age-related changes in the rate of disease transmission: implication for the designing of vaccination programs,* J. Hyg. Camb., 94 (1985), pp. 365–436.

[2] S. BUSENBERG, K. COOKE, AND M. IANNELLI, *Endemic thresholds and stability in a class of age-structured epidemics,* SIAM J. Appl. Math., 48 (1988), pp. 1379–1395.

[3] C. CASTILLO-CHAVEZ, H. W. HETHCOTE, V. ANDREASSEN, S. A. LEVIN, AND W. M. LIU, *Epidemiological models with age-structure, proportionate mixing and cross-immunity,* J. Math. Biol., 27 (1989), pp. 233–258.

[4] M. EL DOMA, *Analysis of nonlinear integro-differential equations arising in age dependent epidemic models,* Ph.D. thesis, Claremont Graduate School, Claremont, CA, 1985.

[5] G. GRIPENBERG, *On a nonlinear integral equation modelling an epidemic in an age-structured population,* J. Reine Angew. Math., 341 (1983), pp. 54–67.

[6] K. HADELER AND K. DIETZ, *Nonlinear hyperbolic partial differential equations for the dynamics of parasite populations,* Comput. Math. Appl., 9 (1983), pp. 415–430.

[7] H. W. HETHCOTE, H.W. STECH, AND P. VAN DEN DRIESSCHE, *Periodicity and stability in epidemic models: a survey,* in Differential Equations and Applications in Ecology, Epidemics and Population Dynamics, S. Busenberg and K. L. Cooke, eds., Academic Press, New York, 1983, pp. 65–82.

[8] F. HOPPENSTEADT, *An age-dependent epidemic model,* J. Franklin Inst., 197 (1974), pp. 325–333.

[9] M. A. KRASNOSEL'SKII, *Positive Solutions of Operator Equations,* Noordhoff, Groningen, the Netherlands, 1964.

[10] W. A. LONDON AND J. A. YORKE, *Recurrent outbreaks of measles, chickenpox and mumps,* I, Amer. J. Epid., 98 (1973), pp. 453–468.

[11] A. MCLEAN, *Dynamics of Childhood Infections in High Birthrate Countries,* Lecture Notes in Biomath., 65, Springer-Verlag, Berlin, 1985, pp. 171–197.

[12] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations,* Lecture Notes in Appl. Math. Sci., 44, Springer-Verlag, Berlin, 1983.

[13] D. SCHENZLE, *An age structured model for pre and post-vaccination measles transmission,* IMA J. Math. Appl. Med. Biol., 1 (1984), pp. 169–191.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1081-1101, July 1991

014

CONVERGENCE FOR STRONGLY ORDER-PRESERVING SEMIFLOWS*

HAL L. SMITH†‡ AND HORST R. THIEME†§

Abstract. Convergence of almost all trajectories of strongly order-preserving semiflows is derived under suitable additional assumptions. These essentially consist in slightly sharpening the strongly order-preserving property and in the continuous differentiability of the flow with respect to the state variable. Required spectral properties of the linearizations of the flow around equilibria usually follow in the same way as the compactness and monotonicity assumptions for the flow itself. The proofs are based on sharpened versions of the limit set dichotomy and the sequential limit set trichotomy.

Key words. monotone dynamical system, strongly order-preserving semiflow, open dense set of (stable) convergent points, limit set dichotomy, nonordering principle, sequential limit set trichotomy, spectral theory of strongly positive linear operators

AMS(MOS) subject classifications. 34C35, 34G20, 47H20

Introduction. In the fundamental paper [7], Hirsch establishes that most orbits of a strongly monotone semiflow on a strongly ordered space X tend to the set E of the equilibria. Somewhat more precisely, Hirsch shows that the set of all $x \in X$ for which $\omega(x)$, the omega (positive) limit set of x, satisfies $\omega(x) \subset E$, is a "large" (open dense, residual, set of full measure) subset of X. Points x for which $\omega(x) \subset E$ are called *quasi-convergent* points; they are called *convergent* points if $\omega(x)$ consists of a single point of E.

The results in [7] extend earlier work of Hirsch [4], [5] for competitive and cooperative ordinary differential equations to infinite-dimensional semiflows. Applications are made to parabolic partial differential equations.

Matano pioneered the monotone dynamical systems approach to parabolic equations in [12] and [13] where he introduces the notion of upward (downward) stability of equilibria and considers the dynamics near such equilibria using monotonicity arguments. In his more recent work [13]-[16], he outlines an alternative approach to monotone dynamical systems, parallel to Hirsch's, which does not require that the space be strongly ordered and is, therefore, more suitable for applications to parabolic equations. One of the main results of Matano's theory, like that of Hirsch's, provides sufficient conditions for "most" points to be quasi-convergent.

In a recent paper [27], Smith and Thieme combine ideas from Hirsch and Matano to obtain a theory which improves several of the results of both authors while at the same time being conceptually simpler. We adopt (a slight generalization of) Matano's idea of a *strongly order-preserving* semiflow, but base our results on modified versions of two fundamental results due to Hirsch, namely, the *nonordering principle* for limit sets [7] and the *limit set dichotomy* [7]. These two principles also hold under Matano's weaker assumptions (see [16]) and can actually be shown by modifying Hirsch's proofs accordingly.

In this paper we present conditions for the existence of a dense open set of convergent points, i.e., any trajectory starting in that dense open set converges to a single equilibrium.

* Received by the editors June 19, 1989; accepted for publication (in revised form) June 28, 1990.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287-1804.

‡ The research of this author was supported in part by National Science Foundation grant DMS 8722279.

§ The research of this author was supported by a Heisenberg scholarship of Deutsche Forschungsgemeinschaft.

Trivial conditions consist in assuming that the set of equilibria E is discrete or totally ordered. Both assumptions guarantee that any quasi-convergent point is convergent: in the first case because the ω limit set is connected, in the second because the ω limit set cannot contain two different order-related points due to the *nonordering principle.*

These two trivial conditions do not harmonize with the spirit of the theory, namely, finding general structural conditions (e.g., strongly order preserving) under which almost all solutions have a very simply asymptotic behavior. For systems of equations and particularly for infinite-dimensional problems, it can be very difficult to verify one of the two trivial conditions above.

Our aim is to obtain general structural conditions that guarantee the existence of a dense open set of convergent points. These conditions do not imply that any quasi-convergent point is convergent (though this holds if we assume that it is a stable point). Essentially they consist in strengthening the notion of a *strongly order-preserving semiflow* (remaining more general than *strongly monotone*) and, more importantly, assuming that the semiflow is C^1, i.e., continuously differentiable in the state variable, at least in a neighborhood of each point of E. We also require spectral conditions for the linearization of the semiflow around equilibria, but these usually follow from the same considerations which establish the strengthened strongly order-preserving properties and the compactness properties of the semiflow.

Our approach has been inspired by the work of Poláčik [20], who finds conditions for abstract semilinear parabolic evolution equations to have a residual set of convergent points. In setting down these conditions, he is able to exploit the special framework for semilinear parabolic equations, as described by Henry [3]. Our approach is more general in the sense that we make no assumptions on the underlying system of equations generating the semiflow. Our abstract results, applied to the case considered by Poláčik, yield stronger conclusions than those in [20], assuming less smoothness but more compactness. We conclude that the set of stable convergent points is open and dense, whereas Poláčik concludes that the set of convergent points is a residual set.

In the remainder of this section we describe some basic ideas and notation and preview several of the main results.

Let X be an ordered metric space with metric d and order relation $\leqq$. We write $x < y$ if $x \leqq y$ and $x \neq y$. Points x and y in X are *ordered* if either $x < y$ or $y < x$. Given two subsets A and B of X we write $A \leqq B$ ($A < B$) whenever $x \leqq y$ ($x < y$) for each choice of $x \in A$ and $y \in B$. If $x < y$ then $[x, y] = \{z \in X : x \leqq z \leqq y\}$. A subset Y of X is *order convex* if $[y_1, y_2] \subset Y$ whenever $y_1, y_2 \in Y$ and $y_1 < y_2$.

We assume that the order and the topology on X are compatible in the sense that $x \leqq y$ whenever $x_n \to x$, $y_n \to y$, and $x_n \leqq y_n$ for all n. If $x \in X$ we say that x can be approximated from below (above) in X if there exists a sequence $\{x_n\}$ in X satisfying $x_n < x_{n+1} < x$ ($x < x_{n+1} < x_n$) for $n \geqq 1$ and $x_n \to x$.

The space X is said to be normally ordered if there exists a constant $k > 0$ such that

$$d(u, v) \leqq k d(x, y)$$

for all x, y, u, v with $u, v \in [x, y]$.

For simplicity here, we assume that Φ is a global semiflow on X. More precisely, let $\Phi : X \times \mathbf{R}^+ \to X$ be a semiflow on X, that is, Φ is continuous and $\Phi_t(x) \equiv \Phi(x, t)$ satisfies $\Phi_0(x) = x$ for every x and $\Phi_t \Phi_s = \Phi_{t+s}$ for every $t, s \geqq 0$. For $x \in X$, let

$$\mathcal{O}^+(x) = \{\Phi_t(x) : t \geqq 0\},$$
$$\omega(x) = \bigcap_{t \geqq 0} \overline{\mathcal{O}^+(\Phi_t(x))}$$

be the orbit initiating at x and the omega (positive) limit set of $\mathcal{O}^+(x)$, respectively. If $\mathcal{O}^+(x)$ has compact closure in X, then $\omega(x)$ is nonempty, compact, connected, and invariant, i.e., $\Phi_t(\omega(x)) = \omega(x)$, $t \geqq 0$, and $\Phi_t(x) \to \omega(x)$ as $t \to \infty$. We let $E = \{x \in X: \Phi_t(x) = x, t \geqq 0\}$ be the set of equilibria. The set of quasi-convergent points is denoted by $Q = \{x \in X: \omega(x) \subset E\}$ and the set of convergent points by $C = \{x \in Q: \omega(x)$ is a singleton set$\}$.

The semiflow Φ is said to be *monotone* provided

$$\Phi_t(x) \leqq \Phi_t(y) \quad \text{whenever } x \leqq y.$$

Following Matano [15], [16], Φ is said to be *strongly order-preserving* if Φ is monotone and whenever $x, y \in X$ with $x < y$, there exist open sets U and V, $x \in U$, $y \in V$, and $t_0 \geqq 0$ such that $\Phi_{t_0}(U) \leqq \Phi_{t_0}(V)$. By monotonicity, it follows that

$$\Phi_t(U) \leqq \Phi_t(V) \quad \text{for } t \geqq t_0.$$

In almost every case in which the strong order-preserving property can be verified, it results from the fact that the semiflow Φ eventually operates as a strongly monotone semiflow on some subset Z of an ordered Banach space $(Y, \|\cdot\|)$ with cone Y_+ and where Int $Y_+ \neq \varnothing$. We write $\leqq_Y$ for the partial order on Y generated by Y_+ and $\ll$ for the strong ordering induced by Int Y_+; that is, $y_1 \ll y_2$ if and only if $y_2 - y_1 \in$ Int Y_+. We assume that Z is a common subset of Y and of X, that the two partial orders $\leqq$ and $\leqq_Y$ induced on Z by the ordering on X and the ordering on Y agree, and that the inclusion $Z \hookrightarrow X$ is continuous, where Z is given the topology induced by the norm on Y. In order to simplify the statement of our results here, we assume that Z is order convex.

Additional hypotheses intertwine the set Z and the semiflow Φ:

(I) $\Phi_t(Z) \subseteq Z$ for $t \geqq 0$.

(J) There exists $\tau \geqq 0$ such that $\Phi_\tau(X) \subseteq Z$ and $\Phi_\tau: X \to Z$ is continuous.

(M) If $x_1, x_2 \in X$ satisfy $x_1 < x_2$, then $\Phi_\tau(x_1) \ll \Phi_\tau(x_2)$.

(D) Φ_τ is continuously differentiable on Z with derivative $\Phi'_\tau(z)$.

(Σ) For any equilibrium e of Φ satisfying $\rho(e) := \operatorname{spr}(\Phi'_\tau(e)) \geqq 1$, $\rho(e)$ is a pole of the resolvent of $\Phi'_\tau(e)$ with finite rank and $N(\rho(e)I - \Phi'_\tau(e))$ is spanned by an element of Int Y_+.

Here we use the notation spr L for the spectral radius of a bounded linear operator L and $N(L)$ for the null space of L. Less restrictive hypotheses are stated with more care and detail in § 3.

The spectral hypothesis (Σ) is crucial to our result. It may appear very restrictive until we realize that the monotonicity of Φ implies that the derivative $\Phi'_\tau(e)$ is a positive linear operator on Y. In applications we usually can establish the compactness of $\Phi'_\tau(e)$ which implies that $\rho(e)$ is a pole of the resolvent with finite rank. The same consideration that checks the strong monotonicity assumption (M) usually provides that the operator $\Phi'_\tau(z)$ is strongly positive, i.e., maps nonzero positive elements into Int Y_+. This implies [22] that the eigenspace associated with $\rho(e)$ is one-dimensional and spanned by a vector in Int Y_+.

Hypotheses (D) and (Σ) can be replaced by an alternative hypothesis that $\Phi'_t(e)$ generates a strongly continuous semigroup for each $e \in E$ and where spectral conditions are assumed to hold for the generator of this semigroup instead of for the semigroup itself. This approach is more appropriate for applications to semilinear parabolic equations.

It turns out that (J) and (M), together with the monotonicity of Φ on X, imply (see Lemma 2.1) that Φ is strongly order preserving on X. The additional assumptions allow a significant improvement of the *limit set dichotomy* in [27] (see Proposition 2.2).

PROPOSITION. *Let Φ be a monotone semiflow on X satisfying* (I), (J), (M), (D), *and* (Σ). *If $x_1, x_2 \in X$ are such that $\mathcal{O}^+(x_1)$ and $\mathcal{O}^+(x_2)$ have compact closure in X and $x_1 < x_2$, then either*

(a) $\omega(x_1) < \omega(x_2)$, *or*

(b) $\omega(x_1) = \omega(x_2) = \{e\}$ *for some equilibrium $e \in E$.*

Armed with this sharpened limit set dichotomy, we improve our sequential limit set trichotomy of [27] in Proposition 2.3 and this, in turn, leads to two of our main results stated below.

THEOREM 1. *Suppose that for each $x \in X$, $\mathcal{O}^+(x)$ has compact closure in X and that x can be approximated either from above or from below in X by a sequence $\{x_n\}$ such that $\bigcup_{n \geq 1} \omega(x_n)$ has compact closure in X. Then* Int C *is dense in X.*

In fact, under minimal additional hypotheses we can show that X contains an open and dense set of stable convergent points (see Theorem 2.6). A point $x \in X$ is called a *stable point* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d(\Phi_t(x), \Phi_t(y)) < \varepsilon$ for $t \geqq 0$ whenever $y \in X$ and $d(x, y) < \delta$.

In many applications, however, it is unreasonable to assume that we have a global semiflow on the space X. We typically have a local semiflow on X (see § 3 for precise definitions) and can show that $\mathcal{O}^+(x)$ has compact closure in X only for some subset X_0 of points of X. In § 3 we show that, by suitably modifying assumptions (I), (J), (M), (D), and (Σ), we can obtain the following result.

THEOREM 2. *Let X_0 be an open subset of X such that every point of X_0 can be approximated both from above and from below in X. Assume either*

(a) *X is normally ordered, every orbit starting in X_0 has compact closure in X, and for any convergent monotone sequence x_n in X_0, $\bigcup_{n \geq 1} \omega(x_n)$ has compact closure in X, or*

(b) *Each point of X_0 belongs to a neighborhood U such that $U \times [0, \infty) \subset \operatorname{Dom} \Phi$ and $\Phi(U \times [t_0, \infty))$ has compact closure in X for some $t_0 > 0$.*

Then X_0 contains a dense open subset of stable convergent points.

The modified versions of assumptions (I), (J), (M), (D), and (Σ) are assumed to hold in Theorem 2.

The compactness assumptions contained in both (a) and (b) of Theorem 2 are rather mild for most applications. However, these assumptions are not made in [7] or [20]. In § 3 we show that neither of these compactness assumptions is required for the weaker result that there are at most a countable number of nonconvergent points on any totally ordered arc of points of X having precompact orbits in X. Such a result for quasi-convergent points was obtained by Hirsch [7, Thm. 7.3].

For applications of the results of monotone dynamical systems theory, we refer the reader to [4]–[6], [23], and [24] for applications to systems of ordinary differential equations, to [7], [12]–[17], [20], and [21] for applications to parabolic initial-boundary value problems, to [25], [26], and [28] for applications to functional differential equations, and to [10] and [11] for applications to systems of parabolic equations with time delays. In these applications the stronger results of this paper are typically obtained by finding additional conditions that guarantee that the induced semiflow is continuously differentiable in the state variable.

The organization of this paper is as follows. In § 1 we establish the key result, Proposition 1.2, leading to the improved limit set dichotomy. Proposition 1.2 states that for a strongly order-preserving discrete dynamical system with certain smoothness and compactness properties, any compact, connected, unordered, and nonempty set

of equilibria which attracts two orbits $\{u_n\}$ and $\{v_n\}$ with $u_n \ll v_n$ must be a singleton. In § 2, we use this result to establish the improved limit set dichotomy (the proposition above) and Theorem 1. In this section we assume that Φ is a global semiflow on X. In § 3 we show how to modify the ideas of § 2 to apply in case the semiflow is a local one. Theorem 2 is a corollary of the main result of this section. Finally, in § 4 we consider two applications of our results. The first application is to semilinear parabolic equations. Here we compare our results to those of Poláčik [20]. The second is to systems of functional differential equations, following [25].

1. Discrete prelude. Let Y be an ordered Banach space with cone Y_+ such that Int $Y_+ \neq \emptyset$. Let $Z \subset Y$ and $S: Z \to Y$ be a mapping which is order preserving and has the following property. If $e \in Z$ is a fixed point of S, i.e., $Se = e$ and l a natural number, then there exists a neighborhood U of e in Z which is order convex in Y such that (i) S^l is defined on U and (ii) S is continuously differentiable on U; that is, there exists a continuous mapping $S': U \to L_+(Y)$ such that

$$S(z) - S(z_0) = S'(z_0)(z - z_0) + \phi(z, z_0)\|z - z_0\|$$

with $\phi(z, z_0) \to 0$ as $z \to z_0$ where $z_0, z \in U$. Here $L_+(Y)$ denotes the cone of bounded linear operators on Y mapping Y_+ into itself. The reader may think of S as the map Φ_τ where Φ is the semiflow in the Introduction.

We assume the existence of a compact connected set K of fixed points of S in Z with the following properties:

(A) For any $e \in K$ with $\rho_e \equiv \operatorname{spr}(S'(e)) \geqq 1$, ρ_e is a pole of the resolvent of $S'(e)$ of order m_e and $N((\rho_e I - S'(e))^{m_e})$ is finite-dimensional. Moreover, $N(\rho_e I - S'(e))$ is spanned by a vector in Int Y_+.

(B) If $e_1, e_2 \in K$ and $e_1 \leqq e_2$, then $e_1 = e_2$, i.e., K contains no pair of distinct order-related points.

If an isolated point α of the spectrum of a closed linear operator L is a pole of the resolvent of L of finite order and $\cup_{k \geqq 1} N((\alpha I - L)^k)$ is finite-dimensional, we say that α is a pole of the resolvent of *finite rank.*

Assumption (A) has several consequences that are conveniently described here. First, we note that $\rho_e \geqq 1$ for all $e \in K$ unless K is a singleton. For if $e \in K$ is a limit point of K, then necessarily one belongs to the spectrum of $S'(e)$. So let us assume that $\rho_e \geqq 1$ for all $e \in K$. It follows that ρ_e depends continuously on $e \in K$. In fact, the upper semicontinuity follows from the well-known upper semicontinuity of the spectrum as a function of the operator and the lower semicontinuity is a result of the lower semicontinuity of isolated parts of the spectrum [8, Thm. 3.1, Remark 3.3, Thm. 3.16, Chap. IV]. As ρ_e is a pole of the resolvent of the order $m = m_e$

$$Y = N((\rho_e I - S'(e))^m) \oplus \operatorname{Im}((\rho_e I - S'(e))^m)$$

where both subspaces are invariant under $S'(e)$ and ρ_e does not belong to the spectrum of the restriction of $S'(e)$ to Im $((\rho_e I - S'(e))^m)$. Let $P(e)$ denote the projection onto $N((\rho_e I - S'(e))^m)$ along Im $((\rho_e I - S'(e))^m)$. Then $P(e)$ depends continuously on $e \in K$ [8, Thm. 3.16, Chap. IV].

Finally, let $z(e)$ denote the unique eigenvector of $S'(e)$ associated with ρ_e of unit norm belonging to Int Y_+. We establish the continuity of $z(e)$ on K as follows. Let $e_n \to e$ as $n \to \infty$ where e_n, $e \in K$, and let $z_n = z(e_n)$, $z = z(e)$. Then $(I - P(e))z_n = [P(e_n) - P(e)]z_n \to 0$ as $n \to \infty$, and $\{P(e)z_n\}$ is precompact in the finite-dimensional

space $N((\rho_e I - S'(e))^m)$. Hence, $\{z_n\}$ is precompact, and, if $z_{n_i} \to u$ for some subsequence $\{n_i\}$, then $S'(e)u = \lim_{i\to\infty} S'(e)z_{n_i} = \lim_{i\to\infty} \rho_{e_{n_i}} z_{n_i} = \rho_e u$. It follows that $u = z$ and that $z_n \to z$ as $n \to \infty$.

PROPOSITION 1.1. *If K is not a singleton, then $\rho_e > 1$ for all $e \in K$.*

Proof. Assume K is not a singleton. Then $\rho_e \geqq 1$ for all $e \in K$ as noted above. Suppose that $\rho_e = 1$ for some $e \in K$. Choose a sequence $e_n \in K$, $e_n \neq e$, $e_n \to e$ as $n \to \infty$. Then

$$e - e_n = S(e) - S(e_n) = S'(e)(e - e_n) + o(\|e - e_n\|).$$

Put $v_n = (e - e_n / \|e - e_n\|)$. We have

$$v_n = S'(e)v_n + r_n, \quad r_n \to 0 \quad \text{as } n \to \infty.$$

In view of our hypothesis (A) and the discussion following it, we may conclude that $I - S'(e)$, restricted to $(I - P(e))Y$, is invertible where $P(e)$ is the spectral projection defined in the discussion preceding the proposition. Let $w_n = (I - P(e))v_n$ and observe that

$$(I - S'(e))w_n = (I - P(e))r_n \to 0$$

which implies that $w_n \to 0$. As in a previous argument, this implies that $\{v_n\}$ is precompact in Y and any limit point belongs to $P(e)Y$. If $v_{n_i} \to v$ as $i \to \infty$ for some subsequence, then $v = S'(e)v$. Thus, $v = \pm z(e)$, and this implies that $e_{n_i} = e - \|e - e_{n_i}\| v_{n_i}$ is related to e for large i, contradicting (B) and completing our proof.

Proposition 1.1 is the only place where we require that $N(\rho_e I - S'(e))$ be spanned by a vector in Int Y_+ (see (A)). The proof of Proposition 1.2 below requires only that the null space $N(\rho_e I - S'(e))$ contain a vector in Int Y_+.

PROPOSITION 1.2. *If there exist sequences $\{u_n\}$ and $\{v_n\}$ in Z, $u_{n+1} = S(u_n)$, $v_{n+1} = S(v_n)$, $u_n \gg v_n$ and $\operatorname{dist}(u_n, K) \to 0$, $\operatorname{dist}(v_n, K) \to 0$ as $n \to \infty$, then K is a singleton.*

Proof. Suppose K is not a singleton set but that there exist two sequences as in the hypotheses. Let $w \in \text{Int } Y_+$ and define a new norm $\|\cdot\|_w$ on Y by $\|y\|_w = \inf\{\lambda > 0 : -\lambda w \leqq y \leqq \lambda w\}$. Since $w \in \text{Int } Y_+$, there exists $\delta > 0$ such that for all $y \in Y$, $y \neq 0$, it follows that $w \pm \delta(y/\|y\|) \geqq 0$. From this, we conclude that $\|y\|_w \leqq \delta^{-1}\|y\|$ for all $y \in Y$, i.e., the w-norm is weaker than the original norm. All convergence and continuity statements below are with respect to the original norm. The $\|\cdot\|_w$ norm is used only for order purposes.

Define $\alpha_n = \sup\{\alpha > 0 : u_n \geqq v_n + \alpha w\}$ and note that $\alpha_n > 0$ and $u_n \geqq v_n + \alpha_n w$. We observe that $\alpha_n \to 0$ as $n \to \infty$. If not, $\alpha_{n_i} \geqq \alpha > 0$ for some subsequence $\{n_i\}$ and, by the compactness of K and the fact that $u_n, v_n \to K$ as $n \to \infty$, we may as well assume that $u_{n_i} \to u \in K$ and $v_{n_i} \to v \in K$ as $i \to \infty$. Then $u \geqq v + \alpha w$ so u and v are distinct order-related points of K in contradiction to our hypothesis (B).

Choose $e_n \in K$ such that $v_n - e_n \to 0$ as $n \to \infty$. Now K is assumed not to be a singleton, so $\rho_e > 1$ for all $e \in K$ by Proposition 1.1. As K is compact and $e \mapsto \rho_e$ is continuous, there exists $r > 1$ such that $\rho_{e_n} \geqq r$ for all n. Let $z_n \in \text{Int } Y_+$, $\|z_n\|_w = 1$ be the normalized positive eigenvector of $S'(e_n)$ corresponding to ρ_{e_n} so $z_n \leqq w$. There is an $\varepsilon > 0$ such that $z_n \geqq \varepsilon w$ holds for all n. This follows from the continuity of the map $e \mapsto z(e) \in \text{Int } Y_+$ assigning to each $e \in K$ the normalized ($\|z(e)\|_w = 1$) positive eigenvector of $S'(e)$ corresponding to ρ_e, which, in turn, implies the continuity of the map $\beta : K \to (0, \infty)$ defined by $\beta(e) = \sup\{\beta > 0 : z(e) \geqq \beta w\}$. As K is compact, there exists $\varepsilon > 0$ such that $\beta(K) \geqq \varepsilon$.

Choose l, a positive integer, such that $r^l \varepsilon > 1$. By (i) and (ii) and the compactness of K there is a cover of K by order-convex open $U \subset Z$ such that S^l is differentiable on U and for each sufficiently large n, the entire segment $v_n + \xi\alpha_n w$, $0 \leqq \xi \leqq 1$, belongs

to one of the U. Hence

$$S^l(v_n+\alpha_n w)-S^l v_n=(S^l)'(e_n)\alpha_n w+\alpha_n\Delta_n$$

where

$$\Delta_n\equiv\int_0^1[(S^l)'(v_n+\xi\alpha_n w)-(S^l)'(e_n)]w\,d\xi.$$

Since $v_n+\alpha_n w-e_n\to 0$, K is compact, and $(S^l)'$ is continuous, it follows that $\delta_n\equiv\|\Delta_n\|_w\to 0$ as $n\to\infty$. Thus we have

$$\begin{aligned}S^l(v_n+\alpha_n w)-S^l v_n&\geqq[S'(e_n)]^l\alpha_n w-\alpha_n\delta_n w\\&\geqq[S'(e_n)]^l\alpha_n z_n-\alpha_n\delta_n w\\&\geqq r^l\alpha_n z_n-\alpha_n\delta_n w\\&\geqq(r^l\varepsilon-\delta_n)\alpha_n w\\&\geqq\alpha_n w\end{aligned}$$

for all large n. We have used that $\varepsilon w\leqq z_n\leqq w$ and $e_n\in E$ in the estimate above. From this last estimate, we obtain

$$\begin{aligned}u_{n+1}=S^l u_n&\geqq S^l(v_n+\alpha_n w)\\&\geqq S^l v_n+\alpha_n w\\&=v_{n+l}+\alpha_n w\end{aligned}$$

for all large n. This implies that

$$\alpha_{n+l}\geqq\alpha_n$$

for large n, contradicting that $0<\alpha_n\to 0$ as $n\to\infty$ and completing our proof.

2. Stable convergence. Returning to the continuous scenario outlined in the Introduction, consider an order-preserving (or monotone) semiflow Φ on an ordered metric space X. We assume that for each $x\in X$, $\mathcal{O}^+(x)$ has compact closure in X.

Suppose that there is an ordered Banach space Y with order $\leqq_Y$ generated by a cone Y_+ having nonempty interior, Int Y_+, and that Y and X have a common subset Z with the following properties:

(i) The metric d_Z induced by the norm in Y on Z is stronger than the metric induced by d, the metric on X, i.e., the embedding from (Z, d_Z) to (Z, d) is continuous.

(ii) The restriction of $\leqq_Y$ to Z agrees with the restriction of $\leqq$ to Z.

In the following, the topology on Z will be assumed to be that induced by d_Z unless otherwise indicated. Additional hypotheses intertwine the set Z and the semiflow Φ:

(I) $\Phi_t(Z)\subseteq Z$ for $t\geqq 0$.

(J) There exists $\tau\geqq 0$ such that $\Phi_\tau(X)\subseteq Z$ and $\Phi_\tau: X\to Z$ is continuous.

(M) If $x_1, x_2\in X$ satisfy $x_1<x_2$, then $\Phi_\tau(x_1)\ll\Phi_\tau(x_2)$ where τ is as in (J) and the strong order $\ll$ is induced by Int Y_+ on Z.

(D) For any $e\in E$, there exists a neighborhood U of e in Z which is order convex in Y and a $\tau>0$ such that Φ_τ is continuously differentiable on U, i.e., there exists a continuous mapping

$$\Phi'_\tau: U\to L_+(Y)$$

such that

$$\Phi_\tau(z)-\Phi_\tau(z_0)=\Phi'_\tau(z_0)(z-z_0)+\phi(z,z_0)\|z-z_0\|_Y$$

for $z, z_0 \in U$ with $\|\phi(z,z_0)\|_Y \to 0$ as $z \to z_0$.

(Σ) For any equilibrium e of Φ satisfying $\rho(e) := \operatorname{spr}(\Phi'_\tau(e)) \geqq 1$, $\rho(e)$ is a pole of the resolvent of $\Phi'_\tau(e)$ with finite rank and $N(\rho(e)I-\Phi'_\tau(e))$ is spanned by an element of Int Y_+, where τ is as in (D) above.

Remark 1. As mentioned in the Introduction, assumption (Σ) is less restrictive than it appears. In applications some compactness property of Φ will typically imply that $\Phi'_\tau(e)$ (or one of its powers) can be represented as a sum of a compact operator and an operator with norm strictly less than 1. This implies that $\rho(e) \geqq 1$ is a pole of the resolvent with finite rank [18]. The same consideration which derives (M) usually shows that $\Phi'_\tau(e)$ is a strongly positive operator, i.e., maps positive nonzero elements into Int Y_+. In particular, $\Phi'_\tau(e)$ is irreducible, implying that the eigenspace of $\Phi'_\tau(e)$ associated with $\rho(e)$ is one-dimensional and spanned by a vector in Int Y_+. See Theorem 3.2 of [22, p. 270]. It is possible to give alternative conditions for (Σ) to hold which are weaker in certain respects (see [2, §§ 8, 9], [18], [22]), but the above appears to be most appropriate in the applications.

Remark 2. In many cases of interest, Z will be order convex, and, in this case, it is possible to verify (M) and (Σ) by establishing the following:

(Σ') For any $z \in Z$, $\Phi'_\tau(z)$ is a strongly positive operator, i.e., maps any positive nonzero element in Y into Int Y_+. Moreover, for $e \in E$, $\Phi'_\tau(e)$ is compact.

Observe that (M) follows from (Σ') by the integral version of Taylor's formula.

Remark 3. In many applications, semilinear parabolic equations, for example, $\Phi'_\tau(e)$, $e \in E$, will usually be a strongly continuous semigroup. In this case, it is more natural to replace the spectral condition (Σ) involving $\Phi'_t(e)$ by one which involves the infinitesimal generator $A(e)$ of $\Phi'_t(e)$. We will show, at the end of this section, how our results may also be obtained by replacing (D) and (Σ) by the assumption (S) below.

(S) For any $e \in E$, $t > 0$, Φ_t can be continuously differentiated in a neighborhood of e in Z which is order convex in Y and $\Phi'_t(e)$, $t \geqq 0$, forms a strongly continuous semigroup on Y. Assume that, for $e \in E$, there exists $\varepsilon > 0$ and a neighborhood U of e such that Φ' is defined and bounded on $U \times [0, \varepsilon]$ as an $L(Y)$-valued map. Assume that the generator $A(e)$ has compact resolvent. If the spectral bound $s(e) = \sup\{\operatorname{Re}\lambda : \lambda \in \sigma(A(e))\}$ of $A(e)$ satisfies $s(e) \geqq 0$, let $N(s(e)I-A(e))$ be spanned by an element $z(e) \in \operatorname{Int} Y^+$.

It turns out that (J) and (M), together with the monotonicity of Φ on X, imply that Φ is strongly order preserving on X.

LEMMA 2.1. *Assuming the hypotheses above, we have*

(a) Φ_t *maps* X *continuously into* Z *for* $t \geqq \tau$.

(b) *For any* $x_1, x_2 \in X$ *with* $x_1 < x_2$, $\Phi_t(x_1) < \Phi_t(x_2)$ *for* $t \geqq 0$ *and* $\Phi_t(x_1) \ll \Phi_t(x_2)$ *for* $t \geqq \tau$.

(c) *Any compact invariant set in* X *is also compact and invariant in* Z.

(d) Φ *is strongly order preserving on* X.

Proof. We prove (d) as the other statements are obvious. If $x_1 < x_2$ then $\Phi_\tau(x_1) \ll \Phi_\tau(x_2)$ so we may find neighborhood U_Y, V_Y in Z with $U_Y \leqq V_Y$ and $\Phi_\tau(x_1) \in U_Y$, $\Phi_\tau(x_2) \in V_Y$. As Φ_τ maps X continuously into Z, we have $\Phi_\tau(U) \subseteq U_Y$, $\Phi_\tau(V) \subseteq V_Y$ for suitable neighborhoods U, V of x_1, x_2 in X.

We next prove the following improvement of the *limit set dichotomy* in [27]. For the remainder of this section, we assume the existence of a space Z satisfying (i) and (ii) above and we assume (I), (J), (M), (D), and (Σ) hold.

PROPOSITION 2.2. *If $x_1, x_2 \in X$ and $x_1 < x_2$, then either*

(a) $\omega(x_1) < \omega(x_2)$, *or*

(b) $\omega(x_1) = \omega(x_2) = \{e\}$ *for some equilibrium $e \in E$.*

Proof. We only need to show that if $\omega(x_1) = \omega(x_2) \subseteq E$, then (b) holds. Let $K = \omega(x_1)$ and note that K is compact in Y by Lemma 2.1(c). By the nonordering principle, K contains no pair of order-related points. Let $v_n = \Phi_{n\tau}(x_1)$, $u_n = \Phi_{n\tau}(x_2)$ and $S: Z \to Z$ be given by $S = \Phi_\tau$. Then K is a compact connected set of fixed points of S, $u_{n+1} = Su_n$, $v_{n+1} = Sv_n$ and $u_n \ll v_n$ holds for $n \geqq 2$. Moreover, $\operatorname{dist}_Z(u_n; K) \to 0$ and $\operatorname{dist}_Z(v_n; K) \to 0$ as $n \to \infty$. In fact, (J) implies that $\Phi_t(x_1) \to K$ in Z as $t \to \infty$. In order to apply Proposition 1.2, it remains only to check (A) of § 1, as the fact that S is continuously differentiable and order preserving is inherited from the corresponding properties of Φ_τ. But (Σ) implies (A) of § 1, so by Proposition 1.2, K is a singleton.

Armed with this sharpened limit set dichotomy, we improve our sequential limit set trichotomy of Proposition 3.1 of [27]. Observe that only alternative (c) of the earlier version is affected. Note also that we include explicitly in the hypotheses only the compactness hypotheses required for the proof, rather than assuming hypothesis (C) as in [27].

PROPOSITION 2.3 (sequential limit set trichotomy). *Let $x_0 \in X$ have the property that it can be approximated from below in X by a sequence $\tilde{x}_n$ such that $\overline{\bigcup_{n\geqq 1} \omega(\tilde{x}_n)}$ is compact in X. Then there exists a subsequence x_n of $\tilde{x}_n$ such that $x_n < x_{n+1} < x_0$, $n \geqq 1$, with $x_n \to x_0$ satisfying one of the following*:

(a) *There exists $u_0 \in E$ such that*

$$\omega(x_n) < \omega(x_{n+1}) < u_0 = \omega(x_0), \qquad n \geqq 1$$

and

$$\lim_{n\to\infty} \operatorname{dist}(\omega(x_n), u_0) = 0.$$

(b) *There exists $u_0 \in E$ such that*

$$\omega(x_n) = u_0 < \omega(x_0), \qquad n \geqq 1.$$

If $u \in E$ and $u < \omega(x_0)$, then $u \leqq u_0$.

(c) *There exists $u_0 \in E$ such that $\omega(x_n) = \omega(x_0) = u_0$ for $n \geqq 1$.*

The special property "$\overline{\bigcup_{n\geqq 1} \omega(\tilde{x}_n)}$ is compact in X" of the sequence $\tilde{x}_n$ approximating x from below will be required for several results of this and the next section. It is convenient to have some notation for it. We say x has property (ω_-) ((ω_+)) if it can be approximated from below (above) by a sequence x_n such that $\overline{\bigcup_{n\geqq 1} \omega(x_n)}$ is compact in X. We say x has property (ω) if it has both property (ω_-) and (ω_+).

Repeating the proof of Theorem 3.3 in [27] and using the stronger sequential limit set trichotomy in Proposition 2.3 in place of Proposition 3.1 in [27], we find that there is an open dense subset of X which consists of convergent points.

THEOREM 2.4. *Suppose each point of X has property (ω_-) or (ω_+). Then* Int C *is dense in X.*

In fact, under minimal additional hypotheses, we can show that X contains an open and dense set of stable convergent points. A point $x \in X$ is called a *stable point* if, for every $\varepsilon > 0$, there exists $\delta > 0$ such that $d(\Phi_t(x), \Phi_t(y)) < \varepsilon$ for $t \geqq 0$ whenever $y \in X$ and $d(x, y) < \delta$. The set of stable points in X is denoted by S.

PROPOSITION 2.5. *Let $x \in S$ be a stable point which has property (ω_-) or (ω_+). Then $x \in C$.*

Proof. Consider an approximating sequence $\{x_n\}$. Then, after passing to a subsequence only alternatives (a) or (c) of Proposition 2.3 are possible because points close to x have limit sets which are close to the limit set of x. Thus $x \in C$.

Proposition 2.5 implies the following conclusion from Theorem 3.9 in [27].

THEOREM 2.6. *Let X be a normally ordered metric space. Suppose that each point $x \in X$ has property (ω_-) or (ω_+) and that X contains an open and dense subset X_0 of points which have property (ω). Then $S \subseteq C$ and* Int S *is dense in X.*

In [27], Theorem 3.13, we give conditions under which the set of asymptotically stable points A is dense in X. This theorem cannot be improved by adding the assumptions of this paper, but note that we now have $A \subseteq C$ instead of $A \subseteq Q$. Recall that $A \subseteq S \subseteq C$ by Proposition 2.5 and that, by its definition, A is open.

We conclude this section by sketching modifications of the arguments in the proofs of Propositions 1.1 and 1.2, now in the continuous setting of the semiflow Φ_t, which are required in order to establish that assumptions (I), (J), (M), and (S) imply the limit set dichotomy (Proposition 2.2). We begin with the analogue of Proposition 1.1.

PROPOSITION 2.7. *Assume* (I), (J), (M), *and* (S) *hold. Let K be a nonempty, compact, connected, unordered set of equilibria of Φ_t which is not a singleton. Then $s(e) > 0$ for all $e \in K$.*

Proof. Let $e \in K$. As in Proposition 1.1, we can find a sequence $e_n \in K$ such that $e_n \neq e$ and $e_n \to e$ as $n \to \infty$. We have

$$e - e_n = \Phi_t(e) - \Phi_t(e_n).$$

Set $v_n = (e - e_n / \|e - e_n\|)$. Then

$$v_n = \Phi_t'(e) v_n + r_n(t) \tag{2.1}$$

where $r_n(t) \to 0$ as $n \to \infty$ pointwise in t. By our boundedness assumption concerning Φ' in (S), we can assume that the

$$r_n(t) = \int_0^1 [\Phi_t'(se + (1-s)e_n) - \Phi_t'(e)] v_n \, ds, \qquad n \geqq 1,$$

are uniformly bounded for t in $[0, \varepsilon]$. Multiplying through (2.1) by $e^{-\lambda t}$ where λ is chosen sufficiently large, and integrating the result from zero to ε, we obtain

$$\begin{aligned} \lambda^{-1}(1 - e^{-\lambda\varepsilon}) v_n &= \int_0^\varepsilon e^{-\lambda t} \Phi_t'(e) v_n \, dt + \int_0^\varepsilon e^{-\lambda t} r_n(t) \, dt \\ &= (\lambda I - A(e))^{-1} v_n - \int_\varepsilon^\infty e^{-\lambda t} \Phi_t'(e) v_n \, dt + \int_0^\varepsilon e^{-\lambda t} r_n(t) \, dt, \end{aligned}$$

where, in the second equality, we have used the Laplace transform formula for the resolvent of the generator of the semigroup $\Phi_t'(e)$ (see [19, proof of Remark 5.4, Chap. 1]). By the uniform boundedness of the $r_n(t)$ on $[0, \varepsilon]$ and the fact that they converge pointwise to zero, $\int_0^\varepsilon e^{-\lambda t} r_n(t) \, dt \to 0$ as $n \to \infty$. Let α denote the measure of noncompactness [9]. Using the compactness of the resolvent and standard properties of α, we obtain from the last formula

$$\begin{aligned} \lambda^{-1}(1 - e^{-\lambda\varepsilon}) \alpha\{v_n\} &\leqq \alpha\left\{ \int_\varepsilon^\infty e^{-\lambda t} \Phi_t'(e) v_n \, dt \right\} \\ &\leqq \frac{M e^{(\omega(\varepsilon) - \lambda)\varepsilon}}{\lambda - \omega(e)} \alpha\{v_n\}, \end{aligned}$$

where we use that $\|\Phi_t'(e)\| \leqq M e^{\omega(e)t}$ for some $M \geqq 1$ and $\omega(e)$. Choosing λ sufficiently large in the inequality above establishes that $\alpha\{v_n\} = 0$. Hence, v_n has a subsequence converging to a point u and, by (2.1),

$$u = \Phi_t'(e)u, \qquad t \geqq 0.$$

Multiplying through by $e^{-\lambda t}$ and integrating over $t \in [0, \infty)$, we get

$$\lambda^{-1}u = (\lambda I - A(e))^{-1}u$$

for large λ. In particular, $r_\lambda(e)$, the spectral radius of $(\lambda I - A(e))^{-1}$, satisfies $r_\lambda(e) \geqq \lambda^{-1}$ for large λ. As $r_\lambda(e) = (\lambda - s(e))^{-1} \geqq \lambda^{-1}$, it follows that $s(e) \geqq 0$. If $s(e) = 0$, then it follows that $r_\lambda(e) = \lambda^{-1}$ and so u is an eigenvector of $A(e)$ associated with $s(e)$. By (S), $u = \pm z(e)$, where $z(e)$ is as in (S). But then $e_n = e - \|e - e_n\| v_n$ is related to e for some large n, contradicting that K is unordered. This proves the proposition.

Now, the proof of Proposition 2.2 can be followed to obtain the sequences $u_{n+1} = Su_n$ and $v_{n+1} = Sv_n$ where $S = \Phi_\tau$, τ as in (M), $u_n \ll v_n$ and $\operatorname{dist}_Z(u_n; K) \to 0$, $\operatorname{dist}_Z(v_n; K) \to 0$ as $n \to \infty$ with $K = \omega(x_1)$, a nonempty compact, connected, unordered set of equilibria. Now we need Proposition 1.2 to conclude that K is a singleton. Let us see how to modify the proof of Proposition 1.2 so that we may apply it. Indeed, we can follow the proof of Proposition 1.2 through the first paragraph without change. For the remainder of the proof we take z_n as the eigenvector of $A(e_n)$ corresponding to $s(e_n) > 0$. Clearly z_n is an eigenvector of $\Phi_\tau'(e_n)$ associated with the eigenvalue $r(e_n) = e^{s(e_n)\tau} > 1$, although $r(e_n)$ need not be the spectral radius of $\Phi_\tau'(e_n)$. We need only to establish the continuity of the maps $e \mapsto z(e)$ and $e \mapsto s(e)$ in order to conclude the existence of $r > 1$ and $\varepsilon > 0$ such that $r(e_n) > r$ and $z(e_n) \geqq \varepsilon w$ hold for all n. The remainder of the proof of Proposition 1.2 can then be followed to obtain the desired contradiction.

In order to establish the continuity of the maps $e \mapsto z(e)$, $e \mapsto s(e)$, we show that for sufficiently large λ, the resolvents $(\lambda I - A(e))^{-1}$ depend continuously on $e \in K$. Indeed, if $e_n \to e$, $A = A(e)$, $A_n = A(e_n)$, $T_n(t) = \Phi_t'(e_n)$, and $T(t) = \Phi_t'(e)$, then from the boundedness of Φ' in $U \times [0, \varepsilon]$, where U is a neighborhood of e (see (S)), it follows that $\|T_n(t)\|, \|T(t)\| \leqq M e^{\omega t}$, $n \geqq 1$, for some $M \geqq 1$ and ω (see the proof of Theorem 2.2 [19, Chap. 1]). Since $T_n(t) \to T(t)$ in the uniform operator topology as $n \to \infty$, we can argue as in Theorem 4.2 of [19, Chap. 3] to conclude that $(\lambda I - A(e_n))^{-1} \to (\lambda I - A(e))^{-1}$ as $n \to \infty$ in the uniform operator topology. Note that part (b) of Theorem 4.2 of [19, Chap. 3] proves that if $T_n(t) \to T(t)$ in the strong topology, then the resolvents converge in the strong topology. We obtain the stronger conclusion above by virtue of the stronger hypothesis that $T_n(t) \to T(t)$ in the uniform topology. The spectral radius $r_\lambda(e)$ of $(\lambda I - A(e))^{-1}$ depends continuously on e by the same arguments establishing the continuity of the spectral radius ρ_e in § 1. Moreover, $r_\lambda(e) = (\lambda - s(e))^{-1}$ and $z(e)$ spans the eigenspace of $(\lambda I - A(e))^{-1}$ corresponding to $r_\lambda(e)$. Thus, the continuity of $s(e)$ and $z(e)$ follows as in the arguments preceding Proposition 1.1. This completes our sketch of the validity of Proposition 2.2 under the hypotheses (I), (J), (M), and (S). In particular, these assumptions can replace (I), (J), (M), (D), and (Σ) in all the results of this section.

3. Stable convergence for local semiflows. In this section we consider the case that the semiflow Φ may not be globally defined on X. We assume only that $\Phi : \operatorname{Dom} \Phi \subseteq X \times \mathbf{R}^+ \to X$ is a local semiflow as defined below. Our aim will be to show that, through slight modification of the ideas in [27] and the previous section, a stable convergence result may still be obtained in this more general setting. As in § 2, we assume that X is an ordered metric space.

Following Hirsch [7], we assume that $\Phi: \operatorname{Dom}\Phi \to X$, where $\operatorname{Dom}\Phi$ is an open subset in $X \times \mathbf{R}^+$ containing $X \times \{0\}$, and Φ is continuous on $\operatorname{Dom}\Phi$ and satisfies

(i) $\Phi_0 = \Phi(\cdot, 0) = \mathrm{id}_X$

(ii) For all $s, t \geqq 0$, $\operatorname{Dom}\Phi_{s+t} = \Phi_t^{-1}(\operatorname{Dom}\Phi_s)$ and

$$\Phi_{s+t} = \Phi_t \circ \Phi_s$$

where $\Phi_t: \operatorname{Dom}\Phi_t \to X$, $\operatorname{Dom}\Phi_t = \{x: (x, t) \in \operatorname{Dom}\Phi\}$ and $\Phi_t(x) = \Phi(x, t)$ for $x \in \operatorname{Dom}\Phi_t$.

For each $x \in X$, $\{t \geqq 0: x \in \operatorname{Dom}\Phi_t\} = [0, \sigma_x)$ is a half-open interval with right-hand endpoint $\sigma_x \leqq +\infty$. The orbit of x is $\mathcal{O}^+(x) = \{\Phi_t(x): 0 \leqq t < \sigma_x\}$. A set $K \subset X$ is positively invariant if it contains the orbit of each of its points. If K is compact and positively invariant, then $\sigma_x = +\infty$ for each $x \in K$. In particular, if $\overline{\mathcal{O}^+(x)}$ is compact in X, then $\sigma_x = +\infty$ and the omega limit set $\omega(x)$ has the properties described in the Introduction.

We assume hereafter that Φ is monotone (order preserving) in the sense that $\Phi_t(x) \leqq \Phi_t(y)$ holds whenever $x \leqq y$ and $0 \leqq t < \min\{\sigma_x, \sigma_y\}$.

As in § 2, we assume that there is an ordered Banach space Y with order $\leqq_Y$ generated by a cone Y_+ with $\operatorname{Int} Y_+ \neq \varnothing$ and that Y and X have a common subset Z with the assumptions (i) and (ii) of § 2 holding.

We modify the assumptions (I)-(M) of § 2 as follows:

(I) $\Phi_t(z) \in Z$, $0 \leqq t < \sigma_z$, for each $z \in Z$.

(J) There exists $\tau \geqq 0$ such that, for any $x \in X$ with $\sigma_x > \tau$, there is a neighborhood U of x in X such that Φ_τ is defined on U and maps U, with the X topology, continuously into Z, with the Y topology.

(M) There exists $\tau > 0$ such that if $x_1, x_2 \in X$ satisfy $x_1 < x_2$ and $\sigma_{x_i} > \tau$, then $\Phi_\tau(x_1) \ll \Phi_\tau(x_2)$, where the strong order $\ll$ is induced by $\operatorname{Int} Y_+$ on Z.

Assumptions (D) and (Σ) of § 2 are assumed to hold as well, where τ in assumption (Σ) is the same as τ in (D). Alternatively, assumption (S) of Remark 3 may be assumed in place of (D) and (Σ).

By taking τ sufficiently large, we may assume that the τ of (J) and (M) are identical. Indeed, if in (J), $\tau < t < \sigma_x$, then there exists a neighborhood V contained in U such that $V \times [0, t] \subset \operatorname{Dom}\Phi$ and Φ_t maps V, with the X topology, continuously into Z, with the Y topology. From (M), it is easily deduced that $\Phi_t(x_1) \ll \Phi_t(x_2)$ holds for all t satisfying $\tau \leqq t < \min\{\sigma_{x_1}, \sigma_{x_2}\}$. As (D) and ($\Sigma$) remain valid with τ replaced by any integral multiple of itself, it is clear that we may assume, by passing to a sufficiently large τ if necessary, that the τ of (J), (M), (D), and (Σ) are identical.

The strong order-preserving property of Φ must be understood in a slightly modified form. Arguing as in Lemma 2.1, we can show that if $x_1, x_2 \in X$ satisfy $\sigma_{x_1}, \sigma_{x_2} > \tau$ and $x_1 < x_2$, then there exist neighborhoods U of x_1 and V of x_2 in X such that $U \times [0, \tau]$ and $V \times [0, \tau]$ are contained in $\operatorname{Dom}\Phi$ and $\Phi_\tau(U) \leqq \Phi_\tau(V)$. Since we will usually assume that the points x with which we work belong to orbits having compact closure in X, this modification of the strong order-preserving property will not have a significant effect.

The reader may check that the limit set dichotomy in Proposition 2.3 of [27] remains valid in the present context provided that we assume that $\overline{\mathcal{O}^+(x)}$ and $\overline{\mathcal{O}^+(y)}$ are compact in X. Similarly, the sequential limit set trichotomy in Proposition 3.1 of [27] is valid if we assume that $\overline{\mathcal{O}^+(x_0)}$ is compact in X, $\overline{\mathcal{O}^+(\tilde{x}_n)}$ is compact in X for each n, corresponding to the sequence $\tilde{x}_n$ approximating x_0 from below in X, and $\overline{\bigcup_{n \geqq 1} \omega(x_n)}$ is compact in X.

It is now easy to check that the improved limit set dichotomy (Proposition 2.2) and the sequential limit set trichotomy (Proposition 2.3) of § 2 remain valid, provided that we assume that the appropriate points (x and y in Proposition 2.2 and x_0 and $\tilde{x}_n$ in Proposition 2.3) belong to orbits having compact closure in X. Note that hypothesis (i) of § 1 is automatically satisfied since for any $t>0$ and $e\in E$ there exists an open set U such that $U\times[0, t)\subset \operatorname{Dom}\Phi$.

We now introduce a local uniform compactness-of-orbits assumption on Φ which will allow us to obtain strong stability results as in Theorem 2.6 without the normality assumption of that theorem. This is important for applications to semilinear parabolic equations where the usual interpolation spaces fail to have the normality property.

We say that orbits of Φ are locally uniformly compact at x_0, provided there exists a neighborhood U of x_0 such that $\sigma_y=+\infty$ for all $y\in U$ and

$$\Phi(U\times[t_0,\infty))$$

has compact closure in X for some $t_0>0$. Observe that this assumption implies that $\overline{\cup_{n\geqq 1}\omega(x_n)}$ is compact in X if x_n is a sequence approximating x_0 from below (above) in X and $x_n\in U$ for $n\geqq 1$.

Let $x\in X$ belong to a neighborhood U in X such that $\overline{\mathcal{O}^+(y)}$ is compact in X for each $y\in U$. Then we say x is stable provided for each $\varepsilon>0$ there exists $\delta>0$ such that $d(\Phi_t(x),\Phi_t(y))<\varepsilon$ for $t\geqq 0$ whenever $y\in X$ satisfies $d(x, y)<\delta$. Let S denote the set of stable points in X. We say that x is asymptotically stable if there is a neighborhood V of x, contained in U, with the property that for every $\varepsilon>0$, there exists $t_\varepsilon>0$ such that $d(\Phi_t(x),\Phi_t(y))<\varepsilon$ if $t\geqq t_\varepsilon$ and $y\in V$. Let A denote the set of asymptotically stable points in X. Note that A is open and that $A\subset S$.

The next two lemmas prepare the way for a main result of this section, Theorem 3.3 below.

LEMMA 3.1. *Assume that x_0 is contained in a neighborhood U such that $\sigma_y=+\infty$ for $y\in U$ and*

$$\Phi(U\times[t_0,\infty))$$

has compact closure in X for some $t_0>0$. Furthermore, assume that $x_0\in X$ can be approximated from below such that either alternative (b) *or* (c) *of the sequential limit set trichotomy holds.*

Then there exists a neighborhood $W\subset U$ of x_0 such that any point $x\in W$, $x<x_0$, is an asymptotically stable point.

Proof. Let (b) hold. From Corollary 3.2 of [27] we find a neighborhood W of x_0 contained in U with the following property: For each $x\in W$, $x<x_0$, there is a neighborhood $V=V_x$ of x in W, a positive integer $N=N_x$ and a time $T=T_x>0$ such that

$$u_0\leqq\Phi_t(V)\leqq\Phi_t(x_N)\quad\text{for } t\geqq T.$$

We claim that, for any $\varepsilon>0$, there is some $t_\varepsilon>0$ such that

$$d(\Phi_t(v), u_0)<\varepsilon,\qquad t>t_\varepsilon,\quad v\in V.$$

If this does not hold, we find sequences $t_j\to\infty$, $j\to\infty$ and $v_j\in V$ such that, for each j,

$$d(\Phi_{t_j}(v_j), u_0)>\varepsilon>0.$$

By our compactness assumption we can assume, possibly after choosing a subsequence, that $\Phi_{t_j}(v_j)$ converges towards an element $w\in X$, $d(w, u_0)\geqq\varepsilon$. As

$$u_0\leqq\Phi_{t_j}(v_j)\leqq\Phi_{t_j}(x_N)\to u_0,\qquad j\to\infty,$$

we conclude that $w=u_0$, a contradiction.

We now consider alternative (c). From Corollary 3.2 of [27] we find a neighborhood W of x_0 contained in U with the following property. For each $x \in W$, $x < x_0$, there is a neighborhood $V = V_x$ of x in W, a positive integer $N = N_x$ and a time $T = T_x > 0$, such that

$$\Phi_t(x_1) \leqq \Phi_t(V) \leqq \Phi_t(x_N) \quad \text{for } t \geqq T.$$

We claim that, for any $\varepsilon > 0$, there is some $t_\varepsilon > 0$ such that

$$d(\Phi_t(v), \Phi_t(x_1)) < \varepsilon, \qquad t > t_\varepsilon, \quad v \in V.$$

If this does not hold, we find sequences $t_j \to \infty$, $j \to \infty$, and $v_j \in V$ such that

$$d(\Phi_{t_j}(v_j), \Phi_{t_j}(x_1)) > \varepsilon > 0.$$

By our compactness assumption we can assume, after possibly choosing subsequences, that

$$\Phi_{t_j}(x_1) \to w_1, \quad \Phi_{t_j}(v_j) \to w_2, \quad \Phi_{t_j}(x_N) \to w_3, \quad j \to \infty$$

and

$$d(w_2, w_1) \geqq \varepsilon > 0, \qquad w_1, w_3 \in \omega(x_1) = \omega(x_N).$$

From the compatibility between order and topology we conclude that

$$w_1 \leqq w_2 \leqq w_3.$$

The nonordering principle implies $w_1 = w_3$; hence $w_1 = w_2$, a contradiction.

LEMMA 3.2. *Let $x_0 \in X$ be such that it can be approximated both from below and from above. Furthermore, assume that x_0 is contained in a neighborhood U such that $\sigma_y = +\infty$ for $y \in U$ and*

$$\Phi(U \times [t_0, \infty))$$

has compact closure in X for some $t_0 > 0$.

(a) *Let alternative* (a) *of the sequential limit set trichotomy hold both below and above x_0. Then x_0 is a stable point.*

(b) *Let alternative* (c) *of the sequential limit set trichotomy hold both below and above x_0. Then x_0 is an asymptotically stable point.*

Proof. We first consider (a). So we have sequences

$$x_n < x_{n+1} < x_0 < y_{n+1} < y_n, \qquad x_n \to x_0, \quad y_n \to x_0, \quad n \to \infty,$$
$$\omega(x_n) < \omega(x_{n+1}) < u_0 = \omega(x_0) < \omega(y_{n+1}) < \omega(y_n),$$

where

$$\lim_{n\to\infty} \operatorname{dist}(\omega(x_n), u_0) = \lim_{n\to\infty} \operatorname{dist}(\omega(y_n), u_0) = 0.$$

By Corollary 3.2 of [27] we have neighborhoods U_n of x_0 in U and times $t_n > 0$ such that

$$\Phi_t(x_n) \leqq \Phi_t(U_n) \leqq \Phi_t(y_n), \qquad t \geqq t_n.$$

If x_0 is not a stable point, we find $z_j \to x_0$, $s_j \to \infty$, such that, for all j,

$$d(\Phi_{s_j}(z_j), \Phi_{s_j}(x_0)) > \varepsilon > 0.$$

After choosing appropriate subsequences we can assume that $s_j > t_j$, $z_j \in U_j$, hence

$$\Phi_{s_j}(x_j) \leqq \Phi_{s_j}(z_j) \leqq \Phi_{s_j}(y_j).$$

By our compactness assumption, we can assume, after choosing subsequences, that

$$\Phi_{s_j}(x_j) \to w_1, \quad \Phi_{s_j}(z_j) \to w_2, \quad \Phi_{s_j}(y_j) \to w_3, \quad j \to \infty.$$

In particular,

$$w_1 \leqq w_2 \leqq w_3, \qquad d(w_2, u_0) \geqq \varepsilon.$$

The monotonicity of the flow implies that, for each j,

$$\omega(x_j) < w_1 \leqq w_2 \leqq w_3 < \omega(y_j).$$

This implies $w_1 = u_0 = w_2 = w_3$, a contradiction.

We consider the situation in (b). So we have the sequences

$$x_n < x_{n+1} < x_0 < y_{n+1} < y_n, \qquad x_n \to x_0, \quad y_n \to x_0, \quad n \to \infty,$$

where

$$\omega(x_n) = \omega(x_0) = \omega(y_n).$$

Using that the semiflow is strongly order preserving we find a neighborhood V of x_0 in U and a time $T > 0$ such that

$$\Phi_t(x_1) \leqq \Phi_t(V) \leqq \Phi_t(y_1), \qquad t > T.$$

Arguing as in the proof of Lemma 3.1 case (c), we find: For any $\varepsilon > 0$ there is some $t_\varepsilon > 0$ such that

$$d(\Phi_t(v), \Phi_t(x_0)) < \varepsilon, \qquad t > t_\varepsilon, \quad v \in V.$$

This completes our proof.

We can now state the main result of this section.

THEOREM 3.3. *Let X_0 be an open subset of X such that every point of X_0 can be approximated both from above and from below in X. Assume that either*

(a) *X is normally ordered, every orbit starting in X_0 has compact closure in X and, for any convergent monotone sequence x_n in X_0, $\bigcup_{n \geq 1} \omega(x_n)$ has compact closure in X, or*

(b) *Each point of X_0 belongs to a neighborhood U such that $U \times [0, \infty) \subset \operatorname{Dom} \Phi$ and*

$$\Phi(U \times [t_0, \infty)) \text{ has compact closure in } X \text{ for some } t_0 > 0.$$

Then $A \subset S \subset C$ and

$$X_0 \cap (A \cup \operatorname{Int} S)$$

is dense in X_0. In particular, X_0 contains a dense open subset of stable convergent points.

Proof. We consider only the case where (b) holds, since the other case follows simply as in Theorem 3.9 of [27]. It is sufficient to show that

$$U \cap (A \cup \operatorname{Int}(S \cap C)) \neq \varnothing$$

for any open subset U of X_0. Assume that $U \cap A = \varnothing$. Let x_0 be an arbitrary element in U. By assumption x_0 can be approximated from below and above. If the alternatives (b) or (c) of the limit set trichotomy hold below x_0 or above x_0, $U \cap A \neq \varnothing$, because x_0 is then approximated by asymptotically stable points. See Lemma 3.1. Hence, alternative (a) has to hold below x_0 and above x_0. Therefore, x_0 is a stable convergent point by Lemma 3.2(a). Since x_0 was an arbitrary point in U we have $U \subset S \cap C$. As U is open we have $U \subset \operatorname{Int}(S \cap C)$. This implies the assertion. Note that $A \subset S \subset C$ follows as in Proposition 2.5.

COROLLARY 3.4. *Let X contain a dense open set X_0 satisfying one of the assumptions* (a) *or* (b) *of the previous theorem. Then $A \cup \operatorname{Int} S$ is open and dense in X.*

Theorem 2.4 and Corollary 3.4 imply that, from a topological point of view, convergence to equilibrium occurs for a "large" set of initial data for strongly order-preserving semiflows enjoying the properties assumed to hold in this paper. Another notion of a "large" set, in finite dimensions, is a set whose complement has Lebesgue measure zero. We can establish that the set of nonconvergent points has measure zero in finite dimensions by showing that the set of such points on any totally ordered arc is at most countable and then applying Fubini's theorem. Thus, a natural extension of this notion of the scarcity of nonconvergent points, to infinite dimensions, is that any totally ordered arc contains at most countably many nonconvergent points. Such a result for quasi-convergent points was obtained by Hirsch [7, Thm. 7.3]. If we restrict ourselves to totally ordered arcs as opposed to sets, this result does not require the separability of the space X. Below we extend this result to convergent points as well as obtain a stronger result with the additional compactness hypothesis required for the sequential limit set trichotomy.

Recall that a totally ordered arc is the continuous image of a nontrivial interval $I \subseteq \mathbf{R}$ under a map $\psi: I \to X$ such that $\psi(s) < \psi(t)$ whenever $s, t \in I$ and $s < t$.

THEOREM 3.5. *Let J be a totally ordered arc of points belonging to orbits with compact closure in X. Then $J \setminus C$ is at most countable. If, furthermore, for every convergent sequence $x_n \in J$, $\overline{\bigcup_{n \geq 1} \omega(x_n)}$ is compact in X, then $J \setminus C$ consists of isolated points of J.*

Proof. Let $W = \overline{\Phi([0, \infty) \times J)}$. Then the continuity of Φ implies that W is a separable metric space which is positively invariant under Φ. Hence, we may as well assume from the beginning that X is separable since we can always pass to the space W.

We next establish that if $x \in J$ and

$$\inf \{\operatorname{dist}(\omega(x), \omega(y)): y \in J, y \neq x\} = 0,$$

then $x \in C$. Choose a sequence $x_n \in J$, $x_n \neq x$, such that $\operatorname{dist}(\omega(x_n), \omega(x)) \to 0$, $n \to \infty$. We may assume that $x_n < x$ for all n. Taking a subsequence, we can conclude from Proposition 2.2 that either $\omega(x_n) = \omega(x)$ holds for some n or $\omega(x_n) < \omega(x)$ holds for all n. In the first case x is convergent. In the second case we choose $y_n \in \omega(x_n)$, $z_n \in \omega(x)$, such that $d(y_n, z_n) \to 0$, $n \to \infty$. As $\omega(x)$ is compact we can assume, possibly by passing to a subsequence, that $y_n, z_n \to z \in \omega(x)$. As $y_n \leqq \omega(x)$, we conclude that $z \leqq \omega(x)$ and $z \in \omega(x)$ implies that $\omega(x) = \{z\}$ by the nonordering principle [27, Prop. 22]. Again, x is a convergent point, establishing our claim.

Now, if $x \in J \setminus C$, we find, by the claim of the previous paragraph, that there exists a neighborhood U_x of $\omega(x)$ such that $U_x \cap \omega(y) = \varnothing$ for every $y \in J$, $y \neq x$. By the axiom of choice we have a mapping $x \mapsto u_x \in \omega(x) \subset U_x$ from $J \setminus C$ into X which is injective. As X is separable, it has a countable base and we may obtain a neighborhood $V_x \subset U_x$ for each $x \in J \setminus C$, where V_x belongs to the countable base. A second application of the axiom of choice gives an injective mapping from $J - C$ into the countable base by $x \mapsto V_x$. This establishes the countability of $J - C$.

Assume the stronger assumption holds that for each convergent sequence $x_n \in J$, $\overline{\bigcup_{n \geq 1} \omega(x_n)}$ is compact in X. If x were not an isolated point of $J - C$ relative to J, then we could approximate x by a monotone sequence of points of $J - C$. But, as $x \in J - C$, this violates the sequential limit set trichotomy.

4. Applications. In this section we apply our theory to an abstract parabolic evolution equation and a concrete parabolic initial/boundary problem as well as to a system of quasi-monotone functional differential equations.

4.1. An abstract parabolic evolution equation. We consider the Cauchy problem

$$\text{(SPE)} \qquad \frac{du}{dt} + Au = f(u)$$

and compare the results provided by our theory with those obtained by Poláčik [20].

We assume that A is a sectorial operator on a Banach space X^0 with compact resolvent $(\lambda I - A)^{-1}$ and $\operatorname{Re}\sigma(A) > 0$, where $\sigma(A)$ denotes the spectrum of A.

For $\nu > 0$, X^ν denotes the domain of the fractional power A^ν with the corresponding graph norm which makes it a Banach space.

We fix some $0 \leqq \alpha < 1$. X^α is going to be our basic state space.

We require the following properties for f:

(f1) $f: X^\alpha \to X$ is C^1.

(f2) f maps bounded sets in X^α onto bounded sets in X.

By Theorems 3.3.3 and 3.3.4 of [3] the solutions u of (SPE) induce a local semiflow Φ on X^α by $\Phi_t(u(0)) = u(t)$. The maximal interval of existence of $\Phi_t(x)$ is denoted by σ_x.

It follows from Corollary 3.4.6 of [3] that Φ is continuously differentiable as a mapping from $X^\alpha \times (0, \infty)$ into X^α on its domain of existence. (SPE) then implies that $\Phi_t(x)$ is a continuous function of t from $(0, \sigma_x)$ into X^1.

As A has a compact resolvent and we have assumed (f2), Φ_t, for $t > 0$, maps bounded sets in X^α into bounded sets in X^β, $\alpha < \beta < 1$, and, hence, into compact sets in X^α. See the proof of Theorem 3.3.6 and Theorem 1.4.8 of [3]. In particular, sets $\Phi(B \times [\varepsilon, \infty))$ with $B \subset X^\alpha$, which are defined and bounded in X^α for $\varepsilon = 0$, are relatively compact in X^α for $\varepsilon > 0$.

We assume that X^0 is an ordered Banach space under a partial ordering $\leqq$. Our essential requirement is that there is some $\beta \in [\alpha, 1)$ such that X^β is strongly ordered by the induced ordering $\leqq$, i.e., the positive cone X^β_+ has nonempty open interior in X^β and that

$$\text{(P)} \qquad \Phi_t(y) - \Phi_t(x) \in \operatorname{Int} X^\beta_+, \qquad t > 0, \quad x < y,$$

$x, y \in X^\alpha$, t in the common interval of existence. (P) is formally weaker than Poláčik's assumption (M4) because it only involves solutions and not subsolutions to (SPE). (P) implies, in particular, assumptions (J) and (M) in § 3.

Φ_t restricts to a continuously differentiable (in the state variable) local semiflow on X^β since f restricts to a C^1 map on X^β. Hence (I) holds.

We verify that (S) of Remark 3, § 2 holds. For $e \in E$, we have already noted that Φ_t is continuously differentiable in an X^β neighborhood of e. From Theorem 3.4.4 of [3], it follows that $\Phi'_t(e)$ defines a strongly continuous semiflow on X^β generated by $A(e) = -A + f'(e)$. $A(e)$ is sectorial and $(\mu I - A(e))^{-1}$ maps X^0 continuously into X^1 and hence restricts to a compact operator on X^β. Thus $A(e)$ has compact resolvent. For each $e \in E$ and $\varepsilon > 0$ there exists a neighborhood U' of e such that $U' \times [0, \varepsilon] \subset \operatorname{Dom}\Phi$. Using again Theorem 3.4.4 of [3], standard Gronwall estimates, and continuity of Φ_t on X^β, we can verify that there is a neighborhood U containing e in U' such that $\Phi'(U \times [0, \varepsilon])$ is a bounded set in $L(X^\beta)$.

The remaining hypotheses in (S) concerning the spectrum of $A(e)$ must be added to our hypothesis:

(SP) For each $e \in E$, the spectral bound $s(e) = \sup\{\operatorname{Re}\lambda : \lambda \in \sigma(A(e))\}$ has the property that if $s(e) \geqq 0$, then $N(s(e)I - A(e))$ is spanned by an element $z(e)$ belonging to $\operatorname{Int} X^\beta$.

We note that (SP) follows immediately from (M3) of Poláčik (see [20, beginning of § 3]).

In the following results we assume that (f1), (f2), (P), and (SP) hold. Theorem 4.1 follows from the first assertion of Theorem 3.5 and Theorem 4.2 from Theorem 3.3 with hypothesis (b).

THEOREM 4.1. *Let J be a totally ordered arc of points in X^α, the orbits of which are defined for all $t \geqq 0$, and are bounded. Then the set of nonconvergent points on J is at most countable.*

THEOREM 4.2. *Let U be an open subset of X^α with the following property. For any $u \in U$ there is a neighborhood V of u in U such that $V \times [0, \infty) \subset \operatorname{Dom} \Phi$ and*

$$\Phi(V \times [0, \infty))$$

is bounded in X^α. Then U contains a dense open subset of stable convergent points.

Theorems 4.1 and 4.2 express that the set of nonconvergent points is small. They should be compared to Poláčik's Corollary 5.3 in [20]. We obtain a stronger conclusion in Theorem 4.2 than [20, Cor. 5.3], namely, that U contains a dense open set of stable convergent points rather than a residual set of convergent points, assuming more compactness. We can dispense with the technical conditions (SM1) and (SM2) in [20] and, more importantly, need $f \in C^1$ rather than $f \in C^2$ as in [20].

The results for abstract parabolic evolution equations can be applied to scalar parabolic initial boundary value problems

$$\begin{aligned} &\partial_t u - \mathcal{A}u = g(x, u, \nabla u), \qquad x \in \Omega, \\ &Bu = 0, \qquad x \in \partial\Omega, \\ &u(0, x) = u_0(x), \qquad x \in \Omega, \end{aligned}$$

where Ω is a bounded domain in $\mathbf{R}^N$ with sufficiently smooth boundary $\partial\Omega$ and $\mathcal{A}$ is a uniformly elliptic second-order operator and B a suitable boundary operator, in essentially the same way as done by Poláčik. Our hypotheses (P) and (SP) can be verified by application of the strong maximum principle for parabolic equations. We need $g \in C^1$ instead of C^2 and can dispense with Poláčik's assumption (6.8) in the case of Dirichlet boundary conditions.

4.2. Quasi-monotone functional differential equations. Consider the functional differential equation

$$\text{(FDE)} \qquad x'(t) = f(x_t)$$

where $f: C_r \to \mathbf{R}^n$ is C^1, $r = (r_1, r_2, \cdots, r_n) > 0$, $C_r = \prod_{i=1}^n C([-r_i, 0], \mathbf{R})$. If $x^i(t)$ is defined and continuous on $[-r_i, \sigma)$, $\sigma > 0$, $1 \leqq i \leqq n$, and $0 \leqq t < \sigma$, then $x_t \in C_r$ is given by $x_t = (x_t^1, x_t^2, \cdots, x_t^n)$ where $x_t^i(\theta) = x_i(t + \theta)$, $-r_i \leqq \theta \leqq 0$. Given $\phi \in C_r$, (FDE) has a unique solution $x(t; \phi)$ on a maximal interval of existence $[0, \sigma_\phi)$ satisfying $x_0^i = \phi^i$. The system (FDE) generates a (local) semiflow Φ on C_r given by

$$\Phi_t(\phi) = x_t(\phi), \qquad 0 \leqq t < \sigma_\phi.$$

Let $C_r^+ = \prod_{i=1}^n C([-r_i, 0], \mathbf{R}_+)$ be the cone of nonnegative functions in C_r and note that $\operatorname{Int} C_r^+$ consists of the functions each of whose components is positive on its domain.

Let U be an open subset of C_r which is positively invariant for (FDE); that is, $\Phi_t(\phi) \in U$ for $0 \leqq t < \sigma_\phi$, for all $\phi \in U$. Following [24], we introduce the following hypotheses:

(K) For all $\psi \in U$ and $\phi \in C_r^+$ with $\phi_i(0) = 0$,

$$f_i'(\psi)(\phi) \geqq 0.$$

The derivative $f_i'(\psi)$, of f_i, appearing in (K) can be represented as

$$f_i'(\psi)(\phi) = \sum_{j=1}^{n} \int_{-r_j}^{0} \phi_j(\theta)\, d\eta_{ij}(\theta)$$

where $\eta_{ij} = \eta_{ij}(\theta; \psi)$: $\mathbf{R} \to \mathbf{R}$ satisfy

$$\eta_{ij}(\theta) = \eta_{ij}(0), \qquad \theta \geqq 0,$$

$$\eta_{ij}(\theta) = 0, \qquad \theta \leqq -r_j,$$

$\eta_{ij} \in BV[-r_j, 0]$ and is continuous from the left on $(-r_j, 0)$.

As noted in [24], (K) is equivalent to assuming that η_{ij} is nondecreasing on $[-r_j, 0]$, $i \neq j$, and η_{ii} is nondecreasing on $[-r_i, 0)$.

In addition, we introduce the following hypotheses.

(R) For each $\psi \in U$, the $n \times n$ matrix

$$(\eta_{ij}(0; \psi))$$

is irreducible.

(S) For every j for which $r_j > 0$, there exists i such that for all $\psi \in U$ and all small $\varepsilon > 0$

$$\eta_{ij}(-r_j + \varepsilon; \psi) > 0.$$

We refer the reader to [24] for additional motivation for the hypotheses above. Recall that a matrix is irreducible if it does not leave invariant any nontrivial coordinate subspaces. If f satisfies (K), (R), and (S) on U, then f is cooperative and irreducible in U.

In order to insure that (FDE) generates a (global) semiflow with appropriate compactness hypotheses we will also assume the following:

(T) f maps bounded subsets of U to bounded sets in $\mathbf{R}^n$. For each $\psi \in U$ the orbit $\mathcal{O}^+(\psi) = \{x_t(\psi): t \geqq 0\}$ exists and is bounded. Moreover, assume that there is a closed bounded subset $B \subset U$ such that $\omega(\psi) \subset B$ for all $\psi \in U$, where $\omega(\psi)$ is the omega limit set of $\mathcal{O}^+(\psi)$.

Hypothesis (T) is easily seen to imply that the compactness assumptions of Theorem 3.3(a) hold. Indeed, if $\{\psi_n\}$ is a convergent sequence, then $\bigcup_{n \geqq 1} \omega(\psi_n)$ is an invariant set contained in the closed and bounded subset B. Since f is bounded on bounded sets, Φ_t is completely continuous for large t [24] so it follows that $\bigcup_{n \geqq 1} \omega(\psi_n)$ has compact closure in B.

In the following result, we summarize several important consequences of our assumptions.

PROPOSITION 4.3. *Let f satisfy* (K), (R), (S), *and* (T). *Then Φ_t is an eventually strongly monotone C^1-semiflow on U; that is, if $\psi, \phi \in U$, and $\phi < \psi$, then*

$$x_t(\phi) \ll x_t(\psi) \quad \textit{for } t \geqq (n+1) \max r_i.$$

If $e \in U$ is an equilibrium point of (FDE), *then $\{T_e(t) \equiv \Phi_t'(e)\}_{t \geqq 0}$ is a strongly continuous semigroup and $y_t = \Phi_t'(e)\phi$ satisfies*

$$y'(t) = L(y_t), \quad L(\psi) = f'(e)\psi, \quad y_0 = \phi.$$

Moreover, $T_e(t)$ is compact and strongly positive for $t \geqq (n+1) \max r_i$.

Proof. The proof follows immediately from Theorem 3.1 of [24].

THEOREM 4.4. *Let f satisfy* (K), (R), (S), *and* (T). *Then U contains an open and dense set of stable convergent points.*

Theorem 4.4 follows from Theorem 3.3(a) using the normality of C_r. Concrete examples where the above result applies appear in [24], [25], and [27].

Acknowledgment. The authors thank an anonymous reviewer who stimulated us to significantly improve our original manuscript.

REFERENCES

[1] H. AMANN, *Dual semigroups and second order linear elliptic boundary value problems*, Israel J. Math., 45 (1983), pp. 225-254.

[2] PH. CLÉMENT, H. J. A. M. HEIJMANS, S. ANGENENT, C. J. VAN DUIJN, AND B. DE PAGTER, *One-Parameter Semigroups*, CWI Monograph 5, North-Holland, Amsterdam, the Netherlands, 1987.

[3] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1981.

[4] M. W. HIRSCH, *Systems of differential equations which are competitive or cooperative* I: *Limit Sets*, SIAM J. Math. Anal., 13 (1982), pp. 167-179.

[5] ———, *Systems of differential equations which are competitive or cooperative* II: *Convergence almost everywhere*, SIAM J. Math. Anal., 16 (1985), pp. 432-439.

[6] ———, *The dynamical systems approach to differential equations*, Bull. Amer. Math. Soc., 11 (1984), pp. 1-64.

[7] ———, *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math., 383 (1988), pp. 1-53.

[8] T. KATO, *Perturbation Theory for Linear Operators*, ed. Springer-Verlag, Berlin, New York, 1976.

[9] R. H. MARTIN, JR., *Nonlinear Operators and Differential Equations in Banach Spaces*, John Wiley, New York, 1976.

[10] R. H. MARTIN, JR. AND H. L. SMITH, *Abstract functional differential equations and reaction-diffusion systems*, Trans. Amer. Math. Soc., 321 (1990), pp. 1-44.

[11] ———, *Reaction diffusion systems with time delays: monotonicity, invariance, comparison and convergence*, J. Reine Angew. Math., to appear.

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

[13] ———, *Existence of nontrivial unstable sets for equilibriums of strongly order-preserving systems*, J. Fac. Sci. Univ. Tokyo, Sect. 1A Math., 30 (1983), pp. 645-673.

[14] H. MATANO AND M. MIMURA, *Pattern formation in competition diffusion systems in nonconvex domains*, Publ. Res. Inst. Math. Sci., 19 (1983), pp. 645-673.

[15] H. MATANO, *Strongly order preserving local semi-dynamical systems—theory and applications*, in Semigroups, Theory and Applications, Vol. I, H. Brezis, M. G. Crandall and F. Kappel, eds., Res. Notes in Math., 141, Longman Scientific & Technical, London, 1986, pp. 178-185.

[16] ———, *Strong comparison principle in nonlinear parabolic equations*, in Nonlinear Parabolic Equations: Qualitative Properties of Solutions, L. Boccardo and A. Tesei, eds., Pitman Res. Notes in Math., 149, Longman Scientific & Technical, London, 1987, pp. 148-155.

[17] H. MATANO, X.-Y. CHEN AND L. VÉRON, *Anisotropic singularities of solutions of nonlinear elliptic equations in* $\mathbf{R}^2$, J. Funct. Anal., 83 (1989), pp. 50-97.

[18] R. D. NUSSBAUM, *Hilbert's projective metric and iterated nonlinear maps*, Mem. Amer. Math. Soc., 75 (1988).

[19] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.

[20] P. POLÁČIK, *Convergence in smooth strongly monotone flows defined by semilinear parabolic equations*, J. Differential Equations, 79 (1989), pp. 89-110.

[21] ———, *Domains of attraction of equilibria and monotonicity properties of convergent trajectories in parabolic systems admitting strong comparison principles*, J. Reine Angew. Math., 400 (1989), pp. 32-56.

[22] H. H. SCHAEFER, *Topological Vector Spaces*, Springer-Verlag, Berlin, New York, 1971.

[23] J. F. SELGRADE, *A Hopf bifurcation in single-loop positive-feedback systems*, Quart. J. Appl. Math., 40 (1982), pp. 347-351.

[24] H. L. SMITH, *Systems of ordinary differential equations which generate an order-preserving flow. A survey of results*, SIAM Rev., 30 (1988), pp. 87–111.

[25] ———, *Monotone semiflows generated by functional differential equations*, J. Differential Equations, 66 (1987), pp. 420–442.

[26] H. L. SMITH AND H. R. THIEME, *Monotone semiflows in scalar nonquasimonotone functional differential equations*, J. Math. Anal. Appl., 150 (1990), pp. 289–306.

[27] ———, *Quasi-convergence and stability for strongly order-preserving semiflows*, SIAM J. Math. Anal., 21 (1990), pp. 673–692.

[28] ———, *Strongly order preserving semiflows generated by functional differential equations*, J. Differential Equations, to appear.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1102-1109, July 1991

015

PERIODIC TRIDIAGONAL COMPETITIVE AND COOPERATIVE SYSTEMS OF DIFFERENTIAL EQUATIONS*

HAL L. SMITH†

Abstract. J. Smillie [*J. Differential Equations*, 64 (1986), pp. 165-194] proves that a bounded solution of a smooth autonomous tridiagonal competitive or cooperative system of differential equations converges to an equilibrium. This result is extended to ω-periodic systems, showing that every bounded solution is asymptotic to an ω-periodic solution, at the same time relaxing Smillie's smoothness requirements. In addition, it is shown that the Floquet multipliers of an ω-periodic solution are positive, simple, and have distinct absolute values.

Key words. periodic competitive system, periodic cooperative system, monotone dynamical systems, integer-valued Lyapunov function

AMS(MOS) subject classifications. 34C25, 34C15

Introduction. We consider the nonlinear periodic tridiagonal system

$$\begin{aligned} y_1' &= f_1(t, y_1, y_2), \\ y_j' &= f_j(t, y_{j-1}, y_j, y_{j+1}), \qquad 2 \leqq j \leqq n-1, \\ y_n' &= f_n(t, y_{n-1}, y_n), \end{aligned} \tag{0.1}$$

where $f=(f_1, f_2, \cdots, f_n)$ is defined on $\mathbf{R}\times\mathcal{O}$, $\mathcal{O}$ a nonempty open subset of $\mathbf{R}^n$. We assume that the f_i and their partial derivatives with respect to the y_j are continuous in $\mathbf{R}\times\mathcal{O}$ and that there exist $\delta_i \in \{-1, +1\}$, $1 \leqq i \leqq n-1$, such that

$$\delta_i \frac{\partial f_i}{\partial y_{i+1}} > 0, \quad \delta_i \frac{\partial f_{i+1}}{\partial y_i} > 0, \qquad 1 \leqq i \leqq n-1 \tag{0.2}$$

holds for all values of the arguments $(t, y) \in \mathbf{R}\times\mathcal{O}$. It will be assumed that there exists $\omega > 0$ such that

$$f(t+\omega, y) = f(t, y), \qquad (t, y) \in \mathbf{R}\times\mathcal{O}. \tag{0.3}$$

Finally, concerning the open set $\mathcal{O}$, we assume that each of the coordinate projections $\mathcal{O}_1 \subseteq \mathbf{R}^2$ of $\mathcal{O}$ onto the (y_1, y_2) plane, $\mathcal{O}_n \subseteq \mathbf{R}^2$ of $\mathcal{O}$ onto the (y_{n-1}, y_n) plane, and $\mathcal{O}_j \subseteq \mathbf{R}^3$ of $\mathcal{O}$ onto the (y_{j-1}, y_j, y_{j+1}) space, $2 \leqq j \leqq n-1$, are nonempty convex subsets.

Autonomous systems of the form (0.1) satisfying (0.2) where f_i was assumed to be $n-1$ times differentiable with respect to the y_j were considered by Smillie [18]. With these hypotheses, Smillie shows that every bounded orbit converges to an equilibrium. We extend this result by showing that every bounded orbit of (0.1) is asymptotic to an ω-periodic solution of (0.1), at the same time relaxing Smillie's smoothness assumptions. In the autonomous case, we obtain his result with only the assumption that $f \in C^1$.

The requirement (0.2) implies that the Jacobian matrix, $\partial f/\partial y$, corresponding to (0.1), is tridiagonal and sign symmetric in the sense that $\partial f_i/\partial y_{i+1}$ and $\partial f_{i+1}/\partial y_i$ have the same sign, namely, δ_i. In case $\delta_i = -1$ for all i, (0.1) is called competitive and if $\delta_i = +1$ for all i then (0.1) is called cooperative. The change of variables $\bar{y}_i = \mu_i y_i$,

* Received by the editors February 20, 1990; accepted for publication (in revised form) July 16, 1990.
† Department of Mathematics, Arizona State University, Tempe, Arizona 85287-1804. This research was supported in part by National Science Foundation grant DMS8722279.

$\mu_i \in \{-1, +1\}$, $1 \leqq i \leqq n$, transforms (0.1) to a new system of the same type, satisfying (0.2), where now $\bar{\delta}_i = \mu_i \mu_{i+1} \delta_i$. It is clear that we may choose the μ_i, $1 \leqq i \leqq n$, so that $\mu_i \mu_{i+1} = \delta_i$ for $1 \leqq i \leqq n-1$. Indeed, set $\mu_1 = 1$, $\mu_i = \delta_{i-1} \mu_{i-1}$, $2 \leqq i \leqq n$. After such a change of variables we may always assume that (0.1) is cooperative. Note also that the time reversed system is again a system of type (0.1) satisfying (0.2) and (0.3) with the signs of δ_i reversed.

System (0.1) with the assumption (0.2) can be viewed as a monotone dynamical system in the sense that the Poincaré (period) map preserves a partial ordering on $\mathbf{R}^n$. There is now an extensive literature on monotone dynamical systems, beginning with the path-breaking work of M. W. Hirsch [9]–[12] for monotone semiflows. The results of Hirsch and later improvements by Matano [15], Smith and Thieme [22], [23], and Poláčik [17] established that most orbits of a strongly order-preserving semiflow converge to the set of equilibria. These results, of course, apply only to autonomous systems. Smillie's result stands out in that all bounded orbits converge.

The theory of discrete-time monotone dynamics is significantly more complicated than its continuous-time counterpart. There has been much recent work following the early studies of Alikakos and Hess [1] and Alikakos, Hess, and Matano [2]. Examples of stable k-cycles, $k \geqq 2$, for strongly order-preserving discrete-time dynamical systems have been given by Takáč [26], [27] and by Dancer and Hess [28]. In the example given in [27], the dynamics is generated by the Poincaré map for a cooperative and irreducible time-periodic four-dimensional vector field. These examples show that the limit set dichotomy of Hirsch [9], [10], [12] for strongly monotone semiflows does not carry over to strongly monotone discrete-time dynamical systems and that we cannot expect to prove that "most orbits" of a strongly monotone discrete-time dynamical system converge to the set of equilibria (fixed points). Convergence for strongly monotone discrete-time dynamical systems has been established by Takáč [25] in the case that all equilibria are Lyapunov stable and additional technical assumptions. This result improved the previously mentioned work of Alikakos and Hess [1] and Alikakos, Hess, and Matano [2].

In a certain sense, our results here are natural generalizations of the results of de Mottoni and Schiaffino [4] and Hale and Somolinos [7], who proved that all solutions of two-dimensional ω-periodic competitive or cooperative systems are asymptotic to ω-periodic solutions. See also [19] and [20] for extensions of this work.

The proof of our main result, namely, that every bounded solution $y(t)$ of (0.1) is asymptotic to an ω-periodic solution, uses the main technique introduced by Smillie and which is also used in a related form by Smith with Mallet-Paret in [13], by Fusco and Oliva in [6], and by Smith in [21]. An integer-valued function is introduced which has the property that it is a Lyapunov function for a certain class of linear systems including the variational equation corresponding to (0.1). In other words, this function is defined for all but an at most finite set of points t along a nontrivial solution of the linear system, is locally constant near points where it is defined, and strictly decreases as t increases through points where it is not defined. The nature of the domain of this Lyapunov function, together with the fact that the solution of the linear system belongs to this domain for almost all values of t, places strong restrictions on the signs of components of the solution and leads to the proof of our main result, Theorem 2.2. The Lyapunov function also places strong restrictions on the Floquet multipliers of an ω-periodic solution of (0.1). In Theorem 1.3 we show that all multipliers are distinct and positive.

The idea of using integer-valued Lyapunov functions in dynamical systems seems to go back to the work of Nickel [16] and later to that of Matano [14], where the

so-called lap number is introduced as a count of the number of sign changes of a solution $u(t, x)$ of a scalar reaction diffusion equation on an interval $0 \leqq x \leqq L$. As t increases this number cannot increase. The lap number has been used by Henry [8] and Angenent [3] to establish the Morse-Smale property for scalar reaction diffusion equations. See also Fiedler and Mallet-Paret [5].

The paper is organized as follows. In § 1 we introduce the Lyapunov function and use it to study certain linear systems which include the variational equation along a solution of (0.1). Theorem 1.3 concerning Floquet multipliers is also proved. In § 2 we prove our main result, Theorem 2.2, by applying the results of § 1 to the difference of two solutions of (0.1), which satisfies an appropriate linear system.

1. Linear systems. Consider the linear system

$$\begin{aligned} x_1' &= a_{11}x_1 + a_{12}x_2, \\ x_j' &= a_{jj-1}x_{j-1} + a_{jj}x_j + a_{jj+1}x_{j+1}, \qquad 2 \leqq j \leqq n-1, \\ x_n' &= a_{nn-1}x_{n-1} + a_{nn}x_n, \end{aligned} \tag{1.1}$$

where the a_{ij} are continuous functions defined on a nontrivial interval I and

$$\begin{aligned} a_{jj+1}(t) &> 0 \quad \text{on } I, \quad 1 \leqq j \leqq n-1, \\ a_{jj-1}(t) &> 0 \quad \text{on } I, \quad 2 \leqq j \leqq n. \end{aligned} \tag{1.2}$$

Following [18], we define the continuous map

$$\sigma : \Lambda \to \{0, 1, 2, \cdots, n-1\}$$

on

$$\begin{aligned} \Lambda = \{v \in \mathbf{R}^n &: v_1 \neq 0,\ v_n \neq 0 \text{ and if } v_i = 0 \\ &\text{for some } i,\ 2 \leqq i \leqq n-1, \text{ then } v_{i-1}v_{i+1} < 0\} \end{aligned}$$

by

$$\sigma(v) = \#\{i : v_i v_{i+1} < 0\}.$$

Here, # denotes the cardinality of the set. Note that Λ is open and dense in $\mathbf{R}^n$ and Λ is the maximal domain on which σ is continuous.

A second integer-valued function was introduced in [13] and in [6]. In [21], Smith extends the applicability of this function to a class of linear systems including (1.1). This map is defined by

$$N : \mathcal{N} \to \{0, 1, 2, \cdots, n\}$$

on

$$\mathcal{N} = \{v \in \mathbf{R}^n : \text{if } v_i = 0 \text{ for some } i,\ 1 \leqq i \leqq n, \text{ then } v_{i-1}v_{i+1} < 0\},$$

where i is to be interpreted modulo n ($0 \sim n$, $n+1 \sim 0$), and

$$N(v) = \#\{i : v_i v_{i-1} < 0\}.$$

It is not hard to see that N is continuous on its domain $\mathcal{N}$, which is open, and that N takes only even values in $\{0, 1, 2, \cdots, n\}$.

Clearly, we have

$$\Lambda \subset \mathcal{N} \tag{1.3}$$

and

$$N(v)=\begin{cases}\sigma(v), & v_1v_n>0,\\ \sigma(v)+1, & v_1v_n<0\end{cases} \tag{1.4}$$

for all $v\in\Lambda$.

A special case of the main result in [21] is the following.

PROPOSITION 1.1. *If $x(t)$ is a nontrivial solution of* (1.1) *on I then*

(i) $x(t)\in\mathcal{N}$ *except possibly for isolated values of t.*

(ii) *If $x(s)\notin\mathcal{N}$ for some $s\in\operatorname{int} I$ then $N(x(s+))<N(x(s-))$.*

Assertion (i) implies that if $x(s)\notin\mathcal{N}$ then there exists $\varepsilon>0$ such that $x(t)\in\mathcal{N}$ for $0<|t-s|<\varepsilon$. The continuity of N on $\mathcal{N}$ implies that $N(x(t))$ is constant on $(s-\varepsilon, s)$ and on $(s, s+\varepsilon)$. $N(x(t))$ decreases by a positive multiple of 2 as t increases through s. The notation int I denotes the interior of the interval I and $N(x(s+))$ denotes the limit of $N(x(t))$ as t approaches s from the right and similarly for $N(x(s-))$.

The next result was proved by Smillie [18] in the case where a_{ij} are $n-2$ times differentiable.

PROPOSITION 1.2. *If $x(t)$ is a nontrivial solution of* (1.1) *on I then*

(i) $x(t)\in\Lambda$ *except possibly for isolated values of t.*

(ii) *If $x(s)\notin\Lambda$ for some $s\in\operatorname{int} I$ then $\sigma(x(s+))<\sigma(x(s-))$.*

Corresponding remarks to those following Proposition 1.1 can be made here as well. In particular, $\sigma(x(t))$ decreases (not necessarily by a multiple of 2) as t increases through a point $s\in\operatorname{int} I$ at which $x(s)\notin\Lambda$. Note that as $\sigma(x(t))$ can assume at most n values for $t\in I\cap x^{-1}(\Lambda)$, it follows that there can be at most $n-1$ values of $t\in I$ for which $x(t)\notin\Lambda$.

Proof of Proposition 1.2. We use Proposition 1.1 in the proof. Let $t_1\in\operatorname{int} I$ be such that $x(t_1)\notin\Lambda$. Suppose first that $x(t_1)\in\mathcal{N}$. Then $x_1(t_1)=0$ or $x_n(t_1)=0$ or both. But $x(t_1)\in\mathcal{N}$ implies that if $x_1(t_1)=0$ then $x_n(t_1)x_2(t_1)<0$ and if $x_n(t_1)=0$ then $x_{n-1}(t_1)x_1(t_1)<0$. We suppose $x_n(t_1)=0$, the other case being similar. Now $x_n'(t_1)=a_{nn-1}x_{n-1}(t_1)\neq 0$ so t_1 is an isolated point at which $x(t_1)\notin\Lambda$ and

$$\left.\frac{d}{dt}\right|_{t=t_1} x_1(t)x_n(t)=a_{nn-1}(t_1)x_{n-1}(t_1)x_1(t_1)<0.$$

By (1.4) and the fact that $N(t)=N(t_1)$ in a neighborhood of t_1, it is clear that

$$\sigma(x(t_1-))=\sigma(x(t_1+))+1.$$

Thus the conclusion of the proposition holds in this case.

Suppose $x(t_1)\notin\mathcal{N}$. Then by Proposition 1.1 there exists $\varepsilon>0$ such that $x(t)\in\mathcal{N}$ for $0<|t-t_1|<\varepsilon$ and $N(x(t_1-))-N(x(t_1+))$ is a positive multiple of 2. We claim that $x_1(t)x_n(t)$ can vanish at at most two points of $0<|t-t_1|<\varepsilon$. At a zero of x_1x_n in $0<|t-t_1|<\varepsilon$, $d/dt\, x_1x_n<0$ since $x\in\mathcal{N}$ at such a point. Thus x_1x_n can vanish at most at one point of $(t_1-\varepsilon, t_1)$ and at most at one point of $(t_1, t_1+\varepsilon)$, proving our claim. Hence we may find $\varepsilon_0\leqq\varepsilon$ such that $x(t)\in\Lambda$ for $0<|t-t_1|<\varepsilon_0$. As N decreases by a positive multiple of 2 as t increases through t_1, (1.4) implies that σ must decrease by at least 1. Our proof is complete. □

As Proposition 1.1 had important implications for Floquet theory for periodic linear monotone cyclic feedback systems in [13], Proposition 1.2 has similar implications for Floquet theory for (1.1) when the system is periodic. Hereafter in this section, we assume that $a_{ij}(t+\omega)=a_{ij}(t)$ holds for all t and all i, j, where $\omega>0$. Let $X(t)$ denote the fundamental matrix solution of (1.1) satisfying $X(0)=I$, where I is the identity matrix. Recall that the eigenvalues of $X(\omega)$ are the Floquet multipliers of

(1.1). The proof of the next result follows closely the proof of related results in [13] and so we merely sketch the proof, making use of results and notation in [13, § 2].

THEOREM 1.3. *The ω-periodic linear system* (1.1) *has n distinct positive Floquet multipliers $\alpha_1, \cdots, \alpha_n$, satisfying*

$$\alpha_1 > \alpha_2 > \cdots > \alpha_{n-1} > \alpha_n > 0. \tag{1.5}$$

If E_{α_i} are the corresponding one-dimensional eigenspaces of $X(\omega)$ then $E_{\alpha_i}\backslash\{0\} \subset \Lambda$ and

$$\sigma(E_{\alpha_i}\backslash\{0\}) = i-1, \qquad 1 \leqq i \leqq n. \tag{1.6}$$

Proof. It is easy to see that Lemmas 2.1 through 2.4 of [13] hold with σ replacing N. Let α be a multiplier of (1.1) and $G_\alpha = \ker (X(\omega) - \alpha I)^m$ be the generalized eigenspace corresponding to α (here m is chosen sufficiently large). By Lemma 2.2 in [13], for each $\sigma_0 > 0$ which is the modulus of some Floquet multiplier of (1.1), we have

$$\mathcal{G}_{\sigma_0}\backslash\{0\} \subseteq \Lambda$$

and σ is constant on $\mathcal{G}_{\sigma_0}\backslash\{0\}$ where

$$\mathcal{G}_{\sigma_0} = \operatorname{Re} \bigoplus_{|\alpha|=\sigma_0} G_\alpha.$$

If $\dim \mathcal{G}_{\sigma_0} \geqq 2$ then we could find linearly independent vectors $v, w \in \mathcal{G}_{\sigma_0}$ such that $\beta v_1 + w_1 = 0$ for a suitable scalar $\beta \neq 0$. But this contradicts that $\beta v + w \in \mathcal{G}_{\sigma_0}\backslash\{0\} \subset \Lambda$. Hence, $\dim \mathcal{G}_{\sigma_0} = 1$ and $\mathcal{G}_{\sigma_0} = G_{\alpha_0} = E_{\alpha_0}$ for some Floquet multiplier α_0. Moreover, α_0 is the only Floquet multiplier with modulus $|\alpha_0| = \sigma_0$. Thus α_0 is real.

Recall that for a Floquet multiplier α of (1.1) there corresponds a nontrivial solution $x(t)$ of (1.1) satisfying $x(\omega) = \alpha x(0)$. If α were negative then necessarily $x_1(t)(x_n(t))$ must vanish for some t, contradicting that $E_\alpha\backslash\{0\} \subseteq \Lambda$. Thus, every Floquet multiplier of (1.1) is positive.

Lemmas 2.4 and 2.5 in [13], where now $\dim \operatorname{span}_{\sigma_0 \in S} \mathcal{G}_{\sigma_0} \leqq 1$ replaces the estimate in Lemma 2.5, by the argument above, imply that σ takes different values on different eigenspaces E_α (more precisely on $E_\alpha\backslash\{0\}$). Lemma 2.3 in [13] then implies (1.6). □

2. Main results. In this section we establish our main result concerning the nonlinear periodic system

$$\begin{aligned} y_1' &= f_1(t, y_1, y_2), \\ y_j' &= f_j(t, y_{j-1}, y_j, y_{j+1}), \qquad 2 \leqq j \leqq n-1, \\ y_n' &= f_n(t, y_{n-1}, y_n), \end{aligned} \tag{2.1}$$

where $f = (f_1, f_2, \cdots, f_n)$ is defined on $\mathbf{R} \times \mathcal{O}$ where $\mathcal{O}$ is a nonempty open set in $\mathbf{R}^n$ with the property that each of the coordinate projections $\mathcal{O}_1 \subseteq \mathbf{R}^2$ of $\mathcal{O}$ onto the (y_1, y_2) plane, $\mathcal{O}_n \subseteq \mathbf{R}^2$ of $\mathcal{O}$ onto the (y_{n-1}, y_n) plane, and $\mathcal{O}_j \subseteq \mathbf{R}^3$ of $\mathcal{O}$ onto the (y_{j-1}, y_j, y_{j+1}) space, $2 \leqq j \leqq n-1$, are convex. We assume that the f_i and each of their partial derivatives with respect to the y_j exist and are continuous on $\mathbf{R} \times \mathcal{O}$ and

$$f(t+\omega, y) = f(t, y) \tag{2.2}$$

for all $(t, y) \in \mathbf{R} \times \mathcal{O}$ for some positive number ω (not necessarily the minimal period). In addition, we assume that f is cooperative:

$$\frac{\partial f_j}{\partial y_{j-1}} > 0, \qquad \frac{\partial f_j}{\partial y_{j+1}} > 0, \tag{2.3}$$

$2 \leqq j \leqq n-1$, and the first inequality holds for $j = n$ and the second holds for $j = 1$. In the introduction we showed that the more general assumption (0.2) can be transformed to (2.3) by a change of variables.

LEMMA 2.1. *Let $y(t)$ and $\bar{y}(t)$ be distinct solutions of* (2.1) *on an interval I. Then $y(t)-\bar{y}(t)\in\Lambda$ except possibly at finitely many values of $t\in I$. $\sigma(y(t)-\bar{y}(t))$ is locally constant and strictly decreases as t increases through a value s at which it is not defined.*

Proof. $x(t)=y(t)-\bar{y}(t)$ satisfies the linear system (1.1), where

$$a_{ij}(t)=\int_0^1 \frac{\partial f_i}{\partial y_j}(t, u_{i-1}(s,t), u_i(s,t), u_{i+1}(s,t))\,ds$$

with $u_i(s,t)=sy_i(t)+(1-s)\bar{y}_i(t)$, $j=i-1, i, i+1$. The result now follows from (2.3) and Proposition 1.2. □

THEOREM 2.2. *Let $\bar{y}(t)$ be a solution of* (2.1) *for $t\geqq 0$ which is not ω-periodic but which is bounded for $t\geqq 0$. Then $\bar{y}(t)$ is asymptotic to an ω-periodic solution of* (2.1).

Proof. Define the Poincaré map corresponding to (2.1) by $Py(0)=y(\omega)$ for those points $y(0)$ for which it is defined. The omega limit set of the orbit $\mathcal{O}^+(\bar{y}(0))=\{\bar{y}(n\omega)=P^n\bar{y}(0)\colon n=0,1,2,\cdots\}$ of $\bar{y}(0)$ under the map P, $\Omega(\bar{y}(0))$, is defined, compact, nonempty, invariant under P, and $\bar{y}(n\omega)\to\Omega(\bar{y}(0))$ as $n\to\infty$. The proof will be complete if we establish that $\Omega(\bar{y}(0))$ is a singleton.

By our hypotheses and (2.2), $\bar{y}(t)$ and $\bar{y}(t+\omega)$ are distinct solutions of (2.1) for $t\geqq 0$. Lemma 2.1 implies that $\bar{y}(t+\omega)-\bar{y}(t)\in\Lambda$ for all large t. This implies that $\bar{y}_1(t+\omega)-\bar{y}_1(t)(\bar{y}_n(t+\omega)-\bar{y}_n(t))$ cannot vanish for all large t and so must be of constant sign, say, $\bar{y}_1(t+\omega)-\bar{y}_1(t)>0$ for $t\geqq m\omega$ where m is some positive integer. Putting $t=k\omega$, $k\geqq m$, into this inequality, we find that

$$\bar{y}_1(k\omega)<\bar{y}_1((k+1)\omega), \qquad k\geqq m.$$

In particular, $\lim_{k\to\infty}\bar{y}_1(k\omega)=\bar{p}_1$ and $\lim_{k\to\infty}\bar{y}_n(k\omega)=\bar{p}_n$ exist.

If $p\in\Omega(\bar{y}(0))$ then it follows that $p_1=\bar{p}_1$ and $p_n=\bar{p}_n$. If $p\neq q$ and $p, q\in\Omega(\bar{y}(0))$, then the solutions $p(t)$ and $q(t)$ of (2.1) with $p(0)=p$ and $q(0)=q$ are defined for all $t\in\mathbf{R}$ and $p(l\omega)$, $q(k\omega)$ belong to $\Omega(\bar{y}(0))$ for all integers l and k. Since $p(t)$ and $q(t)$ are distinct solutions of (2.1), Lemma 2.1 implies that $p(t)-q(t)\in\Lambda$ for all large values of t. In particular, $p_1(k\omega)-q_1(k\omega)\neq 0$ for all large positive integers k. But this contradicts that $p_1(k\omega)=\bar{p}_1=q_1(k\omega)$, which was established above. Thus $\Omega(\bar{y}(0))$ must be a singleton set, establishing the theorem. □

According to Theorem 1.3, if $p(t)=p(t+\omega)$ is an ω-periodic solution of (2.1), its Floquet multipliers, determined by the periodic linear variational system (1.1) where

$$a_{ij}(t)=\frac{\partial f_i}{\partial y_j}(t, p_{i-1}(t), p_i(t), p_{i+1}(t))$$

for $j=i-1, i, i+1$, are positive and distinct.

Theorem 1.3 has an important additional implication if we assume that every ω-periodic solution $p(t)$ of (2.1) is nondegenerate, that is, that one is not a Floquet multiplier. Of course, this implies that $p(t)$ is hyperbolic by Theorem 1.3, that is, there are no Floquet multipliers of unit modulus. In this case, if there is a heteroclinic (homoclinic) solution $z(t)$, asymptotic to an ω-periodic solution $p(t)$ as $t\to+\infty$ and to an ω-periodic solution $q(t)$ as $t\to-\infty$, then $\dim W^s(q(0))<\dim W^s(p(0))$ where $W^s(q(0))$ denotes the stable manifold of the fixed point $q(0)$ for the Poincaré map P corresponding to (2.1). In particular, there can be no homoclinic solution $z(t)$ corresponding to an ω-periodic solution of (2.1) nor a set of heteroclinic solutions forming a cycle. See [29] for an application of this result. In order to see that $\dim W^s(q(0))<\dim W^s(p(0))$, assume $q(0)\neq p(0)$, a similar argument applies if $p(0)=q(0)$, and suppose $z(t)$ is a heteroclinic solution so that $z(n\omega)\to p(0)$ ($z(n\omega)\to q(0)$) as $n\to+\infty$ ($n\to-\infty$). Then $z(t)-q(t)\in\Lambda$ for $|t|$ large and $p(t)-q(t)\in\Lambda$ for all t by Lemma 2.1

and periodicity. It follows that $\sigma(z(t)-q(t))\to\sigma(p(0)-q(0))$ as $t\to+\infty$ and $\sigma(z(t)-q(t))\geqq\sigma(p(0)-q(0))$ by Lemma 2.1 and continuity of σ. Now, a subsequence of $(z(n\omega)-q(0))/|z(n\omega)-q(0)|$ approaches a unit vector in $E^{q(0)}_{\alpha_j}$, the eigenspace corresponding to some Floquet multiplier α_j of the periodic solution $q(t)$, where the multipliers are ordered as in Theorem 1.3. Thus, $\sigma(z(t)-q(t))=j-1$ for large negative t. Hence $j-1\geqq\sigma(p(0)-q(0))$ and note also that dim $W^u(q(0))\geqq j$ since $\alpha_j>1$. Arguing similarly with $p(t)-z(t)$, we obtain $\sigma(p(0)-q(0))\geqq k-1$ where a subsequence of $(p(0)-z(n\omega))/|p(0)-z(n\omega)|$ approaches a unit vector in some $E^{p(0)}_{\alpha_k}$, the eigenspace corresponding to a Floquet multiplier α_k of the periodic solution $p(t)$. Now, $\alpha_k<1$ so dim $W^s(p(0))\geqq n-k+1$. The inequality $j-1\geqq\sigma(p(0)-q(0))\geqq k-1$ implies that dim $W^s(q(0))=n-$dim $W^u(q(0))\leqq n-j\leqq n-k<n-k+1\leqq$ dim $W^s(p(0))$, completing the argument.

In the special case where (2.1) is autonomous, (2.2) holds for all ω, we obtain the result of Smillie [18]. Smillie required f to be $n-1$-times differentiable. Our result removes that restriction.

COROLLARY 2.3 (C^1 version of Smillie's theorem [18]). *Let f be independent of t in* (2.1). *Then every solution $y(t)$ of* (2.1), *bounded on $t\geqq 0$, converges to an equilibrium of* (2.1).

REFERENCES

[1] N. D. ALIKAKOS AND P. HESS, *On stabilization of discrete monotone dynamical systems*, Israel J. Math., 59 (1987), pp. 185–194.

[2] N. D. ALIKAKOS, P. HESS, AND H. MATANO, *Discrete order preserving semigroups and stability for periodic parabolic differential equations*, J. Differential Equations, 82 (1989), pp. 322–341.

[3] S. B. ANGENENT, *The Morse–Smale property for a semi-linear parabolic equation*, J. Differential Equations, 62 (1986), pp. 427–442.

[4] P. DE MOTTONI AND A. SCHIAFFINO, *Competition systems with periodic coefficients: A geometric approach*, J. Math. Biol. (1981), pp. 319–335.

[5] B. FIEDLER AND J. MALLET-PARET, *Connections between Morse sets for delay-differential equations*, J. Reine Angew. Math., 397 (1989), pp. 23–41.

[6] G. FUSCO AND W. M. OLIVA, *Jacobi matrices and transversality*, Proc. Roy. Soc. Edinburgh Sect. A.

[7] J. K. HALE AND A. S. SOMOLINOS, *Competition for fluctuating nutrient*, J. Math. Biol., 18 (1983), pp. 255–280.

[8] D. B. HENRY, *Some infinite dimensional Morse–Smale systems defined by parabolic partial differential equations*, J. Differential Equations, 59 (1985), pp. 165–205.

[9] M. W. HIRSCH, *Systems of differential equations which are competitive or cooperative* I: *limit sets*, SIAM J. Math. Anal., 13 (1982), pp. 167–179.

[10] ———, *Systems of differential equations which are competitive or cooperative*, II: *convergence almost everywhere*, SIAM J. Math. Anal., 16 (1985), pp. 432–439.

[11] ———, *Systems of differential equations which are competitive or cooperative*, III: *competing species*, Nonlinearity, 1 (1988), pp. 51–71.

[12] ———, *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math., 383 (1988), pp. 1–53.

[13] J. MALLET-PARET AND H. L. SMITH, *The Poincaré–Bendixson Theorem for monotone cyclic feedback systems*, J. Dynamics and Differential Equations, to appear.

[14] H. MATANO, *Non increase of the lapnumber of a solution for a one dimensional semi-linear parabolic equation*, J. Fac. Sci. Univ. Tokyo, 29 (1982), pp. 401–441.

[15] ———, *Strong comparison principle in nonlinear parabolic equations*, in Nonlinear Parabolic Equations: Qualitative Properties of Solutions, L. Boccardo and A. Tesei, eds., Pitman Res. Notes in Math. 149, Longman Scientific and Technical, London, 1987, pp. 148–155.

[16] K. NICKEL, *Gestattaussagen über Lösungen parabolischer Differentialgleichungen*, J. Reine Angew. Math., 211 (1962), pp. 78–94.

[17] P. POLÁČIK, *Convergence in smooth strongly monotone flows defined by semilinear parabolic equations*, J. Differential Equations, 79 (1989), pp. 89–110.

[18] J. SMILLIE, *Competitive and cooperative tridiagonal systems of differential equations*, SIAM J. Math. Anal., 15 (1984), pp. 530–534.

[19] H. L. SMITH, *Periodic solutions of periodic competitive and cooperative systems*, SIAM J. Math. Anal., 17 (1986), pp. 1289–1318.

[20] ———, *Periodic competitive differential equations and the discrete dynamics of competitive maps*, J. Differential Equations, 64 (1986), pp. 165–194.

[21] ———, *A discrete Lyapunov function for a class of linear differential equations*, Pacific J. Math., 144 (1990), pp. 345–360.

[22] H. L. SMITH AND H. R. THIEME, *Quasiconvergence and stability for strongly order preserving semiflows*, SIAM J. Math. Anal., 21 (1990), pp. 673–692.

[23] ———, *Convergence for strongly order-preserving semiflows*, SIAM J. Math. Anal., this issue (1991), pp. 1081–1100.

[24] P. TAKÁČ, *Domains of attraction of generic ω-limit sets for strongly monotone discrete-time semigroups*, preprint.

[25] ———, *Convergence to equilibrium on invariant d-hypersurfaces for strongly increasing discrete-time semigroups*, J. Math. Anal. Appl., to appear.

[26] ———, *Linearly stable subharmonic orbits in strongly monotone time-periodic dynamical systems*, preprint.

[27] ———, *Asymptotic behavior of strongly monotone time-periodic dynamical processes with symmetry*, preprint.

[28] E. N. DANCER AND P. HESS, *Stability of fixed points for order-preserving discrete-time dynamical systems*, preprint.

[29] H. L. SMITH, *Convergent and oscillatory activation dynamics for cascades of neural nets with nearest neighbor competitive or cooperative interactions*, Neural Networks, to appear.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1110-1130, July 1991

016

THE SPHERICAL WIENER-PLANCHEREL FORMULA AND SPECTRAL ESTIMATION

JOHN J. BENEDETTO†

Abstract. The d-dimensional spherical analogues of Wiener's "s-function" and difference operator are defined, and the role of the iterated Laplacian is explained. The corresponding spherical Wiener-Plancherel formula is formulated and proved. A recipe for spectral estimation is extracted from the formula.

Key words. Wiener-Plancherel, Tauberian theorems, iterated Laplacian, Besov and Besicovich spaces, spherical mean value, difference operators

AMS(MOS) subject classifications. 42B10, 46F10, 62M15

Introduction. We shall prove a spherical Wiener-Plancherel formula for d-dimensional Euclidean space $\mathbb{R}^d$. This formula is an analogue of the Plancherel formula in the case of functions that are not square-integrable. Wiener developed such a formula for $\mathbb{R}$ [W2], and it became a beacon in his perception and formulation of the statistical theory of communication, e.g., [W3], [Le]. Wiener even chose to have the formula appear on the cover of his autobiography, *I Am a Mathematician.*

What exactly is a Wiener-Plancherel formula? Given a function φ defined on $\mathbb{R}^d$ having Fourier transform $\hat{\varphi}$ defined on $\hat{\mathbb{R}}^d$ $(=\mathbb{R}^d)$. Suppose the distribution $\hat{\varphi}$ is intractable, as is likely for poorly behaved φ. Let s be an operable integral of $\hat{\varphi}$, i.e., suppose that s is a well-behaved function and that $Ls = \hat{\varphi}$, distributionally, for some differential operator L. Wiener's idea was to deal with a computable function s instead of the more esoteric distribution $\hat{\varphi}$, and to relate the quadratic behavior of φ and s. In particular, for the spherical case dealing with balls $B(0, R) = \{t \in \mathbb{R}^d : |t| \leqq R\}$ having volumes $|B(0, R)|$, a Wiener-Plancherel formula has the form

$$\lim_{R\to\infty} \frac{1}{|B(0, R)|} \int_{B(0,R)} |\varphi(t)|^2 \, dt = Q(s),$$

where $Q(s)$ is an explicit quadratic expression and Q, s, and L are interdependent (cf. (1.1) for the exact formula). In Wiener's original result ($d = 1$), Ls can be correctly formulated as a first distributional derivative of s, e.g., [B1, § 2.1], and

$$Q(s) = \lim_{\lambda\to 0} \frac{1}{2\lambda} \int_{-\infty}^{\infty} |s(\gamma+\lambda) - s(\gamma-\lambda)|^2 \, d\gamma$$

(cf. [La], [M]).

The Plancherel formula allows us to define the Fourier transform of a square-integrable function f, and, at certain levels of abstraction, it is considered to characterize what is meant by an harmonic analysis of f. On the other hand, for most applications in $\mathbb{R}^d$, the Plancherel formula assumes the workaday role of an effective tool used to obtain quantitative results. It is this latter role we envisage for Wiener-Plancherel formulas in the non-square-integrable case. After all, distribution theory (in $\mathbb{R}^d$) gives the proper definition of the Fourier transform of tempered distributions. The real issue

* Received by the editors January 7, 1990; accepted for publication (in revised form) July 30, 1990.
† The MITRE Corporation and the Department of Mathematics, University of Maryland, College Park, Maryland 20742.

is to obtain quantitative results for problems where an harmonic analysis of a non-square-integrable function is desired. A host of such problems comes under the heading of an harmonic (spectral) analysis of signals containing non-square-integrable noise and/or random components, whether it be speech recognition, image processing, geophysical modeling, or turbulence in fluid mechanics. Such problems can be attacked by Beurling's profound theory of spectral synthesis, e.g., [B1], as well as by the extensive multifaceted theory of time series, e.g., [P]. Beurling's spectral synthesis does not deal with energy and power considerations, i.e., quadratic criteria, and time series relies on a stochastic point of view. Our goal is to implement Wiener-Plancherel formulas to address the above-mentioned group of problems. These formulas are well suited to deal with energy and power; they provide an analytic device which should dovetail with spectral estimation methods (from time series) developed since Wiener's time.

Our title is misleading in that we do not obtain results in spectral estimation. However, our Wiener-Plancherel formula contains a critical recipe for power spectrum estimation that is discussed in § 7. This recipe is the basis for our forthcoming work on multidimensional spectral estimation.

The paper is organized as follows. In § 1 we state our spherical Wiener-Plancherel formula, viz. (1.1), without going into any detail concerning hypotheses and motivation. We feel that the technicalities in proof are sufficiently complex to warrant a displayed version of our goal at the outset. We also remark on some previous work and the relevance of the spherical case. Section 2 provides the required Fourier analysis on multiplicative groups, including the Tauberian theorem, as well as some examples of special functions that are used in the proof of (1.1). It turns out that the differential operator L is an iterated Laplacian Δ^k in our case, and the Wiener s-function described above is defined by $\Delta^k s = \hat{\varphi}$. Section 3 deals with these notions and the subtleties required to define distributional convolution properly for this setting (cf. (1.2), Theorem 3.8, and the hypotheses of Theorem 5.7). The expression $Q(s)$ combines a spherical mean-value operator and spherical difference operator. These ideas and accompanying technicalities are the subject of § 4. At this point we are ready to prove our spherical Wiener-Plancherel formula. This is accomplished in Theorem 5.7 of § 5. Besides the Tauberian theorem and the material in §§ 3 and 4, we also utilize the space of functions having bounded quadratic means over spheres as well as the Beurling algebra, which is its predual. Because the Tauberian theorem is required, § 6 contains a preliminary result and example concerning zeros of Fourier transforms of relevant special functions.

Besides the usual notation in analysis as found in the books by Hörmander [Hö], Schwartz [S], and Stein and Weiss [SW], we shall use the conventions and notation described at the end of the paper.

1. Statement of the spherical Wiener–Plancherel formula.

Formula 1.1. The spherical Wiener-Plancherel formula is

$$\lim_{R\to\infty} \frac{1}{|B(0,R)|} \int_{B(0,R)} |\varphi(t)|^2\, dt = \lim_{\lambda\to 0} \frac{c(d,k)(2\pi)^{4k}}{\omega_{d-1}\lambda^{4k-d}} \int |D_\lambda s_k(\gamma)|^2\, d\gamma. \tag{1.1}$$

The function s_k is the Wiener s-function

$$s_k = \hat{\varphi} * E_k, \tag{1.2}$$

where $\Delta^k E_k = 1$, ω_{d-1} is the surface area of the unit sphere Σ_{d-1}, $c(d,k)^{-1}$ is the L^1-norm of a special function related to the Fourier transform of the restriction of surface measure σ_{d-1} to Σ_{d-1}, e.g., Example 2.5,

$$D_\lambda s_k = s_k - \mathcal{M}_\lambda s_k,$$

and

$$\mathcal{M}_\lambda s_k(\gamma) = \frac{1}{\omega_{d-1}} \int_{\Sigma_{d-1}} s_k(\gamma + \lambda\theta)\, d\sigma_{d-1}(\theta).$$

The integer k is related to the dimension d, and there must be control of the quadratic means of φ over spheres in order to verify (1.1). The operator L described in the Introduction is the iterated Laplacian Δ^k.

Remark 1.2. (a) In a previous work [BBE], we proved a rectilinear version of (1.1). The rectilinear result is easier to prove than the spherical one, although by no means elementary. Also, in the case of "rectilinear geometry" the operator L is the hyperbolic operator

$$L = \partial_1 \partial_2 \cdots \partial_d,$$

whereas the "spherical geometry" of (1.1) gives rise to the elliptic operator $L = \Delta^k$. This remark indicates there is a range of Wiener-Plancherel formulas according to the number of degrees of freedom available in various convergence criteria.

(b) It is natural to expect significant differences between the rectilinear and spherical cases.

The analogous situation with the convergence problem for multiple Fourier series makes this point clear. There are several natural rectilinear convergence criteria for multiple Fourier series, and there exist positive results in some cases. For example, using the Carleson-Hunt theorem for $d = 1$, Fefferman [F] proved that

$$\lim_{R\to\infty} \sum_{m \in RP \cap \mathbb{Z}^d} a_m e^{2\pi i t \cdot m} = \varphi(t) \quad \text{a.e.} \tag{1.3}$$

for $\varphi \in L^p(\mathbb{R}^d/\mathbb{Z}^d)$, $1 < p \leqq \infty$, where $P \subseteq \mathbb{R}^d$ is a d-dimensional polygon. The rectilinear convergence we used in [BBE] is analogous to the so-called "restricted rectangular" convergence criterion in the theory multiple Fourier series; this criterion is different from that of (1.3). If the polygonal convergence of (1.3) is replaced by spherical convergence, then it is not known whether all the elements of $L^2(\mathbb{R}^d/\mathbb{Z}^d)$, $d > 1$, have a Fourier series representation pointwise almost everywhere. There are negative results if $p < 2$. The problem of multiple Fourier series with spherical convergence criteria is closely related to deep problems associated with Bochner-Riesz multipliers. There are some positive results, and we close this discussion with one such recent theorem written in terms of Fourier transforms [CS]: if $d \geqq 2$, $\alpha > 0$, $2 \leqq p < 2d/(d-1)$, and φ is an element of the Sobolev space $L^p_\alpha(\mathbb{R}^d)$, then

$$\lim_{R\to\infty} \int_{B(0,R)} \hat{\varphi}(\gamma)\, e^{2\pi i t \cdot \gamma}\, d\gamma = \varphi(t) \quad \text{a.e.}$$

Example 1.3. A formula such as (1.1) established a mapping between spaces of functions. For example, if the left side of (1.1) is finite then $\|\varphi | B^2(\mathbb{R}^d)\| < \infty$, where $B^2(\mathbb{R}^d)$ consists of functions having bounded quadratic means over spheres, e.g., (5.1). There is a hierarchy of Besicovich spaces $B(p, q)$ of which $B^2 = B(2, \infty)$. For the right side of (1.1) the corresponding hierarchy $V(p, q)$ is related to Besov spaces. In the case $d = 1$, the mappings $B(2, 1) \to V(2, 1)$ and $B(2, \infty) \to V(2, \infty)$, established by Wiener's original Wiener-Plancherel formula, are topological isomorphisms, e.g., [Be] and [CL], respectively. Taking $d \geqq 1$ and using the rectilinear Wiener-Plancherel formula in [BBE], Heil [He] proved that the mapping $B(2, q) \to V(2, q)$ is a topological isomorphism for $1 \leqq q \leqq \infty$.

2. The Fourier transform on $\mathbb{R}_+^*$. Let $\mathbb{R}_+^*$ denote the set, $\{r \in \mathbb{R}: r \in (0, \infty)\}$, considered as a multiplicative group under ordinary multiplication. $\mathbb{R}_+^*$ is a locally compact Abelian group taken with the usual topology from $\mathbb{R}$. Haar measure μ on $\mathbb{R}_+^*$ is defined by the formula

$$\forall \Theta \in C_c(\mathbb{R}_+^*), \quad \int_{\mathbb{R}_+^*} \Theta \, d\mu = \int_0^\infty \Theta(r) \frac{dr}{r},$$

and convolution on $\mathbb{R}_+^*$ is written as

$$\Theta \otimes \Phi(r) = \int_0^\infty \Theta\left(\frac{r}{s}\right) \Phi(s) \frac{ds}{s}.$$

$(\mathbb{R}_+^*)^\wedge$, the dual group of $\mathbb{R}_+^*$, consists of elements $\rho \in \hat{\mathbb{R}}$ under the mapping $\rho: \mathbb{R}_+^* \to \mathbb{C}$, $r \mapsto r^{i\rho}$.

DEFINITION 2.1. The *Fourier transform* of $\Theta \in L^1(\mathbb{R}_+^*)$ is the function on $(\mathbb{R}_+^*)^\wedge$ defined as

$$\forall \rho \in (\mathbb{R}_+^*)^\wedge, \quad \mathscr{F}(\Theta)(\rho) = \int_0^\infty r^{i\rho} \Theta(r) \frac{dr}{r}.$$

Formally, the *Mellin transform* of Θ, a function on $\mathbb{R}_+^*$, is

$$M(\Theta)(z) = \int_0^\infty r^{z-1} \Theta(r) \, dr, \qquad z \in \mathbb{C}.$$

Thus,

$$\mathscr{F}(\Theta)(\rho) = M(\Theta)(i\rho).$$

Wiener's Tauberian theorem and its generalizations are valid on any locally compact Abelian group [B1]. The form we use in § 5 is the following.

THEOREM 2.2. *Given $\Psi \in L^1(\mathbb{R}_+^*)$ for which $|\mathscr{F}\Psi(\rho)| > 0$ on $(\mathbb{R}_+^*)^\wedge$, let $\Phi \in L^\infty(\mathbb{R}_+^*)$ and assume there is a constant C_Φ for which*

$$\lim_{r \to \infty} \Psi \otimes \Phi(r) = C_\Phi \int_0^\infty \Psi(s) \frac{ds}{s}.$$

Then

$$\forall \Theta \in L^1(\mathbb{R}_+^*), \quad \lim_{r \to \infty} \Theta \otimes \Phi(r) = C_\Phi \int_0^\infty \Theta(s) \frac{ds}{s}.$$

Example 2.3. Define the function

$$K_A(r) = \begin{cases} 0 & \text{for } r \in (0, 1), \\ 1/r^d & \text{for } r \geqq 1, \end{cases}$$

where $d > 0$. The subscript A is used to designate "arithmetic mean" (cf. (2.1)).

First, note that $K_A \in L^1(\mathbb{R}_+^*)$ since $\int_0^\infty K_A(r) \, dr/r = 1/d$. Also, if $\Phi \in L^\infty(\mathbb{R}_+^*)$ then

$$K_A \otimes \Phi(r) = \frac{1}{R^d} \int_0^R r^{d-1} \Phi(r) \, dr \tag{2.1}$$

since $K_A(R/r)=0$ if and only if $r>R$. Finally, $\mathcal{F}(K_A)(\rho)=1/(d-i\rho)$, so that $|\mathcal{F}(K_A)(\rho)|>0$ on $(\mathbb{R}_+^*)^\wedge$.

Example 2.4. Given an integer $d\geqq 2$, define the function

$$K_k(r)=r^{4k-d}\left(1-\frac{2\pi}{\omega_{d-1}}\left(\frac{1}{r}\right)^{-(d-2)/2}J_{(d-2)/2}\left(\frac{2\pi}{r}\right)\right)^2$$

for a given integer $k\geqq 0$. K_k plays a critical role in the definition of the spherical difference operator (§ 4) used in our formulation of the Wiener-Plancherel theorem (§ 5) (cf. Example 6.3). We have

$$\forall r>0 \text{ and } \forall \beta\in\Sigma_{d-1},\quad K_k(r)=r^{4k-d}\left(\frac{\hat{\mu}_{d-1}(0)-\hat{\mu}_{d-1}(\beta/r)}{\omega_{d-1}}\right)^2,$$

where μ_{d-1} is the restriction of surface area measure σ_{d-1} to $\Sigma_{d-1}\subseteq\mathbb{R}^d$.

Observe that $\|K_k|L^1(\mathbb{R}_+^*)\|<\infty$ for $4(k-1)<d<4k$, and that

$$\|K_k|L^1(\mathbb{R}_+^*)\|=\infty \quad \text{for } d\notin(4(k-1),4k).$$

3. The Wiener *s*-function and iterated Laplacian. Let $\chi_R(t)$ be the characteristic function of $B(0,R)^\sim$, $R>0$.

Definition/Remark 3.1. (a) Given $\varphi\in L^1_{\text{loc}}(\mathbb{R}^d)$, assume there is an integer $k\geqq 0$ such that $(\varphi(t)/|t|^{2k})\chi_R(t)\in L^1(\mathbb{R}^d)$ for all $R>0$. The corresponding *k-spherical Wiener s-mean determined by* $R>0$ is defined by

$$s_{k,R}(\gamma)=\left(\frac{i}{2\pi}\right)^{2k}\int e^{-2\pi it\cdot\gamma}\varphi(t)\frac{1}{|t|^{2k}}\chi_R(t)\,dt,\qquad \gamma\in\hat{\mathbb{R}}^d,$$

for each $R>0$. Clearly, $s_{k,R}\in A(\hat{\mathbb{R}}^d)$.

(b) Formally, we have

$$\Delta^k s_{k,R}(\gamma)=\int e^{-2\pi it\cdot\gamma}\varphi(t)\chi_R(t)\,dt, \tag{3.1}$$

where k is a positive integer and Δ^k is the kth iterate of the Laplacian Δ in $\hat{\mathbb{R}}^d$.

(c) Let $\varphi\in\mathcal{S}'(\mathbb{R}^d)\cap L^1_{\text{loc}}(\mathbb{R}^d)$. Because of (3.1) we consider the well-known equation $\Delta^k s_k=\hat{\varphi}$. We write

$$s_k=\hat{\varphi}*E_k, \tag{3.2}$$

where $\Delta^k E_k=\delta$, and formally compute

$$(-1)^k(2\pi)^{2k}|t|^{2k}E_k^\vee(t)=1.$$

The *Wiener s-function* corresponding to φ is the distribution s_k defined by (3.2) when this distributional convolution exists.

The verification of the following result is standard.

THEOREM 3.2. *Given $k\geqq 1$ for which $d>2k$, define*

$$E_k^\vee(t)=\frac{(-1)^k}{(2\pi|t|)^{2k}}.$$

(a) $E_k^\vee\in\mathcal{S}'(\mathbb{R}^d)\cap L^1_{\text{loc}}(\mathbb{R}^d)$ *and its Fourier transform is*

$$E_k(\gamma)=\frac{(-1)^k\lambda(d-2k)}{(2\pi|\gamma|)^{d-2k}}\in L^1_{\text{loc}}(\hat{\mathbb{R}}^d), \tag{3.3}$$

where $\lambda(\alpha)=\pi^{d/2}\Gamma(\alpha/2)/\Gamma((d-\alpha)/2)$, *e.g.*, [St, *Chap.* 3.3.3].

(b) *If* $d < 4k$ *then* $E_k \in L^2_{\mathrm{loc}}(\hat{\mathbb{R}}^d)$ *and, for each* $R > 0$, $E_k^\vee \in L^2(\mathbb{R}^d \setminus B(0, R))$.

(c) *If* $\varphi \in L^\infty_{\mathrm{loc}}(\mathbb{R}^d)$ *and* $\chi_R(t)\varphi(t)/|t|^{2k} \in L^1(\mathbb{R}^d)$ *for each* $R > 0$ *then* $\varphi \in \mathcal{S}'(\mathbb{R}^d)$ *and* $\varphi E_k^\vee \in L^1(\mathbb{R}^d)$.

The hypotheses of Theorem 3.2 were made in light of the requirements to prove the Wiener–Plancherel formula in § 5. In case the convolution (3.2) does not exist for $\varphi \in L^2_{\mathrm{loc}}(\mathbb{R}^d)$, but $\varphi E_k^\vee \in \mathcal{S}'(\mathbb{R}^d)$, then it will be convenient to refer to $(\varphi E_k^\vee)^\wedge$ as the Wiener s-function corresponding to φ. If the hypotheses of Theorem 3.2(c) are satisfied, then

$$s_k = (\varphi E_k^\vee)^\wedge \in A(\hat{\mathbb{R}}^d) \tag{3.4}$$

(cf. Theorem 3.8 and Example 4.6).

Remark 3.3. Suppose $\varphi \in \mathcal{S}'(\mathbb{R}^d) \cap L^1_{\mathrm{loc}}(\mathbb{R}^d)$. Our proof of the Wiener–Plancherel formula requires k for which $4k > d$; $d > 2k$ is required in order that $E_k^\vee$, $E_k \in \mathcal{S}' \cap L^1_{\mathrm{loc}}$. In attempting to define the s-function by means of (3.2), it is easy to see that there are no integers k for which $2k < d < 4k$ in the cases $d = 1, 2, 4$, but that for all other dimensions there are solutions. There is no problem in defining the s-function by means of (3.2) for $d = 1$. In fact, $s = H * \hat{\varphi}$, H the Heaviside function, and Δ^k is replaced by ordinary one-dimensional distributional differentiation. In particular, $s' = \hat{\varphi}$ distributionally [B1].

If $d = 2, 4$ then we can take $E_k(\gamma) = C \log|\gamma|$ by classical distributional methods, e.g., [GS, Chap. 2.3.3 and 2.4.2].

THEOREM 3.4. *If* $\varphi \in L^\infty_{\mathrm{loc}}(\mathbb{R}^d)$, $\chi_R(t)\varphi(t)/|t|^{2k} \in L^1(\mathbb{R}^d)$ *for all* $R > 0$, *and* $d > 2k$ $(k \geqq 1)$ *then*

$$\lim_{R \to 0} \| s_k - s_{k,R} | A(\hat{\mathbb{R}}^d) \| = 0$$

and

$$\forall f \in \mathcal{S}(\hat{\mathbb{R}}^d), \quad \lim_{R \to 0} \Delta^k s_{k,R}(f) = \Delta^k s_k(f) = \hat{\varphi}(f)$$

(*cf.* (3.2)).

Proof. Since s_k, $s_{k,R} \in A(\hat{\mathbb{R}}^d)$ (Theorem 3.2(c) and (3.4)), the norm estimate is a consequence of estimating $\| \varphi(t)\chi_{B(0,R)}(t)/|t|^{2k} | L^1(\mathbb{R}^d) \|$.

If $f \in \mathcal{S}(\hat{\mathbb{R}}^d)$ then distributional differentiation and Fubini's theorem allow us to write

$$\Delta^k s_{k,R}(f) = s_{k,R}(\Delta^k f) = \left(\frac{i}{2\pi}\right)^{2k} \int_{|t| \geqq R} \frac{\varphi(t)}{|t|^{2k}} (\Delta^k f)^\wedge(t)\, dt = \int_{|t| \geqq R} \varphi(t) \hat{f}(t)\, dt,$$

and so

$$\lim_{R \to 0} \Delta^k s_{k,R}(f) = \int \varphi(t) \hat{f}(t)\, dt = \hat{\varphi}(f).$$

We have

$$\lim_{R \to 0} \Delta^k s_{k,R}(f) = \lim_{R \to 0} s_{k,R}(\Delta^k f) = s_k(\Delta^k f) = \Delta^k s_k(f)$$

by distributional differentiation and the weak convergence, $\lim_{R \to 0} s_{k,R}(g) = s_k(g)$, $g \in \mathcal{S}(\hat{\mathbb{R}}^d)$. This weak convergence is a consequence of our $A(\hat{\mathbb{R}}^d)$-norm estimate. □

The situation is more complicated if $d \leqq 2k$, $k \geqq 1$, but there is a result analogous to Theorem 3.4. For example, if $\varphi \in \mathcal{S}'(\mathbb{R}^d)$ is Borel measurable and $\chi_R(t)\varphi(t)/|t|^{2k} \in L^1(\mathbb{R}^d)$ for all $R > 0$ then $\lim_{R\to 0} \Delta^k s_{k,R}(f) = \hat{\varphi}(f)$ for each f in the "moment space" $\mathcal{S}_0(\hat{\mathbb{R}}^d)$. Of course, $\mathcal{S}_0(\hat{\mathbb{R}}^d)^\vee$ is not dense in $L^1(\mathbb{R}^d)$ since $\{0\} \subseteq \hat{\mathbb{R}}^d$ is a set of spectral synthesis, e.g., [B1].

Example 3.5. Suppose $d > 2k$.

(a) Give $p \in [1, \infty)$. It is not true that $f * E_k \in L^p(\mathbb{R}^d)$ for all $f \in C_c^\infty(\hat{\mathbb{R}}^d)$. In fact, if $f \geqq 0$ then, by Fubini's theorem and translation invariance, we compute

$$\|f * E_k | L^p(\hat{\mathbb{R}}^d)\|^p = C\|f|L^p(\hat{\mathbb{R}}^d)\| \int_0^\infty \rho^{(2k-d)p+d-1}\, d\rho = \infty.$$

(b) On the other hand, $f * E_k \in L^\infty(\hat{\mathbb{R}}^d)$ for all $f \in \mathcal{S}(\mathbb{R}^d)$. In fact,

$$|f * E_k(\gamma)| \leqq C \int_{|\lambda|\leqq 1} \frac{1}{|\lambda|^{d-2k}} |f(\gamma - \lambda)|\, d\lambda + C \int_{|\lambda|>1} |f(\gamma-\lambda)|\, d\lambda$$

$$\leqq C\|f|L^\infty(\hat{\mathbb{R}}^d)\|\omega_{d-1} \int_0^1 \rho^{2k-1}\, d\rho + C\|f|L^1(\hat{\mathbb{R}}^d)\|.$$

Because of (3.2) and (3.4) it is desirable to define $\hat{\varphi} * E_k$ for a large class of functions $\varphi \in \mathcal{S}'(\mathbb{R}^d) \cap L^1_{\text{loc}}(\mathbb{R}^d)$. To this end, it is useful to know that $f * \hat{\varphi} \in L^1(\hat{\mathbb{R}}^d)$ for all $f \in \mathcal{S}(\hat{\mathbb{R}}^d)$, in light of the fact that $f * E_k \in L^\infty(\mathbb{R}^d)$ (see, e.g., Definition/Remark 3.7(a)).

Definition/Remark 3.6. (a) $\mathcal{O}_c(\hat{\mathbb{R}}^d)$ is the space of infinitely differentiable functions g for which there is $m \in \mathbb{Z}$ such that

$$\forall \alpha, \quad (1+|\gamma|^2)^m\, \partial^\alpha g(\gamma) \in C_0(\hat{\mathbb{R}}^d).$$

(b) If $\varphi \in \mathcal{S}'(\mathbb{R}^d) \cap L^1_{\text{loc}}(\mathbb{R}^d)$ then $f * \hat{\varphi} \in \mathcal{O}_c(\hat{\mathbb{R}}^d)$ for all $f \in \mathcal{S}(\hat{\mathbb{R}}^d)$ (see, e.g., [H, Prop. 4.11.7]); the exchange formula $(f^\vee \varphi)^\wedge = f * \hat{\varphi}$ is valid [S].

(c) Further, if $\varphi \in \mathcal{S}'(\mathbb{R}^d) \cap L^1_{\text{loc}}(\mathbb{R}^d)$ then $f^\vee \varphi \in L^1(\mathbb{R}^d)$ for all $f \in \mathcal{S}(\hat{\mathbb{R}}^d)$. In fact, $|\varphi|$ is a positive tempered measure so that

$$\int \frac{|\varphi(t)|}{(1+|t|^2)^m}\, dt < \infty$$

for some m, e.g., [S, Thm. 7.7.4]; hence,

$$\int |(f^\vee \varphi)(t)|\, dt = \int |f^\vee(t)|(1+|t|^2)^m \frac{|\varphi(t)|}{(1+|t|^2)^m}\, dt$$

$$\leqq \|f^\vee(t)(1+|t|^2)^m | L^\infty(\hat{\mathbb{R}}^d)\| \int \frac{|\varphi(t)|}{(1+|t|^2)^m}\, dt < \infty.$$

Definition/Remark 3.7. (a) Given $S, T \in \mathcal{S}'(\hat{\mathbb{R}}^d)$, suppose $(S*f)(\tilde{T}*g) \in L^1(\hat{\mathbb{R}}^d)$ for all $f, g \in \mathcal{S}(\hat{\mathbb{R}}^d)$, where $\tilde{T}(g) = T(\gamma)(g(-\gamma))$. Then there is a unique element $S * T \in \mathcal{S}'(\hat{\mathbb{R}}^d)$, the *$\mathcal{S}'$-convolution* of S and T, satisfying the equation

$$\forall f, g \in \mathcal{S}(\hat{\mathbb{R}}^d), \quad ((S*T)*f)(g) = \int (S*f)(\gamma)(\tilde{T}*g)(\gamma)\, d\gamma. \tag{3.5}$$

The fact that (3.5) determines a well-defined element of $\mathcal{S}'(\hat{\mathbb{R}}^d)$ and the subsequent definition of convolution are due to [HO, p. 148].

(b) Note that (3.5) is automatic if $S, T \in \mathcal{S}(\hat{\mathbb{R}}^d)$ and $S * T$ is usual convolution:

$$\begin{aligned}((S*T)*f)(g) &= \int \left(\int S*T(\lambda) f(\gamma-\lambda)\, d\lambda \right) g(\gamma)\, d\gamma \\ &= \int\int\int S(\lambda-\mu) T(\mu) f(\gamma-\lambda) g(\gamma)\, d\mu\, d\lambda\, d\gamma \\ &= \int\int\int S(\lambda-\mu) T(\mu) f(\gamma-\lambda) g(\gamma)\, d\lambda\, d\gamma\, d\mu \\ &= \int\int\int S(\gamma-\mu-\eta) T(\mu) f(\eta) g(\gamma)\, d\eta\, d\gamma\, d\mu \\ &= \int\int\int S(\nu-\eta) T(\mu) f(\eta) g(\nu+\mu)\, d\eta\, d\nu\, d\mu \\ &= \int \left(\int S(\nu-\eta) f(\eta)\, d\eta \right) \left(\int T(\mu) g(\nu+\mu)\, d\mu \right) d\nu \\ &= \int (S*f)(\nu)(\tilde{T}*g)(\nu)\, d\nu.\end{aligned}$$

(c) Given $S, T \in D'(\mathbb{R}^d)$, the *product* ST is defined as

$$\lim_{n\to\infty} (S*\theta_n)(T*\theta_n) = ST, \qquad \sigma(D'(\mathbb{R}^d), C_c^\infty(\mathbb{R}^d)),$$

if the left side exists for each sequence $\{\theta_n\} \subseteq C_c^\infty(\mathbb{R}^d)$ with the properties that $\bigcap \operatorname{supp} \theta_n = \{0\}$ and $\int \theta_n(t)\, dt = 1$ for all n.

THEOREM 3.8. *Suppose $d > 2k$, $k \geqq 1$. Assume $\varphi \in \mathcal{S}'(\mathbb{R}^d)$ and*

$$\forall f \in \mathcal{S}(\hat{\mathbb{R}}^d), \quad (1+|\gamma|^2)^{k+1} (f * \hat{\varphi})(\gamma) \in L^1(\hat{\mathbb{R}}^d). \tag{3.6}$$

Then the $\mathcal{S}'$ convolution $s_k = \hat{\varphi} * E_k$ exists and

$$(\hat{\varphi} * E_k)^\vee = \varphi E_k^\vee. \tag{3.7}$$

Proof. (a) We can verify that

$$\forall f \in \mathcal{S}(\hat{\mathbb{R}}^d), \quad (1+|\gamma|^2)^{-(k+1)} f * E_k(\gamma) \in C_b(\hat{\mathbb{R}}^d), \tag{3.8}$$

noting that $f * E_k(\gamma)$ clearly exists for each $\gamma \in \hat{\mathbb{R}}^d$.

(b) Using (3.6) and (3.8), Definition/Remark 3.7(a) allows us to define the $\mathcal{S}'$-convolution of $\hat{\varphi}$ and E_k.

(c) Take any $\psi \in C_c^\infty(\mathbb{R}^d)$. Then $\psi\varphi \in \mathcal{E}'(\mathbb{R}^d)$ and it is easy to check that

$$(\psi\varphi) * E_k^\vee \in \mathcal{S}'(\mathbb{R}^d).$$

Furthermore, by a closure argument, e.g., [S, Thm. 7.8.15], we have

$$((\psi\varphi) * E_k^\vee)^\wedge = (\hat{\psi} * \hat{\varphi}) E_k$$

and $\hat{\psi} * \hat{\varphi} \in \mathcal{O}_c(\hat{\mathbb{R}}^d)$ (see, e.g., Definition/Remark 3.6(b)).

Using (3.6), there is $g \in L^1(\hat{\mathbb{R}}^d)$ for which

$$(\hat{\psi} * \hat{\varphi})(\gamma) E_k(\gamma) = \frac{g(\gamma)}{(1+|\gamma|^2)^{k+1} |\gamma|^{d-2k}},$$

and, in particular, $(\hat{\psi} * \hat{\varphi}) E_k \in L^1(\hat{\mathbb{R}}^d \setminus N)$, where N is a neighborhood of zero. Moreover, since $\hat{\psi} * \hat{\varphi} \in \mathcal{C}_c(\hat{\mathbb{R}}^d)$, we know it is bounded on N and hence $(\hat{\psi} * \hat{\varphi}) E_k \in L^1(N)$. Thus, $(\hat{\psi} * \hat{\varphi}) E_k \in L^1(\hat{\mathbb{R}}^d)$, and so, $(\psi\varphi) * E_k^\vee \in A(\mathbb{R}^d)$. The continuity allows us to use [IS, Prop. 5] to conclude that $\varphi E_k^\vee$ exists and

$$(\varphi E_k^\vee)(\psi) = ((\psi\varphi) * E_k^\vee)(0).$$

Consequently, we have

$$(\varphi E^\vee)(\psi) = ((\psi\varphi) * E^\vee)(0) = \int \hat{\psi} * \hat{\varphi}(\gamma) E_k(\gamma)\, d\gamma = (\hat{\varphi} * E_k(\lambda))(\hat{\psi}(-\lambda)) = (\hat{\varphi} * E_k)^\vee(\psi),$$

and this is (3.7). □

The procedure we have used to verify the exchange formula (3.7) after showing that the product $\varphi E_k^\vee$ exists is quite general and is due to [IS]. More recent related work is due to members of the Polish school including Antosik, Burzyk, Kaminski, and Wawak. The proof of Theorem 3.8 is easier if we know $\varphi E_k^\vee \in L^1(\mathbb{R}^d)$.

Example 3.9. (a) Condition (3.6) is automatically satisfied if φ is a trigonometric polynomial.

(b) Further, if $\hat{\varphi} = \mu \in M_b(\hat{\mathbb{R}}^d)$ has the property that

$$\int (1+|\gamma|^2)^{k+1}\, d|\mu|(\gamma) < \infty$$

then (3.6) is satisfied for all $f \in \mathcal{S}(\hat{\mathbb{R}}^d)$.

(c) On the other hand, if $\hat{\varphi} = \mu \in M_{b+}(\hat{\mathbb{R}}^d)$ and

$$\int (1+|\gamma|^2)^{k+1}\, d\mu(\gamma) = \infty \tag{3.9}$$

then (3.6) fails for all nonnegative $f \in \mathcal{S}(\hat{\mathbb{R}}^d)$.

It is easy to construct positive-definite elements $\varphi = \mu^\vee \in A(\mathbb{R}^d)$ that satisfy (3.9). For example, if $g(\gamma) = |\gamma|^{-d+1}(1+|\gamma|^2)^{-1}$ then $\|g|L^1(\hat{\mathbb{R}}^d)\| = \pi\omega_{d-1}/2$, whereas the left side of (3.9) for the case $\varphi = g^\vee$ is

$$\omega_{d-1} \int_0^\infty (1+\rho^2)^k\, d\rho = \infty.$$

4. Spherical mean value and difference operators.

Definition/Remark 4.1. (a) The d-dimensional *spherical mean-value operator* $\mathcal{M}_\lambda$ of radius λ is defined as

$$\mathcal{M}_\lambda f(\gamma) = \frac{1}{\omega_{d-1}} \int_{\Sigma_{d-1}} f(\gamma + \lambda\theta)\, d\sigma_{d-1}(\theta), \tag{4.1}$$

where $f: \hat{\mathbb{R}}^d \to \mathbb{C}$, $\gamma \in \hat{\mathbb{R}}^d$, $\lambda > 0$, and $\theta \in \Sigma_{d-1}$. Thus, $\mathcal{M}_\lambda f(\gamma)$ is a mean-value of f at the point γ. In fact, if $d = 2$ and f is analytic within and on a circle of radius λ about γ then the Cauchy integral formula implies that $f(\gamma) = \mathcal{M}_\lambda f(\gamma)$. Furthermore, if $d \geqq 2$ and f is harmonic in a domain containing $\gamma \in \hat{\mathbb{R}}^d$ and the closed ball of radius λ then $f(\gamma) = \mathcal{M}_\lambda f(\gamma)$ [SW, p. 38].

(b) Each operator $\mathcal{M}_\lambda$ is a continuous linear mapping $\mathcal{M}_\lambda: \mathcal{S}(\hat{\mathbb{R}}^d) \to \mathcal{S}(\hat{\mathbb{R}}^d)$. The dual mapping $\mathcal{M}_\lambda^*: \mathcal{S}'(\hat{\mathbb{R}}^d) \to \mathcal{S}'(\hat{\mathbb{R}}^d)$ is well defined by the duality $(\mathcal{M}_\lambda^* f)(\bar{g}) = f(\mathcal{M}_\lambda g)^-$, where $g \in \mathcal{S}(\hat{\mathbb{R}}^d)$.

(c) $\mathcal{M}_\lambda^*: \mathcal{S}'(\hat{\mathbb{R}}^d) \to \mathcal{S}'(\hat{\mathbb{R}}^d)$ is the unique continuous linear mapping on $\mathcal{S}'(\hat{\mathbb{R}}^d)$ that extends $\mathcal{M}_\lambda$ defined on $\mathcal{S}(\hat{\mathbb{R}}^d) \subseteq \mathcal{S}'(\hat{\mathbb{R}}^d)$; as such we designate $\mathcal{M}_\lambda^*$ by $\mathcal{M}_\lambda$ for each $\lambda > 0$.

To see that $\mathcal{M}_\lambda^*$ is an extension of $\mathcal{M}_\lambda$, take $f, g \in \mathcal{S}(\hat{\mathbb{R}}^d)$ and compute

$$(\mathcal{M}_\lambda^* f)(\bar{g}) = f(\mathcal{M}_\lambda g)^- = \int f(\gamma)\overline{(\mathcal{M}_\lambda g)(\gamma)}\, d\gamma$$

$$= \frac{1}{\omega_{d-1}} \int_{\Sigma_{d-1}} \left(\int f(\gamma)\overline{g(\gamma+\lambda\theta)}\, d\gamma \right) d\sigma_{d-1}(\theta)$$

$$= \frac{1}{\omega_{d-1}} \int_{\Sigma_{d-1}} \left(\int f(\gamma+\lambda(-\theta))\overline{g(\gamma)}\, d\gamma \right) d\sigma_{d-1}(\theta)$$

$$= \int \left(\frac{1}{\omega_{d-1}} \int_{\Sigma_{d-1}} f(\gamma+\lambda(-\theta))\, d\sigma_{d-1}(\theta) \right) \overline{g(\gamma)}\, d\gamma = (\mathcal{M}_\lambda f)(\bar{g}).$$

To verify uniqueness, take $f \in \mathcal{S}'(\hat{\mathbb{R}}^d)$ and let $f_n \to f$ in the $\sigma(\mathcal{S}'(\hat{\mathbb{R}}^d), \mathcal{S}(\hat{\mathbb{R}}^d))$ topology, where $\{f_n\} \subseteq \mathcal{S}(\hat{\mathbb{R}}^d)$. Then

$$(\mathcal{M}_\lambda^* f)(\bar{g}) = f(\mathcal{M}_\lambda g)^- = \lim f_n(\mathcal{M}_\lambda g)^- = \lim (\mathcal{M}_\lambda f_n)(\bar{g}).$$

Remark 4.2. In [BBE], we defined a symmetric difference operator Δ_λ which was a unimodular weighted mean-value operator. An analogue for the present situation is

$$\frac{1}{\omega_{d-1}} \int_{\Sigma_{d-1}} w(\theta) f(\gamma+\lambda\theta)\, d\sigma_{d-1}(\theta),$$

where $|w(\theta)| = 1$ and

$$\lim_{\lambda \to 0} \frac{1}{\omega_{d-1}} \int_{\Sigma_{d-1}} w(\theta) f(\gamma+\lambda\theta)\, d\sigma_{d-1}(\theta) = 0.$$

PROPOSITION 4.3. *Let $g \in L^2(\hat{\mathbb{R}}^d)$, $\alpha \in \mathbb{C}$, and $f \in \mathcal{S}'(\hat{\mathbb{R}}^d)$. Assume f satisfies the following conditions: $f^\vee \in \mathcal{S}'(\hat{\mathbb{R}}^d)$ is a Borel measurable function,*

$$\exists R > 0, \quad \text{such that } f^\vee \in L^2(B(0, R)^\sim)$$

and

$$|t|^2 f^\vee(t) \in L^2_{\text{loc}}(\mathbb{R}^d).$$

Then $f - \mathcal{M}_\lambda f \in L^2(\hat{\mathbb{R}}^d)$ and

$$\|g - \alpha(f - \mathcal{M}_\lambda f) | L^2(\hat{\mathbb{R}}^d)\|$$

$$= \left\| g^\vee(t) - \alpha f^\vee(t) \left(1 - \frac{2\pi}{\omega_{d-1}} (|t|\lambda)^{-(d-2)/2} J_{(d-2)/2}(2\pi|t|\lambda) \right) \middle| L^2(\mathbb{R}^d) \right\|. \tag{4.2}$$

Proof. (a) The hypothesis $|t|^2 f^\vee(t) \in L^2_{\text{loc}}(\mathbb{R}^d)$ implies that $f^\vee \in L^2_{\text{loc}}(\mathbb{R}^d \setminus \{0\})$. In fact, if $K \subseteq \mathbb{R}^d$ is compact and $0 \notin K$ then

$$\int_K |f^\vee(t)|^2\, dt = \int_K \frac{|t|^4}{|t|^4} |f^\vee(t)|^2\, dt \leqq C \int_K |t|^4 |f^\vee(t)|^2\, dt < \infty.$$

If we did not assume $f^\vee$ to be a Borel measurable function, then it could contain terms of the form $\partial^\beta \delta$.

(b) We now prove

$$f^\vee(t)\theta_\lambda(t) = f^\vee(t) \left(1 - \frac{2\pi}{\omega_{d-1}} (|t|\lambda)^{-(d-2)/2} J_{(d-2)/2}(2\pi|t|\lambda) \right) \in L^2(\mathbb{R}^d). \tag{4.3}$$

Since $f^\vee(t) \in L^2(B(0,R)^\sim)$ and

$$\sup_{|t|\geqq R} |t|^{-d+2} J_{(d-2)/2}(2\pi|t|\lambda)^2 \leqq C \sup_{|t|\geqq R} |t|^{-d+2}(2\pi|t|\lambda)^{-1} \leqq C_\lambda \sup_{|t|\geqq R} |t|^{-d+1} \leqq C_\lambda R^{-d+1},$$

it is sufficient for (4.3) to dominate

$$I = \int_{B(0,R)} \left| f^\vee(t)\left(1 - \frac{2\pi}{\omega_{d-1}}(|t|\lambda)^{-(d-2)/2} J_{(d-2)/2}(2\pi|t|\lambda)\right)\right|^2 dt. \tag{4.4}$$

This integral is

$$\int_{B(0,R)} \left| f^\vee(t)\left\{\left(1 - \frac{2\pi}{\omega_{d-1}}(|t|\lambda)^{-(d-2)/2}\left(\frac{2\pi|t|\lambda}{2}\right)^{(d-2)/2} \frac{1}{\Gamma(d/2)}\right)\right.\right.$$
$$\left.\left. - \frac{2\pi}{\omega_{d-1}}(|t|\lambda)^{-(d-2)/2}\left(\frac{2\pi|t|\lambda}{2}\right)^{(d-2)/2} \sum_{k=1}^{\infty} \frac{(-(1/4)(2\pi|t|\lambda)^2)^k}{k!\Gamma(((d-2)/2)+k+1)}\right\}\right|^2 dt$$
$$= \int_{B(0,R)} \left| f^\vee(t)\left\{\frac{2\pi^{d/2}}{\omega_{d-1}} \sum_{k=1}^{\infty} \frac{(-\pi^2|t|^2\lambda^2)^k}{k!\Gamma((d/2)+k)}\right\}\right|^2 dt$$

(see, e.g., Example 2.4). Thus, by Minkowski's inequality, we have

$$I^{1/2} \leqq \sum_{k=1}^{\infty} \left\| f^\vee(t) \frac{2\pi^{d/2}}{\omega_{d-1}} \frac{(\pi|t|\lambda)^{2k}}{k!\Gamma((d/2)+k)} \right| L^2(B(0,R))\Big\|$$
$$= \Gamma\left(\frac{d}{2}\right) \sum_{k=1}^{\infty} \frac{(\pi\lambda)^{2k}}{k!\Gamma((d/2)+k)} \|f^\vee(t)|t|^{2k} | L^2(B(0,R))\|,$$

and this is finite since $|t|^2 f^\vee(t) \in L^2_{\text{loc}}(\mathbb{R}^d)$. Consequently, (4.3) is valid.

(c) Distributionally, we have

$$\forall \psi \in \mathcal{S}(\mathbb{R}^d), \quad (f^\vee \theta_\lambda)(\bar{\psi}) = (f^\vee \theta_\lambda)^\wedge(\tilde{\hat{\psi}}).$$

The left-hand side is

$$\int f^\vee(t)(\theta_\lambda \bar{\psi})(t)\, dt = \int f^\vee(t)\overline{(\theta_\lambda(t)\psi(t))}\, dt = \overline{f(\theta_\lambda \psi)^\wedge}$$
$$= f(\gamma)\left(\int \left(1 - \frac{1}{\omega_{d-1}} \int_{\Sigma_{d-1}} e^{-2\pi i t\cdot\lambda\theta}\, d\sigma_{d-1}(\theta)\right)\psi(t)\, e^{-2\pi i t\cdot\gamma}\, dt\right)^-$$
$$= f(\gamma)\left(\hat{\psi}(\gamma) - \frac{1}{\omega_{d-1}} \int_{\Sigma_{d-1}} \left(\int \psi(t)\, e^{-2\pi i t\cdot(\gamma+\lambda\theta)}\, dt\right) d\sigma_{d-1}(\theta)\right)^-$$
$$= f(\hat{\psi} - \mathcal{M}_\lambda \hat{\psi})^- = f(\hat{\psi})^- - (\mathcal{M}_\lambda^* f)(\hat{\psi})^- = (f - \mathcal{M}_\lambda f)(\tilde{\hat{\psi}}).$$

Thus, $f - \mathcal{M}_\lambda f = (f^\vee \theta_\lambda)^\wedge$.

Since $f^\vee \theta_\lambda \in L^2(\mathbb{R}^d)$ we know that $(f^\vee \theta_\lambda)^\wedge \in L^2(\hat{\mathbb{R}}^d)$, and hence, $f - \mathcal{M}_\lambda f \in L^2(\hat{\mathbb{R}}^d)$. Equation (4.2) is a consequence of the Plancherel theorem. □

The operator (on the function f)

$$-\frac{2d}{\lambda^2}(f - \mathcal{M}_\lambda f)$$

corresponds to the Laplacian in $\hat{\mathbb{R}}^d$ in the same way that the difference operator

$$\frac{1}{2\lambda}(\tau_{-\lambda} f - \tau_\lambda f) \tag{4.5}$$

corresponds to the ordinary derivative in $\hat{\mathbb{R}}$ (cf. [BBE] for the rectangular generalization of (4.5) to $\hat{\mathbb{R}}^d$ and § 1 for the corresponding differential operator). Wiener made the following calculation for the case $d=3$ [W1].

Formal calculation 4.4. Given $f\in\mathcal{S}'(\hat{\mathbb{R}}^d)$ for which $\Delta f\in L^2(\hat{\mathbb{R}}^d)$ and $f^\vee$ is Borel measurable, then

$$\lim_{\lambda\to 0}\left\|\Delta f-\left(-\frac{2d}{\lambda^2}\right)(f-\mathcal{M}_\lambda f)\,\middle|\,L^2(\hat{\mathbb{R}}^d)\right\|=0. \tag{4.6}$$

Proof. Since Δf is a convolution of $f\in\mathcal{S}'(\hat{\mathbb{R}}^d)$ and a distribution having compact support, the exchange formula is valid and $\Delta f^\vee(t)=-4\pi^2|t|^2 f^\vee(t)\in\mathcal{S}'(\mathbb{R}^d)$. The hypothesis $\Delta f\in L^2(\hat{\mathbb{R}}^d)$ allows us to conclude that $-4\pi^2|t|^2 f^\vee(t)\in L^2(\mathbb{R}^d)$. In particular, the hypotheses of Proposition 4.3 are satisfied and thus we have

$$\begin{aligned}&\left\|\Delta f-\left(-\frac{2d}{\lambda^2}\right)(f-\mathcal{M}_\lambda f)\,\middle|\,L^2(\hat{\mathbb{R}}^d)\right\|\\&=\left\|f^\vee(t)\left(\frac{2d}{\lambda^2}\left\{1-\frac{2\pi^2|t|^2\lambda^2}{d}-\frac{2\pi}{\omega_{d-1}}(|t|\lambda)^{-(d-2)/2}J_{(d-2)/2}(2\pi|t|\lambda)\right\}\right)\,\middle|\,L^2(\mathbb{R}^d)\right\|.\end{aligned} \tag{4.7}$$

Using the series representation of J_ν, the right side of (4.7) becomes

$$\begin{aligned}&\left\|f^\vee(t)\left(\frac{2d}{\lambda^2}\left\{\left[1-\frac{2\pi}{\omega_{d-1}}(|t|\lambda)^{-(d-2)/2}\left(\frac{2\pi|t|\lambda}{2}\right)^{(d-2)/2}\frac{1}{\Gamma(d/2)}\right]\right.\right.\\&\quad+\left[-\frac{2(\pi|t|\lambda)^2}{d}-\frac{2\pi}{\omega_{d-1}}(|t|\lambda)^{-(d-2)/2}\left(\frac{2\pi|t|\lambda}{2}\right)^{(d-2)/2}\left(\frac{-(\pi|t|\lambda)^2}{\Gamma((d/2)+1)}\right)\right]\\&\quad\left.\left.-\frac{2\pi^{d/2}}{\omega_{d-1}}\sum_{k=2}^{\infty}\frac{(-(\pi|t|\lambda)^2)^k}{k!\Gamma((d/2)+k)}\right\}\right)\,\middle|\,L^2(\mathbb{R}^d)\right\|\\&=2d\Gamma\left(\frac{d}{2}\right)\left\|f^\vee(t)\frac{1}{\lambda^2}\sum_{k=2}^{\infty}\frac{(-(\pi|t|\lambda)^2)^k}{k!\Gamma((d/2)+k)}\,\middle|\,L^2(\mathbb{R}^d)\right\|,\end{aligned} \tag{4.8}$$

where we use the fact that $\Gamma((d/2)+1)=(d/2)\Gamma(d/2)$ to eliminate the $k=1$ term.

The right side of (4.8) *formally* tends to zero as $\lambda\to 0$ since $k\geqq 2$. □

Naturally, the final and only formal step of the preceding calculation is verifiable for many functions $f^\vee$.

Also, in light of the role of iterated Laplacians it is natural to ask if this calculation can be adapted to deal with $\sum_1 a_j\Delta^j f$ instead of Δf. Since $(\sum_1 a_j\Delta^j f)^\vee(t)=\sum_1(-1)^j(2\pi)^{2j}a_j|t|^{2j}$, the cancellation analogous to that indicated in (4.8) is not possible if $\alpha=-2d/\lambda^2$ and $j\geqq 2$. Similar problems arise when α is adjusted. For example, if $\alpha=c/\lambda^{2j}$, $j\geqq 2$, so that the $|t|^{2j}$ terms cancel, then there is no cancellation for the $|t|^{2k}$ terms, $k<j$. Progress can be made when $\mathcal{M}_\lambda$ is replaced by more complicated means.

DEFINITION 4.5. The d-dimensional *spherical difference operator* D_λ of radius λ is defined by

$$D_\lambda f(\gamma)=f(\gamma)-(\mathcal{M}_\lambda f)(\gamma). \tag{4.9}$$

If the hypotheses of Proposition 4.3 are satisfied for a function f then the spherical difference operator gives rise to an $L^2(\mathbb{R}^d)$-valued function of λ, the path of which is a *helix* (in $L^2(\mathbb{R}^d)$) in the sense of Masani's theory of helices (see, e.g., [M, pp. 351–359]). The helical theory is more complicated for our rectangular Wiener–Plancherel formula (see, e.g., [He]).

Because of the Wiener–Plancherel formulas in Theorems 5.6 and 5.7, it is natural to investigate the relationship between $D_\lambda s_k$ and $s_{k,R}$. We do this in the following example.

Example 4.6. Let $\psi(t) = (i/2\pi)^{2k}\varphi(t)/|t|^{2k}$, where $\varphi \in \mathcal{S}'(\mathbb{R}^d) \cap L^1_{\text{loc}}(\mathbb{R}^d)$, and let f and g be functions defined on $\hat{\mathbb{R}}^d$. We have the following formal relations which are valid for a large class of functions:

$$s_{k,R} = \hat{\psi} * \hat{\chi}_R = s_k * (\delta - \hat{\chi}_{B(0,R)}) = s_k - s_k * \hat{\chi}_{B(0,R)}, \tag{4.10}$$

where $\hat{\psi} = s_k$;

$$\lim_{\lambda \to 0} \mathcal{M}_\lambda f = f \tag{4.11}$$

and

$$\mathcal{M}_\lambda(f * g) = (\mathcal{M}_\lambda f) * g. \tag{4.12}$$

Let $R = 1/\lambda$. Because of (4.11), the mean $\mathcal{M}_\lambda$ of the right side of (4.10) is approximated by $s_k - \mathcal{M}_\lambda(s_k * \hat{\chi}_{B(0,R)})$. Using (4.12) and the weak convergence to δ by $\hat{\chi}_{B(0,R)}$, we see that this approximation is in turn an estimation of $D_\lambda s_k$. Thus, $D_\lambda s_k \approx \mathcal{M}_\lambda s_{k,R}$ (cf. Example 6.3).

5. Proof of the spherical Wiener–Plancherel formula.

Definition/Remark 5.1. (a) The space $B^2(\mathbb{R}^d)$ of functions having *bounded quadratic means over spheres* is the set of all functions $\varphi \in L^2_{\text{loc}}(\mathbb{R}^d)$ for which

$$\|\varphi | B^2(\mathbb{R}^d)\| = \sup_{R>0} \left(\frac{1}{|B(0,R)|} \int_{B(0,R)} |\varphi(t)|^2 \, dt \right)^{1/2} < \infty. \tag{5.1}$$

$B^2(\mathbb{R}^d)$ is a Banach space with norm defined by (5.1).

(b) Given $\varphi \in L^2_{\text{loc}}(\mathbb{R}^d)$. The *spherical average of* φ is the function Φ defined on $\mathbb{R}^*_+$ by

$$\Phi(r) = \frac{1}{\omega_{d-1}} \int_{\Sigma_{d-1}} |\varphi(r\theta)|^2 \, d\sigma_{d-1}(\theta). \tag{5.2}$$

(c) A basic property of spherical averages, and one that is relevant for comparison with the classical and rectangular Wiener–Plancherel formulas [W2], [BBE], is that

$$\Phi \in L^\infty(\mathbb{R}^*_+) \quad \text{implies} \quad \varphi \in B^2(\mathbb{R}^d). \tag{5.3}$$

The verification of (5.3) is immediate:

$$\begin{aligned} \frac{1}{|B(0,R)|} \int_{B(0,R)} |\varphi(t)|^2 \, dt &= \frac{\omega_{d-1}}{|B(0,R)|} \int_0^R r^{d-1} \Phi(r) \, dr \\ &\leqq \frac{\|\Phi | L^\infty(\mathbb{R}^*_+)\|}{d|B(0,R)|} \omega_{d-1} R^d = \|\Phi | L^\infty(\mathbb{R}^*_+)\|. \end{aligned}$$

(d) Clearly, $B^2(\mathbb{R}^d) \setminus L^\infty(\mathbb{R}^d) \neq \varnothing$. In fact we can choose a continuous radial element $\varphi \in L^2(\mathbb{R}^d)$ for which $\overline{\lim}_{r \to \infty} |\varphi(r)| = \infty$. This function also shows that the converse of (5.3) fails since $\Phi(r) = |\varphi(r)|^2$. Furthermore, this observation shows that, for the class of radial continuous functions, $\Phi \in L^\infty(\mathbb{R}^*_+)$ if and only if $\varphi \in L^\infty(\mathbb{R}^d)$.

We quote the following important result of Beurling which is related to one of his deep investigations of spectral synthesis [Be].

THEOREM 5.2 [Be, Thms. I, II] (cf. [BDD]). (a) *$B^2(\mathbb{R}^d)$ is the Banach space dual of the convolution Banach algebra $A^2(\mathbb{R}^d)$ defined as follows*:

$$A^2(\mathbb{R}^d) = \bigcup_{\omega\in\Omega} L^2_{1/\omega}(\mathbb{R}^d) \subseteq L^1(\mathbb{R}^d),$$

where $\Omega = \{\omega \in L^1(\mathbb{R}^d)$: $\omega > 0$, $\|\omega | L^1(\mathbb{R}^d)\| = 1$, ω is radial, and $\omega(|t|)$ decreases on $\mathbb{R}^*_+\}$ and $L^2_{1/\omega}(\mathbb{R}^d)$ is the Banach space of Borel measurable functions φ defined by the norm

$$\|\varphi | L^2_{1/\omega}(\mathbb{R}^d)\| = \left(\int \frac{|\varphi(t)|^2}{\omega(t)}\, dt\right)^{1/2}.$$

The Banach space norm of $\varphi \in A^2(\mathbb{R}^d)$ is

$$\|\varphi | A^2(\mathbb{R}^d)\| = \inf_{\omega\in\Omega} \|\varphi | L^2_{1/\omega}(\mathbb{R}^d)\|,$$

and the duality between $A^2(\mathbb{R}^d)$ and $B^2(\mathbb{R}^d)$ is defined by the relation

$$\forall \theta \in A^2(\mathbb{R}^d) \quad \text{and} \quad \forall \varphi \in B^2(\mathbb{R}^d), \quad \varphi(\theta) = \int \varphi(t)\overline{\theta(t)}\, dt.$$

(b) *$B^2(\mathbb{R}^d)$ is characterized by the intersection*

$$B^2(\mathbb{R}^d) = \bigcap_{\omega\in\Omega} L^2_\omega(\mathbb{R}^d). \tag{5.4}$$

DEFINITION 5.3. The spherical *Wiener space* $W(\mathbb{R}^d)$ consists of all functions $\varphi \in L^2_{\text{loc}}(\mathbb{R}^d)$ for which

$$\|\varphi | W(\mathbb{R}^d)\| = \left(\int \frac{|\varphi(t)|^2}{(1+|t|^2)^d}\, dt\right)^{1/2} < \infty$$

(cf. [BBE, Ex. 4.3]). Clearly, $W(\mathbb{R}^d) = \hat{H}^{-d}$, where H^α is the Hilbert Sobolev space $L^2_\alpha(\mathbb{R}^d)$.

We could prove the first inclusion of the following result directly, analogous to the method used in [BBE, Thm. 3.2], or by invoking the well-known inclusion $B^2(\mathbb{R}^d) \subseteq \hat{H}^\alpha$, for each $\alpha < -d/2$. Instead we use (5.4); this application of Beurling's theorem is due to Heil [He].

THEOREM 5.4. $B^2(\mathbb{R}^d) \subseteq W(\mathbb{R}^d) \subseteq \mathcal{S}'(\mathbb{R}^d)$.

Proof. (a) We shall first prove that $B^2(\mathbb{R}^d) \subseteq W(\mathbb{R}^d)$. The weight $\omega(t) = c/(1+|t|^2)^d$ is an element of Ω, for the proper choice of $c > 0$, and $L^2_\omega(\mathbb{R}^d) = W(\mathbb{R}^d)$. The assertion follows from (5.4).

(b) The proof of the inclusion $W(\mathbb{R}^d) \subseteq \mathcal{S}'(\mathbb{R}^d)$ follows from definitions. □

LEMMA 5.5. *Let $\varphi \in L^2_{\text{loc}}(\mathbb{R}^d)$ for which $\Phi \in L^\infty(\mathbb{R}^*_+)$, and suppose we are given k for which $4k > d$. Then*

$$\forall R > 0, \quad \varphi E^\vee_k \in L^2(\mathbb{R}^d \setminus B(0, R)).$$

Proof.

$$\int_{B(0,R)^\sim} |\varphi(t) E^\vee_k(t)|^2\, dt = \frac{1}{(2\pi)^{4k}} \int_R^\infty r^{d-4k-1} \Phi(r)\, dr \leqq \frac{\|\Phi | L^\infty(\mathbb{R}^*_+)\|}{(2\pi)^{4k}(4k-d)} R^{d-4k}. \qquad \square$$

THEOREM 5.6. *Given $\varphi \in L^2_{\text{loc}}(\mathbb{R}^d)$ and an integer k for which $4k > d$, assume $\Phi \in L^\infty(\mathbb{R}^*_+)$. Then*

$$\begin{aligned} \lim_{R\to\infty} \frac{1}{|B(0,R)|} \int_{B(0,R)} |\varphi(t)|^2\, dt &= \lim_{\lambda\to 0} \frac{(4k-d)(2\pi)^{4k}}{\omega^2_{d-1}\lambda^{4k-d}} \int |s_{k,1/\lambda}(\gamma)|^2\, d\gamma \\ &= \lim_{R\to\infty} \frac{(4k-d)R^{4k-d}}{\omega^2_{d-1}} \int_{B(0,R)^\sim} |\varphi(t)|^2 \frac{1}{|t|^{4k}}\, dt, \end{aligned} \tag{5.5}$$

in the sense that if any one limit exists, then the others exist and they are all equal.

Proof. We compute

$$(2\pi)^{4k}\int |s_{k,R}(\gamma)|^2\,d\gamma = \int_{B(0,R)^\sim} |\varphi(t)|^2 \frac{1}{|t|^{4k}}\,dt$$

$$= \omega_{d-1}\int_R^\infty r^{d-4k-1}\left(\frac{1}{\omega_{d-1}}\int_{\Sigma_{d-1}} |\varphi(r\theta)|^2\,d\sigma_{d-1}(\theta)\right)dr$$

$$= \omega_{d-1}\lambda^{4k-d}\int_R^\infty (\lambda r)^{d-4k}\Phi(r)\frac{dr}{r}$$

$$= \omega_{d-1}\lambda^{4k-d}\int_0^\infty \chi_{(0,1)}\left(\frac{R}{r}\right)\left(\frac{R}{r}\right)^{4k-d}\Phi(r)\frac{dr}{r}$$

$$= \omega_{d-1}\lambda^{4k-d}\Theta\otimes\Phi(R),$$

where $\lambda = 1/R$ and $\Theta(s) = \chi_{(0,1)}(s)s^{4k-d} \in L^1(\mathbb{R}_+^*)$. The calculation is valid because of the Plancherel theorem and because the hypotheses allow us to use Lemma 5.5.

Also,

$$\mathscr{F}(\Theta)(\rho) = \int_0^\infty r^{i\rho}\Theta(r)\frac{dr}{r} = \frac{1}{4k-d+i\rho},$$

so that $|\mathscr{F}(\Theta)| > 0$ on $(\mathbb{R}_+^*)\hat{}$.

Next, we have

$$\frac{1}{|B(0,R)|}\int_{B(0,R)} |\varphi(t)|^2\,dt = \frac{1}{|B(0,R)|}\int_0^R r^{d-1}\Phi(r)\,dr$$

$$= \frac{d}{\omega_{d-1}R^d}\int_0^R r^{d-1}\Phi(r)\,dr = \frac{d}{\omega_{d-1}}K_A\otimes\Phi(R),$$

where the last step follows from Example 2.3 and where $|B(0,R)| = \omega_{d-1}R^d/d$. Thus,

$$\frac{1}{|B(0,R)|}\int_{B(0,R)} |\varphi(t)|^2\,dt = \left(\frac{d}{\omega_{d-1}}K_A\right)\otimes\Phi(R).$$

Recall that $|\mathscr{F}(K_A)| > 0$ on $(\mathbb{R}_+^*)\hat{}$.

We now apply the Tauberian theorem, Theorem 2.2. If

$$\lim_{R\to\infty}\left(\frac{d}{\omega_{d-1}}K_A\right)\otimes\Phi(R) = C_\Phi\int_0^\infty \frac{d}{\omega_{d-1}}K_A(r)\frac{dr}{r},$$

then

$$\lim_{R\to\infty}\Theta\otimes\Phi(R) = C_\Phi\int_0^\infty \Theta(r)\frac{dr}{r},$$

and vice versa. Consequently,

$$\frac{1}{\int_0^\infty K_A(r)\,dr/r}\lim_{R\to\infty}K_A\otimes\Phi(R) = \frac{1}{\int_0^\infty \Theta(r)\,dr/r}\lim_{R\to\infty}\Theta\otimes\Phi(R)$$

and so,

$$\lim_{R\to\infty}\frac{\omega_{d-1}}{|B(0,R)|}\int_{B(0,R)}|\varphi(t)|^2\,dt=\lim_{R\to\infty}(4k-d)(2\pi)^{4k}R^{4k-d}\int|s_{k,R}(\gamma)|^2\,d\gamma$$

$$=\lim_{R\to\infty}(4k-d)R^{4k-d}\int_{B(0,R)^{\sim}}|\varphi(t)|^2\frac{1}{|t|^{4k}}\,dt. \qquad \square$$

The limit in (5.5) involving $s_{k,1/\lambda}$ can be replaced by one involving $\mathcal{M}_\lambda s_{k,1/\lambda}$ at least in the case where the first limit of (5.5) is assumed to exist, e.g., Example 6.3.

THEOREM 5.7. *Let $\varphi\in L^2_{\mathrm{loc}}(\mathbb{R}^d)$, which is bounded in a neighborhood of the origin, and given an integer k for which $4(k-1)<d<4k$ and $d>2k$ ($d>4(k-1)$ implies $d>2k$ for $k\geqq 2$). Assume $\Phi\in L^\infty(\mathbb{R}^*_+)$. Then $s_k=(\varphi E_k^\vee)^\wedge\in\mathcal{S}'(\hat{\mathbb{R}}^d)$ and*

$$\lim_{R\to\infty}\frac{1}{|B(0,R)|}\int_{B(0,R)}|\varphi(t)|^2\,dt=\lim_{\lambda\to 0}\frac{c(d,k)(2\pi)^{4k}}{\omega_{d-1}\lambda^{4k-d}}\int|D_\lambda s_k(\gamma)|^2\,d\gamma, \tag{5.6}$$

where $c(d,k)^{-1}=\|K_k|L^1(\mathbb{R}^*_+)\|$. Equation (5.6) signifies that if the left side exists then the right side exists and they are equal.

Proof. (a) The hypotheses allow us to invoke Proposition 4.3. To see this we proceed as follows. The function $s_k^\vee=\varphi E_k^\vee$ is Borel measurable, and $s_k^\vee\in L^2(B(0,R)^{\sim})$ since $\Phi\in L^\infty(\mathbb{R}^*_+)$ and $4k>d$. Furthermore, $|t|^2 s_k^\vee(t)\in L^2_{\mathrm{loc}}(\mathbb{R}^d)$ since

$$\int_{B(0,R)}\frac{|\varphi(t)|^2}{|t|^{4(k-1)}}\,dt\leqq\omega_{d-1}\|\Phi|L^\infty(\mathbb{R}^*_+)\|\int_0^R r^{d-4k+3}\,dr$$

and $4(k-1)<d$.

Finally, we must show $s_k^\vee\in\mathcal{S}'(\mathbb{R}^d)$. Since $d>2k$ we know from Theorem 3.2(a) that $E_k^\vee\in\mathcal{S}'(\mathbb{R}^d)\cap L^1_{\mathrm{loc}}(\mathbb{R}^d)$. For $\theta\in\mathcal{S}(\mathbb{R}^d)$ we have the estimate

$$|s_k^\vee(\theta)|\leqq\frac{C(\varphi,R)}{(2\pi)^{2k}}\int_{B(0,R)}\frac{|\theta(t)|}{|t|^{2k}}\,dt$$

$$+\frac{1}{(2\pi)^{2k}}\|\varphi|W(\mathbb{R}^d)\|\left(\int_{B(0,R)^{\sim}}|\theta(t)|^2\frac{(1+|t|^2)^d}{|t|^{4k}}\,dt\right)^{1/2},$$

where $|\varphi(t)|\leqq C(\varphi,R)$ on $B(0,R)$. The right side is finite by Theorem 5.4, the hypothesis $d>2k$, and the fact $\theta\in\mathcal{S}(\mathbb{R}^d)$. If $\lim_{m\to\infty}\theta_m=0$ in $\mathcal{S}(\mathbb{R}^d)$ then it is clear that the right side of the inequality tends to zero with θ replaced by θ_m. Thus, $s_k^\vee\in\mathcal{S}'(\mathbb{R}^d)$.

Therefore, Proposition 4.3 applies and thus $D_\lambda s_k\in L^2(\hat{\mathbb{R}}^d)$ for all $\lambda>0$, and (4.2) is valid for $f=s_k$.

(b) We can now compute

$$\int|D_\lambda s_k(\gamma)|^2\,d\gamma=\int\left|s_k^\vee(t)\left(1-\frac{2\pi}{\omega_{d-1}}(|t|\lambda)^{-(d-2)/2}J_{(d-2)/2}(2\pi|t|\lambda)\right)\right|^2 dt$$

$$=\frac{1}{(2\pi)^{4k}}\int_0^\infty r^{d-4k}\Phi(r)\left(1-\frac{2\pi}{\omega_{d-1}}(r\lambda)^{-(d-2)/2}J_{(d-2)/2}(2\pi r\lambda)\right)^2\frac{dr}{r}.$$

Therefore, if $R=1/\lambda$ then

$$\frac{(2\pi)^{4k}}{\lambda^{4k-d}}\int|D_\lambda s_k(\gamma)|^2\,d\gamma=\int_0^\infty\Phi(r)\left(\frac{R}{r}\right)^{4k-d}\left(1-\frac{2\pi}{\omega_{d-1}}\left(\frac{r}{R}\right)^{-(d-2)/2}J_{(d-2)/2}\left(\frac{2\pi r}{R}\right)\right)^2\frac{dr}{r}$$

$$=K_k\otimes\Phi(R).$$

(c) Recall from the proof of Theorem 5.6 that

$$\frac{1}{|B(0,R)|}\int_{B(0,R)}|\varphi(t)|^2\,dt=\left(\frac{d}{\omega_{d-1}}K_A\right)\otimes\Phi(R)$$

and $|\mathcal{F}(K_A)|>0$ on $(\mathbb{R}^*_+)^\wedge$. Therefore, we can apply the Tauberian theorem, Theorem 2.2, to conclude that if

$$\lim_{R\to\infty}\frac{\omega_{d-1}}{|B(0,R)|}\int_{B(0,R)}|\varphi(t)|^2\,dt=C_\Phi,$$

then

$$\lim_{\lambda\to 0}\frac{c(d,k)(2\pi)^{4k}}{\lambda^{4k-d}}\int|D_\lambda s_k(\gamma)|^2\,d\gamma=C_\Phi. \qquad \square$$

6. Computations and zeros of Fourier transforms. The following lemma plays a role in Logan's incisive analysis of bandlimited functions [L]; the method used to prove the accompanying proposition is standard. The proposition is of the type that can be used to study the function L_k defined in Example 6.3. The function L_k arises in the Wiener–Plancherel formula involving the mean of $s_{k,R}$.

LEMMA 6.1. *Given $L>0$, there exists a real and even function $\varphi_L\in L^2(\mathbb{R})$ having the following properties: $\varphi_L<0$ on $(-\infty,-(1/2L))\cup((1/2L),\infty)$, $\hat{\varphi}_L>0$ on $(-L,0)\cup(0,L)$, and* supp $\hat{\varphi}_L=[-L,L]$, *e.g.,*

$$\varphi_L(t)=\frac{L}{\pi}\frac{\cos^2(\pi tL)}{1-(2tL)^2}$$

and

$$\hat{\varphi}_L(\gamma)=\begin{cases}\dfrac{1}{4}\sin\dfrac{\pi\gamma}{L}, & |\gamma|<L,\\ 0, & |\gamma|\geqq L.\end{cases}$$

PROPOSITION 6.2. *Given a nonnegative function $\psi\in(L^1(\mathbb{R})\cap L^2(\mathbb{R}))\backslash\{0\}$ and assuming $\psi=0$ on $(-1/2L,1/2L)$, there is $\rho_0\in(-L,L)$ for which* $(\operatorname{Re}\hat{\psi})(\rho_0)=0$.

Proof. Choose φ_L as in Lemma 6.1. By Parseval's formula and hypothesis, we compute

$$\begin{aligned}0&>\int_{|t|>(1/2L)}\psi(t)\varphi_L(t)\,dt=\int\psi(t)\varphi_L(t)\,dt\\&=\int\hat{\psi}(\gamma)\hat{\varphi}_L(-\gamma)\,d\gamma=\int_{-L}^{L}\hat{\psi}(t)\hat{\varphi}_L(-\gamma)\,d\gamma\\&=\int_{-L}^{L}(\operatorname{Re}\hat{\psi}(\gamma))\hat{\varphi}_L(-\gamma)\,d\gamma+i\int_{-L}^{L}(\operatorname{Im}\hat{\psi}(\gamma))\hat{\varphi}_L(-\gamma)\,d\gamma=a+ib.\end{aligned}$$

Clearly, $a<0$ and $b=0$. Also, $\hat{\psi}(0)=\int\psi(t)\,dt>0$ and so $\hat{\psi}(0)=(\operatorname{Re}\hat{\psi})(0)$. Since $a<0$ and $\hat{\varphi}_L(-\gamma)>0$ on $(-L,0)\cup(0,L)$, we can conclude that $\operatorname{Re}\hat{\psi}<0$ on some subset of positive measure of $(-L,L)$. This combines with the continuity of $\hat{\psi}$ and the fact that $(\operatorname{Re}\hat{\psi})(0)>0$ to yield the result. $\square$

Example 6.3. Define the function

$$L_k(r)=\begin{cases}\dfrac{1}{r^{d-4k}}\,I\left(\dfrac{1}{r}\right) & \text{for } r\in(0,1),\\ 0 & \text{for } r\geqq 1,\end{cases}$$

where

$$I(s)=\left|\frac{2\pi}{\omega_{d-1}}\,s^{-(d-2)/2}J_{(d-2)/2}(2\pi s)\right|^2=\frac{1}{\omega_{d-1}^2}\,|\hat{\mu}_{d-1}(2\pi s)|^2.$$

Clearly, $L_k(R/r)=0$ if $r<R$ and

$$\forall r>R,\quad L_k\left(\frac{R}{r}\right)=\left(\frac{2\pi}{\omega_{d-1}}\right)^2\left(\frac{R}{r}\right)^{4k-2}J_{(d-2)/2}\left(\frac{2\pi r}{R}\right)^2. \tag{6.1}$$

(a) We first observe that $L_k\in L^1(\mathbb{R}_+^*)$ if $4k>1$:

$$\int_0^\infty |L_k(r)|\frac{dr}{r}=\int_1^\infty r^{d-4k}|I(r)|\frac{dr}{r}\leqq\int_1^\infty r^{d-4k-1}\left|\frac{2\pi C}{\omega_{d-1}}\,r^{-(d-2)/2}r^{-1/2}\right|^2 dr$$

$$=\left(\frac{2\pi C}{\omega_{d-1}}\right)^2\int_1^\infty r^{d-4k-1}r^{-(d-1)}\,dr=\left(\frac{2\pi C}{\omega_{d-1}}\right)^2\frac{1}{4k-1}<\infty$$

if $4k>1$.

(b) Next, if $\Phi\in L^\infty(\mathbb{R}_+^*)$ and $4k>1$ then $L_k\otimes\Phi(r)$ exists and

$$L_k\otimes\Phi(r)=\int_r^\infty L_k\left(\frac{r}{s}\right)\Phi(s)\frac{ds}{s}=\int_r^\infty\left(\frac{s}{r}\right)^{d-4k}I\left(\frac{s}{r}\right)\Phi(s)\frac{ds}{s}.$$

(c) The reason for considering the function L_k is that the limit in (5.5) involving $s_{k,1/\lambda}$ can be replaced by one involving $\mathcal{M}_\lambda s_{k,1/\lambda}$, at least in the case where the first limit of (5.5) is assumed to exist. This follows from the proof of Theorem 5.6 and the following calculations:

$$\mathcal{M}_{1/R}s_{k,R}(\gamma)=\left(\frac{i}{2\pi}\right)^{2k}\int e^{-2\pi i t\cdot\gamma}\left(\frac{2\pi}{\omega_{d-1}}\left(\frac{|t|}{R}\right)^{-(d-2)/2}J_{(d-2)/2}\left(\frac{2\pi|t|}{R}\right)\frac{\varphi(t)}{|t|^{2k}}\,\chi_R(t)\right)dt$$

and, by Plancherel's theorem,

$$(2\pi)^{4k}\int|\mathcal{M}_{1/R}s_{k,R}(\gamma)|^2\,d\gamma$$

$$=\left(\frac{2\pi}{\omega_{d-1}}\right)^2\int_{B(0,R)^\sim}\left(\frac{R}{|t|}\right)^{d-2}J_{(d-2)/2}\left(\frac{2\pi|t|}{R}\right)^2\frac{|\varphi(t)|^2}{|t|^{4k}}\,dt$$

$$=\left(\frac{2\pi}{\omega_{d-1}}\right)^2\int_R^\infty\frac{r^{-4k+2}}{R^{-d+2}}J_{(d-2)/2}\left(\frac{2\pi r}{R}\right)^2\Phi(r)\frac{dr}{r}$$

$$=\left(\frac{2\pi}{\omega_{d-1}}\right)^2R^{d-4k}\int_R^\infty\left(\frac{R}{r}\right)^{4k-2}J_{(d-2)/2}\left(\frac{2\pi r}{R}\right)^2\Phi(r)\frac{dr}{r}=R^{d-4k}L_k\otimes\Phi(R),$$

where we have used (6.1) in the last step and where Φ is defined in terms of φ as in § 5.

7. Remarks on spectral estimation.

DEFINITION 7.1. Given $\varphi\in L^2_{\mathrm{loc}}(\mathbb{R}^d)$, define

$$\forall R>0,\quad P_{\varphi,R}(t)=\frac{1}{|B(0,R)|}\int_{B(0,R)}\varphi(t+x)\overline{\varphi(x)}\,dx.$$

Suppose that there is a continuous positive definite function P_φ for which $\lim_{R\to\infty} P_{\varphi,R} = P_\varphi$ in the $\sigma(M(\mathbb{R}^d), C_c(\mathbb{R}^d))$ topology, where $M(\mathbb{R}^d)$ is the space of measures on $\mathbb{R}^d$. Then $P_\varphi \in L^\infty(\mathbb{R}^d)$ is the *autocorrelation* of φ, and the positive measure $\mu_\varphi = \hat{P}_\varphi$ is the *power spectrum* of φ.

Remark 7.2. (a) Depending on the particular problem, the weak topology in Definition 7.1 can be replaced by various other convergence criteria, including pointwise convergence.

(b) Given $\varphi \in L^2_{\text{loc}}(\mathbb{R}^d)$ with power spectrum μ_φ, assume there is an increasing function $i(R)$ on $(0,\infty)$ for which $\sup_{|t|\leqq R} |\varphi(t)| \leqq i(R)$ and $\lim_{R\to\infty} i(R)^2/R = 0$. Then we can prove that

$$\forall \psi \in C_c(\mathbb{R}^d), \quad \lim_{R\to\infty} \frac{1}{|B(0,R)|} \int_{B(0,R)} |\psi * \varphi(t)|^2 \, dt = \int |\hat{\psi}(\gamma)|^2 \, d\mu_\varphi(\gamma), \tag{7.1}$$

[B2, § 5].

If we take $\psi = \delta$ in (7.1) then the left side of (7.1) is the arithmetic mean on the left side of our Wiener–Plancherel formula (5.6). Given $\varphi \in L^2_{\text{loc}}(\mathbb{R}^d)$ and combining formulas (5.6) and (7.1), it is then reasonable to expect that

$$\lim_{\lambda\to 0} \frac{c(d,k)(2\pi)^{4k}}{\omega_{d-1}\lambda^{4k-d}} |D_\lambda s_k|^2 = \mu_\varphi \tag{7.2}$$

in some weak topology, e.g., [B1, § 2.1] for the case $d = 1$. In this same spirit we provide the following calculation which Wiener made for the case $d = 1$ [W2, pp. 155–159].

Formal calculation 7.3. Given $\varphi \in L^2_{\text{loc}}(\mathbb{R}^d)$ with autocorrelation P_φ, for $t \in \mathbb{R}^d$,

$$\lim_{\lambda\to 0} \frac{c(d,k)(2\pi)^{4k}}{\omega_{d-1}\lambda^{4k-d}} \int |D_\lambda s_k(\gamma)|^2 \, e^{2\pi i t\cdot\gamma} \, d\gamma = P_\varphi(t). \tag{7.3}$$

Proof. (a) A direct calculation gives

$$\begin{aligned} P_\varphi(t) &= \lim_{R\to\infty} \frac{1}{|B(0,R)|} \int_{B(0,R)} \varphi(t+x)\overline{\varphi(x)} \, dx \\ &= \frac{1}{4} \lim_{R\to\infty} \frac{1}{|B(0,R)|} \int_{B(0,R)} \{|\varphi(t+x)+\varphi(x)|^2 - |\varphi(t+x)-\varphi(x)|^2 \\ &\qquad + i|\varphi(t+x)+i\varphi(x)|^2 - i|\varphi(t+x)-i\varphi(t)|^2\} \, dt \\ &\equiv \frac{1}{4}(K_1 - K_2 + iK_3 - iK_4). \end{aligned} \tag{7.4}$$

(b) Let $\psi(x) = \varphi(t+x) + c\varphi(x)$, where $|c| = 1$, and write $s_k(\theta)(\gamma) = (\theta E_k^\vee)\hat{}\,(\gamma)$, so that $s_k(\varphi) = s_k$. By the Wiener–Plancherel formula we compute

$$\begin{aligned} &\lim_{R\to\infty} \frac{1}{|B(0,R)|} \int_{B(0,R)} |\psi(x)|^2 \, dx \\ &\quad = \lim_{\lambda\to 0} \frac{c(d,k)(2\pi)^{4k}}{\omega_{d-1}\lambda^{4k-d}} \int |D_\lambda s_k(\tau_{-t}\varphi)(\gamma) + D_\lambda s_k(c\varphi)(\gamma)|^2 \, d\gamma \\ &\quad = \lim_{\lambda\to 0} \frac{c(d,k)(2\pi)^{4k}}{\omega_{d-1}\lambda^{4k-d}} \int |D_\lambda s_k(\tau_{-t}\varphi)(\gamma) - e^{2\pi i t\cdot\gamma} D_\lambda s_k(\varphi)(\gamma) \\ &\qquad\qquad + (c + e^{2\pi i t\cdot\gamma}) D_\lambda s_k(\varphi)(\gamma)|^2 \, d\gamma \\ &\quad = E + \lim_{\lambda\to 0} \frac{c(d,k)(2\pi)^{4k}}{\omega_{d-1}\lambda^{4k-d}} \int |D_\lambda s_k(\varphi)(\gamma)|^2 |c + e^{2\pi i t\cdot\gamma}|^2 \, d\gamma, \end{aligned} \tag{7.5}$$

where the "error" E is estimated by

$$\lim_{\lambda\to 0}\frac{c(d,k)(2\pi)^{4k}}{\omega_{d-1}\lambda^{4k-d}}\int |D_\lambda s_k(\tau_{-t}\varphi)(\gamma)-e^{2\pi it\cdot\gamma}D_\lambda s_k(\varphi)(\gamma)|^2\,d\gamma. \tag{7.6}$$

Under natural hypotheses, and implementing Example 2.4 and Proposition 4.3, we can show that the limit in (7.6) vanishes, i.e., $E=0$.

(c) We now combine the right side of (7.4) and (7.5) with $E=0$ for the four cases $c=\pm 1, \pm i$. Thus,

$$\begin{aligned}P_\varphi(t)=\frac{1}{4}\lim_{\lambda\to 0}\frac{c(d,k)(2\pi)^{4k}}{\omega_{d-1}\lambda^{4k-d}}\int |D_\lambda s_k(\gamma)|^2\\ \cdot\{(2+e_{-t}(\gamma)+e_t(\gamma))-(2-e_{-t}(\gamma)-e_t(\gamma))\\ +i(2+ie_{-t}(\gamma)-ie_t(\gamma))-i(2-ie_{-t}(\gamma)+ie_t(\gamma))\}\,d\gamma,\end{aligned}$$

where $e_u(\gamma)=e^{2\pi iu\cdot\gamma}$. Combining terms, we obtain (7.3). □

Formally, (7.2) and (7.3) are compatible. If we are given data φ_S on a set S, these formulas lead us to consider multidimensional *spectral estimators* molded from expressions of the form

$$\frac{c(d,k)(2\pi)^{4k}}{\omega_{d-1}\lambda^{4k-d}}|D_\lambda(\hat{\varphi}_S * E_k)|^2. \tag{7.7}$$

Instead of continuing this section with a quodlibetic discussion of spectral estimation, we shall view (7.7) as the prologue to our forthcoming work on the subject vis a vis classical algorithms, e.g., [IEa-d], [Ma], and evolutionary spectra for nonstationary processes, e.g., [AGT], [P, Chap. 11].

7.1. Notation. If G is a locally compact Abelian group, then $L^1_{\text{loc}}(G)$ is the space of locally integrable functions with respect to Haar measure and $L^1(G)$ is the subspace of integrable functions. $C_c(G)$ consists of all continuous functions on G having compact supports and $C_b(G)$ consists of all continuous bounded functions on G.

The Fourier transform $\hat{\psi}$ of $\psi\in L^1(\mathbb{R}^d)$ is defined by $\hat{\psi}(\gamma)=\int e^{-2\pi it\cdot\gamma}\psi(t)\,dt$, where $\int$ designates integration over $\mathbb{R}^d$; $f^\vee$ is the inverse Fourier transform of f. $A(\hat{\mathbb{R}}^d)$ is the set $\{\hat{\varphi}\colon \varphi\in L^1(\mathbb{R}^d)\}$ and $\mathcal{S}_0(\hat{\mathbb{R}}^d)$ is the set $\{f\in\mathcal{S}(\hat{\mathbb{R}}^d)\colon f^\vee=0$ on a neighborhood of $0\}$. If $\psi\in L^p(\mathbb{R}^d)$, then its usual L^p-norm is denoted by

$$\|\psi|L^p(\mathbb{R}^d)\|.$$

If S is a set then $S^\sim$ is its complement and χ_S is its characteristic function. For $R>0$,

$$\chi_R=\chi_{B(0,R)^\sim}.$$

We denote translation by $(\tau_\lambda f)(\gamma)=f(\gamma-\lambda)$. Finally, the unit sphere in $\mathbb{R}^d$ is $\Sigma_{d-1}=\{t\in\mathbb{R}^d\colon |t|=1\}$.

Acknowledgments. The author thanks two MITRE colleagues and former students, George Benke and Christopher Heil, for sharing their expertise and providing several valuable insights on this material. Also, besides our program on spectral estimation and Wiener-Plancherel formulas, further work on spherical Wiener-Plancherel formulas is forthcoming from George Benke. The referee suggested eliminating calculations in §§ 2 and 3, and kindly pointed out a number of places throughout the paper where Sobolev spaces might be used to advantage over our essentially L^1 approach.

REFERENCES

[AGT] L. AUSLANDER, I. GERTNER, AND R. TOLIMIERI, *The discrete Zak transform application to time-frequency analysis and synthesis of non-stationary signals*, preprint.

[AS] M. ABRAMOVITZ AND L. STEGUN, EDS., *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, 1970.

[B1] J. BENEDETTO, *Spectral Synthesis*, Academic Press, New York, 1975.

[B2] ———, *A multidimensional Wiener-Wintner theorem and spectrum estimation*, Trans. Amer. Math. Soc., to appear.

[BBE] J. BENEDETTO, G. BENKE, AND W. EVANS, *An n-dimensional Wiener-Plancherel formula*, Adv. in Appl. Math., 10 (1989), pp. 457-487.

[BDD] J. BERTRANDIAS, C. DATRY, AND C. DUPUIS, *Unions et intersections d'espaces L^p invariantes par translation ou convolution*, Ann. Inst. Fourier (Grenoble), 28 (1978), pp. 53-84.

[Be] A. BEURLING, *Construction and analysis of some convolution algebras*, Ann. Inst. Fourier (Grenoble), 14 (1964), pp. 1-32.

[CL] Y.-Z. CHEN AND K.-S. LAU, *Harmonic analysis on functions with bounded means*, in Commutative Harmonic Analysis, D. Colella, ed., 1987; Contemp. Math., 91 (1989), pp. 165-175.

[CS] A. CARBERY AND F. SORIA, *Almost-everywhere convergence of Fourier integrals for functions in Sobolev spaces, and an L^2-localization principle*, Math. Sci. Res. Inst., MSRI 05121-88, Berkeley, CA.

[F] C. FEFFERMAN, *On the convergence of multiple Fourier series*, Bull. Amer. Math. Soc., 77 (1971), pp. 744-745.

[GS] I. GEL'FAND AND G. SHILOV, *Generalized Functions, Vol.* 1 (1959), Academic Press, New York, 1964.

[H] J. HORVÁTH, *Topological Vector Spaces and Distributions* I, Addison-Wesley, Reading, MA, 1966.

[HO] Y. HIRATA AND H. OGATA, *On the exchange formula for distributions*, J. Sci. Hiroshima U., 22 (1958), pp. 147-152.

[He] C. HEIL, *Wiener Amalgam Spaces in Generalized Harmonic Analysis and Wavelet Theory*, Ph.D. thesis, University of Maryland, College Park, MD, 1990.

[Hö] L. HÖRMANDER, *The Analysis of Linear Partial Differential Operators, Vol.* I *and* II, Springer-Verlag, Berlin, New York, 1983.

[IE] (a) E. A. ROBINSON, *A historical perspective of spectrum estimation*, Proc. IEEE, 70 (1982), pp. 885-907.

[IE] (b) E. T. JAYNES, *On the rationale of maximum entropy methods*, Proc. IEEE, 70 (1982), pp. 939-953.

[IE] (c) J. MCCLELLAN, *Multidimensional spectral estimation*, Proc. IEEE, 70 (1982), pp. 1029-1039.

[IE] (d) D. J. THOMSON, *Spectrum estimation and harmonic analysis*, Proc. IEEE, 70 (1982), pp. 1055-1096.

[IS] M. ITANO AND R. SHIRAISHI, *On the multiplicative products of distributions*, J. Sci. Hiroshima U., 28 (1964), pp. 223-235.

[L] B. LOGAN, *Extremal problems for positive-definite bandlimited functions*, II, SIAM J. Math. Anal., 14 (1983), pp. 253-257.

[La] K.-S. LAU, *Extensions of Wiener's Tauberian identity and multipliers on the Marcinkiewicz space*, Trans. Amer. Math. Soc., 277 (1983), pp. 489-506.

[Le] Y. LEE, *Statistical Theory of Communication*, John Wiley, New York, 1960.

[M] P. MASANI, *Commentary on the memoire* [30a] *on generalized harmonic analysis*, in Norbert Wiener: Collected Works, P. Masani, ed., Vol. II, MIT Press, Boston, MA, 1979, pp. 333-379.

[Ma] S. L. MARPLE, JR., *A tutorial overview of modern spectral estimation*, Proc. ICASSP, 4 (1989), pp. 2152-2157.

[O] F. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[P] M. PRIESTLEY, *Spectral Analysis and Time Series*, Vols. 1 and 2, Academic Press, New York, 1981.

[S] L. SCHWARTZ, *Théorie des distributions*, Hermann, Paris, 1966.

[SW] E. STEIN AND G. WEISS, *Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.

[St] E. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[T] E. TITCHMARSH, *Theory of Fourier Integrals*, Second ed., Oxford University Press, Oxford, 1967.

[W1] N. WIENER, *Laplacians and continuous linear functionals*, Acta Sci. Math. (Szeged), 3 (1927), pp. 7-16; also in Norbert Wiener: Collected Works, P. Masani, ed., Vol. III, MIT Press, Cambridge, MA, 1981, pp. 718-727.

[W2] ———, *Generalized harmonic analysis*, Acta Math., 55 (1930), pp. 117-258.

[W3] ———, *Time Series* (1949), MIT Press, Cambridge, MA, 1970.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1131-1146, July 1991

017

SPECTRAL THEORY OF JACOBI MATRICES IN $l^2(\mathbb{Z})$ AND THE $su(1,1)$ LIE ALGEBRA*

D. R. MASSON† AND J. REPKA†

Abstract. The connection between orthogonal polynomials, continued fractions, difference equations, and self-adjoint Jacobi matrices acting in $l^2(\mathbb{Z}^+)$ and the extension of these connections to $l^2(\mathbb{Z})$ are reviewed. This yields three different representations for the resolvent of the Jacobi matrix: an integral representation in terms of orthogonal polynomials, a representation in terms of continued fractions, and a representation in terms of the subdominant (or minimal) solution to the associated difference equation. This latter representation is given explicitly in terms of hypergeometric functions for the cases of associated Meixner, Meixner-Pollaczek, and Laguerre polynomials. It is also shown that it is precisely these cases that occur in the unitary irreducible representations of $su(1,1)$ for the resolvent of a real linear combination of the generators of the algebra.

Key words. orthogonal polynomials, continued fractions, difference equations, Jacobi matrices, spectral theory, resolvent, Lie algebra, $su(1,1)$

AMS(MOS) subject classifications. 33A65, 17B15, 30B70, 40A15, 47B39

1. Introduction.

The classical moment problem [2] provides a natural link between properties of Jacobi matrices, continued fractions, and second-order difference equations. The usual setting for these connections is the Hilbert space $l^2(\mathbb{Z}^+)$, the space of all square summable complex sequences. With standard basis vectors $\{e_n\}_{n=0}^{\infty}$, elements of $l^2(\mathbb{Z}^+)$ are written $u=\sum_{n=0}^{\infty} u_n e_n$ and $v=\sum_{n=0}^{\infty} v_n e_n$ and the inner product is $(u, v)=\sum_{n=0}^{\infty} \bar{u}_n v_n$.

Here we will extend these connections to the case of bilaterally infinite Jacobi matrices acting in $l^2(\mathbb{Z})$, but limit our considerations to Jacobi matrices which are essentially self-adjoint on the domain of finite vectors, i.e., the case of deficiency indices (0, 0). We could also consider the cases of deficiency indices (1, 1) or (2, 2) and weighted l^2 spaces. For all of these in the case of Jacobi matrices associated with bilateral birth and death processes, we refer the reader to Pruitt [16], [17] (see also [7], [10]).

We first recall some results associated with unilaterally infinite Jacobi matrices and the classical moment problem. Consider the symmetric Jacobi matrix

$$A := \begin{pmatrix} a_0 & b_1 & & 0 \\ b_1 & a_1 & \ddots & \\ & \ddots & \ddots & \\ 0 & & & \end{pmatrix}, \qquad \begin{aligned} a_n &= (e_n, Ae_n) \in \mathbb{R}, \\ b_n &= (e_{n-1}, Ae_n) > 0, \end{aligned} \tag{1.1}$$

acting in $l^2(\mathbb{Z}^+)$, the associated continued fraction

$$\frac{1}{CF(z)} := \frac{1}{z - a_0 + K_{n=1}^{\infty}(-b_n^2/(z-a_n))}, \tag{1.2}$$

and the difference equation

$$b_{n+1} Y_{n+1}(z) - (z-a_n) Y_n(z) + b_n Y_{n-1}(z) = 0, \qquad n \geqq 0. \tag{1.3}$$

* Received by the editors November 20, 1989; accepted for publication (in revised form) June 11, 1990. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

† Department of Mathematics, University of Toronto, Toronto, Canada, M5S 1A1.

If A is closed, with minimal domain, and self-adjoint (the determined case for the moment problem with moments $c_n = (e_0, A^n e_0)$) then for $z \notin \sigma(A)$, the spectrum of A, [20],

$$(e_0, (zI - A)^{-1} e_0) = \frac{1}{CF(z)}. \tag{1.4}$$

Also

$$\frac{1}{CF(z)} = \lim_{n\to\infty} \frac{Q_n(z)}{P_n(z)}. \tag{1.5}$$

Here $Q_n(z)/P_n(z)$ is the nth approximant of the continued fraction (1.2) (see [21]); $\{P_n\}_{n=0}^{\infty}$, $\{Q_n\}_{n=0}^{\infty}$ are orthogonal polynomial sets (of the first and second kind, respectively) satisfying the difference equations (1.3) subject to the initial conditions $P_0 = Q_1 = 1$, $P_{-1} = Q_0 = 0$. These polynomials are closely related to properties of A. In particular, $\{P_n(z)\}_{n=0}^{\infty}$ are orthogonal with respect to a probability measure $dw(x) = d(e_0, E_A(x)e_0)$, where $E_A(x)$ is the spectral family of orthogonal projections associated with the diagonalization of A. This follows from the fact that (1.1), (1.3) and the initial conditions for P_n imply that

$$e_n = P_n(A)e_0. \tag{1.6}$$

Thus using the functional calculus for self-adjoint operators [20] and (1.6) we obtain

$$(e_m, (zI - A)^{-1} e_n) = \int_{-\infty}^{\infty} \frac{P_m(x)P_n(x)}{z - x}\, dw(x), \qquad z \notin \sigma(A), \tag{1.7}$$

$$w(x) = (e_0, E_A(x)e_0).$$

The leading term in the large z asymptotic expansion of (1.7) gives the claimed orthogonality

$$\delta_{mn} = (e_m, e_n) = \int_{-\infty}^{\infty} P_m(x)P_n(x)\, dw\,(x). \tag{1.7'}$$

Another useful representation for $(zI - A)^{-1}$ can be obtained in terms of a subdominant (minimal) solution to (1.3) (cf. [12]). Let $\{Y_n^{(s)}(z)\}$ be a solution to (1.3) which is in $l^2(\mathbb{Z}^+)$. Since for a determined moment problem with $z \notin \sigma(A)$, we have that such l^2-solutions exist and are unique to within a constant multiple, this implies the subdominant property

$$\lim_{n\to\infty} Y_n^{(s)}(z)/\, Y_n^{(d)}(z) = 0, \qquad z \notin \sigma(A),$$

so long as $Y_n^{(s)}(z)$, $Y_n^{(d)}(z)$ are linearly independent solutions to (1.3). Pincherle's theorem [9] then yields

$$\frac{1}{CF(z)} = \frac{Y_0^{(s)}(z)}{b_0 Y_{-1}^{(s)}(z)}, \qquad z \notin \sigma(A), \tag{1.8}$$

so that (1.4) and (1.7) then give

$$(e_0, (zI - A)^{-1} e_0) = \int_{-\infty}^{\infty} \frac{dw\,(x)}{z - x} = \frac{Y_0^{(s)}(z)}{b_0 Y_{-1}^{(s)}(z)}. \tag{1.9}$$

More generally ([12]; see also Wall [21, Chap. 12]) we have

$$(e_m, (zI-A)^{-1}e_n) = \frac{P_m(z)\,Y_n^{(s)}(z)}{b_0\,Y_{-1}^{(s)}(z)}, \qquad m \leqq n. \tag{1.10}$$

In § 2 we extend the representations (1.4), (1.7), (1.9), and (1.10) for the resolvent of A to the case of bilaterally infinite Jacobi matrices acting in $l^2(\mathbb{Z})$. In § 3 these formulas are made explicit for the bilateral Jacobi matrices connected with associated Meixner, Meixner-Pollaczek, and Laguerre polynomials. In § 4 we give a purely algebraic calculation for the resolvent of a finite Jacobi matrix; it can be combined with Pincherle's theorem to give (1.4) and (1.5). In § 5 these results are applied to the discrete, principal, and complementary (supplementary) series representations of the Lie algebra $su(1,1)$.

2. Resolvent representations in $l^2(\mathbb{Z})$. Consider the complex Hilbert space of square summable bilateral sequences (column vectors)

$$l^2(\mathbb{Z}) := \left\{ u = (u_m) \mid u_m \in \mathbb{C},\ \sum_{m=-\infty}^{\infty} |u_m|^2 < \infty \right\}$$

with inner product

$$(u, v) := \sum_{m=-\infty}^{\infty} \bar{u}_m v_m$$

and standard orthonormal basis $\{e_n\}_{n=-\infty}^{\infty}$, $e_n := (\delta_{nm})$.

Let

$$A := \begin{pmatrix} \ddots & & & & & 0 \\ & a_{-1} & b_0 & & & \\ & b_0 & a_0 & b_1 & & \\ & & b_1 & a_1 & & \\ 0 & & & & \ddots & \end{pmatrix}, \qquad a_n \in \mathbb{R}, \quad b_n > 0 \tag{2.1}$$

be a closed symmetric minimal domain operator acting in $l^2(\mathbb{Z})$. That is, $a_n = (e_n, Ae_n)$, $b_{n+1} = (e_n, Ae_{n+1}) = (e_{n+1}, Ae_n)$, and

$$Ae_n = b_{n+1}e_{n+1} + a_n e_n + b_n e_{n-1}. \tag{2.2}$$

Associated with A we have the difference equation

$$b_{n+1}Y_{n+1}(z) - (z-a_n)Y_n(z) + b_n Y_{n-1}(z) = 0, \qquad n \in \mathbb{Z} \tag{2.3}$$

and the polynomial solutions $\{Y_n = P_n^{(k)}(z)\}_{n=-\infty}^{\infty}$, $k = 0, 1$ satisfying the initial conditions

$$P_n^{(k)}(z) = \delta_{nk}, \qquad n, k = 0, 1.$$

THEOREM 2.1. *A is self-adjoint if and only if*

$$\sum_{n=0}^{\infty} |P_n^{(1)}(z)|^2 = \infty, \qquad \sum_{n=-\infty}^{0} |P_n^{(0)}(z)|^2 = \infty \tag{2.4}$$

for some (and hence all) z such that $\operatorname{Im} z \neq 0$.

Proof. Condition (2.4) implies that the left and right lateral sections of A yield determined moment problems. We show that this is necessary and sufficient in order for A to be self-adjoint. Recall that A is self-adjoint if and only if A has deficiency

indices (0, 0). That is, if and only if $AY = zY$ has no nontrivial solution $Y \in l^2(\mathbb{Z})$ for some, and hence all, z with $\operatorname{Im} z \neq 0$, since by complex conjugation A must have equal deficiency indices [18, Thm. X.3]. Now $AY(z) = zY(z)$, $Y(z) = (Y_n(z))$ is equivalent to (2.3). Hence $Y_n(z) = c_0 P_n^{(0)}(z) + c_1 P_n^{(1)}(z)$. If $\sum_{n=0}^{\infty} |P_n^{(1)}(z)|^2 < \infty$, $\sum_{n=-\infty}^{0} |P_n^{(0)}(z)| < \infty$, then from the classical moment problem [2] it follows that all solutions to (2.2) are in $l^2(\mathbb{Z})$ and A has deficiency indices (2, 2). If $\sum_{n=0}^{\infty} |P_n^{(1)}(z)|^2 < \infty$, $\sum_{n=-\infty}^{0} |P_n^{(0)}(z)|^2 = \infty$, then all solutions to (2.3) are in $l^2(\mathbb{Z}^+)$ and a special choice of c_0, c_1 ($c_1/c_0 = -\lim_{n\to-\infty} P_n^{(0)}(z)/P_n^{(1)}(z)$) yields $Y(z) \in l^2(\mathbb{Z})$ so that A has deficiency indices (1, 1). If $\sum_{n=0}^{\infty} |P_n^{(1)}(z)| = \infty$, $\sum_{n=-\infty}^{0} |P_n^{(0)}(z)| < \infty$, then the special choice $c_0/c_1 = -\lim_{n\to\infty} P_n^{(1)}(z)/P_n^{(0)}(z)$ yields $Y(x) \in l^2(\mathbb{Z})$ so that A again has deficiency indices (1, 1). If $\sum_{n=0}^{\infty} |P_n^{(1)}(z)|^2 = \sum_{n=-\infty}^{0} |P_n^{(0)}(z)|^2 = \infty$, then for no choice of c_0, c_1 can we have $Y \in l^2(\mathbb{Z})$ ($\lim_{n\to\infty} P_n^{(0)}(z)/P_n^{(1)}(z) \neq \lim_{n\to-\infty} P_n^{(0)}(z)/P_n^{(1)}(z)$ since the former limit is $O(1/z)$ while the latter is $O(z)$), and A has deficiency indices (0, 0). □

COROLLARY 2.2. *A is self-adjoint if*

$$\sum_{n=0}^{\infty} \frac{1}{b_n} = \infty, \qquad \sum_{n=-\infty}^{0} \frac{1}{b_n} = \infty. \tag{2.5}$$

Proof. Condition (2.5) is sufficient for the determinate cases for the right and left lateral sections of A [2, p. 24]. □

Assuming that A is self-adjoint, we now derive three different representations for the resolvent of A. The first two have also been derived by Pruitt for the special case of bilaterally infinite birth and death processes [16], [17]. (See also [7], [10].)

2.1. Integral representation.

THEOREM 2.3. *Let A given by* (2.1) *be self-adjoint. Let $E_A(x)$, $x \in \mathbb{R}$, be its spectral family of orthogonal projections. Then*

$$\begin{aligned}(e_m, (zI - A)^{-1} e_n) &= \int_{\mathbb{R}} (z-x)^{-1} \{P_m^{(0)}(x) P_n^{(0)}(x)\, d\mu_{00}(x) + P_m^{(1)}(x) P_n^{(1)}(x)\, d\mu_{11}(x) \\ &\qquad + [P_m^{(0)}(x) P_n^{(1)}(x) + P_m^{(1)}(x) P_n^{(0)}(x)]\, d\mu_{01}(x)\}, \qquad z \notin \sigma(A) \end{aligned} \tag{2.6}$$

$$\mu_{ij}(x) = (e_i, E_A(x) e_j).$$

Proof. From (2.2) and (2.3) it follows that

$$e_n = P_n^{(0)}(A) e_0 + P_n^{(1)}(A) e_1 \tag{2.7}$$

and from the spectral theorem [20]

$$(zI - A)^{-1} = \int_{\mathbb{R}} (z-x)^{-1}\, dE_A(x). \tag{2.8}$$

From (2.7), (2.8) and the functional calculus for self-adjoint operators we obtain (2.6) with $\mu_{01}(x) = \mu_{10}(x)$ since A is real symmetric. □

Note that as $z \to \infty$ the leading $1/z$ term in (2.6) yields a four-term orthogonality

$$\delta_{nm} = (e_m, e_n) = \int_{\mathbb{R}} \sum_{i,j=0}^{1} P_m^{(i)}(x) P_n^{(i)}(x)\, d\mu_{ij}(x) \tag{2.9}$$

involving four sets of polynomials $\{P_n^{(i)}(x)\}$, $i = 0, 1$, $n \geqq 0$, $n < 0$ and four measures $d\mu_{ij}$, $i, j = 0, 1$ (with $d\mu_{00}$, $d\mu_{11}$ positive and two equal-signed measures $d\mu_{01} = d\mu_{10}$).

This is in contrast with the single set of orthogonal polynomials and positive measure associated with a semi-infinite Jacobi matrix and a determined classical moment problem given by (1.7).

2.2. Subdominant representation. Let $Y_n^{(R)}(z)$, $Y_n^{(L)}(z)$ be solutions to (2.3) satisfying the boundary conditions

$$\lim_{n\to\infty} Y_n^{(R)}(z)/Y_n^{(L)}(z)=0,$$

$$\lim_{n\to-\infty} Y_n^{(L)}(z)/Y_n^{(R)}(z)=0.$$

These will be called right and left subdominant solutions, respectively. Their existence and uniqueness to within a constant multiple is assured for $\operatorname{Im} z\neq 0$ when condition (2.4) is satisfied since the moment problem associated with the left and right lateral portions of A are then determined [2].

THEOREM 2.4. *If A given by* (2.1) *is self-adjoint and* $\operatorname{Im} z\neq 0$, *then*

$$(e_m,(zI-A)^{-1}e_n)=\begin{cases}\dfrac{Y_m^{(L)}(z)\,Y_n^{(R)}(z)}{b_{k+1}W(Y_k^{(R)}(z),\,Y_k^{(L)}(z))}, & m\leqq n,\\[2ex] \dfrac{Y_m^{(R)}(z)\,Y_n^{(L)}(z)}{b_{k+1}W(Y_k^{(R)}(z),\,Y_k^{(L)}(z))}, & n\leqq m,\end{cases} \tag{2.10}$$

where

$$W(Y_k^{(R)}(z),\,Y_k^{(L)}(z))=Y_k^{(R)}(z)\,Y_{k+1}^{(L)}(z)-Y_{k+1}^{(R)}(z)\,Y_k^{(L)}(z).$$

Proof. Let $R_{m,n}(z)=(e_m,(zI-A)^{-1}e_n)$. Then since A is symmetric we have $R_{m,n}(z)=R_{n,m}(z)$. From (2.2) we have

$$\begin{aligned} &zR_{m,n}(z)-b_{n+1}R_{m,n+1}(z)-a_nR_{m,n}(z)-b_nR_{m,n-1}(z)\\ &\quad=(e_m,(zI-A)^{-1}(zI-A)e_n)=\delta_{mn}.\end{aligned} \tag{2.11}$$

Thus for $m\neq n$ we have

$$(e_m,(zI-A)^{-1}e_n)=\begin{cases}\displaystyle\sum_{i,j=1}^{2} c_{i,j}\,Y_m^{(i)}(z)\,Y_n^{(j)}(z), & m<n,\\[2ex] \displaystyle\sum_{i,j=1}^{2} c_{i,j}\,Y_n^{(i)}(z)\,Y_m^{(j)}(z), & n<m,\end{cases}$$

where $Y_n^{(i)}(z)$, $i=1,2$, are solutions to (2.3). Now $R_{m,n}$ is a Fourier coefficient with respect to either subscript. That is $\sum_{n=-\infty}^{\infty}|R_{m,n}(z)|^2<\infty$. Since the right and left subdominant solutions are, to within a constant multiple, the unique right and left lateral square summable solutions we must have

$$(e_m,(zI-A)^{-1}e_n)=\begin{cases}cY_m^{(L)}(z)\,Y_n^{(R)}(z), & m<n,\\ cY_n^{(L)}(z)\,Y_m^{(R)}(z), & n<m.\end{cases}$$

The constant c is determined from (2.11) for $m=n$. This yields

$$c(zY_n^{(L)}(z)\,Y_n^{(R)}(z)-b_{n+1}Y_{n+1}^{(L)}(z)\,Y_{n+1}^{(R)}(z)-a_nY_n^{(L)}(z)\,Y_n^{(R)}(z)-b_nY_n^{(R)}(z)\,Y_{n-1}^{(L)}(z))=1.$$

That is, using (2.3) for $Y_n=Y_n^{(L)}(z)$, $c(b_{n+1}Y_n^{(R)}(z)\,Y_{n+1}^{(L)}(z)-b_{n+1}Y_n^{(L)}(z)\,Y_{n+1}^{(R)}(z))=1$. Thus $c=1/b_{n+1}W(Y_n^{(R)}(z),\,Y_n^{(L)}(z))$, which is independent of n. □

2.3. Continued fraction representation. The resolvent elements may be expressed in terms of continued fractions either by taking the limits of the resolvent elements of the finite portions of A derived in § 4 or by applying Pincherle's theorem [9] to the representation (2.10). Here we choose the latter.

Recall that Pincherle's theorem yields a continued fraction representation for a ratio of subdominant terms. Thus

$$-\frac{b_{n+1}Y_{n+1}^{(R)}(z)}{Y_n^{(R)}(z)} = K_{k=1}^{\infty}\left(\frac{-b_{n+k}^2}{z-a_{n+k}}\right), \tag{2.12}$$

$$\frac{b_{n+1}Y_{n+1}^{(L)}(z)}{Y_n^{(L)}(z)} = z-a_n+K_{k=1}^{\infty}\left(\frac{-b_{n+1-k}^2}{z-a_{n-k}}\right), \tag{2.13}$$

where we use the standard notation

$$K_{k=1}^{\infty}\left(\frac{u_k}{v_k}\right) := \lim_{N\to\infty} \cfrac{u_1}{v_1+\cfrac{u_2}{v_2+\ddots\cfrac{}{\cfrac{u_N}{v_N}}}}.$$

THEOREM 2.5. *If A given by* (2.1) *is self-adjoint and* $\operatorname{Im} z \neq 0$, *then*

$$(e_n,(zI-A)^{-1}e_n) = \frac{1}{z-a_n+K_{k=1}^{\infty}(-b_{n+k}^2/(z-a_{n+k}))+K_{k=1}^{\infty}(-b_{n+1-k}^2/(z-a_{n-k}))}. \tag{2.14}$$

Proof. Since A is self-adjoint we have the existence of right and left subdominant solutions $Y_n^{(R)}(z)$, $Y_n^{(L)}(z)$ to (2.3) for $\operatorname{Im} z \neq 0$. From (2.10) with $k=n$ we have

$$(e_n,(zI-A)^{-1}e_n) = \frac{1}{b_{n+1}(Y_{n+1}^{(L)}(z)/Y_n^{(L)}(z)-Y_{n+1}^{(R)}(z)/Y_n^{(R)}(z))}.$$

The representations (2.12), (2.13) then yield (2.14). □

Remarks. (1) The off-diagonal elements of the resolvent are considerably more complicated and involve sums and products of continued fractions. We do not derive them here. However, an expression for the general matrix element of the resolvent for finite portions of A is given in § 4.

(2) Note that for a semi-infinite Jacobi matrix one of the two continued fractions in (2.14) reduces to a finite fraction. In particular, if $b_0=0$ we have the standard expression (see (1.2), (1.4))

$$(e_0,(zI-A)^{-1}e_0) = \frac{1}{z-a_0+K_{k=1}^{\infty}(-b_k^2/(z-a_k))}.$$

(3) Note that (2.14) is the Stieltjes transform of a positive measure. It would be interesting to find the semi-infinite Jacobi matrix

$$A' = \begin{pmatrix} a_0' & b_1' & & 0 \\ b_1' & a_1' & \ddots & \\ & \ddots & \ddots & \\ 0 & & & \end{pmatrix}$$

such that

$$(e_n,(zI-A)^{-1}e_n) = \frac{1}{z-a_0'+K_{k=1}^{\infty}(-b_k'^2/(z-a_k'))}$$

and relate the elements a_n' and b_{n+1}', $n\in\mathbb{Z}^+$ to the original a_n and b_n, $n\in\mathbb{Z}$.

3. Explicit cases. We derive explicit formulas for the resolvent in cases related to associated Meixner, Meixner-Pollaczek, and Laguerre polynomials. That is the cases $a_n = dn$, $b_n^2 = an^2 + bn + c$ with a, b, c, d real, $a, c \neq 0$, and $b_n^2 > 0$.

The resolvent representation (2.10) of § 2 is made explicit through the use of hypergeometric solutions to the associated difference equation and their asymptotic behaviour. These solutions are given in [5] and [12].

3.1. Associated Meixner. Let $a_n = dn$, $b_n^2 = an^2 + bn + c > 0$, $d^2 > 4a > 0$ and consider the difference equation

$$b_{n+1} Y_{n+1}(z) - (z - a_n) Y_n(z) + b_n Y_{n-1}(z) = 0. \tag{3.1}$$

By renormalizing the solutions in [12, eq. (2.3)] we obtain a pair of linearly independent solutions

$$\begin{aligned} Y_{n-1}^{(1),\pm}(z) &= \left(\frac{\pm\sqrt{a}}{\mu}\right)^n \frac{\sqrt{\Gamma(n+\alpha)\Gamma(n+\beta)}}{\Gamma(n+\gamma_\pm(z))}\, {}_2F_1\left(\begin{matrix} n+\alpha, n+\beta \\ n+\gamma_\pm(z) \end{matrix}; \delta_\pm\right), \\ \mu &= \sqrt{d^2-4a}, \qquad an^2+bn+c = a(n+\alpha)(n+\beta), \\ \delta_\pm &= \frac{1 \pm d/\mu}{2}, \qquad \gamma_\pm(z) = (1+\alpha+\beta)\delta_\pm \pm \frac{z}{\mu}. \end{aligned} \tag{3.2}$$

Note that if $d > 0$ then $\delta_- < \frac{1}{2}$. Thus the $Y_n^{(1),-}$ solution is then well defined either through the power series expansion of ${}_2F_1(\delta_-)$ or its analytic continuation. However, the $Y_n^{(1),+}$ solution will have an ambiguity if $\delta_- < 0$ since we then have $\delta_+ > 1$ and evaluation of ${}_2F_1(\delta_+)$ on the branchcut $[1, \infty)$. This ambiguity is resolved by a consistent evaluation of ${}_2F_1$ and does not affect the final result. For $d < 0$ the situation is similar except that we have $\delta_+ < \frac{1}{2}$ and $\delta_- > 1$ if $\delta_+ < 0$.

The $n \to \infty$ asymptotic behaviour of these solutions is obtained through the use of the transformation

$${}_2F_1\left(\begin{matrix} a, b \\ c \end{matrix}; z\right) = (1-z)^{c-a-b}\, {}_2F_1\left(\begin{matrix} c-a, c-b \\ c \end{matrix}; z\right),$$

and the estimate

$${}_2F_1\left(\begin{matrix} a, b \\ c \end{matrix}; z\right) = 1 + O\left(\frac{1}{\operatorname{Re} c}\right), \qquad z \notin [1, \infty), \quad \operatorname{Re} c \to \infty.$$

Thus we obtain as $n \to \infty$

$$Y_n^{(1),\pm}(z) = \left(\pm\frac{\sqrt{a}}{\mu}\right)^{n+1} n^{(\alpha+\beta-2\gamma_\pm)/2} (\delta_\mp)^{\gamma_\pm - (\alpha+\beta+n+1)} \left(1 + O\left(\frac{1}{n}\right)\right). \tag{3.3}$$

A second pair of linearly independent solutions is given by

$$Y_n^{(2),\pm}(z) = \left(\mp\frac{\mu}{\sqrt{a}}\right)^{n+1} \frac{\sqrt{\Gamma(-n-\alpha)\Gamma(-n-\beta)}}{\Gamma(1-n-\gamma_\pm(z))}\, {}_2F_1\left(\begin{matrix} -n-\alpha, -n-\beta \\ 1-n-\gamma_\pm(z) \end{matrix}; \delta_\pm\right). \tag{3.4}$$

This follows from a renormalization of [12, eq. (2.7)] after correcting a misprint (replace $(\pm\mu)^n$ by $(\mp\mu)^n$). The $n \to -\infty$ asymptotic behaviour of these solutions follows similarly and is given by

$$Y_n^{(2),\pm}(z) = \left(\mp\frac{\mu}{\sqrt{a}}\right)^{n+1} (-n)^{(2\gamma_\pm - (\alpha+\beta+1))/2} (\delta_\mp)^{n+\alpha+\beta-\gamma_\pm} \left(1 + O\left(\frac{1}{n}\right)\right). \tag{3.5}$$

From the above asymptotics it is clear that we have right and left subdominant solutions

$$Y_n^{(R)}(z)=\begin{cases} Y_n^{(1),-}(z), & d>0,\\ Y_n^{(1),+}(z), & d<0,\end{cases} \tag{3.6}$$

$$Y_n^{(L)}(z)=\begin{cases} Y_n^{(2),-}(z), & d>0,\\ Y_n^{(2),+}(z), & d<0.\end{cases} \tag{3.7}$$

To apply the resolvent formula (2.10) of § 2 it remains only to calculate $b_{k+1}W(Y_k^{(R)}(z), Y_k^{(L)}(z))$. To this end we use the identity [8, eq. (34), p. 107]

$$u_2=\frac{\Gamma(a+b+1-c)\Gamma(1-c)}{\Gamma(a+1-c)\Gamma(b+1-c)}u_1+\frac{\Gamma(a+b+1-c)\Gamma(c-1)}{\Gamma(a)\Gamma(b)}u_5,$$

$$u_1={}_2F_1\left(\begin{matrix}a,b\\c\end{matrix};z\right),\qquad u_2={}_2F_1\left(\begin{matrix}a,b\\a+b+1-c\end{matrix};1-z\right),$$

$$u_5=z^{1-c}(1-z)^{c-a-b}{}_2F_1\left(\begin{matrix}1-a,1-b\\2-c\end{matrix};z\right)$$

to obtain the connecting formula

$$\begin{aligned} &Y_n^{(1),\pm}(z)=C_\pm Y_n^{(1),\mp}(z)+D_\pm Y_n^{(2),\pm}(z),\\ &C_\pm=\pi^{-1}\Gamma(1+\alpha-\gamma_\pm(z))\Gamma(1+\beta-\gamma_\pm(z))\sin\pi\gamma_\pm(z),\\ &D_\pm=\pi^{-1}\delta_\pm^{1-\gamma_\pm}\delta_\mp^{1-\gamma_\mp}\Gamma(1+\alpha-\gamma_\pm(z))\Gamma(1+\beta-\gamma_\pm(z))\sqrt{\sin\pi\alpha\sin\pi\beta}\,. \end{aligned} \tag{3.8}$$

From (3.5)-(3.8) it now follows that (by calculating $b_{k+1}W(Y_k^{(R)}(z), Y_k^{(L)}(z))$ asymptotically as $k\to\infty$)

$$b_{k+1}W(Y_k^{(R)}(z), Y_k^{(L)}(z))=\begin{cases}\dfrac{\mu\sin\pi\gamma_-(z)}{\sqrt{\sin\pi\alpha\sin\pi\beta}}, & d>0,\\[2ex] \dfrac{-\mu\sin\pi\gamma_+(z)}{\sqrt{\sin\pi\alpha\sin\pi\beta}}, & d<0.\end{cases} \tag{3.9}$$

We summarize these calculations with Theorem 3.1.

THEOREM 3.1. *Let $A=(A_{m,n})$ be a closed symmetric tridiagonal matrix acting in $l^2(\mathbb{Z})$ with $A_{n,n}=dn$, $A_{n,n-1}=A_{n-1,n}=\sqrt{a(n+\alpha)(n+\beta)}$, and $d^2>4a>0$. Then A is self-adjoint and for $m\leqq n$,*

$$\begin{aligned}((zI-A)^{-1})_{m,n}=&\frac{(-1)^{n+1}\mu^{m-n-1}a^{(n-m)/2}\sqrt{\sin\pi\alpha\sin\pi\beta}}{\sin\pi\gamma_-(z)}\\ &\times\frac{\sqrt{\Gamma(-m-\alpha)\Gamma(-m-\beta)}}{\Gamma(1-m-\gamma_-(z))}{}_2F_1\left(\begin{matrix}-m-\alpha,-m-\beta\\1-m-\gamma_-(z)\end{matrix};\delta_-\right)\\ &\times\frac{\sqrt{\Gamma(1+n+\alpha)\Gamma(1+n+\beta)}}{\Gamma(1+n+\gamma_-(z))}{}_2F_1\left(\begin{matrix}1+n+\alpha,1+n+\beta\\1+n+\gamma_-(z)\end{matrix};\delta_-\right),\end{aligned} \tag{3.10}$$

$$\delta_-=\frac{1-d/\mu}{2},\qquad \gamma_-(z)=(1+\alpha+\beta)\delta_--\frac{z}{\mu},$$

where

$$\mu=\begin{cases}\sqrt{d^2-4a}, & d>0,\\ -\sqrt{d^2-4a}, & d<0.\end{cases} \tag{3.11}$$

Proof. Formula (3.10) follows from (2.10), (3.3)-(3.7), and (3.9). □

COROLLARY 3.2. *With the assumptions of Theorem* 3.1, *A has a discrete spectrum of isolated eigenvalues* $\{z_m\}_{m=-\infty}^{\infty}$ *given by*

$$z_m = m\mu + \tfrac{1}{2}(1+\alpha+\beta)(\mu - d), \qquad m \in \mathbb{Z}. \tag{3.12}$$

Proof. We have ${}_2F_1(\begin{smallmatrix}a,b\\c\end{smallmatrix}; z)/\Gamma(c)$ an entire function of c. It follows that the only resolvent singularities of (3.10) are given by the vanishing of $\sin \pi\gamma_-(z)$. That is, $\gamma_-(z) = m$, $m \in \mathbb{Z}$. Solving for z yields (3.12). □

Remarks. (1) It is curious that the eigenvalues given by (3.12) depend only on $\alpha+\beta$ while the off-diagonal elements of A involve both $\alpha+\beta$ and $\alpha\beta$. Thus, in terms of the original parameters a, b, c, d, we have the surprising fact that the eigenvalues depend only on a, b, d and are independent of c.

(2) Note that as $\mu \to 0+$ the eigenvalue spacing tends to zero and we obtain the associated Laguerre case of § 3.3.

3.2. Associated Meixner–Pollaczek. Let $a_n = dn$, $b_n^2 = an^2 + bn + c > 0$, $d^2 < 4a$. The solutions to (3.1) and their asymptotics are again given by (3.2)-(3.5) with $\mu = i\sqrt{4a - d^2}$. However, the right and left subdominant solutions are now changed because $\delta_\pm$ is complex with $\operatorname{Re}\delta_\pm = \frac{1}{2}$ and $|\delta_\pm| = a/(4a-d^2)$. Thus the asymptotic behaviour of (3.4) and (3.5) now dictates

$$Y_n^{(R)}(z) = \begin{cases} Y_n^{(1),-}(z), & \operatorname{Im} z < 0, \\ Y_n^{(1),+}(z), & \operatorname{Im} z > 0, \end{cases} \tag{3.13}$$

$$Y_n^{(L)}(z) = \begin{cases} Y_n^{(2),+}(z), & \operatorname{Im} z < 0, \\ Y_n^{(2),-}(z), & \operatorname{Im} z > 0. \end{cases} \tag{3.14}$$

The connecting formula (3.8) and its asymptotic behaviour now yields

$$b_{k+1} W(Y_k^{(R)}(z), Y_k^{(L)}(z)) = \begin{cases} \pi\mu/\Gamma(1+\alpha-\gamma_-)\Gamma(1+\beta-\gamma_-)\sqrt{\sin\pi\alpha \sin\pi\beta}, & \operatorname{Im} z > 0, \\ -\pi\mu/\Gamma(1+\alpha-\gamma_+)\Gamma(1+\beta-\gamma_+)\sqrt{\sin\pi\alpha \sin\pi\beta}, & \operatorname{Im} z < 0. \end{cases} \tag{3.15}$$

The application of (3.13)-(3.15) to (2.10) then gives Theorem 3.3.

THEOREM 3.3. *Let* $A = (A_{m,n})$ *be a closed symmetric tridiagonal matrix acting in* $l^2(\mathbb{Z})$ *with* $A_{n,n} = dn$, $A_{n,n-1} = A_{n-1,n} = \sqrt{a(n+\alpha)(n+\beta)}$ *and* $d^2 < 4a$. *Then A is self-adjoint and for* $m \leqq n$ *and* $\operatorname{Im} z \neq 0$ *we have*

$$\begin{aligned} ((zI - A)^{-1})_{m,n} = {} & \frac{\Gamma(1+\alpha-\gamma_-(z))\Gamma(1+\beta-\gamma_-(z))\sqrt{\sin\pi\alpha \sin\pi\beta}}{\pi\mu} \\ & \times \left(\frac{\sqrt{a}}{\mu}\right)^{n+1} \frac{\sqrt{\Gamma(1+n+\alpha)\Gamma(1+n+\beta)}}{\Gamma(1+n+\gamma_+(z))} {}_2F_1\left(\begin{matrix}1+n+\alpha, 1+n+\beta \\ 1+n+\gamma_+(z)\end{matrix}; \delta_+\right) \\ & \times \left(\frac{\mu}{\sqrt{a}}\right)^{m+1} \frac{\sqrt{\Gamma(-m-\alpha)\Gamma(-m-\beta)}}{\Gamma(1-m-\gamma_-(z))} {}_2F_1\left(\begin{matrix}-m-\alpha, -m-\beta \\ 1-m-\gamma_-(z)\end{matrix}; \delta_-\right), \end{aligned} \tag{3.16}$$

$$\delta_\pm = \frac{1}{2}\left(1 \pm \frac{d}{\mu}\right), \qquad \gamma_\pm(z) = (1+\alpha+\beta)\delta_\pm \pm \frac{z}{\mu},$$

$$\mu = \begin{cases} i\sqrt{4a-d^2}, & \operatorname{Im} z > 0, \\ -i\sqrt{4a-d^2}, & \operatorname{Im} z < 0. \end{cases}$$

3.3. Associated Laguerre. Let $a_n = dn$, $b_n^2 = an^2 + bn + c > 0$, $d^2 = 4a \neq 0$. Then by renormalizing the solutions (2.16), (2.17), (2.19), and (2.20) in [12] we obtain pairwise linearly independent solutions to (3.1) given by

$$Y_n^{(1)}(z) = (-1)^{n+1}\sqrt{\Gamma(1+n+\alpha)\Gamma(1+n+\beta)}\; U(1+n+\alpha;\, \alpha-\beta+1;\, x), \tag{3.17}$$

$$Y_n^{(2)}(z) = (-1)^{n+1}\frac{\Gamma(1+n+\alpha)}{\sqrt{\Gamma(1+n+\alpha)\tau(1+n+\beta)}}\,{}_1F_1(1+n+\alpha;\, \alpha-\beta+1;\, x), \tag{3.18}$$

$$Y_n^{(3)}(z) = \sqrt{\Gamma(-n-\alpha)\Gamma(-n-\beta)}\; U(-n-\alpha;\, \beta-\alpha+1;\, -x), \tag{3.19}$$

$$Y_n^{(4)}(z) = \frac{\sqrt{\Gamma(-n-\alpha)\Gamma(-n-\beta)}}{\Gamma(-n-\beta)}\,{}_1F_1(-n-\alpha;\, \beta-\alpha+1;\, -x),$$

$$\begin{aligned} &x = -\frac{2z}{d} - \frac{4b}{d^2} - 1, \\ &-x = e^{-i\pi\varepsilon}x, \qquad \varepsilon = \begin{cases} 1, & \arg x > 0, \\ -1, & \arg x \leqq 0, \end{cases} \\ &an^2 + bn + c = a(n+\alpha)(n+\beta). \end{aligned} \tag{3.20}$$

In the above we have used Slater's notation [19] for the confluent hypergeometric functions. Again following Slater, we have confluent hypergeometric functions y_i, $i = 1, \cdots, 8$ with

$$y_1 = {}_1F_1(a;\, b;\, x), \qquad y_4 = x^{1-b}e^x\,{}_1F_1(1-a;\, 2-b;\, -x),$$
$$y_5 = U(a;\, b;\, x), \qquad y_8 = e^x x^{1-b}U(1-b;\, 2-b;\, -x),$$

and $y_3 = y_1$, $y_4 = y_2$, $y_6 = y_5$, $y_7 = e^{-i\pi\varepsilon(1-b)}y_8$. Thus

$$\begin{aligned} Y_n^{(1)} &= (-1)^{n+1}\sqrt{\Gamma(n+1+\alpha)\Gamma(n+1+\beta)}\; y_5, \\ Y_n^{(2)} &= (-1)^{n+1}\frac{\Gamma(1+n+\alpha)}{\sqrt{\Gamma(1+n+\alpha)\Gamma(1+n+\beta)}}\, y_1, \\ Y_n^{(3)} &= \sqrt{\Gamma(-n-\alpha)\Gamma(-n-\beta)}\; e^{-x}x^{\alpha-\beta}y_8, \\ Y_n^{(4)} &= \frac{\sqrt{\Gamma(-n-\alpha)\Gamma(-n-\beta)}}{\Gamma(-n-\beta)}\, e^{-x}x^{\alpha-\beta}y_4 \end{aligned}$$

with confluent parameters $a = 1+n+\alpha$, $b = \alpha-\beta+1$.

From the connecting formula [19, eq. (2.1.54)]

$$y_1 = \frac{\Gamma(b)}{\Gamma(b-a)}\, e^{i\pi\varepsilon a}y_5 + \frac{\Gamma(b)}{\Gamma(a)}\, e^{\varepsilon i\pi(a-b)}y_7,$$

we then have

$$Y_n^{(2)}(z) = \Gamma(\alpha-\beta+1)\; e^{\varepsilon\pi i\alpha}[\sin\pi\beta\, Y_n^{(1)}(z) - e^x x^{\beta-\alpha}\sqrt{\sin\pi\alpha\,\sin\pi\beta}\; Y_n^{(3)}(z)] \tag{3.21}$$

and three other connecting formulas which follow from [19, eqs. (2.1.55)-(2.1.57)].

The large n asymptotic behaviour of $Y_n^{(1)}$, $Y_n^{(2)}$ is obtained from [19, eqs. (4.5.42), (4.6.43)], This yields

$$Y_n^{(2)}(z) = (-1)^{n+1}\; e^{x/2}x^{(\beta-\alpha)/2}\Gamma(\alpha-\beta+1)I_{\alpha-\beta}(2\sqrt{nx})\left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right), \tag{3.22}$$

$$Y_n^{(1)}(z) = 2(-1)^{n+1}\; e^{x/2}x^{(\beta-\alpha)/2}K_{\alpha-\beta}(2\sqrt{nx})\left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) \tag{3.23}$$

in terms of Bessel functions of imaginary argument.

Now for $|\arg u| < \pi/2$, we have

$$I_\nu(u) \sim \frac{e^u}{\sqrt{2\pi u}}, \qquad K_\nu(u) \sim \sqrt{\frac{\pi}{2u}}\, e^{-u}.$$

Thus

$$Y_n^{(R)}(z) = Y_n^{(1)}(z), \qquad -\pi < \arg x < \pi. \tag{3.24}$$

Similarly, we obtain the $n \to -\infty$ asymptotic behaviour of $Y_n^{(3)}$, $Y_n^{(4)}$ so that

$$Y_n^{(L)}(z) = Y_n^{(3)}(z), \qquad -\pi < \arg(-x) < \pi. \tag{3.25}$$

Thus $b_{n+1} W(Y_n^{(R)}(z), Y_n^{(L)}(z)) = b_{n+1} W(Y_n^{(1)}(z), Y_n^{(3)}(z))$, $\operatorname{Im} z \neq 0$, and from (3.21)

$$b_{n+1} W(Y_n^{(1)}(z), Y_n^{(3)}(z)) = \frac{-\pi\, e^{-\pi i \varepsilon \alpha}\, e^{-x} x^{\alpha-\beta}}{\Gamma(\alpha-\beta+1)\sqrt{\sin \pi\alpha \sin \pi\beta}}\, b_{n+1} W(Y_n^{(1)}(z), Y_n^{(2)}(z)).$$

The right-hand side may then be calculated from the asymptotic behaviour of (3.22) and (3.23). This yields

$$b_{n+1} W(Y_n^{(R)}(z), Y_n^{(L)}(z)) = \frac{\pi\sqrt{a}\, e^{-i\pi\varepsilon\alpha}}{\sqrt{\sin \pi\alpha \sin \pi\beta}}. \tag{3.26}$$

The above calculations are summarized by Theorem 3.4.

THEOREM 3.4. *Let $A = (A_{m,n})$ be a closed symmetric tridiagonal matrix acting in $l^2(\mathbb{Z})$ with $A_{n,n} = dn$, $A_{n-1,n} = A_{n,n-1} = \sqrt{a(n+\alpha)(n+\beta)}$, $d^2 = 4a \neq 0$. Then A is self-adjoint and for $m \leqq n$ and* $\operatorname{Im} z \neq 0$,

$$\begin{aligned}((zI - A)^{-1})_{m,n} = {} & 2(-1)^{n+1} \frac{\sqrt{\sin \pi\alpha \sin \pi\beta}}{\pi d}\, e^{i\pi\varepsilon\alpha} \\ & \times \sqrt{\Gamma(-m-\alpha)\Gamma(-m-\beta)}\; U(-m-\alpha; \beta-\alpha+1; -x) \\ & \times \sqrt{\Gamma(1+n+\alpha)\Gamma(1+n+\beta)}\; U(1+n+\alpha; \alpha-\beta+1; x),\end{aligned} \tag{3.27}$$

$$x = -\frac{2z}{d} - \frac{4b}{d^2} - 1,$$

$$-x = e^{-i\pi\varepsilon} x, \qquad \varepsilon = \begin{cases} 1, & \arg x > 0, \\ -1, & \arg x \leqq 0. \end{cases}$$

Remark. We should also be able to obtain explicit expressions for the polynomials and measures in the resolvent representation (2.6). The polynomials are, of course, the associated polynomials of Meixner, Meixner-Pollaczek, or Laguerre. The measures would be obtained by applying the Stieltjes-Perron inversion formula to the resolvent representations obtained above. We do not attempt this here.

4. Finite matrices. In this section we give an algebraic development for the resolvent formula (2.14), based on a calculation of the inverse of a truncated Jacobi matrix. Beginning with the doubly infinite Jacobi matrix A of (2.1), for each $M \leqq N$ we form the finite truncated Jacobi matrix

$$A_N^M = \begin{pmatrix} a_M & b_{M+1} & & 0 \\ b_{M+1} & \ddots & \ddots & \\ & \ddots & \ddots & b_N \\ 0 & & b_N & a_N \end{pmatrix}. \tag{4.1}$$

We calculate its inverse by Cramer's rule. If $i<j$, the ijth minor is

$$\det\begin{pmatrix} \underline{A_{i-1}^M\,|} & & & & & & \\ & b_i & 0 & 0 & & & \\ & b_{i+1} & a_{i+1} & b_{i+2} & & & \\ & 0 & b_{i+2} & \ddots & \ddots & & \\ & & & \ddots & a_{j-2} & b_{j-1} & 0 \\ & & & & b_{j-1} & a_{j-1} & 0 \\ & & & & & b_j & \dfrac{b_{j+1}}{|A_N^{j+1}} \end{pmatrix} \tag{4.2}$$

$$= \left(\prod_{k=i}^{j-1} b_{k+1}\right) \cdot \det(A_{i-1}^M) \cdot \det(A_N^{j+1}).$$

By symmetry, this also equals the jith minor. The iith minor is

$$\det\left(\begin{array}{c|c} A_{i-1}^M & 0 \\ \hline 0 & A_N^{i+1} \end{array}\right) = \det(A_{i-1}^M)\det(A_N^{i+1}),$$

which is consistent with (4.2) if we interpret $\prod_{k=i}^{i-1} b_{k+1}$ as 1.

Expanding by cofactors along the ith column, we find that

$$\det A_N^M = a_i \det(A_{i-1}^M)\cdot\det(A_N^{i+1}) - b_i^2 \det(A_{i-2}^M)\cdot\det(A_N^{i+1}) - b_{i+1}^2 \det(A_{i-1}^M)\cdot\det(A_N^{i+2}).$$

Cramer's rule gives the following formula for the entries of the inverse $(A_N^M)^{-1}$. For $i \leqq j$,

$$(A_N^M)_{ij}^{-1} = (-1)^{i+j}\left(\prod_{k=i}^{j-1} b_{k+1}\right)\cdot\det(A_{i-1}^M)\cdot(A_N^{j+1})\cdot\det(A_N^M)^{-1}$$

$$= (-1)^{i+j}\left(\prod_{k=i}^{j-1} b_{k+1}\right)\left[\frac{a_i \det(A_{i-1}^M)\det(A_N^{i+1}) - b_{i+1}^2\det(A_{i-2}^M)\det(A_N^{i+1}) - b_{i+1}^2\det(A_{i-1}^M)\det(A_N^{i+2})}{\det(A_{i-1}^M)\det(A_N^{j+1})}\right]^{-1} \tag{4.3}$$

$$= (-1)^{i+j}\left(\prod_{k=1}^{j-1} b_{k+1}\right)\left[a_i\frac{\det(A_N^{i+1})}{\det(A_N^{j+1})} - b_i^2\frac{\det(A_{i-2}^M)\det(A_N^{i+1})}{\det(A_{i-1}^M)\det(A_N^{j+1})} - b_{i+1}^2\frac{\det(A_N^{i+2})}{\det(A_N^{j+1})}\right]^{-1}$$

Using the cofactor expansion, it is easy to show the following relations with continued fractions. For $M \leqq i,\ j \leqq N$,

$$\frac{\det(A_N^i)}{\det(A_N^{i+1})} = a_i + K_{k=i}^N\left(\frac{-b_{k+1}^2}{a_{k+1}}\right),$$
$$\frac{\det(A_j^M)}{\det(A_{j-1}^M)} = a_j + K_{k=j}^M\left(\frac{-b_k^2}{a_{k-1}}\right). \tag{4.4}$$

With these formulas it is possible to express the entries of $(A_N^M)^{-1}$ with products of continued fractions. The simplest case is the diagonal entry:

$$[(A_N^M)^{-1}]_{ii} = \frac{1}{a_i - b_i^2(a_{i-1} + K_{k=i-1}^M(-b_k^2/a_{k-1}))^{-1} - b_{i+1}^2(a_{i+1} + K_{k=i+1}^N(-b_{k+1}^2/a_{k+1}))^{-1}}.$$

For the case of interest in this paper, we replace the diagonal entries of (4.1) with $a_i - z$; we then have formulas for the entries of the resolvent $(A_N^M - zI)^{-1}$.

For the case of an infinite matrix we can justify the convergence of the infinite continued fractions using Pincherle's theorem [9]. This gives an alternate proof of formula (2.14).

5. Applications. The explicit cases of § 3 and their classical limits with $b_0^2 = 0$ are closely related to the representation theory of the Lie algebra $so(2, 1)$ or $su(1, 1)$. We detail this connection here.

Let T_j, $j = 1, 2, 3$ be the generators of the Lie algebra $so(2, 1)$ or $su(1, 1)$ with

$$[T_1, T_2] = -iT_3, \quad [T_2, T_3] = iT_1, \quad [T_3, T_1] = iT_2. \tag{5.1}$$

Consider an irreducible unitary representation with T_3 diagonal and the operator

$$T = \alpha_1 T_1 + \alpha_2 T_2 + \alpha_3 T_3, \qquad \alpha_j \in \mathbb{R}. \tag{5.2}$$

For the highest and lowest weight representations (i.e., "discrete series"; cf. [1]) we have T represented as a one-sidedly infinite Jacobi matrix connected with the classical polynomial cases of § 1 having $a_n = dn$ and $b_n^2 = an^2 + bn$, $n \in \mathbb{Z}^+$.

However, for the principal or complementary series representations we have T represented as a bilaterally infinite Jacobi matrix of the form given by the associated polynomial cases of § 3.

The results of § 3 and their $c \to 0$ classical limits allow us to calculate the resolvent and spectral properties of T. Since there are a variety of differential operators which obey the commutation properties (5.1) (cf. [1], [3], [4], [6], [11], [13]-[15]), this also gives a unified approach to the properties of a class of differential operators.

5.1. Discrete series. The lowest-weight representation $D_+(q)$ with lowest weight $q > 0$ has basis $\{\phi_m\}_{n=0}^{\infty}$ and action

$$\begin{aligned} T_3\phi_m &= (q+m)\phi_m, \\ T_1\phi_m &= \frac{1}{2}([(m+1)(2q+m)]^{1/2}\phi_{m+1} + [m(2q+m-1)]^{1/2}\phi_{m-1}), \\ T_2\phi_m &= \frac{1}{2i}([(m+1)(2q+m)]^{1/2}\phi_{m+1} - [m(2q+m-1)]^{1/2}\phi_{m-1}). \end{aligned} \tag{5.3}$$

Relative to this basis the matrix for the operator T in (5.3) is $(T_{m,n})$ with

$$\begin{aligned} T_{m,m} &= \alpha_3(q+m), \\ \bar{T}_{m,m+1} &= T_{m+1,m} = [(m+1)(2q+m)]^{1/2}\left(\frac{\alpha_1 - i\alpha_2}{2}\right). \end{aligned}$$

Letting $e_m = \exp(mi \arg(\alpha_1 - i\alpha_2))\phi_m$, we see that relative to the basis $\{e_m\}_{m=0}^{\infty}$, the operator $T - \alpha_3 qI$ has a one-sided Jacobi matrix representation A, where

$$\begin{aligned} A_{m,m} &= a_m = \alpha_3 m, \\ A_{m,m+1} &= A_{m+1,m} = b_{m+1} = \tfrac{1}{2}[(m+1)(2q+m)]^{1/2}\sqrt{\alpha_1^2 + \alpha_2^2}. \end{aligned} \tag{5.4}$$

In terms of the parameters a, b, c, d of § 3, we then have

$$a = \frac{\alpha_1^2 + \alpha_2^2}{4}, \quad b = \left(\frac{\alpha_1^2 + \alpha_2^2}{4}\right)(2q-1), \quad c = 0, \quad d = \alpha_3. \tag{5.5}$$

This corresponds to the classical Meixner polynomial case when $\alpha_1^2 + \alpha_2^2 - \alpha_3^2 < 0$, to the Meixner-Pollaczek case when $\alpha_1^2 + \alpha_2^2 - \alpha_3^2 > 0$, and to the case of Laguerre polynomials when $\alpha_1^2 + \alpha_2^2 - \alpha_3^2 = 0$.

The highest-weight representation $D_-(q)$ with highest weight $q<0$ has basis $\{\phi_m\}_{m=-\infty}^{0}$ and action

$$T_3\phi_m = (q+m)\phi_m,$$

$$T_1\phi_m = \frac{1}{2}([(m-1)(2q+m)]^{1/2}\phi_{m-1} + [m(2q+m+1)]^{1/2}\phi_{m+1}), \tag{5.6}$$

$$T_2\phi_m = -\frac{1}{2i}([(m-1)(2q+m)]^{1/2}\phi_{m-1} - [m(2q+m+1)]^{1/2}\phi_{m+1}).$$

Letting $e_m = \exp(-mi \arg(\alpha_1 - i\alpha_2))\phi_{-m}$ we see that relative to the basis $\{e_m\}_{m=0}^{\infty}$, the operator $T - \alpha_3 qI$ has a one-sided Jacobi matrix representation A, where

$$A_{m,m} = a_m = \alpha_3 m,$$
$$A_{m,m+1} = A_{m+1,m} = b_{m+1} = \tfrac{1}{2}[(m+1)(m-2q)]^{1/2}\sqrt{\alpha_1^2+\alpha_2^2}, \qquad m \in \mathbb{Z}^+. \tag{5.7}$$

The parameters of § 3 are now

$$a = \frac{\alpha_1^2+\alpha_2^2}{4}, \quad b = \left(\frac{\alpha_1^2+\alpha_2^2}{4}\right)(-2q-1), \quad c = 0, \quad d = -\alpha_3. \tag{5.8}$$

Thus the classical polynomial correspondence, depending on $\alpha^2 = \alpha_1^2+\alpha_2^2-\alpha_3^2$, is the same for both the lowest- and highest-weight cases.

In summary, the discrete series representations correspond precisely to the classical polynomial cases of Meixner ($\alpha^2<0$), Meixner-Pollaczek ($\alpha^2>0$), and Laguerre ($\alpha^2=0$).

5.2. Principal series. The (unitary) principal series representation $D_p(\sigma)$, $\sigma \in \mathbb{R}$ has basis $\{\phi_m\}_{m=-\infty}^{\infty}$ and action

$$T_3\phi_m = \left(m+\frac{1}{2}\right)\phi_m,$$

$$T_1\phi_m = \frac{1}{2}([(m+1)^2+\sigma^2]^{1/2}\phi_{m+1} + [m^2+\sigma^2]^{1/2}\phi_{m-1}), \tag{5.9}$$

$$T_2\phi_m = \frac{1}{2i}([(m+1)^2+\sigma^2]^{1/2}\phi_{m+1} - [m^2+\sigma^2]^{1/2}\phi_{m-1}).$$

The representations $D_p(\sigma)$ and $D_p(-\sigma)$ are clearly equivalent.

Letting $e_m = \exp(im \arg(\alpha_1 - i\alpha_2))\phi_m$ and with T as in (5.2), we see that relative to the basis $\{e_m\}_{m=-\infty}^{\infty}$, the operator $T - \frac{1}{2}\alpha_3 I$ has a bilateral Jacobi matrix representation A, where

$$A_{m,m} = a_m = \alpha_3 m,$$
$$A_{m,m+1} = A_{m+1,m} = b_{m+1} = \tfrac{1}{2}[(m+1)^2+\sigma^2]^{1/2}\sqrt{\alpha_1^2+\alpha_2^2}. \tag{5.10}$$

In the notation of § 3 we then have

$$a = \frac{\alpha_1^2+\alpha_2^2}{4}, \quad b = 0, \quad c = \frac{\alpha_1^2+\alpha_2^2}{4}\sigma^2, \quad d = \alpha_3. \tag{5.11}$$

5.3. Complementary series. The complementary series (sometimes called supplementary series) representations $D_c(s, t)$ are parameterized by $s, t \in \mathbb{R}$ satisfying

$$-1+|t| < s < -|t|. \tag{5.12}$$

Relative to a basis $\{\phi_m\}_{m=-\infty}^{\infty}$ the action is given by

$$
\begin{aligned}
T_3\phi_m &= (t+m)\phi_m,\\
T_1\phi_m &= \frac{1}{2}([s+t+m)(s+t+m+1)]^{1/2}\phi_{m+1}\\
&\qquad +[(s+t+m)(-s+t+m-1)]^{1/2}\phi_{m-1}),\\
T_2\phi_m &= \frac{1}{2i}([(-s+t+m)(s+t+m+1)]^{1/2}\phi_{m+1}\\
&\qquad -[(s+t+m)(-s+t+m-1)]^{1/2}\phi_{m-1}).
\end{aligned}
\tag{5.13}
$$

Letting $e_m = \exp(im \arg(\alpha_1 - i\alpha_2))\phi_m$ and with T as in (5.2) we find that, relative to the basis $\{e_m\}_{m=-\infty}^{\infty}$, the operator $T-\alpha_3 tI$ has a bilateral Jacobi matrix representation A, with

$$
\begin{aligned}
&A_{m,m} = a_m = \alpha_3 m,\\
&A_{m,m+1} = A_{m+1,m} = b_{m+1} = \tfrac{1}{2}[(-s+t+m)(s+t+m+1)]^{1/2}\sqrt{\alpha_1^2+\alpha_2^2}.
\end{aligned}
\tag{5.14}
$$

In the notation of § 3 we now have

$$
\begin{aligned}
&a = \frac{\alpha_1^2+\alpha_2^2}{4}, \qquad b = \frac{\alpha_1^2+\alpha_2^2}{4}(2t-1),\\
&c = \frac{\alpha_1^2+\alpha_2^2}{4}(s+t)(t-s-1), \qquad d = \alpha_3.
\end{aligned}
\tag{5.15}
$$

Thus both the principal and complementary series cases correspond to the bilateral associated polynomial cases of § 3 of Meixner ($\alpha^2 = \alpha_1^2+\alpha_2^2-\alpha_3^2<0$), Meixner-Pollaczek ($\alpha^2>0$), and Laguerre ($\alpha^2=0$).

Other associated polynomial cases of Charlier, Hermite, and Bessel order related to representations of H_4 and T_3 can also be made explicit (compare [12] and [13]).

REFERENCES

[1] B. G. ADAMS, J. ČIŽEK, AND J. PALDUS, *Representation theory of* SO(4, 2) *for the perturbation treatment of hydrogenic-type Hamiltonians by algebraic methods*, Internat. J. Quantum Chemistry, 21 (1982), pp. 153-171.

[2] N. I. AKHIEZER, *The Classical Moment Problem*, Oliver and Boyd, Edinburgh, 1965.

[3] Y. ALHASSID, F. GÜRSEY, AND F. IACHELLO, *Group theory approach to scattering*, Ann. Physics, 148 (1983), pp. 346-380.

[4] ———, *Group theory approach to scattering* II. *The Euclidean connection*, Ann. Physics, 167 (1986), pp. 181-200.

[5] R. ASKEY AND J. WIMP, *Associated Laguerre and Hermite polynomials*, Proc. Roy. Soc. Edinburgh, Sect. A, 96 (1984), pp. 15-37.

[6] A. O. BARUT, A. INOMATA, AND R. WILSON, *The generalized Morse oscillator in the* SO(4, 2) *dynamical group scheme*, J. Math. Phys., 28 (1987), pp. 605-611.

[7] JU. M. BEREZANSKII, *Expansions in Eigenfunctions of Selfadjoint Operators*, American Mathematical Society, Providence, RI, 1968.

[8] A. ERDÉLYI, ED., *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[9] W. GAUTSCHI, *Computational aspects of three-term recurrence relations*, SIAM Rev., 9 (1967), pp. 24-82.

[10] M. E. H. ISMAIL, J. LETESSIER, D. R. MASSON, AND G. VALENT, *Birth and death processes and orthogonal polynomials*, in Proc. NATO Advanced Study Institute on Orthogonal Polynomials and their Applications, 1989, Orthogonal Polynomials: Theory and Practice, P. Nevai, ed., Kluwer, Dordrecht, the Netherlands, 1990, pp. 229-255.

[11] D. R. MASSON, *Schrödinger's equation and continued fractions*, Internat. J. Quantum Chemistry: Quantum Chemistry Symposium, 21 (1987), pp. 699–712.

[12] ———, *Difference equations, continued fractions, Jacobi matrices and orthogonal polynomials*, in Nonlinear Numerical Methods and Rational Approximation 1987, A. Cuyt, ed., D. Reidel, Dordrecht, the Netherlands, 1988, pp. 239–257.

[13] W. MILLER, JR., *On Lie algebras and some special functions of mathematical physics*, Mem. Amer. Math. Soc., 50 (1964), pp. i–43.

[14] P. C. OJHA, SO(2, 1) *Lie algebra and the Jacobi-matrix method for scattering*, Phys. Rev. A, 34 (1986), pp. 969–977.

[15] ———, *The Jacobi-matrix method in parabolic coordinates: expansion of Coulomb functions in parabolic Sturmians*, J. Math. Phys., 28 (1987), pp. 392–396.

[16] W. E. PRUITT, *Bilateral birth and death processes*, Ph.D. dissertation, Stanford University, Stanford, CA, 1960.

[17] ———, *Bilateral birth and death processes*, Trans. Amer. Math. Soc., 107 (1962), pp. 508–525.

[18] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics* II: *Fourier Analysis, Self-Adjointness*, Academic Press, New York, 1975.

[19] L. J. SLATER, *Confluent Hypergeometric Functions*, Cambridge University Press, Cambridge, 1960.

[20] M. H. STONE, *Linear Transformations in Hilbert Space*, American Mathematical Society, Providence, RI, 1932.

[21] H. S. WALL, *Analytic Theory of Continued Fractions*, Van Nostrand, Princeton, NJ, 1943.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1147–1155, July 1991

018

ORTHOGONAL FUNCTIONS FROM GRAM DETERMINANTS*

JAMES A. WILSON†

Abstract. Generalized Gram determinants are used to derive explicit hypergeometric series or basic hypergometric series formulas for orthogonal polynomials and biorthogonal rational functions. The computations involve some new determinant evaluations.

Key words. orthogonal polynomials, biorthogonal rational functions, hypergeometric series, basic hypergeometric series, Gram determinants

AMS(MOS) subject classifications. 33A65, 33A30

1. Introduction and example. It is well known that the orthogonal polynomials p_n for a given distribution $d\mu$ can be expressed as determinants

$$p_n(x) = \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_1 & \cdots & & \mu_{n+1} \\ \vdots & & & \vdots \\ \mu_{n-1} & \cdots & & \mu_{2n-1} \\ 1 & x & \cdots & x^n \end{vmatrix} \tag{1.1}$$

with $\mu_n = \int x^n d\mu(x)$. What is surprising is that many of the usual explicit formulas for orthogonal polynomials can be derived by directly evaluating determinants similar to this. We demonstrate here how simple determinant reductions lead naturally to the classical orthogonal polynomials and the related hypergeometric and basic hypergeometric orthogonal polynomials [1], even the ${}_4F_3$ polynomials in [7] and [8] and their basic analogues in [2]. In fact, the technique is the source of the ${}_9F_8$ and ${}_{10}\phi_9$ biorthogonal functions referred to in [5] and [6], the proof in §2 being the original derivation from my thesis [8].

Let $\{\phi_k\}_{k=0}^{\infty}$ and $\{\psi_k\}_{k=0}^{\infty}$ be sequences of polynomials, ϕ_k and ψ_k of exact degree k. Then an alternative to (1.1) is

$$p_n(x) = C \cdot \begin{vmatrix} \mu_{0,0} & \mu_{0,1} & \cdots & \mu_{0,n} \\ \mu_{1,0} & \cdots & & \mu_{1,n} \\ \vdots & & & \vdots \\ \mu_{n-1,0} & \cdots & & \mu_{n-1,n} \\ \phi_0(x) & \phi_1(x) & \cdots & \phi_n(x) \end{vmatrix} \tag{1.2}$$

with $\mu_{i,j} = \int \psi_i \phi_j d\mu$. Indeed,

$$\int p_n \psi_k d\mu = C \cdot \begin{vmatrix} \mu_{0,0} & \mu_{0,1} & \cdots & \mu_{0,n} \\ \mu_{1,0} & \cdots & & \vdots \\ \vdots & & & \\ \mu_{n-1,0} & \cdots & & \mu_{n-1,n} \\ \mu_{k,0} & \mu_{k,1} & \cdots & \mu_{k,n} \end{vmatrix}$$

and this is zero if $k < n$.

*Received by the editors July 28, 1990; accepted for publication August 27, 1990.
†Department of Mathematics, Iowa State University, Ames, Iowa 50011.

Here and elsewhere in the paper, C represents a factor which affects only the normalization of the orthogonal functions.

To illustrate the technique, we take $d\mu(x) = (1-x)^\alpha(1+x)^\beta\,dx; -1 < x < 1$; $\alpha, \beta > -1$; and derive the explicit series formula for the Jacobi polynomials $P_n^{(\alpha,\beta)}$. Start with the beta integral

$$\int d\mu(x) = \int_{-1}^{1}(1-x)^\alpha(1+x)^\beta\,dx = 2^{\alpha+\beta+1}\Gamma(\alpha+1)\Gamma(\beta+1)/\Gamma(\alpha+\beta+2).$$

We see that good choices for the basis polynomials are $\phi_k(x) = (1-x)^k$ and $\psi_k(x) = (1+x)^k$, since these combine with the weight function in a simple way:

$$\mu_{i,j} = \int \psi_i\phi_j d\mu = \int_{-1}^{1}(1-x)^{\alpha+j}(1+x)^{\beta+i}dx \tag{1.3}$$
$$= 2^{\alpha+\beta+1+i+j}\Gamma(\alpha+1+j)\Gamma(\beta+1+i)/\Gamma(\alpha+\beta+2+i+j).$$

(The computations which follow would be a little longer with $\phi_k(x) = \psi_k(x) = (1 \pm x)^k$.) In using (1.2), we multiply the columns and rows by appropriate factors to simplify the μ_{ij}'s and make 1's in the first column:

$$P_n^{(\alpha,\beta)} = C \cdot \begin{vmatrix} \tilde{\mu}_{0,0} & \tilde{\mu}_{0,1} & \cdots & \tilde{\mu}_{0,n} \\ \tilde{\mu}_{1,0} & \cdots & & \vdots \\ \vdots & & & \\ \tilde{\mu}_{n-1,0} & \cdots & & \tilde{\mu}_{n-1,n} \\ \tilde{\phi}_0 & \tilde{\phi}_1 & \cdots & \tilde{\phi}_n \end{vmatrix}$$

with $\tilde{\mu}_{ij} = 1/(\alpha+\beta+2+i)_j$ and $\tilde{\phi}_j(x) = (1-x)^j/2^j(\alpha+1)_j$. Now expand along the bottom row:

$$P_n^{(\alpha,\beta)} = C \cdot \sum_{k=0}^{n}(-1)^k\tilde{\phi}_k \cdot \det\left[\tilde{\mu}_{i,j}\right]_{\substack{0\le i\le n-1 \\ 0\le j\le n, j\ne k}}. \tag{1.4}$$

The problem now is to evaluate determinants of the form

$$\Delta(A,n,k) = \det\left[1/(A+i)_j\right]_{\substack{0\le i\le n-1 \\ 0\le j\le n, j\ne k}}$$

($\Delta(A,0,0) = 1$). Begin by making zeros in the first column by subtracting row $n-1$ from row n, row $n-2$ from row $n-1$, etc. Then expand along the first column to get

$$\Delta(A,n,k) = \det\left[\frac{1}{(A+i+1)_{j+1}} - \frac{1}{(A+i)_{j+1}}\right]_{\substack{0\le i\le n-2 \\ 0\le j\le n-1, j\ne k-1}}.$$

The general entry in this determinant simplifies to $-(j+1)/(A+i)_{j+2} = -(j+1)/(A+i)(A+i+1)(A+i+2)_j$, so that

$$\Delta(A,n,k) = \left[(-1)^{n-1}\prod_{\substack{j=0 \\ j\ne k}}^{n-1}(j+1) \Big/ \prod_{i=0}^{n-2}(A+i)(A+i+1)\right]\Delta(A+2,n-1,k-1)$$
$$= [(-1)^{n-1}n!/k(A)_{n-1}(A+1)_{n-1}]\,\Delta(A+2,n-1,k-1).$$

Iterate this k times:

$$(1.5)\quad \Delta(A,n,k) = \left\{\frac{1}{k!}\prod_{r=0}^{k-1}\frac{(n-r)!(-1)^{n-r-1}}{(A+2r)_{n-r-1}(A+2r+1)_{n-r-1}}\right\}\Delta(A+2k,n-k,0).$$

In the right-hand determinant, the index j runs from 1 to $n-k$. If we shift the index, we get

$$\begin{aligned}\Delta(A+2k,n-k,0) &= \det\left[1/(A+2k+i)_{j+1}\right]_{\substack{0\le i\le n-k-1\\ 0\le j\le n-k-1}}\\ &= \det\left[1/(A+2k+i)(A+2k+1+i)_j\right]_{\substack{0\le i\le n-k-1\\ 0\le j\le n-k-1}}\\ &= \Delta(A+2k+1,n-k,n-k)\ /\ (A+2k)_{n-k}.\end{aligned}$$

Finally, the last determinant can be evaluated by formula (1.5). These reductions give

$$(1.6)\qquad \Delta(A,n,k) = (\text{factor indep. of } k)\cdot(-1)^k(-n)_k(A+n-1)_k/k!$$

and from (1.4)

$$\begin{aligned}P_n^{(\alpha,\beta)}(x) &= C\cdot\sum_{k=0}^{n}(-1)^k\Delta(\alpha+\beta+2,n,k)\left(\frac{1-x}{2}\right)^k/(\alpha+1)_k\\ &= C\cdot\sum_{k=0}^{n}\frac{(-n)_k(n+\alpha+\beta+1)_k}{(\alpha+1)_k\ k!}\left(\frac{1-x}{2}\right)^k.\end{aligned}$$

This is a well-known and valuable representation of the Jacobi polynomials. (For the usual normalization, put $C=(\alpha+1)_n/n!$.)

The derivations of the explicit hypergeometric or basic hypergeometric formulas for other classical and related orthogonal polynomials differ little from this. Whether $\int d\mu$ is an ordinary integral or a finite or infinite sum, we choose the basis polynomials ϕ_k and ψ_k so that, as in (1.3), multiplying the weight function by $\psi_i\phi_j$ has the effect of shifting parameters. There are interesting determinant evaluations which arise in these computations, more general than $\Delta(A,n,k)$. Some are given explicitly in §3.

2. Rational function biorthogonalities. After seeing orthogonal polynomials derived in this way from series or integrals containing parameters, it is natural to try for something new using more general formulas. Dougall's theorem [3, p. 27] is

$$\begin{aligned}&{}_7F_6\left[\begin{matrix}2a,a+1,a+b,a+c,a+d,a+e,a+f;\ 1\\ a,a-b+1,a-c+1,a-d+1,a-e+1,a-f+1\end{matrix}\right]\\ (2.1)\qquad &= \sum_{k=0}^{N}w_k(a,b,c,d,e,f)\\ &= \frac{(2a+1)_N(1-c-d)_N(1-c-e)_N(1-d-e)_N}{(a-c+1)_N(a-d+1)_N(a-e+1)_N(b+f)_N},\end{aligned}$$

provided $a+b+c+d+e+f=1$ and $a+b=-N$. Here

$$(2.2)\quad w_k(a,b,c,d,e,f) = \frac{(2a)_k(a+k)(a+b)_k(a+c)_k(a+d)_k(a+e)_k(a+f)_k}{(1)_k\ a\ (a-b+1)_k\ \dots\ (a-f+1)_k}.$$

The simplest functions ϕ_j, ψ_i by which we can multiply w_k to shift parameters (while preserving the condition $a+b+c+d+e+f=1$) are not polynomials, but rational functions of a quadratic function of k. We choose

$$\begin{aligned}\phi_j(z^2) &= (c-z)_j(c+z)_j/(1-e+z)_j(1-e-z)_j,\\ \psi_i(z^2) &= (d-z)_i(d+z)_i/(1-f+z)_i(1-f-z)_i.\end{aligned}$$

Then

$$\begin{aligned}\mu_{ij} &= \sum_{k=0}^{N} w_k(a,b,c,d,e,f)\phi_j\left((a+k)^2\right)\psi_i\left((a+k)^2\right)\\ &= \frac{(c-a)_j(c+a)_j(d-a)_i(d+a)_i}{(1-e+a)_j(1-e-a)_j(1-f+a)_i(1-f-a)_i}\\ &\qquad\cdot\sum_{i=0}^{N} w_k(a,b,c+j,d+i,e-j,f-i)\\ &= \text{(factor indep. of } j)\cdot\frac{(a+c)_j(b+c)_j(c+d+i)_j(c+f-i)_j}{(1-e-a)_j(1-e-b)_j(1-d-e-i)_j(1-e-f+i)_j}\end{aligned}$$

by Dougall's theorem. We define for $0 \le n \le N$

$$R_n(z^2) = \begin{vmatrix} \mu_{0,0} & \mu_{0,1} & \cdots & \mu_{0,n} \\ \mu_{1,0} & & & \mu_{1,n} \\ \vdots & & & \vdots \\ \mu_{n-1,0} & \cdots & & \mu_{n-1,n} \\ \phi_0(z^2) & \cdots & & \phi_n(z^2) \end{vmatrix}, \tag{2.3}$$

$$S_n(z^2) = \begin{vmatrix} \mu_{0,0} & \cdots & \mu_{n,0} \\ \mu_{0,1} & \cdots & \mu_{n,1} \\ \vdots & & \vdots \\ \mu_{0,n-1} & \cdots & \mu_{n,n-1} \\ \psi_0(z^2) & \cdots & \psi_n(z^2) \end{vmatrix}$$

so that

$$\begin{aligned}&\sum_{k=0}^{N} w_k R_n\left((a+k)^2\right)\psi_m\left((a+k)^2\right)\\ &\qquad = \sum_{k=0}^{N} w_k S_n\left((a+k)^2\right)\phi_m\left((a+k)^2\right) = 0\end{aligned}$$

for $0 \le m < n \le N$, and we have a biorthogonality relation

$$\sum_{k=0}^{N} w_k R_m\left((a+k)^2\right) S_n\left((a+k)^2\right) = 0$$

for $m, n \in \{0, 1, \cdots, N\}, m \ne n$.

The hope is, of course, that the computation in §1 generalizes to give hypergeometric series formulas for the determinants (2.3).

Multiplying the columns and rows by appropriate factors and expanding along the last row, as in the Jacobi polynomial case, gives

$$\begin{aligned}R_n(z^2) &= C\cdot\sum_{k=0}^{n}\frac{(-1)^k\,(1-e-a)_k(1-e-b)_k}{(a+b)_k(b+c)_k}\\ &\qquad\cdot\phi_k(z^2)\Delta(c+d,c+f,1-e-f,1-e-d,n,k)\end{aligned} \tag{2.4}$$

where

$$\Delta(A,B,C,D,n,k) = \det\left[\frac{(A+i)_j(B-i)_j}{(C+i)_j(D-i)_j}\right]_{\substack{0\le i\le n-1\\ 0\le j\le n, j\ne k}} \tag{2.5}$$

for $0 \le k \le n$. ($\Delta(A,B,C,D,0,0) = 1$.)

We need to evaluate (2.5) when $A + D = B + C$. This time, subtracting rows leads to

$$\begin{aligned}\Delta(A,B,C,D,n,k) &= \det\Bigg[\frac{(A+i+1)_j(B-i)_j}{(C+i)_{j+2}(D-i-1)_{j+2}} \\ &\quad \{(A+i+j+1)(B-i-1)(C+i)(D-i+j) \\ &\qquad -(A+i)(B-i+j)(C+i+j+1)(D-i-1)\}\Bigg]_{\substack{0\le i\le n-2\\ 0\le j\le n-1, j\ne k-1}}.\end{aligned}$$

The expression in braces factors into $(j+1)(B+C+j)(A-C)(A-B+2i+1)$ when $A + D = B + C$. Because of this fortuitous factorization, the procedure used in §1 may be followed to the conclusion

(2.6)
$$\begin{aligned}&\Delta(A,B,C,D,n,k) \\ &= (\text{factor indep. of } k)\frac{(-1)^k(-n)_k(B+C-1)_k(B+C-1+2k)(C+n-1)_k(D)_k}{k!\,(A)_k(B-n+1)_k(B+C+n)_k}.\end{aligned}$$

Upon substituting the value of $\Delta(A,B,C,D,n,k)$ into (2.3), we find

(2.7)
$$R_n(z^2) = C \cdot {}_9F_8\left[\begin{matrix} c-e,(c-e)/2+1,c-z,c+z,1-e-a,1-e-b,1-e-d,n-e-f,-n;1 \\ (c-e)/2,1-e+z,1-e-z,c+a,c+b,c+d,c+f-n+1,c-e+n+1.\end{matrix}\right]$$

A ${}_9F_8$ of this special type satisfies Bailey's transformation identity [3, p. 27]:

(2.8)
$$\begin{aligned}&{}_9F_8\left[\begin{matrix} a,a/2+1,b,c,d,e,f,g,-n;\ 1 \\ a/2,a-b+1,\cdots,a-g+1,a+n+1\end{matrix}\right] \\ &= \frac{(a+1)_n(a'-e+1)_n(a'-f+1)_n(a'-g+1)_n}{(a'+1)_n(a-e+1)_n(a-f+1)_n(a-g+1)_n} \\ &\quad \cdot {}_9F_8\left[\begin{matrix} a',a'/2+1,a'-a+b,a'-a+c,a'-a+d,e,f,-n;1 \\ a'/2,a-b+1,a-c+1,a-d+1,a'-e+1,a'-g+1,a'+n+1\end{matrix}\right]\end{aligned}$$

with $a' = 2a - b - c - d + 1$, provided $b + c + d + e + f + g - n = 3a + 2$. In terms of the rational functions, Bailey's transformation says that $R_n(z^2)$ is symmetric is a, b, c, and d when we put $C = (c+a)_n(c+b)_n(c+d)_n(-c-f)_n/(c-e+1)_n$ in (2.7). In fact, an iterate of (2.8) gives

$$\begin{aligned}R_n(z^2) &= \frac{(a-z)_n(b-z)_n(c-z)_n(d-z)_n(z-f)_n}{(-2z)_n\,(1-e-z)_n} \\ &\cdot {}_9F_8\left[\begin{matrix} 2z-n,z-n/2+1,a+z,b+z,c+z,d+z,e-n+z,f-n+1+z,-n;\ 1 \\ z-n/2,z-a-n+1,z-b-n+1,z-c-n+1,z-d-n+1,z-e+1,z-f,2z+1\end{matrix}\right]\end{aligned}$$

(with the same choice for C). By symmetry, $S_n(z^2)$ is given by the same formula with e and f interchanged.

The biorthogonality just derived may be written

$$\sum_{k=0}^{N} w_k\, R_m\left((a+k)^2;a,b,c,d,e,f\right) R_n\left((a+k)^2;a,b,c,d,f,e\right) = \delta_{m,n}\cdot h_n \tag{2.9}$$

if $a+b+c+d+e+f=1$; $a+b=-N$, $m,n\in\{0,1,\cdots,N\}$, the weights w_k are given by (2.2), and

$$R_n(z^2) = \frac{(a+b)_n(a+c)_n(a+d)_n(-a-f)_n}{(a-e+1)_n} \cdot {}_9F_8\left[\begin{matrix} a-e,(a-e)/2+1,a-z,a+z,1-e-b,1-e-c,1-e-d,n-e-f,-n;1 \\ (a-e)/2,1-e+z,1-e-z,a+b,a+c,a+d,a+f-n+1,a-e+n+1 \end{matrix}\right].$$

Once the formula for $R_n(z^2)$ is given, direct verification of the biorthogonality is much easier than the derivation, and carrying this out is one way to find that

$$h_n = M\cdot\frac{n!(n-e-f)_n(a+b)_n(a+c)_n(a+d)_n(b+c)_n(b+d)_n(c+d)_n}{(1-e-f)_{2n}}$$

with

$$M = \frac{(2a+1)_N(1-c-d)_N(1-c-e)_N(1-d-e)_N}{(a-c+1)_N(a-d+1)_N(a-e+1)_N(b+f)_N}.$$

Interchanging a and b in (2.9) is equivalent to summing in the reverse order.

If we let e and f become infinite while preserving the condition $a+b+c+d+e+f=1$, the biorthogonality (2.9) reduces to a discrete case of the polynomial orthogonality relation for

$$p_n(z^2) = (a+b)_n(a+c)_n(a+d)_n\ {}_4F_3\left[\begin{matrix} -n,n+a+b+c+d-1,a-z,a+z \\ a+b,a+c,a+d \end{matrix};1\right]$$

($a+b=-N, 0\le n\le N$), which polynomials in turn contain the classical polynomials and others as limiting cases [1], [2], [4], [7], [8].

The rational functions $R_n(z^2)$ have basic hypergeometric analogues. We use the notation $(a;q)_k=(1-a)(1-aq)\cdots(1-aq^{k-1})$ if $k\ge 1$; $(a;q)_0=1$; and

$${}_{r+1}\phi_r\left[\begin{matrix} a_0,a_1,\cdots,a_r \\ b_1,\cdots,b_r \end{matrix};q,z\right] = \sum_{k=0}^{\infty}\frac{(a_0;q)_k\cdots(a_r;q)_k}{(q;q)_k(b_1;q)_k\cdots(b_r;q)_k}z^k.$$

The sums we encounter will actually be terminating ones, a_r being q^{-n} for some nonnegative integer n. Suppose $abcdef=q$, and for $n=0,1,\cdots$ let

(2.10)

$$r_n\left((z+z^{-1})/2;a,b,c,d,e,f;q\right) = \frac{(ab;q)_n(ac;q)_n(ad;q)_n(1/af;q)_n}{(aq/e;q)_n} \cdot {}_{10}\phi_9\left[\begin{matrix} a/e,q\sqrt{a/e},-q\sqrt{a/e},a/z,az,q/be,q/ce,q/de,q^n/ef,q^{-n} \\ \sqrt{a/e},-\sqrt{a/e},qz/e,q/ez,ab,ac,ad,q^{1-n}af,q^{n+1}a/e \end{matrix};q,q\right].$$

Then $r_n\left((z+z^{-1})/2\right)$ is a rational function of degree n/n in the variable $(z+z^{-1})/2$, since

$$\frac{(a/z;q)_k(az;q)_k}{(qz/e;q)_k(q/ez;q)_k} = \prod_{j=0}^{k-1}\frac{1-aq^j(z+z^{-1})+a^2q^{2j}}{1-q^{j+1}e^{-1}(z+z^{-1})+q^{2j+2}e^{-2}}$$

and the series terminates with the $k=n$ term. Replacing a,b,c,d,e,f,z by $q^a,\cdots,q^f$, q^z and then letting $q\to 1$ yields the ${}_9F_8$ functions $R_n(z^2)$. In the same way, the biorthogonality for $r_n\left((z+z^{-1})/2\right)$ contains relation (2.9). The basic analogue of

transformation (2.8) is [3, p. 68]

$$
{}_{10}\phi_9\left[\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, b, c, d, e, f, g, q^{-n}; q, q \\ \sqrt{a}, -\sqrt{a}, aq/b, aq/c, \cdots ; aq^{n+1} \end{matrix}\right]
$$

$$
= \frac{(aq;q)_n(aq/ef;q)_n(aq/eg;q)_n(aq/fg;q)_n}{(aq/e;q)_n(aq/f;q)_n(aq/g;q)_n(aq/efg;q)_n}
$$

$$
\cdot {}_{10}\phi_9\left[\begin{matrix} s, q\sqrt{s}, -q\sqrt{s}, sb/a, sc/a, sd/a, e, f, g, q^{-n}; q, q \\ \sqrt{s}, -\sqrt{s}, aq/b, aq/c, aq/d, sq/e, sq/f, sq/g, sq^{n+1} \end{matrix}\right]
$$

with $s = a^2q/bcd$, provided $a^3q^{n+2} = bcdefg$. This says that r_n is symmetric in a, b, c, and d, and an iterate of this transformation gives

$$
r_n\left((z+z^{-1})/2\right) = \frac{(a/z)_n(b/z)_n(c/z)_n(d/z)_n(z/f)_n}{(z^{-2})_n(q/ez)_n}
$$

$$
\cdot {}_{10}\phi_9\left[\begin{matrix} z^2q^{-n}, zq^{1-n/2}, -zq^{1-n/2}, az, bz, cz, dz, eq^{-n}z, fq^{-n+1}z, q^{-n}; q, q \\ zq^{-n/2}, -zq^{-n/2}, zq^{1-n}/a, zq^{1-n}/b, zq^{1-n}/c, zq^{1-n}/d, zq/e, z/f, z^2q \end{matrix}\right].
$$

In place of Dougall's theorem (2.1) we need Jackson's theorem [3, p. 67]:

$$
{}_8\phi_7\left[\begin{matrix} a^2, aq, -aq, ab, ac, ad, ae, af; q, q \\ a, -a, aq/b, aq/c, aq/d, aq/e, aq/f \end{matrix}\right] = \frac{(a^2q;q)_N(q/cd;q)_N(q/ce;q)_N(q/de;q)_N}{(aq/c;q)_N(aq/d;q)_N(aq/e;q)_N(bf;q)_N},
$$

provided $abcdef = q$ and $ab = q^{-N}$. Mimicking the derivation in §3 leads to a biorthogonality relation for $r_n\left((z+z^{-1})/2\right)$ on the $N+1$ points $(z+z^{-1})/2 = (a+a^{-1})/2, (aq+a^{-1}q^{-1})/2, \cdots, (aq^N+a^{-1}q^{-N})/2$. The determinant formulas produced in the computation are given in the next section.

The biorthogonality is again easier to verify directly:

(2.11)

$$
\sum_{k=0}^{N} w_k r_n\left(\frac{aq^k+a^{-1}q^{-k}}{2}; a, b, c, d, e, f; q\right) r_m\left(\frac{aq^k+a^{-1}q^{-k}}{2}; a, b, c, d, f, e; q\right)
$$

$$
= M \cdot \delta_{m,n} \frac{(q;q)_n(q^n/ef;q)_n(ab;q)_n(ac;q)_n(ad;q)_n(bc;q)_n(bd;q)_n(cd;q)_n}{(q/ef;q)_{2n}}
$$

with $abcdef = q$; $ab = q^{-N}$; $m, n \in \{0, 1, \cdots, N\}$;

$$
w_k = \frac{(a^2;q)_k(1-a^2q^{2k})(ab;q)_k(ac;q)_k(ad;q)_k(ae;q)_k(af;q)_k q^k}{(q;q)_k(1-a^2)(aq/b;q)_k(aq/c;q)_k(aq/d;q)_k(aq/e;q)_k(aq/f;q)_k}
$$

and

$$
M = \frac{(aq;q)_N(q/cd;q)_N(q/ce;q)_N(q/de;q)_N}{(aq/c;q)_N(aq/d;q)_N(aq/e;q)_N(bf;q)_N}.
$$

As in (2.9), interchanging a and b in (2.11) is equivalent to summing in reverse order. If we let $e \to 0$ and $f \to \infty$ with $abcdef = q$, this relation reduces to the discrete case of the orthogonality relation for the ${}_4\phi_3$ polynomials in [2].

3. The determinant formulas. New determinant formulas may have value in a variety of contexts. The determinants involved in the preceding computations seem particularly promising because of their intimate connection with hypergeometric and basic hypergeometric series. Therefore, it is worthwhile to record the complete evaluation formulas.

The determinants involved with the ${}_{10}\phi_9$'s are

$$\Delta_{n,k} = \det[(Aq^i;q)_j(Bq^{-i};q)_j/(Cq^i;q)_j(Dq^{-i};q)_j]_{\substack{0\le i\le n-1\\ 0\le j\le n, j\ne k}} \tag{3.1}$$

with $AD = BC$, $0 \le k \le n$. ($\Delta_{0,0} = 1$.) For these, we have

$$\Delta_{n,n} = \prod_{r=0}^{n-1} \frac{q^{-r}(q;q)_r D^r (1-AC^{-1}q^{-n+1+r})^r (AB^{-1}q^{n-r};q^2)_r (ADq^{n-1-r};q^2)_r}{(Cq^r;q)_{n-1}(Dq^{-r};q)_{n-1}}$$

$$= \prod_{r=1}^{n-1} \frac{q^{-r(r+1)/2}(q;q)_r(-B)^r(CA^{-1};q)_r(AB^{-1}q^{n-r};q^2)_r(ADq^{n-1-r};q^2)_r}{(Cq^{n-1-r};q)_{2r}(Dq^{-r};q)_{2r}}$$

(we have given two of many ways to group the factors) and

$$\Delta_{n,k} = \Delta_{n,n} \cdot \pi_k/\pi_n$$

with

$$\pi_k = \frac{(ADq^{-1};q)_k(1-ADq^{2k-1})(D;q)_k(Cq^{n-1};q)_k(q^{-n};q)_k}{(q;q)_k(1-ADq^{-1})(A;q)_k(Bq^{-n+1};q)_k(ADq^n;q)_k}(-q)^k.$$

All the similar generalized Gram determinants we have evaluated are limiting cases of this one. We now give some examples. If A, B, C, D are replaced by q^A, q^B, q^C, q^D, then as $q \to 1$, we get the determinants involved in the $R_n(z^2)$ computation:

$$\Delta_{n,k} = \det\left[\frac{(A+i)_j(B-i)_j}{(C+i)_j(D-i)_j}\right]_{\substack{0\le i\le n-1\\ 0\le j\le n, j\ne k}}, \tag{3.2}$$

$A + D = B + C$. For these, we have

$$\Delta_{n,k} = \Delta_{n,n}\pi_k/\pi_n,$$

$$\Delta_{n,n} = \prod_{r=0}^{n-1} \frac{(-1)^r r!(C-A)_r 2^{2r}\left((A-B+n-r)/2\right)_r\left((A+D+n-1-r)/2\right)_r}{(C+r)_{n-1}(D-r)_{n-1}},$$

$$\pi_k = \frac{(A+D-1)_k(A+D-1+2k)(D)_k(C+n-1)_k(-n)_k(-1)^k}{k!(A+D-1)(A)_k(B-n+1)_k(A+D+n)_k}.$$

It is interesting that, with a change of variable, these formulas evaluate the more symmetric determinants with entries $(a;q)_{i+j}(b;q)_{i'+j}(c;q)_{i+j'}(d;q)_{i'+j'}$ if $ad = bc$, or $(a)_{i+j}\ (b)_{i'+j}\ (c)_{i+j'}\ (d)_{i'+j'}$ if $a + d = b + c$, $i' = n-1-i$, $j' = n-1-j$.

In the limit as $B, D \to \infty$, (3.2) becomes

$$\begin{aligned}\Delta_{n,k} &= \det[(A+i)_j/(C+i)_j]_{\substack{0\le i\le n-1\\ 0\le j\le n, j\ne k}} \\ &= \Delta_{n,n}\pi_k/\pi_n, \\ \Delta_{n,n} &= \prod_{r=0}^{n-1} r!(C-A)_r/(C+r)_{n-1}, \\ \pi_k &= (C+n-1)_k(-n)_k(-1)^k/k!(A)k.\end{aligned} \tag{3.3}$$

The basic analogue is ($B, D \to 0$ in (3.1))

$$\begin{aligned}\Delta_{n,k} &= \det[(Aq^i;q)_j/(Cq^i;q)_j]_{\substack{0\le i\le n-1\\ 0\le j\le n, j\ne k}}\\ &= \Delta_{n,n}\cdot \pi_k/\pi_n,\\ \Delta_{n,n} &= \prod_{r=0}^{n-1}(q;q)_r(C/A;q)_r A^r q^{r(r-1)}/(Cq^r;q)_{n-1},\\ \pi_k &= (Cq^{n-1};q)_k(q^{-n};q)_k(-q)^k/(q;q)_k(A;q)_k.\end{aligned}$$

The symmetrized versions of these determinants have entries $(a)_{i+j}(d)_{i'+j'}$ and $(a;q)_{i+j}$ $(d;q)_{i'+j'}q^{-ij}$.

Other limiting cases are the determinants with entries $(A+i)_j/(D-i)_j$ or $(B-i)_j/(C+i)_j$, or their basic analogues.

The entries in the determinants become rational functions of i and j (or of q^i and q^j in the basic versions) when $A-C$ and $B-D$ are integers (or when A/C and B/D are integer powers of q). For example, with $C = A+1$ in (3.3), we can calculate

$$\det[1/(A+i+j)]_{\substack{0\le i\le n-1\\ 0\le j\le n-1}} = \prod_{r=0}^{n-1}(r!)^2/(A+r)_n.$$

REFERENCES

[1] G. E. ANDREWS AND R. ASKEY, *Classical orthogonal polynomials*, in Polynômes orthogonaux et applications, C. Brezinski et al., eds., Lecture Notes in Math. 1171, Springer-Verlag, Berlin, New York, 1985.

[2] R. ASKEY AND J. A. WILSON, Some basic hypergeometric orthogonal polyomials that generalize Jacobi polynomials, Mem. Amer. Math. Soc., 319, Providence, RI, 1985.

[3] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, 1935.

[4] J. LABELLE, *Tableau d'Askey*, in Polynômes orthogonaux et applications, C. Brezinski et al., eds., Lecture Notes in Math. 1171, Springer-Verlag, Berlin, New York, 1985.

[5] M. RAHMAN, *Families of biorthogonal rational functions in a discrete variable*, SIAM J. Math. Anal., 12 (1981), pp. 355–367.

[6] ———, *An integral representation of a* ${}_{10}\phi_9$ *and continuous biorthogonal* ${}_{10}\phi_9$ *rational functions*, Canadian J. Math., 38 (1986), pp. 605–618.

[7] J. A. WILSON, *Hypergeometric series recurrence relations and some new orthogonal functions*, Ph.D. thesis, University of Wisconsin, Madison, WI, 1978.

[8] ———, *Some hypergeometric orthogonal polynomials*, SIAM J. Math. Anal., 4 (1980), pp. 690–701.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1156-1172, July 1991

019

CONVERGENT FACTORIAL SERIES EXPANSIONS FOR BESSEL FUNCTIONS*

T. M. DUNSTER† AND D. A. LUTZ†

Abstract. Solutions of Bessel's equation and the modified Bessel equation are examined where the argument $z \to \infty$. Convergent series representations are derived for all of the standard Bessel and modified Bessel functions. These representations involve (inverse) factorial series of the form

$$\sum_{m=0}^{\infty} \frac{c_m(\nu) m!}{z(z+1)\cdots(z+m)},$$

and are uniformly and absolutely convergent in the half plane $\mathrm{Re}\,(z) \geqq \varepsilon > 0$ (ε arbitrary). Error bounds, explicit in the Bessel parameter ν, are derived for the difference between the infinite series and the truncated series.

It is shown that the factorial coefficients have the asymptotic behavior $c_m(\nu) = O(\{\ln(m)\}^{\nu-5/2}/m)$ as $m \to \infty$. The coefficients are also shown to satisfy certain recursion relations, which provide a means of calculating them in a numerically satisfactory manner.

Key words. Bessel functions, asymptotic expansions

AMS(MOS) subject classifications. 33A40, 34E05

1. Introduction. A general theory for factorial series expansions of solutions of linear differential equations has long been available, originating with the work of Horn. However, it does not seem to be widely known, nor used often in practice. This is perhaps partly due to the fact that while the general theory implies the convergence of a factorial series (in some half plane), it does not immediately provide all the information needed to analyze the convergence. Moreover, it does not provide a means of determining the order of magnitude of the factorial coefficients in a satisfactory manner.

The purpose of this paper is to investigate factorial series expansions for Bessel functions. We shall determine factorial series expansions for all the standard Bessel functions (§ 2), each of which converges in a half plane. We will also provide numerically satisfactory procedures for calculating the coefficients in these factorial series expansions (§§ 2, 3), give error bounds for the remainder, including explicit dependence on the parameter ν (§ 4), and determine the asymptotic behavior of the coefficients (§ 5).

A natural question regarding these factorial series expansions would be to ask how good the resulting convergence is for various fixed values of z, especially with respect to the parameter ν. We shall investigate questions such as these and, as a result, learn enough about the calculation and convergence of the factorial series so that their numerical effectiveness can be ascertained. We compare calculations based on factorial series with other standard types of convergent and asymptotic expansions in § 6.

The procedure we follow is based on the general theory (see, for example, [3] or [7]) that goes back to Horn and Norlünd, in which the Laplace transform plays a key role. Other types of convergent expansions for Bessel functions have been considered. J. Hadamard gave an expansion for the modified Bessel equation $I_\nu(z)$ which is particularly useful (see [8, pp. 204-205]). Hadamard's series representations for the other Bessel functions, while convergent, determine the solutions only to within the addition of an exponentially small error term.

* Received by the editors October 27, 1989; accepted for publication (in revised form) September 4, 1990.
† Department of Mathematics, San Diego State University, San Diego, California 92182-0314.

In a relatively recent paper Rosser [6] discussed a third type of series expansion, for the modified Bessel function $K_\nu(z)$, which involves quotients of gamma functions (see (5.17) below). The coefficients in such an expansion, which turn out to be somewhat more elusive than the factorial series coefficients, have been conjectured by Rosser to have a similar rate of growth. Using the same kind of argument as in § 5, we are also able to prove Rosser's conjecture, including the form of their complete asymptotic expansion.

2. Factorial representations for unmodified Bessel functions. In this section we examine solutions of Bessel's equation

$$\frac{d^2w}{dz^2}+\frac{1}{z}\frac{dw}{dz}+\left(1-\frac{\nu^2}{z^2}\right)w=0. \tag{2.1}$$

We assume throughout that ν is real and nonnegative, and furthermore is not equal to $\frac{1}{2}$ (in which case the solutions can be expressed in terms of elementary functions). Extensions of the following results to complex ν are feasible, but we shall not pursue this.

We shall derive a factorial series expansion for $H_\nu^{(1)}(z)$, and from this we derive corresponding results for the Bessel functions $H_\nu^{(2)}(z)$, $J_\nu(z)$, and $Y_\nu(z)$. We seek a factorial series representation for $u(\nu, z)$, defined by

$$H_\nu^{(1)}(z)=\left(\frac{2}{\pi z}\right)^{1/2} e^{-\nu\pi i/2-\pi i/4}\, e^{iz}[1+u(\nu, z)]. \tag{2.2}$$

Note that $u(\nu, z)$ possesses the following asymptotic expansion as $z\to\infty$ (e.g., see [5, p. 238]):

$$u(\nu, z)\sim\sum_{s=1}^{\infty} i^s\frac{A_s(\nu)}{z^s} \qquad (-\pi+\delta\leqq\arg z\leqq 2\pi-\delta), \tag{2.3a}$$

where

$$A_s(\nu)=\frac{(4\nu^2-1^2)(4\nu^2-3^2)\cdots(4\nu^2-(2s-1)^2)}{s!8^s}. \tag{2.3b}$$

Following Horn, consider the function $f(\nu, t)$ defined implicitly by

$$u(\nu, z)=\int_0^\infty e^{-zt}f(\nu, t)\,dt. \tag{2.4}$$

We make the following assumptions on $f(\nu, t)$ a priori, and these will be justified below. We assume

$$\lim_{t\to\infty}(tf(\nu, t))'\,e^{-\kappa t}=\lim_{t\to\infty} tf(\nu, t)\,e^{-\kappa t}=0, \qquad \text{arbitrary } \kappa>0. \tag{2.5}$$

We shall also need the following information, which follows from (2.3) and (2.4) and integration by parts:

$$f(\nu, 0)=iA_1(\nu)=i\sigma/2, \qquad \sigma=\nu^2-\tfrac{1}{4} \quad (\neq 0). \tag{2.6}$$

Here we have introduced a new parameter σ for later convenience.

From (2.4)-(2.6), and the following inhomogeneous differential equation for $u(\nu, z)$

$$z^2\frac{d^2u}{dz^2}+2iz^2\frac{du}{dz}-\sigma u=\sigma, \tag{2.7}$$

we find that $f(\nu, t)$ satisfies

$$t(t-2i)\frac{d^2f}{dt^2}+4(t-i)\frac{df}{dt}+(2-\sigma)f=0. \tag{2.8}$$

This is a form of the hypergeometric equation. From (2.6) we see that

$$f(\nu, t)=\frac{i\sigma}{2}F\left(\frac{3}{2}+\nu, \frac{3}{2}-\nu; 2; -\frac{it}{2}\right), \tag{2.9}$$

with the usual notation. This representation is very closely related to some well-known integral transforms (e.g., see [2, p. 212, eqs. (1), (4)]).

The hypergeometric function $f(\nu, t)$ given by (2.9) has a simple pole at $t=2i$ (see [5, p. 166, eq. (10.11)]). Its behavior at the regular singularity at infinity is given by (see [5, p. 167])

$$f(\nu, t)\sim a_0(\nu)t^{\nu-3/2}, \tag{2.10a}$$

where

$$a_0(\nu)=\begin{cases}\dfrac{\sigma\pi 2^{1/2-\nu}\, e^{(2\nu-1)\pi i/4}}{\sin(2\pi\nu)\Gamma(\frac{3}{2}+\nu)\Gamma(\frac{1}{2}+\nu)\Gamma(1-2\nu)} & (2\nu\neq 0, 1, 2, \cdots), \\ \dfrac{\sigma\Gamma(2\nu)2^{1/2-\nu}\, e^{(2\nu-1)\pi i/4}}{\Gamma(\frac{3}{2}+\nu)\Gamma(\frac{1}{2}+\nu)} & (2\nu=1, 2, 3, \cdots),\end{cases} \tag{2.10b}$$

and

$$f(0, t)=O(t^{-3/2}\ln(t)). \tag{2.10c}$$

The cut associated with the branchpoint at infinity runs along the imaginary axis from $t=i\infty$ to $t=2i$.

To derive a factorial series expansion formally, use

$$\int_0^\infty (1-e^{-t})^m e^{-zt}\,dt=\frac{m!}{z(z+1)\cdots(z+m)}, \qquad (m=0, 1, 2, \cdots), \tag{2.11}$$

and expand $f(\nu, t)$ in terms of a new independent variable $\xi(t)$ defined by

$$\xi=1-e^{-t}. \tag{2.12}$$

A region T in the t plane bounded by the parametric curve $t=-\ln(2\cos(\theta/2))+i\theta/2$ $(-\pi<\theta<\pi)$ is mapped to the unit disk $|\xi|\leqq 1$ in the ξ plane; see Fig. 1. The singularities $t=0, 2i, \infty$ are mapped to $\xi=0, 1-e^{-2i}, 1$, respectively. For this section we confine our attention to t lying in T, or correspondingly, ξ lying in the unit disk $|\xi|\leqq 1$.

Regarding $f(\nu, t)$ as a function of ξ, i.e.,

$$f(\nu, t)=f(\nu, -\ln(1-\xi))\equiv\phi(\nu, \xi), \tag{2.13}$$

it follows that $\phi(\nu, \xi)$ is analytic throughout $|\xi|\leqq 1$ except at $\xi=1$, where it has a logarithmic singularity:

$$\phi(\nu, \xi)=O([-\ln(1-\xi)]^{\nu-3/2}), \qquad \xi\to 1^-. \tag{2.14}$$

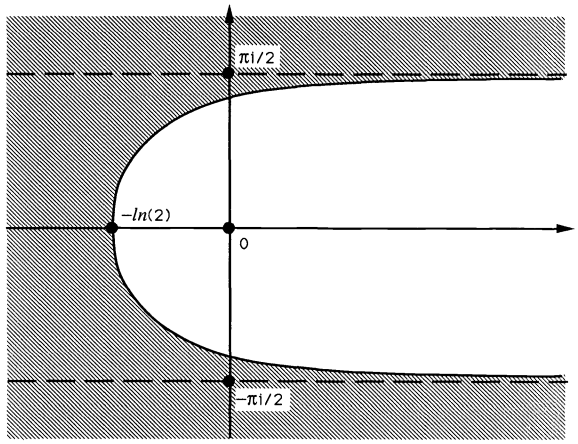


FIG. 1. *Domain T in t plane.*

Therefore, $\phi(\nu, \xi)$ possesses a Maclaurin expansion of the form

$$f(\nu, t) \equiv \phi(\nu, \xi) = \sum_{m=0}^{\infty} c_m(\nu)\xi^m, \tag{2.15}$$

which is convergent for $|\xi| < 1$. On substitution of (2.15) into (2.4), interchanging the summation and integral, and then employing (2.11), we formally arrive at the factorial series representation

$$u(\nu, z) = \sum_{m=0}^{\infty} \frac{c_m(\nu)m!}{z(z+1)\cdots(z+m)}. \tag{2.16}$$

For a justification of this formal process see § 4.

A necessary and sufficient condition for absolute convergence in a half plane $\mathrm{Re}\,(z) \geqq \kappa$ is that the coefficients satisfy

$$c_m(\nu) = O(m^{\kappa-1})$$

as $m \to \infty$; see [3] or [7]. In § 5 we shall show that $c_m(\nu) = o(m^{\varepsilon-1})$ as $m \to \infty$ for arbitrary positive ε. This implies uniform absolute convergence of the series for $\mathrm{Re}\,(z) \geqq \varepsilon > 0$.

One way of determining the Maclaurin coefficients is to employ the differential equation for $\phi(\nu, \xi)$ together with the initial condition $\phi(\nu, 0) = i\sigma/2$. From (2.8) and (2.12) we readily find that the equation is given by

$$p(\xi)\phi_{\xi\xi} + q(\xi)\phi_{\xi} + (2-\sigma)\phi = 0, \tag{2.17}$$

where

$$p(\xi) = (\xi-1)^2 t(t-2i), \qquad q(\xi) = (t^2 - 2it - 4t + 4i)(\xi - 1), \tag{2.18}$$

with $t = -\ln(1-\xi)$. Clearly $p(\xi)$ and $q(\xi)$ possess Maclaurin expansions of the form

$$p(\xi) = \sum_{r=1}^{\infty} p_r \xi^r, \quad q(\xi) = \sum_{r=0}^{\infty} q_r \xi^r, \quad |\xi| < 1.$$

Substituting these series and (2.15) into (2.17) we obtain

$$c_0(\nu)=\frac{i\sigma}{2}, \qquad c_1(\nu)=\frac{\sigma(2-\sigma)}{8}, \tag{2.19a, b}$$

$$\begin{aligned} c_{m+1}(\nu)=&\frac{1}{2i(m+1)(m+2)} \\ &\cdot\left\{q_m c_1(\nu)+\sum_{j=2}^{m}\{jc_j(\nu)[(j-1)p_{m+2-j}+q_{m+1-j}]\}+(2-\sigma)c_m(\nu)\right\} \\ &(m=1,2,\cdots), \end{aligned} \tag{2.20}$$

with the understanding that the summation in (2.20) is null for $m=1$.

Explicit formulas for the coefficients $\{p_r\}$ and $\{q_r\}$ can be derived by differentiating the expression for $p(\xi)$ r-times, and then using Taylor's formula $p_r=p^{(r)}(0)/r!$; the same holds for $q(\xi)$. After simplification using the binomial theorem, we find

$$\begin{aligned} &p_1=-2i, \quad p_2=1+3i, \quad p_3=-1-\frac{2}{3}i, \\ &p_r=\frac{1}{r(r-1)(r-2)}\left\{4\sum_{j=1}^{r-3}\frac{1}{j}-6-4i\right\} \qquad (r\geqq 4); \end{aligned} \tag{2.21}$$

$$\begin{aligned} &q_0=-4i, \quad q_1=4+6i, \quad q_2=-3-i, \\ &q_r=\frac{1}{r(r-1)}\left\{2\sum_{j=1}^{r-2}\frac{1}{j}-6-2i\right\} \qquad (r\geqq 3). \end{aligned} \tag{2.22}$$

The above recursion scheme provides a simple and numerically satisfactory means of evaluating the factorial coefficients $\{c_m(\nu)\}$ $(m=0,1,2,\cdots)$.

The factorial coefficients can also be expressed explicitly in terms of Stirling numbers of the first kind $S_n^{(m)}$ and the coefficients $\{A_s(\nu)\}$ in the asymptotic expansion (2.3) of the Hankel function. Following [7, pp. 329–330] we find

$$c_0(\nu)=iA_1(\nu), \qquad c_m(\nu)=\frac{(-1)^m i}{m!}\sum_{j=1}^{m}(-i)^j A_{j+1}(\nu)S_m^{(j)} \qquad (m=1,2,\cdots). \tag{2.23}$$

These formulas are more difficult to implement in numerical calculations as they require calculation of large individual terms to a high degree of accuracy. Another procedure for recursively calculating the coefficients can be derived by substituting the expansion (2.16) into equation (2.7) and using known formulas for derivatives of factorial series.

Finally, to obtain factorial series representations for the other standard Bessel function we use the well-known relations

$$H_\nu^{(1)}(z)=\overline{H_{\bar\nu}^{(2)}(\bar z)}, \qquad J_\nu(z)=\frac{1}{2}\{H_\nu^{(1)}(z)+H_\nu^{(2)}(z)\},$$

$$Y_\nu(z)=\frac{1}{2i}\{H_\nu^{(1)}(z)-H_\nu^{(2)}(z)\}.$$

We find

$$H_\nu^{(2)}(z)=\left(\frac{2}{\pi z}\right)^{1/2} e^{\nu\pi i/2+\pi i/4}\, e^{-iz}\left[1+\sum_{m=0}^{\infty}\frac{\overline{c_m(\nu)}m!}{z(z+1)\cdots(z+m)}\right], \tag{2.24}$$

$$\begin{aligned} J_\nu(z)=\left(\frac{2}{\pi z}\right)^{1/2}&\left\{\cos(z-\nu\pi/2-\pi/4)\left[1+\sum_{m=1}^{\infty}\frac{\mathrm{Re}\{c_m(\nu)\}m!}{z(z+1)\cdots(z+m)}\right]\right. \\ &\left.-\sin(z-\nu\pi/2-\pi/4)\sum_{m=0}^{\infty}\frac{\mathrm{Im}\{c_m(\nu)\}m!}{z(z+1)\cdots(z+m)}\right\}, \end{aligned} \tag{2.25}$$

$$Y_\nu(z)=\left(\frac{2}{\pi z}\right)^{1/2}\left\{\sin(z-\nu\pi/2-\pi/4)\left[1+\sum_{m=1}^{\infty}\frac{\mathrm{Re}\{c_m(\nu)\}m!}{z(z+1)\cdots(z+m)}\right]\right.$$
$$\left.+\cos(z-\nu\pi/2-\pi/4)\sum_{m=0}^{\infty}\frac{\mathrm{Im}\{c_m(\nu)\}m!}{z(z+1)\cdots(z+m)}\right\}. \tag{2.26}$$

3. Factorial representations for modified Bessel functions. The standard solutions of the modified Bessel equation

$$\frac{d^2w}{dz^2}+\frac{1}{z}\frac{dw}{dz}-\left(1+\frac{\nu^2}{z^2}\right)w=0, \tag{3.1}$$

are the modified Bessel functions $K_\nu(z)$ and $I_\nu(z)$. The purpose of this section is to derive factorial series expansions for these two functions which are uniformly and absolutely convergent in the half plane $\mathrm{Re}(z)\geqq\varepsilon>0$.

Consider $K_\nu(z)$ first. The derivation of a factorial series expansion for this function follows in a similar manner to that of the Hankel functions. We define $\tilde{u}(\nu, z)$ by

$$K_\nu(z)=\left(\frac{\pi}{2z}\right)^{1/2}e^{-z}[1+\tilde{u}(\nu, z)]. \tag{3.2}$$

Proceeding in a manner similar to the previous section we find that

$$\tilde{u}(\nu, z)=\int_0^\infty e^{-zt}\tilde{f}(\nu, t)\,dt, \tag{3.3}$$

where

$$\tilde{f}(\nu, t)=\frac{\sigma}{2}F\left(\frac{3}{2}+\nu, \frac{3}{2}-\nu; 2; -\frac{t}{2}\right). \tag{3.4}$$

The hypergeometric function $\tilde{f}(\nu, t)$ has a simple pole at $t=-2$ and a regular singularity at infinity. It therefore has no finite singularities inside the region T (defined in § 2). Thus, proceeding as before, we find that

$$\tilde{u}(\nu, z)=\sum_{m=0}^{\infty}\frac{d_m(\nu)m!}{z(z+1)\cdots(z+m)}. \tag{3.5}$$

The factorial coefficients $\{d_m(\nu)\}$ are found to be real, with $d_0(\nu)=-ic_0(\nu)$, $d_1(\nu)=-c_1(\nu)$, and the other coefficients satisfying the recursion relations (2.20)-(2.22) with i replaced by -1 in each equation.

We now turn our attention to the modified Bessel function $I_\nu(z)$. Unlike $H_\nu^{(1)}(z)$, $H_\nu^{(2)}(z)$, and $K_\nu(z)$, this function is dominant at infinity for all values of $\arg(z)$. The dominance makes the construction of a factorial series representation for it less straightforward than for the other three Bessel functions. We shall use the following representation:

$$I_\nu(z)=-\frac{i}{\pi}\{K_\nu(z\,e^{-\pi i})-e^{\nu\pi i}K_\nu(z)\}, \tag{3.6}$$

seeking a factorial series expansion for the modified Bessel function $K_\nu(z\,e^{-\pi i})$. Clearly, $K_\nu(z\,e^{-\pi i})$ is recessive at infinity in the sector $\pi/2<\arg(z)<3\pi/2$.

We next define $u^+(\nu, z)$ by

$$K_\nu(z\,e^{-\pi i})=i\left(\frac{\pi}{2z}\right)^{1/2}e^{z}[1+u^+(\nu, z)]. \tag{3.7}$$

Again, we could write $u^+(\nu, z)$ as a Laplace transform

$$u^+(\nu, z) = \int_\Gamma e^{-zt} f(\nu, t)\, dt \qquad (\mathrm{Re}\,(z) \geqq \kappa > 0), \tag{3.8}$$

where the path of integration Γ runs from $t = 0$ to infinity such that $\mathrm{Re}\,(zt) \to +\infty$. For any choice of Γ satisfying these requirements we find that $f(\mu, t)$ is a hypergeometric function which is regular at $t = 0$, but has a simple pole on the real axis at $t = 2$, which lies inside the region T. The presence of this singularity prevents us from proceeding exactly as before. If the pole were lying inside T, but not on the real axis, an appropriate (real) scaling of the variables t and z could reduce the problem to an equivalent one where the pole lies outside T (see [7, p. 326]).

Also we do not have enough information on $u^+(\nu, z)$ to determine the precise path of integration Γ in (3.8). This is because Γ can be taken to lie above or below the pole at $t = 2$ (the integrand remains unchanged in both circumstances). The two choices of Γ result in two different functions for the right-hand side, which differ by the residue of the integrand at $t = 2$. It is seen that this residue is $O(e^{-2z})$ as $\mathrm{Re}\,(z) \to \infty$, and therefore the asymptotic behavior of the integral at $z = \infty$ will be identical for either path. (The path Γ could also be chosen to loop the pole a number of times, or pass through it as a Cauchy principal value.)

The above difficulties arise from the fact that $K_\nu(z\, e^{-\pi i})$ is dominant in the half plane $|\arg(z)| < \pi/2$. In order to overcome the difficulty we seek a factorial series representation for $K_\nu(z\, e^{-\pi i})$ in the half plane $\mathrm{Re}\,(z\, e^{-\pi i/4}) \geqq \varepsilon > 0$. To do this we introduce the scaling factor

$$\omega = \frac{\pi}{4} + i\frac{\pi}{4} = \frac{\pi}{2\sqrt{2}}\, e^{i\pi/4}, \tag{3.9}$$

and seek a factorial series in terms of the variable $z^+ = z/\omega$. In place of (3.8) we therefore define

$$u^+(\nu, z) = \int_0^\infty e^{-z^+ t} f^+(\nu, t)\, dt. \tag{3.10}$$

where the path of integration is along the real axis. The pole of $f^+(\nu, t)$ is then found to be at $t = 2\omega$, which lies outside the region T. The modulus of ω was chosen to ensure this. Then, proceeding as before, we find that

$$u^+(\nu, z) = \sum_{m=0}^{\infty} \frac{d_m^+(\nu) m!}{z(z/\omega + 1) \cdots (z/\omega + m)}, \tag{3.11}$$

where

$$d_0^+(\nu) = -\frac{\sigma}{2}, \qquad d_1^+(\nu) = \frac{\sigma(\sigma - 2)}{8\omega}, \tag{3.12}$$

and where the other coefficients satisfy (2.20)–(2.22) with i replaced by ω, throughout. Thus from (3.6) we have the compound expansion

$$\begin{aligned} I_\nu(z) = \left(\frac{1}{2\pi z}\right)^{1/2} \Bigg\{ & e^z \left[1 + \sum_{m=0}^{\infty} \frac{d_m^+(\nu) m!}{z(z/\omega + 1) \cdots (z/\omega + m)} \right] \\ & + i\, e^{\nu\pi i}\, e^{-z} \left[1 + \sum_{m=0}^{\infty} \frac{d_m(\nu) m!}{z(z+1) \cdots (z+m)} \right] \Bigg\} \\ & \left(\mathrm{Re}\,(z) \geqq \varepsilon > 0, \quad -\frac{\pi}{4} + \delta \leqq \arg(z) \leqq \frac{\pi}{2} \right). \end{aligned} \tag{3.13}$$

A similar expansion, convergent in the conjugate region $\mathrm{Re}\,(z)\geqq\varepsilon>0$, $-\pi/2\leqq\arg z\leqq\pi/4-\delta$, can be derived from (3.13) and the relation $I_\nu(z)=\overline{I_{\bar\nu}(\bar z)}$.

When $z=x$ is real and positive both expansions are valid. A useful formula for $I_\nu(x)$ comes from adding these two formulas and dividing by 2. This yields

$$I_\nu(x)=\left(\frac{1}{2\pi x}\right)^{1/2}\left\{e^x\left[1+\sum_{m=0}^{\infty}\mathrm{Re}\left\{\frac{d_m^+(\nu)m!}{x(x/\omega^++1)\cdots(x/\omega^++m)}\right\}\right]\right.$$

(3.14)

$$\left.-\sin(\nu\pi)\,e^{-x}\left[1+\sum_{m=0}^{\infty}\frac{d_m(\nu)m!}{x(x+1)\cdots(x+m)}\right]\right\}$$

$$(x\geqq\varepsilon>0).$$

4. Error bounds. We seek a bound for $\delta_n(\nu,z)$ defined by

$$u(\nu,z)=\sum_{m=0}^{n-1}\frac{c_m(\nu)m!}{z(z+1)\cdots(z+m)}+\delta_n(\nu,z). \tag{4.1}$$

We concern ourselves here with the unmodified Bessel functions. Bounds for the modified Bessel functions can be derived in a similar manner.

Following [9, pp. 142–144] we derive

$$\delta_n(\nu,z)=\frac{1}{z(z+1)\cdots(z+n-1)}\int_0^1(1-\xi)^{z+n-1}\phi^{(n)}(\nu,\xi)\,d\xi. \tag{4.2}$$

Next, introduce the supremum

$$M(\nu,\alpha)=\sup_{t\in T}\{|f(\nu,t)\,e^{-\alpha t}|\}. \tag{4.3}$$

Because of (2.10a–c) it is seen that the supremum exists for all positive values of α. For our purposes it is necessary to impose the restriction that $0<\alpha<\mathrm{Re}\,(z)$. From (2.12) and (2.13) we find

$$|\phi(\nu,\xi)|\leqq M(\nu,\alpha)|1-\xi|^{-\alpha}\qquad(|\xi|\geqq1). \tag{4.4}$$

Therefore, for $0\leqq\xi<1$, we have from Cauchy's integral formula and (4.4)

$$|\phi^{(n)}(\nu,\xi)|\leqq\frac{M(\nu,\alpha)(1-\rho)^{-\alpha}n!}{\rho^n}\qquad(0<\rho<1-\xi). \tag{4.5}$$

Setting $\rho=(1-\xi)n/(n+1)$ in (4.5) and noting that $(1+n^{-1})^n<e$ $(n>0)$, we obtain from (4.2) and (4.5) the bound

$$|\delta_n(\nu,z)|\leqq\frac{M(\nu,\alpha)\,en!}{|z(z+1)\cdots(z+n-1)|(\mathrm{Re}\,(z)-\alpha)}. \tag{4.6}$$

Incidentally, it can be confirmed from this bound that that factorial series is absolutely convergent for $\mathrm{Re}\,(z)>\alpha$.

Since we know that the convergence becomes slow for large ν we now examine the asymptotic behavior of $M(\nu,\alpha)$ as $\nu\to\infty$. The asymptotic behavior of $f(\nu,t)$ can be established from the differential equation (2.8) written in the normalized form

$$\frac{d^2\hat f}{dt^2}=\left\{\nu^2\frac{1}{t(t-2i)}-\frac{1}{4t(t-2i)}\right\}\hat f, \tag{4.7a}$$

$$\hat f(\nu,t)\equiv t(t-2i)f(\nu,t). \tag{4.7b}$$

This equation is characterized in T by having a simple pole at $t=0$ and a regular singularity at $t=\infty$. To obtain an asymptotic approximation that is uniformly valid at

both singularities we shall apply an asymptotic theory of differential equations having a simple pole in the complex plane [5, Chap. 12, § 9]. *Note that Olver's parameter u corresponds to our parameter ν, and in this application Olver's ν is equal to* 1.

The Liouville transformation given by (see [5, pp. 438-439])

$$\zeta^{1/2}=\int_0^t \frac{d\tau}{\tau^{1/2}(\tau-2i)^{1/2}}=\ln\left[it+1+i(t^2-2it)^{1/2}\right], \tag{4.8}$$

$$W(\nu,\zeta)=\left(\frac{4\zeta}{t(t-2i)}\right)^{1/4}\hat{f}(t), \tag{4.9}$$

takes (4.7a) into the form

$$\frac{d^2W}{d\zeta^2}=\left\{\frac{\nu^2}{4\zeta}-\frac{1}{4\zeta^2}+\frac{\psi(\zeta)}{\zeta}\right\}W, \tag{4.10}$$

where

$$\psi(\zeta)=-\frac{3}{16}\left\{\frac{1}{\zeta}+\frac{1}{t(t-2i)}\right\}. \tag{4.11}$$

With regard to (4.8) and (4.9) we introduce a branchcut along the imaginary t axis from $t=2i$ to $t=i\infty$, and a temporary cut along the negative t axis from $t=0$ to $t=-\infty$. With these cuts the integrand of (4.8) is taken to be positive for $\tau=$ is $(0<s<2)$ and continuous elsewhere. The ζ domain Δ corresponding to the t domain T is depicted in Fig. 2. The asymptote at infinity (dashed line) is the parabola

$$\mathrm{Re}\,(\zeta)=\left(\frac{\mathrm{Im}\,(\zeta)}{\pi}\right)^2-\left(\frac{\pi}{2}\right)^2. \tag{4.12}$$

The $t\leftrightarrow\zeta$ transformation (4.8) is analytic and $1\leftrightarrow1$ at all points in T, including the singularity $t=0$ ($\zeta=0$). The Schwarzian $\psi(\zeta)$ is holomorphic within Δ.

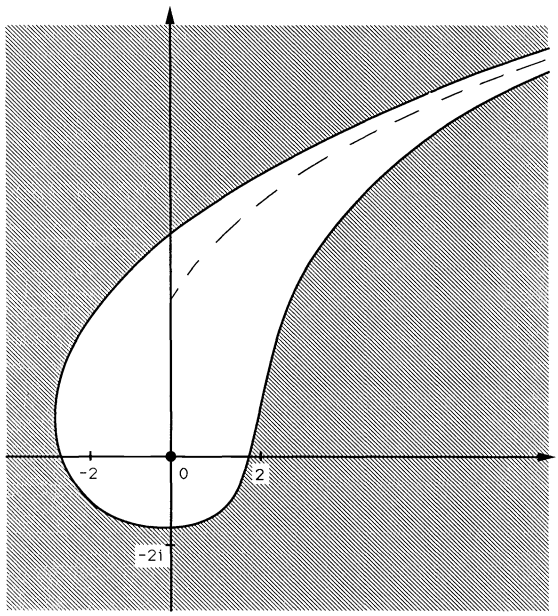


FIG. 2. *Domain Δ in ζ plane.*

The following can be deduced from the above equations:

$$\zeta = 2it + \tfrac{1}{3}t^2 - \tfrac{4}{45}it^3 + O(t^4) \quad \text{as } t \to 0, \tag{4.13}$$

$$\psi(\zeta) = -\tfrac{1}{16} + \tfrac{1}{80}\zeta - \tfrac{1}{504}\zeta^2 + O(\zeta^3) \quad \text{as } \zeta \to 0, \tag{4.14}$$

$$\exp(\zeta^{1/2}) \sim 2it \quad \text{as } t \to \infty. \tag{4.15}$$

We now apply Theorem 9.1 of [5, Chap. 12] to obtain the following solution of (4.10):

$$W(\nu, \zeta) = \zeta^{1/2} I_1(\nu\zeta^{1/2}) + \varepsilon_1(\nu, \zeta), \tag{4.16}$$

where $I_1(z)$ is the modified Bessel function of order 1 and $\varepsilon_1(\nu, \zeta)$ is uniformly bounded for $\zeta \in \Delta$ and $\nu > 0$ by

$$|\varepsilon_1(\nu, \zeta)| \leqq \mu_2 |\zeta|^{1/2} \mathscr{E}_1(\nu\zeta^{1/2}) \mathscr{M}_1(\nu\zeta^{1/2}) \exp\left\{\frac{\mu_1}{\nu} V\right\} \frac{V}{\nu}, \tag{4.17a}$$

with

$$V = \int_{\mathscr{P}} |v^{-1/2}\psi(v)|\, dv. \tag{4.17b}$$

The path of integration $\mathscr{P}$ of (4.18) links zero to ζ in the v plane subject to the first of the conditions (i) and (ii) of [5, p. 457]. The convergence of this integral at $\zeta = 0$ and $\zeta = \infty$ in Δ is readily verified from (4.11), (4.14), and (4.15); compare also equation (4.7a) with Example 4.1 of [5, p. 369].

In (4.17) $\mathscr{E}_1(z)$ and $\mathscr{M}_1(z)$ are auxiliary functions for modified Bessel functions of complex argument, satisfying

$$|I_1(z)| = \mathscr{E}_1(z)\mathscr{M}_1(z) \cos(\theta_1(z)) \qquad (|\arg(z)| \leqq \pi/2), \tag{4.18}$$

where $\theta_1(z)$ is a real function of z. The constants μ_1 and μ_2 are certain suprema involving these auxiliary functions (see [5, pp. 454–456]). The significance of the bound (4.17a, b) is that it establishes the fact that

$$\frac{\varepsilon_1(\nu, \zeta)}{\zeta^{1/2} I_1(\nu\zeta^{1/2})} = \begin{cases} O(\zeta) & \text{as } \zeta \to 0; \\ O(1/\nu) & \text{uniformly in } \Delta \text{ except} \\ & \text{near the zeros of } I_1(\nu\zeta^{1/2}). \end{cases} \tag{4.19}$$

The function (4.16) is the solution of (4.10) that is recessive at $\zeta = 0$. As such it can be directly identified with the hypergeometric function of (2.9), since that function is also recessive at $t = 0$. On comparing then both solutions at $t = 0$ ($\zeta = 0$) we deduce that for $\nu > 0$ ($\nu \neq \frac{1}{2}$), $t \in T$ ($\zeta \in \Delta$),

$$\begin{aligned} f(\nu, t) &= \frac{i\sigma}{2} F\left(\frac{3}{2} + \nu, \frac{3}{2} - \nu; 2; -\frac{it}{2}\right) \\ &= i\left\{\frac{4\nu^2 - 1}{4\nu}\right\} \frac{\zeta^{1/4}}{[t(2i - t)]^{3/4}} [I_1(\nu\zeta^{1/2}) + \zeta^{-1/2}\varepsilon_1(\nu, \zeta)]. \end{aligned} \tag{4.20}$$

In deriving this asymptotic formula we employed (2.6), (4.7b), (4.9), (4.13), (4.16), (4.19), together with the well-known asymptotic behavior of $I_1(z)$ near $z = 0$ (e.g., see [5, p. 435]).

Before applying the above results to bound $M(\nu, \alpha)$ we prove the following result.

LEMMA 1. *For* $\nu > \alpha(\pi/2 + 1) > 0$,

$$\begin{aligned} N(\nu, \alpha) &\equiv \sup_{t \in T} |\exp(\nu\zeta^{1/2} - \alpha t)| \\ &\leqq \max\{2^\nu \nu^\nu \alpha^{-\nu} e^{-\nu + \alpha\pi/2 + \alpha}, 2^\alpha [2(\ln(2) + (\pi/2) + 1)]^\nu\}. \end{aligned} \tag{4.21}$$

Proof. We seek an upper bound on $\mathrm{Re}\,(\nu\zeta^{1/2}-\alpha t)$ for $t\in T$. First, from the parametric equation for the boundary of T (see the paragraph following equation (2.12)) we have for $t\in T$

$$|t|\leqq|x|+\frac{\pi}{2},\qquad x\equiv\mathrm{Re}\,(t). \tag{4.22}$$

Thus, for $t\in T$, $x\geqq 0$,

$$\begin{aligned}\mathrm{Re}\,(\nu\zeta^{1/2}-\alpha t)&=\nu\ln\{|it+1+i(t^2-2it)^{1/2}|\}-\alpha x\\&\leqq\nu\ln\{|t|+1+(|t|^2+2|t|+1)^{1/2}\}-\alpha x\\&\leqq\nu\ln\{2(x+(\pi/2)+1)\}-\alpha x\equiv m(x).\end{aligned} \tag{4.23}$$

The only stationary point of $m(x)$ in $0<x<\infty$ is $x=x_c=\nu/\alpha-(\pi/2)-1$ (note the hypothesis of this lemma). It is readily verified that the absolute maximum of $m(x)$ in $[0,\infty)$ occurs at $x=x_c$. Therefore it follows that

$$\mathrm{Re}\,(\nu\zeta^{1/2}-\alpha t)\leqq m(x)\leqq m\left(\frac{\nu}{\alpha}-\frac{\pi}{2}-1\right)\qquad(t\in T,\ \mathrm{Re}\,(t)\geqq 0).$$

On exponentiation and simplification we arrive at the first of the maxima of the right-hand side of (4.21).

For the complementary case $-\ln(2)<x<0$ we find

$$\begin{aligned}\mathrm{Re}\,(\nu\zeta^{1/2}-\alpha t)&\leqq\nu\ln\{2(|x|+(\pi/2)+1)\}-\alpha x\\&\leqq\nu\ln\{2(\ln(2)+(\pi/2)+1)\}+\alpha\ln(2)\qquad(t\in T,\ \mathrm{Re}\,(t)<0),\end{aligned}$$

noting the monotonicity of the middle expression. Again we exponentiate, and the second of the maxima of the right-hand side of (4.21) is obtained. This completes the proof of Lemma 1.

Finally, from (4.17), (4.18), and (4.20), we derive for $t\in T$ the bound

$$\begin{aligned}|f(\nu,t)\,e^{-\alpha t}|\leqq\left|\left\{\frac{4\nu^2-1}{4}\right\}\left\{\frac{\zeta}{t(2i-t)}\right\}^{3/4}\left\{\frac{\mathscr{E}_1(\nu\zeta^{1/2})\mathscr{M}_1(\nu\zeta^{1/2})\exp\{-\nu\zeta^{1/2}\}}{\nu\zeta^{1/2}}\right\}\right.\\\left.\cdot\{\exp(\nu\zeta^{1/2}-\alpha t)\}\left\{1+\mu_2\exp\left(\frac{\mu_1}{\nu}V\right)\frac{V}{\nu}\right\}\right|.\end{aligned}$$

Hence, from (4.3),

$$M(\nu,\alpha)\leqq\left|\frac{4\nu^2-1}{4}\right|\lambda_1\lambda_2\lambda_3(\nu)N(\nu,\alpha), \tag{4.24}$$

where

$$\lambda_1=\sup_{t\in T}\left|\frac{\zeta}{t(2i-t)}\right|^{3/4}, \tag{4.25}$$

$$\lambda_2=\sup_{\mathrm{Re}\,(z)\geqq 0}\left\{\left|\frac{e^{-z}}{z}\right|\mathscr{E}_1(z)\mathscr{M}_1(z)\right\}, \tag{4.26}$$

$$\lambda_3(\nu)=1+\mu_2\exp\left(\frac{\mu_1}{\nu}V_\infty\right)\frac{V_\infty}{\nu}\qquad\left(V_\infty=\int_0^\infty|v^{-1/2}\psi(v)|\,dv\right), \tag{4.27}$$

and where $N(\nu,\alpha)$ is defined and bounded by (4.21). The existence of λ_1 is seen from (4.13) and (4.15). From (8.09), (8.16), and (8.25) of [5, Chap. 12] it can also be verified that λ_2 exists.

5. Asymptotic behavior of the factorial coefficients. For the convergence of the factorial series derived in § 2, it is helpful to obtain some information on the size of the factorial coefficients $c_m(\nu)$ as $m \to \infty$, including the dependence on the parameter ν. The purpose of this section is to investigate this behavior. We shall suppose that $\nu \neq 0, \frac{1}{2}$.

To determine the asymptotic behavior we use the representation

$$c_m(\nu) = \frac{1}{2\pi i} \oint_{|\xi|=\rho<1} \frac{\phi(\nu, \xi)}{\xi^{m+1}} \, d\xi, \tag{5.1}$$

which comes from (2.15) and Cauchy's formula, or equivalently

$$c_m(\nu) = \frac{1}{2\pi i} \oint_{|t|=\delta>0} \frac{f(\nu, t)\, e^{-t}}{(1-e^{-t})^{m+1}} \, dt, \tag{5.2}$$

where $t = -\ln(1-\xi)$. As it turns out, the asymptotic behavior of $f(\nu, t)$ as $t \to \infty$ determines the behavior of $c_m(\nu)$ as $m \to \infty$.

First deform the contour in (5.2) into a path γ consisting of a finite segment γ_0 from $t = i\pi$ to $t = -i\pi$, and two semi-infinite lines γ^+ and γ^- as depicted in Fig. 3. The integrals along the infinite segments converge because of the behavior of $f(\nu, t)$ near $t = \infty$ (see (2.10a-c)). The only finite singularity of $f(\nu, t)$ is a simple pole at $t = 2i$, and in deforming the contour we have taken into account the location of that pole. Note also that the integrand has an infinite number of poles along the imaginary axis at $t = 2n\pi i$ $(n = 0, \pm 1, \pm 2, \cdots)$.

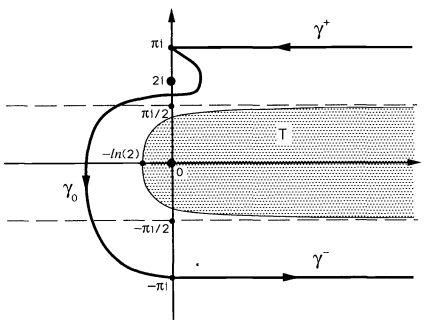


FIG. 3. *Integration path γ in t plane.*

Along the finite segment γ_0 the numerator of the integrand is bounded, while the modulus of the denominator is bounded below by c^m, for some real constant c which is greater than 1. Hence the integral along γ_0 contributes a term of order $O(c^{-m})$ as $m \to \infty$. All such terms of order $O(c^{-m})$, $c > 1$, will be considered from now on as *asymptotically negligible.* The main contributions come from the integrals along γ^+ and γ^-, since we shall see that they are of much larger order of magnitude. Any further information about the value of c is therefore unnecessary.

Note that we could have chosen the segments $\gamma^\pm$ to be semi-infinite lines along $\text{Im}\,(t) = A^\pm$, for any constants $A^+ > 0$ and $A^- < 0$. However, in order that the contribution from the path γ_0 be asymptotically negligible it is necessary that $\pi/2 \leqq |A^\pm| \leqq 3\pi/2$, since otherwise γ_0 would pass through at least one region in which $|1 - e^{-t}| < 1$. Our particular choices $A^\pm = \pm\pi$ were taken for convenience.

Now consider the contribution from the paths $\gamma^{\pm}$. We parameterize $\gamma^{\pm}$ by $t = \pm i\pi + \tau$ $(0 \leqq \tau < \infty)$. Integration along γ^{+} and γ^{-} then gives

$$\frac{1}{2\pi i}\int_0^{\infty} \frac{[f(\nu, i\pi+\tau) - f(\nu, -i\pi+\tau)]\, e^{-\tau}}{(1+e^{-\tau})^{m+1}}\, d\tau. \tag{5.3}$$

In (5.3) we now make the change of variable $1+e^{-\tau} = e^{u}$, which yields

$$\frac{1}{2\pi i}\int_0^{\ln(2)} g(\nu, u)\, e^{-mu}\, du, \tag{5.4}$$

where

$$g(\nu, u) = f(\nu, i\pi - \ln(e^{u}-1)) - f(\nu, -i\pi - \ln(e^{u}-1)). \tag{5.5}$$

The behavior of (5.4) as $m \to \infty$ depends upon the asymptotic behavior of $g(\nu, u)$ as $u \to 0^{+}$, which in turn corresponds to the behavior of $f(\nu, t)$ as $t \to \infty$ along γ^{+} and γ^{-}. For this we use the following representation of $f(\nu, t)$:

$$f(\nu, t) t^{\nu-3/2} \sum_{k=0}^{\infty} a_k(\nu) t^{-k} + t^{-\nu-3/2} \sum_{k=0}^{\infty} b_k(\nu) t^{-k}, \tag{5.6}$$

where both series converge for $|t| > 2$. The coefficients $\{a_k(\nu)\}$ and $\{b_k(\nu)\}$ $(k = 0, 1, 2, \cdots)$ can be determined explicitly from (2.9) and a connection formula (see [5, p. 167, eq. (10.15)]).

Then, setting $t = \pm i\pi + \tau$ in (5.6) and reexpanding yields

$$f(\nu, \pm i\pi + \tau) = \tau^{\nu-3/2} \sum_{k=0}^{\infty} a_k^{\pm}(\nu) \tau^{-k} + \tau^{-\nu-3/2} \sum_{k=0}^{\infty} b_k^{\pm}(\nu) \tau^{-k}, \tag{5.7}$$

which converges for $\tau > \pi$. In this expression $a_0^{\pm}(\nu) = a_0(\nu)$, $b_0^{\pm}(\nu) = b_0(\nu)$, and the other coefficients in (5.7) can be expressed explicitly in terms of those in (5.6).

If we now write

$$\begin{aligned} g(\nu, u) = {} & f\left(\nu, i\pi - \ln(u) + \ln\left\{\frac{u}{e^{u}-1}\right\}\right) \\ & - f\left(\nu, -\pi - \ln(u) + \ln\left\{\frac{u}{e^{u}-1}\right\}\right), \end{aligned} \tag{5.8}$$

and use the fact that $\ln(u/(e^{u}-1))$ is regular at $u = 0$ and vanishes there, we find (observing that the leading terms cancel)

$$\begin{aligned} g(\nu, u) = {} & (-\ln(u))^{\nu-5/2} \sum_{k=0}^{\infty} \frac{\hat{a}_k(\nu)}{(-\ln(u))^{k}} + (-\ln(u))^{-\nu-5/2} \sum_{k=0}^{\infty} \frac{\hat{b}_k(\nu)}{(-\ln(u))^{k}} \\ & + O(u(-\ln(u))^{\nu-5/2}). \end{aligned} \tag{5.9}$$

The series in (5.9) converge uniformly for $0 < u \leqq \varepsilon$, for sufficiently small ε. Each of the coefficients $\hat{a}_k(\nu)$ and $\hat{b}_k(\nu)$ is expressible in terms of the coefficients $\{a_k^{\pm}(\nu)\}$ and $\{b_k^{\pm}(\nu)\}$.

For ε as above let

$$\int_0^{\ln(2)} g(\nu, u)\, e^{-mu}\, du = \int_0^{\varepsilon} g(\nu, u)\, e^{-mu}\, du + \int_{\varepsilon}^{\ln(2)} g(\nu, u)\, e^{-mu}\, du. \tag{5.10}$$

Since the contribution of the second integral is easily seen to be asymptotically negligible, it remains to determine the asymptotic behavior of

$$\int_0^\varepsilon g(\nu,u)\,e^{-mu}\,du=\sum_{k=0}^{\infty}\hat{a}_k(\nu)\int_0^\varepsilon(-\ln(u))^{\nu-5/2-k}\,e^{-mu}\,du$$

$$+\sum_{k=0}^{\infty}\hat{b}_k(\nu)\int_0^\varepsilon(-\ln(u))^{-\nu-5/2-k}\,e^{-mu}\,du \tag{5.11}$$

$$+\int_0^\varepsilon O(u(-\ln(u))^{\nu-5/2})\,e^{-mu}\,du,$$

as $m\to\infty$. Clearly integrals of the form

$$L(m,\lambda,\alpha)=\int_0^\varepsilon(-\ln(u))^{\alpha}u^{\lambda-1}\,e^{-mu}\,du \tag{5.12}$$

play a key role. Such integrals have been discussed by Erdélyi [4] for the parameters $0<\varepsilon<1$, $\lambda>0$, and α real. Erdélyi has shown that

$$L(m,\lambda,\alpha)\sim\textstyle\sum_{n=0}^{\infty}(-1)^n\binom{\alpha}{n}\Gamma^{(n)}(\lambda)m^{-\lambda}(\ln(m))^{\alpha-n}, \tag{5.13}$$

as $m\to\infty$. Hence setting $\lambda=1$ and 2, $\alpha=\pm\nu-\frac{5}{2}-k$ in (5.13), we finally obtain as $m\to\infty$

$$c_m(\nu)\sim\frac{(\ln(m))^{\nu-5/2}}{m}\sum_{k=0}^{\infty}\frac{\alpha_k(\nu)}{(\ln(m))^k}+\frac{(\ln(m))^{-\nu-5/2}}{m}\sum_{k=0}^{\infty}\frac{\beta_k(\nu)}{(\ln(m))^k}$$

$$+O\left(\frac{(\ln(m))^{\nu-5/2}}{m^2}\right), \tag{5.14}$$

where the coefficients $\{\alpha_k(\nu)\}$ and $\{\beta_k(\nu)\}$ are, in principle, computable in terms of those in the expansion for $f(\nu,t)$ given by (5.7) above. For example, the leading coefficient is found to be

$$\alpha_0(\nu)=\tfrac{1}{4}(\nu-\tfrac{3}{2})^2a_0(\nu), \tag{5.15}$$

where $a_0(\nu)$ is given in (2.10b).

By using the same analysis it can be shown that the coefficients $\{d_m(\nu)\}$ appearing in the factorial series expansions of the modified Bessel functions have similar asymptotic representations. Also, when ν is complex the factorial coefficients (in the modified and unmodified cases) have similar asymptotic representations. This can be shown using an analogue of Erdélyi's result (5.13) for α complex, due to Wong and Wyman [10].

The above procedure can also be used to analyze other similarly constructed series expansions and their coefficients. For example, Rosser [6] started with the integral representation

$$K_\nu(z)=\frac{\sqrt{\pi}\,(z/2)^\nu\,e^{-z}}{\Gamma(\nu+\frac{1}{2})}\int_0^1 s^{z-1}(\ln^2(s)-2\ln(s))^{\nu-1/2}\,ds, \tag{5.16}$$

and obtained the convergent (factorial-like) series expansion

$$\int_0^1 s^{z-1}(\ln^2(s)-2\ln(s))^{\nu-1/2}\,ds=\sum_{m=0}^{\infty}E_m(\nu)\frac{\Gamma(z)\Gamma(m+\nu+\frac{1}{2})}{\Gamma(z+m+\nu+\frac{1}{2})}, \tag{5.17}$$

where the coefficients $\{E_m(\nu)\}$ are defined implicitly by means of the auxiliary function

$$h(\nu, t)=\left[\frac{\ln^2(1-t)-2\ln(1-t)}{t}\right]^{\nu-1/2}=\sum_{m=0}^{\infty} E_m(\nu)t^m. \tag{5.18}$$

Rosser conjectured that the coefficients should satisfy

$$E_m(\nu)=O\left(\frac{\{\ln(m)\}^{M(\nu)}}{m}\right) \quad \text{as } m\to\infty, \tag{5.19}$$

where $M(\nu)$ is some number independent of m. With a similar analysis as above it can easily be shown that this is indeed the case. In fact the coefficients $\{E_m(\nu)\}$ have asymptotic expansions of the form

$$E_m(\nu)\sim\frac{(\ln(m))^{2\nu-2}}{m}\sum_{k=0}^{\infty}\frac{e_k(\nu)}{(\ln(m))^k} \quad \text{as } m\to\infty, \tag{5.20}$$

where the coefficients $\{e_k(\nu)\}$ can also be explicitly given. When ν is half an odd integer (5.17) reduces to an ordinary factorial series, but otherwise there do not exist explicit means for computing the coefficients $\{E_m(\nu)\}$. A reason for this is that when ν is not half an odd integer the integral (5.16) does not possess an asymptotic expansion involving only powers of $1/z$.

6. Numerical calculations and conclusions. While the factorial series converge in principle for $\mathrm{Re}\,(z)>0$ and for all $\nu\geqq 0$, in practice the error bounds (§ 4) and the behavior of the coefficients (§ 5) indicate that the convergence becomes quite slow when $\mathrm{Re}\,(z)$ is small, or when ν is large with respect to $\mathrm{Re}\,(z)$. For small $|z|$ the ascending power series converge well, and for moderate to large ν the large-order asymptotic expansions [5, pp. 377, 423–424] are particularly powerful since they are doubly asymptotic: they are uniformly valid at both $z=0$ and at $z=\infty$.

Tables 1(a) ($\nu=0$) and 2(a) ($\nu=2$) contain, for various values of n and real values of z (denoted by x), the following partial factorial sums which appear in the representation for $K_\nu(x)$:

$$\left[1+\sum_{m=0}^{n}\frac{d_m(\nu)m!}{x(x+1)\cdots(x+m)}\right]. \tag{6.1}$$

The coefficients $\{d_m(\nu)\}$ were calculated using the recurrence relations (2.20)–(2.22) with i replaced by -1, and all calculations were performed in double precision. (The rows labeled "A & S" are values of $(2x/\pi)^{1/2}e^xK_\nu(x)$ calculated from tables in Abramowitz and Stegun [1, Chap. 9].) As a comparison, partial sums of the corresponding ascending power series [1, p. 375] for $(2x/\pi)^{1/2}e^xK_0(x)$ are given in Table 1(b).

Tables 1(c) and 2(b) contain calculations for $(2x/\pi)^{1/2}e^xK_\nu(x)$ using asymptotic expansions for large argument, and Table 2(c) contains calculations for $(2x/\pi)^{1/2}e^xK_2(x)$ using the uniform asymptotic expansion for large order; see [1, p. 378, eqs. (9.7.2), (9.7.8)]). In these tables the values in parentheses are the number of terms taken in the expansion to attain the indicated value.

For small $\mathrm{Re}\,(z)$ the convergence of the factorial series, while predictably slow, appears to be numerically stable. For $\mathrm{Re}\,(z)$ large the factorial coefficients are not significantly more difficult to calculate than the coefficients in the large argument asymptotic expansions, and the two series have roughly comparable errors after a moderate number of terms. If better accuracy than available from the asymptotic expansions is required, the convergent factorial series are a reasonable alternative, particularly when ν is small.

TABLE 1

$(2x/\pi)^{1/2}\, e^{x}K_0(x)$: (a) *factorial series*; (b) *ascending power series*; (c) *large argument asymptotic expansion.*

(a)	$x=1$	$x=5$	$x=10$	$x=15$
A & S	0.9131494218	0.9773566865	0.9881392704	0.9919594236
$n=10$	0.9125955022	0.9773565876	0.9881392700	0.9919594236
$n=15$	0.9128097381	0.9773566695	0.9881392704	0.9919594236
$n=20$	0.9129104141	0.9773566819	0.9881392704	0.9919594236
$n=25$	0.9129677449	0.9773566849	0.9881392704	0.9919594236
$n=30$	0.9130043138	0.9773566858	0.9881392704	0.9919594236
$n=40$	0.9130477052	0.97735 66863	0.9881392704	0.9919594236
$n=50$	0.9130722200	0.9773566865	0.9881392704	0.9919594236
(b)	$x=1$	$x=5$	$x=10$	$x=15$
$n=10$	0.9131494218	0.9772056352	−84972.53414	-8.3380×10^{10}
$n=15$	0.9131494218	0.9773566865	−2.889563970	−229477045.2
$n=20$	0.9131494218	0.9773566865	0.9881318057	−26195.53622
$n=25$	0.9131494218	0.9773566865	0.9881394196	0.7192550390
$n=30$	0.9131494218	0.9773566865	0.9881394196	0.9909960580
$n=40$	0.9131494218	0.9773566865	0.9881394196	0.9909960580
(c)	$x=1$	$x=5$	$x=10$	$x=15$
	0.8750000000 (1)	0.9773478565 (7)	0.9881392701 (15)	0.9919594236 (10)

TABLE 2

$(2x/\pi)^{1/2}\, e^{x}K_2(x)$: (a) *factorial series*; (b) *large argument asymptotic expansion*; (c) *large order asymptotic expansion.*

(a)	$x=1$	$x=5$	$x=10$	$x=15$
A & S	3.524072633	1.405741914	1.195422969	1.128560359
$n=10$	3.488120013	1.405736283	1.195422946	1.128560358
$n=15$	3.500191625	1.405740864	1.195422968	1.128560359
$n=20$	3.506299317	1.405741615	1.195422969	1.128560359
$n=25$	3.509968194	1.405741804	1.195422969	1.128560359
$n=30$	3.512408185	1.405741866	1.195422969	1.128560359
$n=40$	3.515441113	1.405741901	1.195422969	1.128560359
$n=50$	3.517244998	1.405741909	1.195422969	1.128560359
(b)	$x=1$	$x=5$	$x=10$	$x=15$
	3.387695313 (3)	1.405726905 (7)	1.195422969 (15)	1.128560359 (9)
(c)	$x=1$	$x=5$	$x=10$	$x=15$
	3.523737330 (13)	1.405735521 (6)	1.195422969 (13)	1.128560359 (7)

In summary, factorial series expansions for Bessel functions are good for small parameter values and moderate to large values of $\mathrm{Re}\,(z)$, $0 \leqq \nu \leqq 2$ and $\mathrm{Re}\,(z) \geqq 10$, say. Special contiguous relations for Bessel functions have been successfully used for their efficient numerical calculation; see [1, pp. 385–388]. Factorial series could possibly be employed in conjunction with these other procedures. For example, as an alternative means of providing starting values in forward recursion schemes or continued fraction iterations [8, p. 153], particularly when the usual normalization procedure is not numerically satisfactory (such as Neumann's series for $K_0(x)$ when x is large [1, p. 377, eq. (9.6.53)]).

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN (1965), *Handbook of Mathematical Functions*, Seventh ed., Dover, New York.

[2] BATEMAN MANUSCRIPT PROJECT (1953), *Tables of Integral Transforms*, Vol. 1, A. Erdélyi, ed., McGraw-Hill, New York.

[3] G. DOETSCH (1955), *Handbuch der Laplace Transformation* II, Birkhaüser-Verlag, Basel, Stuttgart.

[4] A. ERDÉLYI (1961), *General asymptotic expansions of Laplace integrals*, Arch. Rational Mech. Anal., 7, pp. 1–20.

[5] F. W. J. OLVER (1974), *Asymptotics and Special Functions*, Academic Press, New York.

[6] J. B. ROSSER (1975), *Factorial expansions for certain Bessel functions*, MRC Tech. Summary Report No. 1572, University of Wisconsin, Madison, WI.

[7] W. WASOW (1965), *Asymptotic Expansions for Ordinary Differential Equations*, Academic Press, New York.

[8] G. N. WATSON (1944), *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, London, New York.

[9] E. T. WHITTAKER AND G. N. WATSON (1935), *A Course of Modern Analysis*, Fourth ed., Cambridge University Press, London, New York.

[10] R. WONG AND M. WYMAN (1972), *A generalization of Watson's lemma*, Canad. J. Math., 24, pp. 185–208.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1173–1182, July 1991

020

POLYNOMIALS WITH NONNEGATIVE COEFFICIENTS WHOSE ZEROS HAVE MODULUS ONE*

RONALD EVANS† AND JOHN GREENE‡

Abstract. Define $p(z) = \prod_{j=0}^{n-1} \left(z - e^{i(\theta+\alpha j)}\right)\left(z - e^{-i(\theta+\alpha j)}\right)$ for $\alpha > 0$ and $\theta \geq 0$ with $\pi/2 - (n-1)\alpha/2 \leq \theta \leq \pi - (n-1)\alpha/2$. It is proved that if $0 < \alpha < \pi/n$, then the $2n+1$ coefficients of $p(z)$ are all positive. It is also proved that if for some point θ, all coefficients of $p(z)$ are nonnegative, then each coefficient is an increasing function of θ in a neighborhood of this point. A similar result is conjectured for more general polynomials $p(z)$.

Key words. orthogonal polynomials, q-ultraspherical polynomials, absolutely monotonic polynomials

AMS(MOS) subject classifications. 33A65, 30C15

1. Introduction.

For

$$\alpha > 0 \quad \text{and} \quad \theta \geq 0, \tag{1.1}$$

consider the monic polynomial $p(z)$ of degree $2n$ whose zeros consist of the n equally spaced points

$$\exp(i(\theta + \alpha j)), \qquad 0 \leq j \leq n-1, \tag{1.2}$$

along with their n complex conjugates, i.e.,

$$p(z) = \prod_{j=0}^{n-1} \left(z - e^{i(\theta+\alpha j)}\right)\left(z - e^{-i(\theta+\alpha j)}\right). \tag{1.3}$$

We assume *throughout* that the variable θ in (1.3) is restricted to the interval

$$\pi/2 - (n-1)\alpha/2 \leq \theta \leq \pi - (n-1)\alpha/2. \tag{1.4}$$

Equivalently,

$$\pi/2 \leq \theta + (n-1)\alpha/2 \leq \pi, \tag{1.5}$$

so that the geometric mean of the n zeros in (1.2) lies in the second quadrant. Condition (1.5) automatically holds, for example, if each of the n zeros in (1.2) has Argument $\in (0, \pi)$ and the coefficient of z in $p(z)$ is positive; this is easily seen from (2.12) and (2.17). When (1.5) holds, the geometric mean of the n zeros in (1.2) is closer to -1 than to $+1$, and it moves (together with at least half of the zeros of $p(z)$) towards -1 along the unit circle as θ increases.

---

*Received by the editors August 1, 1989; accepted for publication (in revised form) August 21, 1990.

†Department of Mathematics, C-012, University of California, San Diego, La Jolla, California 92093.

‡Department of Mathematics and Statistics, University of Minnesota, Duluth, Minnesota 55812.
The work of this author was supported in part by National Science Foundation grant DMS-8801131.

The coefficients of $p(z)$ are not necessarily increasing functions of θ, even if each of the n zeros in (1.2) has Argument $\in (0, \pi)$ (in which case each of the n quadratic factors in (1.3) has increasing coefficients). For example, if $n = 3$, $\alpha = 5\pi/12$, then the coefficient of z^3 in $p(z)$ is negative and *decreasing* at $\theta = \pi/8$, while $\pi/8$ is in the interval (1.4). However, the following theorem holds for all n. The proof, given in §3, depends on properties of q-ultraspherical polynomials discussed in §2.

THEOREM 1. *If for some nonnegative $\theta = \theta_0$ in the interval* (1.4), *all coefficients of $p(z)$ are nonnegative, then they are each increasing functions of θ for $\theta_0 \leq \theta < \pi - (n-1)\alpha/2$. Except for the coefficients* 1 *of the leading and constant terms, the coefficients are in fact strictly increasing, unless $\alpha = 2\pi/n$.*

For $\alpha = 2\pi/n$, we have

$$p(z) = z^{2n} - 2\cos(\theta n) z^n + 1,$$

which has nonnegative coefficients for $\pi/(2n) \leq \theta \leq \pi/n$, but if $n > 1$, the coefficient of z is zero, which is not strictly increasing. This formula for $p(z)$ is proved in §3 (see (3.10)).

Consider for the moment the general polynomial

$$P(z) = \prod_{j=0}^{n-1} \left(z - e^{i(\theta + a_j)}\right)\left(z - e^{-i(\theta + a_j)}\right) \tag{1.6}$$

where

$$\theta \geq 0, \qquad 0 = a_0 < a_1 < \cdots < a_{n-1}. \tag{1.7}$$

The polynomial $P(z)$ reduces to $p(z)$ when $a_j = j\alpha$, $0 \leq j \leq n-1$. In view of Theorem 1, we might ask if nonnegativity of the coefficients of $P(z)$ for some $\theta = \theta_0$ always implies that the coefficients are increasing for $\theta \geq \theta_0$, when θ is restricted to the interval

$$\pi/2 - (a_1 + \cdots + a_{n-1})/n \leq \theta \leq \pi - (a_1 + \cdots + a_{n-1})/n. \tag{1.8}$$

The answer is no. For example, if $n = 3$, $a_1 = \pi/2$, $a_2 = 7\pi/12$, then the coefficients of $P(z)$ are all positive for $\pi/4 < \theta < 23\pi/36$, yet the coefficients of z^2, z^3, z^4 are each decreasing at $\theta = 2$. However, we believe the following.

CONJECTURE. *If the coefficients of $P(z)$ are all nonnegative for some $\theta = \theta_0 \geq 0$, then they are each increasing functions of θ on the interval $\theta_0 \leq \theta < \pi - a_{n-1}$.*

For convenient application of Theorem 1, we would like to have a simple necessary condition for the nonnegativity of the coefficients of $p(z)$. This is given in Theorem 2.

THEOREM 2. *Suppose that*

$$0 < \alpha < \pi/n. \tag{1.9}$$

Then each coefficient of $p(z)$ is positive (*and hence increasing in θ, by Theorem* 1).

This theorem was motivated by the fact that for sufficiently small α, all zeros of $p(z)$ are closer to -1 than to $+1$ (because of (1.5)), and so all coefficients of $p(z)$ are positive. The question is how small α must be.

For $n > 1$, the upper bound in (1.9) is best possible, i.e., if $\alpha > \pi/n$, the coefficients of $p(z)$ cannot all be positive on the interval (1.4). If $\alpha \geq 2\pi/n$, there is *no* θ

in the interval (1.4) for which all coefficients of $p(z)$ are positive. If $\pi/n \le \alpha < 2\pi/n$, the coefficients of $p(z)$ are all positive only on a subinterval

$$r_\alpha < \theta < \pi - (n-1)\alpha/2 \tag{1.10}$$

of the interval (1.4). These remarks will be proved in §4. Also in §4 we prove Theorem 2 and the following related result.

THEOREM 3. *Let $0 < \alpha < \pi/n$. Then all coefficients of*

$$p(u,v) := \prod_{j=(1-n)/2}^{(n-1)/2} (1 + ue^{i\alpha j} + ve^{-i\alpha j}) \tag{1.11}$$

are positive, i.e.,

$$p(u,v) = \sum_{\substack{0 \le r,s \le n \\ r+s \le n}} a_{rs} u^r v^s, \qquad a_{rs} > 0. \tag{1.12}$$

(The variable j in (1.11) ranges over halves of odd integers if n is even.)

As an application of Theorem 2, we give in §5 a short proof of Theorem 4 below in the special case

$$f(z) = (z^{mk} - 1)/(z^k - 1), \tag{1.13}$$

where m, k are positive integers.

THEOREM 4. *Let $f(z)$ denote a monic polynomial of degree N with nonnegative coefficients and with zeros $z_1, z_2, \cdots, z_N$. For fixed $t \ge 0$, write*

$$f_t(z) = \prod_{\substack{1 \le j \le N \\ |\operatorname{Arg} z_j| > t}} (z - z_j). \tag{1.14}$$

Then if $f(z) \ne f_t(z)$, all coefficients of $f_t(z)$ are positive.

Theorem 4 had been open for several years until a proof was found recently by Barnard et al. [2].

In the special cases $f(z) = z^N + 1$ or $f(z) = 1 + z + \cdots + z^N$, we can say a bit more about the polynomials $f_t(z)$ in (1.14), namely, the following theorem [4].

THEOREM 5. *If $f(z) = z^N + 1$ or $f(z) = 1 + z + \cdots + z^N$, and if $f_t(z) \ne f(z)$, then $f_t(z)$ is a strictly unimodal polynomial. (In particular, all coefficients of $f_t(z)$ are ≥ 1.)*

If $f(z)$ is given by (1.13), it is not generally true that $f_t(z)$ is unimodal when $f_t(z) \ne f(z)$.

2. The coefficients of $p(z)$ in terms of q-ultraspherical polynomials. We will use the following additional notation throughout:

$$q = e^{i\alpha}, \tag{2.1}$$

$$\beta = \theta + (n-1)\alpha/2 - \pi/2, \tag{2.2}$$

and

$$x = \sin\beta. \tag{2.3}$$

Observe that (1.5) is equivalent to

$$0 \le \beta \le \pi/2, \tag{2.4}$$

which implies that

$$0 \le x \le 1, \qquad \frac{dx}{d\theta} \ge 0. \tag{2.5}$$

In order to relate $p(z)$ to q-ultraspherical polynomials (see (2.12)–(2.13)), we begin by replacing j by $j + (n-1)/2$ in (1.3) to obtain

$$p(z) = \prod_{j=(1-n)/2}^{(n-1)/2} \left(z - e^{i(\beta+\alpha j+\pi/2)}\right)\left(z - e^{-i(\beta+\alpha j+\pi/2)}\right). \tag{2.6}$$

Since the range of values of j in (2.6) is symmetric about zero, we have

$$\begin{aligned} p(z) &= \prod_{j=(1-n)/2}^{(n-1)/2} \left(z - e^{i(\beta+\alpha j+\pi/2)}\right)\left(z - e^{-i(\beta-\alpha j+\pi/2)}\right) \\ &= \prod_{j=(1-n)/2}^{(n-1)/2} (z^2 - 2zq^j\cos(\beta+\pi/2) + q^{2j}) \\ &= \prod_{j=(1-n)/2}^{(n-1)/2} (z^2 + 2zq^j\sin\beta + q^{2j}). \end{aligned} \tag{2.7}$$

Replace j by $-j$ and multiply each factor by q^{2j} to obtain

$$p(z) = \prod_{j=(1-n)/2}^{(n-1)/2} (z^2q^{2j} + 2zxq^j + 1). \tag{2.8}$$

Note that the coefficients of $p(z)$ are symmetric about the middle, as

$$z^{2n}p(1/z) = p(z), \tag{2.9}$$

and the leading and constant coefficients of $p(z)$ are 1 for all θ, α.

The generating function for the q-ultraspherical polynomials $C_k(x;t|q)$ is [1, eq. (3.4), p. 179]

$$\sum_{k=0}^{\infty} C_k(x;t|q)w^k = \prod_{k=0}^{\infty} \frac{(1 - 2twxq^k + t^2w^2q^{2k})}{(1 - 2wxq^k + w^2q^{2k})}, \qquad 0 < q < 1. \tag{2.10}$$

In particular, with $t = q^{-n}$,

$$\sum_{k=0}^{\infty} C_k(x; q^{-n}|q) w^k = \prod_{k=-n}^{-1} (1 - 2wxq^k + w^2 q^{2k}). \tag{2.11}$$

The polynomials $C_k(x; q^{-n}|q)$ are well defined by (2.11) for $q = e^{i\alpha}$. Replace w by $-zq^{(n+1)/2}$ in (2.11) and use (2.8) to see that

$$p(z) = \sum_{k=0}^{2n} E_k(x; q^{-n}|q) z^k, \tag{2.12}$$

where

$$E_k := E_k(x) = E_k(x; q^{-n}|q) = (-1)^k q^{k(n+1)/2} C_k(x; q^{-n}|q). \tag{2.13}$$

The $C_k(x; t|q)$ satisfy the recurrence relation [1, eq. (1.1), p. 176]

$$2x(1 - tq^k) C_k(x; t|q) = (1 - q^{k+1}) C_{k+1}(x; t|q) + (1 - t^2 q^{k-1}) C_{k-1}(x; t|q) \tag{2.14}$$

for $k \geq 1$, with

$$C_0(x; t|q) = 1, \qquad C_1(x; t|q) = 2x(1 - t)/(1 - q). \tag{2.15}$$

In view of (2.1) and (2.13)–(2.15), the E_k satisfy the recurrence

$$\begin{aligned} E_k = {} & 2x \frac{\sin((n+1-k)\alpha/2)}{\sin(k\alpha/2)} E_{k-1} \\ & + \frac{\sin((2n+2-k)\alpha/2)}{\sin(k\alpha/2)} E_{k-2} \qquad (k \geq 2) \end{aligned} \tag{2.16}$$

with

$$E_0 = 1, \qquad E_1 = 2x \frac{\sin(n\alpha/2)}{\sin(\alpha/2)}. \tag{2.17}$$

3. Proof of Theorem 1. Theorem 1 is trivial for $n = 1$, so let $n > 1$. For brevity, write

$$A_k = \frac{\sin((n+1-k)\alpha/2)}{\sin(k\alpha/2)}, \quad B_k = \frac{\sin((2n+2-k)\alpha/2)}{\sin(k\alpha/2)}, \quad k \geq 1, \tag{3.1}$$

so by (2.16),

$$E_k = 2xA_k E_{k-1} + B_k E_{k-2}, \qquad k \geq 2. \tag{3.2}$$

By hypothesis, for some x_0 with $0 \leq x_0 < 1$,

$$E_k(x_0) \geq 0 \quad \text{for } 0 \leq k \leq 2n. \tag{3.3}$$

By (2.9), it suffices to show that the polynomials $E_k(x)$ are strictly increasing on $x_0 < x < 1$ for $1 \leq k \leq n$.

Case 1. $\alpha < 2\pi/n$. In this case,

$$A_k > 0 \quad \text{for } 1 \leq k \leq n. \tag{3.4}$$

In particular, the leading coefficient of $E_k(x)$ is positive for each k, $1 \leq k \leq n$.

Suppose there is an integer m with $2 \leq m \leq n$ such that

$$B_m < 0, \tag{3.5}$$

and choose the maximal such m. By (3.1),

$$B_k < 0 \quad \text{for } 2 \leq k \leq m. \tag{3.6}$$

By (3.2) and Favard's theorem [3, Thm. 4.4, p. 21], $E_1, E_2, \cdots, E_m$ are orthogonal polynomials with respect to a positive-definite operator. Thus we can apply the theorem on separation of zeros [3, Thm. 5.3, p. 28] to conclude that the zeros of $E_1, \cdots, E_m$ are all real and simple, and that a zero of E_{k-1} lies strictly between every two consecutive zeros of E_k, $2 \leq k \leq m$.

We proceed to prove by induction on k that if $1 \leq k \leq m$, then the largest zero of E_k is $\leq x_0$. This holds for $k = 1$ since $E_1 = 2A_1 x$ and $0 \leq x_0$. Let $k > 1$. By induction hypothesis, the largest zero of E_{k-1} is $\leq x_0$, so by separation of zeros, x_0 exceeds the second largest zero of E_k. For x between the largest and second largest zeros of E_k, $E_k(x)$ is negative. Thus, by (3.3), the largest zero of E_k is $\leq x_0$, and the induction is complete.

It follows for $1 \leq k \leq m$ that

$$E_k(x) = c_k \prod_{j=1}^{k} (x - \alpha_{jk}) \tag{3.7}$$

with $c_k > 0$ and $\alpha_{jk} \leq x_0$ $(1 \leq j \leq k)$. Thus $E_k(x)$ is strictly increasing on $x_0 < x < 1$ for $1 \leq k \leq m$.

If there is no integer m with $2 \leq m \leq n$ for which (3.5) holds, set $m = 1$. It remains to prove that $E_k(x)$ is strictly increasing on $x_0 < x < 1$ for $n \geq k > m$. This follows from (3.2), since $A_k > 0$ and $B_k \geq 0$.

Case 2. $\alpha = 2\pi/n$. In this case, by (2.16) and (2.17), $E_1(x) = E_2(x) = \cdots = E_{n-1}(x) = 0$. Thus by (2.9) and (2.12),

$$p(z) = z^{2n} + E_n z^n + 1. \tag{3.8}$$

It is easily seen from (1.3) that

$$p(1) = (e^{i\theta n} - 1)(e^{-i\theta n} - 1) = 2 - 2\cos(\theta n). \tag{3.9}$$

By (3.8) and (3.9), $E_n = -2\cos(\theta n)$, so

$$p(z) = z^{2n} - 2\cos(\theta n) z^n + 1. \tag{3.10}$$

For $\pi/(2n) \leq \theta \leq \pi/n$, the coefficients of $p(z)$ are nonnegative and they are increasing functions of θ.

Case 3. $\alpha > 2\pi/n$. In this case, $x_0 > 0$ by (2.2) and (2.3). Moreover, by (1.1) and (1.5), we may suppose that

$$2\pi/n < \alpha < 2\pi/(n-1). \tag{3.11}$$

By (2.17) and (3.11),

$$E_1(x_0) = 2x_0 \sin(n\alpha/2)/\sin(\alpha/2) < 0. \tag{3.12}$$

This contradicts (3.3), so Case 3 is vacuous.

4. Proofs of Theorems 2 and 3.

Proof of Theorem 2. Let $0 < \alpha < \pi/n$. By (2.9) and (2.12), it suffices to prove

$$E_k > 0, \qquad 0 \le k \le n. \tag{4.1}$$

This follows for $k = 0, 1$ by (2.17). For $2 \le k \le n$, all sines in (3.1) are positive, so

$$A_k > 0, \quad B_k > 0 \quad \text{for } 2 \le k \le n. \tag{4.2}$$

Thus (4.1) follows by (3.2) and induction on k.

Proof of Theorem 3. Let $0 < \alpha < \pi/n$. The proof of (4.1) actually yields the stronger result

$$E_k = \sum_{i=0}^{M} b_{ik} x^i, \qquad 0 \le k \le 2n, \tag{4.3}$$

with

$$\begin{aligned} &b_{ik} > 0, \quad \text{if } i \equiv k (\mathrm{mod}\, 2), \\ &b_{ik} = 0, \quad \text{otherwise}, \end{aligned} \tag{4.4}$$

where

$$M = \min(k,\ 2n-k). \tag{4.5}$$

Thus, by (2.8) and (2.12),

$$p(z) = \prod_{j=(1-n)/2}^{(n-1)/2} (z^2 q^j + 2zx + q^{-j}) = \sum_{k=0}^{2n} \sum_{i=0}^{M} b_{ik} x^i z^k. \tag{4.6}$$

Replace x by $x/(2z)$ to get

$$\sum_{k=0}^{2n} \sum_{i=0}^{M} b_{ik} 2^{-i} x^i z^{k-i} = \prod_{j=(1-n)/2}^{(n-1)/2} (z^2 q^j + x + q^{-j}). \tag{4.7}$$

Replace z^2 by z, then x by x^{-1}, and multiply by x^n to get

$$\sum_{k=0}^{2n} \sum_{i=0}^{M} b_{ik} 2^{-i} x^{n-i} z^{(k-i)/2} = \prod_{j=(1-n)/2}^{(n-1)/2} (zxq^j + 1 + xq^{-j}). \tag{4.8}$$

Replace z by z/x to get

$$\sum_{k=0}^{2n}\sum_{i=0}^{M} b_{ik}2^{-i}x^{n-(i+k)/2}z^{(k-i)/2} = \prod_{j=(1-n)/2}^{(n-1)/2}(zq^j + 1 + xq^{-j}). \tag{4.9}$$

Now (1.12) follows easily from (4.9), completing the proof of Theorem 3.

We close this section by proving the remarks made in §1 between the statements of Theorems 2 and 3.

Let $n > 1$. Then the upper bound π/n in (1.9) is best possible. For, if α is slightly larger than π/n, then $E_2 < 0$ for sufficiently small x, since

$$E_2 = 4x^2\frac{\sin(n\alpha/2)\sin((n-1)\alpha/2)}{\sin(\alpha/2)\sin(\alpha)} + \frac{\sin(n\alpha)}{\sin(\alpha)}. \tag{4.10}$$

If $\alpha \geq 2\pi/n$, there is no θ in the interval (1.4) for which all coefficients of $p(z)$ are positive, by (3.11) and (3.12). Finally, suppose that

$$0 < \alpha < 2\pi/n. \tag{4.11}$$

Then all coefficients of $p(z)$ are positive on a small interval (1.10), i.e., for x sufficiently close to 1. To see this, it suffices to show that when $x = 1$ (and (4.11) holds), all coefficients of $p(z)$ are positive.

By (2.8), when $x = 1$,

$$p(z) = \prod_{j=(1-n)/2}^{(n-1)/2}(q^j z + 1)^2, \tag{4.12}$$

so

$$p(z) = \left(\sum_{\nu=0}^{n} C(n,\nu)z^\nu\right)^2, \tag{4.13}$$

where the $C(n,\nu)$ are central Gaussian coefficients (see [5, p. 449]). By (4.11) and Theorem 3 of [5, p. 449], all of the $C(n,\nu)$ are positive. Thus, by (4.13), all coefficients of $p(z)$ are positive when $x = 1$, $0 < \alpha < 2\pi/n$.

5. Application to Theorem 4. Let $f(z), f_t(z)$ be given by (1.13) and (1.14), and suppose that $f(z) \neq f_t(z)$. We will use Theorem 2 to show that all coefficients of $f_t(z)$ are positive.

Case 1. $t < 2\pi/k$. We have

$$f(z) = g(z)/h(z), \tag{5.1}$$

where

$$g(z) = \frac{z^{mk}-1}{z-1}, \qquad h(z) = \frac{1-z^k}{1-z}, \tag{5.2}$$

so

$$f_t(z) = g_t(z)/h_t(z). \tag{5.3}$$

However, in Case 1, $h_t(z) = h(z)$, so by (5.3),

$$f_t(z) = g_t(z)/h(z) = (g_t(z)(1-z))\,(1 + z^k + z^{2k} + \cdots)\,. \tag{5.4}$$

Let

$$d = \text{degree}\,(g_t(z)). \tag{5.5}$$

By Theorem 5 with $N = mk$, $g_t(z)$ is strictly unimodal, so all terms of $g_t(z)(1-z)$ of degree $\leq d/2$ have positive coefficients. Therefore, by (5.4), all terms of $f_t(z)$ of degree $\leq d/2$ have positive coefficients. However, $f_t(z)$ has degree $d-(k-1) \leq d$ by (5.4), so since the coefficients of $f_t(z)$ are symmetric about the middle one, they are all positive.

Case 2. $t \geq 2\pi/k$. If m is even, say $m = 2M$, then

$$f(z) = \frac{z^{Mk}-1}{z^k-1} \cdot (z^{Mk}+1)\,. \tag{5.6}$$

Applying Theorem 5, we could then deduce the result by induction on m. Thus assume that m is odd, so -1 is not a zero of $f(z)$. We have

$$f(z) = \prod_{r=1}^{m-1} A^{(r)}(z), \tag{5.7}$$

where

$$A^{(r)}(z) = \prod_{\substack{0<\nu<mk/2 \\ \nu \equiv r \pmod m}} \left(z - e^{2\pi i\nu/mk}\right)\left(z - e^{-2\pi i\nu/mk}\right). \tag{5.8}$$

Thus,

$$f_t(z) = \prod_{r=1}^{m-1} A_t^{(r)}(z), \tag{5.9}$$

with

$$A_t^{(r)}(z) = \prod_{\substack{mkt/2\pi<\nu<mk/2 \\ \nu \equiv r \pmod m}} \left(z - e^{2\pi i\nu/mk}\right)\left(z - e^{-2\pi i\nu/mk}\right). \tag{5.10}$$

For any fixed r, the zeros of $A_t^{(r)}(z)$ on the upper half of the unit circle can be written in the form

$$\exp(i(\theta_r + \alpha j)), \qquad 0 \leq j \leq n_r - 1, \tag{5.11}$$

where

$$\theta_r > t \geq 2\pi/k = \alpha \tag{5.12}$$

and

$$\theta_r + \alpha(n_r - 1) < \pi < \theta_r + \alpha n_r. \tag{5.13}$$

Therefore $A_t^{(r)}(z)$ has the same form as $p(z)$ in (1.3), and furthermore,

$$\pi/2 < \theta_r + (n_r - 1)\alpha/2 < \pi \tag{5.14}$$

as in (1.5). Since, moreover, $0 < \alpha < \pi/n_r$, Theorem 2 implies that all coefficients of $A_t^{(r)}(z)$ are positive. Thus all coefficients of $f_t(z)$ are positive by (5.9).

Acknowledgment. The authors are very grateful to Richard Askey for helpful ideas supplied at the Bateman Conference in Allerton Park, April 1989.

REFERENCES

[1] R. ASKEY AND M. ISMAIL, *The Rogers q-ultraspherical polynomials*, in Approximation Theory III, E. W. Cheney, ed., Academic Press, New York, 1980, pp. 175–182.

[2] R. BARNARD, W. DAYAWANSA, K. PEARCE, AND D. WEINBERG, *Polynomials with nonnegative coefficients*, Proc. Amer. Math. Soc., to appear.

[3] T. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[4] R. EVANS AND P. MONTGOMERY, *Some unimodal polynomials whose zeros are roots of unity*, Amer. Math. Monthly, 97 (1990), pp. 432–433.

[5] I. J. SCHOENBERG, *On the zeros of the generating functions of multiply positive sequences and functions*, Ann. of Math., 62 (1955), pp. 447–471.

SIAM J. MATH. ANAL.
Vol. 22, No. 4, pp. 1183–1191, July 1991

021

CONVEXITY PRESERVING APPROXIMATION BY FREE KNOT SPLINES*

YINGKANG HU†

Abstract. In this paper the order of shape-preserving approximation of functions f in Sobolev space by free knot splines is studied. The main result is that the k-convexity of f for general k can be preserved, and the optimal order of approximation n^{-r} can be retained at the same time.

Key words. shape preserving, approximation order, free knot spline, nonlinear approximation

AMS(MOS) subject classifications. 41A15, 41A25, 41A29

1. Introduction.

In many applications, it is desired that the mathematical model preserve certain geometric properties of the data, such as monotonicity and convexity. We shall give some results which show it is possible to preserve these properties without deteriorating the order of approximation by free knot splines. More precisely, we shall prove a theorem about the approximation order of k-convex functions in the Sobolev space $\mathbf{W}_1^r[0, 1]$ by k-convex spline functions with n free knots of order r. We say a function $f \in \mathbf{C}[0, 1]$ is k-convex if its k-th forward differences $\Delta_h^k(f, x)$ are nonnegative for all choices of x and $h > 0$ such that $0 \le x < x + kh \le 1$. If f has k continuous derivatives, then f is k-convex if and only if $f^{(k)} \ge 0$. It is clear that k-convexity means f is nonnegative if $k = 0$, nondecreasing if $k = 1$, and convex if $k = 2$. We will develop our results for the interval $[0, 1]$, but they hold also for any finite interval $[a, b]$ by a linear change of variable.

Let Σ_n be the space of all splines of order r on $[0, 1]$ with n knots:

$$\Sigma_n := \Sigma_n[0, 1] := \cup\, \mathcal{S}_r(\mathbf{T}, [0, 1]) \tag{1.1}$$

with the union taken over all knot sequences $\mathbf{T}$ on $[0, 1]$ of length n. For any $f \in \mathbf{C}[0, 1]$ which is k-convex for some $0 \le k < r$, let

$$\sigma_{n,k}^*(f) := \inf\{\|f - s\|_\infty : s \in \Sigma_n, s^{(k)} \ge 0\}. \tag{1.2}$$

We now state the main theorem of this paper, whose proof will be given in §3. In this theorem and throughout the paper, M will denote $\|f^{(r)}\|_1[0, 1]$ exclusively.

THEOREM 1.1. *If f is in the Sobolev space $\mathbf{W}_1^r[0, 1]$, and if $f^{(k)} \ge 0$ for some $0 \le k < r$, then there exists an $s_0 \in \mathbf{C}^{r-2} \cap \Sigma_n$ such that $s_0^{(k)} \ge 0$ and*

$$\|f - s_0\|_\infty \le C\,\frac{M}{n^r} \tag{1.3}$$

with C depending only on r, i.e.,

$$\sigma_{n,k}^* = \mathcal{O}\,(n^{-r})\,.$$

Theorem 1.1 says that there is a free knot approximation to f which preserves k-convexity of f, has the highest smoothness of all splines of order r which are not polynomials, and retains the optimal order of approximation n^{-r} at the same time. We

*Received by the editors November 20, 1989; accepted for publication August 20, 1990.

†Department of Mathematics and Computer Science, Georgia Southern University, Statesboro, Georgia 30460-8093.

should point out that for nonconstrained free knot approximation the optimal order is achievable for a wider class of functions than $\mathbf{W}_1^r$. It would be nice if we could prove our theorem for the same class of functions, but $\mathbf{W}_1^r$ is the largest function space contained in that class whose seminorm is suitable for our constructive proof.

There are many similar results in the literature for the case of fixed knot spline approximation. DeVore [6] proved that the optimal order is achievable by monotone spline approximation with equally spaced knots in $\mathbf{L}_\infty$. He also proved in [7] a similar result about monotone approximation by polynomials. Several years later, Chui, Smith, and Ward [5] gave a simpler proof of DeVore's result in [6] and also proved the corresponding result in $\mathbf{L}_p$, $1 \le p < \infty$. More recently, Beatson gave corresponding results on k-convex approximation for $k = 0$ and 1 with arbitrary fixed knots ([2], [3]), and $k = 2$ with equally spaced knots [1]. But [3] has never been published since he found an elegant short proof to treat the cases $k = 1$ and 2 (for an arbitrary fixed knot sequence) simultaneously [4]. We emphasize again that the distinction with these results and our Theorem 1.1 is that Theorem 1.1 is for free knot spline approximation for which the same order of approximation $n^{-\alpha}$ is achievable for a wider class of functions. The simplest example of this is Kahane's theorem which says that a continuous function f can be approximated by free knot piecewise constants with the optimal order n^{-1} if and only if $f \in \mathbf{BV}$, while in the case of equally spaced knots the condition becomes $f \in \mathrm{Lip}1$, which is obviously smaller than $\mathbf{BV}$ [10].

Theorem 1.1 deals with convexity of a general order k. The proof consists of three steps. In Step 1, we construct an approximating spline $g \ge 0$ of order $r_0 := r - k$ to $f^{(k)}$ with $0 \le f^{(k)} - g = \mathcal{O}(n^{-r_0})$. This step is long and the notation is a little cumbersome but the idea is simple. We cut the interval in an appropriate way, construct some overlapping approximating polynomials on the subintervals obtained, then use a "blending lemma" (Lemma 2.5) to put them together and obtain the spline g. The k-fold integral of g is a k-convex approximation to f, but its approximation order is only n^{-r_0}. In Step 2 we first modify g, then integrate it k times on $[0, 1]$. *The key step in the proof is modifying g so as to prevent the error from building up when integrating back.* In Step 3 we estimate the error. In the proof we need some preparatory theorems and lemmas which we discuss in the section that follows.

2. Preliminary results. We begin with the following result on approximation by splines with free knots.

THEOREM 2.1. *If $f \in \mathbf{W}_1^r[0, 1]$, then for any $n \ge 1$, there exists a partition $\mathbf{T}$: $0 = t_0 < t_1 < \cdots < t_n = 1$, and a piecewise polynomial s of order $\le r$ on $\mathbf{T}$, such that*

$$\|f - s\|_\infty \le \frac{1}{(r-1)!}\frac{M}{n^r}. \tag{2.1}$$

Theorem 2.1 is well known, save for the value of the constant on the right-hand side of (2.1), [9, Thm. 7.2]. Our new proof is based on the following two lemmas which show that a good partition can be obtained by balancing the values of a certain "interval function." This idea will also be important in the proof of Theorem 1.1 in §3.

Let $\mathbf{T} := \{t_i\}$ be a partition of $[0, 1]$, and k a nonnegative integer. Let $I_i := [t_{i-1}, t_i]$ and

$$C(\mathbf{T}) := \left(\sum_{i=1}^{n} \frac{1}{|I_i|^k}\right)^{-1}.$$

We define the interval function

$$F_i := F(t_{i-1}, t_i) := F^{(k)}(t_{i-1}, t_i) := |I_i|^k \|f^{(r)}\|_1(I_i),$$

and the ratio of its largest and smallest values

$$R(\mathbf{T}) := \frac{\max_i F_i}{\min_i F_i}. \tag{2.2}$$

This R is considered 1 whenever all F_i are equal, even if they are all equal to zero.

LEMMA 2.2. *For any partition* $\mathbf{T}$ *of* $[0, 1]$, *we have*

$$C(\mathbf{T}) \leq \frac{1}{n^{k+1}} \tag{2.3}$$

and

$$F_i \leq R(\mathbf{T})C(\mathbf{T})M \leq R(\mathbf{T})\frac{M}{n^{k+1}}. \tag{2.4}$$

Proof. We can assume $k \geq 1$ and $\min_i F_i > 0$, for the cases $k = 0$ or $\min_i F_i = 0$ are trivial. We make use of the following inequality, which is a special case of (2.9.1) in [8]:

$$\left(\frac{1}{n}\sum_{i=1}^{n} a_i{}^{\alpha}\right)^{1/\alpha} \leq \left(\prod_{i=1}^{n} a_i\right)^{1/n} \leq \left(\frac{1}{n}\sum_{i=1}^{n} a_i{}^{\beta}\right)^{1/\beta}, \tag{2.5}$$

which holds whenever $\alpha < 0 < \beta$, $a_i > 0$, $i = 1, \cdots, n$. In (2.5) we set $\alpha = -k$, $a_i = |I_i|$, $\beta = 1$, and obtain

$$\left(\frac{1}{n}\sum_i \frac{1}{|I_i|^k}\right)^{-1/k} \leq \frac{1}{n}\sum_i |I_i| = \frac{1}{n},$$

or

$$\left(\frac{1}{n}\sum_i \frac{1}{|I_i|^k}\right)^{-1} \leq \frac{1}{n^k}.$$

This gives (2.3). For (2.4), we simply note

$$M = \sum_i \|f^{(r)}\|_1(I_i) = \sum_i \frac{F_i}{|I_i|^k} \geq \frac{\min_i F_i}{C(\mathbf{T})}. \tag{2.6}$$

Thus

$$1 \leq \frac{C(\mathbf{T})M}{\min_i F_i},$$

and

$$F_i \leq \max_i F_i \leq \frac{\max_i F_i}{\min_i F_i} C(\mathbf{T})M = R(\mathbf{T})C(\mathbf{T})M. \qquad \square$$

Remark. With $k = r - 1$, the above lemma and Taylor's theorem give

$$\|f - s\|_\infty \leq \frac{R(\mathbf{T})}{(r-1)!}\frac{M}{n^r} \tag{2.7}$$

for some piecewise polynomial s on $\mathbf{T}$. Therefore the ratio $R(\mathbf{T})$ (with $k = r - 1$) measures the quality of a partition $\mathbf{T}$ with respect to this error bound.

LEMMA 2.3. *There exists a partition* $\mathbf{T}^*$: $0 = t_0^* < t_1^* < \cdots < t_n^* = 1$, *such that* $R(\mathbf{T}^*) = 1$, *and*

$$F_i = C(\mathbf{T}^*)M, \qquad i = 1, \cdots, n. \tag{2.8}$$

If $M > 0$ *and* $k \geq 1$, *the partition is unique.*

Proof. The cases $M = 0$ or $k = 0$ are trivial. So we suppose $M > 0$ and $k \geq 1$. We observe that for any t_{i-1} and t_i, $F(t_{i-1}, t_i)$ continuously increases when t_{i-1} decreases and/or t_i increases. If $F(t_{i-1}, t_i) > 0$, it increases strictly.

Now we consider the continuous function $g(t_1, \cdots, t_{n-1}) := \max_i F_i - \min_i F_i$ defined on the compact subset $D := \{(t_1, \cdots, t_{n-1}) : 0 \leq t_1 \leq \cdots \leq t_{n-1} \leq 1\}$ of $\mathbf{R}^{n-1}$, and suppose g assumes its minimum at $(t_1^*, \cdots, t_{n-1}^*)$. We also suppose, towards a contradiction, that this minimum is not zero. Let $A := \max_i F(t_{i-1}^*, t_i^*) > 0$ and $\Lambda := \{j : F_j = A\}$. There must be an index j such that $j \in \Lambda$ but one of its neighbors, say $j - 1$, is not. That is, $F_{j-1} < F_j = A$. We slightly increase the point t_{j-1}^* so that $F_{j-1} \leq F_j < A$. This makes $\#\Lambda$ smaller by one. We keep doing this until Λ is empty, and end up with a new point in D which gives a smaller value for g than that at $(t_1^*, \cdots, t_{n-1}^*)$, i.e., than the minimum of g. This contradiction shows that $\min_D g = 0$, or $R(\mathbf{T}^*) = 1$. It is obvious from the definition that all $F_i > 0$ for such $\mathbf{T}^*$, thus all t_i^* must be distinct. For $\mathbf{T}^*$, we have $|I_i|^k \|f^{(r)}\|_1(I_i) = F_i =: C_0$ with C_0 a constant. Therefore by (2.6) $M = C_0 \sum |I_i|^{-k} = C_0 C(\mathbf{T}^*)^{-1}$ which is (2.8). The uniqueness is due to the (strict) monotonicity of F_i. □

Theorem 2.1 follows from the lemma and (2.7).

In Step 1 of the proof of our main theorem, after finding a partition by balancing subintervals, we shall construct some nonnegative overlapping polynomials which approximate $f^{(k)}$ from below, then construct the spline g by putting these local approximants together. The following lemma from Beatson [2] will be used to guarantee the existence of the local polynomials with the required properties.

LEMMA 2.4 (Beatson, [2, Lemma 2.1]). *Let* $g \in \mathbf{C}^n[a, b]$ *be nonnegative. Let* p^* *be a best restricted uniform approximation to* g *from* $W := \{p \in \mathbf{P}_n[a, b] : 0 \leq p(x) \leq g(x), x \in [a, b]\}$, *and let* $r := g - p^*$. *Then*

$$\|r^{(i)}\|_\infty \leq (b - a)^{n-i} \omega(g^{(n)}, [a, b], b - a), \qquad i = 0, \cdots, n. \tag{2.9}$$

Lemma 2.5 below enables us to put the local polynomials together to get the smooth spline g, which, roughly speaking, approximates the function no worse than those polynomials. The first version of the lemma was established by Beatson in the special case of $g_1 = 0$, and equally spaced knots. DeVore then gave a simpler proof of the special case. The proof of Beatson in [2] is a modification of this one.

LEMMA 2.5 (Beatson, [2, Lemma 3.2]). *Let* $r \geq 2$ *be an integer and* $d = 2(r-1)^2$. *Let* $\mathbf{T} = \{t_i\}_{i=-\infty}^{\infty}$ *be a strictly increasing knot sequence with* $t_0 = a$ *and* $t_d = b$. *Let* g_1, g_2 *be two polynomials of degree* $< r$. *Then there exists a spline* $g \in \mathcal{S}_r(\mathbf{T})$ *such that*

(1) $g(x)$ *is a number between* $g_1(x)$ *and* $g_2(x)$ *for each* $x \in [a, b]$,

(2) $g = g_1$ *on* $(-\infty, a]$ *and* $g = g_2$ *on* $[b, \infty)$.

In Step 2 of the proof of Theorem 1.1, we shall modify the spline g, without destroying its positivity, of course, before integration to prevent the error from building up. This will lead to a system of linear equations with some undetermined coefficients. The undetermined coefficients should be chosen so that the system has a nonnegative solution, but it is not trivial to do so directly. We shall make use of the following Banach perturbation lemma (Theorem 2.6) to change the system into a limiting form

which is relatively easier to solve. The final stage of the proof was greatly helped by DeVore, who pointed out that this is a moment problem and so made it possible for the author to solve it beyond the case of $k \leq 3$. The author then studied the moment theory and related areas, which resulted in the final elegant proof for the case of general k, based on the well-known Gauss–Jacobi quadrature (Theorem 2.7.)

THEOREM 2.6 [11, Thm. 4.3.6]. *Let $\|\cdot\|$ be a matrix norm such that $\|XY\| \leq \|X\| \cdot \|Y\|$ for any $n \times n$ matrices X and Y, and $\|I\| = 1$, where I is the identity matrix. Let A and E be two $n \times n$ matrices with A invertible. If $\|A^{-1}E\| < 1$ then $A + E$ is also invertible and*

$$\frac{\|A^{-1} - (A+E)^{-1}\|}{\|A^{-1}\|} \leq \frac{\|A^{-1}E\|}{1 - \|A^{-1}E\|}.$$

THEOREM 2.7 [12, Thms. 3.3.1, 3.4.1, 3.4.2]. *If $y_1, y_2, \cdots, y_n$ denote the zeros of the nth orthogonal polynomial p_n with respect to the positive distribution $d\alpha$ on $[a, b]$, then*

(1) *They are all real, distinct, and located in (a, b);*

(2) *There exist positive numbers (Christoffel numbers) $\lambda_1, \lambda_2, \cdots, \lambda_n$ such that*

$$\int_a^b p(x)\, d\alpha(x) = \sum_{j=1}^{n} \lambda_j p(y_j)$$

for all polynomials p of degree $\leq 2n-1$. The distribution $d\alpha$ and the integer n uniquely determine these numbers λ_j.

3. Proof of the main theorem. We begin with some comments which will simplify our analysis. We shall define below an integer l, depending only on r and k and bounded by a function of r, and prove that if $n = ml$, $m = 1, 2, \cdots$, then $\sigma^*_{n,k}(f) \leq CMm^{-r}$. The theorem then holds for general n, because if $n < l$, we can enlarge the constant C; if $lm \leq n < l(m+1)$ for some $m \geq 1$, then

$$\begin{aligned}\sigma^*_{n,k}(f) \leq \sigma^*_{lm,k}(f) \leq CMm^{-r} &= \left[C\left(l\frac{m+1}{m}\right)^r\right] M[l(m+1)]^{-r} \\ &< [C(2l)^r]Mn^{-r} =: C'Mn^{-r}.\end{aligned}$$

We can also assume that $M > 0$. Otherwise $f \in \mathbf{P}_{r-1}[0, 1]$ and we simply set $s_0 = f$. As mentioned before, we prove the theorem in three steps. We shall use our previous notation $r_0 := r - k$ for the order of the spline function g below.

Step 1. We shall construct a nonnegative spline g of order r_0 which approximates $f^{(k)}$ *from below* and whose k-fold definite integral has an approximation order n^{-r_0}. Later, in Step 2, we shall make corrections to this g and integrate the resulting spline $\hat{g}$ to obtain the desired approximant to f. The reason we approximate $f^{(k)}$ *from below* is that the corrections will be in the positive direction and the positivity of g will then be automatically preserved.

Case 1. $r_0 = 1$ (i.e., $k = r - 1$). This is a trivial case in which g is a nonnegative piecewise constant approximant to $f^{(r-1)}$. Let $l = k(r_0+1)+1 = 2r-1$. By Lemmas 2.2 and 2.3, there exists a partition $\mathbf{X}$: $0 = x_0 < x_1 < \cdots < x_m = 1$ such that

$$F^{(r-1)}(x_{i-1}, x_i) := |J_i|^{r-1}\|f^{(r)}\|_1(J_i) = C(\mathbf{X})M \leq Mm^{-r},$$

where $J_i := [x_{i-1}, x_i]$, $1 \leq i \leq m$. This estimate, and (3.1) in Case 2, will control the local errors of the final approximant s_0 to f. We apply Lemma 2.4 to $f^{(k)}$ on each J_i. Then there is a constant $\alpha_i \geq 0$ such that on J_i

$$0 \leq f^{(k)} - \alpha_i \leq \omega(f^{(k)}, J_i, |J_i|) \leq \|f^{(r)}\|_1(J_i) = F^{(r_0-1)}(x_{i-1}, x_i).$$

Define $g(x) := \alpha_i$, for $x \in (x_{i-1}, x_i)$, we obtain the desired approximant to $f^{(k)}$.

Case 2. $r_0 \geq 2$. Let $l = 2[2(r_0-1)^2 + 1 + k(r_0+1)] + 1 \leq 6r^2 + 2r + 3$. Then, as in Case 1, there is a partition (which we index by even integers): $0 = x_0 < x_2 < \cdots < x_{2m} = 1$ such that

$$F^{(r-1)}(x_{2i-2}, x_{2i}) := |J_{2i}|^{r-1}\|f^{(r)}\|_1(J_{2i}) \leq Mm^{-r}, \tag{3.1}$$

where $J_i := [x_{i-2}, x_i]$. Since $M > 0$, we have $F^{(r-1)}(x_{2i-2}, x_{2i}) > 0$ (see the proof of Lemma 2.3). We choose the points $x_1, x_3, \cdots, x_{2m-1}$ in the following way. We choose x_1 as the midpoint of $[x_0, x_2]$. We then find x_3 such that $F^{(r-1)}(x_1, x_3) = F^{(r-1)}(x_0, x_2)$. Since $r = r_0 + k \geq 2$, we have $r - 1 \geq 1$ and therefore the F_i are strictly increasing as the first variable decreases and/or the second increases (see the proof of Lemma 2.3 again). We have $x_2 < x_3 < x_4$. We continue in this way to find $x_5, x_7, \cdots, x_{2m-1}$, and end up with a partition $\mathbf{X}$: $0 = x_0 < x_1 < x_2 < \cdots < x_{2m} = 1$ such that $F^{(r-1)}(x_{i-2}, x_i)$ are all equal for $2 \leq i \leq 2m$.

If we apply Lemma 2.4 to $f^{(k)}$ on each $J_i := [x_{i-2}, x_i]$, $2 \leq i \leq 2m$, we know there is a polynomial $p_i \geq 0$ of degree $< r_0$ such that

$$\begin{aligned} 0 \leq f^{(k)} - p_i &\leq |J_i|^{r_0-1}\omega(f^{(r-1)}, J_i, |J_i|) \\ &\leq |J_i|^{r_0-1}\|f^{(r)}\|_1(J_i) = F^{(r_0-1)}(x_{i-2}, x_i). \end{aligned}$$

We then blend these overlapping polynomials as follows: We insert $2(r_0-1)^2 + 1$ simple knots in each $I_i := (x_{i-1}, x_i)$ for $1 \leq i \leq 2m$ in an arbitrary way, and then apply Lemma 2.5 to each pair p_i, p_{i+1}, $2 \leq i \leq 2m-1$, to obtain a spline g_i of order r_0 satisfying the conditions (1) and (2) in the lemma. Define g by setting $g := g_i$ on I_i, $2 \leq i \leq 2m-1$, $g := g_2 = p_2$ on I_1, and $g := g_{2m-1} = p_{2m}$ on I_{2m}. Then g is a spline on the knot sequence consisting of the knots inserted above as well as those in $\mathbf{X}$.

From the construction of g, Lemmas 2.4 and 2.5 show that $g \geq 0$ and on I_i

$$0 \leq f^{(k)} - g \leq \begin{cases} \max\limits_{j \in \{i, i+1\}} F^{(r_0-1)}(x_{j-2}, x_j), & 2 \leq i \leq 2m-1, \\ F^{(r_0-1)}(x_0, x_2), & i = 1, \\ F^{(r_0-1)}(x_{2m-2}, x_{2m}), & i = 2m. \end{cases} \tag{3.2}$$

Step 2 (error correction). Having constructed g, a natural candidate for our s_0 is the solution $\bar{s}_0$ of the following initial value problem:

$$\begin{aligned} &\bar{s}_0^{(k)}(x) = g(x), \\ &\bar{s}_0^{(j)}(0) = f^{(j)}(0), \qquad j = 0, \cdots, k-1. \end{aligned}$$

First we take a look at what kind of errors this choice would give. Denote the error function by $r(x) := f(x) - \bar{s}_0(x)$, then $r^{(k)}(x) = f^{(k)}(x) - g(x)$. The first k derivatives of $r(x)$ at x_1, for example, will be

$$\begin{aligned} r^{(k-1)}(x_1) &= \int_{x_0}^{x_1} r^{(k)}(t)\,dt \quad =: e_0, \\ r^{(k-2)}(x_1) &= \int_{x_0}^{x_1} (x_1 - t)\, r^{(k)}(t)\,dt \quad =: e_1, \\ &\vdots \\ r(x_1) &= \frac{1}{(k-1)!}\int_{x_0}^{x_1} (x_1 - t)^{k-1}\, r^{(k)}(t)\,dt =: \frac{1}{(k-1)!} e_{k-1}. \end{aligned} \tag{3.3}$$

Here we have used the Taylor's expansions at x_0. In the trivial case that $r^{(k)}$ is identically zero on $[x_0, x_1]$ no correction is necessary. Hence, attention can now be restricted to the case where $\int_{x_0}^{x_1} r^{(k)}(t)\,dt > 0$. In this case we have to control all of them or they would build up and end up with, roughly speaking, order $n^{-r_0} = n^{k-r}$.

To correct these errors, we will add some spline $m_1(x) = \sum_{j=1}^{k} \bar{\lambda}_j M_j(x)$ to g, where $\bar{\lambda}_j \geq 0$, M_j are the B-splines of order r_0 on some knots which will be inserted into I_1 later, with $\int M_j(t)\,dt = 1$, such that

$$\int_{x_0}^{x_1} (x_1 - t)^i m_1(t)\,dt = e_i, \qquad i = 0, \cdots, k-1, \tag{3.4}$$

i.e., we will add a nonnegative spline supported on I_1 to g which will eliminate all the errors in the first k derivatives at x_1. We need to show that there exist coefficients $\bar{\lambda}_j$ and knots for the B-splines M_j which satisfy (3.4). For this we use the mean value theorem to rewrite (3.4) as

$$e_i = \sum_{j=1}^{k} \bar{\lambda}_j \int_{x_0}^{x_1} (x_1 - t)^i M_j(t)\,dt = \sum_{j=1}^{k} \bar{\lambda}_j (x_1 - \xi_{i,j})^i \tag{3.5}$$

or

$$\bar{A}\bar{\lambda} = \mathbf{e}, \tag{3.6}$$

where $\bar{\lambda} := (\bar{\lambda}_1, \cdots, \bar{\lambda}_k)^T$, $\mathbf{e} := (e_0, \cdots, e_{k-1})^T$, and $\bar{A}$ is the $k \times k$ matrix $((x_1 - \xi_{i,j})^i)_{i,j}$, with $\xi_{i,j}$ lying in $\mathrm{Supp}(M_j)$. We shall choose in I_1 the supporting knots for M_j such that (3.6) has a *nonnegative* solution $\bar{\lambda}$. This is difficult to solve directly, so we first consider an ideal problem in which the knots of M_j are allowed to coalesce. That is, suppose $M_j = \delta_{y_j}$ are the Dirac functions with a unit mass at some $y_j \in I_1$, then $\xi_{i,j} = y_j$ will depend only on j. In this case it is easy to see that we can choose the points y_j so that (3.5) has a solution $\lambda := (\lambda_1, \cdots, \lambda_k)^T$ with $\lambda_j > 0$:

$$A\lambda = \mathbf{e}, \quad \text{where } A = ((x_1 - y_j)^i)_{i,j}\,. \tag{3.7}$$

Namely, we apply the Gauss–Jacobi quadrature (Theorem 2.7) to the integrals in (3.3) with $n := k$, and $d\alpha(t) := r^{(k)}(t)\,dt \geq 0$, and find

$$e_i = \int_{x_0}^{x_1} (x_1 - t)^i r^{(k)}(t)\,dt = \sum_{j=1}^{k} \lambda_j (x_1 - y_j)^i, \qquad i = 0, \cdots, k-1,$$

for some $y_j \in I_1$ and $\lambda_j > 0$. Since all the y's are distinct, A is a Vandermonde matrix, thus invertible. We can write the solution as $\lambda = A^{-1}\mathbf{e}$.

We can now solve the original system (3.6) as a perturbation of this ideal problem. By Theorem 2.6, if we insert into I_1 for each M_j $r_0 + 1$ simple knots which are sufficiently close to y_j, then $\bar{A}$ will be so close to A (in the ∞-norm) that it will also be invertible, and the solution $\bar{\lambda} = \bar{A}^{-1}\mathbf{e}$ of (3.6) will be so close to λ that it will be nonnegative. This proves the existence of m_1.

We call m_1 a correcting function on I_1 since it completely corrects all the errors in the first k derivatives of g accumulated on I_1. We are now ready to construct the final approximant s_0 to f: we find a correcting function $m_i(x)$ for each I_i, and let $\hat{g}(x) = g(x) + \sum_i m_i(x) \geq 0$. Since $\mathrm{Supp}(m_i) \subset I_i$, we have

$$\hat{g}(x) = g(x) + m_i(x) \quad \text{for } x \in I_i. \tag{3.8}$$

Let $s_0(x)$ be the solution of the initial value problem

$$\begin{aligned} s_0^{(k)}(x) &= \hat{g}(x), \\ s_0^{(j)}(0) &= f^{(j)}(0), \qquad 0 \le j < k. \end{aligned} \tag{3.9}$$

Then s_0 is a k-convex $\mathbf{C}^{r-2}$ spline function of order r on the knot sequence consisting of all the knots we inserted: x_i's (called the test points), those inserted in the blending, and those for M_j's. The only thing that remains to show is the approximation order of s_0, which we will establish in Step 3 below.

Step 3 (error estimate). It is easy to see from our construction that s_0 interpolates f k times at each test point x_i, that is to say, the situation at each x_i is exactly the same as that at $x_0 = 0$. Therefore the errors on different subintervals I_i are independent of each other, and can be estimated locally. We shall do this on I_1 for the case $r_0 \ge 2$ only. Estimates for other I_i and the case $r_0 = 1$ are similar.

Since $r^{(k)}(x) \ge 0$, we have

$$\begin{aligned} 0 \le e_{k-1} &:= \int_{x_0}^{x_1} (x_1 - t)^{k-1} r^{(k)}(t)\, dt \le |I_1|^{k-1} \int_{x_0}^{x_1} (f^{(k)} - g)(t)\, dt \\ &\le |I_1|^{k-1}\, |I_1|\, F^{(r_0-1)}(x_0, x_2) \le |J_2|^{r-1} \|f^{(r)}\|_1(J_2) \le M m^{-r}. \end{aligned}$$

Here (3.2) and (3.1) have been used. Therefore, using the Taylor expansion at x_0, (3.9), and (3.8), we have, for any $x \in [x_0, x_1]$,

$$\begin{aligned} |f(x) - s_0(x)| &= C_0 \left| \int_{x_0}^{x} (x - t)^{k-1} (f^{(k)} - s_0^{(k)})(t)\, dt \right| \\ &= C_0 \left| \int_{x_0}^{x} (x - t)^{k-1} (f^{(k)} - g - m_1)(t)\, dt \right| \\ &\le C_0 \left| \int_{x_0}^{x_1} (x_1 - t)^{k-1} r^{(k)}(t)\, dt \right| + C_0 \left| \int_{x_0}^{x_1} (x_1 - t)^{k-1} m_1(t)\, dt \right|, \end{aligned}$$

where $C_0 := 1/(k-1)!$. The last two integrals are both equal to e_{k-1} by (3.3) and (3.4), therefore

$$|f(x) - s_0(x)| \le 2C_0\, e_{k-1} \le \frac{2}{(k-1)!} \frac{M}{m^r}.$$

This finishes our proof. □

Acknowledgments. I would like to thank Professor Ronald A. DeVore for interesting me in this problem, and for his patient guidance and valuable ideas, especially those for the case of general k.

REFERENCES

[1] R. K. BEATSON, *Convex approximation by splines*, SIAM J. Math. Anal., 12 (1981), pp. 549–559.

[2] ———, *Restricted range approximation by splines and variational inequalities*, SIAM J. Numer. Anal., 19 (1982), pp. 372–380.

[3] ———, *Monotone approximation by splines by means of restricted range approximation to the first derivative*, preprint, 1981.

[4] ———, *Monotone and convex approximation by splines: error estimates and a curve fitting algorithm*, SIAM J Numer. Anal., 19 (1982), pp. 1278–1285.

[5] C. K. CHUI, P. W. SMITH, AND J. D. WARD, *Degree of L_p approximation by monotone splines*, SIAM J. Math. Anal., 11 (1980), pp. 436–447.

[6] R. A. DEVORE, *Monotone approximation by splines*, SIAM J. Math. Anal., 8 (1977), pp. 891–905.

[7] ———, *Monotone approximation by polynomials*, SIAM J. Math. Anal., 8 (1977), pp. 906–921.

[8] G. H. HARDY, J. E. LITTLEWOOD, AND G. PÓLYA, *Inequalities*, Cambridge University Press, Cambridge, London, 1934.

[9] P. P. PETRUSHEV AND V. A. POPOV, *Rational Approximation of Real Functions*, Encyclopedia of Mathematics and Its Applications, Vol. 28, G.-C. Rota, ed., Cambridge University Press, Cambridge, 1987.

[10] L. SCHUMAKER, *Spline Functions, Basic Theory*, John Wiley, New York, 1981.

[11] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, London, 1973.

[12] G. SZEGÖ, *Orthogonal Polynomials*, Fourth ed., American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1193–1221, September 1991

THE LARGE DEFORMATION OF NONLINEARLY ELASTIC TUBES IN TWO-DIMENSIONAL FLOWS*

MASSIMO LANZA DE CRISTOFORIS† AND STUART S. ANTMAN‡

Abstract. This paper treats the large deformation of closed nonlinearly elastic cylindrical tubes (rings) under an external pressure field generated by the steady, irrotational, two-dimensional flow of an incompressible, inviscid fluid. The flow is assumed to have a prescribed velocity U and pressure P at infinity. The deformation of the tubes is described by a geometrically exact theory of rods. The parameters U and P and the deformed shape of the ring uniquely determine the velocity field of the steady flow exterior to the tube. It is shown that velocity of the flow on the tube depends continuously and compactly on the function describing the shape. The pressure field on the ring, depending on U, P, and the velocity field on the tube, is substituted into the equilibrium equations for the tube, yielding a system of ordinary functional-differential equations. These are converted into a fixed-point form, which is analyzed by a global implicit function theorem. Refined results from conformal mapping theory are used to handle serious technical difficulties with regularity, which apparently do not arise in the study of flows past rigid obstacles.

Key words. fluid-solid interactions, two-dimensional perfect flows, nonlinearly elastic tubes, global continuation theory, boundary behavior of conformal mappings

AMS(MOS) subject classifications. 30C99, 34B15, 47H10, 73K10, 73J06, 76B99

1. Introduction.

In this paper we study the large deformation of closed nonlinearly elastic cylindrical shells (tubes) produced by the action of an external pressure field generated by the steady, irrotational, two-dimensional flow of an incompressible, inviscid fluid. The flow is assumed to have a prescribed velocity U and pressure P at infinity. The two-dimensionality of our problem means that the generators of the cylindrical shell in any deformed configuration remain perpendicular to the flow so that every section of the shell perpendicular to the generators suffers the same deformation. The equations for the deformation of a typical section are those for the planar deformation of a ring. We accordingly refer to such a section as a *ring*. We describe the deformation of these rings with a geometrically exact theory of rods (cf. Antman and Rosenfeld (1978)) that accounts for flexure, compression, and shear. We allow the material properties of the ring to be described by a very general class of nonlinear constitutive relations. We limit our attention to problems having a line of symmetry.

We begin our analysis by observing that U, P, and the deformed shape of the ring uniquely determine the velocity field of the steady flow exterior to the ring. Then we

*Received by the editors July 17, 1989; accepted for publication (in revised form) August 22, 1990.

† Dipartimento di Matematica Pura ed Applicata, Università di Padova, Via Belzoni 7, 35131 Padova, Italy. The research of this author was supported in part by the Consiglio Nazionale delle Ricerche of Italy.

‡ Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. The research of this author was supported in part by grants from the National Science Foundation and the Air Force Office of Scientific Research.

show that the velocity of the flow on the ring depends continuously and compactly on the function describing the shape. We use Bernoulli's theorem to express the pressure field on the ring in terms of U, P, and the velocity field on the ring. We substitute this pressure into the equilibrium equations for the elastic ring. We transform these equations into a fixed-point equation for the shape, involving a compact operator and depending parametrically upon U and P. We apply a generalization of the Global Implicit Function Theorem of Alexander and Yorke (1976) to these equations to deduce the existence of connected families of solutions. In this program we encounter serious technical difficulties in showing that the pressure on the rod depends continuously and compactly on an appropriate function describing the shape and in constructing a suitable fixed point equation. To handle these difficulties we determine detailed regularity properties of the operator describing the dependence of the boundary values of the pressure field on the shape of the rod. (These properties can be and, to the best of our knowledge, have been ignored in the study of flows past rigid obstacles. Consequently, the presence of a deformable obstacle raises the fluid mechanical part of our problem to a level of difficulty higher than that for rigid obstacles.) To obtain the requisite properties of the pressure field we exploit refined results from conformal mapping theory. For clarity of exposition we relegate this technical analysis of compactness to §6.)

Thus we replace the coupled problem for the deformation of the ring and the external flow of the fluid with a single problem for the deformation of the ring in which the pressure field on it depends nonlocally on its shape. In this process it is necessary to marry the material (Lagrangian) description, which is natural for the ring, with the spatial (Eulerian) description, which is natural for the fluid. This fact contributes to the complexity of the formulation.

Two of the goals of this paper are to treat the nonlinear deformation of a solid under the pressure field arising from a correctly posed theory of fluid mechanics (rather than by some ad hoc approximation of the pressure) and to develop effective methods for treating well set nonlinear problems from mechanics with nonlocal terms. (Similar problems arise in electromagnetism.) Finnila and Sloss (1966) discussed a problem related to ours for a solid linearly elastic cylinder by reducing it to an integral equation.

Let us define a *solution-parameter pair* to consist of the parameters (U, P) and a set of functions that determine a configuration of the ring. (Each solution-parameter pair determines the flow.) Our fundamental result is Theorem 5.38. Informally it says that there is a connected two-dimensional family of classical solution-parameter pairs of our problem that has some very nice topological properties.

Notation. Partial derivatives are denoted by subscripts and ordinary derivatives by primes. If f and g are functions of u and v, then $\partial(f, g)/\partial(u, v)$ denotes the *matrix* of partial derivatives of f and g with respect to u and v. We denote the closure of a set $\mathcal{E}$ by $\operatorname{cl}\mathcal{E}$ and the set of elements belonging to set $\mathcal{A}$ and not belonging to set $\mathcal{B}$ by $\mathcal{A} \setminus \mathcal{B}$.

The space of m-times continuously differentiable functions on a compact set (usually the interval $[-L, L]$) with the usual norm $\|\cdot\|_m$ is denoted C^m. The domain of the functions under consideration will be evident from the context. The subspace of C^m whose functions have mth derivatives that are Hölder continuous with exponent α are denoted $C^{m,\alpha}$; If u is in $C^{0,\alpha}$ on a region Ω, then its α-Hölder quotient is

$$|u|_\alpha \equiv \sup\left\{\frac{|u(x) - u(y)|}{|x - y|^\alpha},\ x, y \in \Omega,\ x \neq y\right\}.$$

The space $C^{m,\alpha}$ on an interval of $\mathbb{R}$ or on a region of $\mathbb{C}$ is equipped with the usual norm :$\|u\|_{m,\alpha} \equiv \|u\|_m + |u^{(m)}|_\alpha$ where $u^{(m)}$ is the mth derivative of u.

To conduct the analysis effectively, especially in §6, it is necessary to employ a very precise notation. For example, the pressure at a material point s on the ring depends on the whole shape of the ring, defined by ζ, and on U and P. We denote it by $p[\zeta, U, P](s)$. We need such a notation in §6, where we must carry out limit processes simultaneously in the variables ζ and s.

2. Equilibrium equations for nonlinearly elastic rings. Since the problems we treat are two-dimensional, we express all vectorial quantities in complex notation. We describe the deformation of the ring by the planar equations for a rod that can suffer flexure, extension, and shear (cf. Antman and Rosenfeld (1978)). The configuration of such a (Cosserat) rod is a plane curve every point of which is equipped with a coplanar unit vector. The theory is geometrically exact in that no variable characterizing such a configuration (such as the sine of an angle) is ever approximated (by the angle itself or any finite Taylor sum for the sine). The plane curve characterizes the gross behavior of a slender body and can be identified with the deformed shape of some material base curve. The unit vector characterizes the deformed configuration of a material cross-section. Such a body can sustain forces and torques. Indeed, the theory we employ is the most general for which equilibrium can be completely described by equations involving no stress resultants other than the classical contact force and contact couple. (Moreover, it is more than adequate for many practical purposes.)

Formally, a *configuration of a ring* is specified by the pair of continuously differentiable functions

$$[-L, L] \ni s \mapsto \big(\zeta(s),\ \theta(s)\big) \in \mathbb{C} \times \mathbb{R} \tag{2.1a}$$

with $\zeta \equiv x(s) + iy(s)$ a simple closed curve having positive (counterclockwise) orientation so that

$$\zeta(-L) = \zeta(L),\quad \zeta'(-L) = \zeta'(L),\quad \theta(L) = \theta(-L) + 2\pi. \tag{2.1b,c,d}$$

Such a configuration is an intrinsically one-dimensional concept. Solely for the sake of interpretation, we can regard (ζ, θ) as defining a deformation of the two-dimensional body

$$\mathcal{B} \equiv \{Z(s) + \xi e^{i\Theta(s)} \in \mathbb{C} :\ -L \le s \le L,\ 0 \le \xi \le h(s)\} \tag{2.2}$$

where $s \mapsto (Z(s), \Theta(s)) \in \mathbb{C} \times \mathbb{R}$ is a prescribed continuously differentiable function satisfying (2.1b,c,d) and

$$Z' = -ie^{i\Theta} \tag{2.3}$$

and h is a prescribed positive-valued continuous function representing the thickness of the ring. (Equation (2.2) defines the reference (undeformed) configuration of the body.) h must satisfy the relation $h(-L) = h(L)$. (See Fig. 2.4.) A point in $\mathcal{B}$ is called a *material point.* Equation (2.3) requires that $Z'(s)$ be a unit vector normal to $e^{i\Theta(s)}$. This condition is not critical; it can be relaxed at the cost of complicating a few formulas. The requirement that there be a locally one-to-one orientation-preserving correspondence between the material points of $\mathcal{B}$ and pairs (s, ξ) of curvilinear coordinates in $\{(s, \xi) : -L \le s < L,\ 0 \le \xi \le h(s)\}$ is ensured by

$$h(s)\Theta'(s) < 1. \tag{2.5}$$

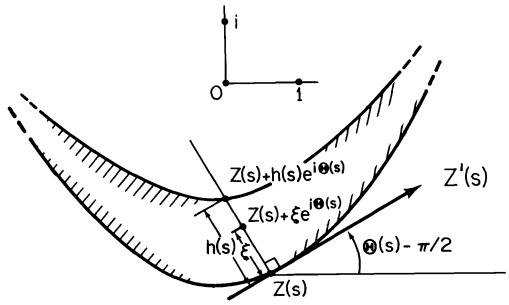


FIG. 2.4. *Reference configuration $\mathcal{B}$ of the ring regarded as a two-dimensional body.*

This condition implies that the curvature vectors from nearby points of Z do not intersect inside $\mathcal{B}$. One could formulate mild additional geometric restrictions on $\mathcal{B}$ to ensure that the correspondence be global. We assume that this is so.

s may be interpreted as the scaled arc length parameter of the bounding curve Z of the ring $\mathcal{B}$ in its undeformed reference configuration. Thus s identifies a *material section* $\xi \mapsto Z(s)+\xi e^{i\Theta(s)}$ of the ring. The configuration (ζ, θ) may then be interpreted as defining the deformation

$$\mathcal{B} \ni Z(s) + \xi e^{i\Theta(s)} \mapsto \zeta(s) + \xi e^{i\theta(s)}. \tag{2.6}$$

Thus $\zeta(s)$ may be interpreted as the deformed position of the material point $Z(s)$ originally on the bounding curve Z of the ring in its reference configuration $\mathcal{B}$, and θ may be interpreted as characterizing the orientation of the deformed configuration of the section s with respect to the positive real axis. $\Theta(s)$ is the given value of $\theta(s)$ in the reference configuration. (It is more common to interpret Z not as a bounding curve of $\mathcal{B}$, but rather as its curve of centroids. Since the material points of Z are those in contact with the fluid, our choice of interpretation proves to be more natural for the problems we study.)

We can regard (2.6) as constraining the deformations of $\mathcal{B}$ to be those for which plane sections remain plane and do not change their length. Since we are not confined to representations of the form (2.6), we can assign more general interpretations to (ζ, θ). A virtue of the abstract definition (2.1a) is that it admits any one of a multitude of such interpretations and is not tied to (2.6).

We denote differentiation with respect to s by a prime. We decompose the vector $\zeta'(s)$ tangent to the curve ζ at $\zeta(s)$ by

$$\zeta'(s) \equiv -i(\nu + i\eta)e^{i\theta} \tag{2.7}$$

where ν and η are real-valued. The function ν can be thought of as measuring elongation. (It actually contributes to the volume change; see (2.10).) The function η

measures shear. We set

$$\psi \equiv \theta - \Theta, \qquad \mu \equiv \psi'(s). \tag{2.8a,b}$$

μ measures the amount of bending. Then

$$\mathbf{q} \equiv (\nu, \eta, \mu) \tag{2.9}$$

is the set of *strains* for our rings.

If we adopt the two-dimensional interpretation (2.6) of (ζ, θ) given above, then the requirement that the local ratio of deformed-to-reference volume be positive leads to the inequality

$$\nu(s) - \xi\theta'(s) > 0 \quad \forall \xi \in [0, h(s)], \tag{2.10}$$

which is equivalent to

$$\nu(s) > \max\{0, h(s)[\mu(s) + \Theta'(s)]\} \equiv \kappa\big(\mu(s), s\big). \tag{2.11}$$

$\kappa(\cdot, s)$ is convex. This is the essential property of κ that we require in our analysis. (This convexity arises no matter what reasonable interpretation is given to (ζ, θ); cf. Antman (1976).) We require that (2.11) hold and for simplicity we take κ to have the specific form shown there. The positivity of ν following from (2.11) ensures that the local ratio of deformed to reference length of ζ at s be positive and that the section at s not be sheared so severely that $e^{i\theta(s)}$ is parallel to $\zeta'(s)$. We require ζ to be simple. We do not bother to impose a global injectivity restriction on (2.6).

Let $-i[\check{N}(s) + i\check{H}(s)]e^{i\theta(s)}$ denote the resultant contact force and $\check{M}(s)$ the resultant contact couple acting across the section s of the ring exerted by the material of $(s, s+\epsilon]$ on the material of $[s-\epsilon, s]$ where ϵ is a small positive number. By definition of contact force and couple, these resultants are independent of ϵ. ($\check{N}$, $\check{H}$, $\check{M}$ are each real-valued.) We assume that the only external force applied to the ring at s is a hydrodynamical pressure of intensity $p(s)$ per unit deformed length acting in the direction $i\zeta'(s)$. (This assumption means that the fluid lies to the right of the oriented curve ζ. A natural sign convention allows for the possibility, remote from the focus of our analysis, that the pressure be negative.) Then the equilibrium equations for the ring are

$$\{[\check{N}(s) + i\check{H}(s)]e^{i\theta(s)}\}' = p(s)\zeta'(s), \tag{2.12}$$

$$\check{M}'(s) - \operatorname{Re}\{\zeta'(s)[\check{N}(s) - i\check{H}(s)]e^{-i\theta(s)}\} = 0. \tag{2.13}$$

We can put these equations into the componential forms

$$\check{N}'(s) = \theta'(s)\check{H}(s) + p(s)\eta(s), \tag{2.14}$$

$$\check{H}'(s) = -\theta'(s)\check{N}(s) - p(s)\nu(s), \tag{2.15}$$

$$\check{M}'(s) = \eta(s)\check{N}(s) - \nu(s)\check{H}(s). \tag{2.16}$$

Let

$$\mathcal{Q}(s) \equiv \{\mathbf{q} \in \mathbb{R}^3 : \nu > \kappa(\mu, s)\}, \tag{2.17}$$

$$\mathcal{Q} \equiv \{(\mathbf{q}, s) : \mathbf{q} \in \mathcal{Q}(s), s \in [-L, L]\}. \tag{2.18}$$

The material of the ring is *elastic* if there are constitutive functions

$$\mathbf{Q} \equiv (N, H, M) : \mathcal{Q} \to \mathbb{R}^3 \tag{2.19}$$

such that

$$\check{N}(s) = N\big(\mathbf{q}(s), s\big), \text{ etc.} \tag{2.20}$$

We assume that the constitutive functions (2.19) are thrice continuously differentiable. In particular, in consonance with (2.1) we assume that $\mathbf{Q}(\mathbf{q}, \cdot)$, when extended to $\mathbb{R}$ by periodicity, is twice continuously differentiable. $\mathbf{Q}$ is required to satisfy the *monotonicity condition*:

$$\text{The matrix } \mathbf{Q}_{\mathbf{q}}(\mathbf{q}, s) \equiv \frac{\partial(N, H, M)}{\partial(\nu, \eta, \mu)}(\mathbf{q}, s) \text{ is positive definite } \forall\ (\mathbf{q}, s) \in \mathcal{Q}. \tag{2.21}$$

This assumption can be shown to be inherited from the strong ellipticity condition of three-dimensional elasticity in the process by which rod theories are constructed from the three-dimensional theory (cf. Antman (1976)). It ensures, e.g., that an increase in the bending strain μ is accompanied by a corresponding increase in the bending couple M. We require that:

$$\eta \mapsto N(\nu, \eta, \mu, s) \text{ is even}, \qquad \eta \mapsto H(\nu, \eta, \mu, s) \text{ is odd}, \tag{2.22a,b}$$

$$\eta \mapsto M(\nu, \eta, \mu, s) \text{ is even}, \tag{2.22c}$$

$$N(1, 0, 0, s) = 0, \qquad M(1, 0, 0, s) = 0. \tag{2.23a,b}$$

Conditions (2.22b), (2.23) ensure that the reference configuration is stress-free.

Remark. There is no obstacle to extending our entire analysis to a rod theory with any level of complexity. (Such theories can be interpreted as replacing (2.6) with any expression having more degrees of freedom.) We would merely have to augment our geometrical and mechanical variables with a larger set, many of whose members would lack simple interpretations; cf. Antman (1976) and Antman and Carbone (1977).

3. Steady, two-dimensional, irrotational flows of incompressible, inviscid fluids. The force acting on the ring is generated by the two-dimensional, steady, irrotational flow of an incompressible inviscid fluid. No body force (such as gravity) acts on the fluid. The velocity and pressure fields are assumed to have limits at infinity (in the complex plane $\mathbb{C}$). In this section we formulate the governing equations for the motion of these fluids. The derivations can be found in standard books on fluid dynamics (cf., e.g., Serrin (1959) and Milne-Thomson (1968)).

Let $\mathcal{F}$ denote the (open) domain of $\mathbb{C}$ occupied by the fluid. We assume that $\mathcal{F}$ is of class C^1. Let $w(z) = u(z) - iv(z)$, $p(z)$, ρ denote the complex velocity, pressure, and density at point $z = x + iy$ in the closure $\mathrm{cl}\,\mathcal{F}$ of $\mathcal{F}$. ρ is a given positive number. u, v, p are real-valued. Under our assumptions, Euler's equations for the flow in $\mathcal{F}$ are the *momentum equations*

$$\rho(uu_x + vu_y) = -p_x, \qquad \rho(uv_x + vv_y) = -p_y \ \text{ in } \mathcal{F}, \tag{3.1}$$

the *incompressibility condition*

$$u_x + v_y = 0 \ \text{ in } \mathcal{F}, \tag{3.2}$$

and the *irrotationality condition*

$$u_y - v_x = 0 \ \text{ in } \mathcal{F}. \tag{3.3}$$

Note that (3.2), (3.3) are the Cauchy–Riemann equations so that w must be a holomorphic function in $\mathcal{F}$.

We assume that

(3.4a,b) $$w(z) \to U, \qquad p(z) \to P \quad \text{as } z \to \infty$$

where U, P are given nonnegative real numbers. Then (3.1), (3.3), (3.4) yield Bernoulli's equation

(3.5) $$p(z) = -\frac{\rho}{2}|w(z)|^2 + \frac{\rho}{2}U^2 + P \quad \text{in } \mathcal{F},$$

which shows that p in $\mathcal{F}$ is determined by w, U, P. This formula can be extended by continuity to $\operatorname{cl}\mathcal{F}$ whenever w can be so extended.

Equations (3.2) and (3.3) imply that on any simply-connected subset of $\mathcal{F}$ there exists a holomorphic function $\Omega = \Phi + i\Psi$ such that

(3.6) $$w(z) = \Omega'(z).$$

Φ is the *velocity potential*, Ψ is the *stream function*, and Ω is the *complex potential*. In §6 we observe that under the assumptions we shall impose on the flow, the *circulation*

(3.7) $$\int_\zeta w(z)\,dz = 0,$$

which is precisely the supplementary condition needed to show that for this problem Ω is also defined on all of $\mathcal{F}$. Thus Ω is holomorphic on $\mathcal{F}$ while Φ and Ψ satisfy the Cauchy–Riemann equations here.

Since the boundary $\partial\mathcal{F}$ of $\mathcal{F}$ is solid, fluid cannot penetrate it, i.e.,

(3.8) $$\operatorname{Re}(nw) = 0 \quad \text{on } \partial\mathcal{F}$$

where n is the unit inner normal to $\partial\mathcal{F}$. It follows from (3.6) and (3.8) that $\Psi =$ const. on $\partial\mathcal{F}$. Since $\mathcal{F}$ is at worst doubly connected, there is no loss in generality in taking this constant to be zero:

(3.9) $$\Psi = 0 \quad \text{on } \partial\mathcal{F}.$$

The solution of (3.1)–(3.4), (3.8) can thus be effected by finding a holomorphic function Ω on $\mathcal{F}$ that together with its derivative admits a continuous extension to $\operatorname{cl}\mathcal{F}$ and that satisfies (3.9) and

(3.10) $$\Omega'(z) \to U \quad \text{as } z \to \infty.$$

Then w and p can be found from (3.6), (3.5).

In the sequel, when it is necessary to emphasize the dependence of $\mathcal{F}$, Ω, etc., on ζ, U, we write $\mathcal{F}[\zeta]$, $\Omega[\zeta, U]$, etc.

4. The symmetrically deformed ring. In this section we give a detailed description of the class of functions in which we seek solutions of our equations. In the next section we carry out a global analysis of the symmetric deformation of a ring in a symmetric flow under the assumption that the pressure field on the ring has the right form. We deduce the properties of the pressure field in §6, where we analyze the flow problem.

We record (2.1b,c,d):

(4.1a,b,c) $$\zeta(-L) = \zeta(L), \quad \zeta'(-L) = \zeta'(L), \quad \theta(L) = \theta(-L) + 2\pi.$$

To ensure that the flow does not apply a net torque to the ring and that the complex potential Ω is well defined, we consider only configurations that are symmetric with respect to the x-axis and have the property that $\operatorname{Re}\zeta(\pm L) > \operatorname{Re}\zeta(0)$ (see Fig. 4.2):

(4.3a,b) $$\zeta(-s) = \overline{\zeta}(s), \qquad \theta(s) = -\theta(-s).$$

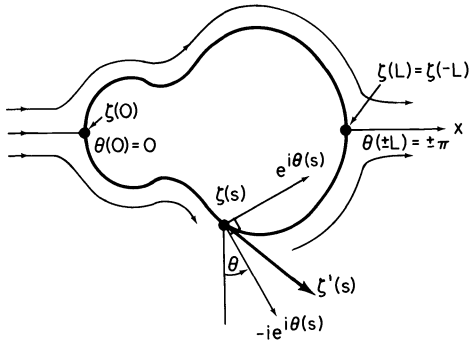


FIG. 4.2. *Flow about a symmetrically deformed ring.*

Using (2.7), we readily deduce from (4.3) that

(4.4a,b,c) $$\nu(s) = \nu(-s), \quad \eta(s) = -\eta(-s), \quad \mu(s) = \mu(-s).$$

By comparing (4.1) with (4.3), we obtain the obvious conditions

(4.5a,b,c) $$y(0) = 0, \quad x'(0) = 0, \quad \theta(0) = 0$$

together with

(4.6a,b,c) $$y(\pm L) = 0, \quad x'(\pm L) = 0, \quad \theta(\pm L) = \pm\pi.$$

Note that (4.3) and (those equations of) (4.6) (with a plus sign) imply (4.1). For (4.3) to be reasonable, we require that the reference configuration satisfy it and that the constitutive functions N, H, M, h be even in s. It is convenient to fix the rigid translation of ζ by requiring that

(4.7) $$\zeta(0) = Z(0).$$

In order to carry out compactness proofs needed to meet the hypotheses of the Global Implicit Function Theorem (stated below), we require a convenient characterization of simple, closed, continuously differentiable ζ's. Let us extend ζ to $\mathbb{R}$ by periodicity. Let

(4.8a) $$\sigma(s_1, s_2) \equiv \min\{|t_1 - t_2| : e^{i\pi t_\alpha/L} = e^{i\pi s_\alpha/L}, \ \alpha = 1, 2\},$$

(4.8b) $$l[\zeta] \equiv \inf\left\{\left|\frac{\zeta(s) - \zeta(t)}{\sigma(s,t)}\right| : s, t \in [-L, L], \ s \neq t\right\}.$$

For $\delta > 0$, we set

(4.9) $$\mathcal{Z}(\delta) \equiv \{\zeta \in C^1 : (4.1a,b) \text{ holds}, \ l[\zeta] > \delta\},$$

(4.10) $$\mathcal{Z} \equiv \bigcup_{\delta>0} \mathcal{Z}(\delta).$$

LEMMA 4.11. *$\mathcal{Z}$ is precisely the set of all simple, closed, continuously differentiable ζ's (with $|\zeta'|$ everywhere positive).*

Proof. It is a straightforward exercise to show that an element of $\mathcal{Z}(\delta)$, and hence of $\mathcal{Z}$, is a simple, closed, continously differentiable curve with a nowhere vanishing derivative. Let us prove the converse. Suppose for contradiction that ζ is such a curve with $l[\zeta] = 0$. Then by the definition of l (and by the Bolzano–Weierstrass theorem) there are sequences$\{s_k\}$ and $\{t_k\}$ in $[-L, L]$ with $s_k \leq t_k$ converging to s_∞ and t_∞ such that

$$\lim_{k\to\infty} \frac{|\zeta(s_k) - \zeta(t_k)|}{|\sigma(s_k, t_k)|} = 0. \tag{4.12}$$

If $s_\infty \neq t_\infty \mod 2L$, then $\zeta(s_\infty) \neq \zeta(t_\infty)$ and (4.12) cannot hold. If $s_\infty = t_\infty$, then (4.12) implies that $\zeta'(s_\infty) = 0$, a contradiction. If $s_\infty = -L$, $t_\infty = L$, then for k sufficiently large, $\sigma(s_k, t_k) = |t_k - s_k - 2L|$. Since $\zeta(t_k) = \zeta(t_k - 2L)$, equation (4.12) implies that $\zeta'(-L) = 0$, another contradiction. □

LEMMA 4.13. *$\mathcal{Z}(\delta)$ and $\mathcal{Z}$ are open in $\{\zeta \in C^1 : (4.1a,b)\ holds\}$.*

Proof. Let $\zeta \in \mathcal{Z}(\delta)$ with $\delta > 0$, let $\chi \in \{\zeta \in C^1 : (4.1a,b)\ \text{holds}\}$, and let$\|\zeta - \chi\|_1 < \frac{1}{2}(l[\zeta] - \delta)$. We show that $\chi \in \mathcal{Z}(\delta)$ by showing that $l[\chi] > \delta$. Definitions (4.8) imply that

$$\begin{aligned} l[\chi] &\geq \inf\left\{\left|\frac{\zeta(s) - \zeta(t)}{\sigma(s,t)}\right| - \left|\frac{(\zeta - \chi)(s) - (\zeta - \chi)(t)}{\sigma(s,t)}\right| : s, t \in [-L, L],\ s \neq t\right\} \\ &\geq l[\zeta] - \|\zeta - \chi\|_1 > \frac{1}{2} l[\zeta] + \frac{\delta}{2} > \delta. \end{aligned} \tag{4.14}$$

In deriving this inequality we have used the inequality for the mean value. □

Remark. When interpretation (2.6) is used, it would be physically appropriate to require the deformed inner boundary curve $s \mapsto \zeta(s) + h(s)e^{i\theta(s)}$ of the annular region $\mathcal{B}$ of (2.2) to be simple, closed, and positively oriented. We do not bother to pursue the technical adjustments needed to accommodate this requirement.

In §6 we prove the following theorem.

THEOREM 4.15. *Let* (4.1a,b) *and* (4.3a) *hold. For $\zeta \in \mathcal{Z}$ let $w[\zeta, U](s)$, which is independent of P, denote the complex velocity of the (symmetric) flow at the point $\zeta(s)$. Then $U^{-1}w[\zeta, U]$ is independent of U (and is thus solely determined by ζ). The function $s \mapsto U^{-1}w[\zeta, U](s)$ is continuous. The operator $C^2 \cap \{\zeta \in \mathcal{Z} :$ (4.3a) holds$\} \ni \zeta \mapsto |U^{-1}w[\zeta, U]| \in C^0$ is continuous. It is compact on $C^2 \cap \{\zeta \in \mathcal{Z}(\delta) :$ (4.3a) holds$\}$ for every $\delta > 0$. The pressure on the ring at $\zeta(s)$ is given by*

$$p[\zeta, U, P](s) = P + \tfrac{\rho}{2} U^2 \{1 - U^{-2}|w[\zeta, U](s)|^2\} \tag{4.16}$$

and satisfies $p[\zeta, U, P](s) = p[\zeta, U, P](-s)$. Moreover,

$$p[\cdot, \cdot, \cdot] : [C^2 \cap \{\zeta \in \mathcal{Z} :\ (4.3a)\ holds\}] \times \mathbb{R}^2 \to C^0 \tag{4.17}$$

is continuous and is compact on $[C^2 \cap \{\zeta \in \mathcal{Z}(\delta) :\ (4.3a)\ holds\}] \times \mathbb{R}^2$ for every $\delta > 0$.

Note that these fluid-dynamical variables are independent of θ, i.e., independent of the state of shear in the rod.

We introduce the Banach space

$$\mathcal{V} \equiv \{(\zeta, \psi) \in C^2 : (4.1a,b),\ (4.3)\ \text{hold}\} \tag{4.18}$$

endowed with the C^2–norm. (We need this norm in order to use Theorem 4.15.) We define the operators $\hat{\mathbf{q}} \equiv (\hat{\nu}, \hat{\eta}, \hat{\mu})$, $\hat{\theta}$ by

$$\hat{\mu}[\psi] \equiv \psi', \quad \hat{\theta}[\psi](s) \equiv \Theta(s) + \int_0^s \psi'(\xi)\, d\xi, \tag{4.19}$$

$$\hat{\nu}[\zeta, \psi] + i\hat{\eta}[\zeta, \psi] \equiv ie^{-i\hat{\theta}[\psi]}\zeta'. \tag{4.20}$$

Let us set

$$\begin{aligned} &\mathcal{Z}^+(\delta) \equiv \{\zeta \in \mathcal{Z}(\delta) : \zeta \text{ has positive orientation}\}, \\ &\mathcal{Z}^+ \equiv \{\zeta \in \mathcal{Z} : \zeta \text{ has positive orientation}\}. \end{aligned} \tag{4.21}$$

We define $\mathcal{Z}^-(\delta)$ and $\mathcal{Z}^-$ analogously for ζ's having negative orientation. Lemma 4.13 holds with $\mathcal{Z}^\pm(\delta)$, $\mathcal{Z}^\pm$ replacing $\mathcal{Z}(\delta)$, $\mathcal{Z}$. We shall seek solutions of our operator equations in the following subset of $\mathcal{V}$:

$$\mathcal{A} \equiv \{(\zeta, \psi) \in \mathcal{V} : \zeta \in \mathcal{Z}^+,\ \operatorname{Re}\zeta(\pm L) > \operatorname{Re}\zeta(0),\ \hat{\mathbf{q}}[\zeta, \psi](s) \in \mathcal{Q}(s)\ \forall s\}. \tag{4.22a}$$

Clearly $\mathcal{A}$ contains the pair $(Z, 0)$ corresponding to the reference configuration (Z, Θ). Note that $\mathcal{A}$ does not account for (4.1c). We accordingly define

$$\mathcal{A}^\sharp \equiv \{(\zeta, \psi) \in \mathcal{A} : (4.1c) \text{ holds}\}. \tag{4.22b}$$

We assume that $\Theta \in C^2$. In the reference configuration $\mathbf{q} = (1, 0, 0)$ (by definition of this configuration). It then follows from (2.7) that Z is in C^3. Our *boundary value problem* (BVP) is to find $(\zeta, \psi) \in \mathcal{A}^\sharp$ such that the strain-configuration relations (4.19), (4.20), the equilibrium equations (2.14)–(2.16) with $p(s)$ replaced with $p[\zeta, U, P](s)$, defined in (4.16), and the constitutive equations (2.20) hold. Note that the reference configuration (Z, Θ) satisfies BVP for $U = 0, P = 0$. In the rest of the paper we regard (2.8a) as an identity and express restrictions on ψ in terms of θ or vice versa.

5. The fixed-point problem. We now recast BVP as a fixed-point problem to which we can apply the following theorem, whose interpretation follows its statement.

THEOREM 5.1 (Global Implicit Function Theorem). *Let $\mathcal{X}$ be a Banach space and let $\{\mathcal{O}(\epsilon), 0 < \epsilon < E\}$ be a family of open sets (not necessarily bounded) in $\mathcal{X} \times \mathbb{R}^m$ for which*

$$(0, 0) \in \mathcal{O}(\epsilon), \qquad 0 < \epsilon < E, \tag{5.2}$$

$$\operatorname{cl}\mathcal{O}(\epsilon_2) \subset \mathcal{O}(\epsilon_1) \quad \text{for } 0 < \epsilon_1 < \epsilon_2 < E. \tag{5.3}$$

Let

$$\mathcal{O} \equiv \bigcup_{0<\epsilon<E} \mathcal{O}(\epsilon). \tag{5.4}$$

Let $F : \mathcal{O} \to \mathcal{X}$ be continuous, let $F(0, 0) = 0$, and let $F : \mathcal{O}(\epsilon) \to \mathcal{X}$ be compact for $0 < \epsilon < E$, where E is a given positive number. Let I denote the identity operator on $\mathcal{X}$. Let the Fréchet derivative $I - F_x(0, 0) : \mathcal{X} \to \mathcal{X}$ of $x \mapsto x - F(x, \lambda)$ at $(0, 0)$ exist and be invertible. Let $\mathcal{S} \equiv \{(x, \lambda) \in \mathcal{O} : x = F(x, \lambda)\}$ and let $\mathcal{S}_0$ be the connected component of $\mathcal{S}$ containing $(0, 0)$. (In a neighborhood of $(0, 0)$ $\mathcal{S}$ agrees with $\mathcal{S}_0$.) Then one of the following statements is true:

(i) *$\mathcal{S}_0$ is bounded and there is an $\epsilon^* \in (0, E)$ such that $\mathcal{S}_0 \subset \mathcal{O}(\epsilon^*)$. There is an essential map (i.e., a continuous map not homotopic to a constant) σ from $\mathcal{S}_0$ onto the m-dimensional sphere $\mathbb{S}^m$ whose restriction to $\mathcal{S}_0 \setminus \{(0, 0)\}$ is inessential. Moreover, $\mathcal{S}_0$ contains a connected subset $\mathcal{S}_{00}$*

that contains $(0,0)$*, that has the same properties as* $\mathcal{S}_0$ *with respect to* σ*, and that has the property that each point of it has Lebesgue dimension at least* m*.*

(ii) $\mathcal{S}_0 \setminus \mathcal{O}(\epsilon) \neq \emptyset$ *for all* $\epsilon \in (0,E)$ *or* $\mathcal{S}_0$ *is unbounded. For each* $\epsilon \in (0,E)$ *there is a modified equation* $x = \varphi(x,\lambda,\epsilon)F(x,\lambda)$ *(cf.* (7.4)*) defined on all of* $\mathcal{X} \times \mathbb{R}^m$ *that agrees with* $x = F(x,\lambda)$ *on* $\mathcal{O}(\epsilon)$*. The one-point compactification* $\mathcal{S}_0^+(\epsilon)$ *of the connected component* $\mathcal{S}_0(\epsilon)$ *containing* $(0,0)$ *of the set of solution pairs of the modified equation has the same properties as* $\mathcal{S}_0$ *in statement* (i)*.*

We prove this theorem, which generalizes results of Alexander and Yorke (1976), in §7. Without comment we shall shift the base point $(0,0)$ in this theorem to any other point. In our problem the role of λ is played by (U,P). In case (i) the theorem states that the set of solution-parameter pairs (x,λ) of the equation $x = F(x,\lambda)$ has a connected subset $\mathcal{S}_{00}$ containing $(0,0)$ and having dimension at least 2 at each point. This means that $\mathcal{S}_{00}$ contains a subset that looks like a surface. The statement about the essential map roughly means that $\mathcal{S}_0$ cannot abruptly terminate and that it looks like a sphere to which are possibly attached further solution pairs.

We set

$$\mathbf{f}(\mathbf{q},p,s) \equiv \big(-N_s+(\Theta'(s)+\mu)H+p\eta, -H_s-(\Theta'(s)+\mu)N-p\nu, -M_s+\eta N-\nu H\big) \tag{5.5}$$

where the arguments of N, N_s, etc., are $\mathbf{q}$, s. By substituting the constitutive equations (2.20) and the pressure equation (4.16) into the equilibrium equations (2.14)–(2.16), and then carrying out the indicated differentiations, we find that the system (2.14)–(2.16), (2.20), (4.16) is formally equivalent to

$$\mathbf{q}(s) = \mathbf{q}(0) + \int_0^s \mathbf{Q}_{\mathbf{q}}\big(\mathbf{q}(\xi),\xi\big)^{-1}\mathbf{f}\big(\mathbf{q}(\xi),p[\zeta,U,P](\xi),\xi\big)\,d\xi \tag{5.6}$$

$$\equiv \mathbf{q}(0) + \mathbf{g}[\mathbf{q},\zeta,U,P](s). \tag{5.7}$$

In consonance with (4.4b) we take

$$\eta(0) = 0. \tag{5.8}$$

We define the operators $\mathbf{q}^\sharp \equiv (\nu^\sharp,\eta^\sharp,\mu^\sharp)$ by

$$\mathbf{q}^\sharp[\zeta,\psi,U,P](s) \equiv \mathbf{g}\big[\hat{\mathbf{q}}[\zeta,\psi],\zeta,U,P\big](s). \tag{5.9}$$

We now use (4.7), (4.19), (4.20) to convert (5.5), (5.6) to

$$\begin{aligned}\zeta(s) &= Z(0) - i\int_0^s e^{i\hat{\theta}(\xi)}\{\nu_0 + \nu^\sharp[\zeta,\psi,U,P](\xi) + i\eta^\sharp[\zeta,\psi,U,P](\xi)\}\,d\xi \\ &\equiv Z(0) + x^\sharp[\zeta,\psi,\nu_0,U,P](s) + iy^\sharp[\zeta,\psi,\nu_0,U,P](s),\end{aligned} \tag{5.10}$$

$$\psi(s) = \mu_0 s + \int_0^s \mu^\sharp[\zeta,\psi,U,P](\xi)\,d\xi, \tag{5.11}$$

where the unknown constants ν_0 and μ_0 are equal to $\hat{\nu}(0)$ and $\hat{\mu}(0)$.

Remark. System (5.10), (5.11) would have a fixed-point form if the constants ν_0 and μ_0 were chosen to depend on (ζ,ψ). To get a useful fixed-point form in $\mathcal{V} \cap \{(\zeta,\psi) : \text{(4.1c) holds}\}$, we should like to choose ν_0 and μ_0 to ensure that the

right-hand sides of (5.10), (5.11) satisfy (4.6), i.e., so that

$$y^\sharp[\zeta,\psi,\nu_0,U,P](L)=0, \tag{5.12a}$$

$$x^\sharp[\zeta,\psi,\nu_0,U,P]'(L)=0, \tag{5.12b}$$

$$\mu_0 L+\int_0^L \mu^\sharp[\zeta,\psi,U,P](\xi)\,d\xi=0. \tag{5.12c}$$

Unfortunately, we have only two constants available to satisfy three relations. The situation is actually somewhat worse: If we compute the matrix of partial derivatives of the left-hand sides of (5.12) with respect to ν_0 and μ_0, we get

$$\begin{pmatrix} -\int_0^L \cos\hat\theta(\xi)\,d\xi & 0\\ \sin\hat\theta(L) & 0\\ 0 & L\end{pmatrix}, \tag{5.13}$$

which reduces to

$$\begin{pmatrix} 0 & 0\\ 0 & 0\\ 0 & L\end{pmatrix} \tag{5.14}$$

in the trivial state with $(U,P)=(0,0)$. Thus there is no way to use the classical implicit function theorem directly to solve (5.12) for ν_0 and μ_0 as functionals of (ζ,ψ) near the trivial solution or to get any such representation for ν_0. We could solve for μ_0 as a functional of (ζ,ψ), by defining it to be the solution of (5.12c) so that (4.6c) would hold. But while this definition of μ_0 is useful for certain problems with *constant* hydrostratic pressure (cf. Antman (1973)), it does not compensate for the degeneracy inherent in (5.14). We could replace the $\hat\theta(\xi)$ in the integrand in (5.10) with $\mu_0\xi+\int_0^\xi \mu^\sharp[\zeta,\psi,U,P](\omega)\,d\omega$ and thus introduce the parameter μ_0 elsewhere in our equations, but such a step does not ameliorate the situation. In short, the complexity of our class of admissible configurations satisfying (2.1) prevents us from setting up our fixed-point problem in the most obvious way. We now exploit the underlying mechanics and geometry of our problem to resolve this difficulty by a less straightforward procedure.

Since we have no assurance that the right-hand sides of (5.10), (5.11) take $\mathcal{A}$ or even $\mathcal{A}^\sharp$ into $\mathcal{V}$, we contrive to cast our problem in a more suitable form. We let α and β play the roles of ν_0 and μ_0. Then in place of (5.10)–(5.12) we consider

$$\begin{aligned}
x(s)&=X(0)+x^\sharp[\zeta,\psi,\alpha,U,P](s)+\frac{2L}{\pi}x^\sharp[\zeta,\psi,\alpha,U,P]'(L)\cos\frac{\pi s}{2L},\\
y(s)&=y^\sharp[\zeta,\psi,\alpha,U,P](s)-y^\sharp[\zeta,\psi,\alpha,U,P](L)\sin\frac{\pi s}{2L},\\
\psi(s)&=\beta s+\int_0^s \mu^\sharp[\zeta,\psi,U,P](\xi)\,d\xi,\\
\alpha&=\alpha+x^\sharp[\zeta,\psi,\alpha,U,P]'(L),\\
\beta&=\beta+y^\sharp[\zeta,\psi,\alpha,U,P](L).
\end{aligned} \tag{5.15a}$$

We abbreviate (5.15a) as

$$(\zeta,\psi,\alpha,\beta)=\mathbf{k}[\zeta,\psi,\alpha,\beta,U,P]. \tag{5.15b}$$

It is easy to see that $\mathbf{k}[\zeta,\psi,\alpha,\beta,U,P]$ satisfies (4.1a,b), (4.3) (and thus $\mathbf{k}$ takes $\mathcal{A}\times\mathbb{R}^4$ into $\mathcal{V}\times\mathbb{R}^2$) and that if $(\zeta,\psi,\alpha,\beta)$ in $\mathcal{A}\times\mathbb{R}^2$ satisfies (5.15), then (ζ,ψ) satisfies all the conditions of BVP, except possibly (4.1c), which is equivalent to (4.6c).

Furthermore, $\alpha = \hat{\nu}(0)$ and $\beta = \hat{\mu}(0)$ and conversely, if (ζ, ψ, U, P) in $\mathcal{A} \times \mathbb{R}^2$ satisfies BVP, except for (4.1c), then $(\zeta, \psi, \hat{\nu}(0), \hat{\mu}(0), U, P)$ satisfies (5.15). We now show that under mild conditions (ζ, ψ) must satisfy (4.6c) as well.

PROPOSITION 5.16. *Let* $(\zeta, \psi, \alpha, \beta) \in \mathcal{A} \times \mathbb{R}^2$ *satisfy* (5.15) *and*

(5.17)
$$\frac{N(\hat{\mathbf{q}}[\zeta,\psi](L), L)}{\hat{\nu}[\zeta,\psi](L)} < \frac{H(\hat{\mathbf{q}}[\zeta,\psi](L), L)}{\hat{\eta}[\zeta,\psi](L)} \equiv \int_0^1 H_\eta\big(\hat{\nu}[\zeta,\psi](L), \alpha\hat{\eta}[\zeta,\psi](L), \hat{\mu}[\zeta,\psi](L), L\big)\, d\alpha$$

(*when* $\hat{\eta}[\zeta,\psi](L) \neq 0$). *Then* (ζ, ψ) *satisfies* BVP.

Remarks. The identity in (5.17), which is used to define the limit of the second quotient as $\eta \to 0$, follows from (2.22b) and the Mean Value Theorem. By virtue of (2.21), condition (5.17) holds in the trivial state, in which its left-hand side is zero, and therefore holds in a neighborhood of this state. It also holds when $\check{N}(L) < 0$, i.e., when the material of the ring at L is in compression. We might expect this condition to hold, at least for moderate deformations, because such a compression of the ring is a likely effect of a positive pressure exerted on it by the surrounding fluid (when $P > 0$). The violation of (5.17), which could occur when the ring is under sufficiently large tension at its trailing point, is associated with a shear instability (cf. Antman and Carbone (1977)). We could show that such a possibility is precluded either by obtaining a suitable a priori estimate or else by elevating (5.17) into the (reasonable) constitutive restriction:

$$\frac{N(\mathbf{q}, L)}{\nu} < \frac{H(\mathbf{q}, L)}{\eta} \quad \forall \mathbf{q} \in \mathcal{Q}(L), \quad \eta \neq 0. \tag{5.18}$$

PROOF OF PROPOSITION 5.16. In view of the remarks preceding the statement of the proposition, it suffices to show that (4.1c) holds. Let ϕ be the angle between ζ' and the negative y-axis:

$$\zeta' = -i|\zeta'|e^{i\phi} \tag{5.19}$$

with $\phi(0) = 0$. Equation (4.20) then implies that

$$\hat{\nu} + i\hat{\eta} = |\zeta'|e^{i(\phi - \hat{\theta})}. \tag{5.20}$$

Since ζ is a simple, closed, positively oriented, continuously differentiable curve, it follows that

$$\phi(\pm L) = \pm\pi. \tag{5.21}$$

Since the membership of (ζ, ψ) in $\mathcal{A}$, defined in (4.22a), implies that (2.11) holds, it follows that $\hat{\nu}(s)$ must be positive for all s. It then follows from (5.20) that $\cos\big(\phi(s) - \hat{\theta}(s)\big) > 0$, and since $\hat{\theta}(0) = 0 = \phi(0)$ by (4.5), it follows from the continuity of $\phi - \hat{\theta}$ that $|\phi(s) - \hat{\theta}(s)| < \pi/2$. Thus (5.21) yields

$$|\hat{\theta}(L) - \pi| < \frac{\pi}{2}. \tag{5.22}$$

Now the integral of (2.12) over $(-L, L)$ yields

$$(\check{N} + i\check{H})e^{i\hat{\theta}}\Big|_{-L}^{L} = \int_{-L}^{L} p[\zeta, U, P](s)\zeta'(s)\, ds. \tag{5.23}$$

The Kutta–Joukowski Theorem stated in §6 implies that the integral on the right-hand side of (5.23), which is the net force exerted by the inviscid fluid on the deformed ring, must vanish, and consequently the left-hand side must vanish. (The fundamental "stress principle" or law of action and reaction embodied in the integrals of (2.12) over arbitrary intervals would directly imply that the left-hand side of (5.23) would vanish, were it not for the distinguished role we have assigned to $\pm L$.) We now use (2.20), (2.22a,b), (4.4a,c), and the evenness of H in s to deduce from (5.23) that

$$N\big(\hat{\mathbf{q}}(L),L\big)\sin\hat{\theta}(L)+H\big(\hat{\mathbf{q}}(L),L\big)\cos\hat{\theta}(L)=0. \tag{5.24}$$

From (2.7), (4.6b) we obtain

$$\hat{\nu}(L)\sin\hat{\theta}(L)+\hat{\eta}(L)\cos\hat{\theta}(L)=0. \tag{5.25}$$

Since (5.24), (5.25), regarded as a linear system for $\sin\hat{\theta}(L)$, $\cos\hat{\theta}(L)$, cannot have a trivial solution, we conclude that

$$\hat{\eta}(L)N\big(\hat{\mathbf{q}}(L),L\big)=\hat{\nu}(L)H\big(\hat{\mathbf{q}}(L),L\big). \tag{5.26}$$

(The arguments of the functions in (5.26) are understood to equal those of (5.17).) Condition (5.17) implies that (5.26) is satisfied if and only if $\hat{\eta}(L)=0$. From (5.20) we thus find that $\sin\big(\phi(L)-\hat{\theta}(L)\big)=0$. It follows from (5.22) that $\hat{\theta}(L)=\phi(L)=\pi$. □

We now introduce the subsets $\mathcal{O}(\epsilon)$ of the domain of definition of $\mathbf{k}$ that are used in Theorem 5.1. Let $(\zeta,\psi)\in\mathcal{A}$. We define

$$\hat{\epsilon}[\zeta,\psi]\equiv\min\big\{l[\zeta],\min\{\hat{\nu}[\zeta,\psi](s)-\kappa\big(\psi'(s),s\big)\}\big\}, \tag{5.27a}$$

$$E\equiv\hat{\epsilon}[Z,0]. \tag{5.27b}$$

For each $\epsilon\in[0,E)$, we define

$$\mathcal{O}(\epsilon)\equiv\{(\zeta,\psi,\alpha,\beta,U,P)\in\mathcal{V}\times\mathbb{R}^4:\hat{\epsilon}[\zeta,\psi]>\epsilon,\ \zeta\text{ has positive orientation}\}. \tag{5.28}$$

(Note that $\mathcal{O}(\epsilon)\subset\mathcal{A}\times\mathbb{R}^4$.) Clearly $\mathcal{O}(\epsilon)$ satisfies (5.2)–(5.4) with $\mathcal{O}=\mathcal{O}(0)=\mathcal{A}\times\mathbb{R}^4$ and with $(0,0)$ replaced with $(Z,0,1,0,0,0)$.

PROPOSITION 5.29. $\mathbf{k}$ *is continuous on* $\mathcal{A}\times\mathbb{R}^4$ *and is continuously Fréchet differentiable with respect to* $(\zeta,\psi,\alpha,\beta,P)$ *if* $U=0$. *The restriction of* $\mathbf{k}$ *to* $\mathcal{O}(\epsilon)$ *is compact for every* $\epsilon\in(0,E)$.

Proof. The continuity and Fréchet differentiability follow immediately from their definitions and from Theorem 4.15. The compactness follows from the Arzelà–Ascoli Theorem and Theorem 4.15 in a straightforward way. (Note that for arbitrary $\epsilon>0$ the set $\{(\nu,\eta,\mu,s)\in\mathcal{Q}:\nu-\kappa(\mu,s)\geq\epsilon,\ \nu+|\eta|+|\mu|\leq 1/\epsilon\}$ is a compact subset of $\mathcal{Q}$.) □

The way we have replaced $\hat{\nu}(0)$, $\hat{\mu}(0)$ with α, β was designed precisely to compensate for the degeneracy inherent in (5.14) and to enable us to prove the following proposition.

PROPOSITION 5.30. *The Fréchet derivative of* $\mathbf{I}-\mathbf{k}$ *with respect to* $(\zeta,\psi,\alpha,\beta)$ *about the trivial state* $(\zeta,\psi,\alpha,\beta,U,P)=(Z,0,1,0,0,0)$ *is an invertible linear mapping of* $\mathcal{V}\times\mathbb{R}^2$ *onto itself.*

Proof. In view of Proposition 5.29, we can use the Open Mapping Theorem and the Fredholm Alternative to reduce the proof to showing that this derivative is injective, i.e., to showing that the linearization of (5.15) about the trivial solution has the unique zero solution. By an argument analogous to that leading up to Proposition 5.16, we can show that the linearization of (5.15) is equivalent to the linearization of BVP. (In particular, the treatment of the linearization of (4.6c) follows that for (4.6c)

itself.) This linearization of (5.15) can be formally obtained by assuming that all the variables depend upon a small parameter τ with the trivial state corresponding to $\tau = 0$, then by differentiating the governing equations with respect to τ, and finally by setting $\tau = 0$. The value of the derivative of a variable such as ν with respect to τ at $\tau = 0$ is denoted $\dot{\nu}$. Variables with dots are the unknowns of the linearization. The values of constitutive functions at the trivial state are identified by the superscript 0. Thus, e.g., $N_\nu^0(s) \equiv N_\nu(1,0,0,s)$. In deriving the linearization we have used (4.4) and the evenness of $\mathbf{Q}(\mathbf{q},\cdot)$ to show that various constitutive functions, such as N_η are zero in the trivial state. We use (4.16) to show that $\dot{p} = 0$. The linearization of BVP is accordingly given by

$$\dot{x}' = \dot{\nu}\sin\Theta + (\dot{\eta}+\dot{\theta})\cos\Theta, \quad \dot{y}' = -\dot{\nu}\cos\Theta + (\dot{\eta}+\dot{\theta})\sin\Theta, \quad \dot{\theta}' = \dot{\mu}, \tag{5.31a}$$

$$\begin{aligned} &\dot{x}(s) = \dot{x}(-s), \quad \dot{y}(s) = -\dot{y}(-s), \quad \dot{\theta}(s) = -\dot{\theta}(-s),\\ &\dot{\nu}(s) = \dot{\nu}(-s), \quad \dot{\eta}(s) = -\dot{\eta}(-s), \quad \dot{\mu}(s) = \dot{\mu}(-s),\\ &\dot{y}(\pm L) = 0, \qquad \dot{x}'(\pm L) = 0, \qquad \dot{\theta}(\pm L) = 0, \end{aligned} \tag{5.31b}$$

$$\dot{N}' = \Theta'\dot{H}, \quad \dot{H}' = -\Theta'\dot{N}, \quad \dot{M}' = -\dot{H}, \tag{5.31c}$$

$$\dot{N} = N_\nu^0\dot{\nu} + N_\mu^0\dot{\mu}, \quad \dot{H} = H_\eta^0\dot{\eta}, \quad \dot{M} = M_\nu^0\dot{\nu} + M_\mu^0\dot{\mu}. \tag{5.31d}$$

From (5.31c) we obtain (the principle of virtual work)

$$\int_{-L}^{L} [\dot{y}(-\dot{N}\cos\Theta + \dot{H}\sin\Theta)' + \dot{x}(\dot{N}\sin\Theta + \dot{H}\cos\Theta)' + \dot{\theta}(\dot{M}' + \dot{H})]\,ds = 0. \tag{5.32}$$

We integrate (5.32) by parts and then use (5.31a,b,d) to obtain

$$\int_{-L}^{L} [(N_\nu^0\dot{\nu} + N_\mu^0\dot{\mu})\dot{\nu} + H_\eta^0\dot{\eta}^2 + (M_\nu^0\dot{\nu} + M_\mu^0\dot{\mu})\dot{\mu}]\,ds = 0. \tag{5.33}$$

Since the integrand of (5.33) is positive definite by (2.21), we find that $\dot{\nu} = \dot{\eta} = \dot{\mu} = 0$, whence the triviality of the solution follows. □

From Propositions 5.29, 5.30 and the Implicit Function Theorem in Banach space we deduce the following theorem.

THEOREM 5.34. *Let $U = 0$. There exists a neighborhood $\mathcal{E}$ of $(\zeta,\psi,\alpha,\beta) = (Z,0,1,0)$ in $\mathcal{A}\times\mathbb{R}^2$ and a number $\tilde{P} > 0$ such that if $|P| < \tilde{P}$, then (5.15) has exactly one solution $\big(\tilde{\zeta}(P)(\cdot),\tilde{\psi}(P)(\cdot),\tilde{\alpha}(P),\tilde{\beta}(P)\big)$ in $\mathcal{E}$. The mapping taking P into this solution is continuously differentiable.*

Since the set of bounded linear invertible mappings is open in the space of bounded linear mappings of a Banach space into itself, we can choose $\tilde{P}$ so small that the Fréchet derivative $\mathbf{I}-\mathbf{k}_{(\zeta,\psi,\alpha,\beta)}\big(\tilde{\zeta}(P)(\cdot),\tilde{\psi}(P)(\cdot),\tilde{\alpha}(P),\tilde{\beta}(P),0,P\big)$ of $\mathbf{I}-\mathbf{k}$ is invertible. Accordingly, we define

$$\begin{aligned} \mathcal{P} \equiv \{P\in\mathbb{R} :\ &(5.15) \text{ has a solution } \big(\zeta^\star(P)(\cdot),\psi^\star(P)(\cdot),\alpha^\star(P),\beta^\star(P)\big) \text{ for } U = 0,\\ &\mathbf{I}-\mathbf{k}_{(\zeta,\psi,\alpha,\beta)}\big(\zeta^\star(P)(\cdot),\psi^\star(P)(\cdot),\alpha^\star(P),\beta^\star(P),0,P\big) \text{ is invertible}\}. \end{aligned} \tag{5.35}$$

By Theorem 5.34 we know that $\mathcal{P}$ contains a nonempty interval containing zero. Let

$$\mathcal{S} \equiv \{(\zeta,\psi,\alpha,\beta,U,P)\in\mathcal{A}\times\mathbb{R}^4 : (5.15) \text{ holds}\}. \tag{5.36}$$

It is the set of solution-parameter pairs of (5.15). Let $P_0 \in \mathcal{P}$ and let

$$(\zeta_0, \psi_0, \alpha_0, \beta_0) = \big(\zeta^\star(P_0)(\cdot), \psi^\star(P_0)(\cdot), \alpha^\star(P_0), \beta^\star(P_0)\big) \tag{5.37}$$

be defined as in (5.35). We define $\mathcal{S}_0$ to be the connected component of $\mathcal{S}$ containing $(\zeta_0, \psi_0, \alpha_0, \beta_0, 0, P_0)$. Proposition 5.29, the membership of P_0 in $\mathcal{P}$, and Theorem 5.1 now yield our fundamental result, which asserts the existence of a two-dimensional connected set of solution-parameter pairs of (5.15) having nice topological properties.

THEOREM 5.38. *At least one of the following statements holds:*

(5.39a) $\exists\ \epsilon_1 \in (0, E)$ *such that* $\mathcal{S}_0 \subset \mathcal{O}(\epsilon_1)$, $\mathcal{S}_0$ *is bounded,*

(5.39b) $\mathcal{S}_0 \cap [\mathcal{O} \setminus \mathcal{O}(\epsilon)] \neq \emptyset \quad \forall\ \epsilon \in (0, E)$ *or* $\mathcal{S}_0$ *is unbounded.*

If (5.39a) *holds, then there is an essential mapping* σ *from* $\mathcal{S}_0$ *to* $\mathbb{S}^2$ *whose restriction to* $\mathcal{S}_0 \setminus \{(\zeta_0, \psi_0, \alpha_0, \beta_0, 0, P_0)\}$ *is inessential. Moreover,* $\mathcal{S}_0$ *contains a subset* $\mathcal{S}_{00}$ *each point of which has topological dimension at least* 2. *The restriction of* σ *to* $\mathcal{S}_{00}$ *is essential. If* (5.39b) *holds, then for each* ϵ *there is a compact and continuous operator* $\mathbf{k}_\epsilon$ *on* $\mathcal{V} \times \mathbb{R}^4$ *such that the equation* $(\zeta, \psi, \alpha, \beta) = \mathbf{k}_\epsilon[\zeta, \psi, \alpha, \beta, U, P]$ *on* $\mathcal{V} \times \mathbb{R}^4$ *agrees with* (5.15b) *on* $\mathcal{O}(\epsilon)$. *The one-point compactification* $\mathcal{S}_0^+(\epsilon)$ *of the connected component* $\mathcal{S}_0(\epsilon)$ *of its solution pairs containing* $(0, 0)$ *has the same properties as* $\mathcal{S}_0$ *in statement* (i).

Remark. We have not yet been able to show that $\mathbf{k}$ is Fréchet differentiable with respect to $(\zeta, \psi, \alpha, \beta)$ when $U \neq 0$. For this reason we cannot invoke the local Implicit Function Theorem in Banach space to assert that $\mathcal{S}_0$ (or $\mathcal{S}$) is a C^1-surface of dimension 2 near $(\zeta_0, \psi_0, \alpha_0, \beta_0, 0, P_0)$. Such a result would be necessary to justify the most primitive of perturbation methods.

We now study solutions of (5.15) with P fixed at a value P_0 in $\mathcal{P}$.

THEOREM 5.40. *Let* $P_0 \in \mathcal{P}$ *and let* (5.37) *hold. Then there is a number* $U_1 > 0$, *depending on* P_0, *such that the set*

$$\mathcal{S}_*(P_0) \equiv \bigcup_{|U| \leq U_1} \{(\zeta, \psi, \alpha, \beta, U, P_0) \in \mathcal{S}\} \tag{5.41}$$

contains a connected subset joining the planes $U = \pm U_1$. (*Thus there is a solution for each* U *with* $|U| \leq U_1$.)

Proof. The existence of a branch bifurcating from (5.37) follows from a one-parameter version of Theorem 5.38. Since $\mathbf{I} - \mathbf{k}_{(\zeta,\psi,\alpha,\beta)}(\zeta_0, \psi_0, \alpha_0, \beta_0, 0, P)$ is invertible, it follows that $(\zeta_0, \psi_0, \alpha_0, \beta_0, 0, P)$ is an isolated solution of (5.15) in the space $\mathcal{V} \times \mathbb{R}^2 \times \{0\} \times \{P_0\}$. Thus the values of U on the branch cannot be confined to zero. Since (5.15) is even in U, the values of U must be symmetrically disposed about $U = 0$. □

Remark. These theorems describe the connectivity properties of solution sextuples $(\zeta, \psi, \alpha, \beta, U, P)$ of the modified problem (5.15). It is easy to use the relations $\alpha = \hat{\nu}(0)$ and $\beta = \hat{\mu}(0)$ to show that the same results apply to the solution quadruples (ζ, ψ, U, P) of the original BVP, except possibly for (4.1c), which can be handled with Proposition 5.16.

Remark. For P held fixed at a positive $P_0 \in \mathcal{P}$, the continuity of $p[\cdot, \cdot, \cdot]$, defined in (4.16), ensures that the pressure on the ring is everywhere positive if U is small enough. If the pressure becomes negative on part of the ring, then cavitation occurs. Note that the speed $|w|$ of the flow has a strict maximum on $\operatorname{cl}\mathcal{F}[\zeta]$ on its boundary ζ (as can be shown by mapping $\mathcal{F}[\zeta]$ onto a bounded region and applying a maximum principle). Thus the second summand in (4.16) is negative somewhere on the ring.

For P held fixed at 0 and $U \neq 0$, there must be a place on the ring where the pressure is negative.

6. The conformal mapping problem for the flow about a symmetric ring. Our main goal in this section is to prove Theorem 4.15. We use the results of §3. Let $\zeta \in C^{1,\alpha} \cap \mathcal{Z}^+$ be a simple closed positively oriented curve satisfying (4.3a). Without loss of generality we assume that 0 lies inside of ζ. Our *flow problem* FP is to find a complex potential $\Omega = \Phi + i\Psi$ holomorphic on $\mathcal{F}[\zeta]$ with Ω' continuous on $\operatorname{cl}\mathcal{F}[\zeta]$ that satisfies (3.9), (3.10), and the symmetry condition

$$\Omega(\overline{z}) = \overline{\Omega}(z) \text{ or, equivalently, } \Phi(x,y) = \Phi(x,-y). \tag{6.1a}$$

By virtue of (3.6a), this condition is in turn equivalent to

$$u(x,y) = u(x,-y), \qquad v(x,y) = -v(x,-y). \tag{6.1b}$$

Condition (6.1b) ensures (3.7), which states that the flow has zero circulation (cf. Brezis and Stampacchia (1976)). The Kutta–Joukowski Theorem states that the resultant force on the ring produced by the flow $w[\zeta,U]$ is $-i\rho U \int_\zeta w[\zeta,U](z)\,dz$. This vanishes when (3.7) holds (in which case we have d'Alembert's paradox that the resultant force on the ring is zero). The Blasius Theorem states that the resultant moment on the ring produced by the flow $w[\zeta,U]$ is $-\frac{\rho}{2}\operatorname{Re}\int_\zeta zw[\zeta,U](z)^2\,dz$. That (6.1) forces this integral to vanish is our main motivation for adopting this restriction. We can prove Theorem 4.15 when the circulation is zero without invoking (6.1), but we do not emphasize this point. At several places our analysis makes contact with standard results on existence and uniqueness of solutions to FP. It is nevertheless necessary for us to reiterate some of the steps in order to show precisely how the boundary values of solutions depend on ζ.

Let

$$\mathcal{D} \equiv \{\omega \in \mathbb{C} : |\omega| < 1\}. \tag{6.2}$$

We shall reduce FP to finding a conformal mapping f from $\mathcal{F}[\zeta]$ to $\mathbb{C}\backslash\operatorname{cl}\mathcal{D}$.

For technical reasons it is convenient to study conformal mappings from bounded sets to bounded sets. For this purpose we express f as a composition of the inversion $\mathcal{F}[\zeta] \ni z \mapsto 1/z$ followed by a mapping $g[\zeta]^{-1}$ and then followed by another inversion $\omega \mapsto 1/\omega$. See Fig. 6.3. Since our aim is to determine precisely how the flow variables depend on ζ, we must determine precisely how $g[\zeta]$ depends on ζ. Nevertheless, for notational simplicity we shall often suppress the dependence of variables on ζ in our computations, but not in the statements of important results.

We now introduce a class of boundary shapes that are compatible with the analysis of the ring problem in §§4 and 5 and that promote the analysis of the conformal mapping problem formulated below. For $0 < \alpha \leq 1$, $0 < h_1 < h_2$, $0 < \delta$ we define

$$\begin{aligned} &\mathcal{Z}^{\pm}(\alpha,h_1,h_2,\delta) \equiv \mathcal{A}(h_1,h_2) \cap C^{1,\alpha} \cap \mathcal{Z}^{\pm}(\delta), \\ &\mathcal{Z}^{\pm}(\alpha,h_1,h_2) \equiv \mathcal{A}(h_1,h_2) \cap C^{1,\alpha} \cap \mathcal{Z}^{\pm}, \\ &\mathcal{A}(h_1,h_2) \equiv \{\zeta : h_1 < |\zeta| < h_2,\ (4.3a) \text{ holds, } \operatorname{Re}\zeta(\pm L) > \operatorname{Re}\zeta(0)\}. \end{aligned} \tag{6.4}$$

DEFINITION 6.5. For $\zeta \in \mathcal{Z}^+(\alpha,h_1,h_2)$, $\mathcal{I}[\zeta]$ is the domain of $\mathbb{C}$ enclosed by the curve $1/\zeta$.

The following result is immediate.

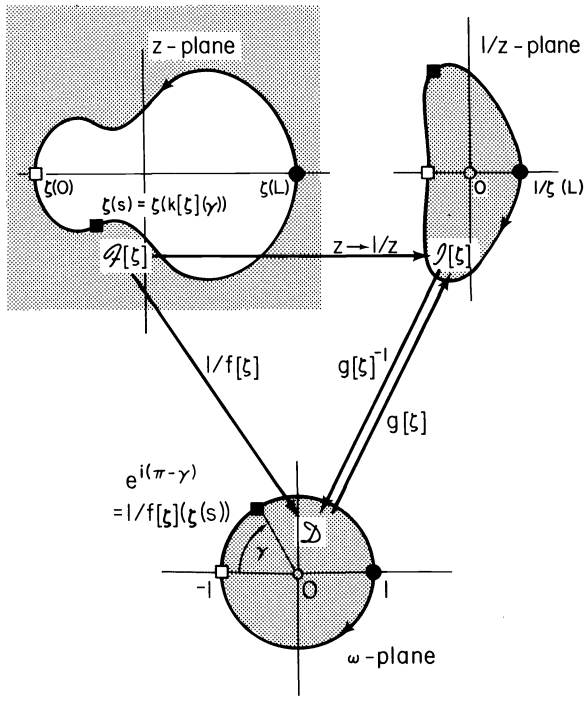


FIG. 6.3. *Conformal mappings. The positive (counterclockwise) orientation we have adopted for ς is a negative orientation for the boundary of $\mathcal{F}$. The counterclockwise orientation of ς is reversed by the inversion $z \mapsto 1/z$ so that the boundaries of $\mathcal{I}$ and $\mathcal{D}$ are shown with clockwise orientations. The parameter γ for $\partial\mathcal{D}$ corresponds to this orientation.*

LEMMA 6.6. *Let $0 < \alpha \leq 1$, $0 < h_1 < h_2$. The mapping $\mathcal{Z}^+(\alpha, h_1, h_2) \ni \zeta \mapsto 1/\zeta \in \mathcal{Z}^-(\alpha, h_2^{-1}, h_1^{-1})$ is continuous in the $C^{1,\alpha}$-norm.*

For $0 < \alpha \leq 1$, $0 < h_1 < h_2$, let $\zeta \in \mathcal{Z}^+(\alpha, h_1, h_2)$. Our *conformal mapping problem* (CMP) is to find a continuously differentiable homeomorphism

$$f : \mathrm{cl}\,\mathcal{F}[\zeta] \to \mathbb{C} \setminus \mathcal{D}, \tag{6.7a}$$

with

$$f(\partial\mathcal{F}[\zeta]) = \partial\mathcal{D}, \tag{6.7b}$$

$$f(\infty) = \infty, \tag{6.7c}$$

$$f'(\infty) \text{ exists and is a positive real number,} \tag{6.7d}$$

$$f(\overline{z}) = \overline{f}(z) \tag{6.7e}$$

such that the restriction of f to $\mathcal{F}[\zeta]$ is conformal. This problem is equivalent to that of finding a continuously differentiable homeomorphism

$$g : \operatorname{cl}\mathcal{D} \to \operatorname{cl}\mathcal{I}[\zeta] \tag{6.8a}$$

with

$$g(\partial\mathcal{D}) = \partial\mathcal{I}[\zeta], \tag{6.8b}$$

$$g(0) = 0, \tag{6.8c}$$

$$g'(0) \text{ is a positive real number,} \tag{6.8d}$$

$$g(\overline{z}) = \overline{g}(z) \tag{6.8e}$$

such that the restriction of g to $\mathcal{D}$ is conformal. We denote the solutions to these problems, each designated CMP, by $f[\zeta]$ and $g[\zeta]$. They are related by

$$f[\zeta](z) = \frac{1}{g[\zeta]^{-1}(1/z)}. \tag{6.9a}$$

We readily find that

$$f[\zeta]'(\infty) = g[\zeta]'(0). \tag{6.9b}$$

THEOREM 6.10. *Let $f[\zeta]$ satisfy* CMP. *Then* FP *has a solution*

$$\Omega[\zeta, U] = \frac{U}{f[\zeta]'(\infty)}\left\{f[\zeta] + \frac{1}{f[\zeta]}\right\}, \tag{6.11}$$

unique to within a real constant.

Proof. All the assertions, except the uniqueness, follow from direct computation. To prove uniqueness, let $\tilde{\Omega}$ be the difference of two solutions of FP. Set

$$H(\omega) \equiv \tilde{\Omega}\big(1/g(\omega)\big), \qquad \omega \in \operatorname{cl}\mathcal{D} \setminus \{0\}. \tag{6.12}$$

Since $\omega \mapsto \tilde{\Omega}'(1/g(\omega))$ is holomorphic on $\mathcal{D} \setminus \{0\}$ and tends to zero as ω tends to zero by (3.4a), it follows that

$$\lim_{\omega\to 0} \omega H'(\omega) = -\lim_{\omega\to 0} \frac{\omega^2 g'(\omega)}{g(\omega)^2} \lim_{\omega\to 0} \frac{\tilde{\Omega}'(1/g(\omega))}{\omega} \in \mathbb{C}. \tag{6.13}$$

Since (6.1a) implies that $\int_{\partial\mathcal{D}} H'(\omega)\,d\omega = \int_{\zeta} \tilde{\Omega}'(z)\,dz = 0$, it follows that H' is holomophic in $\mathcal{D}$. Conditions (3.9) and (6.8b) imply that $\operatorname{Im} H = 0$ on $\partial\mathcal{D}$. Therefore H and consequently $\tilde{\Omega}$ are real constants. □

We now obtain some preliminary lemmas that will enable us not only to solve CMP, but also to determine how solutions depend on the shape ζ of the deformed ring. Our attention is focused on the boundary behavior of g.

LEMMA 6.14. *Let $0 < \alpha \le 1$ and let $\zeta \in C^{1,\alpha}$ satisfy $|\zeta'(s)| > 0$ everywhere. Let $\phi[\zeta](s)$ be (a branch of) the angle from the negative imaginary axis to $\zeta'(s)$. Let $\phi[\zeta]$ be continuous. Then there is a positive constant a depending only on $\phi[\zeta](0)$, L, α (and otherwise independent of ζ) such that*

$$\|\phi[\zeta]\|_{0,\alpha} \le a\big(\|\zeta\|_{1,\alpha} + 1\big)^{4/\alpha}\big(\min|\zeta'|\big)^{-4/\alpha}. \tag{6.15}$$

Proof. $\phi[\zeta](s)$ satisfies $\zeta' = -i|\zeta'|e^{i\phi}$. Without loss of generality we take $\phi[\zeta](0) = 0$. Therefore,

(6.16)
$$|\sin(\phi(s) - \phi(t))| = \left|\mathrm{Re}\left\{-i\frac{\zeta'(s)}{|\zeta'(s)|}\frac{\overline{\zeta}'(t)}{|\zeta'(t)|}\right\}\right| = \left|\mathrm{Re}\left\{i\left[\frac{\zeta'(s)}{|\zeta'(s)|} - \frac{\zeta'(t)}{|\zeta'(t)|}\right]\frac{\overline{\zeta}'(t)}{|\zeta'(t)|}\right\}\right|$$
$$\leq \left|\frac{\zeta'(s)}{|\zeta'(s)|} - \frac{\zeta'(t)}{|\zeta'(t)|}\right| \leq \frac{2\|\zeta\|_{1,\alpha}^2|s-t|^\alpha}{\min|\zeta'(s)|^2}.$$

Let us choose an integer $k[\zeta] \geq 1$ so that

$$\frac{L}{k} \leq 4^{-1/\alpha}\left(\min|\zeta'(s)|\right)^{2/\alpha}\|\zeta\|_{1,\alpha}^{-2/\alpha} < \frac{L}{k-1}. \tag{6.17}$$

Then

$$\left|\sin(\phi(s) - \phi(t))\right| \leq \frac{1}{2} \quad \text{when } |s-t| \leq \frac{L}{k}. \tag{6.18}$$

It follows that ϕ can vary at most by $\pi/3$ on an interval of the form $[jL/k, (j+1)L/k]$, $j = -k, \cdots, k-1$. Therefore,

$$|\phi| \leq \frac{k\pi}{3}. \tag{6.19}$$

Thus

$$|\phi(s) - \phi(t)| = \frac{|\phi(s) - \phi(t)|}{\left|\sin(\phi(s) - \phi(t))\right|}\left|\sin(\phi(s) - \phi(t))\right| \leq \frac{2\pi}{3}\left|\sin(\phi(s) - \phi(t))\right| \tag{6.20}$$
$$\text{when } |s-t| \leq \frac{L}{k}.$$

We combine (6.16) and (6.20) to estimate the Hölder quotient for ϕ when $|s-t| \leq L/k$. Otherwise, we find that $|\phi(s) - \phi(t)| \leq (2k/L^\alpha)\|\phi\|_0|s-t|^\alpha \leq (2k^2\pi/3L^\alpha)|s-t|^\alpha$ by (6.19). We now estimate k from the second inequality of (6.17) to obtain (6.15). □

We now state some results from complex function theory that are needed in the ensuing analysis.

LEMMA 6.21. *Let κ be a simple closed continuously differentiable curve in $\mathbb{C}$ with $|\kappa'|$ everywhere positive. If h is a one-to-one holomorphic function taking $\mathcal{D}$ onto the interior of κ, then h can be extended to $\mathrm{cl}\,\mathcal{D}$ so that the extension is a continuous bijection of $\mathrm{cl}\,\mathcal{D}$ onto $h(\mathrm{cl}\,\mathcal{D})$.*

The proof is given by Hille (1962, p. 367).

LEMMA 6.22 (Warschawski (1932), (1961), (1968)). *Let $\kappa : [-L, L] \to \mathbb{C}$ and h be as in Lemma 6.21. Let $\phi(\sigma)$ be the angle* (mod 2π) *that the tangent vector $\kappa'(s)$ makes with the negative imaginary axis at the point s with arc length parameter*

$$\sigma = \int_0^s |\kappa'(t)|\,dt.$$

If the function ϕ is Hölder continuous with exponent $\alpha \in (0, 1]$, then h' has a continuous extension to $\mathrm{cl}\,\mathcal{D}$ with h' vanishing nowhere on $\mathrm{cl}\,\mathcal{D}$.

LEMMA 6.23 (Radó (1923, p. 182); cf. Gattegno and Ostrowski (1949, p. 29)). *Let κ, $\kappa_n : [-L, L] \to \mathbb{C}$, $n = 1, 2, \cdots$ be simple closed continuously differentiable curves enclosing 0 with with $|\kappa'|$, $|\kappa_n'|$ everywhere positive. Let $h[\kappa]$, $h[\kappa_n]$ be one-to-one holomorphic mappings from $\mathcal{D}$ onto the interiors of κ, κ_n, respectively, with*

$h[\kappa](0) = h[\kappa_n](0) = 0$, $h[\kappa]'(0) > 0$, $h[\kappa_n]'(0) > 0$. *If* κ_n *converges uniformly to* κ *on* $[-L, L]$, *then* $h[\kappa_n]$ *converges uniformly to* $h[\kappa]$ *on* $\operatorname{cl}\mathcal{D}$.

We can now prove our first basic result about the solvability of CMP and the regularity of its solution.

THEOREM 6.24. *Let* $0 < \alpha \leq 1$, $0 < h_1 < h_2$, $\zeta \in \mathcal{Z}^{\pm}(\alpha, h_1, h_2)$. *Then there is a unique conformal mapping* $g[\zeta]$ *satisfying* CMP *such that* $g[\zeta]'$ *can be continuously extended to* $\operatorname{cl}\mathcal{D}$ *and vanishes nowhere on* $\operatorname{cl}\mathcal{D}$. *Moreover, if* $\{\zeta_k\}$ *is a sequence in* $\mathcal{Z}^{\pm}(\alpha, h_1, h_2)$ *converging to* ζ *in the* $C^{1,\alpha}$*-norm, then* $g[\zeta_k]$ *converges uniformly to* $g[\zeta]$ *on* $\operatorname{cl}\mathcal{D}$ *and* $g[\zeta_k]^{-1}(1/\zeta_k(\cdot))$ *converges pointwise to* $g[\zeta]^{-1}(1/\zeta(\cdot))$ *on* $[-L, L]$.

Proof. The existence and uniqueness of a conformal mapping $g[\zeta]$ from $\mathcal{D}$ to $\mathcal{I}[\zeta]$ such that (6.8c,d) hold is ensured by the Riemann Mapping Theorem. Condition (6.8b) follows from Lemma 6.21. By (4.3a), $z \mapsto \overline{g[\zeta](\bar{z})}$ satisfies (6.8a–d). Hence (6.8e) is satisfied by the uniqueness ensured by the Riemann Mapping Theorem. That $g[\zeta]$ enjoys the regularity properties at the boundary follows from Lemmas 6.14, 6.21, 6.22. The uniform convergence of $g[\zeta_k]$ follows from Lemmas 6.6 and 6.23. To prove the last statement, let $\sigma_0 = 1/\zeta(s)$, $s \in [-L, L]$. Then $g[\zeta_k]^{-1}(1/\zeta_k(s)) \in \partial\mathcal{D}$ and is accordingly bounded. By the Bolzano–Weierstrass Theorem, it has a subsequence (denoted the same way) converging to $\omega_0 \in \partial\mathcal{D}$. Now

(6.25)
$$\begin{aligned}|\sigma_0 - g[\zeta](\omega_0)| \leq{}& |\sigma_0 - 1/\zeta_k(s)| + \left|g[\zeta_k]\left(g[\zeta_k]^{-1}(1/\zeta_k(s))\right) - g[\zeta]\left(g[\zeta_k]^{-1}(1/\zeta_k(s))\right)\right| \\ &+ \left|g[\zeta]\left(g[\zeta_k]^{-1}(1/\zeta_k(s))\right) - g[\zeta](\omega_0)\right|.\end{aligned}$$

As $k \to \infty$ the terms on the right of (6.25) approach 0 because $g[\zeta_k]$ and ζ_k converge uniformly to $g[\zeta]$ and ζ and because $g[\zeta]$ is continuous. Hence $\omega_0 = g[\zeta]^{-1}(\sigma_0)$. Since this argument holds for all subsequences, the last statement of the theorem is established. □

We must supplement the results of Theorem 6.24 in order to prove the assertions about compactness made in Theorem 4.15. The next lemma introduces a function k, which intimately connects g and ζ and which intervenes in the explicit formulas, like (6.33) below, for the boundary values of fluid mechanical variables.

LEMMA 6.26. *Let* $0 < \alpha \leq 1$, $0 < h_1 < h_2$, $\zeta \in \mathcal{Z}^{+}(\alpha, h_1, h_2)$. *Let a typical point of* $\partial\mathcal{D}$ *be denoted by* $e^{i(\pi-\gamma)}$ (*cf. Fig.* 6.3). *Then there exists a unique function*

$$k[\zeta] : [-\pi, \pi] \to [-L, L] \tag{6.27}$$

(*which associates the parameter* s *of* ζ *with the angle* γ *parametrizing* $\partial\mathcal{D}$) *such that*

$$g[\zeta]\left(e^{i(\pi-\gamma)}\right) = 1/\zeta\left(k[\zeta](\gamma)\right). \tag{6.28}$$

$k[\zeta]$ *is continuously differentiable and* $k[\zeta]'$ *is everywhere positive. Moreover, if* $\zeta_l \in \mathcal{Z}^{+}(\alpha, h_1, h_2)$ *and if* $\zeta_l \to \zeta$ *in the* $C^{1,\alpha}$*-norm, then* $k[\zeta_l]$ *converges pointwise to* $k[\zeta]$.

Proof. Since ζ is simple, the mapping $[-L, L) \ni s \mapsto 1/\zeta(s) \in \partial\mathcal{I}$ has an inverse $j[\zeta]$. We define $k[\zeta](\gamma) \equiv j[\zeta]\left(g[\zeta](e^{i(\pi-\gamma)})\right)$. Thus (6.28) is satisfied. Since ζ' vanishes nowhere, the classical Implicit Function Theorem enables us to deduce from (6.28) that $k[\zeta]$ is continuously differentiable. Since g' does not vanish on $\partial\mathcal{D}$ by Theorem 6.24, we obtain from (6.28) that k' does not vanish on $[-\pi, \pi]$. Since conformal mappings preserve orientation, we conclude that $k[\zeta]'$ is everywhere positive.

We now turn to the convergence properties of k. We fix γ. Since $\{k[\zeta_l](\gamma)\}$ is bounded, it has a subsequence, denoted the same way, that converges to $s \in [-L, L]$. Now

$$\left|\frac{1}{\zeta_l\left(k[\zeta_l](\gamma)\right)} - \frac{1}{\zeta(s)}\right| \leq \left|\frac{1}{\zeta_l\left(k[\zeta_l](\gamma)\right)} - \frac{1}{\zeta\left(k[\zeta_l](\gamma)\right)}\right| + \left|\frac{1}{\zeta\left(k[\zeta_l](\gamma)\right)} - \frac{1}{\zeta(s)}\right|. \tag{6.29}$$

It follows from the uniform convergence of $\zeta_l \to \zeta$ and from the properties of our subsequence that the right-hand side of (6.29) approaches zero as $l \to \infty$. It then follows from this observation, from (6.28), and from Theorem 6.24 that

$$1/\zeta(s) = \lim g[\zeta_l]\big(e^{i(\pi-\gamma)}\big) = g[\zeta]\big(e^{i(\pi-\gamma)}\big). \tag{6.30}$$

Therefore $s = k[\zeta](\gamma)$. Since our argument holds for all subsequences, we conclude that $\lim k[\zeta_l](\gamma) = k[\zeta](\gamma)$. □

From (6.9) and (6.11) we compute

$$\Omega'(z) = \frac{U\big[1 - g^{-1}(\frac{1}{z})^2\big]}{g'(0)g^{-1}(\frac{1}{z})^2 g'\big(g^{-1}(\frac{1}{z})\big)z^2}. \tag{6.31}$$

Let $z = \zeta(s)$, $s = k(\gamma)$, so that $g^{-1}(\frac{1}{z}) = e^{i(\pi-\gamma)}$. Provided that the hypotheses of Lemma 6.26 hold, we find that the complex velocity at $\zeta(s)$ is

$$\Omega'\big(\zeta(s)\big) = \frac{2U\sin\gamma}{g'(0)\zeta'(s)k'(\gamma)}, \qquad s = k(\gamma). \tag{6.32}$$

From Bernoulli's equation (3.5), we thus obtain a form of (4.16) that makes the role of ζ explicit:

$$p[\zeta, U, P](s) = P + \frac{\rho}{2}U^2\left\{1 - \frac{4\,|\sin k[\zeta]^{-1}(s)|^2}{\big|g'(0)\zeta'(s)k[\zeta]'\big(k[\zeta]^{-1}(s)\big)\big|^2}\right\}. \tag{6.33}$$

To obtain Hölder and C^1 bounds needed to establish Theorem 4.15 we employ the following lemma.

LEMMA 6.34 (Warschawski (1932, pp. 451–452)). *Let κ, h, ϕ be as in Lemmas 6.22 and 6.23 and let $0 < \alpha < 1$. If there exists a positive number A such that*

$$\operatorname{diam} h(\mathcal{D}) < A, \tag{6.35a}$$

$$h(\mathcal{D}) \text{ contains the disk } (1/A)\mathcal{D} \text{ of radius } 1/A \text{ centered at } 0, \tag{6.35b}$$

$$\frac{1}{A} \le \inf\left\{\frac{|z_1 - z_2|}{\lambda(z_1, z_2)} \;:\; z_1, z_2 \in h(\partial\mathcal{D}), \quad z_1 \ne z_2\right\} \tag{6.35c}$$

where $\lambda(z_1, z_2)$ is the length of the shortest arc of κ joining z_1 to z_2,

$$\frac{1}{A} \le \int_{-L}^{L} |\kappa'(s)|\,ds, \tag{6.35d}$$

$$\|\phi\|_{0,\alpha} < A, \tag{6.35e}$$

then there exists a positive number C depending only on A and α (and independent of κ) such that

$$1/C < |h'(\omega)| < C \quad \forall\,\omega \in \operatorname{cl}\mathcal{D}, \tag{6.36a}$$

$$\|h\|_{C^{1,\alpha}(\operatorname{cl}\mathcal{D})} < C. \tag{6.36b}$$

We now convert these results to those for g.

LEMMA 6.37. *Let $0 < \alpha < 1$, $0 < h_1 < h_2$, $c > 0$, $\delta > 0$ and let $\zeta \in \mathcal{Z}^{\pm}(\alpha, h_1, h_2, \delta)$ with $\|\zeta\|_{1,\alpha} \le c$. Then there exists a positive number C depending only on c, α, h_1, h_2, δ (and independent of ζ) such that*

$$1/C < |g[\zeta]'(\omega)| < C \quad \forall\,\omega \in \operatorname{cl}\mathcal{D}, \tag{6.38a}$$

$$\|g[\zeta]\|_{C^{1,\alpha}(\operatorname{cl}\mathcal{D})} < C. \tag{6.38b}$$

Proof. We identify h and κ of Lemma 6.34 with $g[\zeta]$ and $\partial\mathcal{I}[\zeta]$. (In particular, (6.30) holds.) The properties of g listed in Theorem 6.24 imply that (6.38) is a consequence of (6.36). Thus we need only find an A, depending only on c, α, h_1, h_2, δ such that (6.35) holds. Since $\operatorname{diam} g(\mathcal{D}) < \sup\{|\zeta(s)^{-1} - \zeta(t)^{-1}| : s \neq t\} \leq 2ch_1^{-2}$ and since $\int_{-L}^{L} |(1/\zeta)'|\, ds \geq 2L\delta h_2^{-2}$, we ensure (6.35a,d) by taking $A > 2ch_1^{-2}, h_2^2 L^{-1}\delta^{-1}2^{-1}$. We ensure (6.35b) by taking $A > h_2$. To verify (6.35c) we let $s < t$ and observe that

(6.39)
$$\begin{aligned}
&\lambda\big(\zeta(s)^{-1}, \zeta(t)^{-1}\big) \\
&\quad = \min\left\{\int_{-L}^{s} |\zeta'(u)\zeta(u)^{-2}|\, du + \int_{t}^{L} |\zeta'(u)\zeta(u)^{-2}|\, du,\ \int_{s}^{t} |\zeta'(u)\zeta(u)^{-2}|\, du\right\} \\
&\quad \leq \|\zeta'\|_0 h_1^{-2} \min\{2L - (t-s), t-s\} = \|\zeta'\|_0 h_1^{-2}\sigma(t,s)
\end{aligned}$$

where σ is defined in (4.8a). Since $\zeta \in \mathcal{Z}(\delta)$, we have

$$\frac{|\zeta(s)^{-1} - \zeta(t)^{-1}|}{\lambda\big(\zeta(s)^{-1}, \zeta(t)^{-1}\big)} \geq \frac{\delta h_1^2 \sigma(t,s)}{ch_2^2\sigma(t,s)} = \frac{\delta h_1^2}{ch_2^2}. \tag{6.40}$$

Thus we ensure (6.35c) by taking $A > ch_2^2/\delta h_1^2$. We now use (6.15) with ζ replaced by $1/\zeta$ and use the definition of $\mathcal{Z}(\delta)$ to ensure (6.35e) by taking $A > a[h_1^{-1} + ch_1^{-2} + c^3(1 + 2(2L)^{1-\alpha})h_1^{-4} + 1]^{4/\alpha}(\delta h_2^{-2})^{-4/\alpha}$. □

We need the following technical and intrinsically interesting lemma (a proof of which we could not find in the literature).

LEMMA 6.41. *Let $0 < \beta < \alpha \leq 1$ and let m be a nonnegative integer. Let $\{f_k\}$ be a sequence of functions on $[-L, L]$ that is bounded in $C^{m,\alpha}$. If f_k converges to $f \in C^0$ either pointwise almost everywhere or in the sense of distributions on $(-L, L)$, then $f \in C^{m,\alpha}$ and f_k converges to f in the norm of $C^{m,\beta}$.*

Proof. By hypothesis there is a positive number C such that $\|f_k\|_{m,\alpha} \leq C$. Since $C^{m,\alpha}$ is compactly embedded in $C^{m,\beta}$, the sequence $\{f_k\}$ has a subsequence, denoted the same way, that converges to a limit $g \in C^{m,\beta}$ in the $C^{m,\beta}$-norm. (A fortiori f_k approaches g pointwise almost everywhere and in the sense of distributions.) Our bound on $\|f_k\|_{m,\alpha}$ implies that

$$\left|\frac{d^m}{ds^m} f_k(s_1) - \frac{d^m}{ds^m} f_k(s_2)\right| \leq C|s_1 - s_2|^\alpha \quad \forall\, s_1, s_2, \tag{6.42a}$$

whence we obtain

$$\left|\frac{d^m}{ds^m} g(s_1) - \frac{d^m}{ds^m} g(s_2)\right| \leq C|s_1 - s_2|^\alpha \quad \forall\, s_1, s_2. \tag{6.42b}$$

Thus $g \in C^{m,\alpha}$. Since f_k converges to both f and g either pointwise almost everywhere or in the sense of distributions and since both f and g are continuous, we conclude that $f = g$. The full sequence $\{f_k\}$ converges to f in the norm of $C^{m,\beta}$, for if not, there would be a number $\epsilon > 0$ and another subsequence $\{f_k\}$ such that $\|f_k - f\|_{m,\beta} \geq \epsilon$. We apply the preceding argument to this subsequence to derive a contradiction. □

The following lemma is the culmination of our analysis of of the function k, which plays a central role in determining the boundary behavior of the fluid-mechanical variables.

LEMMA 6.43. *Let $c > 0$, $0 < \beta < \alpha \leq 1$. For all positive integers n let ζ_n, $\zeta \in \{\zeta \in \mathcal{Z}^+ : (4.3a)\ holds\} \cap C^{1,\alpha}$; $0 \in \mathcal{I}[\zeta]$, $\mathcal{I}[\zeta_n]$; $\|\zeta_n\|_{1,\alpha}$, $\|\zeta\|_{1,\alpha} < c$. Let $\zeta_n \to \zeta$ in*

the C^1-norm. Then

$$k[\zeta_n]^{-1} \to k[\zeta]^{-1}, \tag{6.44}$$

$$k[\zeta_n]'\big(k[\zeta_n]^{-1}(\cdot)\big) \to k[\zeta]'\big(k[\zeta]^{-1}(\cdot)\big) \tag{6.45}$$

in the $C^{0,\beta}$-norm.

Proof. Since $\zeta_n \to \zeta$ in the C^1-norm, there is a number N such that $\|\zeta_n - \zeta\|_0 < \frac{1}{2}\min_s|\zeta(s)|$ for $n > N$. We define

$$\begin{aligned} h_1[\zeta] &\equiv \min\Big\{\tfrac{1}{2}\min_s|\zeta(s)|,\ \min_s|\zeta_n(s)|,\ n = 1,\cdots,N\Big\} > 0,\\ h_2[\zeta] &\equiv \max\Big\{2\max_s|\zeta(s)|,\ \max_s|\zeta_n(s)|,\ n = 1,\cdots,N\Big\} < \infty. \end{aligned} \tag{6.46}$$

Thus we find that

$$h_1[\zeta] \le |\zeta_n(s)| \le h_2[\zeta] \quad \forall n, s. \tag{6.47a}$$

Using Lemma 4.13 we likewise show that there is a $\delta[\zeta] > 0$ such that

$$l[\zeta],\ l[\zeta_n] > \delta \tag{6.47b}$$

(cf. (4.8)) and that there is an $\epsilon[\zeta] > 0$ such that

$$\min_s|\zeta'(s)|,\quad \min_s|\zeta_n'(s)| > \epsilon. \tag{6.47c}$$

Thus $\zeta,\ \zeta_n \in \mathcal{Z}^+(\alpha, h_1[\zeta], h_2[\zeta], \delta[\zeta])$.

We now show that

$$\|k[\zeta]\|_{1,\alpha},\ \sup_n\big\{\|k[\zeta_n]\|_{1,\alpha}\big\} < \infty. \tag{6.48}$$

Clearly, $\|k[\zeta_n]\|_0 = L$. From (6.28) we obtain

$$\begin{aligned} &\|k[\zeta_n]'\|_0 = \|r_n(\gamma)\|_0 < \infty,\\ &r_n(\gamma) \equiv ig[\zeta_n]'\big(e^{i(\pi-\gamma)}\big)\zeta_n\big(k[\zeta_n](\gamma)\big)^2/\zeta_n'\big(k[\zeta_n](\gamma)\big), \end{aligned} \tag{6.49}$$

the inequality following from (6.38a), (6.47c). From (6.28) we also obtain

$$|k[\zeta_n]'(\gamma_1) - k[\zeta_n]'(\gamma_2)| \le |r_n(\gamma_1) - r_n(\gamma_2)| + |r_n(\gamma_2)||e^{i(\pi-\gamma_1)} - e^{i(\pi-\gamma_2)}|. \tag{6.50}$$

The continuity of $k[\zeta_n]'$ (ensured by Lemma 6.26), the hypothesis that $\|\zeta_n\|_{1,\alpha}, \|\zeta\|_{1,\alpha} < c$, and the inequalities (6.38b), (6.47c) ensure that there is a positive number C independent of n such that

$$|r_n(\gamma_1) - r_n(\gamma_2)| \le C|e^{i(\pi-\gamma_1)} - e^{i(\pi-\gamma_2)}|^\alpha. \tag{6.51}$$

Since $|e^{i(\pi-\gamma_1)} - e^{i(\pi-\gamma_2)}| = 2|\sin\frac{\gamma_2-\gamma_1}{2}| \le |\gamma_2 - \gamma_1|$, we readily deduce that (6.48) holds.

We now prove that (6.44) holds pointwise. Let us assume for contradiction that it does not. Then there would be a number $t \in [-L, L]$, a number $\epsilon > 0$, and a subsequence such that $|k[\zeta_n]^{-1}(t) - k[\zeta]^{-1}(t)| > \epsilon$. Since $\{k[\zeta_n]^{-1}(t)\}$ is bounded, it has a further subsequence converging to $\tau \in [-\pi, \pi]$. Since

$$(6.52)\quad |t - k[\zeta](\tau)| \leq |k[\zeta_n]\big(k[\zeta_n]^{-1}(t)\big) - k[\zeta]\big(k[\zeta_n]^{-1}(t)\big)| + |k[\zeta]\big(k[\zeta_n]^{-1}(t)\big) - k[\zeta](\tau)|,$$

we can use Lemma 6.26, Lemma 6.41, and inequality (6.48) to deduce that $t = k[\zeta](\tau)$, a contradiction. Thus $k[\zeta_n]^{-1}$ converges pointwise to $k[\zeta_n]^{-1}$. That this convergence is actually in the norm of $C^{0,\beta}$ follows from Lemma 6.41 because the inequality

$$(6.53)\quad |k[\zeta_n]^{-1}(s_1) - k[\zeta_n]^{-1}(s_2)| \leq \frac{|s_1 - s_2|}{\min_s k[\zeta_n]'\big(k[\zeta_n]^{-1}(s)\big)} \leq \frac{\max_s |\zeta_n'(s)||s_1 - s_2|}{\min_s |\zeta_n(s)|^2 \min_{\mathrm{cl}\,\mathcal{D}} |g[\zeta_n]'(\omega)|}$$

together with (6.38a), (6.46), and the hypothesis that $\|\zeta_n\|_{1,\alpha},\ \|\zeta\|_{1,\alpha} < c$ imply that there is a number C depending on ζ such that

$$(6.54)\qquad \|k[\zeta_n]^{-1}\|_{0,1},\ \|k[\zeta_n]^{-1}\|_{0,1} \leq C.$$

We now turn to the proof of (6.45). In view of (6.48), we can use Lemmas 6.26 and 6.41 to show that $k[\zeta_n]'$ converges uniformly to $k[\zeta]'$. Since

$$(6.55)\quad \begin{aligned} &|k[\zeta_n]'\big(k[\zeta_n]^{-1}(s)\big) - k[\zeta]'\big(k[\zeta]^{-1}(s)\big)| \\ &\quad \leq |k[\zeta_n]'\big(k[\zeta_n]^{-1}(s)\big) - k[\zeta]'\big(k[\zeta_n]^{-1}(s)\big)| \\ &\qquad + |k[\zeta]'\big(k[\zeta_n]^{-1}(s)\big) - k[\zeta]'\big(k[\zeta]^{-1}(s)\big)| \end{aligned}$$

we can use the continuity of $k[\zeta]'$, the uniform convergence of $k[\zeta_n]'$ to $k[\zeta]'$, and the pointwise convergence of $k[\zeta_n]^{-1}$ to $k[\zeta]^{-1}$ to deduce that (6.45) holds pointwise. Conditions (6.48), (6.54) imply that there is a number C depending on ζ such that

$$(6.56)\qquad \|k[\zeta_n]'\big(k[\zeta_n]^{-1}(\cdot)\big)\|_{0,\alpha},\ \|k[\zeta]'\big(k[\zeta]^{-1}(\cdot)\big)\|_{0,\alpha} \leq C.$$

Property (6.45) now follows from another application of Lemma 6.41. □

Having collected these technical results, we are now ready to put them together in the following lemma, which leads immediately to the proof of Theorem 4.15.

LEMMA 6.57. *Let* $\{\zeta_n, U_n, P_n\}$ *be a sequence bounded in* $[\{\zeta \in \mathcal{Z}^+\colon$ (4.3a) *holds*$\}$ $\cap C^{1,1}] \times \mathbb{R}^2$ *and let there be a number* $\alpha \in (0,1]$ *such that this sequence converges in the norm of* $C^{1,\alpha} \times \mathbb{R}^2$ *to* $\{\zeta, U, P\} \in [\{\zeta \in \mathcal{Z}^+ :$ (4.3a) *holds*$\} \cap C^{1,1}] \times \mathbb{R}^2$. *Then for each* $\beta \in (0,\alpha)$, *the functions* $\Omega'\big(\zeta_n(\cdot)\big)$ *and* $p[\zeta_n, U_n, P_n]$, *defined in* (6.32), (6.33) *converge to* $\Omega'\big(\zeta(\cdot)\big)$ *and to* $p[\zeta, U, P]$ *in the norm of* $C^{0,\beta}$.

Proof. Since $\Omega[\zeta]'(z) = \Omega[\zeta + a]'(z + a)$ for all real a, there is no loss of generality in assuming that $0 \in \mathcal{I}[\zeta],\ \mathcal{I}[\zeta_n]$ for all n. Then just as in the proof of Lemma 6.43, we can show that there are positive numbers h_1, h_2, depending on ζ, such that $\zeta,\ \zeta_n \in \mathcal{Z}^+(\alpha, h_1, h_2)$ for n sufficiently large. In particular, $|\zeta'|,\ |\zeta_n'|$ are everywhere positive. It follows from Theorem 6.24 that $g[\zeta]'(0),\ g[\zeta_n]'(0) \neq 0$ and from Lemma 6.26 that $k[\zeta]',\ k[\zeta_n]'$ are everywhere positive. Consequently the corresponding flow variables can be defined as in (6.32), (6.33). The convergence of the terms involving the function k follows from Lemmas 6.26 and 6.43. Let us identify $g[\zeta_n]$ with f_n of Lemma 6.41. We choose $m = 1$ and replace $[-L, L]$ with $\mathrm{cl}\,\mathcal{D}$. By (6.38b) and Lemma 6.23, f_n meets the hypotheses of Lemma 6.41. We thus conclude that $g[\zeta_n]'(0) \to g[\zeta]'(0)$. □

Proof of Theorem 4.15. Clearly $\mathcal{Z} = \mathcal{Z}^+ \cup \mathcal{Z}^-$, $\mathcal{Z}^+ \cap \mathcal{Z}^- = \emptyset$, and $\mathcal{Z}^+$ and $\mathcal{Z}^-$ are both open and closed in $\mathcal{Z}$ with respect to the C^1-norm. Hence it suffices to prove the statement when $\mathcal{Z}$ is replaced with $\mathcal{Z}^+$. Now $\mathcal{Z}^+$ is the disjoint union of those ζ's

for which $\operatorname{Re}\zeta(\pm L) > \operatorname{Re}\zeta(0)$ and of those ζ's for which the reverse inequality holds. It suffices to consider just the former subset of $\mathcal{Z}^+$.

We must verify that (6.32), (6.33) have the requisite properties. We use (3.6a). That $U^{-1}\Omega'\big(\zeta(\cdot)\big)$ is independent of U follows immediately from (6.32). Let δ be any positive number. The continuity of

(6.58)
$$C^2 \cap \{\zeta \in \mathcal{Z}^+(\delta) : \text{ (4.3) holds, } \operatorname{Re}\zeta(\pm L) > \operatorname{Re}\zeta(0)\} \ni \zeta \mapsto |U^{-1}w[\zeta, U]| \in C^0$$
$$p[\cdot,\cdot,\cdot] : C^2 \cap \{\zeta \in \mathcal{Z}^+(\delta) : (4.3) \text{ holds, } \operatorname{Re}\zeta(\pm L) > \operatorname{Re}\zeta(0)\} \times \mathbb{R}^2 \to C^0$$

follows from Lemma 6.57. To prove the compactness, let $\{\zeta_n, U_n, P_n\}$ be a sequence bounded in $C^{1,1} \cap \{\zeta \in \mathcal{Z}^+(\delta) : (4.3a) \text{ holds, } \operatorname{Re}\zeta(\pm L) > \operatorname{Re}\zeta(0)\} \times \mathbb{R}^2$. Since $C^{1,1}$ is compactly embedded in $C^{1,\alpha}$ for $\alpha \in (0,1)$, there is a subsequence converging in the norm of $C^{1,\alpha} \times \mathbb{R}^2$ to $\{\zeta, U, P\}$. Since $|\zeta_n'(s_1) - \zeta_n'(s_2)| \le \sup \|\zeta_n\|_{1,1} |s_1 - s_2|$ and since $|\zeta_n(s_1) - \zeta_n(s_2)| \ge \delta\sigma(s_1, s_2)$ (cf. (4.8)), we find that $\zeta \in C^{1,1} \cap \mathcal{Z}^+(\delta/2)$. Since $\operatorname{Re}\zeta_n(\pm L) > \operatorname{Re}\zeta_n(0)$, it follows that $\operatorname{Re}\zeta(\pm L) \ge \operatorname{Re}\zeta(0)$. Hence the injectivity of $\zeta \in \mathcal{Z}^+(\delta/2)$ implies that $\operatorname{Re}\zeta(\pm L) > \operatorname{Re}\zeta(0)$.

Lemma 6.57 thus implies that $|U^{-1}w[\zeta_n, U]|$ and $p[\zeta_n, U_n, P_n]$ converge in the norm of $C^{0,\beta}$, which is compactly embedded in C^0. □

7. Proof of the Global Implicit Function Theorem. We deduce the Global Implicit Function Theorem 5.1 from the following theorem.

THEOREM 7.1 (Alexander and Yorke (1976)). *Let $\mathcal{X}$ be a Banach space. Let $F : \mathcal{X} \times \mathbb{R}^m \to \mathcal{X}$ be continuous and compact with $F(0,0) = 0$. Let I denote the identity operator on $\mathcal{X}$. Let the Fréchet derivative $I - F_x(0,0) : \mathcal{X} \to \mathcal{X}$ of $x \mapsto x - F(x,\lambda)$ at $(0,0)$ exist and be invertible. Let $\mathcal{S} \equiv \{(x,\lambda) \in \mathcal{X} \times \mathbb{R}^m : x = F(x,\lambda)\}$, let $\mathcal{S}_0$ be the connected component of $\mathcal{S}$ containing $(0,0)$, and let $\mathcal{S}_0^+$ be the one-point compactification of $\mathcal{S}_0$. Then there is an essential map (i.e., a continuous map not homotopic to a constant) from $\mathcal{S}_0^+$ onto the m-dimensional sphere $\mathbb{S}^m$ whose restriction to $\mathcal{S}_0 \setminus \{(0,0)\}$ is inessential. Moreover, $\mathcal{S}_0^+$ contains a connected subset $\mathcal{S}_{00}^+$ that contains $(0,0)$, that has the same properties as $\mathcal{S}_0^+$ with respect to essential maps onto $\mathbb{S}^m$, and that has the property that each of its points has topological dimension m.*

Remarks. That each point of $\mathcal{S}_{00}^+$ has topological dimension m was observed by Alexander and Antman (1981). Alexander and Yorke (1976) actually stated the somewhat stronger result that the conclusion of Theorem 7.1 holds when the domain of F is a set $\mathcal{U}$ in $\mathcal{X} \times \mathbb{R}^m$ that is homeomorphic to $\mathcal{X} \times \mathbb{R}^m$. (Alexander and Antman (1981) made an analogous assertion.) We could reduce the more general problem to Theorem 7.1 provided that $\mathcal{U}$ and F admit a homeomorphism from $\mathcal{U}$ to $\mathcal{X} \times \mathbb{R}^m$ that preserves (i) the fixed-point form of the equation and the distinguished role of the parameter in $\mathbb{R}^m$, (ii) the compactness of the appropriate operator, and (iii) the existence of an invertible partial Fréchet derivative at a base point. Problems that admit such homeomorphisms must have very special forms. Thus we believe that the stronger result of Alexander and Yorke (and the corresponding result of Alexander and Antman) do not hold without further qualification. Since our problem apparently fails to have the requisite special form, we resort to Theorem 5.1, a corollary of Theorem 7.1, which not only retains much of the force of the stronger version, but also applies to a wider class of domains and operators. Continuation theorems related to Theorem 7.1 are given by Fitzpatrick, Massabò, and Pejsachowicz (1983) and by Ize, Massabò, Pejsachowicz, and Vignoli (1985).

Proof of Theorem 5.1. The result about $\mathcal{S}$ in a neighborhood of $(0,0)$ is a consequence of the classical Implicit Function Theorem in Banach space. Let us first suppose that the first statements of (ii) and (iii) do not hold. Then the first statement of (i) must hold. We prove the rest of statement (i).

Let $E > 0$ be given. We choose ϵ^* accordingly. Since $\mathcal{X} \times \mathbb{R}^m$ is a Banach space and thus a normal topological space, Urysohn's lemma implies that there exists a continuous function

$$\mathcal{X} \times \mathbb{R}^m \ni (x, \lambda) \mapsto \varphi(x, \lambda, \epsilon) \in [0,1] \tag{7.2}$$

such that

$$\begin{aligned} &\varphi(x,\lambda,\epsilon) = 1 \quad \text{for } (x,\lambda) \in \mathcal{O}(\epsilon), \\ &\varphi(x,\lambda,\epsilon) = 0 \quad \text{for } (x,\lambda) \in \mathcal{X} \times \mathbb{R}^m \setminus \mathcal{O}(\epsilon/2), \end{aligned} \tag{7.3}$$

for each $\epsilon \in (0, \epsilon^*)$ (provided that $\mathcal{O}(\epsilon/2) \neq \mathcal{X} \times \mathbb{R}^m$, which we may assume. We shall apply Theorem 7.1 to the mollified equation

$$x = \varphi(x, \lambda, \epsilon) F(x, \lambda). \tag{7.4}$$

Let

$$\mathcal{S}(\epsilon) \equiv \{(x,\lambda) \in \mathcal{X} \times \mathbb{R}^m : (7.4) \text{ holds}\} \tag{7.5}$$

and let $\mathcal{S}_0(\epsilon)$ be the connected component of $\mathcal{S}(\epsilon)$ containing $(0,0)$. We reduce our proof to a sequence of lemmas. The properties of φ imply the following lemmas.

LEMMA 7.6. *Let* $0 < \epsilon < \epsilon^*/4$. *Then* $\mathcal{S} \cap \mathcal{O}(\epsilon^*) = \mathcal{S}(\epsilon) \cap \mathcal{O}(\epsilon^*)$.

LEMMA 7.7. *Let* $\mathcal{A}$ *be an open subset of a metric space* $\mathcal{Y}$ *and let* $\mathcal{C}$ *be a closed connected subset of* $\mathcal{Y}$ *with* $\mathcal{C} \cap \operatorname{cl}\mathcal{A}$ *compact and* $\mathcal{C} \setminus \mathcal{A} \neq \emptyset$. *Let* $x \in \mathcal{C} \cap \mathcal{A}$. *Then there exists a connected subset* $\mathcal{C}_1$ *of* $\mathcal{C} \cap \operatorname{cl}\mathcal{A}$ *such that* $x \in \mathcal{C}_1$, $\mathcal{C}_1 \cap \partial\mathcal{A} \neq \emptyset$.

Proof. The connectedness of $\mathcal{C}$ suppports a simple proof by contradiction that $\mathcal{C} \cap \partial\mathcal{A}$ and x are not separated in $\mathcal{C} \cap \operatorname{cl}\mathcal{A}$. Hence a well-known result from topology (cf. Alexander (1981, Cor. 4)) implies that x and $\mathcal{C} \cap \partial A$ are connected in $\mathcal{C} \cap \operatorname{cl}\mathcal{A}$. □

LEMMA 7.8. *Let* $\mathcal{S}_0$ *be bounded, let* $\mathcal{S}_0 \subset \mathcal{O}(\epsilon^*)$, *and let* $0 < \epsilon < \epsilon^*/4$. *Then* $\mathcal{S}_0 = \mathcal{S}_0(\epsilon)$.

Proof. By Lemma 7.6 $\mathcal{S}_0 \subset \mathcal{S} \cap \mathcal{O}(\epsilon^*) = \mathcal{S}(\epsilon) \cap \mathcal{O}(\epsilon^*) \subset \mathcal{S}(\epsilon)$. The connectivity properties of $\mathcal{S}_0$ and $\mathcal{S}_0(\epsilon)$ then imply that $\mathcal{S}_0 \subset \mathcal{S}_0(\epsilon)$.

We now prove the reverse inclusion. The compactness of F on $\mathcal{O}(\epsilon)$ and the boundedness of $\mathcal{S}_0$ imply that $\mathcal{S}_0$ is compact. Since $\mathcal{S}_0 \subset \mathcal{S} \cap \mathcal{O}(\epsilon^*)$, it follows that

$$d \equiv \operatorname{dist}\{\mathcal{S}_0,\ \mathcal{X} \times \mathbb{R}^m \setminus \mathcal{O}(\epsilon^*)\} > 0. \tag{7.9}$$

Let

$$\mathcal{A} \equiv \{(x,\lambda) \in \mathcal{X} \times \mathbb{R}^m : \operatorname{dist}\{(x,\lambda), \mathcal{S}_0\} < d/4\}. \tag{7.10}$$

Clearly, $\mathcal{A}$ is open and bounded and $\mathcal{S}_0 \subset \mathcal{A} \subset \mathcal{O}(\epsilon^*)$. We now prove by contradiction that $\mathcal{S}_0(\epsilon) \subset \operatorname{cl}\mathcal{A}$. Suppose that there were a $p \in \mathcal{S}_0(\epsilon) \setminus \operatorname{cl}\mathcal{A}$. Since φF is compact on $\mathcal{X} \times \mathbb{R}^m$, we find that $\mathcal{S}_0(\epsilon) \cap \operatorname{cl}\mathcal{A}$ is compact. Lemma 7.7 implies that there would exist a connected subset $\mathcal{S}_1(\epsilon)$ of $\mathcal{S}_0(\epsilon) \cap \operatorname{cl}\mathcal{A}$ such that $(0,0) \in \mathcal{S}_1(\epsilon)$, $\mathcal{S}_1(\epsilon) \cap \partial\mathcal{A} \neq \emptyset$. Since $0 < \epsilon < \epsilon^*/4$, Lemma 7.6 implies that $\mathcal{S}_1(\epsilon) \subset \mathcal{S}_0(\epsilon) \cap \mathcal{O}(\epsilon^*) \subset \mathcal{S}(\epsilon) \cap \mathcal{O}(\epsilon^*) = \mathcal{S} \cap \mathcal{O}(\epsilon^*)$. Thus $\mathcal{S}_1(\epsilon) \subset \mathcal{S}$. Since $(0,0) \in \mathcal{S}_1(\epsilon)$ and $\mathcal{S}_1(\epsilon)$ is connected, it follows that $\mathcal{S}_1(\epsilon) \subset \mathcal{S}_0$. Hence, $\mathcal{S}_0 \cap \partial\mathcal{A} \neq \emptyset$. But this is impossible because $\mathcal{A}$ is an open neighborhood of $\mathcal{S}_0$. Thus $\mathcal{S}_0(\epsilon) \subset \operatorname{cl}\mathcal{A} \subset \mathcal{O}(\epsilon^*)$. Lemma 7.6 now implies that

$\mathcal{S}_0(\epsilon) \subset \mathcal{S}(\epsilon) \cap \mathcal{O}(\epsilon^*) = \mathcal{S} \cap \mathcal{O}(\epsilon^*) \subset \mathcal{S}$. Since $\mathcal{S}_0(\epsilon)$ is connected, contains $(0,0)$, and is contained in $\mathcal{S}$, it follows that $\mathcal{S}_0(\epsilon) \subset \mathcal{S}_0$. □

Since Lemma 7.8 and the compactness of $\mathcal{S}_0$ imply that $\mathcal{S}_0(\epsilon) \cup \{\infty\} = \mathcal{S}_0 \cup \{\infty\} = \mathcal{S}_0^+$ (with ∞ isolated), we can apply Theorem 7.1 to (7.4) to deduce statement (i) of Theorem 5.1. The mollification argument we have used enables us to apply Theorem 7.1 directly to (7.4) in order to obtain statement (ii). □

Acknowledgment. We are grateful to J. C. Alexander and J. A. Hummel for a number of helpful comments. The work reported here represents a significant extension of part of the doctoral dissertation of Lanza de Cristoforis (1987).

REFERENCES

J. C. ALEXANDER (1981), *A primer on connectivity*, in Proc. Conf. on Fixed Point Theory, E. Fadell and G. Fournier, eds., Lecture Notes in Math., 886, Springer-Verlag, Berlin, New York, pp. 455–483.

J. C. ALEXANDER AND S. S. ANTMAN (1981), *Global and local behavior of bifurcating multidimensional continua of solutions for multiparameter nonlinear eigenvalue problems*, Arch. Rational Mech. Anal., 76, pp. 339–354.

J. C. ALEXANDER AND J. A. YORKE (1976), *The implicit function theorem and global methods of cohomology*, J. Funct. Anal., 21, pp. 330–339.

S. S. ANTMAN (1973), *Monotonicity and invertibility conditions in one-dimensional nonlinear elasticity*, in Symposium on Nonlinear Elasticity, R. W. Dickey, ed., Academic Press, New York, pp. 57–92.

———, (1976), *Ordinary differential equations of one-dimensional nonlinear elasticity* I: *Foundations of the theories of nonlinearly elastic rods and shells*, Arch. Rational Mech. Anal., 61, pp. 307–351.

S. S. ANTMAN AND E. R. CARBONE (1977), *Shear and necking instabilities in nonlinear elasticity*, J. Elasticity, 7, pp. 125–151.

S. S. ANTMAN AND G. ROSENFELD (1978), *Global behavior of buckled states of nonlinearly elastic rods*, SIAM Rev., 20, pp. 513–566.

H. BREZIS AND G. STAMPACCHIA (1976), *The hodograph method in fluid dynamics in the light of variational inequalities*, Arch. Rational Mech. Anal., 61, pp. 1–18.

R. FINNILA AND J. M. SLOSS (1966), *Existence and uniqueness theorems for a class of nonlinear singular integral equations with applications to a hydroelastic problem*, J. Math. Mech., 16, pp. 509–534.

P. M. FITZPATRICK, I. MASSABÓ, AND J. PEJSACHOWICZ (1983), *Global several-parameter bifurcation and continuation theorems: a unified approach via complementing maps*, Math. Ann., 263, pp. 61–73.

C. GATTEGNO AND A. OSTROWSKI (1949), *Représentation conforme à la frontière; domaines généraux*, Mémorial des Sciences Mathématiques, fasc. 109, Gauthier-Villars, Paris.

E. HILLE (1962), *Analytic Function Theory*, Vol. II, Ginn, Needham Heights, MA.

J. IZE, I. MASSABÒ, J. PEJSACHOWICZ, AND A. VIGNOLI (1985), *Structure and dimension of global branches of solutions to multiparameter nonlinear equations*, Trans. Amer. Math. Soc., 291, pp. 383–435.

M. LANZA DE CRISTOFORIS (1987), *Nonlinear deformations of structures in perfect flows*, Ph.D. thesis, University of Maryland, College Park, MD.

L. M. MILNE-THOMSON (1968), *Theoretical Hydrodynamics*, Fifth edition, Macmillan, New York.

T. RADÓ (1923), *Sur la représentation conforme de domaines variables*, Acta Univ. Szeged. 1, pp. 180–186.

J. B. SERRIN (1959), *Mathematical principles of classical fluid mechanics*, in Handbuch der Physik, Vol. VIII/1, C. Truesdell ed., Springer-Verlag, Berlin, New York, pp. 125–263.

S. E. WARSCHAWSKI (1932), *Über das Randverhalten der Abbildungsfunktion bei konformer Abbildung*, Math. Z., 35, pp. 321–456.

———, (1961), *On the differentiability at the boundary in conformal mapping*, Proc. Amer. Math. Soc., 12, pp. 614–620.

———, (1968), *On the Hölder continuity at the boundary in conformal maps*, J. Math. Mech., 19, pp. 423–427.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1222-1245, September 1991

002

THE POISSON EQUATION WITH NONAUTONOMOUS SEMILINEAR BOUNDARY CONDITIONS IN DOMAINS WITH MANY TINY HOLES*

SATOSHI KAIZU†

Abstract. This paper discusses asymptotic behaviors of the solution of the Poisson equation in domains with many tiny holes, where the number of holes grows to infinity and the diameter of each hole tends to zero. The solution satisfies a nonautonomous semilinear boundary condition on fragmentary boundaries of many tiny holes. Sufficient conditions are given, under which an extension of the solution to the domain with no hole converges, and the equation satisfied by the limit of the extension of the solution is determined. The convergence properties of Radon measures are applied, and also the convergence properties of the resolvent of subdifferentials together with variational inequalities are applied.

Key words. Poisson equation, semilinear boundary condition, domains with many tiny holes, homogenization, variational inequalities, subdifferentials, asymptotic behavior

AMS(MOS) subject classifications. 35B25, 35B40, 35J05, 35J20

Introduction. Let Ω be a bounded domain in $\mathbb{R}^N$, $N \geqq 2$, with smooth boundary $\partial\Omega$ and let $Y = [-\frac{1}{2}, \frac{1}{2}]^N$. Let T be a starlike, subdomain of Y, with smooth boundary ∂T, such that $T \ni 0$ and, with some $c^* > 1$,

$$c^* T \subset Y. \tag{0.0}_a$$

We simply denote by ε and r_ε the values of each sequence of $\{\varepsilon\}$ and $\{r_\varepsilon\}$ of positive numbers decreasing to zero, such that $\varepsilon/2 \geqq r_\varepsilon$. Let $Y_\varepsilon^i = p_\varepsilon^i + \varepsilon Y$ and $T_\varepsilon^i = p_\varepsilon^i + r_\varepsilon T$, where $p_\varepsilon^i \in \mathbb{R}^N$, $i \in \mathbb{N}$, are lattice points of edge length ε, by measurement in a parallel direction to each coordinate axis, i.e., $\varepsilon\mathbb{Z}^N = \{p_\varepsilon^i;\ i \in \mathbb{N}\}$. Let

$$T_\varepsilon = \bigcup \{T_\varepsilon^i;\ p_\varepsilon^i + c^* r_\varepsilon T \subset \Omega^-\} \quad \text{and} \quad \Omega_\varepsilon = \Omega \setminus T_\varepsilon. \tag{0.0}_b$$

We call T_ε the holes in Ω_ε, and Ω_ε a domain with holes (Fig. 1). Let β be a maximal monotone graph in $\mathbb{R} \times \mathbb{R}$ such that $\beta(0) \ni 0$. Let $f \in L^2(\Omega)$ and $g \in H^{1/2}(\partial T)$. We set $g_\varepsilon(x) = g((x - p_\varepsilon^i)/r_\varepsilon)$ for $x \in \partial T_\varepsilon^i$.

We consider the boundary value problem (P_ε):

$$-\Delta u_\varepsilon = f \quad \text{a.e. in } \Omega_\varepsilon, \tag{0.1}$$

$$\partial u_\varepsilon / \partial \nu + \alpha_\varepsilon \beta(u_\varepsilon) \ni g_\varepsilon \quad \text{on } \partial T_\varepsilon, \tag{0.2}$$

$$u_\varepsilon = 0 \quad \text{on } \partial\Omega, \tag{0.3}$$

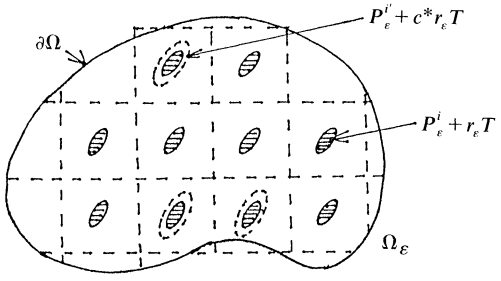


FIG. 1

* Received by the editors October 10, 1989; accepted for publication (in revised form) October 3, 1990.
† Department of Computer Science and Information Mathematics, Faculty of Electro-Communications, The University of Electro-Communications, 1-5-1, Chofugaoka, Chofu-shi, Tokyo, 182, Japan.

where ν denotes the outer unit normal to $\partial\Omega_\varepsilon$ and α_ε denotes a positive constant. We call g_ε a nonautonomous term, which is usually called a nonhomogeneous term in linear cases. Let $V_\varepsilon = \{v \in H^1(\Omega_\varepsilon);\ v|\partial\Omega = 0\}$. Let us recall that there exists a unique weak solution $u_\varepsilon \in V_\varepsilon \cap H^2(\Omega_\varepsilon)$ (see Brezis [4]). In particular, the boundary condition (0.2) includes the following conditions:

(a) Dirichlet's homogeneous boundary condition (an autonomous boundary condition), $\beta_D(\xi) = \varnothing$ for $\xi \neq 0$ and $\beta_D(0) = \mathbb{R}$.

(b) Neumann's boundary condition, $\beta_N(\xi) = \{0\}$.

(c) Robin's boundary condition, $\beta_R(\xi) = \xi$.

$(d)_p$ $\beta(\xi) = \xi|\xi|^{p-1}$ with $p > 0$.

(e) $C_1|\xi| \leqq |\beta(\xi)| \leqq C_2|\xi|$ and $\beta(\xi) \in C(\mathbb{R})$ with positive constants C_1 and C_2.

(f) Heaviside's boundary condition:

$$\beta_H(\xi) = \begin{cases} \{1\} & \text{for } \xi > 0, \\ [0, 1] & \text{for } \xi = 0, \\ \{0\} & \text{for } \xi < 0. \end{cases}$$

(g) Signorini's boundary condition:

$$\beta_S(\xi) = \begin{cases} \phi & \text{for } \xi > 0, \\ [0, \infty) & \text{for } \xi = 0, \\ \{0\} & \text{for } \xi < 0. \end{cases}$$

Thus, condition (0.2) covers a wide class of nonautonomous nonlinear boundary conditions (see Remark 0.1).

Remark 0.1. The type of (0.2) contains the homogeneous Dirichlet boundary condition as seen in the above (a). But it does not contain the nonhomogeneous Dirichlet boundary condition, which is represented by (a)′ below using β_D with $c \neq 0$.

(a)′ Dirichlet's nonhomogeneous boundary condition (a nonautonomous boundary condition), $\beta(\xi) = \beta_D(\xi - c)$ for $\xi - c \in D(\beta_D)$ $(= D(\beta))$ with $c \neq 0$.

For the definition (0.0) of the domain Ω_ε, later we show that there exists a sequence $\{\tilde{u_\varepsilon} \in H_0^1(\Omega)\}$ of extensions of $\{u_\varepsilon\}$, bounded in the space $H_0^1(\Omega)$. Our aim is to seek sufficient conditions, under which $\tilde{u_\varepsilon}$ converges weakly in $H_0^1(\Omega)$ as $\varepsilon \to 0$, and to determine the equation in Ω, satisfied by the weak limit u of $\tilde{u_\varepsilon}$.

For the autonomous case in condition (0.2) with β such as (b), (c), $(d)_p$, and (e) we have three parameters $\bar{\alpha}$, a, and θ_λ, $0 \leqq \lambda \leqq N$, that play an important role in determining the behavior of u_ε. Here

$$\theta_\lambda = \lim_{\varepsilon \to 0} \theta_{\lambda,\varepsilon},$$

where

$$\theta_{\lambda,\varepsilon} = \begin{cases} r_\varepsilon^\lambda / \varepsilon^N & \text{for } 0 < \lambda \leqq N, \\ \varepsilon^{-N} (\log r_\varepsilon^{-1})^{-1+N} & \text{for } \lambda = 0, \end{cases} \tag{0.4}$$

$$\bar{\alpha} = \lim_{\varepsilon \to 0} \bar{\alpha}_\varepsilon \quad \text{and} \quad \bar{\alpha}_\varepsilon = \alpha_\varepsilon r_\varepsilon, \tag{0.5}$$

and

$$a = \lim_{\varepsilon \to 0} a_\varepsilon \quad \text{and} \quad a_\varepsilon = \alpha_\varepsilon \theta_{N-1,\varepsilon}. \tag{0.6}$$

All limit equations on $\tilde{u_\varepsilon}$ are drawn in a picture (see Kaizu [15]).

Thus our question is whether these parameters $\bar{\alpha}$, a, and θ_λ are still available for the nonautonomous case in (0.2) with β such as (b), (c), $(d)_p$, and (e), and whether they are also applicable even for (0.2) with β such as (f) and (g).

Recently, Conca and Donato [10] and Cioranescu and Donato [6] have shown that no different limit equation on $u_\varepsilon^\sim$ appears in the nonautonomous case of (0.2) with β such as (b) and (c), provided that $I_g = 0$, where $I_h = \int_{\partial T} g \, d\sigma$. However, if $I_g \neq 0$, the behavior of $u_\varepsilon^\sim$ differs greatly from that in the autonomous case.

The other purpose of this paper is to understand the works of Conca, Donato, and Cioranescu [6], [10] in a unified way under the boundary condition (0.2).

We review some works around the case of homogeneous linear boundary conditions, i.e., autonomous linear boundary conditions. Systematic works are done by Khruslov in [17] and [18] for the Dirichlet boundary condition, and in [19] and [20] for the Neumann boundary condition. In his work [18], strongly elliptic operators of order $2m$, $m \geqq 1$, are considered. He shows that the parameter θ_{N-2m} plays an important role together with the notion of capacity, in the process $\varepsilon \to 0$. Vanninathan [24], [25] and Cioranescu and Saint Jean Paulin [8] have studied the Neumann boundary condition case. In particular, Vanninathan [24], [25] has considered the behavior of eigenvalues of the Laplacian. The homogenized operator $\mathscr{H}$ is introduced by each study of Cioranescu and Saint Jean Paulin [8], Khruslov [19], and Vanninathan [24], [25] for "the Laplacian in domains with many tiny holes," and there it is shown that θ_N is critical. Different approaches are given by Rauch and Taylor [23] for the Laplacian with a general linear boundary condition, and by Cioranescu and Murat [7] and Ozawa [22] also for the same operator with the Dirichlet boundary condition. For the Robin boundary condition Kaizu considered in [12], under the special form of T (i.e., T is a ball) the author has shown that the parameter θ_{N-1} is important. We see in [15] that it remains true even for a general T.

The Laplacian under autonomous nonlinear boundary conditions is considered by Kaizu [13]–[16]. In [16], Kaizu has shown that the Hausdorff dimension and the Hausdorff measure play some roles, where domains $\{\Omega_\varepsilon\}$ in [16] are more complicated than those presented here. An interesting family of domains has also been studied recently by Damlamian and Donato [11].

Among various methods for our asymptotic problem, we apply and extend those of the notion of epi-convergence of functionals. Several ideas, inequalities, and some special test functions have appeared in Attouch [1], Cioranescu and Donato [6], Cioranescu and Murat [7], Cioranescu and Saint Jean Paulin [8], Conca [9], Conca and Donato [10], and Vanninathan [25].

Notation. We denote by $C, C_1, C_2, \cdots$, positive constants which are independent of ε. By $\|v\|_{p,G}$ we denote the L^p norm, $(\int_G |v|^p \, dx)^{1/p}$. In particular, we simply denote the norm by $\|v\|_G$ if $p = 2$.

1. Theorems. Let $E_\varepsilon : V_\varepsilon \to H_0^1(\Omega)$ be a linear bounded extension operator such that

$$\|\nabla(E_\varepsilon v)\|_{T_\varepsilon} \leqq C\|\nabla v\|_{\Omega_\varepsilon} \quad \text{and} \quad \|E_\varepsilon v\|_{T_\varepsilon} \leqq C(\|\nabla v\|_{\Omega_\varepsilon} + \|c\|_{\Omega_\varepsilon}),$$

where C is independent of ε. For the existence of $\{E_\varepsilon\}$ see Corollary 3.3 (for a special case $p = 2$ and $\varepsilon \sim r_\varepsilon$ as $\varepsilon \to 0$; see also § 4 of Cioranescu and Saint Jean Paulin [8]). All the results in our paper are available for such a family of extension operators. But, to fix the extensions $V_\varepsilon \to H_0^1(\Omega)$ in a simple and quick way, we adopt the following special mapping.

Let $u_\varepsilon^\sim \in H_0^1(\Omega)$ be the extension of the solution u_ε of (P_ε) defined by

$$\Delta u_\varepsilon^\sim = 0 \quad \text{on } T_\varepsilon \quad \text{and} \quad u_\varepsilon^\sim = u_\varepsilon \quad \text{on } \partial T_\varepsilon.$$

We consider the following conditions.

$(\beta\text{-}0)_a$ β is a monotone graph in $\mathbb{R}\times\mathbb{R}$, i.e., for $\eta_i\in\beta(\xi_i)$, $\xi_i\in D(\beta)=\{\xi\in\mathbb{R};\ \beta(\xi)\neq\emptyset\}$, $i=1, 2$, we have

$$(\eta_2-\eta_1)(\xi_2-\xi_1)\geqq 0.$$

$(\beta\text{-}0)_b$ β is maximal in the family of all monotone graphs in $\mathbb{R}\times\mathbb{R}$, and further we suppose that $\beta(0)\ni 0$ and $D(\beta)\neq\{0\}$.

$(\beta\text{-}1)$ β is continuous in $D(\beta)$ except at finite points.

$(\beta\text{-}2)$ Let $\beta^0(\xi)=\eta$, where $\eta\in\beta(\xi)$ and $|\eta|=\min\{|\eta'|;\ \eta'\in\beta(\xi)\}$ for $\xi\in D(\beta)$. We have $C_1\geqq 0$, $C_2>0$, and $r\geqq 0$ such that

$$|\beta^0(\xi)|\leqq C_1+C_2|\xi|^r \quad \text{for all } \xi\in D(\beta).$$

$(\beta\text{-}3)$ We have that $C>0$ and $s>0$ such that

$$|\beta^0(\xi)|\geqq C|\xi|^s \quad \text{for all } \xi\in D(\beta).$$

We do not consider Dirichlet's boundary condition (see (a) in § 0); thus, $D(\beta)\neq\{0\}$ in $(\beta\text{-}0)$ is supposed. The nonnegative value r is regarded as the minimum exponent satisfying $(\beta\text{-}2)$, while we regard s as a certain positive number satisfying $(\beta\text{-}3)$. If $D(\beta)\cap(0,\infty)$ is a bounded interval, C_2 and r are determined by the graph of $\beta^0(\xi)$, $\xi\in D(\beta)\cap(-\infty, 0)$. Now, for a fixed r we simply set

$$\rho=\begin{cases}\rho(r) & \text{for } N\geqq 3,\\ 2 & \text{for } N=2 \text{ and } r=0,\\ a \text{ number of } [c(r), 2) & \text{for } N=2 \text{ and } 0<r<\infty,\end{cases} \tag{1.1}$$

where $\rho(r)=2N(N+r(N-2))^{-1}$ and $c(r)=\max\{1, 2/(1+r)\}$ (see the proof of Lemma 4.2). For $N\geqq 3$ we have $2\geqq\rho(r)\geqq 1$ if and only if $0\leqq r\leqq N/(N-2)$. Let $\rho^*=\rho/(\rho-1)$ with convention $1^*=\infty$, and let

$$\kappa_{\varepsilon,\rho}=\begin{cases}(r_\varepsilon/\varepsilon)^{(\rho-1)N/\rho} & \text{for } 1\leqq\rho<N,\\ (r_\varepsilon/\varepsilon)^{N-1}[\log(c_0\varepsilon/r_\varepsilon)]^{(N-1)/N} & \text{for } \rho=N,\end{cases} \tag{1.2}$$

where $c_0=\max\{N^{1/2}/(2|y|);\ y\in\partial T\}$. We give $r_\varepsilon^{(\lambda)}$, which plays a scale for the size of each hole T_ε^i in Ω_ε, as follows:

$$r_\varepsilon^{(\lambda)}=\varepsilon^{N/\lambda} \quad \text{for } 0<\lambda\leqq N \quad \text{and} \quad r_\varepsilon^{(0)}=\exp\{-\varepsilon^{-N/(N-1)}\}.$$

The inequality $0\leqq\lambda<\mu\leqq N$ implies that $r_\varepsilon^{(\lambda)}\ll r_\varepsilon^{(\mu)}$ as $\varepsilon\to 0$. We get $r_\varepsilon^{(\lambda)}\sim r_\varepsilon$ as $\varepsilon\to 0$ if and only if $0<\theta_\lambda<\infty$, $r_\varepsilon^{(\lambda)}\ll r_\varepsilon$ as $\varepsilon\to 0$ if and only if $\theta_\lambda=\infty$, where θ_λ is determined by (0.4).

1.1. Estimates.

THEOREM A. *We suppose condition* $(\beta\text{-}0)$ *holds. Then we have*

$$\|\nabla u_\varepsilon\|_{\Omega_\varepsilon}\leqq C\left(\|f\|_\Omega+\theta_{N-1,\varepsilon}\left|\int_{\partial T} g\,d\sigma\right|+\|g\|_{\partial T}\kappa_{\varepsilon,2}\right).$$

Let

$$b=\lim_{\varepsilon\to 0} b_\varepsilon \quad \text{and} \quad b_\varepsilon=\theta_{N-1,\varepsilon}/\alpha_\varepsilon^{1/s}. \tag{1.3}$$

THEOREM B. *In Theorem* A *we suppose condition* $(\beta\text{-}3)$ *holds. Let* $\sigma=s+1$ *and* $\sigma'=\min\{s, 1\}+1$. *We further suppose that* $f\in L^{\sigma'*}(\Omega)$, $g\in L^{\sigma^*}(\partial T)$, (0.6) *with* $a=\infty$, (1.3) *with* $b=0$, *and*

$$r_\varepsilon^{(N-\sigma')}\ll r_\varepsilon\leqq\varepsilon \quad \text{as } \varepsilon\to 0.$$

Then $u_\varepsilon^{\sim}$ *converges strongly to zero in* $W_0^{1,\sigma'}(\Omega)$. *Furthermore, we get*

$$\|\nabla u_\varepsilon\|_{\sigma',\Omega_\varepsilon}^2 \leqq C[\|f\|_{\sigma'^*,\Omega}^2 \theta_{N-\sigma',\varepsilon}^{-2/\sigma'} + \|f\|_{\sigma'^*,\Omega}^{\sigma^*} a_\varepsilon^{-1/s} + b_\varepsilon \|g\|_{\sigma^*,\partial T}^{\sigma^*}].$$

THEOREM C. *In Theorem* A *we suppose that* $\int_{\partial T} g\, d\sigma \neq 0$ *and*

$$r_\varepsilon^{(N-1)} \ll r_\varepsilon \leqq \varepsilon \quad \text{as } \varepsilon \to 0.$$

(I) *We have*

$$\|\nabla u_\varepsilon\|_{\Omega_\varepsilon} \leqq C\theta_{N-1,\varepsilon}.$$

(II) *In addition to* (I) *we suppose that* $D(\beta) = \mathbb{R}$, (β-2) *with* $0 \leqq r \leqq 1$, *and* (0.6) *with* $0 \leqq a < \infty$. *Then we have*

$$\|\nabla u_\varepsilon\|_{\Omega_\varepsilon} \geqq C\theta_{N-1,\varepsilon}.$$

(III) *In* (I) *we suppose condition* (β-3) *and* (0.6) *holds with* $a = \infty$, $0 < b \leqq \infty$. *Then we have*

$$\|\nabla u_\varepsilon\|_{\sigma',\Omega_\varepsilon} \leqq C b_\varepsilon^{1/2}, \tag{1.4}$$

$$\|u_\varepsilon\|_{\sigma',\Omega_\varepsilon} \leqq C b_\varepsilon^{1/\sigma'} (\theta_{N-\sigma',\varepsilon}^{-1/\sigma'} + a_\varepsilon^{-1/\sigma}), \tag{1.5}$$

and

$$u_\varepsilon^{\sim}/b_\varepsilon^{1/\sigma'} \xrightarrow{w} 0 \quad \text{in } W_0^{1,\sigma'}(\Omega) \quad \text{as } \varepsilon \to 0.$$

Remark 1.1. For the bounds in (1.4) and Theorem C(I) we want to know $\lim \theta_{N-1,\varepsilon}/b_\varepsilon^{1/\sigma'}$. Let

$$a_{h,\varepsilon} \equiv \alpha_\varepsilon (\theta_{N-1,\varepsilon})^h. \tag{1.6}$$

Then we have $\theta_{N-1,\varepsilon}/b_\varepsilon^{1/\sigma'} = a_{s,\varepsilon}^{1/2s}$ for $s \geqq 1$ and $\theta_{N-1,\varepsilon}/b_\varepsilon^{1/\sigma'} = a_{\sigma'',\varepsilon}^{1/s\sigma'}$, $\sigma'' = s^2$, for $0 < s < 1$.

1.2. Convergence.

THEOREM a. *We suppose* (β-0), (β-1), *and* (β-2) *with*

$$\begin{aligned} &0 \leqq r \leqq 1 \quad \text{for } N \geqq 3, \\ &r = 0 \quad \text{for } N = 2. \end{aligned}$$

In the case where $1 > r \geqq 0$ *with* $N \geqq 3$, *using* (0.5) *and* (1.1) *we suppose that*

$$\bar{\alpha} < \infty, \quad \bar{\alpha}\theta_{N-\rho^*} = 0 \quad \text{and} \quad r_\varepsilon \leqq C r_\varepsilon^{(N-\rho^*)} \quad \text{as } \varepsilon \to 0. \tag{1.7}$$

Let $\tilde{\alpha}_\varepsilon = \bar{\alpha}_\varepsilon (\log r_\varepsilon^{-1})^{(N-1)/N^*}$. *In the case where* $r = 1$ *with* $N \geqq 3$, *we suppose that*

$$\tilde{\alpha} = \lim_{\varepsilon \to 0} \tilde{\alpha}_\varepsilon < \infty, \ \tilde{\alpha}\theta_0 = 0 \quad \text{and} \quad r_\varepsilon \leqq C r_\varepsilon^{(0)} \quad \text{as } \varepsilon \to 0. \tag{1.8}$$

Let $\hat{\alpha}_\varepsilon = \tilde{\alpha}_\varepsilon (\log (c_0 \varepsilon / r_\varepsilon))^{1/2}$. *In the case where* $r = 0$ *with* $N = 2$ *we suppose that*

$$\hat{\alpha} = \lim_{\varepsilon \to 0} \hat{\alpha}_\varepsilon < \infty, \quad \hat{\alpha}\theta_0 = 0, \quad \text{and} \quad r_\varepsilon \leqq C r_\varepsilon^{(0)} \quad \text{as } \varepsilon \to 0. \tag{1.9}$$

Then $u_\varepsilon^{\sim}$ *weakly converges to* u *in* $H_0^1(\Omega)$, *where* u *is the solution of*

$$-\Delta u = f \quad \text{a.e. in } \Omega.$$

THEOREM b. *We suppose conditions* (β-0), (β-1), *and* (β-2) *hold with*

$$\begin{aligned} &0 \leqq r < N/(N-2) \quad \text{for } N \geqq 3, \\ &0 \leqq r < \infty \quad \text{for } N = 2. \end{aligned} \tag{1.10}$$

We suppose (0.6) *holds with* $0 \leqq a \leqq \infty$, *and*

$$r_\varepsilon^{(N-\rho)} \ll r_\varepsilon \leqq C r_\varepsilon^{(N-1)} \quad \text{as } \varepsilon \to 0.$$

Let $S(\partial T)$ *be the surface area of* ∂T. *Then* $u_\varepsilon^{\sim}$ *converges to* u *weakly in* $H_0^1(\Omega)$, *where* u *satisfies*

$$-\Delta u + aS(\partial T)\beta(u) \ni f + \theta_{N-1} \int_{\partial T} g \, d\sigma \quad \text{a.e. in } \Omega, \tag{1.11}$$

for $0 \leqq a < \infty$, *otherwise, if* $a = \infty$, *then*

$$u = 0 \quad \text{a.e. in } \Omega,$$

provided that

$$\int_0^{\pm\delta} \beta(\xi)\, d\xi > 0 \quad \text{for any } \delta > 0. \tag{β-4}$$

1.3. Convergence after renormalization. We have renormalization on u_ε as follows:

$$z_\varepsilon \equiv u_\varepsilon / \theta_{N-1,\varepsilon}.$$

We recall (I) and (II) in Theorem C. It is natural to consider the behavior of z_ε when $\theta_{N-1,\varepsilon} \to \infty$ as $\varepsilon \to 0$.

(A) Let h be a fixed positive number such that $h \geqq r$ and let

$$\beta^\varepsilon(\xi) \equiv \beta(\theta_{N-1,\varepsilon}\xi)/\theta_{N-1,\varepsilon}^h \quad \text{for } \xi \in D(\beta^\varepsilon),$$

where $D(\beta^\varepsilon) = \{\xi \in \mathbb{R};\ \theta_{N-1,\varepsilon}\xi \in D(\beta)\}$. A multivalued function β^ε satisfies (β-0) and (β-1) as β. Each element β^ε of $\{\beta^\varepsilon;\ \theta_{N-1,\varepsilon} \to \infty\}$ also satisfies (β-2) with the same r as β. We have $\beta^\varepsilon \equiv \beta$ for a homogeneous function of degree h. In this section, one of the following conditions, (γ-1), (γ-2), (γ-3), is supposed.

(B) Conditions.

(γ-1) β satisfies (β-2) with (1.10). Besides, $D(\beta^0) = D(\beta^\varepsilon) = \mathbb{R}$ and $|\beta^\varepsilon(\xi) - \beta^0(\xi)| \leqq c_\varepsilon(1+|\xi|)^t$ with $c_\varepsilon \to 0$ as $\varepsilon \to 0$ and $0 \leqq t \leqq r$.

(γ-2) β satisfies (β-2) with (1.10). Besides, $D(\beta^0) = D(\beta^\varepsilon) = \mathbb{R}$ and $\beta^\varepsilon \uparrow \beta^0$ or $\beta^\varepsilon \downarrow \beta^0$ as $\varepsilon \downarrow 0$.

(γ-3) β satisfies (β-2) with $r = 0$, $D(\beta^0) = \liminf_{\varepsilon \to 0} D(\beta^\varepsilon)$ and β^ε converges to β^0 as $\varepsilon \to 0$ pointwise in $D(\beta^0)$ except at discontinuous points of β^0.

(γ-4) β^0 satisfies (β-0) and (β-1).

Example 1.2. Let $\theta_\varepsilon = \theta_{N-1,\varepsilon}$. For $\beta(\xi) = |\xi|^{r-1}\xi$, we get $\beta^\varepsilon(\xi) = \theta_\varepsilon^{r-h}\beta(\xi)$. We have ($\gamma$-1) with $\beta^0 = \beta$ for $h = r$, and $\beta^0 = \beta_N$ for $h > r$. For

$$\beta(\xi) = \begin{cases} 1 & \text{for } \xi \geqq 1, \\ \xi & \text{for } 0 \leqq \xi \leqq 1, \\ 0 & \text{for } \xi \leqq 0, \end{cases}$$

which satisfies (β-0), (β-1), and (β-2) with $r = 0$, we have

$$\beta^\varepsilon(\xi) = \begin{cases} \theta_\varepsilon^{-h} & \text{for } \xi \geqq \theta_\varepsilon^{-1}, \\ \theta_\varepsilon^{1-h}\xi & \text{for } 0 \leqq \xi \leqq \theta_\varepsilon^{-1}, \\ 0 & \text{for } \xi \leqq 0; \end{cases}$$

(γ-2): $\beta^\varepsilon \uparrow \beta_H$ as $\varepsilon \downarrow 0$ for $h=0$; (γ-2): $\beta^\varepsilon \downarrow \beta_N$ as $\varepsilon \uparrow 0$ for $h \geqq 1$; (γ-1): $\|\beta^\varepsilon - \beta_N\|_\infty \to 0$ as $\varepsilon \to 0$ for $0<h<1$. If

$$\beta(\xi)=\begin{cases}\infty & \text{for } \xi \geqq 1,\\ \xi & \text{for } 0 \leqq \xi \leqq 1,\\ 0 & \text{for } \xi \leqq 0\end{cases}$$

is chosen, we have (β-0), (β-1), and (β-2) with $r=0$. Then

$$\beta^\varepsilon(\xi)=\begin{cases}\infty & \text{for } \xi \geqq \theta_\varepsilon^{-1},\\ \theta_\varepsilon^{1-h}\xi & \text{for } 0 \leqq \xi \leqq \theta_\varepsilon^{-1},\\ 0 & \text{for } \xi \leqq 0.\end{cases}$$

Thus (γ-3) is satisfied with $\beta^0=\beta_S$.

(C) The homogenized operator. Let $Y^* = Y \setminus T^*$ and $T^* = \theta_N^{1/N} T$. Let $|Y^*|$ be the Lebesgue measure of Y^*, and let

$$\mathcal{H}=\begin{cases}-\Delta & \text{for } \theta_N=0,\\ -\sum\limits_{i,j=1}^{N} q_{ij}\dfrac{\partial^2}{\partial x_i \partial x_j} & \text{for } 0<\theta_N \leqq 1,\end{cases}$$

where q_{ij}, $1 \leqq i, j \leqq N$, are constants defined by $q_{ij}=|Y^*|\delta_{ij}-\int_{Y^*} \nabla \kappa_i \nabla \kappa_j \, dy$. Here $\kappa_i \in W_0$, is defined by

$$\int_{Y^*} \nabla \kappa_i \nabla v \, dy + \int_{\partial T^*} \nu_i v \, d\sigma = 0 \quad \text{for all } v \in W_0,$$

where $W_0=\{v \in H^1(Y^*); \int_{Y^*} v\,dy=0$ and v is Y-periodic$\}$ and $\nu=\{\nu_i; 1 \leqq i \leqq N\}$ is the outer unit normal to ∂Y^*.

(D) THEOREM c. *In Theorem* C(I) *we further suppose conditions* (β-0), (β-1), *and* (γ-4) *hold and one of the conditions* (γ-1), (γ-2), (γ-3) *holds.*

(I) *We suppose that* $a_h=\lim_\varepsilon a_{h,\varepsilon}<\infty$ *with* (1.6). *Then* $z_\varepsilon^\sim$ *weakly converges in* $H_0^1(\Omega)$ *to* z, *where* z *is a unique solution of*

$$\mathcal{H}z + a_h S(\partial T)\beta^0(z) \ni \int_{\partial T} g \, d\sigma \quad \text{a.e. in } \Omega. \tag{1.12}$$

(II) *If* $a_h=\infty$ *and* (β-4) *is satisfied by* β^0, *then* $z_\varepsilon^\sim$ *converges weakly to zero in* $H_0^1(\Omega)$.

1.4. Convergence for continuous functions β. For continuous functions β we have more results as below for the case where $Cr_\varepsilon^{(N-2)} \geqq r_\varepsilon$ than in Theorem a, and for the case where $r_\varepsilon^{(N-2)} \ll r_\varepsilon \leqq Cr_\varepsilon^{(N-1)}$ than in Theorem b, respectively.

THEOREM a′. *We suppose that* $N \geqq 3$, *and conditions* (β-0), (β-2), *and* (β-3) *hold with* $C_1=0$, $r=1=s$, *and* (β-1)$_0$ *below.*

(β-1)$_0$ $D(\beta)=\mathbb{R}$ *and* β *is continuous in* $\mathbb{R}$.

Besides, we suppose (0.5) *holds with* $\bar{\alpha}=\infty$, *and* $Cr_\varepsilon^{(N-2)} \geqq r_\varepsilon$ *with a fixed* θ_{N-2} *as* $\varepsilon \to 0$. *Then* $u_\varepsilon^\sim$ *converges weakly to* u *in* $H_0^1(\Omega)$ *as* $\varepsilon \to 0$, *where* u *satisfies either*

$$-\Delta u + \theta_{N-2} C_T u = f \quad \text{a.e. in } \Omega \quad \text{for } \theta_{N-2}<\infty,$$

or

$$u=0 \quad \text{a.e. in } \Omega \quad \text{for } \theta_{N-2}=\infty,$$

where C_T denotes the capacity of T defined by

$$C_T = \inf\left\{\int_{\mathbb{R}^N} |\nabla v|^2\, dx;\ v \in H^1(\mathbb{R}^N) \text{ and } v \geqq 1 \text{ on } T\right\}.$$

For $\bar{\alpha} < \infty$, we can have a value $C(\bar{\alpha}, T)$ also like the capacity of T (for more detail see Theorem A.1 in Kaizu [15]).

THEOREM b′. *We suppose* (β-0), (β-3), *and* (β-1)$_1$ *below hold.*

(β-1)$_1$ $D(\beta) = \mathbb{R}$ *and* β *is Lipschitz continuous.*

We also suppose (0.6) *holds with* $0 \leqq a \leqq \infty$, $r_\varepsilon^{(N-2)} \ll r_\varepsilon \leqq C r_\varepsilon^{(N-1)}$. *Then* $u_\varepsilon^{\sim}$ *converges in* $H_0^1(\Omega)$ *weakly to* u, *where* u *is the solution of*

$$-\Delta u + aS(\partial T)\beta(u) = f + \theta_{N-1}\int_{\partial T} g\, d\sigma \quad \text{a.e. in } \Omega \quad \text{for } a < \infty,$$

$$u = 0 \quad \text{a.e. in } \Omega \quad \text{for } a = \infty.$$

Remark 1.3. The range $r_\varepsilon^{(N-2)} \ll r_\varepsilon \leqq C r_\varepsilon^{(N-1)}$ of r_ε in Theorem b′ is slightly wider than that of r_ε in Theorem b.

2. Convergence of subdifferentials. To our asymptotic problem we shall apply the theory of the convergence of a sequence of subdifferentials, described below.

Let H be a real Hilbert space of lower semicontinuous convex functions $\varphi^n : H \to [0, \infty]$ with $\varphi^n(0) = 0$. We set $D(\varphi^n) = \{x \in H;\ \varphi^n(x) < \infty\}$, which we call the effective domain of φ^n. Let H^n be a closed linear subspace which contains the closure of $D(\varphi^n)$ in the topology of H. Let p^n be the projection from H onto H^n (see Example 2.3). Let φ^0 be a lower semicontinuous convex function from H into $[0, \infty]$, with $\varphi^0(0) = 0$. Let $D(\varphi^0)$ be the effective domain of φ^0. We suppose the following conditions:

(c-1) There exists a sequence $\{E^n\}$ of linear mappings E^n: $D(\varphi^n) \to D(\varphi^0)$, satisfying the following properties:
  (i) $p^n E^n = p^n$ and $(x, y) = (E^n x, p^n y)$ for all $x, y \in D(\varphi^n)$;
  (ii) If $\limsup_{n\to\infty} \varphi^n(x^n) < \infty$, then the sequence $\{E^n x^n\}$ is totally bounded.

(c-2) There exists a positive constant η such that the strong convergence $x^n \to x$, implies the weak convergence $p^n x^n \xrightarrow{w} \eta x$.

Now, we write $\varphi^n \xrightarrow{(e)} \varphi^0$, if and only if the following two conditions are satisfied:

(e-1) For any $x \in D(\varphi^0)$ we have $x^n \in D(\varphi^n)$ such that $E^n x^n \xrightarrow{s} x$ and

$$\varphi^0(x) \geqq \limsup_{n\to\infty} \varphi^n(x^n).$$

(e-2) The convergence $E^n x^n \xrightarrow{s} x$, implies the inequality

$$\varphi^0(x) \leqq \liminf_{n\to\infty} \varphi^n(x^n).$$

It is clear that, for the case $E^n \equiv$ the identity map, we have $\varphi^n \xrightarrow{(e)} \varphi^0$ if and only if φ^0 is the τ-epi-limit of φ^n, where τ is the topology of H (see Attouch [1, Prop. 1.14, p. 30]).

The subdifferential $\partial\varphi$ of a lower semicontinuous convex function φ is defined as follows: $\partial\varphi(x) = \{y \in H;\ \varphi(z) \geqq \varphi(x) + (y, z - x) \text{ for all } z \in H\}$. It is well known that we have the equivalent relation between Mosco's convergence and the convergence of the resolvent mappings of the subdifferentials (see [1, Thm. 3.26, p. 305]). Even for the above convergence we still get the following theorem.

THEOREM 2.1. *We suppose conditions* (c-1) *and* (c-2) *hold. Then the following statements are equivalent*:

(I) $\varphi^n \xrightarrow{(e)} \varphi^0$.

(II) *Let* $f^n \xrightarrow{s} f^0$ *and* x^n *be the solution of* $(\lambda\partial\varphi^n + I)x^n \ni f^n$, $\lambda > 0$. *Then* $E^n x^n$ *converges strongly and the limit* x *satisfies* $(\lambda\partial\varphi^0 + \eta I)x \ni \eta f^0$.

(III) *Let* x^n *be the solution of* $(\lambda\partial\varphi^n + I)x^n \ni f$, $\lambda > 0$. *Then* $E^n x^n$ *converges strongly and the limit* x *satisfies* $(\lambda\partial\varphi^0 + \eta I)x \ni \eta f$.

PROPOSITION 2.2. (a) *The following statement* (IV) *is sufficient for the statement* (III).

(IV) *Let* $f^n \xrightarrow{s} f^0$ *and* x^n *be the solution of* $\partial\varphi^n(x^n) \ni f^n$. *Then* $E^n x^n$ *converges strongly and the limit* x *satisfies* $\partial\varphi^0(x) \ni \eta f^0$.

(b) *If* $\{(\partial\varphi^n)^{-1}\}$ *is uniformly bounded, the statement* (IV) *is necessary for* (II).

Example 2.3. Let $H = L^2(\Omega)$ and $V^0 = H_0^1(\Omega)$. We denote by $\{\varepsilon_n\}$ a sequence of positive numbers decreasing to zero. Let Ω^n be a subdomain of Ω defined by (0.0). We consider $H^n = L^2(\Omega^n)$, as a subspace of H, which consists of functions of H, vanishing outside Ω^n. We write $V^n = V_\varepsilon$ with $\varepsilon = \varepsilon_n$. Let φ^n, φ^0: $H \to [0, \infty]$, be lower semicontinuous convex functions defined by

$$\varphi^n(v) = \begin{cases} \displaystyle\int_{\Omega^n} |\nabla v|^2 \, dx/2 & \text{for } v \in V^n, \\ \infty & \text{otherwise,} \end{cases}$$

$$\varphi^0(v) = \begin{cases} \displaystyle\int_{\Omega} (\nabla v, \nabla v)_Q \, dx/2 & \text{for } v \in V^0, \\ \infty & \text{otherwise,} \end{cases}$$

where

$$(\nabla v, \nabla v)_Q = \sum_{i,j=1}^{N} Q_{ij} \frac{\partial v}{\partial x_i} \frac{\partial v}{\partial x_j}$$

and

$$Q_{ij} = \begin{cases} \delta_{ij}, & \text{the Kronecker delta} \quad \text{for } \theta_N = 0, \\ q_{ij}, & \text{constants given in §§ 1.3} \quad \text{for } 0 < \theta_N \leqq 1. \end{cases}$$

For $v \in V^n$ we define $E^n(v) = v^\sim \in V^0$, where $\Delta v^\sim = 0$ on T_ε and $v^\sim = v$ on ∂T_ε with $\varepsilon = \varepsilon_n$ (we can choose another extension family $\{E^n\}$ in Corollary 3.3). Let $p^n v = \chi^n v$, $v \in H$, where χ^n denotes the characteristic function of Ω^n. Then (c-1) is satisfied by combining Corollary 3.3 with the Rellich theorem. We see that $\chi^n \to |Y^*|(=\eta)$ weakly star in $L^\infty(\Omega)$ as $n \to \infty$. We consider (P_ε) with $\alpha_\varepsilon \equiv 0$ and $g \equiv 0$, and have $\partial\varphi^n(u^n) \ni f$. According to Cioranescu and Saint Jean Paulin [8] and Khruslov [19], we have that $u_n^\sim \to u$ weakly in $H_0^1(\Omega)$ as $\varepsilon \to 0$, where

$$-\Delta u = f \quad \text{in } \Omega \quad \text{for } \theta_N = 0,$$

$$\mathcal{H}u = \eta f \quad \text{in } \Omega \quad \text{for } \theta_N > 0.$$

The convergence of $u_n^\sim$ to u remains true, even if f is replaced with f^n, which strongly converges to f, in the equation $\partial\varphi^n(u^n) \ni f$. By Proposition 2.2(a), we get

$$\varphi^n \xrightarrow{(e)} \varphi^0. \tag{2.1}$$

We consider a perturbation of (2.1) as follows.

PROPOSITION 2.4. *We suppose that* $\varphi^n \xrightarrow{(e)} \varphi^0$. *Besides we suppose that* $j^n, j^0 : H \to [0, \infty]$, *satisfy the following conditions*:

(i) $\varphi^n + j^n$ *is a lower semicontinuous convex function for each* $n \in \mathbb{N} \cup \{0\}$.

(ii) *We have maps* $F^n : D(\varphi^n) \to D(\varphi^n + j^n)$, *such that* $\varphi^n(F^n x) \leqq \varphi^n(x)$ *for* $x \in D(\varphi^n)$, *and* $E^n F^n x^n \to x^0$ *for any* $E^n x^n \to x^0 \in D(\varphi^0 + j^0)$ *as* $n \to \infty$.

(iii) $j^n(x^n) \to j^0(x)$ *as* $n \to \infty$ *for any sequence* $\{x^n\}$ *such that* $\sup\{\varphi^n(x^n) + j^n(x^n)\} < \infty$ *and* $E^n x^n \to x$.

Then we get

$$\varphi^n + j^n \xrightarrow{(e)} \varphi^0 + j^0.$$

Proof of Theorem 2.1. It is trivial that (II) $\to$ (III). We show (III) $\to$ (I). Let $J_\lambda^n = (\lambda \partial \varphi^n + I)^{-1}$. The Yosida approximation of φ_λ^n, $\lambda > 0$, is defined by

$$\varphi_\lambda^n(x) = \inf\{\varphi^n(y) + \|y - x\|^2/(2\lambda);\ y \in H\}$$

for all $x \in H$. The following properties are fundamental.

(a) φ_λ^n is of class C^1 and its derivative A_λ^n is Lipschitz continuous with constant $1/\lambda$. Besides we have $|A_\lambda^n x| \leqq |y|$ for all $y \in \partial^n \varphi(x)$, $x \in D(\partial \varphi^n)$.

(b) $\varphi_\lambda^n(x) \uparrow \varphi^n(x)$ as $\lambda \downarrow 0$.

(c) $\varphi_\lambda^n(x) = \lambda \|A_\lambda^n x\|^2/2 + \varphi^n(J_\lambda^n x)$.

(d) $\lambda A_\lambda^n + J_\lambda^n = I$.

(e) $p^n = s - \lim_{\lambda \to 0} J_\lambda^n$.

(f) $A^n J_\lambda^n \supset A_\lambda^n$.

LEMMA 2.5. *Under* (c-2) *and* (III), *we have*

(a) $(A_\lambda^n y, x^n) \to (A_{\lambda/\eta}^0 y, x)$ *as* $n \to \infty$, *for any* $y \in H$, $x^n \in D(\varphi^n)$, $E^n x^n \to x$,

(b) $(A_\lambda^n y, x) - (y - p^n y, x)/\lambda \to (A_{\lambda/\eta}^0 y, x)$ *as* $n \to \infty$, *for any* x *and* $y \in H$,

(c) $(A_\lambda^n y, x) \to (A_{\lambda/\eta}^0 y, x) + (1 - \eta)(y, x)/\lambda$ *as* $n \to \infty$, *for any* x *and* $y \in H$.

Proof. By (c-2) and (III), we have $E^n J_\lambda^n y \to J_{\lambda/\eta}^0 y$, $x^n \xrightarrow{w} \eta x$ and $p^n y \xrightarrow{w} \eta y$ as $n \to \infty$. By (d) we get Lemma 2.5(a) as follows:

$$\begin{aligned}(A_\lambda^n y, x^n) &= [(y, p^n E^n x^n) - (E^n J_\lambda^n y, p^n E^n x^n)]/\lambda \\ &\to \eta[(y, x) - (J_{\lambda/\eta}^0 y, x)]/\lambda = (A_{\lambda/\eta}^0 y, x).\end{aligned}$$

Combining (d), (c-2), and (III) with $(J_\lambda^n y, x) = (p^n E^n J_\lambda^n y, x)$, we get

$$(A_\lambda^n y, x) = [(p^n y - p^n E^n J_\lambda^n y, x) + (y - p^n y, x)]/\lambda,$$

$$(A_\lambda^n y, x) - (y - p^n y, x)/\lambda \to (A_{\lambda/\eta}^0 y, x) \quad \text{as } n \to 0.$$

Lemma 2.5(b) implies (c) together with (c-2). □

LEMMA 2.6. *Under* (c-2), (III) *implies the following*:

(a) $\lim_{n \to \infty} \varphi_\lambda^n(x) - (\|x\|^2 - \lim_{n \to \infty} \|p^n x\|^2)/(2\lambda) = \varphi_{\lambda/\eta}^0(x)$,

(b) $\lim_{n \to \infty} \varphi_\lambda^n(x) = \varphi_{\lambda/\eta}^0(x) + (1 - \eta)\|x\|^2/(2\lambda)$.

Proof. By (a), Lemma 2.5(b), the Lebesgue convergence theorem and $\varphi^n(0) = 0 = \varphi^0(0)$, we get

$$\begin{aligned}&\varphi_\lambda^n(x) - (\|x\|^2 - \|p^n x\|^2)/(2\lambda) \\ &\quad = \int_0^1 [(A_\lambda^n(sx), x) - s\lambda^{-1}(x - p^n x, x)]\, ds \\ &\quad \to \int_0^1 (A_{\lambda/\eta}^0(sx), x)\, dx = \varphi_{\lambda/\eta}^0(x).\end{aligned}$$

This shows (a) of this lemma. So, we get (b) of this lemma. □

First we show (e-2). By (c), (d), and (f) we get

$$\begin{aligned}\varphi^n(x^n) &\geqq \varphi^n(J_\lambda^n x)+(A_\lambda^n x, x^n - J_\lambda^n x)\\ &= (\varphi_\lambda^n(x)-\lambda\|A_\lambda^n x\|^2/2)+(A_\lambda^n x, x-J_\lambda^n x)+(A_\lambda^n x, x^n - x)\\ &= \varphi_\lambda^n(x)+\lambda\|A_\lambda^n x\|^2/2+(A_\lambda^n x, x^n - x)\\ &= \varphi_\lambda^n(x)+(x-J_\lambda^n x, A_\lambda^n x)/2+(A_\lambda^n x, x^n - x) = B_\lambda^n(x).\end{aligned}$$

Applying (d), Lemma 2.6(b), and (a) and (c) of Lemma 2.5 we have

$$B_\lambda^n(x) \to \varphi_{\lambda/\eta}^0(x)+\lambda\|A_{\lambda/\eta}^0 x\|^2/(2\eta)+(1-\eta)(1-\lambda)\|x\|^2/(2\lambda)$$

as $n\to\infty$. Thus, we get

$$\liminf_{n\to\infty} \varphi_\lambda^n(x^n) \geqq \varphi_{\lambda/\eta}^0(x).$$

By (b) we have shown condition (e-2).

Next we show (e-1). It suffices to show that, for any subsequence, still denoted by $\{\varphi^n\}$, there exists a subsequence of $\{\varphi^n\}$ satisfying (e$-$1). By (e) we have λ_n such that $\|x - J_{\lambda/\eta}^0 x\| \leqq (2n)^{-1}$, $\lambda < \lambda_n$, $n\in\mathbb{N}$. By (III) and Lemma 2.6(a) we get $m'(n)$ such that $\|J_{\lambda_n/\eta}^0(x) - E^m J_{\lambda_n}^m(x)\| \leqq (2n)^{-1}$ and $\varphi_{\lambda_n}^m(x)-(\|x\|^2-\|p^m x\|^2)/(2\lambda_n) \leqq \varphi_{\lambda_n/\eta}^0(x)+1/n$ for $m\geqq m'(n)$. For $x^{n'} = J_{\lambda_n}^{m'}(x)$, $n'\equiv m'(n)$, we get $\|E^{n'}x^{n'}-x\|\leqq 1/n$. By (c), we have

$$\begin{aligned}\varphi_{\lambda_n/\eta}^0(x)+\frac{1}{n} &\geqq \varphi_{\lambda_n}^{n'}(x)+(\|p^{n'}x\|^2-\|x\|^2)/(2\lambda_n)\\ &= \varphi^{n'}(x^{n'})+(\|\lambda_n A_{\lambda_n}^{n'}x\|^2+\|p^{n'}x\|^2-\|x\|^2)/(2\lambda_n)\\ &= \varphi^{n'}(x^{n'})+C_\lambda^n(x)/(2\lambda_n).\end{aligned}$$

By (d) we have $C_\lambda^n(x) = \|J_{\lambda_n}^{n'}x\|^2+\|p^{n'}x\|^2-2(x, J_{\lambda_n}^{n'}x)$. Since $p^{n'}x = x+(p^{n'}x-x)$, we get

$$\begin{aligned}C_\lambda^{n'}(x) &= (J_{\lambda_n}^{n'}x, J_{\lambda_n}^{n'}x-x)-(J_{\lambda_n}^{n'}x-p^{n'}x, x)\\ &= \|J_{\lambda_n}^{n'}x-x\|^2-\|(I-p^{n'})x\|^2\\ &= \|J_{\lambda_n}^{n'}x-p^{n'}x\|^2 \geqq 0.\end{aligned}$$

Thus, by (b) we have

$$\varphi^0(x) \geqq \limsup_{n\to\infty}\left(\varphi_{\lambda_n/\eta}^0(x)+\frac{1}{n}\right) \geqq \limsup_{n\to\infty}\varphi^{n'}(x^{n'}).$$

Since $E^{n'}x^{n'}\to x$, we have shown (e-1), and so, (I).

Finally, we show the implication (I)$\to$(II). We have

$$\varphi^n(z) \geqq \varphi^n(x^n)+\lambda^{-1}(f^n-x^n, z-x^n) \tag{2.2}$$

for all $z\in H$. Since J_λ^n is contractive, $\{x^n\}$ is bounded. By $\varphi^n(0)=0$, together with (c-1) and (2.2), $\{E^n x^n\}$ is totally bounded. We extract a subsequence, still denoted by $\{E^n x^n\}$ such that $E^n x^n \to x$ as $n\to\infty$. By (e-1), for $z\in D(\varphi^0)$ we have z^n such that $E^n z^n \to z$ as $n\to\infty$, and $\limsup_{n\to\infty}\varphi^n(z^n)\leqq\varphi^0(z)$. Applying (c-2), we get

$$(f^n-x^n, z^n) = (f^n - E^n x^n, p^n E^n z^n) \to \eta(f^0-x, z),$$

$$\|x^n\|^2 = (p^n x^n, E^n x^n) \to \eta\|x\|^2 \quad \text{as } n\to\infty.$$

Putting $z = z^n$ into (2.2) and applying (e-1) and (e-2), we have

$$\begin{aligned}\varphi^0(z) &\geqq \limsup_{n\to\infty} \varphi^n(z^n)\\ &\geqq \liminf \varphi^n(x^n) + \lambda^{-1}\eta(f^0 - x, z - x)\\ &\geqq \varphi^0(x) + \lambda^{-1}\eta(f^0 - x, z - x).\end{aligned}$$

This shows $J^0_{\lambda/\eta} f^0 = x$. Thus, the full sequence $\{E^n x^n\}$ converges to $J^0_{\lambda/\eta} f^0$ as $n \to \infty$. This means that (I) $\to$ (II). □

Proof of Proposition 2.2. (a) Let x^n be the solution of $(\lambda \partial\varphi^n + I)x^n \ni f$, i.e., $\partial\varphi^n(x^n) \ni (f - x^n)/\lambda$. Since J^n_λ is contractive, $\{x^n\}$ is bounded. By the definition of $\partial\varphi^n$ with (c-1), we see that $\{E^n x^n\}$ is totally bounded. We extract a subsequence such that $E^n x^n \to x$, and we also have that $\partial\varphi^n(x^n) \ni (f - E^n x^n)/\lambda$. By (IV) we get $\partial\varphi^0(x) \ni \eta(f - x)/\lambda$. This means that (III) is true.

(b) Let x^n be the solution of $\partial\varphi^n(x^n) \ni f^n$. Since $\{f^n\}$ is bounded and $\{(\partial\varphi^n)^{-1}\}$ is uniformly bounded, $\{x^n\}$ is bounded. By the defintion of $\partial\varphi^n$ with (c-1), $\{E^n x^n\}$ is totally bounded. We suppose $E^n x^n \to x$. We have $(\partial\varphi^n + I)(x^n) \ni f^n + E^n x^n$ and $f^n + E^n x^n \to f^0 + x$ as $n \to \infty$. By (II) we get $(\partial\varphi^0 + \eta I)(X) \ni \eta(f^0 + x)$. This means that (IV) is true. □

3. Lemmas. Recall that c^* be a constant larger than one given in (0.0). Let $A(T)$ be an annular set $c^*T \setminus T$.

LEMMA 3.1 (Rauch and Taylor [23, Ex. 1, p. 40]).

(a) *For $p \in [1, \infty)$, we have a constant C and extension operators $E_{T,p}: W^{1,p}(A(T)) \to W^{1,p}(c^*T)$, such that*

(i) $\|\nabla(E_{T,p}v)\|_{p,T} \leqq C\|\nabla v\|_{p,A(T)}$,

(ii) $\|E_{T,p}v\|_{p,T} \leqq C(\|\nabla v\|_{p,A(T)} + \|v\|_{p,A(T)})$

for all $v \in W^{1,p}(A(T))$.

(b) *For $p = 2$, we can replace $E_{T,p}v$ with $v^\sim$ in* (i) *and* (ii) *above.*

Remark. The extension operators in Lemma 3.1(b) are used in [15]. There is another way to construct extension operators, satisfying properties in Lemma 3.1, due to Cioranescu and Saint Jean Paulin [8].

Let $r^*_\varepsilon = r_\varepsilon/\varepsilon$, $T^*_\varepsilon = r^*_\varepsilon T$ and $Y^*_\varepsilon = Y \setminus T(\varepsilon)^*$. We often write $T^*(\varepsilon) = T^*_\varepsilon$ and $Y^*(\varepsilon) = Y^*_\varepsilon$.

COROLLARY 3.2.

(a) *For $p \in [1, \infty)$, we have a constant C and an extension operator $E_{\varepsilon,p}: W^{1,p}(Y^*_\varepsilon) \to W^{1,p}(Y)$, such that*

(i) $\|\nabla(E_{\varepsilon,p}v)\|_{p,T^*(\varepsilon)} \leqq C\|\nabla v\|_{p,Y^*(\varepsilon)}$,

(ii) $\|E_{\varepsilon,p}v\|_{p,T^*(\varepsilon)} \leqq C(\|\nabla v\|_{p,Y^*(\varepsilon)} + \|v\|_{p,Y^*(\varepsilon)})$

for all $v \in W^{1,p}(Y^*_\varepsilon)$.

(b) *For $p = 2$, we can replace $E_{\varepsilon,p}v$ with $v^\sim$ in* (i) *and* (ii) *above.*

By the method of scaling together with Lemma 3.1 we see

$$\|\nabla(E_{\varepsilon,p}v)\|_{p,T^*(\varepsilon)} \leqq C\|\nabla v\|_{p,r^*_\varepsilon A(T)}$$

and

$$\|E_{\varepsilon,p}v\|_{p,T^*(\varepsilon)} \leqq C(r^*_\varepsilon\|\nabla v\|_{p,r^*_\varepsilon A(T)} + \|v\|_{p,r^*_\varepsilon A(T)}).$$

Thus, the above corollary is true. By the same way we also obtain the following corollary under (0.0) for our sequence $\{\Omega_\varepsilon\}$.

COROLLARY 3.3 (uniform extension property, [15, § 4]).

(a) *We suppose* (0.0) *holds on* $\{\Omega_\varepsilon\}$. *Then for* $p \in [1, \infty)$, *we have a constant* C *and extensions* $E_\varepsilon : \{v \in W^{1,p}(\Omega_\varepsilon);\ v|\partial\Omega = 0\} \to W_0^{1,p}(\Omega)$, *such that*

(i) $\|\nabla(E_\varepsilon v)\|_{p,T_\varepsilon} \leqq C\|\nabla v\|_{p,\Omega_\varepsilon}$, *and*

(ii) $\|E_\varepsilon v\|_{p,T_\varepsilon} \leqq C(\|\nabla v\|_{p,\Omega_\varepsilon} + \|v\|_{p,\Omega_\varepsilon})$.

(b) *For* $p = 2$ *we can replace* $E_\varepsilon v$ *with* $\tilde{v}$ *in* (i) *and* (ii) *above.*

Remark. Corollary 3.3(a) with $p = 2$ and $\varepsilon \sim r_\varepsilon$ as $\varepsilon \to 0$, is introduced by Cioranescu and Saint Jean Paulin [8] and Khruslov [19]. Their results are also extended to the case where $\varepsilon \gg r_\varepsilon$. The method of scaling for extension mappings in domains with many tiny holes has been previously applied by Rauch and Taylor [23, p. 40].

COROLLARY 3.4. *For* $p \in [1, \infty)$, *we have a constant* C *such that*

$$\|v\|_{p,\Omega_\varepsilon} \leqq C\|\nabla v\|_{p,\Omega_\varepsilon}$$

for all $v \in W^{1,p}(\Omega_\varepsilon)$ *such that* $v|\partial\Omega = 0$.

LEMMA 3.5. *For* $p \in (1, \infty)$, *we have a constant* C *such that*

$$\|v\|_{p,Y^*(\varepsilon)} \leqq C\|\nabla v\|_{p,Y^*(\varepsilon)}$$

for all $v \in W^{1,p}(Y_\varepsilon^*)$ *such that* $\int_{Y^*(\varepsilon)} v\, dy = 0$.

Proof. We suppose the contrary. Let $u_n \in W^{1,p}(Y_n^*)$ and a sequence $\varepsilon_n \downarrow 0$ as $n \uparrow \infty$, such that $\int_{Y^*(n)} u_n\, dy = 0$ and $\|u_n\|_{p,Y^*(n)} = 1 > n\|\nabla u_n\|_{p,Y^*(n)}$ for $n \in \mathbb{N}$, where $Y^*(n) = Y_\varepsilon^*$ with $\varepsilon = \varepsilon_n$. We have extensions $v_n = E_n u_n \in W^{1,p}(Y)$ satisfying (i) and (ii) in Corollary 3.2, where $E_n = E_{p,\varepsilon}$ with $\varepsilon = \varepsilon_n$. Note that $\int_Y v_n\, dy = \int_{T^*(n)} v_n\, dy = c_n$ and

$$\left|\int_Y v_n\, dy\right| \leqq C|T^*(n)|^{1/p^*}\|v_n\|_{p,Y^*(n)} \to 0 \quad \text{as } \varepsilon \to 0.$$

For $v_{n,0} = v_n - c_n$, we have $\int_Y v_{n,0}\, dy = 0$, $\|v_{n,0}\|_{p,Y} \to 1$ and $\|\nabla v_{n,0}\|_{p,Y} \to 0$ as $n \to \infty$. There exists a subsequence, still denoted by $v_{n,0}$, such that $v_{n,0} \to v_0$ in $L^p(Y)$. Then $v_0 \equiv 0$ almost everywhere in Y and this contradicts to $\|v_0\|_{p,Y} = 1$. □

The next lemma appeared as Lemma 6.1 of Conca [9] in the case where $r_\varepsilon = \varepsilon$. It is proved by the same spirit as in [9].

LEMMA 3.6. *Let* $p \in [1, N]$. *We have a constant* C *such that*

$$\|v\|_{p,\Omega_\varepsilon} \leqq C\{\theta_{N-p,\varepsilon}^{-1/p}\|\nabla v\|_{p,\Omega_\varepsilon} + \theta_{N-1,\varepsilon}^{-1/p}\|v\|_{p,\partial T_\varepsilon}\}$$

for all $v \in V_\varepsilon$.

LEMMA 3.7 (Conca and Donato [10, Lemma 2.1]). *For* $p \in [1, \infty)$ *we have a constant* C *such that*

$$\|v\|_{p,\partial T_\varepsilon} \leqq C\{c(r_\varepsilon, \varepsilon)^{1/p}\|\nabla v\|_{p,\Omega_\varepsilon} + \theta_{N-1,\varepsilon}^{1/p}\|v\|_{p,\Omega_\varepsilon}\}.$$

Here $c(r_\varepsilon, \varepsilon) = r_\varepsilon^{N-1}\varepsilon^{p-N}$ *for* $p > N$, $c(r_\varepsilon, \varepsilon) = c_1(r_\varepsilon)$ *for* $p \leqq N$,

$$c_1(r_\varepsilon) = \frac{\theta_{N-1,\varepsilon}}{\theta_{N-p,\varepsilon}} = \begin{cases} r_\varepsilon^{p-1} & \text{for } 1 \leqq p < N, \\ (r_\varepsilon \log r_\varepsilon^{-1})^{N-1} & \text{for } p = N. \end{cases}$$

We consider test functions introduced first in the case where $r_\varepsilon \sim \varepsilon$ with $p=2$, by Vanninathan [25]. Let $\nu=\{\nu_i; 1\leqq i\leqq N\}$ be the unit outer normal to $\partial(Y\setminus T_\varepsilon^*)$ and let

$$Av=-\sum_{i=1}^{N}\partial\left(\left|\frac{\partial v}{\partial y_i}\right|^{p-2}\frac{\partial v}{\partial y_i}\right)\Big/\partial y_i\in L^{p^*}(Y_\varepsilon^*),$$

$$Bv=\sum_{i=1}^{N}\nu_i\left|\frac{\partial v}{\partial y_i}\right|^{p-2}\frac{\partial v}{\partial y_i}\in L^{p^*}(\partial T_\varepsilon^*)$$

for $v\in W^{1,p}(Y_\varepsilon^*)$ (note that the mapping $v\in L^p(Y_\varepsilon^*)\to|v|^{p-2}v\in L^{p^*}(Y_\varepsilon^*)$, is the singly valued duality mapping). We consider (Q_ε^*):

$$A\Phi_\varepsilon=-C(\varepsilon),\quad \text{a constant, in } Y_\varepsilon^*,$$

$$(B\Phi_\varepsilon)(y)=g(y/r_\varepsilon^*)\quad \text{for } y\in\partial T_\varepsilon^*,$$

$$\int_{Y^*(\varepsilon)}\Phi_\varepsilon\,dy=0,$$

$$\Phi_\varepsilon\quad \text{is } Y \text{ periodic},$$

where $C(\varepsilon)$ satisfies the compatibility condition, i.e., $C(\varepsilon)=(r_\varepsilon^*)^{N-1}|Y_\varepsilon^*|^{-1}\int_{\partial T}g\,d\sigma$. Let $X=\{v\in W^{1,p}(Y_\varepsilon^*);\ v \text{ is } Y \text{ periodic}\}$. We consider (Q_ε^*) in a general form.

We consider the following problem. For $h_1\in L^{p^*}(Y_\varepsilon^*)$ and $h_2\in L^{p^*}(\partial T_\varepsilon^*)$, find $v\in X/\mathbb{R}$ such that

$$Av=h_1\quad \text{in } Y_\varepsilon^*\quad \text{and}\quad Bv=h_2\quad \text{on } \partial T_\varepsilon^*. \tag{3.1}$$

When there exists a solution v, we get a compatibility condition

$$\int_{Y_\varepsilon^*}h_1\,dy+\int_{\partial T_\varepsilon^*}h_2\,d\sigma=0. \tag{3.2}$$

We show that there exists a unique solution $v\in X/\mathbb{R}$ of (3.1) under (3.2).

Let X^* be the dual space of X, which is reflexive. Generally, for a fixed reflexive (respectively, uniformly convex) normed space E, we can say that any closed subspace of E, or any quotient space modulo a closed subspace is also reflexive (respectively, uniformly convex). Let $F=\{l\in X^*;\langle l,1\rangle=0\}$. We have $\mathbb{R}=\{v\in X;\langle l,v\rangle=0 \text{ for all } l\in F\}$ and $\{X/\mathbb{R},F\}$ as a dual pairing. Let φ be a lower semicontinuous convex function, $X/\mathbb{R}\to[0,\infty]$, defined by

$$\varphi(v)=p^{-1}\int_{Y_\varepsilon^*}|\nabla v|^p\,dy.$$

Since $\|\nabla v\|_{p,Y_\varepsilon^*}$ is a norm on $X/\mathbb{R}$, we have $\varphi(v)/\|\nabla v\|_{p,Y_\varepsilon^*}\to\infty$ as $\|\nabla v\|_{p,Y_\varepsilon^*}\to\infty$. We have $A(=\partial\varphi):X/\mathbb{R}\to F$ with its range, $R(A)=F$ and have the bounded inverse of A, A^{-1} (see, for example, Barbu [2, Prop. 2.6, p. 56]). Generally, if we denote by $|\cdot|$ a norm on a reflexive space E with its dual E^*, we have the duality mapping $E\ni x\to G(x)=\partial(|\cdot|^p/p)(x)\in E^*$, $1<p<\infty$. When E^* is strictly convex, G is single-valued and demicontinuous (see [2, pp. 13, 53]). In our case $X/\mathbb{R}$ is isometrically isomorphic to a closed subspace Z of $L^p(Y_\varepsilon^*)^N$ and the dual space F is isometrically isomorphic to the space $L^q(Y_\varepsilon^*)^N/Z^\perp$, where $Z^\perp$ is the polar set of Z in $L^q(Y_\varepsilon^*)^N$. Thus, F is uniformly convex, so, strictly convex. We see that A and A^{-1} are single-valued.

The operator A with $2\leqq p<\infty$, is frequently considered, while A with general p, $1<p<\infty$, and the boundary operator B is not so often seen. Here, for this point we cite only Proposition 4.1 of Lions [21, p. 205].

Therefore, we have a solution Φ_ε of (Q_ε^*), which satisfies

$$\int_{Y^*(\varepsilon)} \sum_{i=1}^{N} \left|\frac{\partial \Phi_\varepsilon}{\partial y_i}\right|^{p-2} \frac{\partial \Phi_\varepsilon}{\partial y_i} \frac{\partial v}{\partial y_i}\, dy = -C(\varepsilon) \int_{Y^*(\varepsilon)} v\, dy + \int_{\partial T^*(\varepsilon)} g\left(\frac{y}{r_\varepsilon^*}\right) v(y)\, d\sigma(y) \tag{3.3}$$

for all $v \in X$.

Let

$$\tilde{\kappa}_{\varepsilon,p} = \begin{cases} \kappa_{\varepsilon,p} & \text{for } p \in [1, N], \\ (r_\varepsilon^*)^{N-1} & \text{for } p \in (N, \infty). \end{cases}$$

LEMMA 3.8. *Let $I_g = \int_{\partial T} g\, d\sigma$ and $g \in L^{p^*}(\partial T)$ with $1 < p < \infty$.*

(a) *We have a constant C such that, for all $v \in W^{1,p}(Y_\varepsilon^*)$,*

$$\left| \int_{Y^*} \sum_{i=1}^{N} \left|\frac{\partial \Phi_\varepsilon}{\partial Y_i}\right|^{p-2} \frac{\partial \Phi_\varepsilon}{\partial y_i} \frac{\partial v}{\partial y_i}\, dy \right| \leqq C(|I_g| + \|g\|_{p^*,\partial T}) \tilde{\kappa}_{\varepsilon,p} \|\nabla v\|_{p,Y^*}.$$

(b) *Let $\gamma_\varepsilon = \{\gamma_\varepsilon^i;\ 1 \leqq i \leqq N\}$ be a vector function defined by*

$$\gamma_\varepsilon^i(x) = \left(\left|\frac{\partial \Phi_\varepsilon}{\partial y_i}\right|^{p-2} \frac{\partial \Phi_\varepsilon}{\partial y_i} \right)(y)$$

where $y = (x - p_\varepsilon^i)/\varepsilon$ and $x \in p_\varepsilon^i + (\varepsilon Y \setminus r_\varepsilon T)$ for all $i \in \mathbb{N}$. Then, for $v \in W_0^{1,p}(\Omega_\varepsilon)$, we have

$$\left| \int_{\Omega_\varepsilon} \gamma_\varepsilon \nabla v\, dx \right| \leqq C(|I_g| + \|g\|_{p^*,\partial T}) \tilde{\kappa}_{\varepsilon,p} \|\nabla v\|_{p,\Omega_\varepsilon}.$$

To show this we need the lemma below, which is proved as in the proof of Lemma 3.7.

LEMMA 3.9. *For $q \in [1, \infty)$ we have a constant C such that*

$$\|w\|_{q,\partial T^*(\varepsilon)}^q \leqq C\{\tilde{c}(r_\varepsilon^*) \|\nabla w\|_{q,Y^*(\varepsilon)}^q + (r_\varepsilon^*)^{N-1} \|w\|_{q,Y^*(\varepsilon)}^q\}$$

for all $w \in W^{1,q}(Y_\varepsilon^*)$, where $\tilde{c}(r_\varepsilon^*) = c(r_\varepsilon^*, 1)$, $c(\cdot,\cdot)$ is as in Lemma 3.7.

Proof of Lemma 3.8(a). We suppose $v \in X$. By (3.3), Lemmas 3.5 and 3.9 with the Hölder inequality applied to $\int_{Y^*} v\, dy$ and $\int_{\partial T^*(\varepsilon)} g(y/r_\varepsilon^*) v(y)\, d\sigma(y)$, we get

$$\begin{aligned} &\left| \int_{Y^*} \sum_{i=1}^{N} \left|\frac{\partial \Phi_\varepsilon}{\partial y_i}\right|^{p-2} \frac{\partial \Phi_\xi}{\partial y_i} \frac{\partial v}{\partial y_i}\, dy \right| \\ &\quad \leqq C \|\nabla v\|_{p,Y^*(\varepsilon)} [|C(\varepsilon)| + \|g\|_{p^*,\partial T} \{(r_\varepsilon^*)^{N-1} + (r_\varepsilon^*)^{(N-1)/p^*} \tilde{c}(r_\varepsilon^*)^{1/p}\}] \\ &\quad \leqq C(|I_g| + \|g\|_{p^*}) \tilde{\kappa}_{\varepsilon,p} \|\nabla v\|_{p,Y^*}. \end{aligned}$$

Substituting $v = \Phi_\varepsilon$ into the above inequality, we have

$$\|\nabla \Phi_\varepsilon\|_{p,Y^*(\varepsilon)}^{p-1} \leqq C(|I_g| + \|g\|_{p^*}) \tilde{\kappa}_{\varepsilon,p}. \tag{3.4}$$

For $v \in W^{1,p}(Y_\varepsilon^*)$, the inequality in (a) also stands. □

Proof of Lemma 3.8(b). *This directly follows* (3.4). □

Let $\mathcal{T}_\varepsilon = \cup \{T_\varepsilon^i;\ i \in \mathbb{N}\}$ and $\mathcal{T}_\varepsilon^c = \mathbb{R}^N \setminus \mathcal{T}_\varepsilon$. Let $g_\varepsilon \delta_\varepsilon$ be a signed measure defined by

$$\langle g_\varepsilon \delta_\varepsilon, \zeta \rangle = \int_{\partial \mathcal{T}_\varepsilon} g_\varepsilon \zeta\, d\sigma \quad \text{for } \zeta \in C_0(\mathbb{R}^N).$$

Using Φ_ε we introduce test functions $\varphi_\varepsilon \in W^{1,p}_{\mathrm{loc}}(\mathcal{T}_\varepsilon)$ and $\gamma_\varepsilon \in L^{p^*}_{\mathrm{loc}}(\mathcal{T}_\varepsilon)^N$ by $\varphi_\varepsilon(x) = \Phi_\varepsilon((x-p^i_\varepsilon)/\varepsilon)$, $\gamma_\varepsilon(x) = (\nabla\varphi_\varepsilon)((x-p^i_\varepsilon)/\varepsilon)$ for all $x \in p^i_\varepsilon + (\varepsilon Y \setminus r_\varepsilon T)$, $i \in \mathbb{N}$. We get

$$\varepsilon^p A\varphi_\varepsilon = -C(\varepsilon) \quad \text{in } \mathcal{T}_\varepsilon,$$

$$\varepsilon^{p-1} B\varphi_\varepsilon = g_\varepsilon(x) \quad \text{on } \partial\mathcal{T}_\varepsilon,$$

$$\varphi_\varepsilon \quad \text{is } \varepsilon Y \text{ periodic, and}$$

$$\langle g_\varepsilon \delta_\varepsilon, v\rangle = \theta_{N-1,\varepsilon} I_g \int_{\mathcal{T}^c_\varepsilon} v\,dx/|Y^*_\varepsilon| + \int_{\mathcal{T}^c_\varepsilon} \gamma_\varepsilon \nabla v\,dx \tag{3.5}$$

for all $v \in W^{1,p}(\mathbb{R}^N)$ such that supp v is compact (for (3.5) with $p=2$ and $\varepsilon \sim r_\varepsilon$ see (11.6) of Vanninathan [25]).

LEMMA 3.10. Let $p \in (1, \infty)$.

(a) *In the case where $p > N$ we suppose only that $\varepsilon \to 0$. In the case where $p \in [1, N]$, we suppose that $\theta_{N-p} = \infty$, as $\varepsilon \to 0$. Then we have*

$$g_\varepsilon \delta_\varepsilon / \theta_{N-1,\varepsilon} \xrightarrow{s} I_g\,dx \quad \text{in } W^{-1,p^*}(\Omega) \quad \text{as } \varepsilon \to 0. \tag{3.6}$$

(b) *If $\alpha_\varepsilon \tilde{\kappa_\varepsilon} \to 0$ and $a < \infty$ as $\varepsilon \to 0$, then we have*

$$\alpha_\varepsilon \delta_\varepsilon \to a\,dx \quad \text{as } \varepsilon \to 0 \quad \text{in } W^{-1,p^*}(\Omega). \tag{3.7}$$

(c) *Let $\langle g_\varepsilon \bar{\delta}_\varepsilon, \zeta\rangle = \int_{\partial T_\varepsilon} g_\varepsilon \zeta\,d\sigma$ for $\zeta \in C^\infty_0(\Omega)$. Then we have all the results in* (a), (b), *even if δ_ε is replaced with $\bar{\delta}_\varepsilon$.*

Assertions (a) and (b) follow from (3.5) with Lemma 3.8(b), while (c) follows from the estimate $|\Omega \cap (\mathcal{T}_\varepsilon \setminus T_\varepsilon)| \leqq C|\partial\Omega|\varepsilon$.

It is a generalization of Theorem 2.2 of [15] from a case where $g \equiv 1$ to $g \in L^{p^*}(\partial T)$. Another generalization is considered in [16].

4. Proofs of theorems. Let $I_g = \int_{\partial T} g\,d\sigma$, $\theta_\varepsilon = \theta_{N-1,\varepsilon}$ and $\theta = \theta_{N-1}$. Let $B(\xi) = \int_0^\xi \beta^0(\xi')\,d\xi'$ for $\xi \in D(\beta)^-$ and $B(\xi) = \infty$ for $\xi \in \mathbb{R} \setminus D(\beta)^-$ (for β^0 see § 1; β^0 is single-valued, monotonously increasing in $D(\beta)$). Let $m_\beta = \inf\{\xi \in D(\beta)\}$ and $M_\beta = \sup\{\xi \in D(\beta)\}$. We can define B^ε, m^ε_β, M^ε_β, and B^0, m^0_β, M^0_β, in the same way as B, m_β, and M_β, using β^ε, β^0, instead of using β in the integrand above or the region in which the infimum and the supremum are taken. We see $\theta_\varepsilon m^\varepsilon_\beta = m_\beta$, $\theta_\varepsilon M^\varepsilon_\beta = M_\beta$, and $B^\varepsilon(\xi) = B(\theta_\varepsilon\xi)/\theta^{h+1}_\varepsilon$ (see §§ 1.3).

Let $K^\varepsilon = \{v \in V_\varepsilon;\ B(v)|\partial T_\varepsilon \in L^1(\partial T_\varepsilon)\}$. The weak solution $u_\varepsilon \in K^\varepsilon$ of (P_ε), satisfies the next inequality:

$$\int_{\Omega_\varepsilon} [\nabla u_\varepsilon \nabla(v-u_\varepsilon) - (v-u_\varepsilon)f]\,dx - \langle g_\varepsilon \bar{\delta}_\varepsilon, v-u_\varepsilon\rangle + \alpha_\varepsilon\langle \bar{\delta}_\varepsilon, B(v) - B(u_\varepsilon)\rangle \geqq 0 \quad \text{for all } v \in K^\varepsilon. \tag{4.1}$$

4.1. Estimate. Theorems A, B, C, follow from (4.1) combining with estimates in § 3.

(A) *Proof of Theorem* A. Let $\kappa_\varepsilon = \kappa_{\varepsilon,2}$. After substituting $v = 0$ into (4.1) and applying (3.5) and Lemma 3.8(b) we get

$$\|\nabla u_\varepsilon\|^2_{\Omega_\varepsilon} + \alpha_\varepsilon\langle \bar{\delta}_\varepsilon, B(u_\varepsilon)\rangle \leqq \|f\|_\Omega \|u_\varepsilon\|_{\Omega_\varepsilon} + C\theta_\varepsilon |I_g|\,\|u_\varepsilon\|_{\Omega_\varepsilon} + (|I_g| + \|g\|_{\partial T})\kappa_\varepsilon \|\nabla u_\varepsilon\|_{\Omega_\varepsilon}. \tag{4.2}$$

Corollary 3.4 implies that

$$\|\nabla u_\varepsilon\|_{\Omega_\varepsilon} \leqq C(\|f\|_\Omega + \theta_\varepsilon |I_g| + \kappa_\varepsilon |I_g| + \kappa_\varepsilon \|g\|_{\partial T})$$

and

$$(4.2)_b \qquad \alpha_\varepsilon \langle \delta_\varepsilon, B(u_\varepsilon)\rangle \leqq C(\|f\|_\Omega + \theta_\varepsilon |I_g| + \kappa_\varepsilon |I_g| + \kappa_\varepsilon \|g\|_{\partial T})^2$$

(B) *Proof of Theorem* B. By Hölder's inequality we get

$$(4.3)_a \qquad \|u_\varepsilon\|^2_{\sigma',\Omega_\varepsilon} \leqq C \|u_\varepsilon\|^2_{\Omega_\varepsilon}$$

and

$$(4.3)_b \qquad \theta_\varepsilon^{-(\sigma-\sigma')/\sigma'} \|u_\varepsilon\|^\sigma_{\sigma',\partial T_\varepsilon} \leqq C \|u_\varepsilon\|^\sigma_{\sigma,\partial T_\varepsilon},$$

where $u_\varepsilon \in L^\sigma(\partial T_\varepsilon)$. By (4.1) with $v=0$, together $(4.3)_a$, (β-3) with Lemma 3.6 we get

$$\begin{aligned}\|\nabla u_\varepsilon\|^2_{\sigma',\Omega_\varepsilon} + \alpha_\varepsilon \|u_\varepsilon\|^\sigma_{\sigma,\partial T_\varepsilon} \leqq C(&\|f\|_{\sigma'^*,\Omega_\varepsilon} \|\nabla u_\varepsilon\|_{\sigma',\Omega_\varepsilon} / \theta^{1/\sigma'}_{N-\sigma',\varepsilon} \\ &+ \|f\|_{\sigma'^*,\Omega_\varepsilon} \|u_\varepsilon\|_{\sigma',\partial T_\varepsilon} / \theta_\varepsilon^{1/\sigma'} \\ &+ \theta_\varepsilon^{1/\sigma^*} \|g\|_{\sigma^*,\partial T} \|u_\varepsilon\|_{\sigma,\partial T_\varepsilon}).\end{aligned}$$

Young's inequality with $(4.3)_b$ implies

$$(4.4) \qquad \begin{aligned}&\|\nabla u_\varepsilon\|^2_{\sigma',\Omega_\varepsilon} + \alpha_\varepsilon \|u_\varepsilon\|^\sigma_{\sigma,\partial T_\varepsilon} \\ &\leqq C(\|f\|^2_{\sigma'^*,\Omega} / \theta^{2/\sigma'}_{N-\sigma',\varepsilon} + \|f\|^{\sigma^*}_{\sigma'^*,\Omega} / a_\varepsilon^{\sigma^*/\sigma} + \|g\|^{\sigma^*}_{\sigma^*,\partial T} \theta_\varepsilon / \alpha_\varepsilon^{\sigma^*/\sigma}).\end{aligned}$$

By Corollary 3.3 we have $u_\varepsilon^\sim \xrightarrow{s} 0$ in $W_0^{1,\sigma'}(\Omega)$ as $\varepsilon \to 0$. □

(C) *Proof of Theorem* C(I). This directly follows Theorem A with $\theta = \infty$. □

(D) *Proof of Theorem* C(II). Note that $\beta = \partial B$. Let γ_λ be the Yosida approximation of $(\partial B)_\lambda$ and let v_λ be the solution of

$$-\Delta v_\lambda = f \quad \text{in } \Omega_\varepsilon,$$

$$\frac{\partial v_\lambda}{\partial \nu} + \gamma_\lambda(v_\lambda) = g_\varepsilon \quad \text{on } \partial T_\varepsilon \quad \text{and} \quad v_\lambda = 0 \quad \text{on } \partial\Omega.$$

By a slight modification of the proof of Théorème I.10 of Brezis [4], to a nonautonomous boundary condition on ∂T_ε, we have $v_\lambda \xrightarrow{W} u_\varepsilon$ in $H^2(\Omega_\varepsilon)$ as $\lambda \to +0$. Thus, we can extract $\{v_\varepsilon\}$ such that $\|\nabla(v_\varepsilon - u_\varepsilon)\|_{\Omega_\varepsilon} + \|v_\varepsilon - u_\varepsilon\|_{\Omega_\varepsilon} \to 0$ as $\varepsilon \to 0$, where $v_\varepsilon = v_\lambda$ with $\lambda = \lambda_\varepsilon$. It suffices to show that $\|\nabla v_\varepsilon\|_{\Omega_\varepsilon} \geqq C\theta_\varepsilon$. We get

$$\int_{\Omega_\varepsilon} [\nabla v_\varepsilon \nabla \zeta - f\zeta]\, dx + \alpha_\varepsilon \langle \bar{\delta}_\varepsilon, \gamma_\varepsilon(v_\varepsilon)\zeta\rangle = \langle g_\varepsilon \bar{\delta}_\varepsilon, \zeta\rangle \quad \text{for } \zeta \in C_0^\infty(\Omega),$$

where $\gamma_\varepsilon = \gamma_\lambda$ with $\lambda = \lambda_\varepsilon$. For ζ, $0 \leqq \zeta \leqq 1$, and $\zeta \equiv 1$ on a nonempty open set $G \Subset \Omega$, we have C such that $\int_{\Omega_\varepsilon} \zeta\, dx \geqq C > 0$. Property (a) in § 2 implies that $|\gamma_\varepsilon(\xi)| \leqq |\beta^0(\xi)|$ for all $\xi \in \mathbb{R}$. Thus, (β-2) with $0 \leqq r \leqq 1$, $\beta_m = -\infty$, and $\beta_M = \infty$, give an estimate,

$$|\alpha_\varepsilon \langle \bar{\delta}_\varepsilon, \gamma_\varepsilon(v_\varepsilon)\zeta\rangle| \leqq C_1 \alpha_\varepsilon \langle \bar{\delta}_\varepsilon, |v_\varepsilon| + 1\rangle.$$

Applying (3.5) to $\alpha_\varepsilon \bar{\delta}_\varepsilon$ and $g_\varepsilon \bar{\delta}_\varepsilon$ together with Lemma 3.10(b), (c), we get

$$\theta_\varepsilon |I_g| \leqq C_2(\|\nabla v_\varepsilon\|_{\Omega_\varepsilon} + a_\varepsilon \|v_\varepsilon\|_{\Omega_\varepsilon} + \alpha_\varepsilon \kappa_\varepsilon \|\nabla v_\varepsilon\|_{\Omega_\varepsilon} + \|f\|_\Omega + \alpha_\varepsilon S(\partial T_\varepsilon) + \kappa_\varepsilon).$$

But $\alpha_\varepsilon S(\partial T_\varepsilon) \leqq a_\varepsilon |\partial T|$ and $\lim_{\varepsilon \to 0} \alpha_\varepsilon \kappa_\varepsilon = 0$. Using Corollary 3.4, we then have

$$\theta_\varepsilon |I_g| \leqq C_3 \|\nabla v_\varepsilon\|_{\Omega_\varepsilon} + C_4.$$

But $\theta = \infty$ and so we get $\theta_\varepsilon |I_g| \leqq C_5 \|\nabla v_\varepsilon\|_{\Omega_\varepsilon}$. □

(E) *Proof of Theorem* C(III). The estimate (4.4) with $\theta_{N-\sigma'}=\infty=a$, implies (1.4) and $\|u_\varepsilon\|^{\sigma}_{\sigma,\partial T_\varepsilon}\leqq Cb_\varepsilon/\alpha_\varepsilon$. By $(4.3)_b$ we get $\|u_\varepsilon\|^{\sigma'}_{\sigma',\partial T_\varepsilon}\leqq C\|u_\varepsilon\|^{\sigma'}_{\sigma,\partial T_\varepsilon}\theta_\varepsilon^{(\sigma-\sigma')/\sigma}$. Lemma 3.6 implies

$$\begin{aligned}\|u_\varepsilon\|_{\sigma',\Omega_\varepsilon}&\leqq C_1\|\nabla u_\varepsilon\|_{\sigma',\Omega_\varepsilon}/\theta^{1/\sigma'}_{N-\sigma',\varepsilon}+C_2\|u_\varepsilon\|_{\sigma',\partial T_\varepsilon}/\theta^{1/\sigma'}_\varepsilon\\&\leqq C_1 b_\varepsilon^{1/2}\theta^{-1/\sigma'}_{N-\sigma',\varepsilon}+C_2(b_\varepsilon/\alpha_\varepsilon)^{1/\sigma}\theta_\varepsilon^{(\sigma-\sigma')/\sigma\sigma'-1/\sigma'}\\&\leqq C_3 b_\varepsilon^{1/\sigma'}(\theta^{-1/\sigma'}_{N-\sigma',\varepsilon}+a_\varepsilon^{-1/\sigma}).\end{aligned}$$

Thus we get (1.5). The estimates (1.4), (1.5), with Corollary 3.3 imply $\tilde{u_\varepsilon}/b_\varepsilon^{1/\sigma'}\xrightarrow{w}0$ in $W_0^{1,\sigma'}(\Omega)$ as $\varepsilon\to 0$. □

4.2. Convergence. Theorems from §§ 1.2–1.4 follow Theorem 2.1 combining with Propositions 2.2, 2.4 and Lemma 3.10.

Problem (P_ε) is rewritten with a subdifferential. Let $H=L^2(\Omega)$. We then use the terminology in Example 2.3. Let $\varphi^\varepsilon=\varphi^n$ and $\mathbf{j}^\varepsilon=\mathbf{k}^\varepsilon+\mathbf{m}^\varepsilon$, $\varepsilon=\varepsilon_n$, where $\mathbf{k}^\varepsilon(\xi)=-\langle g_\varepsilon\bar{\delta}_\varepsilon,\xi\rangle$ for $\xi\in L^2(\partial T_\varepsilon)$ and

$$m^\varepsilon(\xi)=\begin{cases}\alpha_\varepsilon\langle\bar{\delta}_\varepsilon, B(\xi)\rangle & \text{for } \xi\in L^2(\partial T_\varepsilon) \text{ and } B(\xi)\in L^1(\partial T_\varepsilon),\\ \infty & \text{for others } \xi \text{ in } L^2(\partial T_\varepsilon).\end{cases}$$

We see that $D(\partial\mathbf{j}^\varepsilon)=\{\xi\in L^2(\partial T_\varepsilon)$; there exists $\xi'\in L^2(\partial T_\varepsilon)$ such that $\xi'(x)\in\alpha_\varepsilon\beta(\xi(x))-g_\varepsilon(x)$ almost everywhere on $\partial T_\varepsilon\}$. Let $\Phi^\varepsilon: H\to[0,\infty]$ be

$$\Phi^\varepsilon(v)=\begin{cases}\varphi^\varepsilon(v)+\mathbf{j}^\varepsilon(v|\partial T_\varepsilon) & \text{for } v\in K^\varepsilon,\\ \infty & \text{otherwise.}\end{cases}$$

We set $\tilde{\mathbf{j}}^\varepsilon=\mathbf{j}^\varepsilon\,|\,H^{1/2}(\partial T_\varepsilon)$ as in Brezis [4, p. 32]. A basic fact below follows the argument in I.2.1, Théorème I.10 of [4, p. 33] and Theorem 12 of [5, p. 112].

PROPOSITION 4.1. (I) *For $f\in L^2(\Omega_\varepsilon)$ and $g\in L^2(\partial T_\varepsilon)$, there exists a unique solution u_ε of* (4.1). *This u_ε is also the solution of the problem*:

$$-\Delta u_\varepsilon=f \quad \text{in } \mathcal{D}(\Omega)',$$

$$\frac{\partial u_\varepsilon}{\partial\nu}+\partial\tilde{\mathbf{j}}^\varepsilon(u_\varepsilon)\ni 0 \quad \text{in } H^{-1/2}(\partial T_\varepsilon),$$

$$u_\varepsilon=0 \quad \text{on } \partial\Omega,$$

where $\mathcal{D}(\Omega)'$ is the space of distributions.

(II) *In addition to* (I), *if we suppose $g\in H^{1/2}(\partial T)$, then $u_\varepsilon\in H^2(\Omega_\varepsilon)\cap V_\varepsilon$, and u_ε satisfies* (0.2).

(III) *If $g\in H^{1/2}(\partial T)$, then we have that $D(\partial\Phi^\varepsilon)=\{v\in H^2(\Omega_\varepsilon)\cap V_\varepsilon$; v satisfies* (0.2)*\} and $\partial\Phi^\varepsilon(v)=-\Delta v$ for $v\in D(\partial\Phi^\varepsilon)$.*

The condition $g\in H^{1/2}(\partial T)$ in the Introduction implies

$$\partial\Phi^\varepsilon(u_\varepsilon)\ni f. \tag{4.5}$$

As seen in (2.1), we have

$$\varphi^\varepsilon\xrightarrow{(e)}\varphi^0, \tag{4.6}$$

if $\varepsilon\to 0$ with a fixed value θ_N. When $\{\Phi^\varepsilon\}$ and a certain Φ^0 satisfy the conditions in Proposition 2.4 and $\{(\partial\Phi^\varepsilon)^{-1}\}$ is uniformly bounded, Proposition 2.2(b) is applied to $\{\Phi^\varepsilon\}$.

LEMMA 4.2. (a) *Let* $F^\varepsilon : v \in V_\varepsilon \to F^\varepsilon v = \max\{\min\{v, c_\varepsilon\}, C_\varepsilon\}$, c_ε, $C_\varepsilon \in \mathbb{R}$. *Then* $\varphi^\varepsilon(F^\varepsilon v) \leqq \varphi^\varepsilon(v)$ *and* $(F^\varepsilon v_\varepsilon)^\sim \xrightarrow{w} v$ *for* $v_\varepsilon^\sim \xrightarrow{w} v$ *in* $H_0^1(\Omega)$, $c_0 \leqq v \leqq C_0$ *provided that* $c_\varepsilon \to c_0$ *and* $C_\varepsilon \to C_0$ *as* $\varepsilon \to 0$.

(b) *We suppose* (β-1) *and* (β-2) *with* (1.10). *Then,* $B(v_\varepsilon) \to B(v_0)$ *weakly in* $W_0^{1,\rho}(\Omega)$ *as* $\varepsilon \to 0$, *for any* $v_\varepsilon \xrightarrow{w} v_0$ *in* $H_0^1(\Omega)$ *as* $\varepsilon \to 0$ *such that* $\beta_m \leqq v_\varepsilon(x) \leqq \beta_M$ *almost everywhere in* Ω, *where for* ρ *given in* (1.1).

Proof. We can easily see (a). The boundedness of $\{B(v_\varepsilon)\}$ in $W_0^{1,\rho}(\Omega)$ follows the growth condition of β together with the Hölder inequality. By this argument we get $\rho = 2q(q+2r)^{-1}$, where q is taken as the injection, $H_0^1(\Omega) \to L^q(\Omega)$, is bounded. The continuity of B in (b) follows the growth condition on $B: |B(\xi)| \leqq C(1+|\xi|^{r+1})$, combining with the compact imbedding, $H_0^1(\Omega) \to L^{q'}(\Omega)$, $1 \leqq q' < q = 2N/(N-2)$ for $N \geqq 3$ and $1 \leqq q' < \infty$ for $N = 2$. The mapping, $B: L^{q'}(\Omega) \to L^\rho(\Omega)$, is continuous, if $r+1 \leqq q'/\rho < r + q'/2$, which implies $q \geqq 2$ for $N = 2$ (from here we get $c(r) \leqq \rho$). Thus, we have (1.1) for ρ (see the proof of [15, Lemma B.1].

(A) *Proofs of Theorems* a *and* b. Before giving the proofs we consider the relation between the value r and the value $\alpha_\varepsilon \kappa_{\varepsilon,\rho}$, which plays a key role in Lemma 3.10(b). The definition of ρ implies $\rho^* > 1$. In Theorem a we see that $r \leqq 1$ if and only if $\rho^* \leqq N$ for $N \geqq 3$. For $N = 2$, we get $r = 0$ if and only if $\rho^* = N(=2)$. Using $\bar{\alpha}_\varepsilon$, $\tilde{\alpha}_\varepsilon$, and $\hat{\alpha}_\varepsilon$, we have $(a)_1$ $\alpha_\varepsilon \kappa_{\varepsilon,\rho} = \bar{\alpha}_\varepsilon \theta_{N-\rho^*,\varepsilon}^{1/\rho^*}$ for $N \geqq 3$ and $0 \leqq r < 1$, $(a)_2$ $\alpha_\varepsilon \kappa_{\varepsilon,\rho} \sim \tilde{\alpha}_\varepsilon \theta_{0,\varepsilon}^{1/N}$ as $\varepsilon \to 0$, for $N \geqq 3$ and $r = 1$, $(a)_3$ $\alpha_\varepsilon \kappa_{\varepsilon,\rho} \sim \hat{\alpha}_\varepsilon \theta_{0,\varepsilon}^{1/2}$ as $\varepsilon \to 0$, for $N = 2$ and $r = 0$. These relations are used for Theorem a. In Theorem b we again consider the relation between r and the value $\alpha_\varepsilon \kappa_{\varepsilon,\rho}$. For the case $N \geqq 3$ and $r \geqq 0$, we always have $\rho < N$ and note that $1 \leqq \rho$ if and only if $N/(N-2) \geqq r$. For the case $N = 2$ and $r \geqq 0$ we have $2 \geqq \rho \geqq 1$. The relation, (b) $\alpha_\varepsilon \kappa_{\varepsilon,\rho} = a_\varepsilon \theta_{N-\rho}^{-1/\rho}$, holds and is useful in Theorem b. The relations $(a)_1$, $(a)_2$, and $(a)_3$ work well for the case where $r_\varepsilon \leqq C r_\varepsilon^{(N-2)}$ and we get Theorem a, while the relation (b) is applied to the case $r_\varepsilon \gg r_\varepsilon^{(N-2)}$ and we get Theoerem b. In more detail, we describe the proofs as follows.

Proof of Theorem A. Since $r_\varepsilon \leqq C r_\varepsilon^{(N-\rho^*)} \ll r_\varepsilon^{(N-1)}$, Theorem A, and Corollary 3.4 hold true, $\{(\partial\Phi^\varepsilon)^{-1}\}$ is uniformly bounded. By Proposition 2.2(b) it suffices to show that

$$(4.6)_1 \qquad \Phi^\varepsilon \xrightarrow{(e)} \varphi^0.$$

To show this we take F^n in Proposition 2.4 as F^ε in Lemma 4.2. Besides for $\varepsilon = \varepsilon_n$ we put $j^n(v) = \mathbf{j}^\varepsilon(v|\partial T_\varepsilon)$ for $v \in K^\varepsilon$. The condition $\alpha_\varepsilon \kappa_\varepsilon \to 0$ in Lemma 3.10(b) follows any one of (1.7), (1.8), or (1.9). We get $a = 0 < \infty$ from $N - \rho^* \leqq N - 2$ and $a_\varepsilon = \bar{\alpha}_\varepsilon \theta_{N-2,\varepsilon}$. The assertion $(4.6)_1$ follows (4.6), Propositions 2.2(b) and 2.4, Lemma 3.10(b), (c), and Lemma 4.2.

We prove Theorem b. For $a < \infty$ let $\Phi^0 = \varphi^0 + j^0$ and $j^0 = k^0 + m^0$, where the functionals k^0, $m^0 : H \to [0, \infty]$, are defined by

$$k^0(v) = -\theta I_g \int_\Omega v \, dx \quad \text{for } v \in L^2(\Omega)$$

and

$$m^0(v) = \begin{cases} aS(\partial T) \displaystyle\int_\Omega B(v)\, dx & \text{for } v \in L^2(\Omega) \text{ and } B(v) \in L^1(\Omega), \\ \infty & \text{otherwise.} \end{cases}$$

The uniform boundedness of $\{(\partial\Phi^\varepsilon)^{-1}\}$ follows the same reason as in the previous argument. As in the proof of Theorem a, the convergence (4.6), Propositions 2.2(b) and 2.4, and Lemmas 3.10(b) and 4.2 imply that $\Phi^\varepsilon \xrightarrow{(e)} \Phi^0$. This shows Theorem b in the case where $a < \infty$.

In the above argument, the point is as follows. For any $m_\beta \leqq v_\varepsilon \leqq M_\beta$, $v_\varepsilon \overset{w}{\to} v$ in $H_0^1(\Omega)$ we have $B(v_\varepsilon)$, $B(v) \in W_0^{1,\rho}(\Omega)$, $B(v_\varepsilon) \overset{w}{\to} B(v)$ in $W_0^{1,\rho}(\Omega)$ and $\alpha_\varepsilon \bar{\delta}_\varepsilon \overset{s}{\to} a\,dx$ in $W^{-1,\rho^*}(\Omega)$.

Now, it remains to prove Theorem b with $a = \infty$. Even in this case, $\{u_\varepsilon^\sim\}$ is bounded in $H_0^1(\Omega)$. After dividing both sides of $(4.2)_b$ by a_ε and applying (3.6), we get $\int_\Omega B(u)\,dx = 0$. By (β-4) we see that $u = 0$ almost everywhere in Ω. □

(B) *Proof of Theorem* c. We suppose $a_h < \infty$. First recall the definition of m_β^ε, M_β^ε and B^ε in the first paragraph in this section. We take $w \in K_*^\varepsilon \equiv \{v \in V_\varepsilon;\ B^\varepsilon(w) \in L^1(\partial T_\varepsilon)\}$; i.e., $m_\beta^\varepsilon \leqq w \leqq M_\beta^\varepsilon$ on ∂T_ε. After substituting $v = \theta_\varepsilon w$ into (4.1) and dividing both sides of (4.1) by θ_ε^2, we get

$$\int_{\Omega_\varepsilon} [\nabla z_\varepsilon \nabla(w - z_\varepsilon) - (w - z_\varepsilon)\theta_\varepsilon^{-1} f]\,dx + a_{h,\varepsilon}\langle \theta_\varepsilon^{-1}\bar{\delta}_\varepsilon, B^\varepsilon(w) - B^\varepsilon(z_\varepsilon)\rangle - \langle g_\varepsilon \bar{\delta}_\varepsilon, w - z_\varepsilon\rangle / \theta_\varepsilon \geqq 0. \tag{4.7}$$

Let $\mathbf{k}_*^\varepsilon$ and $\mathbf{m}_*^\varepsilon$ be the new functionals defined by

$$\mathbf{k}_*^\varepsilon(\xi) = -\theta_\varepsilon^{-1}\langle g_\varepsilon \bar{\delta}_\varepsilon, \xi\rangle \quad \text{for } \xi \in L^2(\partial T_\varepsilon),$$

$$\mathbf{m}_*^\varepsilon(\xi) = \begin{cases} a_{h,\varepsilon}\theta_\varepsilon^{-1}\langle \bar{\delta}_\varepsilon, B^\varepsilon(\xi)\rangle & \text{for } \xi \in L^2(\partial T_\varepsilon),\quad B(\xi) \in L^1(\partial T_\varepsilon), \\ \infty & \text{otherwise.} \end{cases}$$

Let $\mathbf{j}_*^\varepsilon : L^2(\partial T_\varepsilon) \to [0, \infty]$ and $j_*^\varepsilon : L^2(\Omega_\varepsilon) \to [0, \infty]$ defined by $\mathbf{j}_*^\varepsilon = \mathbf{k}_*^\varepsilon + \mathbf{m}_*^\varepsilon$ and $j_*^\varepsilon(v) = \mathbf{j}_*^\varepsilon(v|\partial T_\varepsilon)$ for $v \in V_\varepsilon$ with $\mathbf{j}_*^\varepsilon(v|\partial T_\varepsilon) < \infty$, $\mathbf{j}_*^\varepsilon(v) = \infty$, otherwise. Furthermore,

$$\Phi_*^\varepsilon(v) = \begin{cases} \varphi^\varepsilon(v) + j_*^\varepsilon(v) & \text{for } v \in K_*^\varepsilon, \\ \infty & \text{otherwise.} \end{cases}$$

We set $k_*^0(v) = -I_g \int_\Omega v\,dx$ for $v \in H$, $m_*^0(v) = a_h S(\partial T) \int_\Omega B_*^0(v)\,dx$ for $v \in H_0^1(\Omega)$, and $B_*^0(v) \in L^1(\Omega)$ and $m_*^0(v) = \infty$, otherwise. Let $j_*^0 = k_*^0 + m_*^0$ and $\Phi_*^0 = \varphi^0 + j_*^0$. Inequality (4.7) and (1.12) are rewritten by using subdifferentials as $\partial\Phi_*^\varepsilon(z_\varepsilon) \ni \theta_\varepsilon^{-1} f$ and $\partial\Phi_*^0(z) \ni 0$, respectively. It suffices to see the convergence of $w_\varepsilon^\sim$ to w in H, where $\partial\Phi_*^\varepsilon(w_\varepsilon) \ni f_\varepsilon$, $\partial\Phi_*^0(w) \ni \eta f$ and $f_\varepsilon \to f$ as $\varepsilon \to 0$, where $\eta = |Y^*|$. $\{(\partial\Phi_*^\varepsilon)^{-1}\}$ is uniformly bounded by the inequality obtained from (4.7) replacing f/θ_ε with f_ε, where f_ε belongs to a bounded set in H. By Proposition 2.2(b) it suffices to show that

$$\Phi_*^\varepsilon \xrightarrow{(e)} \Phi_*^0. \tag{4.8}$$

It suffices to show the conditions in Proposition 2.4. By the definition of m_β^ε and M_β^ε we see that $m_\beta^\varepsilon \to m_\beta^0$ and $M_\beta^\varepsilon \to M_\beta^0$ as $\varepsilon \to 0$. We regard F^ε in Lemma 4.2 with $c_\varepsilon = m_\beta^\varepsilon$ and $C_\varepsilon = M_\beta^\varepsilon$ as F^n in Proposition 2.4 with $\varepsilon = \varepsilon_n$. Lemma 4.2(a) implies (ii) of Proposition 2.4. We have to show (iii) of Proposition 2.4.

For this aim we need some p such that

$$B_*^\varepsilon(v_\varepsilon), B_*^0(v) \in W_0^{1,p}(\Omega),\ B_*^\varepsilon(v_\varepsilon) \overset{w}{\to} B_*^0(v) \quad \text{in } W_0^{1,p}(\Omega) \tag{4.8}_1$$

and

$$a_{h,\varepsilon}\theta_\varepsilon^{-1}\bar{\delta}_\varepsilon \overset{s}{\to} a_h|\partial T|\,dx \quad \text{in } W^{-1,p^*}(\Omega) \tag{4.8}_2$$

for any $\{v_\varepsilon\}$ such that

$$m_\beta^\varepsilon \leqq v_\varepsilon \leqq M_\beta^\varepsilon \quad \text{and} \quad v_\varepsilon \overset{w}{\to} v \quad \text{in } H_0^1(\Omega) \quad \text{as } \varepsilon \to 0.$$

LEMMA 4.3. (i) *We have* $(4.8)_1$ *with* $p=\rho$ *provided that* $(\gamma\text{-}1)$ *or* $(\gamma\text{-}2)$ *are satisfied; in this case,* $m_\beta^\varepsilon=-\infty$ *and* $M_\beta^\varepsilon=\infty$.

(ii) *We also have* $(4.8)_1$ *with* $p=2$ *provided that* $(\gamma\text{-}3)$ *are satisfied.*

Proof. The functions β^ε and β^0 satisfy $(\beta\text{-}2)$ with the same r and coefficients C_1 and C_2, as those of β, because $\theta_\varepsilon\to\infty$ and $h\geqq r$.

For β satisfying $(\beta\text{-}1)$ or $(\beta\text{-}2)$ with (1.10), the inclusion, $B^\varepsilon(v_\varepsilon)$, $B^0(v)\in W_0^{1,\rho}(\Omega)$, and the boundedness of $\{B^\varepsilon(v_\varepsilon)\}$ in $W_0^{1,\rho}(\Omega)$, is derived from the definition of $B^\varepsilon(v_\varepsilon)$ using the Sobolev imbedding theorem together with Hölder's inequality as in the proof of Lemma 4.2. The convergence of $\{B^\varepsilon(v_\varepsilon)\}$ in $L^\rho(\Omega)$ under $(\gamma-1)$, is implied by the growth condition of β^0 (cf. [15, Lemma B.1]).

The case $(\gamma\text{-}2)$ is proved by analogy with the proof of Dini's theorem for a monotone sequence of continuous functions defined on a compact set as follows.

We get $B^0(v_\varepsilon)\overset{w}{\to}B^0(v)$ in $W_0^{1,\rho}(\Omega)$ by Lemma 4.2(b). It suffices to show that

$$\|B^\varepsilon(v_\varepsilon)-B^0(v_\varepsilon)\|_{\rho,\Omega}\to 0 \quad \text{as } \varepsilon\to 0. \tag{4.9}$$

We see that $\{B^\varepsilon(\cdot)-B^0(\cdot)\}$ is a continuous mapping from $L^{q'}(\Omega)$ into $L^\rho(\Omega)$, where $2\leqq q'\leqq 2N/(N-2)$ for $N\geqq 3$, and $2\leqq q'<\infty$ for $N=2$, because of $|B^\varepsilon(\xi)|\leqq C(1+|\xi|^{r+1})$ with (1.10) (see [15, Lemma B.1, p. 78]. Suppose that (4.9) is not true. Let $B^n=B^\varepsilon$ with $\varepsilon=\varepsilon_n$ and $v_m=v_\varepsilon$ with $\varepsilon=\varepsilon_m$, $n, m\in\mathbb{N}$. There exists $\{v_m\}$ such that $\|B^n(v_m)-B^0(v_m)\|_{\rho,\Omega}\geqq\delta>0$ for all $m, n\in\mathbb{N}$ with some $\delta>0$, where the sequence $\{v_m\}$ converges to v_0 in $L^{q'}(\Omega)$. Let $I_n=\{v_m, m\in\mathbb{N}, v_0;\ \|B^n(v_m)-B^0(v_m)\|_{\rho,\Omega}\geqq\delta/2\}$. By $(\gamma\text{-}2)$ we get $I_n\supset I_{n+1}$. Since $\{v_m, m\in\mathbb{N}, v_0\}$ is compact, B^n-B^0 is continuous, and $\bigcap_{n=1}^\infty I_n=\varnothing$, we have n_0 such that $I_{n_0}=\varnothing$. This is a contradiction.

We suppose $(\gamma\text{-}3)$ holds. Then we have $B^\varepsilon(v_\varepsilon)-B^\varepsilon(v_0)\in L^2(\Omega)$, $\partial B^\varepsilon(v_\varepsilon)/\partial x_i=\beta^\varepsilon(v_\varepsilon)\partial v_\varepsilon/\partial x_i\in L^2(\Omega)$, and $B^\varepsilon(v_\varepsilon)-B^\varepsilon(v_0)\in H_0^1(\Omega)$. The class $\{B^\varepsilon\}$ is equicontinuous because $\|B^\varepsilon(v_\varepsilon)-B^\varepsilon(v_0)\|_\Omega^2\leqq C\|v_\varepsilon-v_0\|_\Omega^2$. On the other hand, we have

$$\|B^\varepsilon(v_0)-B^0(v_0)\|_\Omega\to 0 \quad \text{as } \varepsilon\to 0. \tag{4.10}$$

In fact, since

$$\partial(B^\varepsilon(v_0)-B^0(v_0))/\partial x_i=(\beta^\varepsilon(v_0)-\beta^0(v_0))\partial v_0/\partial x_i,$$
$$|(\beta^\varepsilon(v_0)-\beta^0(v_0))\partial v_0/\partial xi|^2\leqq C|\partial v_0/\partial x_i|^2,$$

and $(\beta^\varepsilon(v_0)-\beta^0(v_0))\partial v_0/\partial x_i\to 0$ as $\varepsilon\to 0$, we have $\|\nabla(B^\varepsilon(v_0)-B^0(v_0))\|_\Omega\to 0$ as $\varepsilon\to 0$ by the Lebesgue convergence theorem. Thus, (4.10) follows from the Poincaré inequality. We have shown the lemma. □

Since $\theta_{N-\rho}=\infty=\theta_{N-2}$ for $r_\varepsilon\gg r_\varepsilon^{(N-1)}$, $(4.8)_2$ is implied from Lemma 3.10(a) with $a_h<\infty$, where $p=\rho$ for $(\gamma\text{-}1)$ and $(\gamma\text{-}2)$ and $p=2$ for $(\gamma\text{-}3)$. Thus, we have (4.8).

For the case where $a_h=\infty$, $\{z_\varepsilon^\sim\}$ is bounded in $H_0^1(\Omega)$. We take a subsequence, denoted by $\{z_\varepsilon^\sim\}$ again, such that $z_\varepsilon^\sim\overset{w}{\to}z$ in $H_0^1(\Omega)$ as $\varepsilon\to 0$. We substitute $w=0$ into (4.7), and divide both sides of the inequality by $a_{h,\varepsilon}$. We apply Lemma 3.10(a) with $\theta_{N-2}=\infty$, and get $\int_\Omega B_*^0(z)\,dx=0$. By $(\beta\text{-}4)$ we get $z=0$. This concludes the proof of Theorem c.

(C) *Proof of Theorem* a′. The proof is done in the same way as the proof of Theorem A.2 of [15], because, in that proof the differentiability of β is not used at all. Thus, we can omit the proof of Theorem a′.

(D) *Proof of Theorem* b′. This proof is also derived directly from the proof of Theorem B in [15] together with Lemma 3.10. Thus, we omit it.

5. Asymptotic behavior in $\mathbb{R}^2$. Theorem a′ says nothing for the case where $N=2$, nor does Theorem a cover the case where

$$N=2, \quad r_\varepsilon\leqq Cr_\varepsilon^{(0)} \quad \text{with } \hat{\alpha}=\infty. \tag{5.1}$$

Let $\hat{\alpha}_\varepsilon = \tilde{\alpha}_\varepsilon(\log r_\varepsilon^{-1})^{1/2}$ and $\hat{\alpha}_* = \lim \hat{\alpha}_\varepsilon$. We have $\hat{\alpha} < \infty$ if and only if $\hat{\alpha}_* < \infty$. Let $c_\varepsilon = \theta_{0,\varepsilon}/a_\varepsilon$ and $c = \lim c_\varepsilon$. We also have

$$\hat{\alpha}_\varepsilon = 2(\log r_\varepsilon^{-1})^{1/2}/c_\varepsilon.$$

Here, $r_\varepsilon T$ is supposed to be a ball $\{x \in \mathbb{R}^N; |x| \leqq r_\varepsilon/2\}$ together with $\beta(\xi) = \xi$ in (0.2). We are interested in the case where $0 < c < \infty$, and the case where $c = 0$. The state $c = 0$ follows both

(i) $\theta_0 \in [0, \infty)$ and $a = \infty$,

and

(ii) $\theta_0 = 0$ and $a \in (0, \infty]$.

The state $0 < c < \infty$ follows

(iii) θ_0, $a \in (0, \infty)$.

The remaining case of (θ_0, a) such that $\theta_0 = 0$ and $a = 0$ may imply various values of c.

The test function in [12], $w_\varepsilon(r) = [1 - \alpha_\varepsilon \bar{r}_\varepsilon(\log(r_\varepsilon/\varepsilon) - \log(r/\bar{\varepsilon}))]/(1 - \alpha_\varepsilon \bar{r}_\varepsilon \log(r_\varepsilon/\varepsilon))$ with $r = |x|$, $\bar{\varepsilon} = \varepsilon/2$, and $\bar{r}_\varepsilon = r_\varepsilon/2$, is useful. Let $\theta_{0,\varepsilon}^{\sim} = (\varepsilon^2 \log(\varepsilon/r_\varepsilon))^{-1}$. The formula $\partial w_\varepsilon/\partial r|_{r=\varepsilon} = \bar{\varepsilon}\theta_{0,\varepsilon}^{\sim}(1 + \theta_{0,\varepsilon}^{\sim}/a_\varepsilon)^{-1}$, together with the method in Cioranescu and Murat [7], imply the proposition below.

PROPOSITION 5.1. *We suppose* $r_\varepsilon \leqq Cr_\varepsilon^{(0)}$, (0.4) *with a fixed* $\theta_0 < \infty$ *and a fixed* $c \in [0, \infty]$ *as* $\varepsilon \to 0$. *Then* $u_\varepsilon^{\sim} \to u$ *weakly in* $H_0^1(\Omega)$, *where* u *is determined by the equation*

$$-\Delta u + 2\pi\theta_0(c+1)^{-1}u = f \quad \text{a.e. in } \Omega. \tag{5.2}$$

Remark 5.2. If instead of (0.2) we consider the Dirichlet boundary condition, we get (5.2) with $c = 0$, using the method in Cioranescu and Murat [7].

6. Applications.

Application 6.1. The theorems in § 1 are applied to monotone functions β such as (b), (c), $(d)_p$, (e), (f), and (g) described in the Introduction. We give Table 1 for $N \geqq 3$.

Application 6.2. Let C_1, $C_2 > 0$. We suppose $\int_{\partial T} g\, d\sigma \neq 0$,

$$\beta(\xi) = \begin{cases} C_1\xi & \text{for } \xi \leqq 0, \\ C_2\xi & \text{for } \xi \geqq 0, \end{cases}$$

TABLE 1

Theorem	To be applicable	Not to be applicable
A	(b) (c) $(d)_p$ (e) (f) (g)	(a)
B	(c) $(d)_p$ (e)	(a) (b) (f) (g)
C (I)	(b) (c) $(d)_p$ (e) (f) (g)	(a)
C (II)	(b) (c) $(d)_p$ (e) (f)	(a) (g)
C (III)	(c) $(d)_p$ (e)	(a) (b) (f) (g)
a	(b) (c) $(d)_p$ (e) (f) (g)	(a)
b	(b) (c) $(d)_p$ (e) (f) (g)	(a)
c	(b) (c) $(d)_p$ (e) (f) (g)	(a)
a′	(c) $\equiv$ $(d)_p$ $(p=1)$ (e)	(a) (b) (f) (g)
b′	(c) $\equiv$ $(d)_p$ $(p=1)$ (e)	(a) (b) (f) (g)

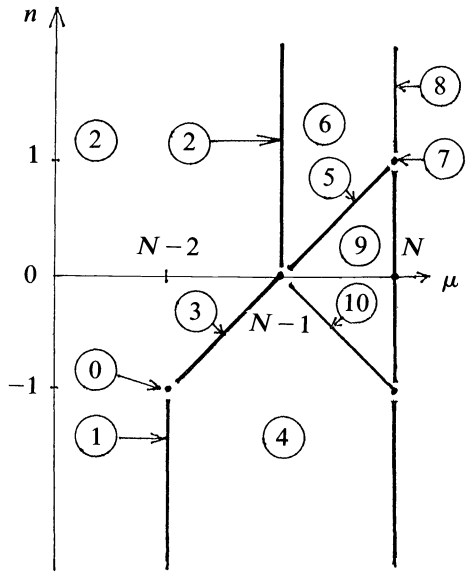


FIG. 2

TABLE 2

Number	Equations or behavior	Number	Equations or behavior
0	Unknown	1	$-\Delta u + C_T u = f$
2	$-\Delta u = f$	3	$-\Delta u + S(\partial T)\beta(u) = f$
4	$u_\varepsilon^{\sim} \to u = 0$	5	$-\Delta z + S(\partial T)\beta(z) = I_g$
6	$-\Delta z = I_g$	7	$\mathcal{H}z + S(\partial T)\beta(z) = I_g$
8	$\mathcal{H}z = I_g$	9	$u_\varepsilon^{\sim} / b_\varepsilon^{1/q'} \to 0$
10	$u_\varepsilon^{\sim} \to 0$		

and relations among ε, r_ε, and α_ε as follows:

$$\varepsilon = r_\varepsilon^m, \qquad m \in (0, 1],$$
$$\alpha_\varepsilon = r_\varepsilon^n, \qquad n \in \mathbb{R}.$$

Let $\mu = Nm \in (0, N]$. Then all the limit equations are displayed as points having coordinates (μ, n) in $(0, N] \times \mathbb{R}$. Theorem a′ derives limit equations number 1, 2 in Fig. 2, for points of $\{(\mu, n);\ \mu \leqq N-2 \text{ and } n \in \mathbb{R}\}$. Theorem b′ also derives the equations number 2, 3, 4, which are settled in the region, $\{(u, n);\ N-2 < \mu \leqq N-1 \text{ and } n \in \mathbb{R}\}$. Theorems B and c concern the set $\{(\mu, n);\ N-2 < \mu \leqq N,\ n < \mu - N + 1 \text{ and } n < -\mu + N - 1\}$ and $\{(\mu, n);\ N-1 < \mu \leqq N \text{ and } n \geqq \mu - N + 1\}$, respectively (number 4, 5, 6, 7, and 8). Last, by Theorem C(III) we get limit equations (number 9, 10) for points of $\{(\mu, n);\ N-1 < \mu \leqq N,\ n < \mu - N + 1 \text{ and } n \geqq -\mu + N - 1\}$. Thus, limit equations are drawn in one picture as Fig. 2 in [15]. Unfortunately, by our theorems we can show no limit equation corresponding to a point $(N-2, -1)$ (number 0), when $C_1 \neq C_2$. The exact equation corresponding to each number is described in Table 2.

REFERENCES

[1] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, Boston, MA, 1984.
[2] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, 1976.

[3] A. BENSOUSSAN, J. L. LIONS AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[4] H. BREZIS, *Problèmes unilatéraux*, J. Math. Pures Appl., 51 (1972), pp. 1-168.

[5] ———, *Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations*, in Contributions to Nonlinear Functional Analysis, E. Zarantonello, ed., Academic Press, New York, 1971.

[6] D. CIORANESCU AND P. DONATO, *Homogénéisation du problème de Neumann non homogène dans des ouverts perforés*, Asymptotic Anal., 1 (1988), pp. 115-138.

[7] D. CIORANESCU AND F. MURAT, *Un term étrange venu d'aillerus*, I and II, Nonlinear partial differential equations, Res. Notes in Math., 60 and 70, Pitman, Boston, 1982, Boston, pp. 98-138, 154-178.

[8] D. CIORANESCU AND J. SAINT JEAN PAULIN, *Homogenization in open sets with holes*, J. Math. Anal. Appl., 71 (1979), pp. 590-607.

[9] C. CONCA, *On the application of the homogenization theory to a class of problems arising in fluid mechanics*, J. Math. Pures Appl., 64 (1985), pp. 31-75.

[10] C. CONCA AND P. DONATO, *Non-homogeneous Neumann's problems in domains with many small holes*, Math. Model. Numer. Anal., 22 (1988), pp. 561-607.

[11] A. DAMLAMIAN AND P. DONATO, *Homogenization with small shape-varying perforations*, preprint.

[12] S. KAIZU, *The Robin problems on domains with many tiny holes*, Proc. Japan Acad. Ser. A Math. Sci., 61 (1985), pp. 39-42.

[13] ———, *A monotone boundary condition for a domain with many tiny spherical holes*, Proc. Japan Acad. Ser. A Math. Sci., 61 (1985), pp. 141-143.

[14] ———, *An average effect of many tiny holes in nonlinear boundary values problems with monotone boundary conditions*, Proc. Japan Acad. Ser. A Math. Sci., 62 (1986), pp. 133-136.

[15] ———, *The Poisson equation with semilinear boundary conditions in domains with many tiny holes*, J. Fac. Sci. Univ. Tokyo Sect. IA, Math., 36 (1989), pp. 43-86.

[16] ———, *Behavior of solutions of the Poisson equation under fragmentation of the boundary of the domain*, Japan J. Appl. Math., 7 (1990), pp. 77-102.

[17] Ē. YA. KHRUSLOV, *The method of orthogonal projections and the Dirichlet problem in domains with a fine-grained boundary*, Math. USSR-Sb., 17 (1972), pp. 37-59.

[18] ———, *The first boundary value problem in domains with a complicated boundary for higher order equations*, Math. USSR-Sb., 32 (1977), pp. 535-549.

[19] ———, *The asymptotic behavior of solutions of the secondary boundary value problem under fragmentation of the boundary of the domain*, Math. USSR-Sb., 35 (1979), pp. 266-282.

[20] ———, *Convergence of solutions of the secondary boundary value problems in weakly connected domains*, Naukova Dumka 191, Kiev, 1981, pp. 129-173.

[21] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod Gauthier-Villars, Paris, 1969.

[22] S. OZAWA, *Point interaction potential approximation for* $(-\Delta + U)^{-1}$ *and eigenvalues of the Laplacian on wildly perturbed domains*, Osaka J. Math., 20 (1983), pp. 923-937.

[23] J. RAUCH AND M. TAYLOR, *Potential and scattering theory on wildly perturbed domains*, J. Funct. Anal., 18 (1975), pp. 27-59.

[24] M. VANNINATHAN, *Homogénéisation des valeur propres dans les milieux perforés*, C.R. Acad. Sci. Paris 287(A) (1987), pp. 403-406.

[25] ———, *Homogenization of eigenvalue problems in perforated domains*, Proc. Indian Acad. Sci. Math. Sci., 90 (1981), pp. 239-271.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1246-1255, September 1991

003

ON THE NEWTONIAN POTENTIAL OF A HETEROGENEOUS ELLIPSOID*

HENRIK SHAHGHOLIAN†

Abstract. In this note the Newtonian potential of a heterogeneous ellipsoid in $\mathbb{R}^n$ is calculated. It turns out that for polynomial densities the potential is also a polynomial in the interior of the ellipsoid. As an application, it is shown that the solution of $\Delta u = P$ near ∂E and $u = |\nabla u| = 0$ on ∂E, where E is an ellipsoid and P a polynomial, has a harmonic continuation to $\mathbb{R}^n \setminus E_0$, where E_0 is the focal ellipsoid of E.

Key words. Newtonian potential, Schwarz potential, ellipsoid

AMS(MOS) subject classifications. 31C99, 35J05

Introduction. This paper consists of two parts:

(1) Calculation of the Newtonian potential of a heterogeneous ellipsoid, at an internal point. Here we prove that the Newtonian potential of an ellipsoid in $\mathbb{R}^n$ with polynomial density is a polynomial in the interior of the ellipsoid. This was proved in $\mathbb{R}^3$ by Ferrers (see [Fe]) using special techniques of ellipsoidal coordinates; it seems difficult to extend this method to more than three dimensions.

In any case, it seems of interest to give a new proof of this fundamental result in $\mathbb{R}^n$. We do this by adapting the celebrated method of Dirichlet [Di], based on Fourier's integral, combined with Ferrer's [Fe] idea that any polynomial can be expressed by a finite linear combination of P, $P^2, \cdots$ and their derivatives of different degrees (see Lemma 1.4).

(2) Application to the Cauchy problem for the Laplace equation. Here we use part (1) to solve the following Cauchy problem:

$$\begin{aligned} &\Delta u = 0 \quad \text{near } \partial E \\ &u \equiv P \quad \text{on } \partial E \end{aligned} \qquad E \text{ is an ellipsoid in } \mathbb{R}^n.$$

It turns out that the solution of this problem is harmonic in $\mathbb{R}^n \setminus E_0$.

Notation. We will use the following notation.

(i) By $g \equiv f$, we mean $g = f$ and $\nabla g = \nabla f$, where f, $g \in C^1(\mathbb{R}^n)$ and ∇ denotes gradient.

(ii) The solution of the following Cauchy problem, when it exists, will be called the *Schwarz potential* U_Γ of Γ, where Γ is a hypersurface in $\mathbb{R}^n$,

$$\begin{aligned} &\Delta u = 0 \quad \text{near } \Gamma, \\ &u \equiv \tfrac{1}{2}|x|^2 \quad \text{on } \Gamma. \end{aligned}$$

(iii) $K_m(x, y)$ is defined as follows. Set $\psi(x-y) = \omega_n^{-1}|x-y|^{2-n}$, where $\omega_n = (2-n) \cdot A_n$; A_n is the surface area of the $(n-1)$-dimensional unit sphere. For $a \in \mathbb{R}^n$, $x \in \mathbb{R}^n$, and $x \neq a$ set

$$K_m^a(x, y) := \psi(x-y) - \sum_{|\alpha|<m} \frac{(y-a)^\alpha}{\alpha!} \frac{\partial^\alpha}{\partial z^\alpha} \psi(x-z)\big|_{z=a},$$

$$K_m(x, y) := K_m^0(x, y).$$

* Received by the editors November 28, 1989; accepted for publication (in revised form) October 3, 1990.
† Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden.

(iv) Let $\Omega \subset \mathbb{R}^n$ and $\bar{\Omega} \neq \mathbb{R}^n$. Then we define the *generalized Newtonian potential* with polynomial density R, with respect to $a \in \mathbb{R}^n$ by

$$\int_\Omega K_m^a(x, y) R(x)\, dx, \quad \text{where } m = \deg R + 3 \text{ and } a \notin \bar{\Omega}.$$

(v) For $a_1 \geqq a_2 \geqq \cdots \geqq a_n > 0$ we define E and E_0:

$$E = \left\{ x \in \mathbb{R}^n : \sum_{j=1}^n \frac{x_j^2}{a_j^2} \leqq 1 \right\},$$

$$E_0 = \left\{ x \in \mathbb{R}^n : \sum_{j=1}^{n-1} \frac{x_j^2}{a_j^2 - a_n^2} \leqq 1 \text{ and } x_n = 0 \right\}.$$

E_0 is the so-called *focal ellipsoid* of E.

(vi) We also define φ, S, and P as follows:

$$\varphi^2(s) = (s + a_1^2)(s + a_2^2) \cdots (s + a_n^2),$$

$$S(s, y) = \sum_{j=1}^n \frac{y_j^2}{a_j^2 + s} \quad \text{and} \quad P(x) = 1 - \sum_{j=1}^n \frac{x_j^2}{a_j^2}.$$

1. Calculation of the Newtonian potential of a heterogeneous ellipsoid, at an internal point. Here we calculate the potential of E (ellipsoid) with polynomial density. To do this we begin with the following theorem.

THEOREM 1.1. *Let $m \geqq 0$; then*

$$\int_E \frac{P^m(x)\, dx}{|x - y|^{n-2}} = \frac{c_n}{m+1} \int_\lambda^\infty \left(1 - \sum_{j=1}^n \frac{y_j^2}{s + a_j^2} \right)^{m+1} \varphi(s)^{-1}\, ds$$

where $c_n = \pi^{n/2} a_1 \cdots a_n \cdot (n-2)/2\Gamma(n/2)$, and λ is the largest of the ellipsoidal coordinates of $y \notin E$ and $\lambda = 0$ if $y \in E$. Consequently, if m is a nonnegative integer then the Newtonian potential of E with density $P^m(x)$ is a polynomial of degree $m+2$ in the interior of E.

Proof. Set

$$V(y) = \int_E \frac{P^m(x)}{|x - y|^{n-2}}\, dx \quad \text{and} \quad V_k := \frac{\partial V}{\partial y_k}(y).$$

We first calculate $V_k(y)$ for $y \in \mathbb{R}^n$, and $m > 0$. Define

$$f(\xi) = \begin{cases} (1 - |\xi|)^m, & |\xi| < 1, \\ 0, & |\xi| \geqq 1, \end{cases} \quad \xi \in \mathbb{R}.$$

Then

$$f(\xi) = \operatorname{Re} \int_0^\infty \frac{\hat{f}(t)}{\pi} e^{it\xi}\, dt, \tag{1}$$

where

$$\hat{f}(t) = \int_{-\infty}^\infty f(\xi)\, e^{-it\xi}\, d\xi$$

is in $L^1(\mathbb{R})$, for $m > 0$. We have

$$\frac{V_k}{n-2} = \int_{\mathbb{R}^n} f\left(\sum_{j=1}^n \frac{x_j^2}{a_j^2} \right) (x_k - y_k) |x - y|^{-n}\, dx.$$

Here we want to use (1) with $\xi = \sum_{j=1}^{n} (x_j^2/a_j^2)$ and then change the order of integration, but since $(x_k - y_k)|x-y|^n$ is not integrable over $\mathbb{R}^n$, we first introduce a convergence factor $\exp(-\delta|x|^2)$ which gives

$$\frac{V_k}{n-2} = \lim_{\delta \to 0} \int_{\mathbb{R}^n} f\left(\sum_{j=1}^{n} \frac{x_j^2}{a_j^2}\right)(x_k - y_k)|x-y|^{-n} e^{-\delta|x|^2}\, dx. \tag{2}$$

Now we insert (1) into (2) which implies that (2) becomes

$$\operatorname{Re}\left[\lim_{\delta \to 0} \int_{\mathbb{R}^n} \int_0^\infty \frac{\hat{f}(t)}{\pi} I_1(x, y, t)\, dt\, dx\right], \tag{2'}$$

where $I_1(x, y, t) = (x_k - y_k)|x-y|^{-n} \exp[it \sum_{j=1}^{n} (x_j^2/a_j^2) - \delta|x|^2]$. Here we change the order of integration, so (2') becomes

$$\lim_{\delta \to 0} \operatorname{Re} \int_0^\infty \frac{\hat{f}(t)}{\pi} \int_{\mathbb{R}^n} I_1(x, y, t)\, dx\, dt. \tag{2''}$$

The idea is to split the integration over $\mathbb{R}^n$ into n products of one-dimensional integrals. For this we will use the following known formula:

$$\int_0^\infty \exp[-(\varepsilon - i\beta)\nu] \cdot \nu^{r-1}\, d\nu = (\varepsilon - i\beta)^{-r} \cdot \Gamma(r),$$

where $r > 0$, $\varepsilon > 0$, $\beta \in \mathbb{R}^n$, and Γ is the gamma function (see [GH, p. 62]). To make the formula applicable to our purposes, we set $\beta = |x-y|^2$ and $r = n/2$. Then

$$\begin{aligned} &\int_{\mathbb{R}^n} I_1(x, y, t)\, dx \\ &= \lim_{\varepsilon \to 0} (-i)^{n/2} \int_{\mathbb{R}^n} [\varepsilon - i|x-y|^2]^{-n/2}(x_k - y_k) \exp\left[it \sum_{j=1}^{n} \frac{x_j^2}{a_j^2} - \delta|x|^2\right] dx \qquad (3) \\ &= \lim_{\varepsilon \to 0} (-i)^{n/2} (\Gamma(n/2))^{-1} \int_{\mathbb{R}^n} \int_0^\infty \exp(I_2) \cdot (x_k - y_k) \cdot \nu^{(n-2)/2}\, d\nu\, dx, \end{aligned}$$

where

$$\begin{aligned} I_2 &= -(\varepsilon - i|x-y|^2)\nu + it \sum_{j=1}^{n} \frac{x_j^2}{a_j^2} - \delta|x|^2 \\ &= -\varepsilon\nu + i|y|^2\nu + \sum_{j=1}^{n} \left[\left(i\nu + \frac{it}{a_j^2} - \delta\right)x_j^2 - 2iy_j\nu x_j\right]. \end{aligned}$$

Set

$$P_j = \left(\nu + \frac{t}{a_j^2}\right) \quad \text{and} \quad Q_j = -y_j\nu, \quad j = 1, 2, \cdots, n;$$

then

$$I_2 = -\varepsilon\nu + i|y|^2\nu + \sum_{j=1}^{n} [(iP_j - \delta)x_j^2 + 2iQ_j x_j].$$

Now changing the order of integration in (3) we can reduce it to

$$\lim_{\varepsilon \to 0} (-i)^{n/2} (\Gamma(n/2))^{-1} \int_0^\infty \nu^{(n-2)/2} \exp[-\varepsilon\nu + i|y|^2\nu] G(t, y, \nu)\, d\nu, \tag{3'}$$

where

$$G(t, y, \nu) = \int_{\mathbb{R}^n} \exp \sum_{j=1}^{n} [(iP_j - \delta)x_j^2 + 2iQ_j x_j] \cdot (x_k - y_k)\, dx$$
$$= \left(\prod_{\substack{j=1 \\ j \neq k}}^{n} \int_{-\infty}^{\infty} \exp [(iP_j - \delta)x_j^2 + 2iQ_j x_j]\, dx_j \right)$$
$$\cdot \int_{-\infty}^{\infty} (x_k - y_k) \exp [(iP_k - \delta)x_k^2 + 2iQ_k x_k]\, dx_k.$$

By calculating these integrals we find

$$\int_{-\infty}^{\infty} \exp [(iP_j - \delta)x_j^2 + 2iQ_j x_j]\, dx_j = \frac{\exp [-Q_j^2/(\delta - iP_j)]}{\sqrt{\delta - iP_j}} \cdot \sqrt{\pi}$$

and

$$\int_{-\infty}^{\infty} (x_k - y_k) \exp [(iP_k - \delta)x_k^2 + 2iQ_k x_k]\, dx_k$$
$$= \left(\frac{iQ_k}{\delta - iP_k} - y_k \right) \frac{\exp [-Q_k^2/(\delta - iP_k)] \cdot \sqrt{\pi}}{\sqrt{\delta - iP_k}}$$
$$= \frac{-\delta + (it/a_k^2)}{(\delta - iP_k)} y_k \frac{\exp [-Q_k^2/(\delta - iP_k)] \cdot \sqrt{\pi}}{\sqrt{\delta - iP_k}},$$

i.e.,

$$G(t, y, \nu) = \pi^{n/2} \frac{(-\delta + (it/a_k^2))}{\delta - iP_k} \cdot y_k \cdot \frac{\exp [-\sum_{j=1}^{n} Q_j^2/(\delta - iP_j)]}{\sqrt{(\delta - iP_1) \cdots (\delta - iP_n)}}. \tag{4}$$

Now, inserting (4) in (3′) and making the variable changes $s = t/\nu$ we get, after recalling the definition of P_j and Q_j, that (3′) is reduced to

$$\lim_{\varepsilon \to 0} d_n \int_0^{\infty} \frac{((-ia_k^2\delta/t) - 1)y_k}{R_k} \cdot \frac{\exp [-\varepsilon t/s + i \sum_{j=1}^{n} ((i\delta a_j^2 + t)/R_j)y_j^2]}{\sqrt{R_1 \cdot R_2 \cdots R_n}}\, ds$$
$$= d_n \int_0^{\infty} \frac{((-ia_k^2\delta/t) - 1)y_k}{R_k} \cdot \frac{\exp [i \sum_{j=1}^{n} ((i\delta a_j^2 + t)/R_j)y_j^2]}{\sqrt{R_1 \cdots R_n}}\, ds, \tag{3″}$$

where $R_j = ((i\delta a_j^2 s/t) + a_j^2 + s)$ and

$$d_n = \pi^{n/2} \cdot (\Gamma(n/2))^{-1} \cdot a_1 \cdots a_n = \frac{2c_n}{n-2}.$$

Summing up we arrive at

$$\int_{\mathbb{R}^n} I_1(x, y, t)\, dx = d_n \int_0^{\infty} F\, ds,$$

where

$$F = F(y, s, t, \delta) = \frac{((-ia_k^2\delta/t) - 1)y_k}{R_k} \cdot \frac{\exp [i \sum_{j=1}^{n} ((i\delta a_j^2 + t)/R_j)y_j^2]}{\sqrt{R_1 \cdots R_n}}.$$

Now we put this into (2″); then we have

$$\frac{V_k}{n-2} = \operatorname{Re} \lim_{\delta \to 0} d_n \int_0^{\infty} \int_0^{\infty} \frac{\hat{f}(t)}{\pi} F\, ds\, dt.$$

Since $\int_0^\infty |F|\,ds<\infty$ we can change the order of integration and let $\delta\to 0$. This gives

$$\frac{V_k}{n-2}=d_n\int_0^\infty \operatorname{Re}\int_0^\infty \frac{\hat f(t)}{\pi}F_0\,dt\,ds,$$

where

$$F_0=F(x,y,t,0)=\frac{-y_k\exp\left[it\sum_{j=1}^n y_j^2/(a_j^2+s)\right]}{(a_k^2+s)\cdot[(a_1^2+s)\cdots(a_n^2+s)]^{1/2}}.$$

Recall the definition of S and φ; then we observe

$$V_k=(2-n)d_n y_k\int_0^\infty[\varphi(s)(s+a_k^2)]^{-1}\int_{-\infty}^\infty\frac{\hat f(t)}{2\pi}\exp\left[itS(s,y)\right]dt\,ds. \tag{5}$$

Now let λ be the positive root of $S(s,y)=1$ for fixed y, i.e., $\lambda=\lambda(y)$. Then the inner integral in (5) is zero for $s<\lambda$ and is equal to $f(S(s,y))$ for $s\geqq\lambda$. Therefore

$$\begin{aligned}V_k&=(2-n)d_n y_k\int_\lambda^\infty\frac{f(S(s,y))}{\varphi(s)(s+a_k^2)}\,ds\\&=(2-n)d_n y_k\int_\lambda^\infty\left(1-\sum_{j=1}^n\frac{y_j^2}{s+a_j^2}\right)^m\cdot[\varphi(s)(s+a_k^2)]^{-1}\,ds,\end{aligned}\qquad k=1,2,\cdots,n \tag{5'}$$

and, integrating,

$$V(y)=\frac{(n-2)d_n}{2(m+1)}\int_\lambda^\infty\left(1-\sum_{j=1}^n\frac{y_j^2}{s+a_j^2}\right)^{m+1}\varphi^{-1}(s)\,ds+c. \tag{6}$$

For $y\in E$, λ is replaced by zero in the above integral. The case $m=0$ could be obtained by letting m go to zero in (6). Thus the proof is completed.

Remark. In (6) we see, by letting $|y|\to\infty$, that $V(y)\to 0$ and the absolute value of the integral on the right is bounded by

$$A\int_\lambda^\infty\varphi^{-1}(s)\,ds,\quad\text{where } A=\frac{(n-2)d_n}{2(m+1)}$$

and this goes to zero, as $|y|\to\infty$ (since $\lambda=\lambda(y)\to\infty$). Thus $c=0$.

Remark. Theorem 1.1 is a special case of a much more general theorem, which we state without proof (the proof is similar to the proof of Theorem 1.1). Let g have compact support in $(0,\infty)$ and be (for example) piecewise continuous. Set

$$G(t)=\int_t^\infty g(\tau)\,d\tau.$$

Then

$$\int_{\mathbb{R}^n}\frac{g(\sum_{j=1}^n(x_j^2/a_j^2))\,dx}{(n-2)A_n}=\frac{\varphi(0)}{4}\int_0^\infty\frac{G(\sum_{j=1}^n(y_j^2/(a_j^2+s))\,ds}{\varphi(s)}.$$

We also want to mention that the case $m=0$ (for $n=3$) is the goal of classical papers such as those by Lagrange, Gauss, Chasles, and Dirichlet. For references see § 3.

COROLLARY 1.2. *Let P be as in the notation, and let m be a nonnegative integer. Then the Newtonian potential of E with density P^m is harmonic in the exterior of E and has a harmonic continuation into $E\setminus E_0$.*

Proof. Define

$$W(y)=\frac{c_n}{m+1}\int_\lambda^\infty\varphi^{-1}(s)\left(1-\sum_{j=1}^n\frac{y_j^2}{a_j^2+s}\right)^{m+1}ds,$$

where $c_n = \pi^{n/2} a_1 \cdots a_n (n-2)/2\Gamma(n/2)$, $\lambda = \lambda(y)$ is such that $\sum_{j=1}^{n} (y_j^2/(a_j^2+\lambda)) = 1$. Then it is easy to show that W is real analytic in $\mathbb{R}^n \setminus E_0$. Moreover, by Theorem 1.1, we have $W = V$ in $\mathbb{R}^n \setminus E$, where $V = \int_E (P^m(x)/|x-y|^{n-2})\, dx$. Thus V can be continued into $E \setminus E_0$ as a harmonic function.

THEOREM 1.3. *The Newtonian potential of the ellipsoid E in $\mathbb{R}^n (n \geq 3)$ with polynomial density Q of degree m is a polynomial of degree $m+2$ in the interior of E.*

To prove this theorem we need the following two lemmas, which are essentially due to Ferrers [Fe].

LEMMA 1.4. *Define $\mathcal{B}$ to be*

$$\mathcal{B} = \operatorname{span} \left\{ \sum_{j=0}^{m'} \sum_{|\alpha|=m-2j} b_\alpha \partial^\alpha P^{m-j} \right\}_{m=0}^{\infty},$$

whence

$$m' = \begin{cases} m/2 & m \text{ even}, \\ (m-1)/2 & m \text{ odd}, \end{cases}$$

$$b_\alpha \in \mathbb{R}, \quad \partial^\alpha = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \cdots \partial_n^{\alpha_n}, \quad \partial_j = \frac{\partial}{\partial x_j}$$

and P is as in the notation. Set

$$\mathcal{P} = \{\textit{polynomials in } n \textit{ variables with real coefficients}\}.$$

Then $\mathcal{B} = \mathcal{P}$.

LEMMA 1.5. *Let $m \geqq 0$ be an integer. Then for $|\alpha| \leqq m$,*

$$\int_{P>0} \frac{\partial^\alpha P^m(x)}{\partial x^\alpha} |x-y|^{2-n}\, dx = \frac{\partial^\alpha}{\partial y^\alpha} \int_{P>0} \frac{P^m}{|x-y|^{n-2}}\, dx \quad \forall y \in \mathbb{R}^n. \tag{7}$$

The lemmas will be proved by induction.

Remark. That the derivative in (7) exists depends on the fact that the function $\int_{P>0} P(x)^m |x-y|^{2-n}\, dx$ is of class $C^m(\mathbb{R}^n)$. This in turn depends on the following. Set $g = P^m$ for $P > 0$ and $g = 0$ for $P > 0$. Then $\int_{\mathbb{R}^n} g(x)|x-y|^{2-n}\, dx = \int_{P>0} P^m(x)|x-y|^{2-n}\, dx$ and $g \in C_0^{m-1}(\mathbb{R}^n)$, which implies that $\int_{\mathbb{R}^n} g(x)|x-y|^{2-n}\, dx$ is of class $C^m(\mathbb{R}^n)$.

Proof of Lemma 1.4. Define

$$\mathcal{P}_k = \{\text{polynomials of degree} \leqq k \text{ in } n \text{ variables with real coefficients}\}:$$

(i) $\mathcal{P}_0 \subset \mathcal{B}$ is clear.

(ii) Suppose $\mathcal{P}_k \in \mathcal{B}$. Then we will show that every monomial of the form $x_j x^\gamma$ for $|\gamma| = k$ and $j = 1, 2, \cdots, n$ is in $\mathcal{B}$. Set (without loss of generality) $j = 1$, then $x^\gamma \in \mathcal{B}$ implies

$$x_1 x^\gamma = x_1 \sum_{j=0}^{k'} \sum_{|\alpha|=k-2j} b_\alpha \partial^\alpha P^{k-j} = \sum_{j=0}^{k'} \sum_{|\alpha|=k-2j} b_\alpha [x_1 \partial^\alpha P^{k-j}], \tag{8}$$

where

$$k' = \begin{cases} k/2 & \text{if } k \text{ even}, \\ (k-1)/2 & \text{if } k \text{ odd}. \end{cases}$$

We introduce the notation $\beta = (\alpha_1 + 1, \alpha_2, \cdots, \alpha_n)$, $\sigma = (\alpha_1 - 1, \alpha_2, \cdots, \alpha_n)$. Then by the Leibniz formula

$$-x_1 \partial^\alpha P^{k-j} = \frac{a_1^2}{2(k+1-j)} \partial^\beta P^{k+1-j} + \alpha_1 \partial^\sigma P^{k-j}. \tag{9}$$

Substituting (9) into (8) we have

$$-x_1x^\gamma = \sum_{j=0}^{k'} \sum_{|\alpha|=k-2j} b_\alpha \left(\frac{a_1^2}{2(k+1-j)} \partial^\beta P^{k+1-j} + \alpha_1 \partial^\sigma P^{k-j} \right),$$

i.e., $x_1x^\gamma \in \mathcal{B}$ for all $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_n)$ and $|\gamma| = k$, which gives that $\mathcal{P}_{k+1} \in \mathcal{B}$, and the proof is completed.

Proof of Lemma 1.5. For $m = 0$ there is nothing to prove. Now let $m > 0$ and $|\alpha| \leqq m$. If $|\alpha| = 0$, then there is nothing to prove, so let $|\alpha| > 0$. Then $\alpha_j > 0$ for some j. So let $j = 1$ and $\alpha' = (\alpha_1 - 1, \alpha_2, \cdots, \alpha_n)$, then

$$\begin{aligned} &\int_{P>0} \frac{\partial^\alpha P^m(x)}{\partial x^\alpha} |x-y|^{2-n}\, dx \\ &\quad = \int_{P>0} \frac{\partial}{\partial x_1} \frac{\partial^{\alpha'} P^m(x)}{\partial x^{\alpha'}} \cdot |x-y|^{2-n}\, dx \\ &\quad = \int_{P>0} \frac{\partial}{\partial x_1} \left(\frac{\partial^{\alpha'} P^m(x)}{\partial x^{\alpha'}} \cdot |x-y|^{2-n} \right) dx - \int_{P>0} \frac{\partial^{\alpha'} P^m(x)}{\partial x^{\alpha'}} \cdot \frac{\partial}{\partial x_1} |x-y|^{2-n}\, dx \\ &\quad = \frac{\partial}{\partial y_1} \int_{P>0} \frac{\partial^{\alpha'} P^m(x)}{\partial x^{\alpha'}} \cdot |x-y|^{2-n}\, dx. \end{aligned} \tag{10}$$

Summing up,

$$\int_{P>0} \frac{\partial^\alpha P^m(x)}{\partial x^\alpha} |x-y|^{2-n}\, dx = \frac{\partial}{\partial y_1} \int_{P>0} \frac{\partial^{\alpha'} P^m(x)}{\partial x^{\alpha'}} \cdot |x-y|^{2-n}\, dx.$$

Proceeding in this way we can move the differential operator ∂^α outside the integral sign, and this will complete the proof of the lemma.

Proof of Theorem 1.3. By Lemma 1.4

$$Q(x) = \sum_{j=0}^{m'} \sum_{|\alpha|=m-2j} b_\alpha \partial^\alpha P^{m-j}$$

and

$$m' = \begin{cases} m/2 & m \text{ even}, \\ (m-1)/2 & m \text{ odd}. \end{cases}$$

Thus

$$\begin{aligned} \int_E Q(x)|x-y|^{2-n}\, dx &= \sum_{j=0}^{m'} \sum_{|\alpha|=m-2j} b_\alpha \int_E \frac{\partial^\alpha P^{m-j}}{\partial x^\alpha} \cdot |x-y|^{2-n}\, dx \\ &= \sum_{j=0}^{m'} \sum_{|\alpha|=m-2j} b_\alpha \frac{\partial^\alpha}{\partial y^\alpha} \int P^{m-j} |x-y|^{2-n}\, dx \quad \text{(by Lemma 1.5).} \end{aligned}$$

Now by Theorem 1.1 the last integral is a polynomial of degree $2m-2j+2$ in the interior of E, and since $|\alpha| = m-2j$, $(\partial^\alpha/\partial y^\alpha) \int P^{m-j}(x)|x-y|^{2-n}\, dx$ is a polynomial of degree $m+2$ in the interior of E. Thus $\int_E Q(x)|x-y|^{2-n}\, dx$ is a polynomial of degree $m+2$ in the interior of E.

Remark. In [DF] the authors prove the following theorem.

THEOREM [DiBenedetto, Friedman]. *Let Ω be a bounded domain in $\mathbb{R}^n$ and suppose that the Newtonian potential of Ω with constant density is a polynomial in the interior of Ω. Then Ω is an ellipsoid.*

It is very natural to ask whether or not this is true when the density is a polynomial. In the following proposition we will show that this is generally not true. But we still have the following problem.

PROBLEM. Give conditions on R (= polynomial) such that

$$\int_\Omega \frac{R(x)}{|x-y|^{n-2}}\,dx = \text{polynomial for } y\in\Omega \Rightarrow \Omega \text{ is an ellipsoid.}$$

PROPOSITION 1.6. *Let $\Omega\subset\mathbb{R}^n$ be a bounded domain and suppose Q is a polynomial such that*

$$Q\equiv 0 \quad \text{on } \partial\Omega. \tag{11}$$

Then the Newtonian potential of Ω with density ΔQ is a constant multiple of Q.

Proof. The proof is an easy consequence of the fact that $\Delta_x|x-y|^{n-2} = c\delta_y$ where δ is the Dirac distribution, and that $Q\equiv 0$ (on $\partial\Omega$). Thus

$$\int_\Omega \frac{\Delta Q}{|x-y|^{n-2}}\,dx = \omega_n Q(y), \qquad y\in\Omega.$$

ω_n is defined in the notation.

2. Application to Cauchy's problem for the Laplace equation. In [KS] the authors conjecture the following.

CONJECTURE [KS]. Let Γ be an analytic hypersurface in $\mathbb{R}^n$ whose Schwarz potential $U_\Gamma(x)$ is real analytic in the domain Ω. Then, for any polynomial $R(x) = R(x_1,\cdots,x_n)$ the solution of the following initial value problem can be continued analytically into Ω:

$$\begin{aligned} &\Delta u = 0 \quad \text{near } \Gamma,\\ &u\equiv R \quad \text{on } \Gamma. \end{aligned} \tag{12}$$

Here we want to prove this conjecture in some special cases, namely $\Gamma=\partial\Omega$, and Ω is such that there exists an increasing sequence of ellipsoids E_i with

$$\Omega = \mathbb{R}^n \Big\backslash \bigcup_{i=1}^{\infty} E_i.$$

(In particular, this holds if Γ is an ellipsoid.) It is not difficult to show that $\partial\Omega$ is a hyperplane or a (elliptic or parabolic) quadratic hypersurface if $\partial\Omega$ is not empty.

First we prove the following lemma.

LEMMA 2.1. *Let E_i be an increasing sequence of ellipsoids such that*

$$\mathbb{R}^n \neq \overline{\bigcup_{i=1}^{\infty} E_i}.$$

Set $\Omega' := \bigcup_{i=1}^{\infty} E_i$ and let (without loss of generality) $0^\delta \notin \bar{\Omega}'$. Then the generalized Newtonian potential of Ω' with respect to $a=0$ with polynomial density R is a polynomial of degree deg $R+2$, *in the interior of Ω'.*

Remark. Since $K_m^a(x,y) = O(|y|^m/|x|^{m+n-2})$ for $|x|$ large, the generalized Newtonian potential is well defined.

Proof. Define $N_i(y) = \int_{E_i} K_m(x,y)R(x)\,dx$ and $N(y) = \int_\Omega K_m(x,y)R(x)\,dx$. Then by Theorem 1.3, $N_i(y)$ is a polynomial in E_i and $N_i(y)$ converges uniformly to $N(y)$ on each compact subset of Ω'. Thus, $N(y)$ is a polynomial in the interior of Ω'.

THEOREM 2.2. *Let E_i, $i=1,2,\cdots$, and Ω' be as in Lemma* 2.1. *Then the solution of* (12) *is real analytic in $\Omega := \mathbb{R}^n\backslash\Omega'$.*

Proof (without loss of generality $0 \notin \bar{\Omega}'$). Let R be the polynomial in (12). Then, by Lemma 2.1, the generalized Newtonian potential of Ω' with respect to zero with density ΔR is a polynomial in the interior of $\mathbb{R}^n \setminus \Omega = \Omega'$. Define Q to be the polynomial such that

$$Q(y) = \int_{\Omega'} K_m(x, y) \Delta R \, dx \quad \forall y \in \Omega' = \mathbb{R}^n \setminus \Omega.$$

Set $u_0 = Q(y) - \int_{\Omega'} K_m(x, y) \Delta R \, dx$ in Ω, and let u be the solution of (12) with $\Gamma = \partial\Omega$, i.e.,

$$\begin{aligned} \Delta u &= 0 \quad \text{near } \partial\Omega, \\ u &\equiv R \quad \text{on } \partial\Omega. \end{aligned}$$

Since $Q(y) \equiv \int_{\Omega'} K_m(x, y) \Delta R(x) \, dx$ on $\partial\Omega$ we have

$$\begin{aligned} \Delta u_0 &= \Delta R \quad \text{in } \Omega, \\ u_0 &\equiv 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{13}$$

Define

$$v = \begin{cases} R - u_0 & \text{in } \Omega, \\ u & \text{in } \Omega' = \mathbb{R}^n \setminus \Omega. \end{cases}$$

Since $u \equiv R - u_0$ on $\partial\Omega$, u is harmonic near $\partial\Omega$ and $R - u_0$ is harmonic in Ω. They are the harmonic continuations of each other (see [Ke, p. 261]), and v will be harmonic in a neighborhood of $\partial\Omega$. Now, by the uniqueness part of the Cauchy-Kowalevsky theorem, $u = v$, that is, $u = R - u_0$ in Ω. Then by (13), $\Delta u = 0$ in Ω, i.e., u is harmonic in Ω and the proof is completed.

Remark. In Theorem 2.2 when $E_i = E$ for all i we obtain $u_0 = Q(y) - \omega_n^{-1} \int_E \Delta R \cdot |x - y|^{2-n} \, dx$, where Q is a polynomial such that

$$Q(y) = \omega_n^{-1} \int_E |x - y|^{2-n} \Delta R(x) \, dx$$

for all $y \in E$. Now define

$$v = \begin{cases} R - u_0 & \text{in } \Omega = \mathbb{R}^n \setminus E, \\ u & \text{in } E. \end{cases}$$

Then as in the above theorem, $u = v$ in $\mathbb{R}^n$, and u_0 is analytic in $\Omega = \mathbb{R}^n \setminus E$, and by Corollary 1.2, in conjunction with Lemmas 1.4 and 1.5, it has a continuation into $E \setminus E_0$. Thus the Schwarz potential of ∂E is harmonic in $\mathbb{R}^n \setminus E_0$.

Remark. For a ball B centered at x_0, we get that the solution of (12) with

Γ = the boundary of B has a harmonic continuation into $\mathbb{R}^n \setminus \{x_0\}$.

3. Concluding remarks. The kernel K_m appears, for a general elliptic operator, earlier in L. Nirenberg and H. F. Walker [NW].

The following are some references mentioned earlier. A. Wangerin, ed., *Über die Anziehung homogener Ellipsoide, Abhandlungen von Laplace* (1782), *Ivory* (1809), *Gauss* (1813), *Chasles* (1838) *and Dirichlet* (1839) [W].

As for the case $n = 2$, the only reason we left it out is the passage of formula (5′) to (6). In (6) we obtain that

$$v(0) = \frac{c_n}{m+1} \int_\lambda^\infty \frac{ds}{\varphi(s)},$$

where $\varphi^2(s) = (a_1^2 + s)(a_2^2 + s)$, i.e., $\varphi^{-1} \notin L^1(\lambda, \infty)$, hence $v(0) = \infty$. Thus this formula ceases to be true for $n = 2$. But at any rate we do have an explicit formula for v_k (which

is (5')). All other results here are true for $n = 2$, and we just have to replace the potential $|x-y|^{2-n}$ by $\log|x-y|^{-1}$.

Acknowledgments. I thank Professor Harold S. Shapiro for suggesting the problem to me and encouraging me during this work. I also thank Lavi Karp for showing me his (unpublished) work on the generalized Newtonian potential and D. Khavinson for suggesting some improvements in an earlier version of this report.

REFERENCES

[Di] L. KRONEKER, *Dirichlet's Werke*, part I, pp. 375-391, part II, pp. 3-17, Chelsea, New York, 1969.

[DF] E. DIBENEDETTO AND A. FRIEDMAN, *Bubble growth in porous media*, Indiana Univ. Math. J., 35 (1986), pp. 573-606.

[Fe] M. N. FERRERS, *On the potentials of ellipsoids, ellipsoidal shells, elliptic laminae, and elliptic rings, of variable densities*, Quart. J. Pure Appl. Math., 14 (1877), pp. 1-22.

[GH] W. GRÖBNER AND N. HOFREITER, *Integraltafel*, Springer-Verlag, Berlin, New York, 1958.

[KS] D. KHAVINSON AND H. S. SHAPIRO, *The Schwarz potential in $\mathbb{R}^n$ and Cauchy's problem for the Laplace equation*, preprint.

[Ke] O. D. KELLOGG, *Foundations of Potential Theory*, Ungar, New York, 1970 (4th printing).

[NW] L. NIRENBERG, AND H. F. WALKER, *The null spaces of elliptic partial differential operator in $\mathbb{R}^n$*, J. Math. Anal. Appl., 42 (1973), pp. 271-301.

[W] A. WANGERIN, ED, *Über die Anziehung homogener Ellipsoide, Abhandlungen von Laplace* (1782), *Ivory* (1809), *Gauss* (1813), *Chasles* (1838) *and Dirichlet* (1839), Ostwalds Klassiker der Exakt. Wiss. Nr. 77, Leipzig, Germany.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1256-1283, September 1991

004

ARTIFICIAL BOUNDARY CONDITIONS FOR INCOMPLETELY PARABOLIC PERTURBATIONS OF HYPERBOLIC SYSTEMS*

LAURENCE HALPERN†

Abstract. Artificial boundary conditions are devised for small incompletely parabolic perturbations of hyperbolic systems, which are local, consistent with the hyperbolic equation, well posed, and produce weak boundary layers. The general strategy is applied to the Navier-Stokes system.

Key words. artificial boundary conditions, Navier-Stokes equations

AMS(MOS) subject classifications. 35B25, 35Q10, 35G10

Introduction. A general model for a fluid motion is the following time-dependent compressible Navier-Stokes system:

$$\frac{\partial \rho}{\partial t} + \operatorname{div} \rho v = 0,$$

$$\frac{\partial \rho v}{\partial t} + \operatorname{div} (\rho v \cdot v + pI) = \rho g + \operatorname{div} \mu\tau,$$

$$\frac{\partial e}{\partial t} + \operatorname{div} (e+p)v = \rho g \cdot v + \operatorname{div} (K \operatorname{grad} T + \mu v \cdot \tau),$$

where ρ represents the density, p the pressure, T the temperature, and v the velocity of the fluid. τ is the momentum flux density tensor due to friction: $\tau = -\frac{2}{3} I \operatorname{div} v + \operatorname{grad} v + (\operatorname{grad} v)^{\tau}$. μ and K are the coefficients of viscosity and heat conductivity, respectively. An equation of state relating ρ, p, and T is added to close the system. Those equations are a special case of a class of equations called incompletely parabolic equations.

Although the mathematical analysis of these nonlinear equations is not entirely satisfactory, and due to the increasing complexity of the physical problems involved, the Navier-Stokes model is more and more widely used in today's computational fluid dynamics.

In many problems of interest, the computational domain is infinite, so that an important task is the design and analysis of reliable numerical boundary conditions. Very often the Euler equations have replaced the Navier-Stokes system in computations (i.e., assuming the viscosity and heat conductivity coefficients negligible). In that case stable boundary conditions are provided by prescribing the entering characteristic quantities (see, for instance, [OS]). For better accuracy strategies were described in [EM1], [EM2], and [BT1], [BT2], [BT3], which led to higher-order differential operators on the boundary.

For the Navier-Stokes system, it is well known that more boundary conditions are needed to ensure the well posedness. Considering the Navier-Stokes equation as a perturbation of the Euler system, it has been suggested that extra boundary conditions

* Received by the editors August 7, 1989; accepted for publication (in revised form) August 28, 1990. This work was completed while the author was visiting the University of California, Berkeley, California, and was supported by Office of Naval Research grant N00014-86-K-0691.

† Ecole Polytechnique, Centre de Mathématiques appliquées, 91128 Palaiseau Cedex, France, and Département de Mathématiques et Informatique, Centre Scientifique et Polytechnique, Université Paris Nord, 93430 Villetaneuse, France.

be added to those derived for the Euler system [OS]. The artificial boundary is usually set in a "smooth" region, where the equations can be linearized about a regular state (in general, it is supposed to be constant). The derivation and analysis can then be carried out for the linear equation. In [GS] boundary conditions were built by adding conditions on the normal derivatives to the "hyperbolic" boundary conditions to produce dissipation. In [RS1] and [RS2] "hyperbolic" boundary conditions were tested for a flow over a flat plate to force the convergence to the steady state. More recently in [ABL] Abarbanel, Bayliss, and Lustman worked directly on the Navier-Stokes equation for the flow past an airplane. They decoupled the domain into the boundary layer region and the hyperbolic region, and in the former region used a modal expansion and an approximation of the solution. This approximation is made in the regime of long wavelength.

We develop here a general strategy for the derivation of artificial boundary conditions for incompletely parabolic perturbations of hyperbolic systems. Because of the remark above we shall consider linear systems with constant coefficients. Using the Fourier transform as an essential tool, we shall write artificial boundary conditions for a half-space in such a way that the well posedness and the convergence to the hyperbolic equation are ensured by the well posedness of a reduced hyperbolic problem. The strategy has been introduced in [H] and [HS] for incompressible flows and consists of expanding the modes in terms of the small parameter ν. For the analysis of these boundary conditions we shall rely on the results by Strikwerda in [S] on the well posedness of incompletely parabolic systems, and by Michelson in [M] on the boundary layer expansion and convergence to the "inviscid" equation. This strategy theoretically allows for a convergence up to any accuracy, but the well posedness is not guaranteed (note that in the hyperbolic case, no well-posedness proof is available for general artificial boundary condition; see [EM1]).

Consider an incomplete singular perturbation of a hyperbolic system, i.e.,

$$\frac{\partial w}{\partial t} = \sum_{j=1}^{N} A^{(j)} \frac{\partial w}{\partial x_j} + \nu \sum_{j,k=1}^{N} P^{(jk)} \frac{\partial^2 w}{\partial x_j \, \partial x_k} + F(x, t), \tag{0.1}$$

where the $n \times n$ matrices $P^{(jk)}$ are assumed of the form

$$P^{(jk)} = \begin{pmatrix} \bar{P}^{(jk)} & 0 \\ 0 & 0 \end{pmatrix} \tag{0.2}$$

with rank $\bar{P}^{(jk)} = r$, $\bar{P}^{(jk)}$ is nonsingular, and $P^{(jk)} = P^{(kj)}$. The matrices $A^{(j)}$ are partitioned in the same way:

$$A^{(j)} = \begin{pmatrix} B^{(j)} & C^{(j)} \\ D^{(j)} & \bar{A}^{(j)} \end{pmatrix}. \tag{0.3}$$

We require the operator $\partial_t - \sum_{j=1}^{N} A^{(j)} \partial_j$ to be hyperbolic, the partial operator $\partial_t - \nu \sum_{j=1}^{N} \bar{P}^{(jk)} \partial_{jk}$ to be Petrovski parabolic, and the reduced operator $\partial_t - \sum_{j=1}^{N} \bar{A}^{(j)} \partial_j$ to be strictly hyperbolic. These assumptions ensure the well posedness of the Cauchy problem. In order to consider an initial boundary value problem in a half-space $x_1 > 0$ or $x_1 < 0$, we shall assume that the boundary $\Gamma = \{x_1 = 0\}$ is noncharacteristic, i.e., that $A^{(1)}$ is nonsingular. Its eigenvalues are denoted by $\lambda_1, \cdots, \lambda_n$ where $\lambda_1, \cdots, \lambda_m$ are negative and $\lambda_{m+1}, \cdots, \lambda_n$ are positive. The corresponding eigenvectors are $\Lambda^1, \cdots, \Lambda^n$. For convenience and simplicity, we shall assume that $\bar{A}^{(1)}$ is a diagonal matrix, with p negative eigenvalues:

$$\bar{A}^{(1)} = \begin{pmatrix} \bar{A}^{(1)^-} & 0 \\ 0 & \bar{A}^{(1)^+} \end{pmatrix} \tag{0.4}$$

where

$$\bar{A}^{(1)^-} = \begin{pmatrix} \bar{\lambda}_{1+r} & & \\ & \ddots & \\ & & \bar{\lambda}_{p+r} \end{pmatrix} < 0, \qquad \bar{A}^{(1)^+} = \begin{pmatrix} \bar{\lambda}_{p+r+1} & & \\ & \ddots & \\ & & \bar{\lambda}_{n} \end{pmatrix} > 0.$$

We further assume the existence of a symmetrizer S for the full operator

$$Q = \sum_{j=1}^{N} A^{(j)} \frac{\partial}{\partial x_j} + \nu \sum_{j=1}^{N} P^{(jk)} \frac{\partial^2}{\partial x_j \, \partial x_k}, \tag{0.5}$$

which implies in particular that the symbol of Q,

$$Q(i\xi) = i \sum_{j=1}^{N} A^{(j)} \xi_j - \nu \sum_{j=1}^{N} P^{(jk)} \xi_j \xi_k, \tag{0.6}$$

is diagonalizable through a transformation analytic in ξ. S is a symmetric positive-definite matrix. We shall denote by $\tilde{A}^{(j)}$ and $\tilde{P}^{(jk)}$ the symmetrized matrices $\tilde{A}^{(j)} = SA^{(j)}$, $\tilde{P}^{(jk)} = SP^{(jk)}$. Both the Navier-Stokes and shallow-water systems fulfill all the conditions above.

In § 1 we shall recall the modal analysis for the Cauchy problem. Most results in this section are known (see, for instance, [YS] for Navier-Stokes, [S] for the general case), but we need to set our notation clearly.

In § 2 we derive the local and nonlocal boundary conditions for a half-space. The transparent boundary condition is first written in terms of generalized eigenvalues and eigenfunctions for the system. It is then approximated with respect to the small parameter ν we shall call viscosity for obvious reasons. This yields boundary conditions that are differential of first order in the normal direction, but still integral in time and the tangential derivatives (like the transparent boundary condition for the pure hyperbolic problem). Those boundary conditions are, in turn, approximated by differential operators which are of order zero in time and one in the tangential direction, using the strategy in [EM1].

In § 3, necessary and sufficient conditions for the well posedness of the corresponding initial boundary value problem are set. The same conditions ensure the convergence to the unperturbed hyperbolic problem, with an error estimate. These results are a direct application of the general analysis in [M].

In § 4 the construction above is carried out explicitly for the two-dimensional compressible Navier-Stokes system.

Finally in § 5, we indicate how to produce more accurate boundary conditions. For the sake of clarity, explicit calculations are made in the special case of the two-dimensional linearized shallow water equation. Nevertheless, the construction carries over to any incompletely parabolic system provided the diagonalizability assumption (0.6) is fulfilled.

1. The Cauchy problem.

1.1. Normal modes for the Cauchy problem. The following analysis can be partly found in [S], but we include it here in order to set our notation and to study more particularly the eigenmodes as functions of the parameter ν.

The normal modes are the solutions of (0.1) with $F \equiv 0$, of the type

$$w = e^{st + \xi x_1 + i\eta \cdot y} \Phi, \qquad \text{Re } s \geqq 0,$$

where

$$x=(x_1,\cdots,x_N),\qquad y=(x_2,\cdots,x_N).$$

They satisfy the equation

$$(Q(\xi, i\eta)-sI)\Phi=0. \tag{1.1}$$

Here s and $i\eta$ are the independent variables and ξ is considered as a function of (s,η). The equation in ξ is of order $n+r$. By an abuse of notation, but for simplicity, we shall often refer to ξ as a "generalized eigenvalue."

We shall first need a general lemma on matrices.

LEMMA 1.1. *Let M and S be two matrices of same order n. If M is nonnegative, if S is symmetric positive definite, and moreover if SM is symmetric, then SM is nonnegative.*

Let us recall that a matrix M is nonnegative if for any u, $(Mu,u)\geqq 0$, where $(\cdot,\cdot)$ denotes the usual scalar product. M is positive definite if there exists a constant $\alpha>0$ such that for any u, $(Mu,u)\geqq\alpha\|u\|^2$.

Proof. For convenience, we choose a basis where S is diagonal: $S=\operatorname{diag}(s_1,\cdots,s_n)$, $M=(m_{ij})$. Using the identity $SM=M^TS$, we can write $(Mu,u)=(Nv,v)$, where $u=Sv$, and N is defined by

$$n_{ij}=\frac{1}{2}s_is_j\left(1+\frac{s_i}{s_j}\right)m_{ij},\qquad i\leqq j,$$

$$n_{ij}=n_{ji},\qquad i>j.$$

On the other hand, we can express SM as

$$(SMu,u)=2(Cu,u),$$

where

$$c_{ij}=\frac{1}{s_i+s_j}n_{ij}.$$

We know that a matrix defines a nonnegative bilinear form if and only if all the principal minors are nonnegative. For the matrix $1/(s_i+s_j)$, they are equal for $k\leqq n$ to

$$\frac{\prod_{i<j\leqq k}(s_i-s_j)^2}{2^k\Pi s_i\prod_{i<j\leqq k}(s_i+s_j)^2}.$$

We now use the following classical result: if A and B are two symmetric nonnegative matrices, the matrix whose general term is $a_{ij}b_{ij}$ is also nonnegative. Hence C is nonnegative and the proof is complete.

LEMMA 1.2. *For* $\operatorname{Re} s>0$, *there are precisely $(r+p)$ generalized eigenvalues ξ with a negative real part, and $n-p$ with a positive real part.*

Proof. We first prove that there are no purely imaginary generalized eigenvalues. Let us assume that $\xi=i\zeta$, where ζ is a real number. We apply S to (1.1) and multiply by Φ. We now take the real part of the Hermitian product:

$$\nu\left(\left(\sum\tilde{P}^{(jk)}\psi_j\psi_k\right)\Phi,\Phi\right)+\operatorname{Re}s(S\Phi,\Phi)=0,\qquad \psi=(\zeta,\eta_2,\cdots,\eta_n).$$

By assumption, the matrix $\sum P^{(jk)}\psi_j\psi_k$ is nonnegative, and so is $\sum\tilde{P}^{(jk)}\psi_j\psi_k$ by Lemma 1.1. If $\operatorname{Re}s>0$, since $(S\Phi,\Phi)$ is positive, we obtain the contradiction.

Now let N_+ be the number of eigenvalues ξ such that $\operatorname{Re}\xi<0$, and N_- the number of eigenvalues ξ such that $\operatorname{Re}\xi>0$. We have $N_+ + N_- = r+n$, from above. N_+ and N_- are constant functions of η and s. Furthermore, they are constant functions of A and P, as long as, for instance, $P^{(11)} \neq 0$.

Let us choose $\eta = 0$ and $B^{(1)} = C^{(1)} = D^{(1)} = 0$. Equation (1.1) for ξ reads

$$\det[\nu P^{(11)}\xi^2 + A^{(1)}\xi - sI] = 0 \quad \text{or}$$

$$\det[\nu \bar{P}^{(11)}\xi^2 - sI] \cdot \det[\bar{A}^{(1)}\xi - sI] = 0.$$

N_+ (respectively, N_-) is then the number of solutions with positive real part (respectively, negative) of the equations

$$\det[\nu \bar{P}^{(11)}\xi^2 - sI] = 0, \qquad \det[\bar{A}^{(1)}\xi - sI] = 0.$$

The first equation is even in ξ. Hence there are r solutions with positive real part and r with negative real part. Moreover, the second equation reduces to $\xi = s/\bar{\lambda}_i$, which gives p values with a negative real part and $n-r-p$ with a positive real part. So

$$N^+ = r+p, \qquad N^- = r+n-r-p = n-p.$$

We now turn to the behavior of these generalized eigenvalues ξ as the parameter ν tends to zero.

THEOREM 1.1. *If* $\operatorname{Re} s>0$, *as* ν *tends to zero,* r *values of* ξ *tend to infinity as* $1/\nu$, *and* n *values have a finite limit.*

Proof. Let us write $\xi = \alpha + 0(\nu)$, where α is a solution to

$$\det\left[A^{(1)}\alpha + \sum_{j\neq 1} iA^{(j)}\eta_j - sI\right] = 0. \tag{1.2}$$

By assumption this equation has n solutions, denoted by $\alpha_1(s,\eta), \cdots, \alpha_n(s,\eta)$, and

$$\begin{aligned} &1 \leqq j \leqq m, \qquad \operatorname{Re}\frac{\alpha_j}{s} \leqq 0, \\ &m+1 \leqq j \leqq n, \qquad \operatorname{Re}\frac{\alpha_j}{s} \geqq 0 \end{aligned} \tag{1.3}$$

to any α_j is associated an eigenvector $\Pi^j(s,\eta)$, and $\Pi^1, \cdots, \Pi^n$ span $\mathbf{R}^n$

$$\left(A^{(1)}\alpha_k + \sum_{j\neq 1} iA^{(j)}\eta_j - sI\right)\Pi^k = 0. \tag{1.4}$$

If now $\xi = \theta/\nu + O(1)$, θ is a solution of

$$\det(P^{(11)}\theta + A^{(1)}) = 0, \tag{1.5}$$

which is an equation of degree r in θ, and has r roots $\theta_1 \cdots \theta_r$, then $\theta_1, \cdots, \theta_{r+p-m}$ are such that $\operatorname{Re}\theta_j < 0$, and $\theta_{r+p-m+1}, \cdots, \theta_r$ are such that $\operatorname{Re}\theta_j > 0$. The corresponding generalized eigenvectors $\Theta^1 \cdots \Theta^r$ are defined by

$$(P^{(11)}\theta_j + A^{(1)})\Theta^j = 0. \tag{1.6}$$

In summary, every solution (ξ, Φ) of $(Q(\xi, i\eta) - sI)\Phi = 0$ for $\operatorname{Re} s>0$ is such that either

$$\xi(s,\eta,\nu) = \alpha(s,\eta) + O(\nu), \qquad \Phi(s,\eta,\nu) = \Pi(s,\eta) + O(\nu),$$

where

$$\left[A^{(1)}\alpha+\sum_{j\neq 1} i\eta_j A^{(j)}-sI\right]\Pi=0$$

or

$$\xi(s,\eta,\nu)=\frac{1}{\nu}\theta+O(1),\qquad \Phi(s,\eta,\nu)=\Theta+O(\nu),$$

where

$$(P^{(11)}\theta+A^{(1)})\Theta=0.$$

We shall denote by $\xi_1\cdots\xi_m$ the values of ξ of the first form with negative real part, i.e., corresponding to "propagating modes," and $\xi_{m+1}\cdots\xi_{r+p}$ the values of ξ of the second form with negative real parts. We define $\zeta_j(s,\eta)$ and $\psi_j(s,\eta)$ as follows:

$$\begin{aligned} &1\leqq j\leqq m, \qquad \xi_j(s,\eta,\nu)=\zeta_j(s,\eta)+O(\nu),\\ &m+1\leqq j\leqq r+p, \qquad \xi_j(s,\eta,\nu)=\frac{1}{\nu}\zeta_j(s,\eta)+O(1) \end{aligned} \tag{1.7}$$

(ζ_j does not actually depend on s and η if $j\geqq m+1$) and

$$1\leqq j\leqq r+p, \qquad \Phi^j(s,\eta,\nu)=\Psi^j(s,\eta)+O(\nu), \tag{1.8}$$

so that

$$\begin{aligned} &1\leqq j\leqq m, \quad \zeta_j(s,\eta)=\alpha_j(s,\eta), \quad \Psi^j(s,\eta)=\Pi^j(s,\eta),\\ &m+1\leqq j\leqq r+p, \quad \zeta_j(s,\eta)=\theta_{j-m}, \quad \Psi^j(s,\eta)=\Theta^{j-m}. \end{aligned} \tag{1.9}$$

1.2. The transmission conditions. Let us first write a weak formulation of (0.1) in a domain Ω with smooth boundary $\partial\Omega$. For any v sufficiently smooth, we multiply (0.1) by S, apply it to v, and integrate on Ω. Using the Green formulas

$$\int_\Omega\left(\tilde{A}^{(j)}\frac{\partial w}{\partial x_j},v\right)+\int_\Omega\left(\tilde{A}^{(j)}w,\frac{\partial v}{\partial x_j}\right)=\int_{\partial\Omega}(\tilde{A}^{(j)}w,v)n_j,$$

$$\sum_{j,k=1}^N\int_\Omega\left(\tilde{P}^{(j,k)}\frac{\partial^2 w}{\partial x_j\,\partial x_k},v\right)+\sum_{j,k=1}^N\int_\Omega\left(\tilde{P}^{(j,k)}\frac{\partial w}{\partial x_j},\frac{\partial v}{\partial x_k}\right)$$
$$=\sum_{j,k=1}^N\int_{\partial\Omega}\left(\tilde{P}^{(j,k)}\frac{\partial w}{\partial x_j},v\right)n_k,$$

where n is the normal exterior to $\partial\Omega$, we get

$$\int_\Omega\left(S\frac{\partial w}{\partial t},v\right)+\nu\sum_{j,k=1}^N\int_\Omega\left(\tilde{P}^{(j,k)}\frac{\partial w}{\partial x_j},\frac{\partial v}{\partial x_k}\right)$$
$$+\frac{1}{2}\sum_{j=1}^N\int_\Omega\left[\left(\tilde{A}^{(j)}\frac{\partial v}{\partial x_j},w\right)-\left(\tilde{A}^{(j)}\frac{\partial w}{\partial x_j},v\right)\right]$$
$$=\nu\int_\Gamma\sum_{j,k=1}^N\left(\tilde{P}^{(jk)}\frac{\partial w}{\partial x_j},v\right)n_k+\frac{1}{2}\int_\Gamma\sum_{j=1}^N(A^{(j)}w,v)n_j+\int_\Omega Fv.$$

We define two bilinear forms a and p, by

$$
\begin{aligned}
a(v, w) &= \frac{1}{2}\int_\Omega \sum_{j=1}^{N}\left[\left(\tilde{A}^{(j)}\frac{\partial v}{\partial x_j}, w\right)-\left(\tilde{A}^{(j)}\frac{\partial w}{\partial x_j}, v\right)\right] dx,\\
p(v, w) &= \int_\Omega \sum_{j,k=1}^{N}\left(\tilde{P}^{(j,k)}\frac{\partial w}{\partial x_j}, \frac{\partial v}{\partial x_k}\right),
\end{aligned}
\tag{1.10}
$$

where a is antisymmetric and p is symmetric, nonnegative (Lemma 1.1).

S being symmetric definite positive, defines a scalar product

$$
s(w, v) = \int_\Omega (Sw, v)\, dx \tag{1.11}
$$

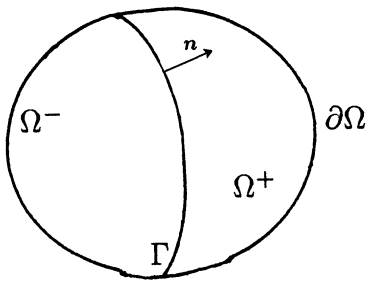
and we can write

$$
s\left(\frac{\partial w}{\partial t}, v\right) + \nu p(v, w) + a(v, w) - \int_{\partial\Omega} (S\mathscr{E}w, v)\, d\gamma = \int_\Omega Fv, \tag{1.12}
$$

where $\mathscr{E}w$ is the normal stress

$$
\mathscr{E}w = \sum_{k=1}^{n}\left(\nu\sum_j P^{(jk)}\frac{\partial w}{\partial x_j} + \frac{1}{2}A^{(k)}w\right)n_k. \tag{1.13}
$$

Suppose now that $\Omega = \Omega^- \cup \Omega^+$,



the orientation of n being from Ω^- to Ω^+. We define a^-, a^+, p^-, p^+, s^-, s^+ as we did for a, p, s, but the integral being taken over Ω^- and Ω^+, respectively. Let v be compactly supported in Ω. We thus have

$$
s\left(\frac{\partial w}{\partial t}, v\right) + \nu p(v, w) + a(v, w) = \int_\Omega Fv, \tag{1.14}
$$

but

$$
\begin{aligned}
p(v, w) &= p^+(v, w) + p^-(v, w),\\
a(v, w) &= a^+(v, w) + a^-(v, w),\\
s\left(\frac{\partial w}{\partial t}, v\right) &= s^+\left(\frac{\partial w}{\partial t}, v\right) + s^-\left(\frac{\partial w}{\partial t}, v\right),
\end{aligned}
$$

and we can write

$$
s^\pm\left(\frac{\partial w^\pm}{\partial t}, v\right) + \nu p^\pm(w^\pm, v) + a^\pm(w^\pm, v) - \int_{\partial\Omega^\pm} (S\mathscr{E}w^\pm, v)\, d\gamma = \int_{\Omega^\pm} F^\pm v, \tag{1.15}^\pm
$$

where $F^\pm = F/\Omega^\pm$ and $w^\pm = w/\Omega^\pm$, so that adding $(1.15)^+$ and $(1.15)^-$, and subtracting

from (1.14) we get

$$\int_{\partial\Omega^+} (S\mathscr{E}w, v)\, d\gamma + \int_{\partial\Omega^-} (S\mathscr{E}w, v)\, d\gamma = 0.$$

Since v is compactly supported in Ω, $\partial\Omega^{\pm}$ reduces to Γ and $(\mathscr{E}w)^- = (\mathscr{E}w^-)|_\Gamma$ if w is defined in Ω^-, $(\mathscr{E}w)^+ = (\mathscr{E}w^+)|_\Gamma$ if w is defined in Ω^+, and the normal on Γ is exterior to Ω^- (thus interior to Ω^+). The transmission conditions then read

$$\begin{aligned} &(\mathscr{E}w)^- = (\mathscr{E}w)^+ \\ &w^- = w^+ \end{aligned} \quad \text{on } \Gamma, \tag{1.16}$$

or

$$\sum_{k=1}^{N} \left(\nu \sum_j P^{(jk)} \frac{\partial w^-}{\partial x_j} + \frac{1}{2} A^{(k)} w^- \right) n_k = \sum_{k=1}^{N} \left(\nu \sum_j P^{(jk)} \frac{\partial w^+}{\partial x_j} + \frac{1}{2} A^{(k)} w^+ \right) n_k. \tag{1.17}$$

In particular, if $\Omega = \mathbf{R}^n$, $\Omega^- = \{x_1 < 0\}$, $\Omega^+ = \{x_1 > 0\}$, so that $\Gamma = \{x_1 = 0\}$, then the transmission conditions are

$$\begin{cases} \nu \sum_{j=1}^{N} \bar{P}^{(j1)} \dfrac{\partial w^-}{\partial x_j} = \nu \sum_{j=1}^{N} \bar{P}^{(j1)} \dfrac{\partial w^+}{\partial x_j} \\ w^- = w^+ \end{cases} \quad \text{on } \Gamma. \tag{1.18}$$

Condition (1.18) is equivalent to (1.16). The Green formula with the constraints is more useful when we want to prove the well posedness through energy estimates, and it is the reason we include it here. Again, (1.18) can also be written

$$w^- = w^+, \qquad \frac{\partial w^{-I}}{\partial x_1} = \frac{\partial w^{+I}}{\partial x_1}, \tag{1.19}$$

where (w^I, w^{II}) corresponds to the decomposition of the matrices $P^{(jk)}$, i.e.,

$$w^I = (w_1, \cdots, w_r), \qquad w^{II} = (w_{r+1}, \cdots, w_n), \tag{1.20}$$

but we prefer to use the form (1.18), which seems more fitted to the multidimensional case.

2. Derivation of the artificial boundary conditions. We shall use the transmission conditions we wrote above to derive the transparent boundary condition. Let F and w^0 be compactly supported in Ω^-; consider the Cauchy problem:

$$\begin{cases} \dfrac{\partial w}{\partial t} = \sum_{j=1}^{N} A^{(j)} \dfrac{\partial w}{\partial x_j} + \nu \sum_{j,k=1}^{N} P^{(jk)} \dfrac{\partial^2 w}{\partial x_j\, \partial x_k} + F(x, t), \\ w(0) = w^0. \end{cases} \tag{2.1}$$

It is equivalent to the transmission problem

$$\begin{cases} \dfrac{\partial w^-}{\partial t} - Qw^- = F(x, t) \quad \text{in } \Omega^-, \\ w^-(0) = w^0, \end{cases} \tag{2.2}$$

$$\begin{cases} \dfrac{\partial w^+}{\partial t} - Qw^+ = 0 \quad \text{in } \Omega^+, \\ w^+(0) = 0 \end{cases} \tag{2.3}$$

with the transmission conditions

$$\begin{cases} \nu \sum_1^N \bar{P}^{(j1)} \dfrac{\partial w^-}{\partial x_j} = \nu \sum_1^N \bar{P}^{(j1)} \dfrac{\partial w^+}{\partial x_j} \\ w^- = w^+ \end{cases} \quad \text{on } \Gamma. \tag{2.4}$$

2.1. The transparent boundary condition. We introduce the initial boundary value problem in Ω^+:

$$\frac{\partial w}{\partial t} - Qw = 0 \quad \text{in } \Omega^+, \tag{2.5a}$$

$$w(t=0) = 0, \tag{2.5b}$$

$$\begin{pmatrix} w_1 \\ \vdots \\ w_{r+p} \end{pmatrix} = g \quad \text{on } \Gamma. \tag{2.5c}$$

THEOREM 2.1. *The boundary value problem* (2.5) *is strongly well posed. The solution is given in Fourier variables* (η, s) *by*

$$\hat{w}(x_1, \eta, s) = \sum_{i=1}^{r+p} \lambda_i e^{\xi_i x_1} \Phi^i, \tag{2.6}$$

where (ξ_i, Φ^i) *are defined in* (1.1) *and the coefficients* λ_i *are determined by the boundary conditions.*

Proof. According to Strikwerda [S], the problem is strongly well posed if and only if the two initial boundary value problems

$$\begin{cases} \dfrac{\partial w^I}{\partial t} = \nu \sum_{j,k=1}^N \bar{P}^{(jk)} \dfrac{\partial^2 w^I}{\partial x_j \, \partial x_k}, \\ w^I = g^I \quad \text{on } \Gamma, \end{cases}$$

and

$$\begin{cases} \dfrac{\partial w^{II}}{\partial t} = \sum_{j=1}^N \bar{A}^{(j)} \dfrac{\partial w^{II}}{\partial x_j}, \\ \begin{pmatrix} w_{r+1} \\ \vdots \\ w_{r+p} \end{pmatrix} = \begin{pmatrix} g_{r+1} \\ \vdots \\ g_{r+p} \end{pmatrix} \quad \text{on } \Gamma \end{cases}$$

are strongly well posed. The first problem is a strongly parabolic problem with a Dirichlet boundary condition, and hence is strongly well posed. As for the second one, since $\bar{A}^{(j)}$ is diagonal, the boundary condition reduces to specifying the entering characteristics which is, again, a strongly well posed problem.

Let us now consider Fourier-Laplace transform (2.5a) with respect to t and y. The corresponding variables are (s, η), with $\operatorname{Re} s > 0$. We then get a second-order ordinary differential equation, whose solution is

$$\hat{w} = \sum \lambda_i e^{\xi_i x_1} \Phi^i,$$

where (ξ_i, Φ^i) are given in (1.1), since we supposed that $Q(i\xi)$ was diagonalizable. In order for $\hat{w}$ to be in L^2, the coefficients λ_i must vanish when $\operatorname{Re} \xi_i \geqq 0$. We thus are led to (2.6).

Remark. We have assumed that Q was diagonalizable, so the $(\Phi_k^i)_{1 \leqq i,k \leqq r+p}$ is a nonsingular matrix and thus the boundary condition determines the λ_i's.

THEOREM 2.2. *The transparent boundary condition at* Γ *for the half-space* Ω^- *is*

$$\nu \sum_{j=1}^{N} \bar{P}^{(1j)} \frac{\partial \hat{w}^I}{\partial x_j} = \nu \sum_{j=1}^{r+p} \bar{P}^{(11)} \sum_{i=1}^{r+p} \xi_i M_{ij}^{-1} \Phi^{iI} \hat{w}_j + \nu \sum_{l \neq 1} i\eta_l \bar{P}^{(l1)} \hat{w}^I, \tag{2.7a}$$

$$\hat{w}_k = \sum_{i=1}^{r+p} \sum_{j=1}^{r+p} M_{ij}^{-1} \hat{w}_j \Phi_k^i, \qquad k = r+p+1, \cdots, n, \tag{2.7b}$$

where (M_{ij}) *is the* $(r+p)\times(r+p)$ *matrix defined by*

$$M_{ij} = \Phi_i^j, \quad \textit{and} \quad M^{-1} \textit{ is the inverse of } M. \tag{2.8}$$

Proof. w^+ is the solution in Ω^+ of the initial boundary problem (2.5) with $g_k = w_k^-$, $1 \leqq k \leqq r+p$. Theorem 2.1 then enables us to calculate explicitly w_k^+, $r+p+1 \leqq k \leqq n$ and $\partial w^{+I}/\partial x_j$. The transmission conditions then give the result

$$\nu \sum_{j=1}^{N} \bar{P}^{(1j)} \frac{\partial \hat{w}^I}{\partial x_j} = \nu \bar{P}^{(11)} \frac{\partial \hat{w}^I}{\partial x_1} + \nu \sum_{j \neq 1} \bar{P}^{(1j)} i\eta_j \hat{w}^I.$$

From (2.6) we deduce that

$$\frac{\partial \hat{w}^+}{\partial x_1} = \sum_{i=1}^{r+p} \lambda_i \xi_i \Phi^i \quad \text{on } \Gamma,$$

so that

$$\nu \sum_{j=1}^{N} \bar{P}^{(1j)} \frac{\partial \hat{w}^{+I}}{\partial x_j} = \nu \bar{P}^{(11)} \sum_{j=1}^{r+p} \lambda_j \xi_j \Phi^{jI} + \nu \sum_{j \neq 1} \bar{P}^{(1j)} i\eta_j \sum_{i=1}^{r+p} \lambda_i \Phi^{iI},$$

$$\hat{w}_k^+ = \sum_{i=1}^{r+p} \lambda_i \Phi_k^i, \qquad k = r+p+1, \cdots, n.$$

The coefficients λ_j are determined by

$$\sum_{i=1}^{r+p} \lambda_i \Phi_j^i = \hat{w}_j^-, \qquad j = 1, \cdots, r+p.$$

So, if the matrix M is defined by (2.8), we have

$$\lambda_i = \sum_{j=1}^{r+p} M_{ij}^{-1} \hat{w}_j^-$$

and finally $\hat{w}^-$ satisfies (2.7a), (2.7b): (2.7a), (2.7b) is actually the transparent boundary condition.

Remark. If $\bar{A}^{(1)}$ were not diagonal, the same study could be carried over, by choosing an admissible boundary condition (2.5c).

Remark. If $n = r+p$, the transparent boundary condition reduces to (2.7a): the "hyperbolic" part does not require any boundary conditions. (It is the case for instance for Navier-Stokes equation when the flow is supersonic and the boundary is on the outflow.)

2.2. Nonlocal approximate boundary condition. Since we are seeking boundary conditions that are consistent with the hyperbolic problem (i.e., $\nu = 0$), we shall approximate the transparent boundary condition (2.7) with respect to the parameter ν. We thus shall obtain boundary condition relating w and $\partial w^I/\partial x_1$, whose kernel is a nonrational function of s and η, and thus integral in the time variable and the boundary variables. This boundary condition will eventually be approximated by local

boundary conditions in § 2.3, using the techniques in [EM1] for hyperbolic problems.

Let us consider the limits as ν tends to zero of the various terms in the right-hand side of (2.7). By (1.8) the vectors Φ^i tend to the corresponding Ψ^i, and hence the matrix M tends to N, where

$$N_{ij} = \Psi^j_i, \qquad 1 \leqq i, j \leqq r+p. \tag{2.9}$$

As for the coefficients $\nu\xi_i$, by (1.9) if $1 \leqq i \leqq m$, $\nu\xi_i$ tends to zero, and if $m+1 \leqq i \leqq n$, $\nu\xi_i$ tends to a finite limit $\zeta_i(s, \eta)$ (which actually does not depend on s, η). Taking the limits in the right-hand side of (2.7) as described, we are led to the boundary condition

$$\nu \sum_{j=1}^{N} \bar{P}^{(1,j)} \frac{\partial \hat{w}^I}{\partial x_j} = \sum_{i=m+1}^{r+p} \bar{P}^{(11)} \sum_{j=1}^{r+p} \zeta_i N^{-1}_{ij} \Psi^{iI} \hat{w}_j, \tag{2.10a}$$

$$\hat{w}_k = \sum_{i=1}^{r+p} \sum_{j=1}^{r+p} N^{-1}_{ij} \hat{w}_j \Psi^i_k, \qquad r+p+1 \leqq k \leqq n. \tag{2.10b}$$

We shall see in the next section how this boundary condition leads to a well-posed problem in the left half-space Ω^-, whose solution converges as ν tends to zero toward the restriction to Ω^- of the solution of the full hyperbolic problem in $\mathbf{R}^n$. The latter will be done using a boundary layer expansion and the criterion in [M]. Before carrying over the analysis we shall write local boundary conditions. In order to make the mechanisms clear, and since we shall need it later, we shall first recall the derivation of transparent and approximate boundary conditions for the hyperbolic problem.

2.3. Absorbing boundary conditions for the hyperbolic problem. Here we will follow the lines drawn in [EM1]. We keep the notation and assumptions set in the first section. The hyperbolic system is

$$\frac{\partial w}{\partial t} = \sum_{j=1}^{N} A^{(j)} \frac{\partial w}{\partial x_j} + F. \tag{2.11}$$

By Laplace-Fourier transform in t and y, the solutions of this equation in the full space when $F = 0$ are given by

$$\hat{w} = \sum_{i=1}^{N} \lambda_i e^{\alpha_i x_1} \Pi^i, \tag{2.12}$$

where (α_i, Π^i) are the eigenvalues and eigenvectors defined in § 1

$$\left(A^{(1)} \alpha_k + \sum_{j \neq 1} i A^{(j)} \eta_j - sI \right) \Pi^k = 0. \tag{2.13}$$

If $(\operatorname{Re} s \operatorname{Re} \alpha_j) \leqq 0$ (respectively, $(\operatorname{Re} s \operatorname{Re} \alpha_j) \geqq 0$), the corresponding mode in (2.12) propagates in the $(x_1 > 0)$-direction (respectively, $x_1 < 0$).

The transparent boundary condition at $x_1 = 0$ for the half-space Ω^- expresses that no wave can propagate from the boundary toward the interior of Ω^-, i.e.,

$$i = m+1, \cdots, n, \qquad \lambda_i = 0. \tag{2.14}$$

Let us define T as the matrix of the eigenvectors:

$$T_{ij}(s, \eta) = \Pi^j_i(s, \eta). \tag{2.15}$$

By (2.15), (2.14) can be rewritten as

$$\forall i = m+1, \cdots, n \quad (T^{-1} \hat{w})_i = 0. \tag{2.16}$$

This is the transparent boundary condition at $x_1 = 0$ for the half-space Ω^-, i.e., the

equivalent of (2.7) for $\nu = 0$. We shall see in the next section that (2.7) (or (2.10)) actually reduces to (2.16) when ν tends to zero, so that the solution of (0.1) coupled with (2.7) (or (2.10)) tends to the solution of (2.11) with the boundary condition (2.16). This boundary condition is nonlocal in time and space. Following [EM1], we shall make an approximation with respect to the angle of incidence of the wave on the boundary. This is easily achieved by letting $\eta = 0$ in (2.13), so that α_k is nothing but s/λ_k where λ_k is an eigenvalue of $A^{(1)}$, and $\Pi^k = \Lambda^k$.

Thus, the first absorbing boundary condition for (2.11) in Ω^- is

$$\forall i = m+1, \cdots, n \quad (\tilde{T}^{-1} w)_i = 0, \tag{2.17}$$

where

$$\tilde{T}_{ij} = \Lambda_i^j, \qquad 1 \leqq i, \quad j \leqq n,$$

which is simply writing that the entering characteristics of the system are prescribed the value zero on the boundary.

2.4. Local boundary conditions for the full problem. We thus want to make the same kind of approximation in (2.10), and it is now clearer: for $m+1 \leqq j \leqq r+p$, neither ζ_j nor Ψ^j depends on (s, η) and therefore both remain unchanged. For $1 \leqq j \leqq m$, ζ_j and Ψ^j are approximated by s/λ_j and Λ^j, respectively.

We then define the vector $\tilde{\Psi}^j$, $1 \leqq j \leqq r+p$ by

$$\begin{aligned} &\tilde{\Psi}^j = \Lambda^j, \qquad 1 \leqq j \leqq m, \\ &\tilde{\Psi}^j = \Psi^j, \qquad m+1 \leqq j \leqq r+p \end{aligned} \tag{2.18}$$

and the matrix $\tilde{N}$ by

$$\tilde{N}_{ij} = \tilde{\Psi}_i^j, \qquad 1 \leqq i, \quad j \leqq r+p. \tag{2.19}$$

The approximate boundary condition takes the form

$$\begin{cases} \nu \sum_{j=1}^{N} \bar{P}^{(1,j)} \dfrac{\partial w^I}{\partial x_j} = \sum_{i=m+1}^{r+p} \bar{P}^{(11)} \sum_{j=1}^{r+p} \zeta_i \tilde{N}_{ij}^{-1} \Psi^{i^I} w_j, \\ w_k = \sum_{i=1}^{r+p} \sum_{j=1}^{r+p} \tilde{N}_{ij}^{-1} \tilde{\Psi}_k^i w_j, \qquad r+p+1 \leqq k \leqq n. \end{cases} \tag{2.20}$$

This is our local boundary condition. It is of first order in x, and zero order in time. For (2.7) and (2.16), we shall see that it converges to (2.17) when ν tends to zero, so that the solution of (0.1) coupled with (2.20) tends to the solution of (2.11) coupled with the first absorbing boundary condition (2.17).

We shall discuss in § 5 further approximations to these boundary conditions, with respect to either parameters ν or the angle of incidence on the boundary. These boundary conditions will involve higher derivatives in time and the tangential variables.

3. Analysis of the approximate boundary conditions. We shall use here the analysis by Michelson [M] of the well posedness and boundary layer for initial boundary value problems related to parabolic perturbations of hyperbolic equations.

3.1. Well posedness of the boundary value problems. As already pointed out by Strikwerda [S], the well posedness of the initial boundary value problem for (0.1) is equivalent to the well posedness of the purely parabolic problem for $\bar{P}$ and purely hyperbolic problem for $\bar{A}$, provided the boundary conditions satisfy certain decoupling conditions, which are automatically satisfied for boundary conditions of the form (2.7).

Furthermore, if the problem satisfies a uniform Lopatinski condition stated by Michelson in [M], then we can get estimates uniform in ν. Let us define in Ω^- the weighted norms:

$$\|u\|^2_{m_1,m_2,\eta} = \sum_{\substack{|\beta|\leqq m_2 \\ |\alpha|+|\beta|\leqq m_1}} \|(\nu D, \nu\eta)^\alpha (\chi D_{x_1}, D_y, D_t, \eta)^\beta e^{-\eta t} u\|^2_{L^2}, \tag{3.1}$$

where $\chi = \chi(x_1)$ is a fixed smooth nondecreasing function of x_1 such that $\chi(x_1) = x_1$ for $|x_1| < \frac{1}{2}$, and $\chi(x_1) = 1$ for $|x_1| > 1$. Denote by $|u(x_1, \cdot)|^2_{m_1,m_2,\eta}$ the obvious restriction of the above norm to the hyperplane $x_1 = \text{const}$. Let σ be the pseudodifferential operator with symbol $\operatorname{Re}(1+\nu s+|\nu\eta|^2)^{1/2}$ $(s = i\omega + \eta)$. If w is partitioned in the natural way mentioned before $w = \binom{w^I}{w^{II}}$, we define v by

$$v = \begin{pmatrix} v^I \\ v^{II} \end{pmatrix}, \quad v^{II} = w^{II}, \quad v^I = \begin{pmatrix} \sigma^{-1}\nu D_{x_1} w^I \\ \sigma w^I \end{pmatrix}. \tag{3.2}$$

We start by writing the decoupled problems.

The parabolic problem is

$$\frac{\partial w^I}{\partial t} = \sum_{j,k=1}^{N} \bar{P}^{(jk)} \frac{\partial^2 w^I}{\partial x_j \, \partial x_k} + F^I, \tag{3.3a}$$

$$\begin{aligned} &\sum_{j=1}^{N} \bar{P}^{(1j)} \frac{\partial w^I}{\partial x_j} = 0 \quad \text{on } \Gamma, \\ &w^I(t=0) = 0, \end{aligned} \tag{3.3b}$$

and the hyperbolic problem is

$$\frac{\partial w^{II}}{\partial t} = \sum_{j=1}^{N} \bar{A}^{(j)} \frac{\partial w^{II}}{\partial x^j} + F^{II}, \tag{3.4a}$$

$$\begin{aligned} &\hat{w}_k = \sum_{i=1}^{r+p} \sum_{j=r+1}^{r+p} N^{-1}_{ij} \hat{w}_j \Psi^i_k, \qquad k = r+p+1, \cdots, n \quad \text{on } \Gamma, \\ &w^{II}(t=0) = 0, \end{aligned} \tag{3.4b}$$

for the boundary conditions (2.10). For the boundary conditions (2.20) N and Ψ must be replaced by $\tilde{N}$ and $\tilde{\Psi}$, respectively.

Then we have Theorem 3.1.

THEOREM 3.1. *The boundary value problem* (0.1) *coupled with either boundary condition* (2.10) *or* (2.20) *is well posed if and only if the reduced hyperbolic problem* (3.4) *is well posed. Furthermore, if* (3.4) *is well posed in the sense of Kreiss, let integers* $m_1 \geqq m_2 \geqq 0$ *be such that* $m_1 - m_2 \geqq 1$. *Then there exist positive constants* k, ν_0, η_0 *such that for all* $\eta > \eta_0$ *and* $0 \leqq \nu < \nu_0$ *the following a priori estimate holds*:

$$\begin{aligned} &\eta\|w\|^2_{m_1,m_2,\eta} + \nu\|D_x w^I\|^2_{m_1,m_2,\eta} + |v(0,\cdot)|^2_{m_1,m_2,\eta} + |\sigma^{1/2} v^I(0,\cdot)|^2_{m_1,m_2,\eta} \\ &\quad + \sup_{x_1} \left(|w(x_1,\cdot)|^2_{m_1-1,m_2-1,\eta} + |\nu D_x w^I(x_1,\cdot)|^2_{m_1-1,m_2-1,\eta}\right) \\ &\quad \leqq k\eta^{-1}\|F\|^2_{m_1,m_2,\eta}. \end{aligned} \tag{3.5}$$

Proof. The first assertion is a mere consequence of the result by Strikwerda in [S]. As for the second a priori estimate, it follows directly from the general theory on parabolic perturbations for hyperbolic systems by Michelson in [M].

The a priori estimate justifies the boundary layer expansion and proves the

convergence of the initial boundary value problems (0.1) coupled with either of the boundary conditions (2.10) or (2.11) as described in § 3.2.

3.2. Boundary layer; Convergence results. A physical phenomenon related to incompletely parabolic approximations of hyperbolic equations with a small parameter ν is the formation of a boundary layer. It is mathematically represented by a formal expansion

$$w(x, t, \nu) = \sum_{i \geqq 0} \nu^i \left(w_i^{(1)}(x, t) + w_i^{(2)}\left(\frac{x_1}{\nu}, y, t\right)\right). \tag{3.6}$$

The functions $w_i^{(1)}$ represent the smooth part of the solution, while the functions $w_i^{(2)}$ represent the boundary layer: they are exponentially decreasing in x_1/ν. Michelson proved in [M] that under the same hypothesis as in Theorem 3.1, the expansion (3.6) was actually valid. We shall apply this result to our particular case.

THEOREM 3.2. *Let $w(x, t, \nu)$ be the solution of* (0.1), (2.10) (*respectively*, (2.20)) *with a sufficiently smooth F. Suppose, as in Theorem* 3.2, *that the reduced hyperbolic problem* (3.4) *is well posed. Then, as ν tends to zero, w converges to the solution u of the hyperbolic problem* (2.11), (2.16) (*respectively*, (2.17)). *More precisely, if $m_1 \geqq m_2 \geqq m_3 \geqq 0$ and $m_1 - m_2 \geqq 1$, we have*

$$\| w(x, t, \nu) - u(x, t) \|_{m_1, m_2, m_3, \eta} \leqq c(\nu + \nu^{(3/2)-m_3}), \tag{3.7}$$

where the norm above is defined by

$$\| u \|^2_{m_1, m_2, m_3, \eta} = \sum_{i=0}^{m_3} \| D^i_{x_1} u \|^2_{m_1 - i, m_2 - i, \eta}. \tag{3.8}$$

Remarks. (1) This result tells even more about the boundary layer: it says that in expansion (3.6) the first term $w_0^{(1)}$ is indeed the solution of the associated hyperbolic problem, and the first term $w_0^{(2)}$ vanishes: the boundary layer is "weak."

(2) Boundary condition (2.16) is actually the transparent boundary condition for the hyperbolic problem, so that the solution of (0.1), (2.10) converges to the solution of the Cauchy problem for (2.1).

Proof of Theorem 3.2. According to Michelson [M], the following estimate holds:

$$\| w(x, t, \nu) - w_0^{(1)}(x, t) - w_0^{(2)}(x_1/\nu, y, t) \|_{m_1, m_2, m_3, \eta} \leqq c(\nu + \nu^{(3/2)-m_3}). \tag{3.9}$$

We merely need to check that $w_0^{(1)}$ is u and $w_0^{(2)}$ is zero. These terms are obtained by substituting the expansion (3.6) into the equation and the boundary condition, separating the scales x_1 and x_1/ν, and equating to zero the successive coefficients of the resulting series:

From the equation we deduce that $w_0^{(1)}$ and $w_0^{(2)}$ are solutions of the following equations:

$$A^{(1)} w_0^{(2)} + P^{(11)} \frac{\partial w_0^{(2)}}{\partial (x_1/\nu)} = 0, \tag{3.10}$$

$$\frac{\partial w_0^{(1)}}{\partial t} = \sum_{j=1}^{N} A^{(j)} \frac{\partial w_0^{(1)}}{\partial x_j} + F, \tag{3.11}$$

and $w_i^{(1)}$ and $w_i^{(2)}$ are solutions of

$$\frac{\partial w_i^{(1)}}{\partial t} = \sum_{j=1}^{N} A^{(j)} \frac{\partial w_i^{(1)}}{\partial x_j} + \sum_{j,k=1}^{N} P^{(jk)} \frac{\partial^2 w_{i-1}^{(1)}}{\partial x_j \, \partial x_k}, \tag{3.12}$$

$$
\begin{aligned}
&A^{(1)}\frac{\partial w_i^{(2)}}{\partial(x_1/\nu)}+P^{(11)}\frac{\partial^2 w_i^{(2)}}{\partial(x_1/\nu)^2}\\
&=\frac{\partial w_{i-1}^{(2)}}{\partial t}-\sum_{j\neq 1}A^{(j)}\frac{\partial w_{i-1}^{(2)}}{\partial x_j}-2\sum_{j\neq 1}P^{(1j)}\frac{\partial^2 w_{i-1}^{(2)}}{\partial(x_1/\nu)\partial x_j}-\sum_{j,k\neq 1}P^{(jk)}\frac{\partial^2 w_{i-2}^{(2)}}{\partial x_j\,\partial x_k}
\end{aligned}
\tag{3.13}
$$

with the convention that $w_{i-2}^{(2)}\equiv 0$ if $i=1$. We shall assume here the boundary conditions (2.10) are imposed. The calculations are the same for (2.20). For $x_1=0$ we have

$$
\frac{\partial \hat{w}_0^{(2)I}}{\partial(x_1/\nu)}=\sum_{i=m+1}^{r+p}\sum_{j=1}^{r+p}\zeta_i N_{ij}^{-1}\Psi^{iI}(\hat{w}_{0,j}^{(1)}+\hat{w}_{0,j}^{(2)}),
\tag{3.14}
$$

$$
\hat{w}_{0,k}^{(1)}+\hat{w}_{0,k}^{(2)}=\sum_{i,j=1}^{r+p}N_{ij}^{-1}\Psi_k^i(\hat{w}_{0,j}^{(1)}+\hat{w}_{0,j}^{(2)}),\qquad r+p+1\leqq k\leqq n.
\tag{3.15}
$$

For $l\geqq 1$,

$$
\begin{aligned}
&\bar{P}^{(11)}\left[\frac{\partial \hat{w}_l^{(2)I}}{\partial(x_1/\nu)}-\sum_{i=m+1}^{r+p}\sum_{j=1}^{r+p}\zeta_i N_{ij}^{-1}\Psi^{iI}(\hat{w}_{lj}^{(1)}+\hat{w}_{lj}^{(2)})\right]\\
&\quad+\sum_{j\neq 1}\bar{P}^{(1j)}\left(\frac{\partial \hat{w}_{l-1}^{(1)}}{\partial x_j}+\frac{\partial \hat{w}_{l-1}^{(2)}}{\partial x_j}\right)+\bar{P}^{(11)}\frac{\partial \hat{w}_{l-1}^{(1)}}{\partial x_1}=0,
\end{aligned}
\tag{3.16}
$$

$$
\hat{w}_{l,k}^{(1)}+\hat{w}_{l,k}^{(2)}=\sum_{i,j=1}^{r+p}N_{ij}^{-1}\Psi_k^i(\hat{w}_{l,j}^{(1)}+\hat{w}_{l,j}^{(2)}),\qquad r+p+1\leqq k\leqq n.
\tag{3.17}
$$

Let us start with (3.10). From this form, we deduce that $w_0^{(2)}$ is a linear combination of the "exponential modes" defined in (1.6). Here $w_0^{(2)}$ is supposed to be exponentially decreasing in Ω^-, so that

$$
w_0^{(2)}=\sum_{j=r+p-m+1}^{r}\lambda_j\Theta^j e^{\theta_j x_1},\qquad x_1\leqq 0.
$$

We substitute into (3.14), remembering that for $i=m+1,\cdots,r+p$, (ζ_i,Ψ^i) is actually $(\theta_{i-m},\Theta^{j-m})$. We thus get

$$
\sum_{j=r+p-m+1}^{r}\theta_j\lambda_j\Theta^{jI}=\sum_{j=1}^{r+p-m}\sum_{i=1}^{r+p}\theta_j N_{j+m,i}^{-1}(\hat{w}_{0,1}^{(1)}+\hat{w}_{0,i}^{(2)})\Theta^{jI}.
$$

This amounts to stating that there exist coefficients α_k such that $\sum_{k=1}^{r}\alpha_k\Theta^{kI}=0$. It implies that $\sum_{k=1}^{r}\alpha_k\Theta^k=0$, for equation (1.6) can be written

$$
\begin{aligned}
&\bar{P}^{(11)}\theta_k\Theta^{kI}+B^{(1)}\Theta^{kI}+\mathbf{C}^{(1)}\Theta^{kII}=0,\\
&D^{(1)}\Theta^{kI}+\bar{A}^{(1)}\Theta^{kII}=0.
\end{aligned}
$$

And if $\sum\alpha_k\Theta^{kI}=0$, then $\sum\alpha_k D^{(1)}\Theta^{kI}=0$, so that $\sum\alpha_k\bar{A}^{(1)}\Theta^{kII}=0$, and since $\bar{A}^{(1)}$ is nonsingular, the result follows. From the assumptions (0.5) on the operator Q, the Θ^k's are independent, and hence the α_k's vanish for any k.

Then $\lambda_j=0$ for $r+p-m+1\leqq j\leqq r$, and thus $w_0^{(2)}$ vanishes identically in Ω^-. We substitute into (3.14) and (3.15), which indicates that $w_0^{(1)}$ is a solution of the following problem in Ω^-:

$$
\frac{\partial u}{\partial t}=\sum_{j=1}^{n}A^{(j)}\frac{\partial u}{\partial x_j}+F,\qquad x\in\Omega^-
$$

with the boundary conditions

$$
\sum_{j=1}^{r+p}N_{ij}^{-1}\hat{u}_j=0,\qquad i=m+1,\cdots,r+p,
\tag{3.18a}
$$

(3.18b) $$\hat{u}_k = \sum_{i,j=1}^{r+p} N_{ij}^{-1}\Psi_k^i \hat{u}_j, \qquad k = r+p+1, \cdots, n.$$

We will now prove that (3.18) implies the transparent boundary condition (2.16) $(T^{-1}\hat{u})_i = 0$, $m+1 \leqq i \leqq n$:

$$(T^{-1}\hat{u})_i = \sum_{j=1}^{n} T_{ij}^{-1}\hat{u}_j = \sum_{j=1}^{r+p} T_{ij}^{-1}\hat{u}_j + \sum_{j=r+p+1}^{n} T_{ij}^{-1}\hat{u}_j.$$

In the second term of the right-hand side we substitute (3.18b):

$$\sum_{k=r+p+1}^{n} T_{ik}^{-1}\hat{u}_k = \sum_{k=r+p+1}^{n} T_{ik}^{-1} \sum_{l=1}^{r+p}\sum_{j=1}^{r+p} N_{lj}^{-1}\Psi_k^l \hat{u}_j,$$

$$\sum_{k=r+p+1}^{n} T_{ik}^{-1}\hat{u}_k = \sum_{k=1}^{n}\sum_{l=1}^{r+p}\sum_{j=1}^{r+p} N_{lj}^{-1}T_{ik}^{-1}\Psi_k^l \hat{u}_j - \sum_{k=1}^{r+p}\sum_{l=1}^{r+p}\sum_{j=1}^{r+p} N_{lj}^{-1}T_{ik}^{-1}\Psi_k^l \hat{u}_j,$$

but

$$\sum_{l=1}^{r+p} N_{lj}^{-1}\Psi_k^l = \delta_{kj} \quad \text{and} \quad \sum_{k=1}^{r+p}\sum_{l=1}^{r+p}\sum_{j=1}^{r+p} N_{lj}^{-1}\Psi_k^l T_{ik}^{-1}\hat{u}_j = \sum_{j=1}^{r+p} T_{ij}^{-1}\hat{u}_j,$$

so that

$$(T^{-1}\hat{u})_i = \sum_{k=1}^{n}\sum_{l=1}^{r+p}\sum_{j=1}^{r+p} N_{lj}^{-1}T_{ik}^{-1}\Psi_k^l \hat{u}_j.$$

On account of (3.18), the latter reduces to

$$(T^{-1}\hat{u})_i = \sum_{l=1}^{m}\sum_{j=1}^{r+p} N_{lj}^{-1}\hat{u}_j \sum_{k=1}^{n} T_{ik}^{-1}\Psi_k^l.$$

For $1 \leqq l \leqq m$, Ψ^l corresponds to the hyperbolic part of Q, so that $\sum_{k=1}^{n} T_{ik}^{-1}\Psi_k^l = \delta_{il}$ and

$$(T^{-1}\hat{u})_i = 0 \quad \text{for } m+1 \leqq i \leqq n. \qquad \square$$

4. Application to Navier–Stokes equations. We consider here the two-dimensional compressible Navier-Stokes equations:

(4.1a) $$\rho\frac{du_i}{dt} = -\frac{\partial p}{\partial x_i} + \sum_{j=1}^{2}\frac{\partial}{\partial x_j}\left[\mu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \frac{2}{3}\delta_{ij}\operatorname{div} u\right)\right], \qquad i = 1, 2,$$

(4.1b) $$\rho\frac{dc_v T}{dt} = -p \operatorname{div} u + \mu\sum_{i,j=1}^{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \frac{2}{3}\delta_{ij}\operatorname{div} u\right)\frac{\partial u_i}{\partial x_j} + \sum_{j=1}^{2}\frac{\partial}{\partial x_j}\left(k\frac{\partial T}{\partial x_j}\right),$$

(4.1c) $$\frac{d\rho}{dt} = -\rho \operatorname{div} u.$$

Here ρ is the density, u_i is the velocity component, p is the pressure, T is the temperature, μ and k are the coefficients of viscosity and heat conductivity, respectively, and c_v is the specific heat at constant volume. The pressure p is related to ρ and T by $p = \rho RT$, where R is the gas constant. We shall introduce γ as the ratio of specific heats, i.e., $\gamma = c_p/c_v$ (recall that $R = c_p - c_v$), and the Prandtl number of the gas $P_r = \mu c_p/k$. P_r is supposed to be constant here. As usual $d/dt = \partial/\partial t + u_1(\partial/\partial x_1) + u_2(\partial/\partial x_2)$. We shall assume that the artificial boundary is sufficiently far from any turbulent regime, so that we can consider (u, ρ, T, p) as a small perturbation of a smooth regime (u, ρ, T, p). Since in our analysis the lower-order derivatives are not of much importance, and the

results of Michelson carry over to variable coefficients as well by freezing the coefficients, we shall concentrate here on the case where the reference regime is constant as a function of time and space. Let us call $(\tilde{u}, \tilde{\rho}, \tilde{T}, \tilde{p})$ the perturbation. It is a solution of the following problem:

$$\rho\frac{d\tilde{u}_1}{dt}+RT\frac{\partial\tilde{\rho}}{\partial x_1}+\rho R\frac{\partial\tilde{T}}{\partial x_1}=\mu\left[\frac{4}{3}\frac{\partial^2\tilde{u}_1}{\partial x_1^2}+\frac{\partial^2\tilde{u}_1}{\partial x_2^2}+\frac{1}{3}\frac{\partial^2\tilde{u}_2}{\partial x_1\,\partial x_2}\right], \tag{4.2a}$$

$$\rho\frac{d\tilde{u}_2}{dt}+RT\frac{\partial\tilde{\rho}}{\partial x_2}+\rho R\frac{\partial\tilde{T}}{\partial x_2}=\mu\left[\frac{\partial^2\tilde{u}_2}{\partial x_1^2}+\frac{4}{3}\frac{\partial^2\tilde{u}_2}{\partial x_2^2}+\frac{1}{3}\frac{\partial^2\tilde{u}_1}{\partial x_1\,\partial x_2}\right], \tag{4.2b}$$

$$\rho\frac{d\tilde{T}}{dt}+(\gamma-1)\rho T\ \mathrm{div}\ \tilde{u}=\gamma P_r^{-1}\mu\Delta\tilde{T}, \tag{4.2c}$$

$$\frac{d\tilde{\rho}}{dt}+\rho\ \mathrm{div}\ \tilde{u}=0. \tag{4.2d}$$

We shall normalize these equations by redefining $\tilde{\rho}$ as $\tilde{\rho}/\rho$ and introducing the undisturbed kinematic viscosity $\nu=\mu/\rho$. So that the equations can be written in the form (0.1)

$$\frac{\partial U}{\partial t}=A^{(1)}\frac{\partial U}{\partial x_1}+A^{(2)}\frac{\partial U}{\partial x_2}+\nu\sum_{j,k=1}^{2}P^{(jk)}\frac{\partial^2 u}{\partial x_j\,\partial x_k}+F(x,t), \tag{4.3}$$

where $U=(\tilde{u}_1, \tilde{u}_2, \tilde{T}, \tilde{\rho})$

$$A^{(1)}=\begin{pmatrix} -u_1 & 0 & -R & -RT \\ 0 & -u_1 & 0 & 0 \\ -(\gamma-1)T & 0 & -u_1 & 0 \\ -1 & 0 & 0 & -u_1 \end{pmatrix},$$

$$A^{(2)}=\begin{pmatrix} -u_2 & 0 & 0 & 0 \\ 0 & -u_2 & -R & -RT \\ 0 & -(\gamma-1)T & -u_2 & 0 \\ 0 & -1 & 0 & -u_2 \end{pmatrix},$$

$$P^{(11)}=\begin{pmatrix} \frac{4}{3} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \gamma P_r^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$P^{(22)}=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{4}{3} & 0 & 0 \\ 0 & 0 & \gamma P_r^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$P^{(12)}=\begin{pmatrix} 0 & \frac{1}{6} & 0 & 0 \\ \frac{1}{6} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

They satisfy all the assumptions we made in the Introduction. We shall write here the precise formulation for the local boundary conditions (2.20). We shall start by studying

the eigenvalues of $A^{(1)}$, and recalling the corresponding boundary conditions for the Euler equations (cf., for instance, [EM1]).

4.1. The Euler system. It is well known that the eigenvalues of $A^{(1)}$ are

$$\lambda_1 = -u_1 - c, \quad \lambda_2 = \lambda_3 = -u_1, \quad \lambda_4 = -u_1 + c. \tag{4.4}$$

The corresponding eigenvectors are

$$\Lambda^1 = \begin{pmatrix} c \\ 0 \\ (\gamma-1)T \\ 1 \end{pmatrix}, \qquad \Lambda^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$
$$\Lambda^3 = \begin{pmatrix} 0 \\ 0 \\ T \\ -1 \end{pmatrix}, \qquad \Lambda^4 = \begin{pmatrix} c \\ 0 \\ -(\gamma-1)T \\ -1 \end{pmatrix}. \tag{4.5}$$

So that $A^{(1)}$ is diagonalizable and the matrix $\tilde{T}$ in (2.17) is

$$\tilde{T} = \begin{pmatrix} c & 0 & 0 & c \\ 0 & 1 & 0 & 0 \\ (\gamma-1)T & 0 & T & -(\gamma-1)T \\ 1 & 0 & -1 & -1 \end{pmatrix} \tag{4.6}$$

and

$$A^{(1)} = \tilde{T} \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \lambda_3 & \\ 0 & & & \lambda_4 \end{pmatrix} \tilde{T}^{-1}.$$

The number of boundary conditions required by the system at $x_1 = 0$ in Ω^- depends on whether the flow is super or subsonic and the boundary is inflow or outflow, as summarized below.

The **subsonic** case: $|u_1| < c$.

- **inflow:** $u_1 < 0$.

$$\lambda_1 < 0, \lambda_2, \lambda_3, \lambda_4 > 0, \quad m = 1: \quad \textbf{3 boundary conditions.}$$
$$(\tilde{T}^{-1}u)_i = 0, \qquad i = 2, 3, 4.$$

- **outflow:** $u_1 > 0$.

$$\lambda_1, \lambda_2, \lambda_3 < 0, \quad \lambda_4 > 0, \quad m = 3: \quad \textbf{1 boundary condition.}$$
$$(\tilde{T}^{-1}u)_4 = 0.$$

The **supersonic case:** $|u_1| > c$.

- **inflow:** $u_1 < 0$.

$$\lambda_1, \lambda_2, \lambda_3, \lambda_4 > 0, \quad m = 0: \quad \textbf{4 boundary conditions.}$$
$$u_i = 0, \qquad i = 1, 2, 3, 4.$$

- **outflow:** $u_1 > 0$.

$$\lambda_1, \lambda_2, \lambda_3, \lambda_4 > 0, \quad m = 4: \quad \textbf{0 boundary conditions.}$$

The matrix $\tilde{T}^{-1}$ is given by

$$\tilde{T}^{-1}=\begin{pmatrix} \frac{1}{2c} & 0 & \frac{1}{2\gamma T} & \frac{1}{2\gamma} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{\gamma T} & -\frac{\gamma-1}{\gamma} \\ \frac{1}{2c} & 0 & -\frac{1}{2\gamma T} & -\frac{1}{2\gamma} \end{pmatrix} \tag{4.7}$$

so that the boundary conditions are as follows.

- *Subsonic inflow*:

$$\frac{\tilde{u}_1}{c}-\frac{1}{\gamma}\frac{\tilde{T}}{T}-\frac{1}{\gamma}\frac{\tilde{\rho}}{\rho}=0,$$

$$\tilde{u}_2=0, \tag{4.8a}$$

$$\frac{1}{\gamma}\frac{\tilde{T}}{T}-\frac{\gamma-1}{\gamma}\frac{\tilde{\rho}}{\rho}=0.$$

- *Subsonic outflow*:

$$\frac{\tilde{u}_1}{c}-\frac{1}{\gamma}\frac{\tilde{T}}{T}-\frac{1}{\gamma}\frac{\tilde{\rho}}{\rho}=0. \tag{4.8b}$$

- *Supersonic inflow*:

$$\tilde{u}_1=\tilde{u}_2=\tilde{T}=\tilde{\rho}=0. \tag{4.8c}$$

- *Supersonic outflow*:

no boundary condition.

We reintroduced here the density ρ for the sake of consistency. These boundary conditions are stable in the sense of Kreiss (see [EM1]).

4.2. The Navier–Stokes system. Here the number of boundary conditions is $n-p$, where p is the number of negative eigenvalues of $\bar{A}^{(1)}$. $\bar{A}^{(1)}$ reduces to $-u_1$, so that we must distinguish only between the inflow and the outflow cases:

- **Inflow** boundary: $u_1<0$.

$$p=0: \quad \textbf{4 boundary conditions.}$$

- **Outflow** boundary: $u_1>0$

$$p=1: \quad \textbf{3 boundary conditions.}$$

We must determine the other part of the family (Ψ^j), i.e., the θ_j and Θ_j solutions of

$$[P^{(1,1)}\theta+A^{(1)}]\Theta=0.$$

The cubic equation for θ has an immediate root we shall call θ_3:

$$\theta_3=u_1. \tag{4.9}$$

The two other roots are solutions of the quadratic equation

$$-u_1(\tfrac{4}{3}\theta-u_1)(\alpha\theta-u_1)-RT(\alpha\theta-\gamma u_1)=0, \tag{4.10}$$

where, for simplicity, we set

$$\alpha=\gamma P_r^{-1}.$$

We have

$$\theta_1\theta_2=\frac{3}{4\alpha}(u_1^2-c^2), \qquad \theta_1+\theta_2=\frac{3}{4}\frac{[\alpha(u_1^2-c^2/\gamma)+\frac{4}{3}u_1^2]}{\alpha u_1}$$

with $c^2=\gamma RT$.

Recalling that $\gamma>1$, we can determine the signs of the roots. We order θ_1 and θ_2 so that $\theta_1<\theta_2$ (θ_1 cannot be equal to θ_2).

Subsonic case: $\theta_1<0<\theta_2$.

- Inflow case: $\theta_3<0$;
- Outflow case: $\theta_3>0$.

Supersonic case.

- Inflow case: $\theta_1<\theta_2<0$, $\theta_3<0$;
- Outflow case: $0<\theta_1<\theta_2$, $\theta_3>0$.

The corresponding generalized eigenvalues are

$$\Theta^i=\begin{pmatrix} u_1 \\ 0 \\ (\gamma-1)Tu_1/(\alpha\theta_i-u_1) \\ -1 \end{pmatrix}, \quad i=1,2, \qquad \Theta^3=\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \tag{4.11}$$

We now have all the elements to (2.20).

- *Subsonic inflow*: $m=1$, $p=0$, $r+p=3$.

$$\zeta_1=\lambda_1, \quad \tilde{\Psi}^1=\Lambda^1, \quad \zeta_2=\theta_1, \quad \tilde{\Psi}^2=\Theta^1, \quad \zeta_3=\theta_3, \quad \tilde{\Psi}^3=\Theta^3$$

and

$$\tilde{N}=\begin{pmatrix} c & u_1 & 0 \\ 0 & 0 & 1 \\ (\gamma-1)T & U_1 & 0 \end{pmatrix}, \qquad \tilde{N}^{-1}=\frac{1}{\det\tilde{N}}\begin{pmatrix} -U_1 & 0 & u_1 \\ (\gamma-1)T & 0 & -c \\ 0 & \det\tilde{N} & 0 \end{pmatrix},$$

where $U_i=(\gamma-1)Tu_1/(\alpha\theta_i-u_1)$, $\det\tilde{N}=(\gamma-1)Tu_1-U_1c$

$$\begin{aligned}
&\nu\frac{\partial\tilde{u}_1}{\partial x_1}+\frac{1}{8}\nu\frac{\partial\tilde{u}_2}{\partial x_2}=\frac{\theta_1}{1-c/(\gamma P_r^{-1}\theta_1-u_1)}\left(\tilde{u}_1-\frac{c}{\gamma-1}\frac{\tilde{T}}{T}\right),\\
&\nu\frac{\partial\tilde{u}_2}{\partial x_1}+\frac{1}{6}\nu\frac{\partial\tilde{u}_1}{\partial x_2}=u_1\tilde{u}_2,\\
&\nu\frac{\partial\tilde{T}}{\partial x_1}=\frac{(\gamma-1)T\theta_1}{\gamma P_r^{-1}\theta_1-u_1-c}\left(\tilde{u}_1-\frac{c}{\gamma-1}\frac{\tilde{T}}{T}\right),\\
&\frac{\tilde{\rho}}{\rho}=\frac{-1}{u_1(1-c/(\gamma P_r^{-1}\theta_1-u_1))}\left[\frac{\gamma P_r^{-1}\theta_1}{\gamma P_r^{-1}\theta_1-u_1}\tilde{u}_1-\frac{u_1+c}{\gamma-1}\frac{\tilde{T}}{T}\right].
\end{aligned} \tag{4.12a}$$

This system reduces to (4.8a) when $\nu=0$.

- *Subsonic outflow*: $m=3$, $p=1$, $r+p=4$.

$$\zeta_i=\lambda_i,\quad \tilde{\Psi}^i=\Lambda^i,\quad i=1,2,3,\quad \zeta_4=\theta_1,\quad \tilde{\Psi}^4=\Theta^1,$$

$$\tilde{N}=\begin{pmatrix} c & 0 & 0 & u_1 \\ 0 & 1 & 0 & 0 \\ (\gamma-1)T & 0 & T & U_1 \\ 1 & 0 & -1 & -1 \end{pmatrix},$$

$$\tilde{N}^{-1}=\frac{1}{TZ}\begin{pmatrix} T-U_1 & 0 & u_1 & u_1T \\ 0 & TZ & 0 & 0 \\ T-U_1-\gamma T & 0 & u_1+c & T(u_1+c)-ZT \\ +\gamma T & 0 & -c & -cT \end{pmatrix},$$

where $Z=\gamma u_1-c(U_1/T-1)$. Condition (2.20) reduces to

$$\nu\sum_{j=1}^{2}\bar{P}^{(1j)}\frac{\partial w^I}{\partial x_j}=\bar{P}^{(11)}\theta_1\left(\sum_{j=1}^{4}\tilde{N}^{-1}_{4j}w_j\right)\Theta^{1I},$$

or

$$\nu\frac{\partial\tilde{u}_1}{\partial x_1}+\frac{1}{8}\nu\frac{\partial\tilde{u}_2}{\partial x_2}=\frac{c\theta_1u_1}{Z}\left(\frac{\gamma}{c}\tilde{u}_1-\frac{\tilde{T}}{T}-\frac{\tilde{\rho}}{\rho}\right),$$

$$\nu\frac{\partial\tilde{u}_2}{\partial x_1}+\frac{1}{6}\nu\frac{\partial\tilde{u}_1}{\partial x_2}=0, \tag{4.12b}$$

$$\nu\frac{\partial\tilde{T}}{\partial x_1}=\frac{c\theta_1U_1}{Z}\left(\frac{\gamma}{c}\tilde{u}_1-\frac{\tilde{T}}{T}-\frac{\tilde{\rho}}{\rho}\right).$$

Again, these equations reduce to (4.8b) when $\nu=0$.

- *Supersonic inflow*: $m=0$, $p=0$, $r+p=3$.

$$\zeta_i=\theta_i,\quad \tilde{\Psi}^i=\Theta^i,\quad i=1,\cdots,3,$$

and

$$\tilde{N}=\begin{pmatrix} u_1 & u_1 & 0 \\ 0 & 0 & 1 \\ U_1 & U_2 & 0 \end{pmatrix}\quad \tilde{N}^{-1}=\frac{1}{u_1(U_1-U_2)}\begin{pmatrix} -U_2 & 0 & u_1 \\ +U_1 & 0 & -u_1 \\ 0 & u_1(U_1-U_2) & 0 \end{pmatrix}.$$

The boundary conditions are

$$\nu\frac{\partial\tilde{u}_1}{\partial x_1}+\frac{1}{8}\nu\frac{\partial\tilde{u}_2}{\partial x_2}=\frac{3}{4u_1}\left(u_1^2-\frac{c^2}{\gamma}\right)\tilde{u}_1+\frac{3c^2}{4\gamma}\frac{\tilde{T}}{T},$$

$$\nu\frac{\partial\tilde{u}_2}{\partial x_1}+\frac{1}{6}\nu\frac{\partial\tilde{u}_1}{\partial x_2}=u_1\tilde{u}_2, \tag{4.12c}$$

$$\nu\frac{\partial\tilde{T}}{\partial x_1}=\frac{(\gamma-1)}{\alpha}Tu_1\left(\frac{\tilde{u}_1}{u_1}+\frac{1}{\gamma-1}\frac{\tilde{T}}{T}\right),$$

$$\frac{\tilde{\rho}}{\rho}=-\frac{\tilde{u}_1}{u_1}.$$

If $\nu = 0$, the system reduces to $\tilde{u}_1 = \tilde{u}_2 = \tilde{T} = \tilde{\rho} = 0$.

- *Supersonic outflow*: $m = 4$, $p = 1$, $r + p = 4$.

$$\zeta_i = \lambda_i, \qquad i = 1, \cdots, 4$$

so that the boundary condition is

$$\begin{aligned} &\frac{\partial \tilde{u}_1}{\partial x_1} + \frac{1}{8}\frac{\partial \tilde{u}_2}{\partial x_2} = 0, \\ &\frac{\partial \tilde{u}_2}{\partial x_1} + \frac{1}{6}\frac{\partial \tilde{u}_1}{\partial x_2} = 0, \\ &\frac{\partial \tilde{T}}{\partial x_1} = 0. \end{aligned} \tag{4.12d}$$

The results in § 3 apply to these equations as follows.

THEOREM 4.1. *The initial boundary value problem for* (4.2) *and the zero-order boundary conditions* (4.12) *is well posed in* Ω^-. *As the viscosity* ν *tends to zero, the solution converges to the solution of the Euler equation with the corresponding boundary conditions* (4.8), *the* L^2 *norm of the error decreases linearly in* ν.

Proof. We merely need to check that the reduced hyperbolic problem is well posed, which is extremely simple here since $\bar{A}^{(1)} = -u_1$. The boundary condition then reduces to $\tilde{\rho} = 0$ (when there is a boundary condition for $\tilde{\rho}$) and the following problem:

$$\begin{aligned} &\frac{\partial \tilde{\rho}}{\partial t} = -u_1 \frac{\partial \tilde{\rho}}{\partial x_1} - u_2 \frac{\partial \tilde{\rho}}{\partial x_2}, \qquad x_1 \leqq 0, \\ &\tilde{\rho} = 0, \qquad x_1 = 0 \end{aligned}$$

for $u_1 < 0$ is obviously well posed.

Remark. In [GS] the authors introduced for the Navier–Stokes compressible equation artificial boundary conditions by requiring them to be dissipative. Furthermore, these boundary conditions produce a weak boundary layer. Therefore Theorem 4.1 also holds in that case. We have not been able to decide whether our boundary conditions are dissipative or not. However for more general systems or higher dimensions, it seems difficult to extend their techniques which consist of studying the boundary form $(\mathscr{E}w, w)$ and matching coefficients of the boundary condition to get the right sign. It does not allow for higher-order boundary conditions either.

5. Higher-order boundary conditions. We discussed earlier the goals of our work: provide boundary conditions which would be (1) local and (2) consistent with the Euler equation. A first step was made in § 3 by an approximation of order zero of the right-hand side in (2.10). We now want to increase the accuracy of our boundary conditions. This means, from our point of view, expand first the transparent boundary condition (2.10) up to higher order in ν. By doing this, we shall keep terms like $\alpha_i(s, \eta)$, for $1 \leqq i \leqq m$, where α_i is the traveling mode defined in (1.3). These will correspond to pseudodifferential operators of order 1 on the boundary, which are, of course, far from being local. We thus shall in turn approximate these modes with the techniques described in [EM]. The first realistic approximation is similar to (2.17): we shall set $\eta = 0$, and approximate the quantities in (2.7) to first order in ν.

We shall restrict ourselves here to the particular case of the viscous linearized shallow-water system, though the procedure carries over without modification to more general systems, provided they possess a symmetrizer. This property ensures that the eigenvalues ξ_i for the system have an expansion $\xi_i(s, \eta, \nu) = \zeta_i(s, \eta) + \nu\chi_i(s, \eta) + O(\nu^2)$,

$1 \leqq i \leqq m$, and $\xi_i(s, \eta, \nu) = (\zeta_i + \nu\chi_i + O(\nu^2))/\nu$, $m+1 \leqq i \leqq r+p$, with a corresponding expansion for the eigenvectors Ψ_j.

Let us consider the shallow-water equations, linearized about the steady-state $(U, 0)$:

$$\frac{\partial w}{\partial t} = A^{(1)} \frac{\partial w}{\partial x_1} + A^{(2)} \frac{\partial w}{\partial x_2} + \nu \left(P^{(11)} \frac{\partial^2 w}{\partial x_1^2} + P^{(22)} \frac{\partial^2 w}{\partial x_2^2} \right), \tag{5.1}$$

where $w = (u_1, u_2, \varphi)$.

$$A^{(1)} = \begin{pmatrix} -U & 0 & -1 \\ 0 & -U & 0 \\ -c^2 & 0 & -U \end{pmatrix}, \tag{5.2}$$

$$A^{(2)} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -c^2 & 0 \end{pmatrix}, \tag{5.3}$$

$$P^{(11)} = P^{(22)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{5.4}$$

with $c > 0$.

In our notation of § 0, $r = 2$, $\bar{P}^{(ij)} = \delta_{ij} I_2$, and $\bar{A}^{(1)} = -U$. The eigenvalues of $A^{(1)}$ are

$$\lambda_1 = -U - c, \quad \lambda_2 = -U, \quad \lambda_3 = -U + c \tag{5.5a}$$

and the corresponding eigenvectors are

$$\Lambda_1 = \begin{pmatrix} 1 \\ 0 \\ c \end{pmatrix}, \quad \Lambda_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \Lambda_3 = \begin{pmatrix} 1 \\ 0 \\ -c \end{pmatrix}. \tag{5.5b}$$

The solutions of $(P^{(1,1)}\theta + A^{(1)})\Theta = 0$ are

$$\theta_1 = \frac{U^2 - c^2}{U}, \qquad \theta_2 = U, \tag{5.6a}$$

and the corresponding generalized eigenvectors are

$$\Theta_1 = \begin{pmatrix} U \\ 0 \\ -c^2 \end{pmatrix}, \qquad \Theta_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}. \tag{5.6b}$$

The signs of the λ's and θ's depend on whether the flow is sub or supersonic, and ingoing or outgoing, as in the case of the full Navier–Stokes equations.

We shall approximate the generalized eigenvalues and eigenvectors $\xi_i(s, 0, \nu)$ and $\Phi^i(s, 0, \nu)$ up to first order in ν.

For $1 \leqq i \leqq m$:

$$\begin{aligned} &\xi_i(s, \nu) = \zeta_i(s) + \nu\chi_i s^2 + O(\nu^2), \\ &\zeta_i(s) = \frac{s}{\lambda_i}, \end{aligned} \tag{5.7a}$$

$$\begin{aligned} &\Phi^i(s, \nu) = \Psi^i + \nu \Xi^i s + O(\nu^2), \\ &\Psi^i = \Lambda^i; \end{aligned} \tag{5.7b}$$

For $m+1 \leqq i \leqq r+p$:

$$\xi^i(s,\nu) = \frac{1}{\nu}\zeta_i + \chi_{i-m}s + O(\nu),$$

(5.8a)

$$\zeta_i = \theta_{i-m},$$

$$\Phi^i(s,\nu) = \Psi^i + \nu\Xi^{i-m}s + O(\nu^2),$$

(5.8b)

$$\Psi^i = \Theta^{i-m}$$

(in the formulas above, the variable η, being zero, has been omitted).

The χ_i's and Ξ_i's are obtained by substitution of the expressions above in formula (11) for $\eta = 0$:

$$(\xi A^{(1)} + \nu\xi^2 P^{(11)} - sI)\Phi = 0.$$

To λ_1, λ_2, λ_3 are associated three values of ξ and vectors Φ:

$$\tilde{\xi}_1 = \frac{s}{\lambda_1} - \frac{\nu s^2}{2\lambda_1^3} + O(\nu^2),$$

(5.9a)

$$\tilde{\xi}_2 = \frac{s}{\lambda_2} - \frac{\nu s^2}{\lambda_2^3} + O(\nu^2),$$

$$\tilde{\xi}_3 = \frac{s}{\lambda_3} - \frac{\nu s^2}{2\lambda_2^3} + O(\nu^2),$$

$$\tilde{\Phi}^1 = \Lambda^1 + \nu s\begin{pmatrix} -1/2\lambda_1 c \\ 0 \\ 0 \end{pmatrix} + O(\nu^2) = \begin{pmatrix} 1 - \nu s/2\lambda_1 c \\ 0 \\ c \end{pmatrix} + O(\nu^2),$$

(5.9b)

$$\tilde{\Phi}_2 = \Lambda^2 + O(\nu^2) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + O(\nu^2),$$

$$\tilde{\Phi}_3 = \Lambda^3 + \nu s\begin{pmatrix} -1/2\lambda_3 c \\ 0 \\ 0 \end{pmatrix} + O(\nu^2) = \begin{pmatrix} 1 - \nu s/2\lambda_3 c \\ 0 \\ -c \end{pmatrix} + O(\nu^2),$$

and to θ_1, θ_2 are associated two values of ξ and Φ:

$$\tilde{\tilde{\xi}}_1 = \frac{U^2 - c^2}{\nu U} + s\frac{U^2 + c^2}{U(U^2 - c^2)} + O(\nu),$$

(5.10a)

$$\tilde{\tilde{\Phi}}^1 = \Theta^1 + \nu s\begin{pmatrix} 0 \\ 0 \\ c^2/(U^2 - c^2) \end{pmatrix} + O(\nu^2)$$

$$= \begin{pmatrix} U \\ 0 \\ -c^2(1 - \nu s/(U^2 - c^2)) \end{pmatrix} + O(\nu^2),$$

$$\tilde{\tilde{\xi}}_2 = \frac{U}{\nu} + \frac{s}{U} + O(\nu),$$

(5.10b)

$$\tilde{\tilde{\Phi}}^2 = \Theta^2 + O(\nu^2) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + O(\nu^2).$$

We cannot go any further without dividing the analysis into four cases: subsonic or supersonic, inflow or outflow.

—*Subsonic inflow case*: $-c < U < 0$, $p = 0$.

$$\lambda_1 < 0, \quad \lambda_2, \lambda_3 > 0: \quad m = 1,$$
$$\theta_1 > 0, \qquad \theta_2 < 0.$$

p is equal to zero. We have three boundary conditions

$$\xi_1 = \tilde{\xi}_1, \qquad \Phi^1 = \tilde{\Phi}^1,$$
$$\xi_2 = \tilde{\tilde{\xi}}^2, \qquad \Phi^2 = \tilde{\tilde{\Phi}}^2.$$

By a zero-order approximation of the ξ_i's and Φ^i's, we get the first set of approximated boundary conditions:

$$(5.11)_0 \qquad \begin{cases} \nu \dfrac{\partial u_1}{\partial x_1} = 0, \\ \nu \dfrac{\partial u_2}{\partial x_1} - U u_2 = 0, \\ \varphi - c u_1 = 0. \end{cases}$$

By a first-order approximation in ν, we obtain a new set of boundary conditions, which contains differentiation in time:

$$(5.11)_1 \qquad \begin{cases} \nu \dfrac{\partial u_1}{\partial x_1} = -\dfrac{\nu}{U+c} \dfrac{\partial u_1}{\partial t}, \\ \nu \dfrac{\partial u_2}{\partial x_1} = U u_2 + \dfrac{\nu}{U} \dfrac{\partial u_2}{\partial t}, \\ \varphi = c u_1 - \dfrac{\nu}{2(U+c)} \dfrac{\partial u_1}{\partial t}. \end{cases}$$

—*Subsonic outflow case*: $0 < U < c$, $p = 1$.

$$\lambda_1, \lambda_2 < 0, \quad \lambda_3 > 0: \quad m = 2,$$
$$\theta_1 < 0, \qquad \theta_2 > 0.$$

The generalized eigenvalues and eigenvectors with negative real parts (when $\operatorname{Re} s > 0$) are

$$\xi_1 = \tilde{\xi}_1, \quad \xi_2 = \tilde{\xi}_2, \quad \xi_3 = \tilde{\tilde{\xi}}_1,$$
$$\Phi^1 = \tilde{\Phi}^1, \quad \Phi^2 = \tilde{\Phi}^2, \quad \Phi^3 = \tilde{\tilde{\Phi}}^1.$$

The analogue to $(5.11)_0$ becomes

$$(5.12)_0 \qquad \nu \frac{\partial u_1}{\partial x_1} = \frac{c-U}{c}(-c u_1 + \varphi), \qquad \nu \frac{\partial u_2}{\partial x_1} = 0,$$

and the first-order boundary condition is

$$(5.12)_1 \qquad \begin{cases} \nu \dfrac{\partial u_1}{\partial x_1} = \dfrac{c-U}{c}(-c u_1 + \varphi) + \dfrac{\nu}{2c^2(c-U)} \dfrac{\partial}{\partial t}(-c^2 u_1 + U\varphi), \\ \nu \dfrac{\partial u_2}{\partial x_1} = -\dfrac{\nu}{U} \dfrac{\partial u_2}{\partial t}. \end{cases}$$

In the supersonic case, the calculations are much easier.

—*Supersonic inflow case*: $U<-c$, $p=0$.

$$\lambda_1, \lambda_2, \lambda_3>0: \quad m=0,$$
$$\theta_1, \theta_2<0.$$

We have three boundary conditions:

$(5.13)_0$
$$\begin{cases} \nu\dfrac{\partial u_1}{\partial x_1}=\dfrac{U^2-c^2}{U}u_1; \\ \nu\dfrac{\partial u_2}{\partial x_1}=Uu_2, \\ \varphi=-\dfrac{c^2}{U}u_1, \end{cases}$$

and

$(5.13)_1$
$$\begin{cases} \nu\dfrac{\partial u_1}{\partial x_1}=\dfrac{U^2-c^2}{U}u_1+\nu\dfrac{U^2+c^2}{U(U^2-c^2)}\dfrac{\partial u_1}{\partial t}, \\ \nu\dfrac{\partial u_2}{\partial x_1}=Uu_2+\dfrac{\nu}{U}\dfrac{\partial u_2}{\partial t}, \\ \varphi=-\dfrac{c^2}{U}u_1+\dfrac{\nu c^2}{U(U^2-c^2)}\dfrac{\partial u_1}{\partial t}. \end{cases}$$

—*Supersonic outflow case*: $U>c$, $p=1$.

$$\lambda_1, \lambda_2, \lambda_3<0: \quad m=3,$$
$$\theta_1, \theta_2>0.$$

Here we have two boundary conditions:

$(5.14)_0$
$$\nu\frac{\partial u_1}{\partial x_1}=0, \qquad \nu\frac{\partial u_2}{\partial x_1}=0,$$

$(5.14)_1$
$$\begin{cases} \nu\dfrac{\partial u_1}{\partial x_1}=\dfrac{\nu}{U^2-c^2}\dfrac{\partial}{\partial t}(-Uu_1+\varphi), \\ \nu\dfrac{\partial u_2}{\partial x_1}=-\dfrac{\nu}{U}\dfrac{\partial u_2}{\partial t}. \end{cases}$$

We shall not repeat here the calculations for the inviscid case, i.e., $\nu=0$. They are identical to those for the full Euler equations. The local boundary condition (2.17) is in this case:

Subsonic inflow case:

(5.15) $$u_2=0, \qquad \varphi-cu_1=0;$$

Subsonic outflow case:

(5.16) $$\varphi-cu_1=0;$$

Supersonic inflow case:

(5.17) $$u_1=u_2=\varphi=0;$$

Supersonic outflow case:

(5.18) no boundary conditions.

The analogue of Theorem 4.1 holds.

THEOREM 5.1. *The initial boundary value problem for* (5.1) *and any of the boundary conditions* $(5.11+i)_0$, $i=0,\cdots,3$ *is well posed in* Ω^-. *As the viscosity* ν *tends to zero, the solution converges to the solution of the inviscid equation with the corresponding boundary condition* $(5.14+i)$, *and the* L^2*-norm of the error decreases as* $O(\nu)$.

The proof is exactly the same as for Theorem 4.1.

Remark. In this case, the well posedness in the classical sense can be expressed by energy estimates, using the variational formula (1.14). Let us denote by $E(t)$ the quantity defined by

$$E(t)=\frac{1}{2}\int\int_{\Omega^-}[c^2(u_1^2+u_2^2)+\varphi^2]\,dx \tag{5.19}$$

and the analogue on Γ:

$$E_\Gamma(t)=\frac{1}{2}\int_\Gamma[c^2(u_1^2+u_2^2)+\varphi^2]\,dx_2. \tag{5.20}$$

The energy equality then reads

$$\begin{aligned}&\frac{dE}{dt}+\nu\int\int_{\Omega^-}(\nabla u_1^2+\nabla u_2^2)\,dx\\&\quad=-UE_\Gamma(t)+c^2\int_\Gamma\left(\nu u\cdot\frac{\partial u}{\partial x_1}-u_1\varphi\right)dx_2+\int\int_{\Omega^-}Fu\,dx.\end{aligned} \tag{5.21}$$

It can be easily checked in each case that the quantity integrated on Γ is negative.

Unfortunately, the decoupling conditions prescribed in [M] to obtain the well posedness and the error estimates do not apply to our higher-order boundary conditions $(5.11+i)_1$. We have not been able to establish a priori estimates in this case either. However, the formal expansion (3.6) is still available. It is an easy matter to check that for the higher-order boundary conditions $(5.11+i)_1$, the next term in the expansion vanishes. For instance, in the subsonic inflow case, it is due to the fact that $(\partial/\partial x_1+1/(U+c)\,\partial/\partial t)(w_0^{(1)})=0$. *So the boundary layer is weaker than in the former case, and the solution of the corresponding initial boundary value problem converges formally to the solution of the inviscid equation with boundary condition* $(5.14+i)$, *the error being* $O(\nu^2)$.

Remark. Consider the boundary conditions derived from $(5.11+i)_1$ in the four cases by neglecting certain terms:

$$\begin{cases}\nu\dfrac{\partial u_1}{\partial x_1}=-\dfrac{\nu}{U+c}\dfrac{\partial u_1}{\partial t},\\[2ex]\nu\dfrac{\partial u_2}{\partial x_1}=Uu_2+\dfrac{\nu}{U}\dfrac{\partial u_2}{\partial t},\qquad \varphi=cu_1,\end{cases} \tag{5.22}$$

$$\nu\frac{\partial u_1}{\partial x_1}=\frac{c-U}{c}(-cu_1+\varphi),\qquad \nu\frac{\partial u_2}{\partial x_1}=-\frac{\nu}{U}\frac{\partial u_2}{\partial t}, \tag{5.23}$$

$$\begin{cases}\nu\dfrac{\partial u_1}{\partial x_1}=\dfrac{U^2-c^2}{U}u_1+\nu\dfrac{U^2+c^2}{U(U^2-c^2)}\dfrac{\partial u_1}{\partial t},\\[2ex]\nu\dfrac{\partial u_2}{\partial x_1}=u_2+\dfrac{\nu}{U}\dfrac{\partial u_2}{\partial t},\qquad \varphi=-\dfrac{c^2}{U}u_1,\end{cases} \tag{5.24}$$

$$\nu\frac{\partial u_1}{\partial x_1}=0,\qquad \nu\frac{\partial u_2}{\partial x_1}=-\frac{\nu}{U}\frac{\partial u_2}{\partial t}. \tag{5.25}$$

These boundary conditions are well posed in the classical sense: we have neglected the terms which could prevent the energy from decreasing in time. Furthermore, they still give an approximation to the inviscid problem with boundary conditions $(5.14+i)$ in $O(\nu^2)$: the relevant equations are unchanged. However, the last statement remains formal, since the decoupling conditions still do not hold.

So far we have considered approximations to the inviscid equations with the "zero-order" boundary conditions (2.17), which are those used in practice. It is tempting to try to approximate the Euler problem better. This adds new important difficulties: as pointed out in [EM] the choice of the "good" boundary condition in the hyperbolic case is not canonical, and furthermore, it is not clear whether or not it is well posed in the sense of Kreiss. An analysis of such boundary conditions, together with numerical experiments, will be presented in a forthcoming paper.

REFERENCES

[ABL] S. ABARBANEL, A. BAYLISS, AND L. LUSTMAN, *Non-reflecting boundary conditions for the compressible Navier-Stokes equations*, ICASE Report 86.9, ICASE-NASA Langley Research Center, Hampton, VA, 1986.

[BT1] A. BAYLISS AND E. TURKEL, *Radiation boundary conditions for wave-like equations*, Comm. Pure Appl. Math., 33 (1980), pp. 707-727.

[BT2] ——, *Outflow boundary conditions for fluid dynamics*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 250-259.

[BT3] ———, *Far field boundary conditions for compressible flows*, J. Comput. Phys., 48 (1982), pp. 182-199.

[EM1] B. ENGQUIST AND A. MAJDA, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp., 31 (1977), pp. 629-651.

[EM2] ———, *Radiation boundary conditions for acoustic and elastic wave calculation*, Comm. Pure Appl. Math., 32 (1979), pp. 313-357.

[GS] B. GUSTAFSSON AND A. SUNDSTROM, *Incompletely parabolic problems in fluid dynamics*, SIAM J. Appl. Math., 35 (1986), pp. 425-439.

[H] L. HALPERN, *Artificial boundary conditions for the linear advection diffusion equation*, Math. Comp., 46 (1986), pp. 425-439.

[HS] L. HALPERN AND M. SCHATZMAN, *Artificial boundary conditions for incompressible viscous flows*, SIAM J. Math. Anal., 20 (1989), pp. 308-353.

[K] H. O. KREISS, *Initial boundary value problems for hyperbolic systems*, Comm. Pure Appl. Math., 22 (1970), pp. 277-298.

[KL] H. O. KREISS AND J. LORENZ, *Initial Boundary Value Problems and the Navier-Stokes Equations*, Academic Press, New York, 1989.

[M] D. MICHELSON, *Initial-Boundary Value Problems for Incomplete Singular Perturbations of Hyperbolic Systems*, Lecture Notes in Appl. Math. 22, Springer-Verlag, Berlin, New York, 1985.

[OS] J. OLIGER AND A. SUNDSTROM, *Theoretical and practical aspects of some initial boundary value problems in fluid dynamics*, SIAM J. Appl. Math., (1978), pp. 419-447.

[RS1] D. H. RUDY AND J. C. STRIKWERDA, *A nonreflecting outflow boundary condition for subsonic Navier-Stokes calculations*, J. Comput. Phys., 36 (1980), pp. 55-70.

[RS2] ———, *Boundary conditions for subsonic compressible Navier-Stokes calculations*, Comput. & Fluids, 9 (1981), pp. 327-338.

[S] J. C. STRIKWERDA, *Initial boundary value problems for incompletely parabolic systems*, Comm. Pure Appl. Math., 33 (1977), pp. 797-822.

[YS] T. YTREHUS AND J. SMOLDEREN, *Structure of linearized Navier-Stokes equations for compressible flows*, Phys. Fluids, 25 (1982), pp. 429-435.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1284-1292, September 1991

005

ADMISSIBILITY CONDITIONS FOR SHOCKS IN CONSERVATION LAWS THAT CHANGE TYPE*

BARBARA LEE KEYFITZ†

Abstract. Systems of conservation laws which are not of classical hyperbolic type appear in models for some complex flows. To formulate admissibility conditions for shocks, a single shock is regularized between a point where the system is hyperbolic and a point where it is not, by the addition of a higher-order term ("viscosity matrix"), to obtain an admissibility criterion based on the existence of connecting orbits in a related dynamical system. This criterion is more sensitive to the structure of the viscosity matrix than in the classical hyperbolic case.

Key words. conservation laws, viscous profiles, shock admissibility

AMS(MOS) subject classifications. 35L65, 35L67, 35M05, 58F14

1. Introduction.

We consider a pair of conservation laws,

$$w_t + h_x \equiv w_t + A(w)w_x = 0, \tag{1.1}$$

where

$$h = h(w), \quad w = (u, v), \quad h = (f, g).$$

The characteristic speeds λ_i satisfy

$$\lambda^2 - \operatorname{tr} A\lambda + \det A = 0.$$

They are ordered, by convention, $\lambda_1 < \lambda_2$, when they are real. The system is hyperbolic or elliptic according to whether

$$D(w) = (\operatorname{tr} A)^2 - 4 \det A$$

is positive or negative. We assume that there are open sets

$$\mathscr{H} = \{w \mid D(w) > 0\}, \qquad \mathscr{E} = \{w \mid D(w) < 0\}$$

representing each type, and that their common boundary,

$$\mathscr{B} = \{w \mid D(w) = 0\}, \tag{1.2}$$

is a smooth curve.

Systems with this feature appear as models for propagation of phase boundaries in elasticity [10], in fluids of van der Waals type near the critical temperature [19], and in the dynamics of some models for shape-memory alloys or austenitic to martensitic transitions in solids [1]. Change of type has also been noted in models for pressure-driven, convection-dominated, three-phase flow in porous media [3]. Some unusual models of flows have also led to systems that change type; these include a kinematic model for two-directional traffic flow [4] and an ecological model [9].

The theory we formulate in this work indicates that there are at least two classes of flows that change type, and we observe that this corresponds to a distinction in the observed phenomena. In models for the dynamics of phase transitions, the presence

* Received by the editors February 27, 1990; accepted for publication (in revised form) December 7, 1990. This research was supported in part by the Air Force Office of Scientific Research under grant 86-0088, and the Air Force Office of Scientific Research and the National Science Foundation under grant DMS-89-03768.

† Department of Mathematics, University of Houston, Houston, Texas 77204-3476.

of an elliptic region is coexistent with a physically unstable range of the order variable (density or strain), as manifested by the behavior of the constitutive relation. A typical system is

$$u_t - v_x = 0, \qquad v_t - \sigma(u)_x = 0. \tag{1.3}$$

This system is a nonlinear wave equation whose characteristic speeds $\pm\sqrt{\sigma'}$ are real and opposite when $\sigma(u)$ is an increasing function of u, and complex, with real part zero, when $\sigma(u)$ is decreasing. The symmetry of the wave speeds expresses Galilean invariance of the system: this system is of *wave equation type*. Symmetry is absent in three-phase porous medium flow, where the pressure differential breaks the reflectional symmetry; it is absent also in the kinematic model of traffic flow and in abstract models, such as general perturbations of quadratic flux functions [8]. Also, there is no obvious physical instability connected with the elliptic region in these models. The theory presented here finds the second class to be *generic*; all wave-equation-type examples exhibit a degeneracy.

Change of type occurs also in steady transonic flow; the model equations are of the form

$$h_x + k_y \equiv A(w)w_x + B(w)w_y = 0, \tag{1.4}$$

and the characteristic directions (ξ, η), which are the roots of

$$\det|\xi A(w) + \eta B(w)| = 0,$$

are real on one side of the sonic line and complex on the other.

There are significant mathematical differences between (1.4) and (1.1).

An attempt to formulate admissibility conditions for systems of mixed type based on the model (1.4) was made by Mock [17]. Although there is some overlap, the results in [17] cannot be applied without modification to (1.1). Conversely, the normal form analysis which is the main result of the present paper can be applied to (1.4), but a different normal form is required [16].

2. Conservation law considerations. We study the admissibility of *uniform planar shocks*: solutions of (1.1) of the form

$$w(x, t) = \begin{cases} w_0, & x < st, \\ w_1, & x > st \end{cases} \tag{2.1}$$

where w_0 and w_1 are constant states, related to the shock speed s by the *Rankine-Hugoniot relation*:

$$s[w] - [h] \equiv s(w_1 - w_0) - (h(w_1) - h(w_0)) = 0, \tag{2.2}$$

which makes (2.1) a weak solution of (1.1). For fixed w_0, the set of values w such that (2.2) is satisfied for some s forms the *wave locus*,

$$W(w_0) = \{w \mid \exists s \ni s(w - w_0) = h(w) - h(w_0)\}. \tag{2.3}$$

Locally, W is a curve, parameterized by s.

Not every configuration (2.1) that satisfies (2.2) is an admissible solution. In the equations of isentropic gas dynamics in Lagrangian coordinates (equation (1.3) with σ a monotone function) W comprises four branches, which may be labeled S_i and S_i^*, $i = 1$ or 2. Each extends from w_0 to infinity; S_i and S_i^* are tangent to r_i, the right eigenvector of A corresponding to λ_i, and $s \to \lambda_i$ as $w \to w_0$ along each branch. In addition, s is monotonically decreasing on S_i^*, and increases along S_i, as $w \to w_0$. Admissibility is determined by the *Lax geometric entropy condition* (LGEC):

$$\lambda_i(w_0) > s > \lambda_i(w)$$

for $i=1$ or 2, where w is the state on the right in (2.1). Figure 1 illustrates the situation schematically: points on the S_i branches correspond to admissible shocks in (2.1), while if $w_1 \in S_i^*(w_0)$, then an admissible shock is obtained if left and right are reversed in (2.1). (See [20] or [11] for a discussion.)

The LGEC cannot be applied to equations that change type. In this paper we apply a well-known method that is not sensitive to type, the *viscous profile* criterion: admissible shocks are those solutions of (2.1) that are limits as $\varepsilon \to 0$ of self-similar solutions of

$$w_t + h_x = \varepsilon (M w_x)_x, \tag{2.4}$$

$$w = w(\xi) \equiv w\left(\frac{x-st}{\varepsilon}\right), \qquad w(\xi) \to \begin{cases} w_0, & \xi \to -\infty, \\ w_1, & \xi \to \infty. \end{cases} \tag{2.5}$$

Such a $w(\xi)$ is a heteroclinic orbit of the vector field

$$M\frac{dw}{d\xi} = h(w) - sw + c, \tag{2.6}$$

where c is the value $sw_0 - h(w_0) = sw_1 - h(w_1)$, from (2.2). The matrix M is called the *viscosity matrix.*

3. The Hugoniot loop. The fact that there are solutions of (2.2) with real s for which one state is in $\mathcal{H}$ and the other in $\mathcal{E}$ is well known from many examples. The first theorem of this section is a local result which gives a qualitative description of the wave locus $W(w_0)$ when w_0 is near $\mathcal{B}$.

A nondegeneracy hypothesis $\nabla_w D(w) \neq 0$ on $\mathcal{B}$, which guarantees that $\mathcal{B}$ is a smooth curve, will be assumed throughout. This implies that

$$N(w) = A(w) - \lambda(w) I$$

is a nilpotent matrix of rank one on $\mathcal{B}$ [15]. Thus N has a unique right eigenvector r and left eigenvector l on $\mathcal{B}$. We impose a second nondegeneracy condition

$$l^T d^2 h(r, r) > 0 \tag{3.1}$$

on $\mathcal{B}$. (Points where this condition fails are interesting, and are discussed in [15].) If (3.1) holds, then (1.1) is genuinely nonlinear in $\mathcal{H}$, sufficiently close to $\mathcal{B}$, and also

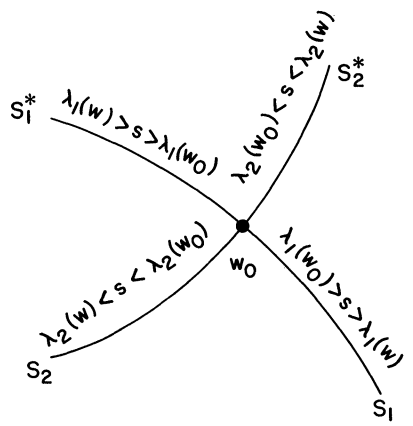


FIG. 1. *Admissibility for hyperbolic systems.*

satisfies the Smoller-Johnson interaction condition [11]:

$$l_j^T d^2h(r_i, r_i) > 0 \quad \text{for } i, j = 1, 2.$$

Here r_i and l_i are right and left unit eigenvectors of $A = dh$ corresponding to λ_i, normalized by $r_i \cdot \nabla\lambda_i > 0$ and $l_i \cdot r_i > 0$. Figure 2 illustrates $W(w_0)$ in a neighborhood of w_0 when w_0 is near $\mathcal{B}$. (Necessarily, S_1^* and S_2 are the branches that turn toward $\mathcal{B}$.) It can be seen that

$$\lim r_1(w) = r(w_*), \quad \lim r_2(w) = -r(w_*), \quad \lim l_i(w) = l(w_*)$$

as $w \to w_* \in \mathcal{B}$ through points in $\mathcal{H}$. Also, writing N as $\beta\alpha^T$, we have $r = \beta/|\beta|$ and $l = -\alpha/|\alpha|$. Finally, r is directed strictly into $\mathcal{E}$.

THEOREM 3.1. *Let w_* be a point in $\mathcal{B}$ where* (3.1) *holds. Then for w_0 in a neighborhood $\mathcal{N}$ of w_*, $W(w_0)$ has the following structure:*

— *If $w_0 \in \mathcal{B}$, then W forms a cusp opening into $\mathcal{H}$, with axis tangent to $r(w_0)$. The value of s at w_0 is $\lambda(w_0)$, and $ds/d\mu = \mathcal{O}(1/\sqrt{\mu})$ near w_0, where μ is arclength on W.*

— *If $w_0 \in \mathcal{E}$, then W is the union of w_0 itself and a smooth curve lying entirely in $\mathcal{H}$ (and therefore disconnected from w_0); s is monotonic on W.*

— *If $w_0 \in \mathcal{H}$, then W consists of a loop which crosses $\mathcal{B}$, and two segments which leave $\mathcal{N}$ in the opposite direction; thus, W is a curve with a self-intersection; s is monotonic along the entire curve.*

Remark. The first and third possibilities are illustrated in Fig. 2. We shall refer to the segment between the two self-intersections as the *Hugoniot loop.*

Proof. The zero-set of (2.2),

$$G(w, w_0, s) \equiv h(w) - h(w_0) - s(w - w_0) = 0,$$

can be analyzed as a bifurcation problem, using the singularity theory approach of Golubitsky and Schaeffer and the notion of t-equivalence [5, p. 129]. This type of equivalence reflects the persistence of the trivial solution $w = w_0$ for all values of the bifurcation parameter s. Since dG has rank greater than or equal to 1, G can be reduced to a one-state-variable problem; we use a Lyapunov-Schmidt reduction that preserves the t-equivalence. Fixing $w_0 = w_* \in \mathcal{B}$, $dG|_0 = N(w_*)$; a basis for $\ker N$ is r and a basis for $(\text{Range } N)^\perp$ is l. We express w implicitly in terms of a reduced variable x as

$$w = xr + w_p(x, \nu) + w_0$$

where $\nu = s - \lambda$ is the bifurcation parameter. A straightforward calculation shows that G is t-equivalent to

$$\eta x^2 + \delta x \nu^2 = 0,$$

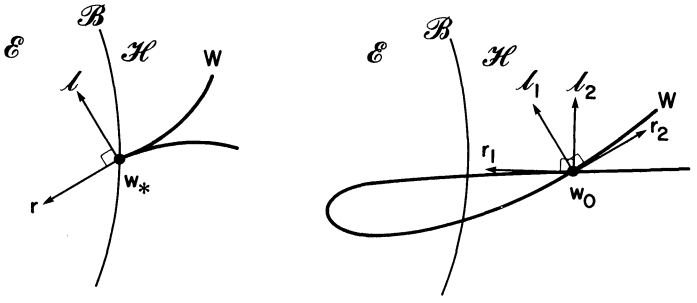


FIG. 2. *The Hugoniot loop.*

where

$$\eta = l^T d^2 h(r, r) > 0, \qquad \delta = \frac{2}{|\alpha||\beta|} > 0.$$

This normal form has t-codimension 1, and an unfolding,

$$\eta x^2 + \delta x \nu^2 + ax = 0, \tag{3.2}$$

is obtained by perturbing w_0 away from $\mathcal{B}$, say

$$w_0 = w_* + ar.$$

The conclusions of the theorem now follow upon examining the nontrivial solution of (3.2), $x = -\nu^2 - a$, and noting that x and a refer to distances along the vector r.

When $a < 0$, so that $w_0 \in \mathcal{H}$, the solution x has two zeros at $\nu = \pm\sqrt{(-a)}$ corresponding to the two ends of the loop. We can parameterize the loop by arclength so that s is monotonically decreasing on the loop. At $\nu = 0$, $x > 0$, so that the loop contains a subinterval inside $\mathcal{E}$.

A similar theorem was proved by Mock [17], using expansions of h near $\mathcal{B}$. Theorem 3.1 clarifies the hypotheses and shows that the size of $\mathcal{N}$ depends on the region of validity of the Lyapunov-Schmidt procedure, which can be determined in specific examples.

To study the vector field (2.6) for w_0 in $\mathcal{N}$ and w_1 in $W(w_0)$, we again choose a degenerate case and unfold it, this time using vector field unfoldings.

THEOREM 3.2. *Under assumption* (3.1) *and an additional nondegeneracy condition*

$$l^T d^2 h(r, l) + r^T d^2 h(r, r) \neq 0; \tag{3.3}$$

then (2.6) *at* $w_0 = w_1 = w_* \in \mathcal{B}$, *and* $M = I$ *is equivalent* (*in the sense of vector field equivalence*) *to*

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= Ax^2 + 2Bxy, \end{aligned} \qquad A, B, \neq 0. \tag{3.4}$$

Furthermore, varying w_0, w_1, *and* M *in a neighborhood of these values gives a universal unfolding of the vector field.*

Remarks. (1) System (3.4) is the Takens-Bogdanov normal form. It has codimension two, in the sense of vector field unfoldings, and is discussed in detail by Guckenheimer and Holmes [6, p. 365]. For the application to (2.6), we wish to unfold under t-equivalence, so that the solution $w_0 = w_1$ is preserved. The unfolding in this case was worked out by Hirschberg and Knobloch [7]. Unfolding theory for vector fields (as in [6] and [7]) does not consider a distinguished parameter, so an interpretation in the spirit of Theorem 3.1 requires further discussion. This follows the proof.

(2) It is not necessary to vary M to obtain the universal unfolding, but it is useful to recognize that $M \neq I$ is included in the application of this theorem.

(3) The additional nondegeneracy condition (3.3) is necessary to obtain the normal form (3.4), and it is not always satisfied. For example, the nonlinear wave equation (1.3) is always degenerate. On the other hand, the model equation devised in [12] to explain some features of a three-phase porous medium flow system and quadratic models, such as [4], [8], and [9], satisfy this condition, except for special values of the parameters. For this reason, such systems were called generic in the Introduction.

Proof. Without loss of generality, let $c=0$ and $w_0=w_1=w_*=0$ in (2.6). By subtracting $\lambda(0)w$ from h we may assume $\lambda(0)=0$ also, and so $s=0$. The linearization of (2.6) at the critical point $w=0$ is

$$\frac{dw}{d\xi}=dh(0)w\equiv Nw, \tag{3.5}$$

where N is the nilpotent matrix previously encountered. By performing a linear change of coordinates, $w\mapsto Pw$, if necessary, we may assume $r=(-1,0)^T$, $l=(0,1)^T$ and

$$N=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}.$$

Expanding h through quadratic powers of w now yields

$$h(w)=Nw+d^2h(w,w).$$

Write $w=xr-yl$, and expand d^2h, which is bilinear in its arguments. Equation (2.6) (with signs reversed) now agrees with (3.4) through linear terms; the higher-order terms must be removed by a nonlinear change of variables. Noting that

$$d^2h(w,w)=x^2d^2h(r,r)-2xyd^2h(r,l)+y^2d^2h(l,l)$$

and that multiplication by r and $-l$, respectively, gives the contribution to the first and second equations, we obtain

$$A=-l^Td^2h(r,r),\qquad B=l^Td^2h(r,l)+r^Td^2h(r,r) \tag{3.6}$$

upon carrying out the nonlinear change of variables explicitly.

A universal unfolding of (3.4) that respects the trivial solution but has no other symmetries is [7]

$$\dot{x}=y,\qquad \dot{y}=Ax^2+2Bxy+\mu_1x+\mu_2y. \tag{3.7}$$

The fixed points are $(0,0)$ and $(-\mu_1/A,0)$, and the linearization at zero is

$$\begin{pmatrix}0 & 1\\ \mu_1 & \mu_2\end{pmatrix}.$$

In identifying unfolding variables, we may choose $\mu_1=(w_1-w_0)\cdot r$. If we specify $\mu_1\gtrless 0$, we avoid some redundancy. Now w_0 is the origin, and is a saddle in the dynamics, while $w_1\in W(w_0)$ is to the left of it in the Hugoniot loop. As discussed earlier, we must have $w_0\in\mathscr{H}$ if $w_1\to w_0$, and then the approach is along a 1- or 2-shock branch. If $w_0\notin\mathscr{B}$, then the steady-state bifurcation as w_1 crosses w_0 is transcritical, and μ_2 is any parameter that measures the nondegeneracy of this bifurcation: for example, $\mu_2=\lambda_2(w_0)-\lambda_1(w_0)$ is adequate locally. As $w_0\to\mathscr{B}$, we use a smooth determination $\mu_2=\lambda_i(w_0)-\lambda_j(w_0)$ of the eigenvalues. We did not attempt to find an unfolding using the minimum number of variables.

If $M\neq I$, the linearized system (3.5) is associated with the matrix $M^{-1}N$, which may not have a double zero eigenvalue. However, if M is close to I, there will be a codimension-2 bifurcation point near $w_1=w_0=w_*$: perturbing M is yet another way of unfolding the singularity.

This completes the proof of Theorem 3.2. We now turn to a qualitative analysis of the vector field (2.6) for w_0 and w_1 in $\mathscr{N}$. Assume w_0 to be in $\mathscr{H}$, and w_1 in the Hugoniot loop of w_0; allow w_0 to represent either the forward or the backward limit in (2.5). We describe the sequence of vector field bifurcations as w_1 traverses the Hugoniot loop, introducing s as a distinguished parameter; this gives a path

$(\mu_1(s), \mu_2(s))$ through unfolding space. We refer to Fig. 3, which follows [7]. The coordinate axes are μ_1 and μ_2. The Hugoniot loop traces a path in this space beginning on the negative μ_2 axis (when $w_1 = w_0$ and $s = \lambda_2$) and ending on the positive μ_2 axis ($w_1 = w_0$ and $s = \lambda_1$), and traversing three open regions in the right half plane. For s near λ_2, there is a connecting orbit: a solution of (2.4), (2.5) with limit (2.1) as $\varepsilon \to 0$. In this region, states w_1 near w_0 are on the classical S_2 branch of $W(w_0)$, and these are classical 2-shocks as long as w_1 remains in $\mathcal{H}$. The transition of w_1 into $\mathcal{E}$ does not affect the existence or qualitative properties of the orbit. It is true that if $M = I$, then $w_1 \in \mathcal{B}$ marks a change of type of the unstable critical point, from a node to a spiral, but this transition occurs at a different value of w_1 if $M \neq I$.

The other end of the loop, near the positive μ_2 axis, corresponds to a state w_1 in S_1^*: the admissible shock wave represented by this pair is

$$w(x, t) = \begin{cases} w_1, & x < st, \\ w_0, & x > st. \end{cases}$$

This orbit also persists as w_1 crosses $\mathcal{B}$ into $\mathcal{E}$.

Two curves B_h and B_{sc}, which separate the S_2 and S_1^* branches, enclose an open region in which there is no orbit connecting the fixed points of (3.7). According to the sign of B in (3.7), B_h and B_{sc} are as shown here ($B < 0$) or interchanged. At a point on B_h, a Hopf bifurcation occurs at w_1. For $M = I$, a simple calculation in (2.6) shows that this occurs when

$$s(w_1, w_0) = \operatorname{Re} \lambda_1(w_1) = \operatorname{Re} \lambda_2(w_1), \tag{3.8}$$

and hence that w_1 is in $\mathcal{E}$ at this point. For any other choice of M, the point may be similarly calculated; this is always a local condition.

In the wedge between B_h and B_{sc}, w_1 is a stable spiral, and an unstable limit cycle separates all orbits containing w_1 from any orbit reaching w_0. Geometrically, this limit cycle grows in size as the point $(\mu_1(s), \mu_2(s))$ moves away from B_h. The curve B_{sc} represents a line of saddle connections: the limit cycle coincides with a homoclinic loop at w_0, which disappears on the other side of B_{sc} (the "blue sky catastrophe"). The existence of B_{sc} could be inferred from the necessity of continuous transitions of the vector field along the Hugoniot loop, but the fact that it is unique, under the

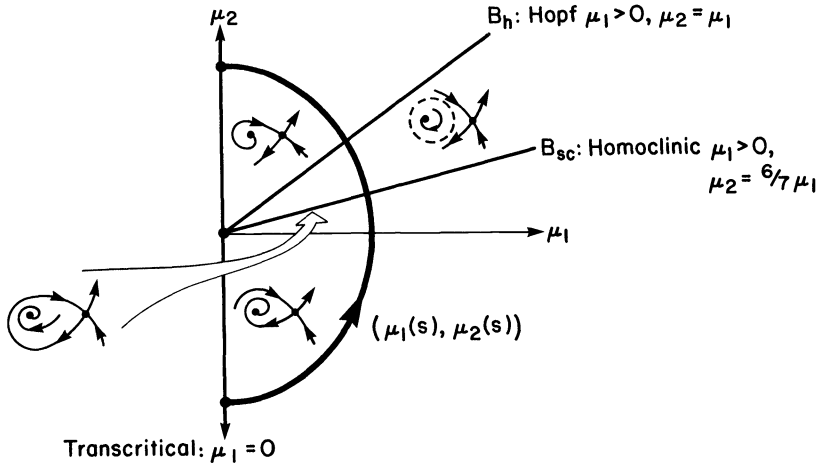


FIG. 3. *The vector field bifurcations.*

nondegeneracy conditions holding for (3.4) and its unfolding, is a consequence of the theory in [6] and [7]. Furthermore, it is theoretically quite difficult to find the point on the Hugoniot loop at which $w_1 \in B_{sc}$, since this, unlike (3.8), is not a local condition.

4. The viscous profile criterion and shock admissibility. We restate the conclusion of Theorem 3.2 as it applies to a shock configuration on the Hugoniot loop. (See Fig. 4.)

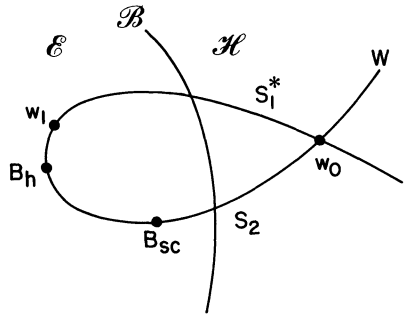


FIG. 4. *Admissibility conditions on the Hugoniot loop.*

COROLLARY TO THEOREM 3.2. *Under the nondegeneracy conditions* $\nabla D(w) \neq 0$ *at* w_* *on* $\mathscr{B}$, *and* (3.1) *and* (3.3), *and for* w_0 *and* w_1 *in a neighborhood* $\mathscr{N}(w_*)$ *in which the reduction to normal form is valid, the Hugoniot loop of* w_0 *is divided by two points,* B_{sc} *and* B_h, *into three open intervals: in* S_2 *and* S_1^*, *there are connecting orbits that give shock profiles that extend the* 2- *and* 1-*shocks, respectively. These profiles are not necessarily monotone.*

Between B_h and B_{sc} is an open interval where no connecting orbit exists.

For a given value of w_0, the locations of B_{sc} and B_h depend on the choice of M. When M is near I, both points are in $\mathscr{E}$.

Existence of traveling wave orbits for (2.6) can be related to other admissibility criteria. In [13], Keyfitz suggested an extension of the LGEC for states in $\mathscr{E}$: for a 2-shock with $w_1 \in \mathscr{E}$,

$$\lambda_2(w_0) > s > \operatorname{Re} \lambda_2(w_1), \quad \lambda_1(w_0) < s, \quad \operatorname{Re} \lambda_1(w_1) < s. \tag{4.1}$$

This condition is insufficient: it does not take into account B_{sc} and the possibility of a limit cycle. Furthermore, while the analogous condition for a 1-shock appears correct for the case where $M = I$ and $B < 0$, it is incorrect when $M \neq I$. A simple geometric condition equivalent to the viscous profile criterion does not exist in this case. Since the LEGC is based on properties of initial boundary-value problems, we can pose such a problem when change of type occurs. We have some preliminary results in [2] that are consistent with the present paper.

The results in § 3 explicitly exclude system (1.3), for which $B = 0$. For $M = I$, the curve B_h exists and along this curve the system is Hamiltonian. Perturbation to $M \neq I$ will introduce both the cases $B > 0$ and $B < 0$. The normal form for this singularity requires third-order terms.

For some model problems, such as (1.3), a singular viscosity matrix may be appropriate. For example, Shearer used this as an admissibility condition in [18]. We have excluded singular viscosities from the theorems by requiring M to be near I, but have examined them in numerical experiments. We conjecture that they describe a limiting case for the S_2 and S_1^* intervals on the Hugoniot loop.

Acknowledgments. In applying bifurcation theory, both steady state and dynamic, to find and describe the local form of the wave locus and orbit structure, I benefited from discussions with Marty Golubitsky and Bill Farr. These results were first presented at the GAMM International Conference in honor of Jack Hale on Problems Involving Change of Type; a summary appears in [14].

REFERENCES

[1] H. W. ALT, K.-H. HOFFMAN, M. NIEZGÓDKA, AND J. SPREKELS, *A numerical study of structural phase transitions in shape memory alloys*, Preprint No. 90, Institute for Math., University of Augsburg, Augsburg, Germany, 1985.

[2] K. A. AMES AND B. L. KEYFITZ, *Stability of shocks in systems that change type: the linear approximation*, in Proc. Third Internat. Conference on Hyperbolic Problems, Engquist and Gustafsson, eds., to appear.

[3] J. B. BELL, J. A. TRANGENSTEIN, AND G. R. SHUBIN, *Conservation laws of mixed type describing three-phase flow in porous media*, SIAM J. Appl. Math., 46 (1986), pp. 1000–1023.

[4] J. H. BICK AND G. F. NEWELL, *A continuum model for two-directional traffic flow*, Quart. Appl. Math., 18 (1960), pp. 191–204.

[5] M. GOLUBITSKY AND D. G. SCHAEFFER, *Singularities and Groups in Bifurcation Theory*, Springer-Verlag, New York, 1985.

[6] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[7] P. HIRSCHBERG AND E. KNOBLOCH, *An unfolding of the Takens-Bogdanov singularity*, Quart. J. Appl. Math. (1991), to appear.

[8] H. HOLDEN AND L. HOLDEN, *On the Riemann problem for a prototype of a mixed type conservation law*, II, in Current Progress in Hyperbolic Systems: Riemann Problems and Computations, W. B. Lindquist, ed., Contemporary Math. 100, American Mathematical Society, Providence, RI, 1989, pp. 331–367.

[9] L. HSIAO AND P. DE MOTTONI, *Existence and uniqueness of Riemann problem for nonlinear system of conservation laws of mixed type*, Amer. Math. Soc. Trans. (1991), to appear.

[10] R. D. JAMES, *The propagation of phase boundaries in elastic bars*, Arch. Rational Mech. Anal., 73 (1980), pp. 125–158.

[11] B. L. KEYFITZ AND H. C. KRANZER, *Existence and uniqueness of entropy solutions to the Riemann problem for hyperbolic systems of two nonlinear conservation laws*, J. Differential Equations, 27 (1978), pp. 444–476.

[12] B. L. KEYFITZ, *An analytic model for change of type in three-phase flow*, in Numerical Simulation in Oil Recovery, M. F. Wheeler, ed., IMA 11, Springer-Verlag, New York, 1988, pp. 149–160.

[13] ———, *Change of type in three-phase flow: a simple analogue*, J. Differential Equations, 80 (1989), pp. 280–305.

[14] B. L. KEYFITZ, *The use of vectorfield dynamics in formulating admissibility conditions for shocks in systems that change type*, in Problems Involving Change of Type, K. Kirchgassner, ed., Springer-Verlag, Berlin, 1990, pp. 141–150.

[15] ———, *A criterion for certain wave structures in systems that change type*, in Current Progress in Hyperbolic Systems: Riemann Problems and Computations, W. B. Lindquist, ed., Contemporary Math. 100, American Mathematical Society, Providence, RI, 1989, pp. 203–213.

[16] ———, *Shocks near the sonic line: a comparison between steady and unsteady models for change of type*, in Nonlinear Evolution Equations that Change Type, B.L. Keyfitz and M. Shearer, eds., Springer-Verlag, New York, 1990, pp. 89–106.

[17] M. S. MOCK, *Systems of conservation laws of mixed type*, J. Differential Equations, 37 (1980), pp. 70–88.

[18] M. SHEARER, *Admissibility criteria for shock wave solutions of a system of conservation laws of mixed type*, Proc. Royal Soc. Edinburgh Sect. A, 93 (1983), pp. 233–244.

[19] M. SLEMROD, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Rational Mech. Anal., 81 (1983), pp. 301–315.

[20] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1293-1302, September 1991

006

BEST POSSIBLE UPPER BOUND FOR BLOWUP SOLUTIONS OF THE QUASILINEAR HEAT CONDUCTION EQUATION WITH SOURCE*

VICTOR A. GALAKTIONOV†

Abstract. This article studies the behavior of the weak solution of the Cauchy problem to quasilinear degenerate parabolic equation $u_t = (u^\sigma u_x)_x + u^{\sigma+1}$, where $\sigma > 0$ is a fixed constant, with a nonnegative bounded compactly supported initial function. Let T_0 be the finite blowup time for solution $u(x, t)$. The estimate $\sup_x u(x, t) \leqq M(t)$ for all $t \in (0, T_0)$, where $M(t)$ is defined from some nonlinear ordinary differential equation, is proved by the comparison of intersection methods (based on the theory of lap number or zero set) with explicit noninvariant blowup solution.

Key words. quasilinear heat conduction equation with source, blowup solution, localization, noninvariant explicit solution, upper bound

AMS(MOS) subject classification. 35K

1. Introduction. Main results. In this paper we consider the Cauchy problem for the quasilinear degenerate parabolic equation

$$u_t = (u^\sigma u_x)_x + u^{\sigma+1} \quad \text{for } x \in \mathbf{R} = (-\infty, \infty), \quad t > 0, \tag{1.1}$$

$$u(x, 0) = u_0(x) \quad \text{in } \mathbf{R}. \tag{1.2}$$

Here $\sigma > 0$ is a fixed constant. We assume that the initial function satisfies the hypotheses

$$\begin{aligned} &u_0 \geqq 0 \quad \text{in } \mathbf{R}, \quad u_0 \not\equiv 0, \quad \sup u_0 < \infty, \\ &u_0 \text{ has compact support } \operatorname{supp} u_0 = \{x \in \mathbf{R} \mid u_0(x) > 0\}, \\ &u_0^\sigma \text{ is uniformly Lipschitz continuous in } \mathbf{R}. \end{aligned} \tag{1.3}$$

Equation (1.1) describes the evolution of initial temperature u_0 in a medium with nonlinear heat conduction coefficient $k(u) = u^\sigma$ and heat source $Q(u) = u^{\sigma+1}$ depending on the temperature $u = u(x, t) \geqq 0$.

It is well known [4], [20, p. 208] that for any nonzero, nonnegative, initial function problem (1.1), (1.2) has no global (in time) solution and $u(x, t)$ blows up in finite time T_0, i.e., the weak solution $u(x, t)$ exists in $\mathbf{R} \times [0, T_0)$ and

$$\overline{\lim_{t \to T_0^-}} \sup_{x \in \mathbf{R}} u(x, t) = +\infty. \tag{1.4}$$

Here $T_0 = T_0(u_0)$, depending upon the initial function u_0, is called the finite blowup time of the unbounded solution $u(x, t)$.

In this paper some global properties of the unbounded solution of the Cauchy problem (1.1), (1.2) on the whole time interval $(0, T_0)$ are investigated. Detailed information concerning different properties of unbounded blowup solutions of (1.1) and of the equation with power nonlinearities of a more general kind,

$$u_t = (u^\sigma u_x)_x + u^\beta, \qquad x \in \mathbf{R}, \quad t > 0, \tag{1.5}$$

where $\sigma > 0$ and $\beta > 1$ are arbitrary fixed constants, are given in [20, Chaps. IV, V, VII]; see also the list of references given there. Properties of blowup solutions of the

* Received by the editors November 20, 1989; accepted for publication (in revised form) June 24, 1990.
† Keldysh Institute of Applied Mathematics, Academy of Sciences of the Union of Soviet Socialist Republics, Miusskaya Square 4, 125047 Moscow, Union of Soviet Socialist Republics.

Cauchy problem for (1.5) are quite different for three cases: (i) so-called LS evolution of blowup solution, $\beta > \sigma+1$, when $u(x, t)$ may go to infinity as $t \to T_0$ in a domain of zero measure or single point blowup (see [7], [20, p. 253]); (ii) S evolution, $\beta = \sigma+1$, when under the hypothesis (1.3) the spatial blowup set for the solution $u(x, t)$ is a bounded domain of nonzero measure (see [20, p. 230], [3]); (iii) HS evolution, $\beta < \sigma+1$, when $u(x, t) \to +\infty$ as $t \to T_0$ for any $x \in \mathbf{R}$ (see [20, p. 238]). This classification (LS, S, and HS evolutions) of different types of unbounded blowup solutions was introduced approximately twenty years ago. See a full list of references in [20].

Thus the case $\beta = \sigma+1$ is the interesting "boundary" one between the region $\{\beta > \sigma+1\}$ of single point localization of blowup solutions and region $\{\beta < \sigma+1\}$ of nonlocalized solutions of total blowup.

We now state the main result of the paper.

THEOREM 1. *Let the hypotheses* (1.3) *hold and let* $T_0 \in (0, \infty)$ *be the finite blowup time of the solution to* (1.1), (1.2). *Then*

$$\|u(\cdot, t)\|_{L^\infty(\mathbf{R})} \equiv \sup_{x \in \mathbf{R}} u(x, t) \leqq \{\varphi(t)[\psi(t)+1]\}^{1/\sigma}, \qquad t \in (0, T_0), \tag{1.6}$$

where the function $\psi(t) \in (-1, 1)$ *satisfies the ordinary differential equation*

$$\psi' = \sigma(\sigma+1)^{-1} C_0 (1-\psi^2)^{-\sigma/2}, \quad t \in (0, T_0), \quad \psi(0) = -1, \tag{1.7}$$

$$C_0 = C_0(T_0) \equiv [(\sigma+1)/\sigma] T_0^{-1} B(1+\sigma/2, \tfrac{1}{2}) \tag{1.8}$$

($B(p, q)$ *is the beta-function*), *and*

$$\varphi(t) = C_0[1-\psi^2(t)]^{-(\sigma+2)/2}, \qquad t \in (0, T_0). \tag{1.9}$$

Note that the upper bound (1.6) holds for the arbitrary blowup solution and the right side of (1.6) depends only on the finite blowup time T_0 and does not depend on the initial function satisfying (1.3) (except through the value of T_0).

Estimate (1.6) allows for the possibility that the amplitude of the solution may be nonmonotonic in time. For example, $\sup_x u(x, t)$ may decrease for small $t > 0$ and $\sup_x u(x, t) \to \infty$ as $t \to T_0$. Under some additional conditions on the initial function ($u_0(-x) = u_0(x)$ in $\mathbf{R}$ and $u_0 \not\equiv$ const. is a nonincreasing function in $\mathbf{R}_+ = (0, \infty)$) the same result was proved in [3]. In order to avoid these restrictive conditions on u_0 and to prove the estimate (1.6) for sufficiently arbitrary initial functions satisfying (1.3), we use the idea of the so-called "comparison of intersection" of blowup solutions from [5] and [6] (see also [10]).

If we define asymptotics of the function

$$M(t) := \{\varphi(t)[\psi(t)+1]\}^{1/\sigma}, \qquad t \in (0, T_0) \tag{1.10}$$

as $t \to 0$ and $t \to T_0$ using (1.6) we get the following estimates.

COROLLARY 1. *Under the hypotheses of Theorem* 1 *there exists* $\tau_1 > 0$, *depending on* σ *and* T_0, *such that*

$$\|u(\cdot, t)\|_{L^\infty(\mathbf{R})} \leqq a_0 t^{-1/(\sigma+2)}(1+o(1)), \qquad 0 < t < \tau_1, \tag{1.11}$$

where

$$a_0 = a_0(T_0) \equiv 2^{-1/\sigma}[(\sigma+1)/\sigma]^{1/\sigma}(\sigma+2)^{-1/(\sigma+2)} \cdot T_0^{-2/\sigma(\sigma+2)}[B(1+\sigma/2, \tfrac{1}{2})]^{2/\sigma(\sigma+2)}. \tag{1.12}$$

COROLLARY 2. *Under the hypotheses of Theorem* 1 *there exists* $\tau_2 > 0$, *depending on* σ *and* T_0, *such that*

$$\|u(\cdot, t)\|_{L^\infty(\mathbf{R})} \leqq [A_0^{-1}(T_0 - t)]^{-1/\sigma}(1 + o(1)), \qquad T_0 - \tau_2 < t < T_0, \tag{1.13}$$

where $A_0 = 2(\sigma+1)/\sigma(\sigma+2)$.

We shall show that the upper estimates (1.11) and (1.13) of the asymptotic behavior of the solution near $t = 0$ and near finite blowup time $t = T_0$ are the best possible.

The plan of the paper is as follows. In § 2 we give some preliminary information and results. In § 3 some basic properties of the noninvariant explicit solution of the equation (1.1) [8], [9] are described. The main result is obtained by the comparison of intersection (see [20, Chap. IV]) of the solution $u(x, t)$ with the above-mentioned explicit solution which has the same finite blowup time. Necessary preliminary information about the comparison of such an intersection is given in § 4. Theorem 1 is proved in § 5.

The explicit solution from § 3 shows the existence of the weak solution of the Cauchy problem for (1.1) with singular initial function

$$u(x, 0) = \delta(x) \quad \text{in } \mathbf{R} \tag{1.14}$$

($\delta(x)$ is Dirac's measure). It is shown that the weak solution $u_*(t, x)$ of (1.1), (1.14) exists in $\mathbf{R} \times [0, T_*)$ where

$$T_* = 2^{-1}\sigma^{-(\sigma+1)}(\sigma+1)^{(\sigma+2)/2}B(1+\sigma/2, \tfrac{1}{2})B^\sigma(1+1/\sigma, \tfrac{1}{2}) \tag{1.15}$$

and $u_*(x, t) \to +\infty$ as $t \to T_*^-$ in the localization domain $\omega_L = (-L_s/2, L_s/2)$, where $L_s = 2\pi(\sigma+1)^{1/2}/\sigma$ is the so-called fundamental length of the nonlinear medium described by (1.1) [20, Chap. IV], and $u_*(x, t) = 0$ in $(\mathbf{R} \setminus \omega_L) \times [0, T_*)$.

The explicit solution $u_*(x, t)$ gives another general property of the localization of blowup solutions to (1.1). It was shown in [3] by the comparison of intersection that the right front $h_+(t) \equiv \sup\{x \in \mathbf{R} \mid u(x, t) > 0\}$ of any unbounded solution satisfies the estimate

$$h_+(t) \leqq h_+(0) + (L_s/2\pi) \cdot [\pi/2 + \arcsin \psi(t)] < h_+(0) + L_s/2$$

in $(0, T_0)$, where the function $\psi(t)$ is as defined in (1.7), (1.8). Hence the right heat front does not move more than a distance $L_s/2$ from its initial position. This upper estimate is exact since the right front of the explicit solution $u_*(x, t)$ moves exactly through the distance $L_s/2$ as $t \to T_*^-$. Therefore L_s is really a fundamental characteristic of the nonlinear equation (1.1), since the property of the heat front mentioned above does not depend on the initial function.

The explicit solution (from § 3) illustrates several of the most general properties of a wide class of blowup solutions, i.e., it possesses the "fundamental" characteristics of the nonlinear equation (1.1).

2. Preliminary results. In the general case, the Cauchy problem for the quasilinear degenerate parabolic equation (1.1) admits a weak blowup solution.

A bounded continuous function $u(x, t)$ satisfying (1.2) is said to be a weak solution of the problem (1.1), (1.2) in $Q_T = \mathbf{R} \times [0, T]$ if there exists a weak derivative $(u^{\sigma+1})_x \in L^2_{\mathrm{loc}}(Q_T)$, and if the identity

$$\iint_{Q_T} \{(\sigma+1)^{-1}(u^{\sigma+1})_x f_x - u^{\sigma+1} f - u f_t\}\, dx\, dt = 0 \tag{2.1}$$

holds for arbitrary test functions $f \in C_0^1(Q_T)$. The function $u(x, t)$ is said to be a blowup weak solution of the problem (1.1), (1.2) in $\mathbf{R} \times [0, T_0)$ if it is a weak solution in Q_T for all $T \in (0, T_0)$ and if (1.4) holds.

Local existence, uniqueness, and the comparison theorems for quasilinear degenerate parabolic equations of type (1.1) are proved, for instance, in [17]-[19] and [14]. See the detailed survey in [11]. For any points where $u > 0$ and where the equation is uniformly parabolic with smooth coefficients, the weak solution is classical and has derivatives of arbitrary order. Nonexistence of the global weak solution to the problem (1.1), (1.2) for $u_0 \not\equiv 0$ was proved in [4]. See also [20, p. 208]. We shall use the well-known property of finite speed of propagation of disturbances for the degenerate equation (1.1), i.e., under hypotheses (1.3), $u(x, t)$ has compact support for any $t \in (0, T_0)$. See the references in [11].

3. Explicit solution. It has been shown in [8] and [9] that for arbitrary $\sigma > 0$ the quasilinear parabolic equation (1.1) admits an explicit weak blowup solution of the form

$$
\begin{aligned}
&u_*(x, t) = \{\varphi(t)[\psi(t) + \cos(2\pi x / L_s)]\}^{1/\sigma} \quad \text{for } x \in (-L_s/2, L_s/2), \\
&u_* \equiv 0 \quad \text{for } x \in \mathbf{R} \setminus (-L_s/2, L_s/2), \quad t \in (0, T_0),
\end{aligned} \tag{3.1}
$$

where the function $\psi(t)$ satisfies the ordinary differential equation (1.7), (1.8) and the function $\varphi(t)$ is defined in (1.9). It can easily be shown that this explicit solution (for C_0 from (1.8)) has the finite blowup time T_0.

Consider the main evolution properties of the explicit solution (3.1). It describes the behavior of the blowup heat structure with nonmonotone spatial amplitude

$$
\|u_*(\cdot, t)\|_{L^\infty(\mathbf{R})} \equiv u_*(0, t) = M(t) \equiv \{\varphi(t)[\psi(t) + 1]\}^{1/\sigma}, \qquad t \in (0, T_0), \tag{3.2}
$$

and $u_*(0, t) = M(t) \to +\infty$ as $t \to T_0$. Using (1.7)-(1.9), we have that

$$
u_*(0, t) = M(t) = a_0 t^{-1/(\sigma+2)}(1 + o(1)) \to +\infty \quad \text{as } t \to 0, \tag{3.3}
$$

$$
u_*(0, t) = [A_0^{-1}(T_0 - t)]^{-1/\sigma}(1 + o(1)) \to +\infty \quad \text{as } t \to T_0^-, \tag{3.4}
$$

where constants a_0, A_0 are as defined in Corollaries 1 and 2.

Thus, the formulas above show that the spatial amplitude $M(t)$ of the solution $u_*(x, t)$ is a nonmonotone function. It is easy to calculate that

$$
\min_{(0, T_0)} M(t) = T_0^{-1/\sigma} B^{1/\sigma}(1 + \sigma/2, \tfrac{1}{2})[\sigma(\sigma+2)/(\sigma+1)^2]^{-(\sigma+2)/2\sigma}
$$

for

$$
t = t_m = T_0 B^{-1}(1 + \sigma/2, \tfrac{1}{2}) \int_{1/(\sigma+1)}^{1} (1 - z^2)^{\sigma/2}\, dz \in (0, T_0/2).
$$

For any $t \in (0, T_0)$ this explicit solution has the compact support with the right front

$$
x = h_+^*(t) \equiv g(t) \equiv (L_s/2\pi)[\pi/2 + \arcsin \psi(t)], \qquad t \in (0, T_0). \tag{3.5}
$$

Therefore $u_*(x, t)$ is the localized solution, since $h_+^*(t) < L_s/2$ for all $t \in (0, T_0)$ and $h_+^*(T_0^-) = L_s/2$. Note that (3.5) leads to the following behavior of the right heat front:

$$
h_+^*(t) = b_0^{1/2} t^{1/(\sigma+2)}(1 + o(1)) \quad \text{as } t \to 0, \tag{3.6}
$$

where

$$
b_0 = b_0(T_0) \equiv (\sigma+1)\sigma^{-2}(\sigma+2)^{2/(\sigma+2)} T_0^{-2/(\sigma+2)}[B(1 + \sigma/2, \tfrac{1}{2})]^{2/(\sigma+2)}.
$$

Consider now a more detailed asymptotic space-time structure of $u_*(x, t)$ near $t=0$. Such results are important for comparison of intersection. First, we note that $u_*(x, t)$ is the nonnegative bounded weak solution of (1.1) in $\mathbf{R}\times[\tau, T]$ for arbitrary fixed $\tau\in(0, T_0)$ and $T\in(\tau, T_0)$. From (1.7) and (1.9) it follows that $\psi(t)=-1+b_0t^{2/(\sigma+2)}(1+o(1))$ and $\varphi(t)=A_0t^{-1}(1+o(1))$ as $t\to 0$. Therefore

$$u_*(x, t)=(A_0t^{-1}(1+o(1))[b_0t^{2/(\sigma+2)}(1+o(1))+\cos(2\pi x/L_s)-1])_+^{1/\sigma} \quad \text{as } t\to 0.$$

Since $1-\cos(2\pi x/L_s)=2\sin^2(\pi x/L_s)=2\pi^2x^2/L_s^2+o(x^2)$ as $x\to 0$ we get

$$u_*(x, t)=a_0t^{-1/(\sigma+2)}([(1+\lambda_1(t, x))-(x/b_0^{1/2}t^{1/(\sigma+2)})^2](1+\lambda_2(t, x)))_+^{1/\sigma}, \tag{3.7}$$

where $\lambda_1(t, x)$ and $\lambda_2(t, x)$ are sufficiently smooth functions such that $\lambda_i(t, x)=o(1)$ as $t+|x|\to 0$, $i=1, 2$.

Hence, for $T_0=T_*$ (see (1.15)) the explicit solution $u_*(x, t)$ satisfies the initial condition (1.14) in the weak sense:

$$\lim_{t\to 0}\int_{-\infty}^{\infty}u_*(x, t)\xi(x)\,dx=\xi(0)$$

for arbitrary test functions $\xi\in C_0^\infty(\mathbf{R})$.

4. Comparison of intersection. Main preliminary statements. We now give some preliminary statements concerning the comparison of intersection of the solution of the problem (1.1), (1.2) with the explicit solution $u_*(x, t)$.

Let $u(x, t)$ be the solution of the Cauchy problem (1.1), (1.2) in $\mathbf{R}\times(0, T_0)$ with the initial function satisfying (1.3), and let T_0 be the finite blowup time. We suppose that $\operatorname{supp}_x u(x, t)\equiv\{x\in\mathbf{R}\,|\,u(x, t)>0\}=(h_-(t), h_+(t))$, i.e., the support of $u(x, t)$ is the finite connected interval for any $t\in[0, T_0)$. Then $h_\pm(t)$ are continuous functions in $[0, T_0)$, $h_+(t)$ is a nondecreasing function, and $h_-(t)$ is a nonincreasing one. See the references in [11]. By the results of localization [3] (see also [20, p. 230]) we have the estimate $\operatorname{supp}_x u(x, t)\subset(h_-(0)-L_s/2, h_+(0)+L_s/2)$ for all $t\in(0, T_0)$.

Let $v(x, t)$ be the weak solution of the Cauchy problem for (1.1) in $\mathbf{R}\times(0, T_0)$ with some initial function $v(0, x)=v_0(x)$ satisfying the same hypotheses. We suppose that $v(x, t)$ has bounded connected support for any $t\in(0, T_0)$.

Proof of the main results of this paper are based on the following preliminary lemmas concerning the comparison of intersection method of different solutions $u(x, t)$ and $v(x, t)$ of the equation (1.1) [20, Chap. IV].

The functions $u(x, t_0)$ and $v(x, t_0)$ for fixed $t_0\in[0, T]$ are said to be intersected on the finite interval $[a, b]$, if the difference $w(x, t_0)\equiv u(x, t_0)-v(x, t_0)=0$ in $[a, b]$, $w(x, t_0)\not\equiv 0$ in $[a-\varepsilon, a]$ and $[b, b+\varepsilon]$ and $w(x, t_0)$ changes sign in the interval $[a-\varepsilon, b+\varepsilon]$ for arbitrarily small $\varepsilon>0$. In other words, the finite interval $[a, b]$ is said to be an intersection if $w(x, t_0)\equiv 0$ on $[a, b]$, and for any small $\varepsilon>0$ there exist $x_1\in[a-\varepsilon, a]$, $x_2\in[b, b+\varepsilon]$, such that $w(x_1, t_0)w(x_2, t_0)<0$. If $a=b$ then this is the point of intersection. However, for the quasilinear degenerate parabolic equation (1.1) with weak compactly supported solutions there exist intervals of intersection.

We shall denote by $N(t)$, the number of intersections for fixed $t\in[0, T_0)$. Evidently, $N(t)$ is equal to the number of sign changes in $\mathbf{R}$ of the function $w(x, t)$.

The first statement of the comparison of intersection is the most general one.

LEMMA 1 (see [20, p. 240]). *Let $N(0)<\infty$. Then $N(t)$ is a nonincreasing function and in particular*

$$N(t)\leqq N(0) \quad \textit{for all } t\in(0, T_0). \tag{4.1}$$

This result is well known for linear and quasilinear parabolic equations (see other such results in [1], [2], [5], [6], [10], [15], [16], [21]).

The second statement is a specific comparison theorem for different weak blowup solutions with the same finite blowup times.

LEMMA 2 (see [20, p. 231]). *Let T_0 be the finite blowup time for $v(x, t)$. Then*

$$\{t \in [0, T_0) \mid \overline{\operatorname{supp}}_x v(x, t) \subset \operatorname{supp}_x u(x, t) \text{ and } v(x, t) \leqq u(x, t) \text{ in } \mathbf{R}\} = \varnothing. \tag{4.2}$$

From the simple properties of weak solutions it follows that if (4.2) is not valid and for some $t_0 \in [0, T_0)$

$$\overline{\operatorname{supp}}_x v(x, t_0) \subset \operatorname{supp}_x u(x, t_0), \quad v(x, t_0) \leqq u(x, t_0) \quad \text{in } \mathbf{R} \tag{4.3}$$

then blowup time for $v(x, t)$ is greater than blowup time for $u(x, t)$. It is easy to show by the strong maximum principle and comparison theorems that under conditions (4.3) there exist small $\tau_1 > 0$, $\tau_2 > 0$ such that $v(x, t_0 + \tau_1 + \tau_2) \leqq u(x, t_0 + \tau_1)$ in $\mathbf{R}$, and hence by the usual comparison theorem $v(x, t + \tau_2) \leqq u(x, t)$ in $\mathbf{R} \times [t_0 + \tau_1, T_0)$ (see [20, p. 231]). Letting $t = T_0 - \tau_2$, we get $v(x, T_0) \leqq u(x, T_0 - \tau_2) < +\infty$ in $\mathbf{R}$, i.e., T_0 is not the blowup time for $v(x, t)$, which leads to a contradiction of the hypothesis of Lemma 2.

5. Proof of Theorem 1. We shall compare the solution of the Cauchy problem (1.1), (1.2) with the explicit solution (3.1) (more precisely, with the two-parametric family of such explicit solutions).

Consider, for fixed arbitrary $\varepsilon > 0$ and $\delta \in \mathbf{R}$, the function $v(x, t; \varepsilon, \delta) \equiv u_*(x - \delta, t + \varepsilon)$, where $\psi(t)$, $\varphi(t)$ satisfy (1.7), (1.9), and $C_0 = C_0(T_0 + \varepsilon)$ is given in (1.8). Then $v(x, t; \varepsilon, \delta)$ is the weak blowup solution in $\mathbf{R} \times (0, T_0)$ of the Cauchy problem for (1.1) with the initial function $v(x, 0; \varepsilon, \delta) \equiv u_*(x - \delta, \varepsilon)$ satisfying hypotheses (1.3). Clearly, $v(x, t; \varepsilon, \delta)$ is the continuous function with respect to x, t and ε, δ. By construction, for any fixed $\varepsilon > 0$ and $\delta \in \mathbf{R}$, solutions $u(x, t)$ and $v(x, t; \varepsilon, \delta)$ have the same blowup time T_0.

Let $N(t; \varepsilon, \delta)$ be the number of intersections of the functions $u(x, t)$ and $v(x, t; \varepsilon, \delta)$.

LEMMA 3. *There exists sufficiently small $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0]$, $\delta \in \mathbf{R}$*

$$\sup_{x \in \mathbf{R}} v(x, 0; \varepsilon, \delta) \equiv v(\delta, 0; \varepsilon, \delta) \equiv u_*(0, \varepsilon) > \sup_{x \in \mathbf{R}} u_0(x) \tag{5.1}$$

and

$$N(0; \varepsilon, \delta) \leqq 2. \tag{5.2}$$

Proof. Estimates (5.1), (5.2) for arbitrary fixed $\varepsilon \in (0, \varepsilon_0]$, $\delta \in \mathbf{R}$ follow from asymptotic behavior of $u_*(x, t)$ near $t = 0$, given in (3.7).

From (5.2) and Lemma 1 it follows that

$$N(t; \varepsilon, \delta) \leqq 2 \quad \text{for } 0 < t < T_0 \quad \text{for any } \varepsilon \in (0, \varepsilon_0], \quad \delta \in \mathbf{R}. \tag{5.3}$$

We now prove that

$$\sup_{x \in \mathbf{R}} u(x, t) \leqq \sup_{x \in \mathbf{R}} v(x, t; \varepsilon, \delta) \equiv v(\delta, t; \varepsilon, \delta) \tag{5.4}$$

for all $t \in [0, T_0)$.

Let

$$t_* = \sup \{s \in (0, T_0) \mid (5.4) \text{ hold for all } 0 \leqq t \leqq s\} < T_0 \tag{5.5}$$

($t_* > 0$ by (5.1)). Then there exists $x_* \in \mathbf{R}$ such that

$$\sup_{x \in \mathbf{R}} u(x, t_*) = u(x_*, t_*)$$

and hence for $\delta = x_*$

$$\sup_{x \in \mathbf{R}} v(x, t_*; \varepsilon, x_*) \equiv v(x_*, t_*; \varepsilon, x_*),$$

i.e., $x = x_*$ is the maximum point for $u(x, t_*)$ and $v(x, t_*; \varepsilon, x_*)$. Without loss of generality we suppose that $x = x_*$ is the point of strict maximum for $u(x, t_*)$, i.e., $u(x, t_*) < u(x_*, t_*)$ in a small deleted neighborhood of $x = x_*$. Evidently, the same assertion holds for the explicit solution $v(x, t_*; \varepsilon, x_*)$. This assumption about zeros or maximum points is valid for arbitrary classical solutions of parabolic equations with sufficiently smooth coefficients. See different results in this direction [2], [5], [15], [21]. Note that all coefficients of (1.1) are smooth (and analytic) functions for $u > 0$ and therefore we may assume that solutions of the Cauchy problems considered are C^∞ (and analytic in x) functions at any point of positivity of the solution [12], [13] (see also [1] where analyticity results were obtained for a wide class of nonlinear degenerate parabolic equation including our quasilinear equation (1.1)). Under these hypotheses any positive maximum of $u(x, t_*)$ is the isolated point. The same holds for any intersection of $u(x, t_*)$ and $v(x, t_*; \varepsilon, \delta)$ if these solutions are strictly positive near the intersection. But the above-mentioned assumptions are not the principal parts of our proof.

Two cases are possible. Certainly, we may assume that $u(x, t_*) \not\equiv v(x, t_*; \varepsilon, x_*)$.

Case 1. $x = x_*$ is the point of *inflection* (see [6], [10]) of functions $u(x, t_*)$ and $v(x, t_*; \varepsilon, x_*)$, i.e., $x = x_*$ is the isolated point of intersection and the function $w(x, t_*; \varepsilon, x_*) \equiv u(x, t_*) - v(x, t_*; \varepsilon, x_*)$ changes sign in any small neighborhood of the point $x = x_*$ and

$$w(x_*, t_*; \varepsilon, x_*) = w_x(x_*, t_*; \varepsilon, x_*) = 0. \tag{5.6}$$

Since $w(x, t_*; \varepsilon, x_*)$ is not of constant sign near $x = x_*$, we suppose that $w(x, t_*; \varepsilon, x_*) > 0$ in a small left neighborhood of $x = x_*$ and $w(x, t_*; \varepsilon, x_*) < 0$ in a small right one. Then there exist $x_1 < x_* < x_2$ such that

$$w(x_1, t_*; \varepsilon, x_*) > 0, \qquad w(x_2, t_*; \varepsilon, x_*) < 0. \tag{5.7}$$

Fix small $\lambda > 0$ and consider the function $v(x, t_*; \varepsilon, x_* - \lambda)$. Since $x = x_*$ is the point of strict maximum for functions $u(x, t_*)$ and $v(x, t_*; \varepsilon, x_*)$ we can see that the difference $w(x, t_*; \varepsilon, x_* - \lambda) \equiv u(x, t_*) - v(x, t_*; \varepsilon, x_* - \lambda)$ satisfies

$$w(x_*, t_*; \varepsilon, x_* - \lambda) > 0, \qquad w(x_* - \lambda, t_*; \varepsilon, x_* - \lambda) < 0. \tag{5.8}$$

The function $v(x, t_*; \varepsilon, x_* - \lambda)$ is continuous with respect to λ. Hence, inequalities (5.7) hold for any small $\lambda > 0$, i.e.,

$$w(x_1, t_*; \varepsilon, x_* - \lambda) > 0, \qquad w(x_2, t_*; \varepsilon, x_* - \lambda) < 0. \tag{5.9}$$

From (5.8), (5.9) we see that $w(x, t_*; \varepsilon, x_* - \lambda)$ has at least three sign changes in the interval (x_1, x_2) near the point $x = x_*$, i.e., $N(t_*; \varepsilon, x_* - \lambda) \geqq 3$, which contradicts (5.3) for $\delta = x_* - \lambda$.

Case 2. $x = x_*$ is the point of *tangency* of the functions $u(x, t_*)$ and $v(x, t_*; \varepsilon, x_*)$, and in particular $w(x_*, t_*; \varepsilon, x_*) = 0$ but $w(x, t_*; \varepsilon, x_*)$ does not change sign near $x = x_*$.

If $v(x, t_*; \varepsilon, x_*) \geqq u(x, t_*)$ near $x = x_*$ and $v(x, t_*; \varepsilon, x_*) \not\equiv u(x, t_*)$, then by the strong maximum principle and the usual comparison theorem the same inequality

holds for any $t > t_*$, where $t - t_*$ is sufficiently small, which contradicts the choice of t_* in (5.5).

Thus, it remains to consider the case where $v(x, t_*; \varepsilon, x_*) \leqq u(x, t_*)$ and $v(x, t_*; \varepsilon, x_*) \not\equiv u(x, t_*)$ in a small neighborhood of $x = x_*$. Then we can choose $x_1 < x_* < x_2$ such that

$$w(x_1, t_*; \varepsilon, x_*) > 0, \qquad w(x_2, t_*; \varepsilon, x_*) > 0. \tag{5.10}$$

If $v(x, t_*; \varepsilon, x_*) \leqq u(x, t_*)$ in $\mathbf{R}$ and $\overline{\operatorname{supp}}_x v(x, t_*; \varepsilon, x_*) \subset \operatorname{supp}_x u(x, t_*)$, then we get a contradiction to Lemma 2 since, by construction, T_0 is the finite blowup time for $u(x, t)$ and $v(x, t; \varepsilon, x_*)$.

Consider now the case where

$$v(x, t_*; \varepsilon, x_*) \leqq u(x, t_*) \quad \text{in } \mathbf{R}, \tag{5.11}$$

i.e., $N(t_*; \varepsilon, x_*) = 0$ (the difference $w(x, t_*; \varepsilon, x_*)$ does not change sign in $\mathbf{R}$), but the condition $\overline{\operatorname{supp}}_x v(x, t_*; \varepsilon, x_*) \subset \operatorname{supp}_x u(x, t_*)$ is not valid. Clearly, (5.11) means that $\operatorname{supp}_x v(x, t_*; \varepsilon, x_*) \subseteq \operatorname{supp}_x u(x, t_*)$ and one or two boundary points of the supports $\operatorname{supp}_x v(x, t_*; \varepsilon, x_*) \equiv (x_* - g(t_*), x_* + g(t_*))$ (see (3.5)) and $\operatorname{supp}_x u(x, t_*) \equiv (h_-(t_*), h_+(t_*))$ coincide. For instance, let the right heat fronts of $v(x, t_*; \varepsilon, x_*)$ and $u(x, t_*)$ coincide, i.e., $x_* + g(t_*) = h_+(t_*)$. Then we can obtain that for any small $\lambda > 0$ the number of intersections $N(t_*; \varepsilon, x_* + \lambda)$ of the functions $v(x, t_*; \varepsilon, x_* + \lambda)$ and $u(x, t_*)$ satisfies the inequality $N(t_*; \varepsilon, x_* + \lambda) \geqq 3$. Evidently, since $\operatorname{supp}_x v(x, t_*; \varepsilon, x_* + \lambda) = (x_* + \lambda - g(t_*), x_* + \lambda + g(t_*))$ and hence $w(h_+(t_*), t_*; \varepsilon, x_* + \lambda) \equiv -v(h_+(t_*), t_*; \varepsilon, x_* + \lambda) < 0 = u(h_+(t_*), t_*)$, one new intersection arises near the right heat front $x = h_+(t_*)$ of the function $u(x, t_*)$. Using the second inequality (5.10), we get that for any small $\lambda > 0$ there exists at least one intersection of the functions $v(x, t_*; \varepsilon, x_* + \lambda)$ and $u(x, t_*)$ on the interval $(x_2, h_+(t_*))$. Two new intersections arise near the point $x = x_*$. Since $x = x_*$ is the point of strict maximum of the functions $v(x, t_*; \varepsilon, x_*)$ and $u(x, t_*)$, it follows that

$$w(x_* + \lambda, t_*; \varepsilon, x_* + \lambda) < 0. \tag{5.12}$$

By continuity of $w(x, t_*; \varepsilon, x_* + \lambda)$ with respect to λ, inequalities (5.10) remain valid for all small $\lambda > 0$, i.e.,

$$w(x_1, t_*; \varepsilon, x_* + \lambda) > 0, \qquad w(x_2, t_*; \varepsilon, x_* + \lambda) > 0. \tag{5.13}$$

From (5.12), (5.13) we get that $w(x, t_*; \varepsilon, x_* + \lambda)$ has at least two sign changes on the interval (x_1, x_2).

Thus, $N(t_*; \varepsilon, x_* + \lambda) \geqq 3$ for all sufficiently small $\lambda > 0$ and we arrive at a contradiction to (5.3) for $\delta = x_* + \lambda$.

Consider the last situation where $v(x, t_*; \varepsilon, x_*) \leqq u(x, t_*)$ in a small neighborhood of the point $x = x_*$ and $N(t_*; \varepsilon, x_*) \geqq 1$. Then we can see by the same "λ-translation the x-axis of the function $v(x, t_*; \varepsilon, x_*)$" that two new intersections arise near the maximum point $x = x_*$. Together with the intersections for $\lambda = 0$ which do not disappear for any small $\lambda > 0$ this leads to the estimate $N(t_*; \varepsilon, x_* + \lambda) \geqq 3$, which contradicts (5.3) for $\delta = x_* + \lambda$.

Thus, $t_* = T_0$, and (5.4) holds for all $t \in [0, T_0)$. Letting $\varepsilon \to 0$ and $\delta = 0$, we get (1.6) which completes the proof of Theorem 1.

Using (3.3), (3.4), we obtain results for Corollaries 1 and 2. Evidently, these estimates are the best possible since the explicit solution $u_*(x, t)$ satisfies (1.11) and (1.13) with equality signs instead of inequality signs.

Remark 1. Equation (1.1) admits the weak blowup self-similar solution of the form (see, e.g., [20, p. 175])

$$u_A(x, t) = (T_0 - t)^{-1/\sigma} \cdot \theta(x),$$

where

$$\theta(x) = \left(\frac{2(\sigma+1)}{\sigma(\sigma+2)} \cos^2\left(\frac{\pi x}{L_s}\right)\right)^{1/\sigma} \quad \text{for } |x| < L_s/2 = \pi(\sigma+1)^{1/2}/\sigma,$$

$$\theta(x) = 0 \quad \text{for } |x| \geqq L_s/2.$$

This solution is localized in the bounded domain $\{|x| < L_s/2\}$, $u_A(x, t) = 0$ in $\mathbf{R}\backslash\{|x| < L_s/2\}$ for all $t \in (0, T_0)$ and $u_A(0, t) = [A_0^{-1}(T_0 - t)]^{-1/\sigma}$ for $t \in (0, T_0)$. The last equality coincides with the right-hand side of (1.13).

Remark 2. It follows from the proof of the Theorem 1 that the estimate (1.6) holds for any blowup solution $u(x, t)$ such that $\sup_x u(x, t)$ is attained at a finite point $x = x_* \in \mathbf{R}$ for arbitrarily fixed $t \in [0, T_0)$. Hence, for instance, Theorem 1 is valid for an arbitrary initial function u_0 such that $u(x, t) \to 0$ as $|x| \to \infty$ for any fixed $t \in [0, T_0)$, i.e., (1.6) holds for more general initial functions u_0 than those with compact support.

Acknowledgments. The author thanks the referees for useful suggestions.

REFERENCES

[1] S. ANGENENT, *Local existence and regularity for a class of degenerate parabolic equations*, Math. Ann., 280 (1988), pp. 465-482.

[2] ———, *The zero set of a solution of a parabolic equation*, J. Reine Angew. Math., 390 (1988), pp. 79-96.

[3] V. A. GALAKTIONOV, *On exact upper estimate of amplitude and support of unbounded solutions for the nonlinear heat equation with source*, Keldysh Inst. Appl. Math. Acad. Sci. USSR, 72, 1989. (In Russian.)

[4] V. A. GALAKTIONOV, S. P. KURDJUMOV, A. P. MIKHAILOV, AND A. A. SAMARSKII, *On unbounded solutions of the Cauchy problem for parabolic equation* $u_t = \nabla(u^\sigma \nabla u) + u^\beta$, Dokl. Akad. Nauk SSSR, 252 (1980), pp. 1362-1364. (In Russian.)

[5] V. A. GALAKTIONOV AND S. A. POSASHKOV, *New variants of using of strong maximum principle for parabolic equations and some applications*, Keldysh Inst. Appl. Math. Acad. Sci. USSR, 167, 1985. (In Russian.)

[6] ———, *Applications of new comparison theorems for unbounded solutions of nonlinear parabolic equations*, Differentsial'nye Uravneniya, 22 (1986), pp. 1165-1173. (In Russian.)

[7] ———, *On some method of investigation of unbounded solutions of quasilinear parabolic equations*, J. Vychislit. Mat. Mat. Fiz., 28 (1988), pp. 842-854. (In Russian.)

[8] ———, *Blow-up explicit solution of the nonlinear heat equation with source*, Keldysh Inst. Appl. Math. Acad. Sci. USSR, 42, 1988. (In Russian.)

[9] ———, *On new explicit solutions of parabolic equations with quadratic nonlinearities*, J. Vychislit. Mat. Mat. Fiz., 29 (1989), pp. 497-506. (In Russian.)

[10] ———, *Any large solution of nonlinear heat conduction equation becomes monotonic in time*, Proc. Roy. Soc. Edinburgh (1991), to appear.

[11] A. S. KALASHNIKOV, *Some questions of qualitative theory of nonlinear degenerate parabolic equations of the second order*, Uspekhi Mat. Nauk, 42 (1987), pp. 135-176. (In Russian.)

[12] G. KINDERLEHRER AND L. NIRENBERG, *Analyticity at the boundary of solutions of nonlinear second-order parabolic equations*, Comm. Pure Appl. Math., 31 (1978), pp. 283-338.

[13] G. KOMATSU, *Analyticity up to the boundary of solutions of nonlinear parabolic equations*, Comm. Pure Appl. Math., 32 (1979), pp. 669-720.

[14] H. A. LEVINE AND P. E. SACKS, *Some existence and nonexistence theorems for solutions of degenerate parabolic equations*, J. Differential Equations, 52 (1984), pp. 135-161.

[15] H. MATANO, *Nonincrease of the lap number of a solution for a one-dimensional semi-linear parabolic equation*, J. Fac. Sci. Univ. Tokyo, Sect. IA Math., 29 (1982), pp. 401-441.

[16] K. NICKEL, *Gestaltaussagen über Lösungen parabolischer Differentialgleichungen*, J. Reine Angew. Math., 211 (1962), pp. 78–94.

[17] O. A. OLEINIK, *On equations of the type of nonstationary filtration*, Dokl. Akad. Nauk SSSR, 113 (1957), pp. 1210–1213. (In Russian.)

[18] O. A. OLEINIK, A. S. KALASHNIKOV, AND CHOU YU-LIN, *The Cauchy problem and boundary problems for equations of the type of nonstationary filtration*, Izv. Acad. Nauk SSSR, Ser. Mat., 22 (1958), pp. 667–704. (In Russian.)

[19] P. E. SACKS, *The initial and boundary value problem for a class of degenerate parabolic equations*, Comm. Partial Differential Equations, 8 (1983), pp. 693–733.

[20] A. A. SAMARSKII, V. A. GALAKTIONOV, S. P. KURDJUMOV, AND A. P. MIKHAILOV, *Blow-up in Problems for Quasilinear Parabolic Equations*, Nauka, Moscow, 1987. (In Russian.)

[21] D. H. SATTINGER, *On the total variation of solutions of parabolic equations*, Math. Ann., 183 (1969), pp. 78–92.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1303–1331, September 1991

007

INVERSE PROBLEMS IN MULTIDIMENSIONS*

LI-YENG SUNG† AND A. S. FOKAS†

Abstract. The direct and inverse problems associated with off-diagonal $N \times N$ matrix-valued potentials in $\mathbb{R}^{n+1}$ are studied rigorously. For $N = 2$ and $n = 1$, they are related to the Davey–Stewartson II equation. The components of the potentials are assumed to be rapidly decreasing Schwartz functions. Under small norm conditions, both the direct and the inverse problems are shown to be uniquely solvable. The inverse data are characterized by certain nonlinear equations. Furthermore, the general problem of reconstructing the potential is reduced to the reconstruction of 2×2 potentials.

Key words. inverse problems, multidimensions

AMS(MOS) subject classification. 35R30

1. Introduction.

In this paper we shall study rigorously an inverse problem formally discussed in [5]. Let $Q(x_0, x)$ be an off-diagonal $N \times N$-matrix-valued function defined on $\mathbb{R}^{n+1}$. To facilitate the discussion, we shall assume that the components of Q are rapidly decreasing Schwartz functions. The space of such matrix-valued functions will be denoted by $S^o_{N\times N}$. We want to reconstruct Q from solutions of the following system:

$$\psi_{x_0} + \sigma \sum_{\ell=1}^{n} J_\ell \psi_{x_\ell} = Q\psi, \qquad \sigma = \sigma_R + i\sigma_I, \quad \sigma_I \neq 0, \quad n > 1, \tag{1.1}$$

where the J_ℓ's are constant real diagonal $N \times N$ matrices and ψ is an $N \times N$-matrix-valued function.

We assume that $n \leq N$, otherwise (1.1) can be reduced to a problem with fewer variables. In order to reconstruct the potential Q we look for bounded solutions of (1.1) in the following form:

$$\psi(x_0, x, k) = \mu(x_0, x, k) \exp\left[i \sum_{\ell=1}^{n} k_\ell (x_\ell I - \sigma x_0 J_\ell) \right], \tag{1.2}$$

where $k \in \mathbb{C}^n$ and I is the $N \times N$ identity matrix. Direct substitution yields the following equation for the Jost function μ:

$$\mu_{x_0} + \sigma \sum_{\ell=1}^{n} (J_\ell \mu_{x_\ell} + i k_\ell [J_\ell, \mu]) = Q\mu. \tag{1.3}$$

In view of the decay of Q, it is natural to require that

$$\lim_{|x_0|+|x|\to\infty} \mu = I. \tag{1.4}$$

*Received by the editors June 4, 1990; accepted for publication November 23, 1990. This work was partially supported by the Office of Naval Research grant N00014-88K-0447, National Science Foundation grant DMS-8803471, and Air Force Office of Scientific Research grant 87-0310.

†Department of Mathematics and Computer Science, Clarkson University, Potsdam, New York 13699.

In §2 we consider the direct problem: In Proposition 2.1 we show that if $Q \in S^o_{N\times N}$ and certain norms of Q are small enough (see (2.11)), then (1.3) with the boundary condition (1.4) has a unique bounded solution for each $k \in \mathbb{C}^n$. In Propositions 2.2 and 2.3 we study the regularity of μ and its asymptotic behavior for large k. The inverse data T is introduced in Definition 2.2, and in Proposition 2.4 we derive the $\bar\partial$ equations expressing $\partial\mu/\partial\bar k_p$ in terms of μ and T. We show in Proposition 2.5 that T satisfies the characterization equations, i.e., $2(n-1)$ constraints. In Definitions 2.3 and 2.4 we define the Fréchet space $D^o_{N\times N}$ of inverse data and the scattering map § that takes Q to T. The results of §2 are summarized in Theorem 2.1.

In §3 we study the inverse problem: In Proposition 3.1 we show that if $T \in D^o_{N\times N}$ and certain norms of T are small enough (see (3.8)), then the $\bar\partial$ equations have unique bounded solutions μ_p $(1 \le p \le n)$. The regularity of μ_1 is discussed in Proposition 3.2. In Proposition 3.3 we show that $\mu_1 = \mu_2 = \cdots = \mu_n$, and hence the $\bar\partial$ equations have a unique simultaneous solution μ. The asymptotics of μ for large k is studied in Proposition 3.4. In Propositions 3.5 and 3.6 and Corollary 3.1 we show that if Q_p is defined in terms of μ via

$$Q_p = \frac{i\sigma}{\pi}\left[J_p, \int_{\mathbb{R}^2} \frac{\partial\mu}{\partial\bar k_p}(x_0, x, k)\, dk_{p_R}\, dk_{p_I}\right], \qquad 1 \le p \le n,$$

then $Q_1 = Q_2 = \cdots = Q_n = Q$ (Q is therefore independent of k), $Q \in S^o_{N\times N}$, and Q and μ solve (1.3)–(1.4). We define the inverse scattering map $\hat{\mathbf{S}}$, which takes T to Q in Definition 3.2. The results of §3 are summarized in Theorem 3.1.

In §4 we study the relation between $\mathbf{S}$ and $\hat{\mathbf{S}}$: We show in Proposition 4.1 that $\mathbf{S}$ and $\hat{\mathbf{S}}$ are inverses of each other. As a result, we obtain Theorem 4.1, which states that $\mathbf{S}$ is a homeomorphism from a neighborhood of $0 \in S^o_{N\times N}$ onto a neighborhood of $0 \in D^o_{N\times N}$.

In §5 we consider the problem of reducing the reconstruction of $N \times N$ potentials to the case of 2×2 potentials: In Proposition 5.1 and Corollary 5.1 we show that the 2×2 matrix

$$\hat\mu = \lim_{k\to\infty} \begin{pmatrix} \mu_{aa} & \mu_{ab} \\ \mu_{ba} & \mu_{bb} \end{pmatrix}$$

(the limit is taken in a certain direction determined by the indices a and b) satisfies (1.3)–(1.4) where the coefficient matrices $\hat J_\ell$ and the potential $\hat Q$ are the restrictions of J_ℓ and Q to the aa, ab, ba, and bbth components, and the inverse data $\hat T$ of the 2×2 system can be obtained from the limits of the inverse data of the original system. In Proposition 5.2 we obtain an equivalent characterization of inverse data. An efficient way of reconstructing the potential is then discussed.

There exist several physically significant cases of the above problem: (1) Equation (1.1) with $\sigma = i$, $N = 2$, $n = 1$, and $Q_{21} = \pm\overline{Q}_{12}$ can be used to integrate the Davey–Stewartson (DS) II equation. This equation is a two-spatial dimensional generalization of the celebrated nonlinear Schrödinger equation. DSII was discussed formally in [4] and [6] and rigorously in [1]. Furthermore, it is shown in [2] that for the case of $Q_{21} = \overline{Q}_{12}$, the small norm assumption for both the direct and the inverse problem can be relaxed. (2) Equation (1.1) with $\sigma = 1$, $N = 2$, $n = 1$, and $Q_{21} = \pm\overline{Q}_{12}$ is associated with the DSI equation while (1.1) with $\sigma = 1$, $n = 1$, and arbitrary N is associated with the N-wave interaction equations (see [4], [9]). Although the case $\sigma = 1$ can be considered as the limit $\sigma \to 1 + i0^+$, it is more convenient to study this case directly. This problem is rigorously discussed in [19].

The novelty associated with inverse problems in greater than two spatial dimensions stems from the fact that the inverse data depends on more variables than the potential. Indeed, in the case studied here the potential $Q(x_0, x)$ depends on $n+1$ real variables, while the inverse data $T(k_R, k_I, m)$, $k_R \in \mathbb{R}^n$, $k_I \in \mathbb{R}^n$, $m \in \mathbb{R}^{n-1}$, depends on $3n-1$ variables. This has important implications: (1) The inverse data must be appropriately constrained. Actually, the "characterization" of the inverse data now becomes the central problem. Such a characterization problem was first considered by Faddeev [3] in connection with the multidimensional Schrödinger equation, and later by Newton [13]–[17], who introduced the so-called miracle condition. This problem has recently been studied using the $\bar{\partial}$ approach [10], [11], [18], which is the approach followed here. (2) The existence of redundant scattering parameters can be used to simplify the problem of reconstruction. Indeed, in the case of the Schrödinger equation, this redundancy yields the well-known Born "approximation," which implies that the potential can be reconstructed in closed form. The novelty associated with (1.1) is that, although the reconstruction problem can be simplified, it remains nontrivial: We reduce the general problem of reconstructing an $N \times N$ potential Q in $n+1$ dimensions to one with $N = n = 2$. Henkin and Novikov [7] have subsequently shown that a similar situation arises in a number of physically important cases.

Equation (1.1) was also discussed in [12], where the characterization problem was formally solved but the reconstruction of Q remained open (see the discussion in [5]).

2. The direct problem. We will make the following nondegeneracy assumptions on the coefficient matrices J_ℓ.

$$J_1^a \neq 0 \quad \text{for } 1 \leq a \leq N. \tag{2.1}$$

$$J_\ell^a - J_\ell^b \neq 0 \quad \text{for } 1 \leq \ell \leq n \text{ and } 1 \leq a < b \leq N. \tag{2.2}$$

For the moment we shall concentrate on the abth component of (1.3). It can be simplified by the following change of variables:

$$\begin{aligned} x_0 &= 2y_0, \\ x_1 &= 2J_1^a[\sigma_R y_0 + \sigma_I y_1], \\ x_\ell &= y_\ell + 2J_\ell^a[\sigma_R y_0 + \sigma_I y_1], \quad 2 \leq \ell \leq n. \end{aligned} \tag{2.3}$$

In the y-coordinates, the equation for the abth component in (1.3) is

$$\frac{1}{2}\left[\frac{\partial}{\partial y_0} + i\frac{\partial}{\partial y_1}\right]\tilde{\mu}_{ab} + i\sigma k_{ab}\tilde{\mu}_{ab} = [\widetilde{Q\mu}]_{ab}, \tag{2.4}$$

where $k_{ab} = \sum_{\ell=1}^n k_\ell \left(J_\ell^a - J_\ell^b\right)$. Note that the partial differential operator on the left-hand side of (2.4) is just the $\bar{\partial}$ operator for the complex variable $z = y_0 + iy_1$.

There exists a bounded fundamental solution

$$G(z, k_{ab}) = \frac{1}{\pi z} \cdot \exp\left[-i\left(\sigma k_{ab}\bar{z} + \overline{\sigma k_{ab}} z\right)\right] \tag{2.5}$$

for the operator $\partial/\partial\bar{z} + i\sigma k_{ab}$. Observe that

$$\lim_{|z|\to\infty} G(z, k_{ab}) = 0. \tag{2.6}$$

Therefore, if (2.4) is satisfied (in the sense of distributions) by a continuous and bounded μ, then it will follow from (2.6), (1.4), and Liouville's theorem that

$$\tilde{\mu}_{ab}(y_0,y,k)=\delta^{ab}+\int_{\mathbb{R}^2}G(\xi,k_{ab})[\widetilde{Q\mu}]_{ab}(y_0-\xi_0,y_1-\xi_1,y_2,\cdots,y_n,k)d\xi_0d\xi_1, \tag{2.7}$$

where $\xi=\xi_0+i\xi_1$ and δ^{ab} is the Kronecker symbol.

In the original coordinates, we have

$$\mu_{ab}(x_0,x,k)=\delta_{ab}+\int_{\mathbb{R}^2}G(\xi,k_{ab})[Q\mu]_{ab}(x_0-2\xi_0,x_1-J_1^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\cdots, x_n-J_n^a(\sigma\bar{\xi}+\bar{\sigma}\xi),k)d\xi_0d\xi_1. \tag{2.8}$$

In summary, if μ is a bounded continuous (generalized) solution of (1.3) with boundary conditions (1.4), then μ satisfies

$$\mu=I+\mathbf{N}_{Q,k}\mu, \tag{2.9}$$

where $\mathbf{N}_{Q,k}: C^b_{N\times N}(\mathbb{R}^{n+1})\longrightarrow C^b_{N\times N}(\mathbb{R}^{n+1})$ is defined by

$$[\mathbf{N}_{Q,k}(\phi)]_{ab}=\int_{\mathbb{R}^2}G(\xi,k_{ab})[Q\phi]_{ab}\Big(x_0-2\xi_0,x_1-J_1^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\cdots, x_n-J_n^a(\sigma\bar{\xi}+\bar{\sigma}\xi)\Big)d\xi_0d\xi_1. \tag{2.10}$$

Here $C^b_{N\times N}(\mathbb{R}^{n+1})$ is the Banach space of bounded continuous $N\times N$-matrix-valued functions on $\mathbb{R}^{n+1}$ equipped with the norm defined by

$$||\phi||_\infty=\max_{1\le a,b\le N}\sup_{\mathbb{R}^{n+1}}|\phi_{ab}(x_0,x)|.$$

Conversely, if μ is a solution of (2.9) in $C^b_{N\times N}(\mathbb{R}^{n+1})$, then μ is a generalized solution of (1.3). Moreover, the decay of $Q(x_0,x)$ and (2.6) imply that μ will also satisfy the boundary condition (1.4).

PROPOSITION 2.1. (Existence and uniqueness of the solution of the direct problem.) *If Q satisfies*

$$||Q||_\infty\cdot\max_{1\le a,b\le N}\sup_{\mathbb{R}^{n+1}}\int_{\mathbb{R}^2}|Q_{ab}\big(x_0-2\xi_0,x_1-J_1^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\cdots, x_n-J_n^a(\sigma\bar{\xi}+\bar{\sigma}\xi)\big)|d\xi_0d\xi_1<\frac{\pi}{8(N-1)^2}, \tag{2.11}$$

then equation (2.9) *has a unique solution in* $C^b_{N\times N}(\mathbb{R}^{n+1})$.

Proof. For each (a,b),

$$\sup_{\mathbb{R}^{n+1}}|[\mathbf{N}_{Q,k}\phi]_{ab}|\le||\phi||_\infty\sum_{j=1}^N\int_{\mathbb{R}^2}|G(\xi,k_{ab})Q_{aj}\big(x_0-2\xi_0,x_1-J_1^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\cdots, x_n-J_n^a(\sigma\bar{\xi}+\bar{\sigma}\xi)\big)|d\xi_0d\xi_1.$$

Note that there are only $(N-1)$ terms on the right-hand side of the inequality because $Q(x_0,x)$ is off-diagonal. For simplicity of expression, we shall suppress the arguments of the integrand Q_{aj}. From (2.5), we have

$$\int_{\mathbb{R}^2} |G(\xi,k_{ab})Q_{aj}|d\xi_0 d\xi_1 \le \int_{|\xi|\le r} \frac{|Q_{aj}|}{\pi|\xi|}d\xi_0 d\xi_1 + \int_{|\xi|\ge r} \frac{|Q_{aj}|}{\pi|\xi|}d\xi_0 d\xi_1$$
$$\le 2A\cdot r + \frac{B}{\pi r},$$

where

$$A = ||Q||_\infty \quad \text{and}$$
$$B = \max_{1\le a,b\le N}\left[\sup_{\mathbb{R}^{n+1}}\int_{\mathbb{R}^2} |Q_{ab}(x_0-2\xi_0, x_1 - J_1^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\cdots,$$
$$x_n - J_n^a(\sigma\bar{\xi}+\bar{\sigma}\xi))|d\xi_0 d\xi_1\right].$$

The minimum of $2Ar + \frac{B}{\pi r}$ is attained at $r = \sqrt{B/2A\pi}$. For this choice of r, we have $\int_{\mathbb{R}^2} |G(\xi,k_{ab})Q_{aj}|d\xi_0 d\xi_1 \le \sqrt{8AB/\pi}$ and the proposition follows from the contraction mapping theorem. □

From now on we assume that (2.11) holds. Therefore there exists a number τ such that

$$||\mathbf{N}_{Q,k}|| \le \tau < 1 \quad \text{for all } k\in\mathbb{C}^n. \tag{2.12}$$

So far k is acting as a parameter, but we can also consider $\mu(x_0,x,k)$ as the unique solution in $C^b_{N\times N}(\mathbb{R}^{n+1}\times\mathbb{C}^n)$ of

$$\mu = I + \mathbf{N}_Q\mu, \tag{2.13}$$

where $\mathbf{N}_Q : C^b_{N\times N}(\mathbb{R}^{n+1}\times\mathbb{C}^n) \to C^b_{N\times N}(\mathbb{R}^{n+1}\times\mathbb{C}^n)$ is defined by $(\mathbf{N}_Q\phi)(x_0,x,k) = [\mathbf{N}_{Q,k}\phi(\cdot,\cdot,k)](x_0,x)$. In view of (2.12), $\mathbf{N}_Q$ is also a contraction on $C^b_{N\times N}(\mathbb{R}^{n+1}\times\mathbb{C}^n)$.

In fact, $\mathbf{N}_Q\phi$ makes sense for a wide range of spaces.

DEFINITION 2.1. For each nonnegative integer ℓ, let $X^\ell_{N\times N} = \{\phi : \phi$ is an $N\times N$ matrix-valued function on $\mathbb{R}^{n+1}\times\mathbb{C}^n$ such that (i) ϕ is uniformly continuous on $K\times\mathbb{C}^n$ for any compact subset K of $\mathbb{R}^{n+1}$ and (ii) $\{\sup_{\mathbb{R}^{n+1}\times\mathbb{C}^n}[|\phi_{ab}(x_0,x,k)| \cdot(1+|x_0|+|x|)^{-\ell}] < \infty\}$ for $1\le a,b\le N$. The norm on $X^\ell_{N\times N}$ is defined by $\|\phi\|_\ell = \max_{1\le a,b\le N}\sup_{\mathbb{R}^{n+1}\times\mathbb{C}^n}[|\phi_{ab}(x_0,x,k)|(1+|x_0|+|x|)^{-\ell}]$.

Remark 2.1. It is easy to see that (i) $X^\ell_{N\times N}$ is a Banach space, (ii) $X^0_{N\times N}$ is a closed subspace of $C^b_{N\times N}$, (iii) $X^0_{N\times N}$ contains I and hence μ (see Lemma 2.1 below), and (iv) the natural embedding of $X^\ell_{N\times N}$ into $X^{\ell+1}_{N\times N}$ is bounded.

Moreover, if a bounded sequence $\{\phi_n\}$ in $X^\ell_{N\times N}$ converges uniformly to ϕ on sets of the form $K\times\mathbb{C}^n$ where K is a compact subset of $\mathbb{R}^{n+1}$, then $\phi\in X^\ell_{N\times N}$ and $\lim_{n\to\infty}\phi_n = \phi$ in $X^{\ell+1}_{N\times N}$.

LEMMA 2.1.

(i) $\mathbf{N}_W$ *is a bounded map from* $X^\ell_{N\times N}$ *into* $X^0_{N\times N}$ *for any* $W\in S^o_{N\times N}$.

(ii) $p(x_0,x)\mathbf{N}_W\phi \in X^0_{N\times N}$ *for any* $W\in S^o_{N\times N}$, $\phi\in X^\ell_{N\times N}$ *and polynomial* p *with degree* ≤ 1.

(iii) $(\mathbf{I}-\mathbf{N}_Q) : X^\ell_{N\times N}\to X^\ell_{N\times N}$ *is invertible and* $(\mathbf{I}-\mathbf{N}_Q)^{-1}\phi = \sum_{m=0}^\infty \mathbf{N}_Q^m\phi$.

Proof. (i) and (ii) follow from the decay of W and the definition of $G(\xi,k_{ab})$. Since $\mathbf{N}_Q\phi$ is in $X^0_{N\times N}$ by (i), $\sum_{m=1}^\infty \cdot\mathbf{N}_Q^m\phi$ converges in $X^0_{N\times N}$. Therefore $\sum_{m=1}^\infty$

$\mathbf{N}_Q^m\phi$, and hence $\sum_{m=0}^{\infty}\mathbf{N}_Q^m\phi$ also converges in $X^{\ell}_{N\times N}$. It is then straightforward to show that $(\mathbf{I}-\mathbf{N}_Q)\sum_{m=0}^{\infty}\mathbf{N}_Q^m\phi=\phi=\sum_{m=0}^{\infty}\mathbf{N}_Q^m[(\mathbf{I}-\mathbf{N}_Q)\phi]$. □

LEMMA 2.2. *Let $W_j\in S^o_{N\times N}$, $\partial^\alpha\phi_j\in X^{\ell-1}_{N\times N}$ $(\ell\geq 1)$ for $|\alpha|\leq 1$, and $p_j(x_0,x)$ be a polynomial of degree $\leq\ell$ for $1\leq j\leq m$. Suppose that $\theta\in X^{\ell-1}_{N\times N}$ satisfies*

$$\theta=\sum_{j=1}^{m}p_j\mathbf{N}_{W_j}\phi_j+\mathbf{N}_Q\theta; \tag{2.14}$$

then the following statements hold.

(i) *θ is C^1 on $\mathbb{R}^{n+1}\times\mathbb{C}^n$.*

(ii) *If χ is a first-order derivative of θ in the (x_0,x) variables, then $\chi\in X^{\ell-1}_{N\times N}$ and it satisfies an equation of the form*

$$\chi=\sum_{j=1}^{m'}p'_j\mathbf{N}_{W'_j}\phi'_j+\mathbf{N}_Q\chi, \tag{2.15}$$

where $W'_j\in S^o_{N\times N}$, $\phi'_j\in X^{\ell-1}_{N\times N}$ is either θ or a derivative of the ϕ_j's of order ≤ 1, and $\deg p'_j\leq\ell$.

(iii) *If ψ is a first-order derivative of θ in the k variables, then $\psi\in X^{\ell}_{N\times N}$ and it satisfies an equation of the form*

$$\psi=\sum_{j=1}^{\tilde m}\tilde p_j\mathbf{N}_{\tilde W_j}\tilde\phi_j+\mathbf{N}_Q\psi, \tag{2.16}$$

where $\tilde W_j\in S^o_{N\times N}$, $\tilde\phi_j\in X^{\ell-1}_{N\times N}$ is either θ or a derivative of the ϕ_j's of order ≤ 1, and $\deg\tilde p_j\leq\ell+1$.

Proof. We will only discuss the case of the k_{1_R}-derivative of θ. The same argument can be applied to the other derivatives.

Let $\delta_\epsilon F=\frac{1}{\epsilon}\big[F(x_0,x,k+\epsilon e_1)-F(x_0,x,k)\big]$ for any function $F(x_0,x,k)$, where $e_1=(1,0,\cdots,0)\in\mathbb{C}^n$. From (2.14), we have

$$\delta_\epsilon\theta=\sum_{j=1}^{m}p_j\delta_\epsilon\big(\mathbf{N}_{W_j}\phi_j\big)+\frac{1}{\epsilon}\big(\mathbf{N}_{Q,k+\epsilon e_1}-\mathbf{N}_{Q,k}\big)\theta(x_0,x,k+\epsilon e_1)+\mathbf{N}_Q\delta_\epsilon\theta.$$

Therefore,

$$\delta_\epsilon\theta=\big(\mathbf{I}-\mathbf{N}_Q\big)^{-1}\Psi_\epsilon, \tag{2.17}$$

where $\Psi_\epsilon=[\sum_{j=1}^{m}p_j\delta_\epsilon\big(\mathbf{N}_{W_j}\phi_j\big)+\frac{1}{\epsilon}\big(\mathbf{N}_{Q,k+\epsilon e_1}-\mathbf{N}_{Q,k}\big)\theta(x_0,x,k+\epsilon e_1)]$.

By the decay of W_j and Q, $\{\Psi_\epsilon:|\epsilon|\leq 1\}$ is a bounded subset of $X^{\ell}_{N\times N}$. Let K be a compact subset of $\mathbb{R}^{n+1}$. From (2.5), the decay of W_j and Q, and the assumptions on ϕ_j and its first-order derivatives, it follows that Ψ_ϵ converges uniformly on $K\times\mathbb{C}^n$ as $\epsilon\to 0$ to a sum of the form $\sum_{j=1}^{\tilde m}\tilde p_j\mathbf{N}_{\tilde W_j}\tilde\phi_j$. Each $\tilde W_j$ is the product of one of the W_ℓ's with a polynomial in (x_0,x_1) of degree ≤ 1, $\tilde\phi_j$ is either θ or a derivative of the ϕ_ℓ's of order ≤ 1, and $\deg\tilde p_j\leq\ell+1$. Therefore, Ψ_ϵ converges to $\sum_{j=1}^{\tilde m}\tilde p_j\mathbf{N}_{\tilde W_j}\tilde\phi_j$ in $X^{\ell+1}_{N\times N}$ by Remark 2.2. Part (iii) of Lemma 2.1 and (2.17)

then implies that $\delta_\epsilon\theta$ converges to $\left(\mathbf{I}-\mathbf{N}_Q\right)^{-1}\left(\sum_{j=1}^{\tilde m}\tilde p_j\mathbf{N}_{\tilde W_j}\tilde\phi_j\right)$ in $X^{\ell+1}_{N\times N}$. Hence $\partial\theta/\partial k_{1_R}$ exists and is continuous on $\mathbb{R}^{n+1}\times\mathbb{C}^n$. Since $\sum_{j=1}^{\tilde m}\tilde p_j\mathbf{N}_{\tilde W_j}\tilde\phi_j\in X^{\ell}_{N\times N}$ by part (ii) of Lemma 2.3, $\partial\theta/\partial k_{1_R}$ actually belongs to $X^{\ell}_{N\times N}$. Moreover, $\partial\theta/\partial k_{1_R}$ satisfies

$$\frac{\partial\theta}{\partial k_{1_R}}=\sum_{j=1}^{\tilde m}\tilde p_j\mathbf{N}_{\tilde W_j}\tilde\phi_j+\mathbf{N}_Q\frac{\partial\theta}{\partial k_{1_R}}. \qquad \square$$

We are now ready to prove results for the regularity and growth of $\mu(x_0,x,k)$.

PROPOSITION 2.2. (Regularity and growth of μ.) *Let μ solve* (2.13), *where Q satisfies* (2.11). *Then*

(i) *μ is C^∞ on $\mathbb{R}^{n+1}\times\mathbb{C}^n$, and*

(ii) $\partial^\alpha_{(x_0,x)}\partial^\beta_k\mu\in X^{\max(|\beta|-1,0)}_{N\times N}$.

Proof. We can rewrite (2.13) as

$$(\mu-I)=\mathbf{N}_QI+\mathbf{N}_Q(\mu-I).$$

Lemma 2.2 implies that $\mu-I\in X^0_{N\times N}$. Repeated applications of Lemma 2.2 yield that $\partial^\alpha_{(x_0,x)}\partial^\beta_k\mu\in X^{\max(|\beta|-1,0)}_{N\times N}$ and it satisfies an equation of the form

$$\partial^\alpha_{(x_0,x)}\partial^\beta_k\mu=\sum_{j=1}^{m}p_j\mathbf{N}_{W_j}\phi_j+\mathbf{N}_Q\left(\partial^\alpha_{(x_0,x)}\partial^\beta_k\mu\right), \tag{2.18}$$

where $\deg p_j\le|\beta|$, $W_j\in S^o_{N\times N}$ is the product of a polynomial in (x_0,x_1) of degree $\le|\beta|$ with a derivative of Q of order $\le|\alpha|+|\beta|$ and ϕ_j is either I or a derivative of μ of order $\le|\alpha|+|\beta|-1$. $\square$

Remark 2.2. By tracing the dependence of $\partial^\alpha_{(x_0,x)}\partial^\beta_k\mu$ on Q (cf. (2.18)), it is easy to see that $\|\partial^\alpha_{(x_0,x)}\partial^\beta_k\mu\|_{\max(|\beta|-1,0)}$ only depends on finitely many seminorms of Q of the form $\sup_{\mathbb{R}^{n+1}}\left[(1+|x_0|+|x|)^m|\partial^\gamma Q(x_0,x)|\right]$.

PROPOSITION 2.3. (Asymptotics of μ for large k.) *For fixed (x_0,x) and $(k_1,\cdots,k_{p-1},k_{p+1},\cdots,k_n)$, we have*

$$\lim_{k_p\to\infty}\mu(x_0,x,k)=I, \tag{2.19}$$

$$\lim_{k_p\to\infty}\frac{\partial\mu}{\partial x_\ell}(x_0,x,k)=0,\qquad 0\le\ell\le n, \tag{2.20}$$

$$\mu(x_0,x,k)=I+\frac{1}{k_p}\mu_p+o\left(\frac{1}{|k_p|}\right), \tag{2.21}$$

where $(\mu_p)_{ab}=\left[Q_{ab}(x_0,x)\right]/\left[i\sigma(J^a_p-J^b_p)\right]$ for $a\ne b$.

Proof. Recall that $\mu=\sum_{j=0}^\infty\mathbf{N}^j_QI$ and the convergence is uniform on $\mathbb{R}^{n+1}\times\mathbb{C}^n$. To prove (2.19) it therefore suffices to show that $\lim_{k_p\to\infty}\mathbf{N}^j_QI=0$ for $j\ge1$.

By (2.10) and the Riemann–Lebesgue lemma, $\lim_{k_p\to\infty}\mathbf{N}_QI=0$. The other limits follow inductively from the Lebesgue dominated convergence theorem.

The limit (2.20) can be proved similarly.

We now prove (2.21). From (2.8) we have, for $a \neq b$,

$$\begin{aligned}
&k_p[\mu - I]_{ab} \\
&= \frac{k_p}{i\sigma k_{ab}} \int_{\mathbb{R}^2} i\sigma k_{ab} G(\xi, k_{ab})[Q\mu]_{ab} d\xi_0 d\xi_1 \\
&= \frac{k_p}{i\sigma k_{ab}} \left\{ \int_{\mathbb{R}^2} \left[\left(\frac{\partial}{\partial \bar{\xi}} + i\sigma k_{ab} \right) G(\xi, k_{ab}) \right] [Q\mu]_{ab} d\xi_0 d\xi_1 \right. \\
&\qquad \left. + \int_{\mathbb{R}^2} G(\xi, k_{ab}) \frac{\partial}{\partial \bar{\xi}} \left([Q\mu]_{ab} \right) d\xi_0 d\xi_1 \right\} \\
&= \frac{k_p}{i\sigma k_{ab}} [Q\mu]_{ab}(x_0, x, k) + \frac{k_p}{i\sigma k_{ab}} \int_{\mathbb{R}^2} G(\xi, k_{ab}) \frac{\partial}{\partial \bar{\xi}} \left([Q\mu]_{ab} \right) d\xi_0 d\xi_1.
\end{aligned}$$

It is obvious that $\lim_{k_p \to \infty} (k_p / i\sigma k_{ab})[Q\mu]_{ab}(x_0, x, k) = \left[Q_{ab}(x_0, x) \right] / \left[i\sigma (J_p^a - J_p^b) \right]$. (Recall that $k_{ab} = \sum_{\ell=1}^{n} k_\ell (J_\ell^a - J_\ell^b)$.) On the other hand, the integral

$$\begin{aligned}
\int_{\mathbb{R}^2} G(\xi, k_{ab}) \frac{\partial}{\partial \bar{\xi}} \big([Q\mu]_{ab} \big(& x_0 - 2\xi_0, x_1 - J_1^a(\sigma\bar{\xi} + \bar{\sigma}\xi), \\
& \cdots, x_n - J_n^a(\sigma\bar{\xi} + \bar{\sigma}\xi) \big) \big) d\xi_0 d\xi_1
\end{aligned}$$

converges to zero as $k_p \to \infty$, by (2.19), (2.20), and the Riemann–Lebesgue lemma. Hence we have

$$[\mu - I]_{ab} = \frac{1}{k_p} \frac{Q_{ab}(x_0, x)}{i\sigma (J_p^a - J_p^b)} + o\left(\frac{1}{|k_p|} \right).$$

The asymptotic expansion (2.21) now follows from (2.8) and the fact that $Q(x_0, x)$ is off-diagonal. □

We next introduce the inverse data for problem (1.3)–(1.4) by relating $\partial\mu / \partial \bar{k}_p$ and μ.

By differentiating (2.13), we have

$$\frac{\partial \mu}{\partial \bar{k}_p} = H_p + \mathbf{N}_Q \frac{\partial \mu}{\partial \bar{k}_p}, \tag{2.22}$$

where

$$\frac{\partial}{\partial \bar{k}_p} = \frac{1}{2} \left[\frac{\partial}{\partial k_{p_R}} + i \frac{\partial}{\partial k_{p_I}} \right],$$

and

$$\begin{aligned}
[H_p]_{ab} = \frac{1}{i} \frac{\bar{\sigma}(J_p^a - J_p^b)}{\pi} \int_{\mathbb{R}^2} & \exp\left[-i \left(\sigma k_{ab} \bar{\xi} + \bar{\sigma} \bar{k}_{ab} \xi \right) \right] \\
& \cdot [Q\mu]_{ab}\big(x_0 - 2\xi_0, x_1 - J_1^a(\sigma\bar{\xi} + \bar{\sigma}\xi), \cdots, \\
& x_n - J_n^a(\sigma\bar{\xi} + \bar{\sigma}\xi), k \big) d\xi_0 d\xi_1.
\end{aligned}$$

By a change of variables, we can rewrite $[H_p]_{ab}$ in the following form.

$$\begin{aligned}
[H_p]_{ab} = & \gamma^a (J_p^a - J_p^b) \int_{\mathbb{R}^2} \exp\left[i\beta^a (x_0 - \xi_0, x_1 - \xi_1, k_{ab}) \right] \\
& \cdot [Q\mu]_{ab} \left(\xi_0, \xi_1, x_2 - \frac{J_2^a}{J_1^a}(x_1 - \xi_1), \cdots, x_n - \frac{J_n^a}{J_1^a}(x_1 - \xi_1), k \right) d\xi_0 d\xi_1,
\end{aligned} \tag{2.23}$$

where

$$\beta^a(x_0,x_1,\kappa)=\frac{1}{\sigma_I}\left[|\sigma|^2\kappa_I x_0-\frac{(\sigma\kappa)_I}{J_1^a}x_1\right]\quad\text{and}\quad\gamma^a=\frac{\bar\sigma}{4\pi i J_1^a|\sigma_I|}. \tag{2.24}$$

As a consequence of Proposition 2.2, $[Q\mu]$ is rapidly decreasing in (x_0,x) for fixed k. Therefore, by the Fourier inversion formula, $[H_p]_{ab}$ can be represented as

$$\begin{aligned}[H_p]_{ab}=&\frac{\gamma^a(J_p^a-J_p^b)}{(2\pi)^{n-1}}\int_{\mathbb{R}^{n-1}}\int_{\mathbb{R}^{n+1}}\exp\left[i\beta^a(x_0-\xi_0,x_1-\xi_1,k_{ab})\right.\\&\left.+i\alpha^a(x_1-\xi_1,x^{(2)}-\eta,m)\right]\cdot[Q\mu]_{ab}(\xi_0,\xi_1,\eta,k)d\xi_0d\xi_1d\eta dm,\end{aligned} \tag{2.25}$$

where $m=(m_2,\cdots,m_n)$, $x^{(2)}=(x_2,\cdots,x_n)$, $\eta=(\eta_2,\cdots,\eta_n)$, $d\eta=d\eta_2\cdots d\eta_n$, $dm=dm_2\cdots dm_n$, and

$$\alpha^a(x_1,x^{(2)},m)=\sum_{\ell=2}^n m_\ell\left(x_\ell-x_1\frac{J_\ell^a}{J_1^a}\right). \tag{2.26}$$

DEFINITION 2.2. (Inverse data.) T is an off-diagonal $N\times N$ matrix-valued function defined on $\mathbb{C}^n\times\mathbb{R}^{n-1}$ such that

$$\begin{aligned}T_{ab}(k,m)=&\int_{\mathbb{R}^{n+1}}\exp\left[-i\beta^a(\xi_0,\xi_1,k_{ab})-i\alpha^a(\xi_1,\eta,m)\right]\\&\cdot[Q\mu]_{ab}(\xi_0,\xi_1,\eta,k)d\xi_0d\xi_1d\eta\quad\text{for } a\neq b,\end{aligned} \tag{2.27}$$

where β^a, α^a are defined in (2.24) and (2.26).

Remark 2.3. In view of the decay of Q and Proposition 2.2, $T(k,m)$ is C^∞ on $\mathbb{C}^n\times\mathbb{R}^{n-1}$. For each multi-index α with $3n-1$ components, $\sup_{\mathbb{C}^n\times\mathbb{R}^{n-1}}[(1+|k_{ab}|+|m|)^j|\partial^\alpha T_{ab}(k,m)|]<\infty$ for $j=0,1,2,\cdots$ and $1\le a,b\le N$.

PROPOSITION 2.4. ($\bar\partial$ equation associated with inverse data.) *For* $1\le p\le n$,

$$\begin{aligned}\frac{\partial\mu}{\partial\bar k_p}(x_0,x,k)=&\sum_{a,b=1}^N\frac{\gamma^a(J_p^a-J_p^b)}{(2\pi)^{n-1}}\exp\left[i\beta^a(x_0,x_1,k_{ab})\right]\\&\cdot\int_{\mathbb{R}^{n-1}}\exp\left[i\alpha^a(x,m)\right]T_{ab}(k,m)\mu\big(x_0,x,\lambda^{ab}(k,m)\big)E^{ab}dm,\end{aligned} \tag{2.28}$$

where

$$\lambda_1^{ab}(k,m)=k_1-\left[\frac{(\sigma k_{ab})_I}{\sigma_I J_1^a}+\sum_{\ell=2}^n m_\ell\frac{J_\ell^a}{J_1^a}\right], \tag{2.29}$$

$$\lambda_\ell^{ab}(k,m)=k_\ell+m_\ell\quad\textit{for } 2\le\ell\le n. \tag{2.30}$$

Proof. By using (2.25) and the definition of T_{ab}, $[H]_{ab}$ can be expressed in the following simple form:

$$[H]_{ab}=\frac{\gamma^a(J_p^a-J_p^b)}{(2\pi)^{n-1}}\int_{\mathbb{R}^{n-1}}\exp\left[i\beta^a(x_0,x_1,k_{ab})+i\alpha^a(x,m)\right]T_{ab}(k,m)dm. \tag{2.31}$$

From (2.22), we have

$$\frac{\partial \mu}{\partial \bar{k}_p} = (\mathbf{I} - \mathbf{N}_Q)^{-1} H = \sum_{a,b} (\mathbf{I} - \mathbf{N}_Q)^{-1} (H_{ab} E^{ab}), \tag{2.32}$$

where E^{ab} is the $N \times N$ matrix with zeros in all its entries except the abth, which equals 1. Note that the summation is actually only over $a \neq b$, since $H_{aa} = 0$.

Since $(\mathbf{I} - \mathbf{N}_Q)^{-1}$ is a linear operator in the (x_0, x, k) variable, it commutes with integration in m. Therefore, in order to compute H_{ab}, it suffices to find $F^{ab} = (\mathbf{I} - \mathbf{N}_Q)^{-1} (\exp[i\beta^a(x_0, x_1, k_{ab}) + i\alpha^a(x, m)] E^{ab})$, which is the unique solution of the following equation.

$$F^{ab} = \exp\left[i\beta^a(x_0, x_1, k_{ab}) + i\alpha^a(x, m)\right] E^{ab} + \mathbf{N}_Q F^{ab}. \tag{2.33}$$

Note that the integral equation (2.33) is equivalent to the following boundary value problem:

$$(F^{ab})_{x_0} + \sigma \sum_{\ell=1}^{n} \{J_\ell (F^{ab})_{x_\ell} + i k_\ell [J_\ell, F^{ab}]\} = Q\mu, \tag{2.34}$$

$$\lim_{|x_0|+|x| \to \infty} \exp\left[-i\beta^a(x_0, x, k_{ab}) - i\alpha^a(x, m)\right] F^{ab}(x_0, x, k) = E^{ab}. \tag{2.35}$$

Since F^{ab} satisfies the same equation as μ but with a different boundary condition, it is natural to look for a solution in the form

$$F^{ab} = \exp\left[i\beta^a(x_0, x_1, k_{ab}) + i\alpha^a(x, m)\right] \mu(x_0, x, \lambda^{ab}(k, m)) E^{ab}, \tag{2.36}$$

where the function $\lambda^{ab}(k, m)$ will compensate for the oscillatory factor $\exp[i(\beta^a + \alpha^a)]$. Note that the boundary conditions (2.33) are automatically satisfied since $\lim_{|x_0|+|x| \to \infty} \mu = I$.

Observe that the only nonzero column in $\mu(x_0, x, \lambda^{ab}(k, m)) E^{ab}$ is the bth one with entries identical to those of the ath column of $\mu(x_0, x, \lambda^{ab}(k, m))$. If we substitute (2.36) into (2.34) and use (1.3), we see that (2.34) is reduced to

$$\begin{aligned} \Bigg\{ &\frac{\partial}{\partial x_0} \beta^a(x_0, x_1, k_{ab}) + \sigma \sum_{\ell=1}^{n} J_\ell^r \frac{\partial}{\partial x_\ell} \left[\beta^a(x_0, x_1, k_{ab}) + \alpha^a(x, m)\right] \\ &+ \sigma \sum_{\ell=1}^{n} k_\ell (J_\ell^r - J_\ell^b) - \sigma \sum_{\ell=1}^{n} \lambda_\ell^{ab}(k, m) (J_\ell^r - J_\ell^a) \Bigg\} \\ &\cdot \mu_{ra}(x_0, x, \lambda^{ab}(k, m)) = 0 \quad \text{for } 1 \leq r \leq N. \end{aligned} \tag{2.37}$$

From (2.24) and (2.26) we obtain

$$\begin{aligned}
&\frac{\partial}{\partial x_0}\beta^a(x_0,x_1,k_{ab})+\sigma\sum_{\ell=1}^{n}J_\ell^r\frac{\partial}{\partial x_\ell}\left[\beta^a(x_0,x_1,k_{ab})+\alpha^a(x,m)\right]\\
&\quad=\frac{1}{\sigma_I}|\sigma|^2(k_{ab})_I-\sigma\left[\frac{(\sigma k_{ab})_I}{\sigma_I J_1^a}+\sum_{\ell=2}^{n}m_\ell\frac{J_\ell^a}{J_1^a}\right]J_1^r+\sigma\sum_{\ell=2}^{n}m_\ell J_\ell^r\\
&\quad=\sigma\Bigg\{\frac{\bar\sigma}{\sigma_I}(k_{ab})_I-\left[\frac{(\sigma k_{ab})_I}{\sigma_I J_1^a}+\sum_{\ell=2}^{n}m_\ell\frac{J_\ell^a}{J_1^a}\right]J_1^a+\sum_{\ell=2}^{n}m_\ell J_\ell^a\\
&\qquad\qquad-\left[\frac{(\sigma k_{ab})_I}{\sigma_I J_1^a}+\sum_{\ell=2}^{n}m_\ell\frac{J_\ell^a}{J_1^a}\right](J_1^r-J_1^a)+\sum_{\ell=2}^{n}m_\ell(J_\ell^r-J_\ell^a)\Bigg\}\\
&\quad=\sigma\Bigg\{\frac{\bar\sigma}{\sigma_I}(k_{ab})_I-\frac{(\sigma k_{ab})_I}{\sigma_I}-\left[\frac{(\sigma k_{ab})_I}{\sigma_I J_1^a}+\sum_{\ell=2}^{n}m_\ell\frac{J_\ell^a}{J_1^a}\right](J_1^r-J_1^a)+\sum_{\ell=2}^{n}m_\ell(J_\ell^r-J_\ell^a)\Bigg\}\\
&\quad=\sigma\Bigg\{-k_{ab}-\left[\frac{(\sigma k_{ab})_I}{\sigma_I J_1^a}+\sum_{\ell=2}^{n}m_\ell\frac{J_\ell^a}{J_1^a}\right](J_1^r-J_1^a)+\sum_{\ell=2}^{n}m_\ell(J_\ell^r-J_\ell^a)\Bigg\}.
\end{aligned}$$

Recall that $k_{ab}=\sum_{\ell=1}^{n}k_\ell(J_\ell^a-J_\ell^b)$. Therefore $-k_{ab}+\sum_{\ell=1}^{n}k_\ell(J_\ell^r-J_\ell^b)=\sum_{\ell=1}^{n}k_\ell(J_\ell^r-J_\ell^a)$. At this point, we see that (2.37) follows from (2.29) and (2.30).

We have just proved that

$$\begin{aligned}
&(\mathbf{I}-\mathbf{N}_Q)^{-1}\left(\exp\left[i\beta^a(x_0,x_1,k_{ab})+i\alpha^a(x,m)\right]E^{ab}\right)\\
&\qquad=\exp\left[i\beta^a(x_0,x_1,k_{ab})+i\alpha^a(x,m)\right]\mu\left(x_0,x,\lambda^{ab}(k,m)\right)E^{ab}.
\end{aligned}\tag{2.38}$$

Equation (2.28) now follows from (2.31), (2.32), and (2.38). □

From Proposition 2.4, Remark 2.3, and (1.4), we have the following two corollaries.

COROLLARY 2.1. *$\partial\mu/\partial\bar k_p$ is rapidly decreasing in k_p, with all the other variables fixed.*

COROLLARY 2.2. (T as asymptotics of μ in large x_0.) *For $1\le a,b\le N$ and $a\ne b$,*

$$\begin{aligned}
&\lim_{x_0\to\infty}\left\{\exp[-i\beta^a(x_0,0,k_{ab})]\frac{\partial\mu_{ab}}{\partial\bar k_p}(x_0,0,x_2,\cdots,x_n,k)\right\}\\
&\qquad=\frac{\gamma^a(J_p^a-J_p^b)}{(2\pi)^{n-1}}\int_{\mathbb{R}^{n-1}}\exp\left(\sum_{\ell=2}^{n}m_\ell x_\ell\right)T_{ab}(k,m)\,dm.
\end{aligned}\tag{2.39}$$

COROLLARY 2.3. (Reconstruction of Q.)

$$Q(x_0,x)=\frac{i\sigma}{\pi}\left\{J_p,\int_{\mathbb{R}^2}\frac{\partial\mu}{\partial\bar k_p}(x_0,x,k_1,\cdots,k_{p-1},k_p,k_{p+1},\cdots,k_n)dk_{p_R}dk_{p_I}\right\}\tag{2.40}$$

for $p=1,\cdots,n$.

Proof. From Corollary 2.1 and the inversion formula for the $\bar\partial$ operator, we have

$$\begin{aligned}
&\mu(x_0,x,k)\\
&\quad=I+\frac{1}{\pi}\int_{\mathbb{R}^2}\frac{(\partial\mu/\partial\bar k_p)(x_0,x,k_1,\cdots,k_{p-1},k'_p,k_{p+1},\cdots,k_n)}{k_p-k'_p}dk'_{p_R}dk'_{p_I}\\
&\quad=I+\frac{1}{k_p}\left[\frac{1}{\pi}\int_{\mathbb{R}^2}\frac{\partial\mu}{\partial\bar k_p}(x_0,x,k_1,k_{p-1},\cdots,k'_p,k_{p+1},\cdots,k_n)dk'_{p_R}dk'_{p_I}\right]\\
&\qquad+o\left(\frac{1}{|k_p|}\right).
\end{aligned}\tag{2.41}$$

Equation (2.40) follows by comparing (2.41) with (2.21). □

PROPOSITION 2.5. (Characterization of T.) *For any (a,b) such that $1 \le a,b, \le N$ and $a \ne b$, we have*

$$\mathbf{L}_r^{ab} T_{ab} = \mathbf{L}_p^{ab} T_{ab}, \qquad 1 \le r, \quad p \le n, \tag{2.42}$$

where

$$\begin{aligned} &\big(\mathbf{L}_\ell^{ab} T_{ab}\big)(k,m) \\ &\quad = \frac{1}{J_\ell^a - J_\ell^b} \frac{\partial T_{ab}}{\partial \bar{k}_\ell}(k,m) - \sum_{j=1}^{N} \frac{\gamma^j}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} \Big(\frac{J_\ell^j - J_\ell^b}{J_\ell^a - J_\ell^b}\Big) \\ &\qquad \cdot T_{jb}(k,M) T_{aj}\big(\lambda^{jb}(k,m), m - M\big) dM. \end{aligned} \tag{2.43}$$

Proof. Define

$$\mathbf{L}_{rp}^{ab} = \big(J_p^a - J_p^b\big) \frac{\partial}{\partial \bar{k}_r} - \big(J_r^a - J_r^b\big) \frac{\partial}{\partial \bar{k}_p} \tag{2.44}$$

and

$$\begin{aligned} \varepsilon^{ab}(x_0,x,k,m) &= \beta^a(x_0,x_1,k_{ab}) + \alpha^a(x,m) \\ &= \frac{1}{\sigma_I} \left[|\sigma|^2 (k_{ab})_I x_0 - \frac{(\sigma k_{ab})_I}{J_1^a} x_1 \right] + \sum_{\ell=2}^{n} m_\ell \Big(x_\ell - x_1 \frac{J_\ell^a}{J_1^a} \Big). \end{aligned} \tag{2.45}$$

It follows from (2.29), (2.30), and the definition of k_{ab} that

$$\mathbf{L}_{rp}^{ab} \varepsilon^{ab} = 0, \tag{2.46}$$

and

$$\mathbf{L}_{rp}^{ab} \Big[\mu\big(x_0,x,\lambda^{ab}(k,m)\big) \Big] = \Big[\mathbf{L}_{rp}^{ab} \mu \Big] \big(x_0,x,\lambda^{ab}(k,m)\big). \tag{2.47}$$

Applying (2.28), (2.46), and (2.47), we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial \bar{k}_r} \Big(\frac{\partial \mu}{\partial \bar{k}_p} \Big) - \frac{\partial}{\partial \bar{k}_p} \Big(\frac{\partial \mu}{\partial \bar{k}_r} \Big) \\ &= \sum_{a,b} \frac{\gamma^a}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} \exp\big[i \varepsilon^{ab}(x_0,x,k,m) \big] \\ &\quad \cdot \Big[\big(\mathbf{L}_{rp}^{ab} T_{ab} \big)(k,m) \mu\big(x_0,x,\lambda^{ab}(k,m)\big) \\ &\qquad + T_{ab}(k,m) \big(\mathbf{L}_{rp}^{ab} \mu \big) \big(x_0,x,\lambda^{ab}(k,m)\big) \Big] E^{ab} dm\,. \end{aligned}$$

Therefore, (2.28) implies that

$$\begin{aligned} 0 = &\sum_{a,b} \frac{\gamma^a}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} \exp\big[i \varepsilon^{ab}(x_0,x,k,m) \big] \big(\mathbf{L}_{rp}^{ab} T_{ab} \big)(k,m) \\ &\qquad \cdot \mu\big(x_0,x,\lambda^{ab}(k,m)\big) E^{ab} dm \\ &+ \sum_{a,b} \frac{\gamma^a}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} \exp\big[i \varepsilon^{ab}(x_0,x,k,m) \big] T_{ab}(k,m) \\ &\quad \cdot \Big\{ \sum_{j=1}^{n} \frac{\gamma^j}{(2\pi)^{n-1}} \Big[\big(J_p^a - J_p^b\big)\big(J_r^j - J_r^a\big) - \big(J_r^a - J_r^b\big)\big(J_p^j - J_p^a\big) \Big] \\ &\qquad \cdot \int_{\mathbb{R}^{n-1}} \exp\big[i \varepsilon^{ja}(x_0,x,\lambda^{ab}(k,m),M) \big] T_{ja}\big(\lambda^{ab}(k,m),M\big) \\ &\qquad\quad \cdot \mu\Big(x_0,x,\lambda^{ja}\big(\lambda^{ab}(k,m),M\big)\Big) E^{jb} dM dm \Big\}. \end{aligned} \tag{2.48}$$

It is straightforward (cf. [5]) to show that

$$\lambda^{ja}(\lambda^{ab}(k,m),M) = \lambda^{jb}(k,m+M), \tag{2.49}$$

$$\varepsilon^{ab}(x_0,x,k,m) + \varepsilon^{ja}\big(x_0,x,\lambda^{ab}(k,m),M\big) = \varepsilon^{jb}(x_0,x,k,m+M). \tag{2.50}$$

If we multiply (2.48) by $(2\pi)^{n-1}(\gamma^a)^{-1}\exp\big[-i|\sigma|^2\sigma_I^{-1}(k_{ab})_I x_0\big]$ and then let $x_1 = 0$ and $x_0 \to \infty$ in the resulting equation, we have by looking at the abth component the following equation:

$$\begin{aligned} 0 = {} & \int_{\mathbb{R}^{n-1}} \exp\Big[i\sum_{\ell=2}^{n} m_\ell x_\ell\Big]\big(\mathbf{L}_{rp}^{ab}T_{ab}\big)(k,m)dm \\ & + \sum_{j=1}^{N} \frac{\gamma^j}{(2\pi)^{n-1}}\Big[\big(J_r^a - J_r^j\big)\big(J_p^j - J_p^b\big) - \big(J_p^a - J_p^j\big)\big(J_\gamma^j - J_\gamma^b\big)\Big] \\ & \cdot \int_{\mathbb{R}^{n-1}}\int_{\mathbb{R}^{n-1}} \exp\Big[i\sum_{\ell=2}^{n}(m_\ell + M_\ell)x_\ell\Big]T_{jb}(k,m) \\ & \cdot T_{aj}\big(\lambda^{jb}(k,m),M\big)dMdm. \end{aligned} \tag{2.51}$$

We have used (2.45), (2.49), (2.50), and the boundary conditions (1.4) in deriving (2.51). From the Fourier inversion formula and (2.51), we obtain

$$\begin{aligned} 0 = {} & \big(\mathbf{L}_{rp}^{ab}T_{ab}\big)(k,m) \\ & + \sum_{j=1}^{N} \frac{\gamma^j}{(2\pi)^{n-1}}\Big[\big(J_r^a - J_r^j\big)\big(J_p^j - J_p^b\big) - \big(J_p^a - J_p^j\big)\big(J_r^j - J_r^b\big)\Big] \\ & \cdot \int_{\mathbb{R}^{n-1}} T_{jb}(k,M)T_{aj}\big(\lambda^{jb}(k,M),m-M\big)dM. \end{aligned} \tag{2.52}$$

Since

$$\begin{aligned} & \frac{1}{\big(J_p^a - J_p^b\big)\big(J_r^a - J_r^b\big)}\sum_{j=1}^{N}\Big[\big(J_r^a - J_r^j\big)\big(J_p^j - J_p^b\big) - \big(J_p^a - J_p^j\big)\big(J_r^j - J_r^b\big)\Big] \\ & \quad = \sum_{\ell=1}^{N}\Big[\frac{J_p^\ell - J_p^b}{J_p^a - J_p^b} - \frac{J_r^\ell - J_r^b}{J_r^a - J_r^b}\Big], \end{aligned}$$

equation (2.52) is equivalent to (2.42). □

DEFINITION 2.3. Let $D_{N\times N}^{o}$ be the space of off-diagonal $N\times N$-matrix-valued functions $\Phi(k,m)$ that satisfy the following conditions.

(i) $\Phi(k,m)$ is C^∞ on $\mathbb{C}^n\times\mathbb{R}^{n-1}$.

(ii) For each multi-index α with $3n-1$ components, $\sup_{\mathbb{C}^n\times\mathbb{R}^{n-1}}\big(1+|k_{ab}|+|m|\big)^j\cdot\ |\partial^\alpha\Phi_{ab}(k,m)| < \infty$ for $j = 0,1,2,\cdots$ and $1\le a,b\le N$.

(iii) $\mathbf{L}_r^{ab}\Phi_{ab} = \mathbf{L}_p^{ab}\Phi_{ab}$, $1\le r,\ p\le n$.

$D_{N\times N}^{o}$ becomes a Fréchet space if we equip it with the seminorms

$$|\Phi|_{(a,b,\alpha,j)} = \sup_{\mathbb{C}^n\times\mathbb{R}^{n-1}}\big(1+|k_{ab}|+|m|\big)^j\cdot\ |\partial^\alpha\Phi_{ab}(k,m)|.$$

DEFINITION 2.4. (Scattering map.)

$$\mathbf{S}(Q) = T. \tag{2.53}$$

Remark 2.4. $\mathbf{S}$ is defined on a neighborhood of $0 \in S^o_{N\times N}$. In view of Remark 2.3 and Proposition 2.5, the range of $\mathbf{S}$ is a subset of $D^o_{N\times N}$. By (2.27) and Remark 2.2, $\mathbf{S}$ is continuous with respect to the Fréchet space topologies on $S^o_{N\times N}$ and $D^o_{N\times N}$.

The results of this section are summarized in the following theorem.

THEOREM 2.1. *The direct problem is uniquely solvable if Q satisfies* (2.11). *If the inverse data T is defined by* (2.27), *then $T \in D^o_{N\times N}$ (cf. Definition* 2.3) *and the $\bar\partial$ equation* (2.28) *is valid. The scattering map $\mathbf{S}$ that takes Q to T is a continuous map from a neighborhood of $0 \in S^o_{N\times N}$ into $D^o_{N\times N}$. Both the potential Q and the inverse data T can be expressed directly in terms of the Jost function μ via* (2.39) *and* (2.40).

Remark 2.5. $\mathbf{S}$ is actually infinitely Fréchet differentiable.

3. The inverse problem. Let $T \in D^o_{N\times N}$. The fundamental equations ($1 \le p \le n$) for the inverse problem are

$$\begin{aligned}\frac{\partial \mu_p}{\partial \bar k_p}(x_0,x,k) = \sum_{a,b} \frac{\gamma^a(J^a_p - J^b_p)}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} \exp\left[i\varepsilon^{ab}(x_0,x,k,m)\right] \\ \cdot T_{ab}(k,m)\mu_p\big(x_0,x,\lambda^{ab}(k,m)\big)E^{ab}dm,\end{aligned} \tag{3.1}$$

with boundary conditions

$$\lim_{k_p\to\infty} \mu_p(x_0,x,k) = I \quad \text{for fixed } (x_0,x,k_1,\cdots,k_{p-1},k_{p+1},\cdots,k_n). \tag{3.2}$$

Note that ε^{ab} is defined in (2.45). It is useful to observe that any first-order derivative of ε^{ab} in the (x_0,x) variables is a linear combination of $\{(k_{ab})_R,(k_{ab})_I, m_2,\cdots,m_n\}$ and that each of the (x_0,x) variables can be represented as a linear combination of $\{(\partial\varepsilon^{ab}/\partial k_{p_R}),(\partial\varepsilon^{ab}/\partial k_{p_I}),(\partial\varepsilon^{ab}/\partial m_2),\cdots,(\partial\varepsilon^{ab}/\partial m_n)\}$.

Let the differential-integral operator $\mathbf{L}^p_{(x_0,x)}$ be defined by

$$\begin{aligned}\left[\mathbf{L}^p_{(x_0,x)}\nu\right](k) = \frac{\partial \nu}{\partial \bar k_p}(k) - \sum_{a,b} \frac{\gamma^a(J^a_p - J^b_p)}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} \exp\left[i\varepsilon^{ab}(x_0,x,k,m)\right] \\ \cdot T_{ab}(k,m)\nu\big(\lambda^{ab}(k,m)\big)E^{ab}dm,\end{aligned} \tag{3.3}$$

so that (3.1) can be written in the simple form $\mathbf{L}^p_{(x_0,x)}\mu(x_0,x,\cdot) = 0$.

The goal of the inverse problem is to find a potential whose scattering data is T. We start with the solvability of (3.1)–(3.2).

Henceforth we shall denote $(k_1,\cdots,k_{p-1},k'_p,k_{p+1},\cdots,k_n)$ by $k'_{(p)}$. The following lemma is a direct consequence of the inversion formula for the $\bar\partial$ operator and Liouville's theorem.

LEMMA 3.1. *$\mu_p(x_0,x,k)$ is a bounded continuous generalized solution of* (3.1) *and* (3.2) *if and only if it is a solution of the following integral equation:*

$$\begin{aligned}\mu_p = I + \sum_{a,b} \frac{\gamma^a\big(J^a_p - J^b_p\big)}{\pi(2\pi)^{n-1}} \int_{\mathbb{R}^{n+1}} \frac{1}{k_p - k'_p} \exp\left[i\varepsilon^{ab}(x_0,x,k'_{(p)},m)\right] \\ \cdot T_{ab}(k'_{(p)},m)\mu_p\big(x_0,x,\lambda^{ab}(k'_{(p)},m)\big)E^{ab}dm dk'_{p_R}\,dk'_{p_I}.\end{aligned} \tag{3.4}$$

Let $\mathbf{P}^p_{(x_0,x)}$ be the operator on $C^b_{N\times N}(\mathbb{C}^n)$ defined by

$$[\mathbf{P}^p_{(x_0,x)}\nu](k) = \sum_{a,b} \frac{\gamma^a(J^a_p - J^b_p)}{\pi(2\pi)^{n-1}} \int_{\mathbb{R}^{n+1}} \frac{1}{k_p - k'_p} \exp\left[i\varepsilon^{ab}(x_0, x, k'_{(p)}, m)\right] \cdot T_{ab}(k'_{(p)}, m)\nu\big(\lambda^{ab}(k'_{(p)}, m)\big)E^{ab}dm dk'_{p_R} dk'_{p_I}. \tag{3.5}$$

Remark 3.1. Equation (3.4) is therefore just $\mu_p = I + \mathbf{P}^p_{(x_0,x)}\mu_p$. Note that by property (ii) of Definition 2.3, $\mathbf{P}^p_{(x_0,x)}$ is also defined on any continuous function $\nu(k)$ with polynomial growth. Moreover, for such ν we have

$$\frac{\partial}{\partial \bar{k}_p}(\mathbf{P}^p_{(x_0,x)}\nu)(k) = \sum_{a,b} \frac{\gamma^a(J^a_p - J^b_p)}{(2\pi)^{n-1}} \exp\left[i\beta^a(x_0, x_1, k_{ab})\right] \cdot \int_{\mathbb{R}^{n-1}} \exp\left[i\alpha^a(x,m)\right] T_{ab}(k,m)\nu\big(\lambda^{ab}(k,m)\big)E^{ab}dm, \tag{3.6}$$

and

$$\lim_{k_j\to\infty} \mathbf{P}^p_{(x_0,x)}\nu = 0 \ (1 \le p \le n) \tag{3.7}$$

for fixed $(x_0, x, k_1, \cdots, k_{j-1}, k_{j+1}, \cdots, k_n)$.

A sufficient condition for (3.4) (equivalently (3.1)–(3.2)) to be uniquely solvable is stated in the next proposition whose proof is omitted due to its similarity to the proof of Proposition 2.1.

PROPOSITION 3.1. (Contracting property of $\mathbf{P}^p_{(x_0,x)}$.) *Assume that* $T \in D^o_{N\times N}$ *and that it satisfies*

$$\int_{\mathbb{R}^{n-1}} \sup_{\mathbb{C}^n} |T(k,m)|dm \cdot \int_{\mathbb{R}^{n-1}} \left[\int_{\mathbb{R}^2} \sup_{\mathbb{C}^{n-1}} |T(k,m)|dk_{p_R} dk_{p_I}\right] dm < \frac{\pi}{8\omega_p^2(N-1)^2}, \tag{3.8}$$

where $\omega_p = \max_{a,b}\left(|\gamma^a||J^a_p - J^b_p|\right)/(2\pi)^{n-1}$ *and* $|T(k,m)| = \max_{1\le a,b\le N}|T_{ab}(k,m)|$. *Then* $\mathbf{P}^p_{(x_0,x)}$ *is a contraction on* $C^b_{N\times N}(\mathbb{C}^n)$ *with*

$$\|\mathbf{P}^p_{(x_0,x)}\| \le \tau_p < 1 \quad \forall (x_0, x) \in \mathbb{R}^{n+1}. \tag{3.9}$$

From now on we shall assume that (3.8) holds for $1 \le p \le n$ and denote by $\mu_p(x_0, x, k)$ the unique solution in $C^b_{N\times N}(\mathbb{C}^n)$ of

$$\mu_p = I + \mathbf{P}^p_{(x_0,x)}\mu_p. \tag{3.10}$$

Since the operator $\mathbf{P}^p_{(x_0,x)}$ depends continuously on (x_0, x), μ_p is bounded and continuous on $\mathbb{R}^{n+1}\times\mathbb{C}^n$. It is a solution of (3.1) in the sense of distribution, and we also have by (3.6)

$$\lim_{k_j\to\infty} \mu_p(x_0, x, k) = I, \qquad 1 \le p \le n, \tag{3.11}$$

for fixed $(x_0, x, k_1, \cdots, k_{j-1}, k_{j+1}, \cdots, k_n)$.

By the same argument in the proof of (2.19), we obtain

$$\lim_{|x_0|+|x|\to\infty} \mu_p(x_0,x,k) = I \tag{3.12}$$

for fixed k.

PROPOSITION 3.2. (Properties of μ_1.) $\mu_1 \in C^\infty(\mathbb{R}^{n+1}\times\mathbb{C}^n)$. *All the derivatives of μ_1 that only involve the (x_0,x) variables are bounded on $\mathbb{R}^{n+1}\times\mathbb{C}^n$. All the derivatives that involve the k variables are bounded on $\mathbb{C}^n$ for fixed (x_0,x). All the first-order derivatives of $\mu_1 \to 0$ as $k_p \to \infty$ with all the other variables fixed.*

Proof. We will employ techniques similar to those in the proof of Lemma 2.2 and Proposition 2.2. But special care must be taken for the k_1-derivatives because of the function $\lambda_1^{ab}(k,m)$ (cf. (2.29)) that appears in the definition of $P^1_{(x_0,x)}$.

Starting with (3.1) for $p=1$, it is easy to show by induction that

$$\left(\frac{\partial}{\partial\bar{k}_1}\right)^\ell \mu_1(x_0,x,k) \quad \text{is continuous on } \mathbb{R}^{n+1}\times\mathbb{C}^n \tag{3.13}$$

for $\ell = 1,2,\cdots$, and

$$\left(\frac{\partial}{\partial\bar{k}_1}\right)^\ell \mu_1(x_0,x,k) \in C^b_{N\times N}(\mathbb{C}^n) \tag{3.14}$$

for $\ell = 1,2,\cdots$ and fixed (x_0,x). Here the derivatives are taken in the sense of distributions.

By elliptic regularity (cf. [8]), $\partial\mu_1/\partial k_{1_R}$ and $\partial\mu_1/\partial k_{1_I}$ exist in the classical sense. Let $\tilde{\mathbf{L}} = \sigma_R(\partial/\partial k_{1_R}) - \sigma_I(\partial/\partial k_{1_I})$. It follows from (2.29) that

$$\tilde{\mathbf{L}}\Big[\mu_1\big(x_0,x,\lambda^{ab}(k,m)\big)\Big] = \big(\tilde{\mathbf{L}}\mu_1\big)\big(x_0,x,\lambda^{ab}(k,m)\big). \tag{3.15}$$

By the technique in the proof of Lemma 2.2, $\tilde{\mathbf{L}}\mu_1$ is the solution in $C^b_{N\times N}(\mathbb{C}^n)$ of

$$\tilde{\mathbf{L}}\mu_1 = W + \mathbf{P}^1_{(x_0,x)}\big(\tilde{\mathbf{L}}\mu_1\big), \tag{3.16}$$

where

$$W = \sum_{a,b} \frac{\gamma^a\big(J_1^a - J_1^b\big)}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n+1}} \tilde{\mathbf{L}}\Big\{ \exp\big[i\varepsilon^{ab}(x_0,x,k_1-k_1',k_2,\cdots,k_n,m)\big] \cdot T_{ab}(k_1-k_1',k_2,\cdots,k_n,m)\Big\} \cdot \frac{1}{k_1'}\mu_1\big(x_0,x,\lambda^{ab}(k_1-k_1',k_2,\cdots,k_n,m)\big)E^{ab}dm dk_{1_R}' dk_{1_I}'.$$

Since W depends continuously on (x_0,x), it follows from (3.16) that $\tilde{\mathbf{L}}\mu_1$ is continuous on $\mathbb{R}^{n+1}\times\mathbb{C}^n$. As $\partial\mu_1/\partial k_{1_R}$ and $\partial\mu_1/\partial k_{1_I}$ can be represented as linear combinations of $\partial\mu_1/\partial\bar{k}_1$ and $\tilde{\mathbf{L}}\mu_1$, we see that both $\partial\mu_1/\partial k_{1_R}$ and $\partial\mu_1/\partial k_{1_I}$ are continuous on $\mathbb{R}^{n+1}\times\mathbb{C}^n$ and bounded on $\mathbb{C}^n$ for fixed (x_0,x).

The other first-order derivatives can now be handled easily. The same arguments are then applied inductively as in the proof of Proposition 2.2 to higher-order derivatives.

Let ζ be one of the first-order derivatives of μ_1. Then ζ satisfies an equation of the form

$$\zeta = K + \mathbf{P}^1_{(x_0,x)}\zeta,$$

where by property (ii) of Definition 2.3, $\lim_{k_p\to\infty} K = 0$ with $(x_0, x, k_1, \cdots, k_{p-1}, k_{p+1}, \cdots, k_n)$ fixed. Since ζ is bounded on $\mathbb{C}^n$ for fixed (x_0, x), we also have in view of (3.7) that $\lim_{k_p\to\infty} P^1_{(x_0,x)}\zeta = 0$ for fixed $(x_0, x, k_1, \cdots, k_{p-1}, k_{p+1}, \cdots, k_n)$. Hence $\lim_{k_p\to\infty}\zeta = 0$. □

Our next goal is to prove that $\mu_1 = \mu_2 = \cdots = \mu_n$, actually.

LEMMA 3.2. *If $\nu(k) \in C^2(\mathbb{C}^n)$ and $\partial^\alpha\nu$ is bounded on $\mathbb{C}^n$ for $|\alpha| \le 1$, then*

$$\left[\mathbf{L}^p_{(x_0,x)}, \mathbf{L}^r_{(x_0,x)}\right]\nu = 0. \tag{3.17}$$

Proof. The commutativity of $\mathbf{L}^p_{(x_0,x)}$ and $\mathbf{L}^r_{(x_0,x)}$ is the consequence of the constraint $\mathbf{L}^{ab}_r T_{ab} = \mathbf{L}^{ab}_p T_{ab}$ in the definition of the space $D^o_{N\times N}$ (cf. (iii) of Definition 2.3). The following computations are similar to those in the proof of Proposition 2.5 and we shall use the notation introduced there.

From (2.46) and (2.47), we have

$$\begin{aligned}
&\left[\mathbf{L}^p_{(x_0,x)}, \mathbf{L}^r_{(x_0,x)}\right]\nu \\
&\quad = \mathbf{L}^p_{(x_0,x)}\mathbf{L}^r_{(x_0,x)}\nu - \mathbf{L}^r_{(x_0,x)}\mathbf{L}^p_{(x_0,x)}\nu \\
&\quad = \sum_{a,b}\frac{\gamma^a}{(2\pi)^{n-1}}\int_{\mathbb{R}^{n-1}}\exp\left[i\varepsilon^{ab}(x_0,x,k,m)\right]\left[\mathbf{L}^{ab}_{rp}T_{ab}\right]\nu\left(\lambda^{ab}(k,m)\right)E^{ab}dm \\
&\qquad + \sum_{a,b}\left\{\sum_j \frac{\gamma^{j+a}}{(2\pi)^{2n-2}}\left[\left(J^j_p - J^b_p\right)\left(J^a_r - J^j_r\right) - \left(J^j_r - J^b_r\right)\left(J^a_p - J^j_p\right)\right]\right. \\
&\qquad\quad \cdot\int_{\mathbb{R}^{2n-2}}\exp\left[i\varepsilon^{jb}(x_0,x,k,M) + i\varepsilon^{aj}\left(x_0,x,\lambda^{jb}(k,M),m\right)\right]T_{jb}(k,M) \\
&\qquad\quad \left.\cdot T_{aj}\left(\lambda^{jb}(k,M),m\right)\nu\left(\lambda^{aj}\left(\lambda^{jb}(k,M),m\right)\right)E^{ab}dmdM\right\}.
\end{aligned}$$

Using (2.49) and (2.50), we can rewrite the second sum as

$$\begin{aligned}
&\sum_{a,b}\frac{\gamma^a}{(2\pi)^{n-1}}\int_{\mathbb{R}^{n-1}}\exp\left[i\varepsilon^{ab}(x_0,x,k,m)\right] \\
&\quad\cdot\left\{\sum_j \frac{\gamma^j\left[\left(J^j_p - J^b_p\right)\left(J^a_r - J^j_r\right) - \left(J^j_r - J^b_r\right)\left(J^a_p - J^j_p\right)\right]}{(2\pi)^{n-1}}\int_{\mathbb{R}^{n-1}} T_{jb}(k,M)\right. \\
&\qquad\left.\cdot T_{aj}\left(\lambda^{jb}(k,M), m-M\right)\nu\left(\lambda^{ab}(k,m)\right)E^{ab}dM\right\}dm.
\end{aligned}$$

The proof of the lemma is then completed by observing that (2.52) holds because $\mathbf{L}^{ab}_r T_{ab} = \mathbf{L}^{ab}_p T_{ab}$. □

PROPOSITION 3.3. (Existence and uniqueness of the solution of the inverse problem.) *Assuming that* (3.8) *holds for $1 \le p \le n$, there is a unique μ which satisfies* (3.1) *and* (3.2) *simultaneously for $1 \le p \le n$.*

Proof. Since $\mathbf{L}^1_{(x_0,x)}\mu_1 = 0$, Lemma 3.2 implies that for $2 \le p \le n$

$$\mathbf{L}^1_{(x_0,x)}\left[\mathbf{L}^p_{(x_0,x)}\mu_1\right] = 0.$$

For fixed $(x_0, x, k_2, \cdots, k_n)$, $\lim_{k_1\to\infty} \partial\mu_1/\partial\bar{k}_p = 0$ by Proposition 3.2. Since the integral in the definition of $\mathbf{L}^p_{(x_0,x)}$ also tends to zero as $k_1 \to \infty$ by the decay of T, we have

$$\lim_{k_1\to\infty} \mathbf{L}^p_{(x_0,x)}\mu_1 = 0$$

for fixed $(x_0, x, k_2, \cdots, k_n)$.

Therefore the inversion formula for the $\bar{\partial}$ operator and Liouville's theorem imply that

$$\mathbf{L}^p_{(x_0,x)}\mu_1 = \mathbf{P}^1_{(x_0,x)}\left[\mathbf{L}^p_{(x_0,x)}\mu_1\right].$$

Hence $\mathbf{L}^p_{(x_0,x)}\mu_1 = 0$ for $2 \le p \le n$, by the contraction property of $\mathbf{P}^1_{(x_0,x)}$. In view of (3.11), μ_1 satisfies the boundary condition (3.2) for $2 \le p \le n$ which means that μ_1 also satisfies (3.4) for $2 \le p \le n$. By the contraction property of $\mathbf{P}^p_{(x_0,x)}$, we have $\mu_1 = \mu_p$ for $2 \le p \le n$. □

PROPOSITION 3.4. (Asymptotics of μ for large k.) *For fixed* $(x_0, x, k_1, \cdots, k_{p-1}, k_{p+1}, \cdots, k_n)$ *and* $1 \le p \le n$,

$$\mu(x_0, x, k) = I + \frac{1}{\pi k_p}\int_{\mathbb{R}^2} \frac{\partial\mu}{\partial\bar{k}_p}(x_0, x, k)\,dk_{p_R}\,dk_{p_I} + o\left(\frac{1}{|k_p|}\right). \tag{3.18}$$

Proof. We have

$$\begin{aligned} k_p(\mu - I) &= \sum_{a,b} \frac{\gamma^a\left(J^a_p - J^b_p\right)}{\pi(2\pi)^{n-1}} \int_{\mathbb{R}^{n+1}} \exp\left[i\varepsilon^{ab}\left(x_0, x, k'_{(p)}, m\right)\right] \\ &\quad \cdot T_{ab}(k'_{(p)}, m)\mu\left(x_0, x, \lambda^{ab}(k'_{(p)}, m)\right)E^{ab}\,dm\,dk'_{p_R}\,dk'_{p_I} \\ &\quad + \sum_{a,b} \frac{\gamma^a\left(J^a_p - J^b_p\right)}{\pi(2\pi)^{n-1}} \int_{\mathbb{R}^{n+1}} \frac{k'_p}{k_p - k'_p} \exp\left[i\varepsilon^{ab}\left(x_0, x, k'_{(p)}, m\right)\right] \\ &\quad \cdot T_{ab}(k'_{(p)}, m)\mu\left(x_0, x, \lambda^{ab}(k'_{(p)}, m)\right)E^{ab}\,dm\,dk_{p_R}\,dk_{p_I}. \end{aligned}$$

The second sum goes to zero as $k_p \to \infty$, by the decay of T. In view of (3.1), the first sum is just $(1/\pi)\int_{\mathbb{R}^2}(\partial\mu/\partial\bar{k}_p)(x_0, x, k)\,dk_{p_R}\,dk_{p_I}$. □

In view of (2.40), it is natural to define the potential Q by

$$Q = \frac{i\sigma}{\pi}\left[J_p, \int_{\mathbb{R}^2} \frac{\partial\mu}{\partial\bar{k}_p}(x_0, x, k)\,dk_{p_R}\,dk_{p_I}\right].$$

The problem with this definition is twofold: (i) it is not clear that Q is a function of (x_0, x) alone and (ii) Q apparently depends on the choice of p. Therefore we begin instead with the matrix-valued functions Q_p $(1 \le p \le n)$ defined by

$$Q_p = \frac{i\sigma}{\pi}\left[J_p, \int_{\mathbb{R}^2} \frac{\partial\mu}{\partial\bar{k}_p}(x_0, x, k)\,dk_{p_R}\,dk_{p_I}\right], \tag{3.19}$$

or equivalently (using (3.1)),

$$Q_p = \frac{i\sigma}{\pi} \sum_{a,b} \frac{\gamma^a (J_p^a - J_p^b)}{\pi (2\pi)^{n-1}} \Bigg[J_p, \int_{\mathbb{R}^{n+1}} \exp\left[i\varepsilon^{ab}(x_0, x, k, m) \right] \cdot T_{ab}(k, m) \mu\big(x_0, x, \lambda^{ab}(k, m)\big) E^{ab} \Bigg] dm dk_{p_R} dk_{p_I}. \tag{3.20}$$

Hence Q_p is C^∞ and bounded on $\mathbb{R}^{n+1} \times \mathbb{C}^n$. Note that Q_p is independent of k_p.

Let $\mathbf{L}_k$ be the differential operator defined by

$$\mathbf{L}_k \theta = \frac{\partial \theta}{\partial x_0} + \sigma \sum_{\ell=1}^{n} \left(J_\ell \frac{\partial \theta}{\partial x_\ell} + i k_\ell [J_\ell, \theta] \right), \tag{3.21}$$

where θ is an $N \times N$-matrix-valued function defined on $\mathbb{R}^{n+1}$. Note that (1.3) can be rewritten as $\mathbf{L}_k \mu = Q\mu$.

LEMMA 3.3. *Let $\nu \in C^2(\mathbb{R}^{n+1} \times \mathbb{C}^n)$ and $\partial^\alpha_{(x_0,x)} \nu$ be bounded for $|\alpha| \le 1$. Then*

$$\big[\mathbf{L}^p_{(x_0,x)}, \mathbf{L}_k \big] \nu = 0. \tag{3.22}$$

Proof. It is clear that $\partial(\mathbf{L}_k \mu)/\partial \bar{k}_p = \mathbf{L}_k (\partial \mu / \partial \bar{k}_p)$. On the other hand, the decay of T allows differentiation under the integral sign:

$$\begin{aligned} \mathbf{L}_k \Bigg\{ \sum_{a,b} \frac{\gamma^a (J_p^a - J_p^b)}{\pi (2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} & \exp\left[i\varepsilon^{ab}(x_0, x, k, m) \right] \\ & \cdot T_{ab}(k, m) \nu\big(x_0, x, \lambda^{ab}(k, m)\big) E^{ab} dm \Bigg\} \\ = \sum_{a,b} \frac{\gamma^a (J_p^a - J_p^b)}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} & T_{ab}(k, m) \mathbf{L}_k \Big\{ \exp\left[i\varepsilon^{ab}(x_0, x, k, m) \right] \\ & \cdot \nu\big(x_0, x, \lambda^{ab}(k, m)\big) E^{ab} \Big\} dm. \end{aligned}$$

Recall (cf. proof of Proposition 2.4) that $\lambda^{ab}(k, m)$ is defined so that

$$\begin{aligned} \mathbf{L}_k \Big\{ \exp\big[i\varepsilon^{ab}(x_0, x, k, m) \big] \nu(x_0, x, \lambda^{ab}(k, m)) E^{ab} \Big\} \\ = \exp\left[i\varepsilon^{ab}(x_0, x, k, m) \right] (\mathbf{L}_k \nu)\big(x_0, x, \lambda^{ab}(k, m)\big). \end{aligned}$$

The proof of the lemma is therefore complete. □

PROPOSITION 3.5. (Equivalence of the reconstruction formulas.)

$$Q_1 = Q_2 = \cdots = Q_n.$$

Proof. From (3.22) we have $\mathbf{L}^p_{(x_0,x)}(\mathbf{L}_k \mu) = 0$ for $1 \le p \le n$, and it follows from (3.18) that

$$\lim_{k_p \to \infty} i\sigma \sum_{\ell=1}^{n} k_\ell [J_\ell, \mu] = Q_p \tag{3.23}$$

for fixed $(x_0, x, k_1, \cdots, k_{p-1}, k_{p+1}, \cdots, k_n)$. Hence Proposition 3.2 implies that

$$\lim_{k_p \to \infty} \mathbf{L}_k \mu = Q_p \tag{3.24}$$

for fixed $(x_0, x, k_1, \cdots, k_{p-1}, k_{p+1}, \cdots, k_n)$.

Since $\mathbf{L}_k \mu$ has polynomial growth in k, we have by Remark 3.1 the inversion formula for the $\bar{\partial}$ operator and Liouville's theorem that

$$\mathbf{L}_k \mu = Q_p + \mathbf{P}^p_{(x_0,x)}(\mathbf{L}_k \mu) \quad \text{for } 1 \le p \le n. \tag{3.25}$$

Recall that Q_p is independent of k_p. On the other hand, if we keep $(x_0, x, k_1, \cdots, k_{p-1}, k_{p+1}, \cdots, k_n)$ fixed and then let $k_p \to \infty$ in the equation $\mathbf{L}_k \mu = Q_j + \mathbf{P}^j_{(x_0,x)} \cdot (\mathbf{L}_k \mu)$, it follows from (3.7) that

$$Q_p = \lim_{k_p \to \infty} \mathbf{L}_k \mu = \lim_{k_p \to \infty} Q_j. \tag{3.26}$$

Therefore Q_p is also independent of k_j. Since j is arbitrary, Q_p is a function of (x_0, x) alone, for $1 \le p \le n$. But then (3.26) implies that $Q_p = Q_j$ for any p and j, i.e., $Q_1 = Q_2 = \cdots = Q_n$. □

COROLLARY 3.1. (Differential equation relating Q and μ.)

$$\mu_{x_0} + \sigma \sum_{\ell=1}^{n} (J_\ell \mu_{x_\ell} + i k_\ell [J_\ell, \mu]) = Q\mu. \tag{3.27}$$

Proof. Since Q is a function of (x_0, x) alone, it follows from (3.10) that

$$Q\mu = Q + \mathbf{P}^p_{(x_0,x)}(Q\mu) \quad \text{for } 1 \le p \le n. \tag{3.28}$$

Equation (3.27) then follows by comparing (3.28) with (3.25). □

It is clear that Q is off-diagonal and C^∞ on $\mathbb{R}^{n+1}$. To study the decay of Q, we exploit the following integral equation, which is a consequence of (3.1), (3.4), and (2.30).

$$\begin{aligned} &\frac{\partial \mu}{\partial \bar{k}_1}(x_0, x, k) \\ &= G + \sum_{a,b} \frac{\gamma^a (J_1^a - J_1^b)}{\pi (2\pi)^{n-1}} \int_{\mathbb{R}^{n+1}} \exp\left[i\varepsilon^{ab}(x_0, x_1, k, m)\right] \frac{1}{\lambda_1^{ab}(k,m) - k_1'} \\ &\quad \cdot T_{ab}(k,m) \frac{\partial \mu}{\partial \bar{k}_1'}(x_0, x, k_1', k_2 + m_2, \cdots, k_n + m_n) E^{ab} dk_{1_R}' dk_{1_I}' dm, \end{aligned} \tag{3.29}$$

where

$$G = \sum_{a,b} \frac{\gamma^a (J_1^a - J_1^b)}{\pi (2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} \exp\left[i\varepsilon^{ab}(x_0, x, k, m)\right] T_{ab}(k,m) E^{ab} dm. \tag{3.30}$$

DEFINITION 3.1. Let $L^1_{N\times N}(\mathbb{C})$ be the Banach space of $N \times N$ matrices whose entries belong to $L^1(\mathbb{C})$. If $\nu \in L^1_{N\times N}(\mathbb{C})$, then $\|\nu\|_{L^1_{N\times N}} = \max_{1\le a,b\le N} \|\nu_{ab}\|_{L^1(\mathbb{C})}$. Let $C^b(\mathbb{C}^{n-1}, L^1_{N\times N}(\mathbb{C}))$ be the Banach space of bounded continuous func-

tions on $\mathbb{C}^{n-1}$ with value in $L^1_{N\times N}(\mathbb{C})$. If $\nu(k_2,\cdots,k_n)\in C^b\big(\mathbb{C}^{n-1},L^1_{N\times N}(\mathbb{C})\big)$, $|||\nu|||=\sup_{\mathbb{C}^{n-1}}\|\nu(k_2,\cdots,k_n)\|_{L^1_{N\times N}}$. We shall also write $\nu(k_2,\cdots,k_n)(k_1)$ as $\nu(k_1,k_2,\cdots,k_n)$.

Let $\mathbf{R}_{(x_0,x)}$ be the operator on $C^b\big(\mathbb{C}^{n-1},L^1_{N\times N}(\mathbb{C})\big)$ defined by

$$\begin{aligned}\Big[\mathbf{R}_{(x_0,x)}\nu\Big](k)&\\ =\sum_{a,b}\frac{\gamma^a\big(J_1^a-J_1^b\big)}{\pi(2\pi)^{n-1}}\int_{\mathbb{R}^{n+1}}&\exp\big[i\varepsilon^{ab}(x_0,x,k,m)\big]\frac{1}{\lambda_1^{ab}(k,m)-k_1'}T_{ab}(k,m)\\ &\cdot\nu\big(k_1',k_2+m_2,\cdots,k_n+m_n\big)E^{ab}dk_{1_R}'dk_{1_I}'dm\,.\end{aligned}\tag{3.31}$$

It is clear from (3.1) that $(\partial\mu/\partial\bar{k}_1)\in C^b\big(\mathbb{C}^{n-1},L^1_{N\times N}(\mathbb{C})\big)$. We can therefore regard (3.29) as an equation in $C^b\big(\mathbb{C}^{n-1},L^1_{N\times N}(\mathbb{C})\big)$ and rewrite it as

$$\frac{\partial\mu}{\partial\bar{k}_1}=G+\mathbf{R}_{(x_0,x)}\Big(\frac{\partial\mu}{\partial\bar{k}_1}\Big).\tag{3.32}$$

LEMMA 3.4. *$\mathbf{R}_{(x_0,x)}$ is a contraction on $C^b\big(\mathbb{C}^{n-1},L^1_{N\times N}(\mathbb{C})\big)$ if (3.8) holds for $p=1$.*

Proof. Note that by (2.29),

$$\big[\lambda_1^{ab}(k_1,0)\big]_I=k_{1_I}\tag{3.33}$$

and

$$\big[\lambda_1^{ab}(k_1,0)\big]_R=\frac{J_1^b}{J_1^a}k_{1_R}-\frac{\sigma_R\big(J_1^a-J_1^b\big)}{\sigma_I J_1^a}k_{1_I}\,.\tag{3.34}$$

From (2.24),

$$\begin{aligned}\omega_1&=\max_{a,b}\frac{|\gamma^a||J_1^a-J_1^b|}{(2\pi)^{n-1}}\\ &=\max_{a,b}\frac{|\gamma^a||J_1^a-J_1^b||J_1^a|}{(2\pi)^{n-1}|J_1^b|}\,.\end{aligned}\tag{3.35}$$

Using (3.33)–(3.35) and an argument similar to the one in the proof of Proposition 2.1, we find

$$\begin{aligned}&\Big\|\Big[\mathbf{R}_{(x_0,x)}\nu\Big]_{ab}(\cdot,k_2,\cdots,k_n)\Big\|_{L^1_{N\times N}}\\ &\quad\le\sum_{j=1}^N\frac{|\gamma^j||J_1^j-J_1^b|}{\pi(2\pi)^{n-1}}\int_{\mathbb{R}^{n+3}}\frac{1}{|\lambda_1^{jb}(k,m)-k_1'|}|T_{jb}(k,m)|\\ &\qquad\cdot|\nu_{aj}(k_1',k_2+m_2,\cdots,k_n+m_n)|dk_{1_R}'dk_{1_I}'dmdk_{1_R}dk_{1_I}\\ &\quad\le\frac{(N-1)8^{\frac12}\omega_1}{\pi^{\frac12}}|||\nu|||\int_{\mathbb{R}^{n-1}}\Big[\sup_{\mathbb{C}^n}|T(k,m)|\Big]^{\frac12}\Big[\sup_{\mathbb{C}^{n-1}}\int_{\mathbb{R}^2}|T(k,m)|dk_{1_R}dk_{1_I}\Big]^{\frac12}dm\\ &\quad\le\frac{(N-1)8^{\frac12}\omega_1}{\pi^{\frac12}}|||\nu|||\Big[\int_{\mathbb{R}^{n-1}}\sup_{\mathbb{C}^n}|T(k,m)|dm\Big]^{\frac12}\\ &\qquad\cdot\Big[\int_{\mathbb{R}^{n-1}}\Big[\sup_{\mathbb{C}^{n-1}}\int_{\mathbb{R}^2}|T(k,m)|dk_{1_R}dk_{1_I}\Big]dm\Big]^{\frac12}.\end{aligned}$$

□

Since we assume that (3.8) holds for $1 \leq p \leq n$, we can control the norm of $R_{(x_0,x)}$ by the same constant that appears in (3.9):

$$||\mathbf{R}_{(x_0,x)}|| \leq \tau_1 < 1 \quad \forall\, (x_0,x) \in \mathbb{R}^{n+1}. \tag{3.36}$$

PROPOSITION 3.6. (Decay of Q.)

$$Q \in S^o_{N\times N}.$$

Proof. From (3.32), we obtain

$$\begin{aligned}(1+|x_0|^2+|x|^2)^j \frac{\partial \mu}{\partial \bar{k}_1} &= (1+|x_0|^2+|x|^2)^j G \\ &\quad + \mathbf{R}_{(x_0,x)}\left((1+|x_0|^2+|x|^2)^j \frac{\partial \mu}{\partial \bar{k}_1}\right)\end{aligned} \tag{3.37}$$

for $j = 0, 1, 2, \cdots$.

Using integration by parts and (3.30), property (ii) of Definition 2.3 implies that there exists $C_j > 0$ such that

$$|||(1+|x_0|^2+|x|^2)G(x_0,x,k)||| \leq C_j \quad \forall\, (x_0,x) \in \mathbb{R}^{n+1}. \tag{3.38}$$

It follows from (3.36)–(3.38) that

$$|||(1+|x_0|^2+|x|^2)^j \frac{\partial \mu}{\partial \bar{k}_1}(x_0,x,k)||| \leq \frac{C_j}{1-\tau_1} \quad \forall\, (x_0,x) \in \mathbb{R}^{n+1}. \tag{3.39}$$

Inductively, we have

$$\partial^\alpha_{(x_0,x)} \frac{\partial \mu}{\partial \bar{k}_1} = G_\alpha + \mathbf{R}_{(x_0,x)}\left(\partial^\alpha_{(x_0,x)} \frac{\partial \mu}{\partial \bar{k}_1}\right), \tag{3.40}$$

where

$$|||(1+|x_0|^2+|x|^2)^j G_\alpha(x_0,x,k)||| \leq C_{\alpha,j} \quad \forall\, (x_0,x) \in \mathbb{R}^{n+1}. \tag{3.41}$$

Hence

$$|||(1+|x_0|^2+|x|^2)^j \partial^\alpha_{(x_0,x)} \frac{\partial \mu}{\partial \bar{k}_1}(x_0,x,k)||| \leq \frac{C_{\alpha,j}}{1-\tau_1} \quad \forall\, (x_0,x) \in \mathbb{R}^{n+1}. \tag{3.42}$$

Since $Q = (i\sigma/\pi)\left[J_1, \int_{\mathbb{R}^2} \frac{\partial \mu}{\partial k_1}(x_0,x,k) dk_{1_R} dk_{1_I}\right]$, the proposition follows from (3.42). □

DEFINITION 3.2. (Inverse scattering map.)

$$\hat{\mathbf{S}}(T) = Q. \tag{3.43}$$

Remark 3.2. If we trace the dependence of Q on T through the arguments in the proof of Proposition 3.6, we see that $\hat{\mathbf{S}}$ is a continuous map from a neighborhood of $0 \in D^o_{N\times N}$ into $S^o_{N\times N}$.

The results of this section are summarized in the following theorem.

THEOREM 3.1. *The inverse problem has a unique simultaneous solution* μ *if* (3.8) *holds for* $1 \le p \le n$. *The reconstruction formulas in* (3.19) *are equivalent, and the potential* $Q(x_0, x)$ *defined by these formulas belongs to* $S^o_{N\times N}$. *Moreover,* (1.3)–(1.4) *are satisfied by* Q *and* μ. *The inverse scattering map* $\hat{\mathbf{S}}$ *that takes* T *to* Q *is a continuous map from a neighborhood of* $0 \in D^o_{N\times N}$ *into* $S^o_{N\times N}$.

Remark 3.3. $\hat{\mathbf{S}}$ is actually infinitely Fréchet differentiable.

4. Relation between S and $\hat{\mathbf{S}}$.

PROPOSITION 4.1.

$$\hat{\mathbf{S}} \circ \mathbf{S} = \mathbf{I} \quad \textit{on the domain of} \quad \hat{\mathbf{S}} \circ \mathbf{S}, \tag{4.1}$$

and

$$\mathbf{S} \circ \hat{\mathbf{S}} = \mathbf{I} \quad \textit{on the domain of} \quad \mathbf{S} \circ \hat{\mathbf{S}}. \tag{4.2}$$

Proof. Given $Q \in S^o_{N\times N}$ belonging to the domain of $\hat{\mathbf{S}} \circ \mathbf{S}$, let $T = \mathbf{S}(Q)$ and $Q' = \hat{\mathbf{S}}(T)$. Let μ be the unique solution of (1.3)–(1.4). Then μ is also the unique solution of the $\bar\partial$ problem (3.1)–(3.2) by Proposition 2.4. In view of Corollary 2.3, Q' defined by

$$Q' = \frac{i\sigma}{\pi}\left[J_p, \int_{\mathbb{R}^2} \frac{\partial \mu}{\partial \bar k_p}(x_0, x, k) dk_{p_R} dk_{p_I}\right], \qquad 1 \le p \le n,$$

must coincide with Q.

Given $T \in D^o_{N\times N}$ belonging to the domain $\mathbf{S} \circ \hat{\mathbf{S}}$, let $Q = \hat{\mathbf{S}}(T)$ and $T' = \mathbf{S}(Q)$. Let μ be the unique solution of (3.1)–(3.2). Then it follows immediately that

$$\begin{aligned} \lim_{x_0\to\infty} &\left\{\exp[-i\beta^a(x_0, 0, k_{ab})]\frac{\partial \mu_{ab}}{\partial \bar k_p}(x_0, 0, x_2, \cdots, x_n, k)\right\} \\ &= \frac{\gamma^a(J^a_p - J^b_p)}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} \exp\left(\sum_{\ell=2}^n m_\ell x_\ell\right) T_{ab}(k, m)\, dm. \end{aligned} \tag{4.3}$$

By Corollary 3.1 and (3.12), μ is also the unique solution of (1.3)–(1.4). Hence Corollary 2.2 implies that

$$\begin{aligned} \lim_{x_0\to\infty} &\left[\exp[-i\beta^a(x_0, 0, k_{ab})]\frac{\partial \mu_{ab}}{\partial \bar k_p}(x_0, 0, x_2, \cdots, x_n, k)\right] \\ &= \frac{\gamma^a(J^a_p - J^b_p)}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} \exp\left(\sum_{\ell=2}^n m_\ell x_\ell\right) T'_{ab}(k, m)\, dm. \end{aligned} \tag{4.4}$$

Comparing (4.3) and (4.4), we conclude that $T = T'$. □

Combining Remarks 2.4 and 3.2 and Proposition 4.1, we obtain the following theorem.

THEOREM 4.1. *The scattering map* $\mathbf{S}$ *is a homeomorphism from a neighborhood of* $0 \in S^o_{N\times N}$ *onto a neighborhood of* $0 \in D^o_{N\times N}$.

Remark 4.1. $\mathbf{S}$ is actually a diffeomorphism.

5. Reduction to 2×2 potentials. We shall show in this section how the reconstruction of the potential from the scattering data can be reduced to the special

case of 2×2 potentials and 2×2 inverse data. As mentioned in the Introduction, this means that the number of variables involved can be reduced to 3, i.e., $n = 2$.

Throughout this section we assume appropriate *smallness* conditions so that all the equations involved are uniquely solvable. We first take a quick look at the simplification of the theory developed in §§2 and 3 when $N = 2$. Let $\hat{Q}(x_0, x)$ be a 2×2 potential. The equations for the direct problem are

$$\frac{\partial \hat{\mu}_{11}}{\partial x_0} + \sigma \sum_{\ell=1}^{n} \hat{J}_\ell^1 \frac{\partial \hat{\mu}_{11}}{\partial x_\ell} = \hat{Q}_{12}\hat{\mu}_{21}, \tag{5.1}$$

$$\frac{\partial \hat{\mu}_{12}}{\partial x_0} + \sigma \sum_{\ell=1}^{n} \hat{J}_\ell^1 \frac{\partial \hat{\mu}_{12}}{\partial x_\ell} + i\sigma \left[\sum_{\ell=1}^{n} k_\ell (\hat{J}_\ell^1 - \hat{J}_\ell^2) \right] \hat{\mu}_{12} = \hat{Q}_{12}\hat{\mu}_{22}, \tag{5.2}$$

$$\frac{\partial \hat{\mu}_{21}}{\partial x_0} + \sigma \sum_{\ell=2}^{n} \hat{J}_\ell^2 \frac{\partial \hat{\mu}_{21}}{\partial x_\ell} + i\sigma \left[\sum_{\ell=1}^{n} k_\ell (\hat{J}_\ell^2 - \hat{J}_\ell^1) \right] \hat{\mu}_{21} = \hat{Q}_{21}\hat{\mu}_{11}, \tag{5.3}$$

and

$$\frac{\partial \hat{\mu}_{22}}{\partial x_0} + \sigma \sum_{\ell=1}^{n} \hat{J}_\ell^2 \frac{\partial \hat{\mu}_{22}}{\partial x_\ell} = \hat{Q}_{21}\hat{\mu}_{12} \tag{5.4}$$

with the boundary condition

$$\lim_{|x_0|+|x| \to \infty} \hat{\mu} = I. \tag{5.5}$$

It is clear that there is really only one complex variable involved, namely, $\chi = \sum_{\ell=1}^{n} k_\ell (\hat{J}_\ell^1 - \hat{J}_\ell^2)$. In particular, the characterization problem does not exist.

$(\partial \hat{\mu}/\partial \bar{\chi})$ and $\hat{\mu}$ are related through

$$\begin{aligned} \frac{\partial \hat{\mu}}{\partial \bar{\chi}} = \frac{\bar{\sigma}}{4\pi i |\sigma_I| (2\pi)^{n-1}} \Bigg\{ & \frac{1}{\hat{J}_1^1} \int_{\mathbb{R}^{n-1}} \exp\left[i\hat{\beta}^1(x_0, x_1, \chi) + i \sum_{\ell=2}^{n} m_\ell \left(x_\ell - x_1 \frac{\hat{J}_\ell^1}{\hat{J}_1^1} \right) \right] \\ & \cdot \hat{T}_{12}(\chi, m) \hat{\mu}\big(x_0, x, \hat{\lambda}^1(\chi, m)\big) E^{12} dm \\ & + \frac{-1}{\hat{J}_1^2} \int_{\mathbb{R}^{n-1}} \exp\left[i\hat{\beta}^2(x_0, x_1, -\chi) + i \sum_{\ell=2}^{n} m_\ell \left(x_\ell - x_1 \frac{\hat{J}_\ell^2}{\hat{J}_1^2} \right) \right] \\ & \cdot \hat{T}_{21}(\chi, m) \hat{\mu}\big(x_0, x, \hat{\lambda}^2(\chi, m)\big) E^{21} dm \Bigg\}, \end{aligned} \tag{5.6}$$

where

$$\hat{\beta}^j(x_0, x_1, \chi) = \frac{1}{\sigma} \left[|\sigma|^2 \chi_I x_0 - \frac{(\sigma\chi)_I}{\hat{J}_1^j} x_1 \right], \qquad j = 1, 2, \tag{5.7}$$

$$\begin{aligned} \hat{T}_{12}(\chi, m) = \int_{\mathbb{R}^{n+1}} & \exp\left[-i\hat{\beta}^1(\xi_0, \xi_1, \chi) - i \sum_{\ell=2}^{n} m_\ell \left(\eta_\ell - \xi_1 \frac{\hat{J}_\ell^1}{\hat{J}_1^1} \right) \right] \\ & \cdot \hat{Q}_{12}(\xi_0, \xi_1, \eta) \hat{\mu}_{22}(\xi_0, \xi_1, \eta, \chi) d\xi_0 d\xi_1 d\eta, \end{aligned} \tag{5.8}$$

$$\begin{aligned} \hat{T}_{21}(\chi, m) = \int_{\mathbb{R}^{n+1}} & \exp\left[-i\hat{\beta}^2(\xi_0, \xi_1, -\chi) - i \sum_{\ell=2}^{n} m_\ell \left(\eta_\ell - \xi_1 \frac{\hat{J}_\ell^2}{\hat{J}_1^2} \right) \right] \\ & \cdot \hat{Q}_{21}(\xi_0, \xi_1, \eta) \hat{\mu}_{11}(\xi_0, \xi_1, \eta, \chi) d\xi_0 d\xi_1 d\eta, \end{aligned} \tag{5.9}$$

and

$$\hat{\lambda}^j(\chi,m)=\frac{\chi}{\hat{J}_1^1-\hat{J}_1^2}+\sum_{\ell=2}^{n}m_\ell\frac{\hat{J}_\ell^1\hat{J}_1^2-\hat{J}_1^1\hat{J}_\ell^2}{\hat{J}_1^j\left(\hat{J}_1^1-\hat{J}_1^2\right)}-(-1)^j\frac{(\sigma\chi)_I}{\sigma_I\hat{J}_1^j},\qquad j=1,2. \tag{5.10}$$

Equation (5.6) is essentially the same as (2.28) except for the minor changes due to the fact that the derivative is taken with respect to $\bar{\chi}$ instead of the individual $\bar{k}_p$.

In the inverse problem, $\hat{\mu}$ is obtained by solving (5.6) with boundary condition

$$\lim_{\chi\to\infty}\hat{\mu}(x_0,x,\chi)=I. \tag{5.11}$$

The potential is then given by

$$\hat{Q}_{12}(x_0,x)=\frac{i\sigma}{\pi}\int_{\mathbb{R}^2}\frac{\partial\hat{\mu}_{12}}{\partial\bar{\chi}}(x_0,x,\chi)d\chi_R d\chi_I \tag{5.12}$$

and

$$\hat{Q}_{21}(x_0,x)=\frac{i\sigma}{\pi}\int_{\mathbb{R}^2}\frac{\partial\hat{\mu}_{21}}{\partial\bar{\chi}}(x_0,x,\chi)d\chi_R d\chi_I. \tag{5.13}$$

We now show how the general $N\times N$ case can be reduced to the special case above. We shall impose the following additional nondegeneracy condition.

$$\left(J_1^a-J_1^b\right)\left(J_p^c-J_p^b\right)\neq\left(J_1^c-J_1^b\right)\left(J_p^a-J_p^b\right) \tag{5.14}$$

for all distinct a,b,c and for $p\neq 1$.

Let μ satisfy (1.3) and (1.4). For fixed $\hat{a}$ and $\hat{b}$ such that $\hat{a}<\hat{b}$, let $\hat{Q}_{12}(x_0,x)=Q_{\hat{a}\hat{b}}(x_0,x),\hat{Q}_{21}(x_0,x)=Q_{\hat{b}\hat{a}}(x_0,x),\hat{J}_\ell^1=J_\ell^a$, and $\hat{J}_\ell^2=J_\ell^b$ for $1\leq\ell\leq n$.

PROPOSITION 5.1. *The reduced 2×2 system*

$$\hat{\mu}_{x_0}+\sum_{\ell=1}^{n}\sigma\left(\hat{J}_\ell\hat{\mu}_{x_\ell}+ik_\ell\left[\hat{J}_\ell,\hat{\mu}\right]\right)=\hat{Q}\hat{\mu} \tag{5.15}$$

with boundary condition

$$\lim_{|x_0|+|x|\to\infty}\hat{\mu}=I \tag{5.16}$$

is satisfied (uniquely) by

$$\hat{\mu}_{11}=\lim_{k_p\to\infty}\mu_{\hat{a}\hat{a}},\qquad \hat{\mu}_{12}=\lim_{k_p\to\infty}\mu_{\hat{a}\hat{b}},$$
$$\hat{\mu}_{21}=\lim_{k_p\to\infty}\mu_{\hat{b}\hat{a}},\qquad \hat{\mu}_{22}=\lim_{k_p\to\infty}\mu_{\hat{b}\hat{b}}. \tag{5.17}$$

Here the limit is taken with fixed $(x_0,x,\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)$, where $\chi=\sum_{\ell=1}^{n}k_\ell\left(\hat{J}_\ell^1-\hat{J}_\ell^2\right)=k_{\hat{a}\hat{b}}$ and $p\geq 2$.

Proof. Observe that (5.14) implies that

$$\lim_{k_p\to\infty} k_{ab} = \infty \tag{5.18}$$

for $b = \hat{a}$ or $\hat{b}$, $1 \le a \le N$, and $(a,b) \notin \{(\hat{a},\hat{a}),(\hat{a},\hat{b}),(\hat{b},\hat{a}),(\hat{b},\hat{b})\}$.

Assume that as $k_p \to \infty$, $\mu_{ab}(x_0,x,\chi,k_2,\cdots,k_n)$ converges uniformly on compact subsets of $\mathbb{R}^{n+1}$ to $\psi_{ab}(x_0,x,\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)$ for $b = \hat{a}$ or $\hat{b}$ and for $1 \le a \le N$. By (2.8), (5.18), the boundedness of μ, the decay of Q, and the Riemann–Lebesgue lemma, we find for $(a,b) \notin \{(\hat{a},\hat{a}),(\hat{a},\hat{b}),(\hat{b},\hat{a}),(\hat{b},\hat{b})\}$

$$\begin{aligned}
&\psi_{ab}(x_0,x,\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)\\
&\quad = \lim_{k_p\to\infty} \mu_{ab}(x_0,x,\chi,k_2,\cdots,k_n)\\
&\quad = \lim_{k_p\to\infty} \int_{\mathbb{R}^2} G(\xi,k_{ab})[Q\psi]_{ab}(x_0-2\xi_0, x_1 - J_1^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\cdots,\\
&\qquad\qquad x_n - J_n^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\chi,k_2,\cdots,k_n)d\xi_0 d\xi_1 \qquad (5.19)\\
&\quad = 0\,.
\end{aligned}$$

It also follows from (2.8), (5.18) and (5.19) that

$$\begin{aligned}
&\psi_{\hat{a}\hat{a}}(x_0,x,\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)\\
&\quad = 1 + \int_{\mathbb{R}^2} G(\xi,0)[Q_{\hat{a}\hat{b}}\psi_{\hat{b}\hat{a}}](x_0-2\xi_0, x_1 - J_1^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\cdots, \qquad (5.20)\\
&\qquad\qquad x_n - J_n^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)d\xi_0 d\xi_1,\\
&\psi_{\hat{a}\hat{b}}(x_0,x,\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)\\
&\quad = \int_{\mathbb{R}^2} G(\xi,\chi)[Q_{\hat{a}\hat{b}}\psi_{\hat{b}\hat{b}}](x_0-2\xi_0, x_1 - J_1^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\cdots, \qquad (5.21)\\
&\qquad\qquad x_n - J_n^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)d\xi_0 d\xi_1,\\
&\psi_{\hat{b}\hat{a}}(x_0,x,\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)\\
&\quad = \int_{\mathbb{R}^2} G(\xi,-\chi)[Q_{\hat{b}\hat{a}}\psi_{\hat{a}\hat{a}}](x_0-2\xi_0, x_1 - J_1^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\cdots, \qquad (5.22)\\
&\qquad\qquad x_n - J_n^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)d\xi_0 d\xi_1,\\
&\psi_{\hat{b}\hat{b}}(x_0,x,\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)\\
&\quad = 1 + \int_{\mathbb{R}^2} G(\xi,0)[Q_{\hat{b}\hat{a}}\psi_{\hat{a}\hat{b}}](x_0-2\xi_0, x_1 - J_1^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\cdots, \qquad (5.23)\\
&\qquad\qquad x_n - J_n^a(\sigma\bar{\xi}+\bar{\sigma}\xi),\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)d\xi_0 d\xi_1.
\end{aligned}$$

In other words, $\hat{\mu} = \begin{pmatrix} \psi_{\hat{a}\hat{a}} & \psi_{\hat{a}\hat{b}} \\ \psi_{\hat{b}\hat{a}} & \psi_{\hat{b}\hat{b}} \end{pmatrix}$ is the unique solution of (5.15) and (5.16). In particular, ψ_{ab} only depends on (x_0,x,χ).

It remains to show that $\mu_{ab}(x_0,x,\chi,k_2,\cdots,k_n)$ actually converges uniformly on compact subsets of $\mathbb{R}^{n+1}$ for $1 \le a \le N$, $b = \hat{a}$ or $\hat{b}$, and for fixed $(\chi,k_2,\cdots,k_{p-1},k_{p+1},\cdots,k_n)$. Since $\partial^\alpha \mu \in C^b_{N\times N}(\mathbb{R}^{n+1}\times C^n)$ for $|\alpha| \le 1$, the Arzela–Ascoli theorem implies that given any sequence $\kappa_j \to \infty$, $\mu_{ab}(x_0,x,\chi,k_2,\cdots,k_{p-1},\kappa_j,k_{p+1},\cdots,k_n)$ has a subsequence that converges uniformly on compact subsets of $\mathbb{R}^{n+1}$. But we have shown that these limit functions must satisfy (5.19)–(5.23), i.e., they are uniquely determined.

Therefore, $\mu_{ab}(x_0, x, \kappa, k_2, \cdots, k_n)$ converges uniformly on compact subsets of $\mathbb{R}^{n+1}$ as the continuous variable $k_p \to \infty$. □

Let T be the inverse data of Q, and let

$$\hat{T}(\chi, m) = \begin{pmatrix} 0 & \hat{T}_{12}(\chi, m) \\ \hat{T}_{21}(\chi, m) & 0 \end{pmatrix}$$

be the inverse data of $\hat{Q}$. By comparing (5.8) and (5.9) with (2.27), we have by (5.19) the following corollary.

COROLLARY 5.1. (Relation between the inverse data.)

$$\hat{T}_{12}(\chi, m) = \lim_{k_p \to \infty} T_{\hat{a}\hat{b}}(k, m), \tag{5.24}$$

and

$$\hat{T}_{21}(\chi, m) = \lim_{k_p \to \infty} T_{\hat{b}\hat{a}}(k, m), \tag{5.25}$$

where the limit is taken with fixed $(\chi, k_2, \cdots, k_{p-1}, k_{p+1}, \cdots, k_n)$.

The characterization equations in (2.42) can be rewritten as

$$\frac{1}{J_p^a - J_p^b} \frac{\partial T_{ab}}{\partial \bar{k}_p}(k, m) - \frac{1}{J_1^a - J_1^b} \frac{\partial T_{ab}}{\partial \bar{k}_1}(k, m) = N_{1p}^{ab}[T], \qquad 2 \le p \le n, \tag{5.26}$$

where

$$\begin{aligned} &N_{1p}^{ab}[T](k, m) \\ &= \sum_{j=1}^{N} \frac{\gamma^j}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} \left[\frac{J_p^j - J_p^b}{J_p^a - J_p^b} - \frac{J_1^j - J_1^b}{J_1^a - J_1^b} \right] T_{jb}(k, M) \\ &\qquad \cdot T_{aj}(\lambda^{jb}(k, m), m - M)\, dM. \end{aligned} \tag{5.27}$$

Denote by $\tilde{T}_{ab}$ and $\tilde{N}_{1p}^{ab}[\tilde{T}]$ the functions representing T_{ab} and $N_{1p}^{ab}[T]$ in the $(k_{ab}, k_2, \cdots, k_n, m)$ coordinates. Then (5.26) is equivalent to

$$\begin{aligned} &\frac{\partial \tilde{T}_{ab}}{\partial \bar{k}_p}(k_{ab}, k_2, \cdots, k_n, m) = (J_p^a - J_p^b) \tilde{N}_{1p}^{ab}[\tilde{T}](k_{ab}, k_2, \cdots, k_n, m), \\ &\qquad\qquad 2 \le p \le n. \end{aligned} \tag{5.28}$$

PROPOSITION 5.2. (Equivalent characterization of the inverse data.) *Given any* a *and* b *such that* $1 \le a \ne b \le n$,

$$\begin{aligned} &\tilde{T}_{ab}(k_{ab}, k_2, \cdots, k_n, m) - \frac{1}{\pi} \int_{\mathbb{R}^2} \frac{(J_p^a - J_p^b)}{k_p - k_p'} \\ &\qquad \cdot \tilde{N}_{1p}^{ab}[\tilde{T}](k_{ab}, k_2, \cdots, k_{p-1}, k_p', k_{p+1}, \cdots, k_n, m)\, dk_{p_k}'\, dk_{p_I}' \end{aligned} \tag{5.29}$$

represent the same function $\Phi_{ab}(k_{ab}, m)$ *for* $2 \le p \le n$.

Proof. By the inversion formula for the $\bar{\partial}$ operator and Liouville's theorem, (5.28) is equivalent to

$$\begin{aligned} &\tilde{T}_{ab}(k_{ab}, k_2, \cdots, k_n, m) \\ &= \lim_{k_p \to \infty} \tilde{T}_{ab}(k_{ab}, k_2, \cdots, k_n, m) + \frac{1}{\pi} \int_{\mathbb{R}^2} \frac{(J_p^a - J_p^b)}{k_p - k_p'} \\ &\qquad \cdot \tilde{N}_{1p}^{ab}[T](k_{ab}, k_2, \cdots, k_{p-1}, k_p', k_{p+1}, \cdots, k_n, m)\, dk_{p_k}'\, dk_{p_I}'. \end{aligned} \tag{5.30}$$

Here the limit is taken with $(k_{ab}, k_2, \cdots, k_{p-1}, k_{p+1}, \cdots, k_n, m)$ fixed. Corollary 5.1 implies that $\lim_{k_p \to \infty} \tilde{T}_{ab}(k_{ab}, k_2, \cdots, k_n, m) = \lim_{k_r \to \infty} \tilde{T}_{ab}(k_{ab}, k_2, \cdots, k_n, m)$ for $2 \leq p, r \leq n$. Their common value is the inverse data of the reduced system, and therefore only depends on (k_{ab}, m).

Conversely, if (5.29) holds for $1 \leq a, b \leq N$, then (5.28) and hence the characterization equations (2.42) are obviously satisfied. □

Combining these results we can treat the reconstruction problem by the following procedures. Let $T(k, m)$ be an off-diagonal $N \times N$-matrix-valued function that satisfies conditions (i) and (ii) of Definition 2.3. If condition (5.29) is satisfied, then $T \in D^o_{N \times N}$ and T belongs to the range of $\mathbf{S}$ if it is small enough. Let $1 \leq \hat{a} < \hat{b} \leq N$. Proposition 5.1 and Corollary 5.1 imply that the $\hat{a}\hat{b}$th and $\hat{b}\hat{a}$th component of the potential Q can be reconstructed through (5.6), (5.12), and (5.13) by letting $\hat{T}_{12}(\chi, m) = \Phi_{\hat{a}\hat{b}}(\chi, m)$ and $\hat{T}_{21}(\chi, m) = \Phi_{\hat{b}\hat{a}}(\chi, m)$.

In other words, we can replace condition (iii) of Definition 2.3 by (5.29). If (5.29) is satisfied, we can reconstruct the components of Q two at a time by solving a 2×2 inverse problem, and the inverse data of such 2×2 systems are already computed in the checking of condition (5.29).

REFERENCES

[1] R. BEALS AND R. R. COIFMAN, *Multidimensional inverse scattering and nonlinear partial differential equations*, Proc. Sympos. Pure Math., 43, American Mathematical Society, Providence, RI, 1985, pp. 45–70.

[2] ———, *The spectral problem for the Davey–Stewartson and Ishimori hierarchies*, Proc. Conf. on Nonlinear Evolution Equations: Integrability and Spectral Methods, Como, University of Manchester, Manchester, UK, 1988.

[3] L. D. FADDEEV, *Inverse problem of quantum scattering theory*, II, J. Sov. Math., 5 (1976), pp. 334–396.

[4] A. S. FOKAS, *Inverse scattering of first-order systems in the plane related to nonlinear multidimensional equations*, Phys. Rev. Lett., 51 (1983), pp. 3–6.

[5] ———, *An inverse problem for multidimensional first-order systems*, J. Math. Phys., 27 (1986), pp. 1737–1746.

[6] A. S. FOKAS AND M. J. ABLOWITZ, *On the inverse scattering transform of multidimensional nonlinear equations related to first-order systems in the plane*, J. Math. Phys., 25 (1984), pp. 2494–2505.

[7] G. M. HENKIN AND R. G. NOVIKOV, *A multidimensional inverse problem in quantum and acoustic scattering*, Inverse Problems, 4 (1988), pp. 103–121.

[8] L. HÖRMANDER, *The Analysis of Linear Partial Differential Operators* II, Springer-Verlag, Berlin, 1983.

[9] D. J. KAUP, *The inverse scattering solution for the full three dimensional three-wave resonant interaction*, Physica D, 1 (1980), pp. 45–67.

[10] R. B. LAVINE AND A. I. NACHMAN, *The inverse problem in the coupling constant*, in Inverse Problems, P. G. Sabatier, ed., Academic Press, New York, 1982.

[11] A. I. NACHMAN AND M. J. ABLOWITZ, *A multidimensional inverse-scattering method*, Stud. Appl. Math., 71 (1984), pp. 243–250.

[12] ———, *Mutidimensional inverse scattering for first-order systems*, Stud. Appl. Math., 71 (1984), pp. 251–262.

[13] R. G. NEWTON, *Inverse scattering* II. *Three dimensions*, J. Math. Phys., 21 (1980), pp. 1698–1715.

[14] ———, *Inverse scattering* III. *Three dimensions, continued*, J. Math. Phys., 22 (1981), pp. 2191–2200.

[15] ———, *Inverse scattering* IV. *Three dimensions: Generalized Marchenko construction with bound states, and generalized Gel'fand–Levitan equations*, J. Math. Phys., 23 (1982), pp. 594–604.

[16] R. G. NEWTON, *Erratum: Inverse Scattering* II. *Three Dimensions* and *Erratum: Inverse Scattering* III. *Three Dimensions, Continued*, J. Math. Phys., 23 (1982), p. 693.

[17] ———, *Scattering Theory of Waves and Particles*, Springer-Verlag, New York, 1982.

[18] R. G. NOVIKOV AND G. M. HENKIN, *The $\bar{\partial}$-equation in the mutidimensional inverse scattering problem*, Usp. Mat. Nauk, 42 (1987), pp. 93–152; Russian Math. Surveys, 42 (1987), pp. 109–180.

[19] L. Y. SUNG AND A. S. FOKAS, *Inverse scattering in the plane and the Davey–Stewartson equations*, preprint.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1332–1348, September 1991

008

THERMOSTAT SYSTEMS HAVING STABLE PERIODIC SOLUTIONS WITH SHORT PERIODS*

GUSTAF GRIPENBERG†

Abstract. The local stability of periodic solutions of a system controlled by a thermostat is studied in the case where the periods are arbitrarily short. If the system is modeled by the equation $y(t) = \int_{-\infty}^{t} a(t-s)u(s)\,ds$ and the kernel a satisfies certain crucial smoothness assumptions, then such solutions exist if (and in most cases only if) $\hat{a}(w) \neq 0$ for $\Re w \geq 0$ and either $a(0) > 0$ or $a(0) = 0$, $a'(0) > 0$, and $a''(0) < 0$.

Key words. thermostat control, short period, stability, hysteresis

AMS(MOS) subject classifications. 34K35, 45D05, 45M05

1. Introduction and statement of results.

The purpose of this paper is to study the local stability of periodic solutions of the equation

$$y(t) = \int_{-\infty}^{t} a(t-s)u(s)\,ds, \qquad t \in \mathbf{R}, \tag{1}$$

when the period approaches zero. Here $u : \mathbf{R} \to \{0, 1\}$ is determined by the requirement that, if $u(t) = 1$ and y reaches an upper level θ_{high} at time t, then $u(t+) = 0$, and conversely, if $u(t) = 0$ and y reaches a lower level θ_{low} at time t, then $u(t+) = 1$, or, in other words, the system functions like a thermostat. Thus we see that if the period is short, then $\theta_{\text{high}} - \theta_{\text{low}}$ must be small and $y(t)$ must stay close to the interval $[\theta_{\text{low}}, \theta_{\text{high}}]$. This is, of course, the reason why we would want to have a short period.

The main motivation for equation (1) is that it is a reasonable model for a system controlled by a thermostat. For example, consider a situation where a heater is turned on if the temperature at some fixed point drops to a level θ_{low} and is turned off if the temperature rises to a level θ_{high} where we have $\theta_{\text{low}} < \theta_{\text{high}}$. It is convenient to normalize the equation in such a way that without heating the temperature is zero and with uninterrupted heating it is 1. The function a is determined by the requirement that if the heating is turned on at time zero for the first time, then the temperature should be $\int_0^t a(s)\,ds$ at time $t > 0$. Some external influences can also be added to (1).

Models of this kind have earlier been studied in, e.g., [2]–[9] and [13], but in many of these references the state equation is not explicitly written in the form (1). In [4] and [5], the main emphasis is on questions of existence of solutions and on a precise description of the heating process involving diffusion. In [9] this work is continued and some sufficient conditions for the existence of periodic solutions are given. In [3] thermostat control in a diffusion model is considered, and the existence—and in some cases uniqueness—of periodic solutions is established. In [2] the question of how to choose parameters optimally (e.g., θ_{low} and θ_{high}) is considered. In [8] the existence and asymptotic behavior of solutions of a diffusion problem is studied, and in [13] it is shown that there are infinitely many periodic solutions of a heat-control problem where $\theta_{\text{low}} = \theta_{\text{high}}$.

*Received by the editors April 16, 1990; accepted for publication September 6, 1990.

†Department of Mathematics, University of Helsinki, Regeringsgatan 15, 00100 Helsingfors, Finland.

In [1] and [10]–[12] a class of related switching systems is considered. In [1] the state equation is written as $L(y, s) = 0$ where L is a differential operator (for example, the heat operator) with respect to y and s is a switching variable and in [10]–[12] it is of the form $y'(t) = f_j(y(t))$ where the index j, and hence the state equation, is changed when the solution reaches a "switching surface." In [12] sufficient conditions for the existence of periodic solutions for this kind of system are given.

Equation (1) is an example of a system involving hysteresis, but in this paper no results from the general theory of such systems will be used.

In [6] the existence of periodic solutions of equation (1) is established, but there the dependence of u on y is in general a slightly weaker form of thermostat control (although the difference appears only in the case where y has a local maximum equal to θ_{high} or a local minimum equal to θ_{low}). In [7] a criterion for the local stability of these periodic solutions is given, and the present paper is based on these results.

Here we will use the following notion of thermostat control (which could also be formulated in many other equivalent ways).

DEFINITION 1. If $I \subset \mathbf{R}$ is an interval and $y : I \to \mathbf{R}$ is a continuous function, then a function $u : I \to \{0, 1\}$ is (strictly) *thermostat controlled* by y with respect to the higher limit θ_{high} and the lower limit θ_{low} on the interval I provided that u is left-continuous with right-hand limits on I and the following conditions hold for all $t \in I$:

(i) $u(t) = 1$ if $y(t) < \theta_{\text{low}}$,
(ii) $u(t) = 0$ if $y(t) > \theta_{\text{high}}$,
(iii) If $y(t) = \theta_{\text{high}}$ and $u(t) = 1$, then $u(t+) = 0$, and if $u(t) - u(t+) = 1$, then $y(t) = \theta_{\text{high}}$,
(iv) If $y(t) = \theta_{\text{low}}$ and $u(t) = 0$, then $u(t+) = 1$, and if $u(t) - u(t+) = -1$, then $y(t) = \theta_{\text{low}}$.

The question that is asked and (at least to some extent) answered in this paper is the following. When are there arbitrarily small positive numbers T and S such that there exists a locally stable periodic solution (y_*, u_*) of (1) with $u_* = 1$ on intervals of length T, $u_* = 0$ on intervals of length S, and u_* is strictly thermostat controlled by y_* on $\mathbf{R}$ with respect to θ_{high} and θ_{low}.

In order to define what we mean by stability, we consider the equation

$$y(t) = \int_{-\infty}^{t} a(t-s)u(s)\,ds + e(t), \qquad t \geq 0, \tag{2}$$

where u is given on $(-\infty, 0]$ and u is strictly thermostat controlled by y on $(0, \infty)$. To fix the notation we define the sequences $\{v_n(u)\}$ and $\{w_n(u)\}$ by the requirement that

$$u(t) = \begin{cases} 1, & t \in (v_n(u), w_n(u)], \\ 0, & t \in (w_n(u), v_{n+1}(u)], \end{cases} \qquad n \in \mathbf{Z}.$$

No generality is lost by requiring that $v_0(u) = 0$, i.e., $u(0) = 0$ and $u(0+) = 1$. We define the numbers $T_n(u)$ and $S_n(u)$ by

$$T_n(u) = w_n(u) - v_n(u),$$
$$S_n(u) = v_{n+1}(u) - w_n(u),$$

so that u is 1 on intervals of length $T_n(u)$ and zero on intervals of length $S_n(u)$.

DEFINITION 2. If (y_*, u_*) is a periodic solution of (1) on $\mathbf{R}$ with $u_* = 1$ on intervals of length T, $u_* = 0$ on intervals of length S, and u_* is strictly thermostat

controlled by y_* on $\mathbf{R}$ (with respect to θ_{low} and θ_{high}), then (y_*, u_*) is locally stable provided that for each $\epsilon > 0$ there exists $\delta > 0$ such that, if $e \in C(\mathbf{R}^+, \mathbf{R})$, if $u : \mathbf{R} \to \{0, 1\}$ is left-continuous with right-hand limits, $u(0) = 0$, $u(0+) = 1$, and if u is strictly thermostat controlled by the solution y of (2) on $(0, \infty)$, then the inequalities

$$\begin{aligned} &|T_i(u) - T| + |S_i(u) - S| \le \delta, \qquad i < 0, \\ &|e(t)| \le \delta, \qquad t \ge 0, \end{aligned}$$

imply that for every $n \ge 0$,

$$|T_n(u) - T| + |S_n(u) - S| \le \epsilon,$$

and in the case where $\lim_{t\to\infty} e(t) = 0$ it follows that

$$T_n(u) \to T \quad \text{and} \quad S_n(u) \to S \quad \text{as } n \to \infty.$$

In [7] it is shown that the following condition is a sufficient condition for local stability (in fact, a slightly stronger result is obtained): The function

$$\begin{aligned} f(z) \overset{\text{def}}{=} {} & \frac{1}{z}\sum_{j=1}^{\infty} a\big(j(T+S) - S\big)z^j \sum_{j=1}^{\infty} a\big(j(T+S) - T\big)z^j \\ & - \left(\sum_{j=1}^{\infty}\Big(a\big(j(T+S)\big)z^j - a\big(j(T+S)\big) + a\big(j(T+S) - T\big)\Big)\right) \\ & \times \left(\sum_{j=1}^{\infty}\Big(a\big(j(T+S)\big)z^j - a\big(j(T+S)\big) + a\big(j(T+S) - S\big)\Big)\right) \end{aligned} \tag{3}$$

has a simple zero at $z = 1$ and no other zeros in $\{\, z \in \mathbf{C} \mid |z| \le 1 \,\}$.

Therefore, the main part of the paper consists of a detailed analysis of this function f when T and S are close to zero.

Some of the assumptions that we use are unnecessarily strong, but the main weakness in the assumptions is that, kernels a of the form $t^\alpha b(t)$ with $b(0) \ne 0$ and $\alpha > -1$ not equal to an integer they do not include.

THEOREM 3. *Let $\theta_* \in (0, 1)$, let $m \ge 0$ be an integer and assume that*

(i) *$a \in L^1(\mathbf{R}^+; \mathbf{R})$, $\int_0^\infty a(t)\,dt = 1$, $a^{(j)} \in BV(\mathbf{R}^+; \mathbf{R}) \bigcap AC(\mathbf{R}^+; \mathbf{R})$ for $j = 0, \cdots, m$, $a^{(j)}(0) = 0$ for $j = 0, \cdots, m-1$, $a^{(m)}(0) \ne 0$, and $\int_0^\infty t^2 |a'(t)|\,dt < \infty$;*

(ii) *If $m = 0$ or 1, then $a^{(j)} \in BV(\mathbf{R}^+; \mathbf{R}) \bigcap AC(\mathbf{R}^+; \mathbf{R})$ for $j = 0, \cdots, 2m+1$, and $\int_0^\infty t\big|a^{(2m+2)}(t)\big|\,dt < \infty$;*

(iii) *There are positive numbers T and S such that there exists a periodic solution (y_*, u_*) of (1) with $u_* = 1$ on intervals of length T, $u_* = 0$ on intervals of length S, $u_*(0) = 0$, $u_*(0+) = 1$, and u_* is (strictly) thermostat controlled by y_* on $\mathbf{R}$ with respect to θ_{low} and θ_{high}, satisfying $0 < \theta_{\text{low}} < \theta_* < \theta_{\text{high}} < 1$.*

Then the following conclusions hold.

(a) *If $\hat{a}(w) \ne 0$ for $w \in \mathbf{C}$ with $\Re w \ge 0$ and either $m = 0$, or $m = 1$ and $a''(0) < 0$, then there exists a number $\mu > 0$ such that (y_*, u_*) is locally stable provided $T + S < \mu$.*

(b) *If $\hat{a}(w_0) = 0$ for some $w_0 \in \mathbf{C}$ with $\Re w_0 > 0$, or $m = 1$ and $a''(0) > 0$, or $m > 1$, then there exists a number $\mu > 0$ such that (y_*, u_*) is not locally stable provided $T + S < \mu$.*

We see from the proof that in some of the cases where instability is claimed, it is due to the fact that there are no periodic solutions at all, so that the assumptions of the theorem are not satisfied. Note also that if $m = 0$ and $a(0) < 0$, or $m = 1$ and $a'(0) < 0$, then there exists a number $\sigma > 0$ such that $\hat{a}(\sigma) = 0$, but in these cases it is also true that there are no periodic solutions with short periods.

When proving part (b) we have to extend some of the results found in [7]. Therefore we consider the discrete equation

$$\sum_{j=-\infty}^{n} \alpha_{n-j}\xi_j = \Phi_n\left(\{\xi_{n-j}\}_{j=0}^{\infty}\right) + \varphi_n, \qquad n \geq 0; \tag{4}$$

$$\xi_i = \psi_i, \qquad i < 0.$$

The criterion for stability of the linear part of (4) is $\det[\sum_{n=0}^{\infty} z^n\alpha_n] \neq 0$ when $|z| \leq 1$, so it is no surprise that the trivial solution is unstable if there is a zero inside the unit circle. (If there are zeros on the unit circle, then we can at least show as in [7] that a certain stronger kind of stability cannot hold.)

For completeness we state and prove the result that we need concerning (4). Let us denote the set of integers by $\mathbf{Z}$, the set of negative integers by $\mathbf{Z}_-$, and the set of natural numbers, i.e., the nonnegative integers, by $\mathbf{N}$. By $|\cdot|$ we denote some norm in $\mathbf{R}^m$, and also the corresponding matrix norm.

PROPOSITION 4. *Assume that the following conditions hold:*

(i) $\alpha \in l^1(\mathbf{N}; \mathbf{R}^{m\times m})$;

(ii) *For each* $j \geq 0$ *the mapping* $\zeta \in l^\infty(\mathbf{N}; \mathbf{R}^m) \mapsto \Phi_j(\zeta) \in \mathbf{R}^m$ *is continuous and* $\|\Phi(\zeta)\|_{l^\infty(\mathbf{N})} = o\big(\|\zeta\|_{l^\infty(\mathbf{N})}\big)$ *as* $\|\zeta\|_{l^\infty(\mathbf{N})} \to 0$;

(iii) $\varphi \in l^\infty(\mathbf{N}; \mathbf{R}^m)$;

(iv) $\psi \in l^\infty(\mathbf{Z}_-; \mathbf{R}^m)$;

(v) $\det[\sum_{n=0}^{\infty} z_0^n\alpha_n] = 0$ *for some* $z_0 \in \mathbf{C}$ *with* $|z_0| < 1$.

Then there exists a number $\epsilon > 0$ *such that there is no* $\delta > 0$ *for which the inequalities*

$$|\psi_i| \leq \delta, \qquad i < 0, \tag{5}$$
$$|\varphi_n| \leq \delta, \qquad n \geq 0$$

imply that, if ξ *is a solution of* (4), *then*

$$|\xi_n| \leq \epsilon, \qquad n \geq 0. \tag{6}$$

2. Proof of Theorem 3(a). It is a consequence of the assumptions that $\lim_{\sigma\to\infty} \sigma^{m+1}\hat{a}(\sigma) = a^{(m)}(0)$ and $\hat{a}(0) = 1$. Since, moreover, $\hat{a}(\sigma) \neq 0$ when $\sigma \geq 0$, it follows that

$$a^{(m)}(0) > 0. \tag{7}$$

Let us introduce the following abbreviations:

$$\mathcal{T} = T + S,$$
$$q = \frac{T}{T+S},$$
$$t_j(q) = (j - q)(T + S).$$

Because $\int_0^\infty a(t)\,dt = 1$ it follows from (iii) that $\lim_{\mathcal{T}\downarrow 0}\bigl(\int_{-\infty}^t a(t-s)u_*(s)\,ds - q\bigr) = 0$, uniformly for $t \in \mathbf{R}$. Hence q is close to θ_* when $\mathcal{T}$ is small, and this guarantees that q and $1-q$ remain bounded away from zero and 1. We may assume that $u_*(0) = 0$ and $u_*(0+) = 1$.

It follows from the results in [7] that, if we want to show that the periodic solution is stable, then we have to prove that (3) holds. Let $\mathcal{D} = \{\, z \in \mathbf{C} \mid |z| \le 1 \,\}$. First we consider the case where

$$z \in \mathcal{D}_{\mathcal{T}} \stackrel{\text{def}}{=} \{\, z \in \mathcal{D} \mid |z-1| \le \gamma(\mathcal{T}) \,\},$$

where $\gamma : \mathbf{R}^+ \to [0, \frac{1}{2}]$ is some continuous function with $\gamma(0) = 0$ that will be fixed later. Define

$$w = -\frac{\log(z)}{\mathcal{T}} \quad \text{and} \quad \mathcal{H}_{\mathcal{T}} = \{\, w \in \mathbf{C} \mid \exp(-\mathcal{T}w) \in \mathcal{D}_{\mathcal{T}} \,\}.$$

We let

$$A(t,w) = a(t)\mathrm{e}^{-wt}, \qquad t \ge 0, \quad \Re w \ge 0,$$

and by $A'(t,w)$ we denote the derivative with respect to t.

Using an integration by parts we get the following result for $p \in [0,1)$:

$$\begin{aligned} z^{-p}\sum_{j=1}^{\infty} a\bigl(t_j(p)\bigr)z^j &= \sum_{j=1}^{\infty} a\bigl(t_j(p)\bigr)\mathrm{e}^{-t_j(p)w} \\ &= \frac{1}{\mathcal{T}}\hat{a}(w) + \frac{1}{\mathcal{T}}\sum_{j=1}^{\infty}\int_{t_{j-1}(0)}^{t_j(0)} \Bigl(a\bigl(t_j(p)\bigr)\,\mathrm{e}^{-t_j(p)w} - a(s)\,\mathrm{e}^{-sw}\Bigr)\,ds \\ &= \frac{1}{\mathcal{T}}\hat{a}(w) + \sum_{j=1}^{\infty} z^{j-1}\int_{t_{j-1}(0)}^{t_j(0)} h_0\Bigl(\frac{s-t_{j-1}(0)}{\mathcal{T}}, p\Bigr)A'(s,w)\,ds, \end{aligned} \tag{8}$$

where h_0 is defined by

$$h_0(t,p) = t - \chi_{[1-p,1]}(t), \qquad t \in [0,1], \quad p \in [0,1]. \tag{9}$$

For $k \ge 0$ we let

$$\begin{aligned} c_k(p) &= \int_0^1 h_k(s,p)\,ds, \\ h_{k+1}(t,p) &= -\int_0^t \bigl(h_k(s,p) - c_k(p)\bigr)\,ds, \end{aligned} \tag{10}$$

and then we define

$$R_k(p,\mathcal{T}) = \sum_{j=1}^{\infty}\int_{t_{j-1}(0)}^{t_j(0)} h_k\Bigl(\frac{s-t_{j-1}(0)}{\mathcal{T}}, p\Bigr)A^{(k+1)}(s,w)\,ds. \tag{11}$$

Provided $a^{(j)}$ is integrable for $0 \le j \le k+2$ we get from an integration by parts that

$$R_k(p,z,\mathcal{T}) = -c_k(p)A^{(k)}(0,w) + \mathcal{T}R_{k+1}(p,z,\mathcal{T}). \tag{12}$$

By (8) we therefore have

$$z^{-p}\sum_{j=1}^{\infty} a\big(t_j(p)\big)z^j = \frac{1}{\mathcal{T}}\hat{a}(w) - \sum_{k=0}^{n} c_k(p)A^{(k)}(0,w)\mathcal{T}^k + \mathcal{T}^{n+1}R_{n+1}(p,z,\mathcal{T}), \tag{13}$$

where $n = m-1$ if $m > 1$ and $n = 2m$ if $m = 0$ or 1.

Using an approximation argument we conclude that

$$R_k(p,z,\mathcal{T}) \to -c_k(p)A^{(k)}(0,w) \text{ as } \mathcal{T} \downarrow 0 \text{ uniformly for } w \in \mathcal{H}_{\mathcal{T}} \text{ and } p \in [0,1], \tag{14}$$

for $0 \le k \le m$ if $m > 1$, and $0 \le k \le 2m+1$ otherwise. Since $\mathcal{T}w/(z-1) \to -1$ as $\mathcal{T} \downarrow 0$ we see from a similar argument that when $m = 0$ or 1, and $k = 0, \cdots, 2m+1$,

$$\frac{\mathcal{T}}{z-1}\big(R_k(p,z,\mathcal{T}) - R_k(p,1,\mathcal{T})\big) \to c_k(p)\frac{1}{w}\big(A^{(k)}(0,w) - A^{(k)}(0,0)\big) \tag{15}$$
$$\text{as } \mathcal{T} \downarrow 0 \text{ uniformly for } w \in \mathcal{H}_{\mathcal{T}} \text{ and } p \in [0,1].$$

Note that the moment assumption in (ii) is needed for the case $k = 2m+1$ and once this is established the remaining cases follow from (12).

It is an immediate consequence of definitions (9) and (10) that

$$h_{k+1}(t,p) = \frac{(-1)^{k+1}}{k!}\int_0^t (t-s)^k h_0(s,p)\,ds - \sum_{j=0}^{k} c_j(p)\frac{(-t)^{k-j+1}}{(k-j+1)!},$$

and since $h_{k+1}(1,p) = 0$ by the definition of $c_k(p)$, we get

$$\sum_{j=0}^{k} c_j(p)\frac{(-1)^{k-j+1}}{(k-j+1)!} = \frac{(-1)^{k+1}}{(k+2)!} - \frac{(-p)^{k+1}}{(k+1)!}. \tag{16}$$

From this equation it is easy to solve the coefficients $c_j(p)$ recursively.

It follows from (3), (11), and (13) that

$$\begin{aligned} f(z) &= \Big[\frac{\hat{a}(w)}{\mathcal{T}} + R_0(q,z,\mathcal{T})\Big]\Big[\frac{\hat{a}(w)}{\mathcal{T}} + R_0(1-q,z,\mathcal{T})\Big] \\ &\quad - \Big[\frac{\hat{a}(w)}{\mathcal{T}} + \Big(R_0(0,z,\mathcal{T}) - R_0(0,1,\mathcal{T}) + R_0(q,1,\mathcal{T})\Big)\Big] \\ &\quad \times \Big[\frac{\hat{a}(w)}{\mathcal{T}} + \Big(R_0(0,z,\mathcal{T}) - R_0(0,1,\mathcal{T}) + R_0(1-q,1,\mathcal{T})\Big)\Big] \\ &= \frac{\hat{a}(w)}{\mathcal{T}}D(q,z,\mathcal{T}) + E(q,z,\mathcal{T}) + F(q,z,\mathcal{T}), \end{aligned} \tag{17}$$

where

$$\begin{aligned} D(q,z,\mathcal{T}) &\overset{\text{def}}{=} R_0(q,z,\mathcal{T}) + R_0(1-q,z,\mathcal{T}) - 2R_0(0,z,\mathcal{T}) \\ &\quad - R_0(q,1,\mathcal{T}) - R_0(1-q,1,\mathcal{T}) + 2R_0(0,1,\mathcal{T}), \\ E(q,z,\mathcal{T}) &\overset{\text{def}}{=} \big(R_0(q,z,\mathcal{T}) - R_0(q,1,\mathcal{T})\big)R_0(1-q,z,\mathcal{T}) \\ &\quad + \big(R_0(1-q,z,\mathcal{T}) - R_0(1-q,1,\mathcal{T})\big)R_0(q,1,\mathcal{T}), \end{aligned}$$

and

$$F(q,z,\mathcal{T}) \stackrel{\text{def}}{=} \Big(R_0(0,z,\mathcal{T}) - R_0(0,1,\mathcal{T})\Big)^2 - \Big(R_0(0,z,\mathcal{T}) - R_0(0,1,\mathcal{T})\Big)\Big(R_0(q,1,\mathcal{T}) + R_0(1-q,1,\mathcal{T})\Big).$$

Assume first that $m = 0$. If $z \in \mathcal{D}_{\mathcal{T}}$, then $w\mathcal{T} \to 0$ as $\mathcal{T} \to 0$. On the other hand, we know that $w\hat{a}(w) \to a(0)$ as $|w| \to \infty$ with $\Re w \geq 0$. Therefore it follows from the fact that $\hat{a}(w) \neq 0$ for $\Re w \geq 0$ that

$$\inf_{w \in \mathcal{H}_{\mathcal{T}}} \left| \frac{\hat{a}(w)}{\mathcal{T}} \right| \to \infty \quad \text{as } \mathcal{T} \downarrow 0. \tag{18}$$

From (12), (15), and (16) we get

$$\begin{aligned} \frac{D(q,z,\mathcal{T})}{z-1} &\to \big(c_1(q) + c_1(1-q) - 2c_1(0)\big)\frac{1}{w}\big(A'(0,w) - A'(0,0)\big) \\ &= a(0)q(1-q) \quad \text{as } \mathcal{T} \downarrow 0 \text{ uniformly for } w \in \mathcal{H}_{\mathcal{T}} \text{ and } p \in [0,1]. \end{aligned} \tag{19}$$

Because $A(0,w) = a(0)$ we get from (12) that

$$R_0(p,z,\mathcal{T}) - R_0(p,1,\mathcal{T}) = \mathcal{T}\big(R_1(p,z,\mathcal{T}) - R_1(p,1,\mathcal{T})\big). \tag{20}$$

Therefore it follows from (14) and (15) that

$$\frac{E(q,z,\mathcal{T})}{z-1} = O(1) \quad \text{as } \mathcal{T} \downarrow 0 \text{ uniformly for } w \in \mathcal{H}_{\mathcal{T}}. \tag{21}$$

Moreover, since $c_0(p) = \frac{1}{2} - p$, we see from (12) that

$$R_0(q,1,\mathcal{T}) + R_0(1-q,1,\mathcal{T}) = \mathcal{T}\big(R_1(q,1,\mathcal{T}) + R_1(1-q,1,\mathcal{T})\big), \tag{22}$$

and hence we conclude from (14), (15), and (20) that

$$\frac{F(q,z,\mathcal{T})}{z-1} = O(\mathcal{T}) \quad \text{as } \mathcal{T} \downarrow 0 \text{ uniformly for } w \in \mathcal{H}_{\mathcal{T}}.$$

Combining this result with (19) and (21) we get from (17) and (18) that

$$\frac{f(z)}{z-1} = \big(1 + o(1)\big)\frac{\hat{a}(w)}{\mathcal{T}} \quad \text{as } \mathcal{T} \downarrow 0, \quad w \in \mathcal{H}_{\mathcal{T}}. \tag{23}$$

Thus we see that if $m = 0$ then $f(z)/(z-1)$ has no zeros in a neighborhood of the point 1.

Next we consider the case where $m = 1$. By using the fact that now $a(0) = 0$ and $c_2(q) + c_2(1-q) - 2c_2(0) = 0$ we get from an argument similar to the one used above that

$$\begin{aligned} \frac{D(q,z,\mathcal{T})}{z-1} &\to \mathcal{T}^2\Big(c_3(q) + c_3(1-q) - 2c_3(0)\Big)\frac{1}{w}\Big(A^{(3)}(0,w) - A^{(3)}(0,0)\Big) \\ &= \mathcal{T}^2\big(-a''(0) + wa'(0)\big)\frac{q^2(1-q)^2}{4} \\ &\qquad \text{as } \mathcal{T} \downarrow 0 \text{ uniformly for } w \in \mathcal{H}_{\mathcal{T}} \text{ and } p \in [0,1]. \end{aligned}$$

Since $a(0) = 0$ we get instead of (20) that

$$R_0(p,z,\mathcal{T}) - R_0(p,1,\mathcal{T}) = \mathcal{T}^2\big(R_2(p,z,\mathcal{T}) - R_2(p,1,\mathcal{T})\big), \tag{24}$$

and also that $R_0(p,z,\mathcal{T}) = \mathcal{T}R_1(p,z,\mathcal{T})$. This implies that

$$\frac{E(q,z,\mathcal{T})}{z-1} = O(\mathcal{T}^2) \quad \text{as } \mathcal{T} \downarrow 0 \text{ uniformly for } w \in \mathcal{H}_{\mathcal{T}}.$$

From (14), (22), and (24) it also follows that

$$\frac{F(q,z,\mathcal{T})}{z-1} = O(\mathcal{T}^2) \quad \text{as } \mathcal{T} \downarrow 0 \text{ uniformly for } w \in \mathcal{H}_{\mathcal{T}}.$$

Now we know that $w^2\hat{a}(w) \to a'(0)$ as $|w| \to \infty$ with $\Re w \geq 0$ and that $-a''(0) + wa'(0) \neq 0$ when $\Re w \geq 0$. Therefore we have by an argument similar to the one used in deriving (18) that

$$\inf_{w \in \mathcal{H}_{\mathcal{T}}} \left| \frac{\hat{a}(w)}{\mathcal{T}}\big(-a''(0) + wa'(0)\big) \right| \to \infty \quad \text{as } \mathcal{T} \downarrow 0.$$

Thus we see that

$$\frac{f(z)}{z-1} = \big(1 + o(1)\big)\frac{\mathcal{T}}{4}\hat{a}(w)\big(-a''(0) + wa'(0)\big) \quad \text{as } \mathcal{T} \downarrow 0 \text{ and } w \in \mathcal{H}_{\mathcal{T}}. \tag{25}$$

This completes the proof in the case $m = 1$ as well.

Next we consider the case where z does not lie in a neighborhood of 1 in the unit disk. Define the function b by

$$b(t) = \begin{cases} a\Big(\frac{t\mathcal{T}}{2\pi}\Big)\mathrm{e}^{t\log(z)/(2\pi)}, & t \geq 0, \\ 0, & t < 0 \end{cases}$$

and let

$$B(t) = \sum_{k=-\infty}^{\infty} b(t + 2k\pi), \qquad t \in [0, 2\pi].$$

It follows from our assumptions that B is of bounded variation and therefore we have (note that all sums $\sum_{n=-\infty}^{\infty}$ are to be interpreted as $\lim_{N\to\infty}\sum_{n=-N}^{N}$)

$$\frac{1}{2}\big(B(t-) + B(t+)\big) = \sum_{n=-\infty}^{\infty} \mathrm{e}^{int}\tilde{B}(n), \tag{26}$$

where the Fourier coefficients $\tilde{B}(n)$ are given by

$$\tilde{B}(n) = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{e}^{-int} \sum_{k=-\infty}^{\infty} b(t + 2k\pi)\,dt = \frac{1}{2\pi}\int_0^{\infty} \mathrm{e}^{-int}b(t)\,dt. \tag{27}$$

Choosing $t = 2\pi(1-p)$ in (26) we get (assuming for the moment that $p \in (0,1)$)

$$\begin{aligned} B(t) &= \sum_{k=0}^{\infty} a\big((1-p)\mathcal{T} + k\mathcal{T}\big)z^k z^{1-p} \\ &= z^{-p}\sum_{j=1}^{\infty} a\big(t_j(p)\big)z^j = \sum_{n=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}2n\pi(1-p)}\tilde{B}(n). \end{aligned} \tag{28}$$

Invoking the definition of b and performing a change of variables we conclude from (27) that the Fourier coefficients $\tilde{B}(n)$ are given by

$$\begin{aligned}\tilde{B}(n) =&\frac{1}{2\pi}\int_0^\infty \mathrm{e}^{-\mathrm{i}nt}a\Big(\frac{t\mathcal{T}}{2\pi}\Big)\mathrm{e}^{t\log(z)/(2\pi)}\,dt\\ =&\frac{1}{\mathcal{T}}\int_0^\infty \mathrm{e}^{-t(\mathrm{i}2n\pi-\log(z))/\mathcal{T}}a(t)\,dt=\frac{1}{\mathcal{T}}\hat{a}\Big(\frac{\mathrm{i}2n\pi-\log(z)}{\mathcal{T}}\Big).\end{aligned}\tag{29}$$

Since we assume that $a^{(j)}\in BV(\mathbf{R}^+;\mathbf{R})$ for $j=0,\cdots,m$, we get

$$\hat{a}(w)=\frac{\widehat{a^{(m+1)}}(w)}{w^{m+1}}+\sum_{j=0}^{m}\frac{a^{(j)}(0)}{w^{j+1}}.\tag{30}$$

Let

$$H_j(w,p)=\frac{\mathrm{d}^j}{\mathrm{d}w^j}\frac{\mathrm{e}^{w(1-p)}}{1-\mathrm{e}^w},\qquad j\geq 0,\quad p\in[0,1].\tag{31}$$

For $p\in(0,1)$ we have

$$H_j(w,p)=j!\sum_{n=-\infty}^{\infty}\frac{\mathrm{e}^{\mathrm{i}2n\pi(1-p)}}{(\mathrm{i}2n\pi-w)^{j+1}}.$$

Because $a^{(m)}\in AC(\mathbf{R}^+;\mathbf{R})$ we see from the Riemann–Lebesgue lemma that we have $\widehat{a^{(m+1)}}(w)\to 0$ as $|w|\to\infty$ with $\Re w\geq 0$. Therefore, from (28)–(30) and the assumptions that $a^{(j)}(0)=0$ for $j=0,\cdots,m-1$ (and $a'\in BV(\mathbf{R}^+;\mathbf{R})$ when $m=0$), it follows that there exists a continuous function γ with $\gamma(0)=0$ such that

$$\begin{aligned}\sum_{j=1}^{\infty}a\big(t_j(p)\big)z^j=z^p a^{(m)}(0)H_m(\log(z),p)\frac{\mathcal{T}^m}{m!}+o\big(\mathcal{T}^m\big)\\ \text{as } \mathcal{T}\downarrow 0 \text{ and } |z|\leq 1,\quad |z-1|\geq\gamma(\mathcal{T}).\end{aligned}\tag{32}$$

It is easy to check that this formula holds in the case $p=0$ as well. (The function B may have a discontinuity at zero, but the expression for $H_0(w,0)$ is modified in a corresponding way.)

Now let us consider the stability condition (3). In order to simplify the notation we define

$$d_k(p)=k!\big(c_k(0,1)-c_k(p,1)\big),\qquad k\geq 0,\quad p\in[0,1].\tag{33}$$

If we use (13), (14), and (32), then we see that the stability condition (3) becomes

$$f(z)=a^{(m)}(0)^2P_m(z,q)\mathcal{T}^{2m}+o\big(\mathcal{T}^{2m}\big)\neq 0,\tag{34}$$

where

$$\begin{aligned}P_m(z,q)=&H_m(\log(z),q)H_m(\log(z),1-q)\\ &-\big(H_m(\log(z),0)+d_m(q)\big)\big(H_m(\log(z),0)+d_m(1-q)\big).\end{aligned}\tag{35}$$

A simple calculation shows that

$$P_0(z,q) = -q(1-q) \quad \text{and} \quad P_1(z,q) = -\tfrac{1}{4}q^2(1-q)^2.$$

We conclude from (34) that, if $m = 0$ or 1 and if $\mathcal{T}$ is sufficiently small, then $f(z) \neq 0$ when $z \in \mathcal{D} \setminus \mathcal{D}_{\mathcal{T}}$.

Now the conclusion of part (a) follows from the results in [7] once we have proved that $y_*'(0-) < 0$ and $y_*'(T-) > 0$. Clearly, $y_*'(0-) = \sum_{j=1}^{\infty}\big(a(t_j(0) - a(t_j(q))\big)$ and $y_*'(T-) = \sum_{j=1}^{\infty}\big(a(t_j(1-q) - a(t_j(0))\big)$. Thus we see that it suffices to show that

$$\sum_{j=1}^{\infty}\big(a(t_j(0) - a(t_j(p))\big) < 0, \qquad p = q, \quad 1-q. \tag{36}$$

By (13) and (14) we have

$$\operatorname{sign}\left(\sum_{j=1}^{\infty}\big(a(t_j(0) - a(t_j(p))\big)\right) = -\operatorname{sign}\big(a^{(m)}(0)\big)\operatorname{sign}\big(d_m(p)\big) \tag{37}$$

for sufficiently small $\mathcal{T}$. Because $d_0(p) = p$ and $d_1(p) = p(1-p)/2$ we see that (36) follows from (7) and the proof is completed.

3. Proof of Theorem 3(b). We use the same notation as in the proof of part (a). First we prove that if $\mathcal{T}$ is sufficiently small, then either there exists a complex number z_0 with $|z_0| < 1$ such that $f(z_0) = 0$ or there is no periodic solution (y_*, u_*) such that u_* is thermostat controlled by y_*.

When $m = 0$ or 1 there exists a complex number w_0 with $\Re w_0 > 0$ such that $\hat{a}(w_0) = 0$ if $m = 0$ and such that $\hat{a}(w_0)(-a''(0) + w_0 a'(0)) = 0$ if $m = 1$. In the argument above it was assumed that $\hat{a}(w) \neq 0$ or $\hat{a}(w)(-a''(0) + wa'(0)) \neq 0$ when $\Re w \geq 0$, but it is easy to see that (23) or (25) holds if w is restricted to some compact set in the right half plane on which this expression does not vanish. Hence we can invoke the argument principle to show that the desired point z_0 exists.

Let $m > 1$. First we formulate some results on the coefficients $d_k(p)$ defined in (33).

LEMMA 5. *For each integer $k \geq 1$ and each $p \in (0,1)$,*

$$\begin{aligned}
&d_{2k}(p) = -d_{2k}(1-p),\\
&d_{2k-1}(p) = d_{2k-1}(1-p),\\
&d_{2k}(p)d_{2k}(1-p) < 0, \qquad p \neq \tfrac{1}{2},\\
&d_{4k-1}(p) < 0,\\
&d_{4k+1}(p) > 0.
\end{aligned}$$

Proof. It is a consequence of (16) and the definition of $d_j(p)$ that

$$\sum_{j=0}^{\infty} d_j(p)\frac{s^j}{j!} = \frac{1 - \mathrm{e}^{-ps}}{1 - \mathrm{e}^{-s}} \stackrel{\text{def}}{=} \varphi(p,s).$$

The first two claims follow from the fact that

$$\varphi(1-p,s) = 1 - \varphi(p,-s).$$

It is easy to see that the functions $d_k(p)$ can be expressed with the aid of Bernoulli polynomials in the following way:

$$d_k(p) = \frac{(-1)^k}{k+1}\big(B_{k+1}(p) - B_{k+1}(0)\big), \qquad k \geq 0.$$

Since the Bernoulli polynomials satisfy $B_n'(p) = nB_{n-1}(p)$, $\int_0^1 B_n(p)\,dp = 0$, and $B_1(p) = p - \frac{1}{2}$ we can use an induction argument to prove that for all $k \geq 1$ and all $p \in (0,1)$ we have

$$\begin{aligned}
&d_k(0) = d_k(1) = 0,\\
&d_{4k-3}(p) > 0,\\
&\big(p - \tfrac{1}{2}\big)d_{4k-2}(p) < 0, \qquad p \neq \tfrac{1}{2},\\
&d_{4k-1}(p) < 0,\\
&\big(p - \tfrac{1}{2}\big)d_{4k}(p) > 0, \qquad p \neq \tfrac{1}{2}.
\end{aligned}$$

This gives the last three claims and the proof of Lemma 5 is complete.

It follows from the argument principle and (34) that, if there exists a point z_1 such that $|z_1| < 1$ and $P_m(z_1, q) = 0$, then there exists, at least when $\mathcal{T}$ is sufficiently small, a point z_0 with $|z_0| < 1$ such that $f(z_0) = 0$.

In the case where m is an even positive number and $q \neq \frac{1}{2}$, it follows from Lemma 5, that $d_m(q)$ and $d_m(1-q)$ have opposite signs. This implies, in view of (37) that, for sufficiently small $\mathcal{T}$ there cannot be any periodic solutions satisfying the conditions of thermostat control, and hence no locally stable ones either. If, on the other hand, $q = \frac{1}{2}$, then we use the fact that by Lemma 5 we have $d_m(\frac{1}{2}) = 0$ and therefore $P_m(0, \frac{1}{2}) = 0$. It follows that there exists a zero of f in the interior of the unit disc when $\mathcal{T}$ is sufficiently small.

In order to treat the remaining cases where $m > 1$ is odd, we note that, because

$$\frac{\mathrm{e}^{-(1-p)w}}{1 - \mathrm{e}^{-w}} = -\frac{\mathrm{e}^{pw}}{1 - \mathrm{e}^{w}},$$

it follows that

$$H_m(-w, p) = (-1)^{m+1} H_m(w, 1-p).$$

When we combine this result with Lemma 5 we get, since $H_m(w,0) = H_m(w,1)$ when $m > 0$, that

$$P_m(z, q) = P_m\Big(\frac{1}{z}, q\Big). \tag{38}$$

Next we observe that by (31)

$$H_k(w,q) = \frac{\mathrm{d}^k}{\mathrm{d}w^k}\Big(\mathrm{e}^{-qw}\frac{\mathrm{e}^w}{1-\mathrm{e}^w}\Big) \quad \text{and} \quad H_k(w,1-q) = \frac{\mathrm{d}^k}{\mathrm{d}w^k}\Big(\mathrm{e}^{qw}\Big(\frac{\mathrm{e}^w}{1-\mathrm{e}^w} + 1\Big)\Big).$$

Moreover,

$$\frac{\mathrm{d}^k}{\mathrm{d}w^k}\frac{\mathrm{e}^w}{1-\mathrm{e}^w} = g_k\Big(\frac{\mathrm{e}^w}{1-\mathrm{e}^w}\Big),$$

where the functions g_j are defined by $g_0(u) = u$ and $g_{j+1}(u) = g_j'(u)(u^2+u)$. Thus we see from (35) that

$$P_m(z,q) = \left[\sum_{j=0}^{m}\binom{m}{j}(-q)^j g_{m-j}\Big(\frac{z}{1-z}\Big)\right]\left[q^m + \sum_{j=0}^{m}\binom{m}{j}q^j g_{m-j}\Big(\frac{z}{1-z}\Big)\right] - \Big[g_m\Big(\frac{z}{1-z}\Big) + d_m(q)\Big]\Big[g_m\Big(\frac{z}{1-z}\Big) + d_m(1-q)\Big]. \tag{39}$$

By induction we easily prove that

$$g_j(u) = u + (2^j - 1)u^2 + \cdots.$$

Therefore we conclude from (39) that when m is odd

(40)

$$\begin{aligned} P_m(z,q) = &-\Big[d_m(q)^2 + \Big(2d_m(q) - q^m(1-q)^m\Big)\frac{z}{1-z} \\ &+ \Big(2(2^m-1)d_m(q) + 1 - q^m(2-q)^m + q^m(1-q)^m \\ &-(1-q^2)^m\Big)\Big(\frac{z}{1-z}\Big)^2 + \cdots + c_n\Big(\frac{z}{1-z}\Big)^n\Big] \\ = &-\frac{1}{(1-z)^n}\Big[d_m(q)^2 - \Big(nd_m(q)^2 + q^m(1-q)^m - 2d_m(q)\Big)z \\ &+ \Big(\frac{n(n-1)}{2}d_m(q)^2 + (n-1)\big(q^m(1-q)^m - 2d_m(q)\big) + 2(2^m-1)d_m(q) \\ &+ 1 - q^m(2-q)^m + q^m(1-q)^m - (1-q^2)^m\Big)z^2 + \cdots\Big]. \end{aligned}$$

It is clear that $n \le 2m+1$ since the degree of g_j is $j+1$ and at least the terms of degree $2m+2$ cancel. By (38) we know that if there is no zero of $P_m(z)$ inside the unit circle, then all the zeros are on the unit circle. If $p(z)$ is a polynomial of degree n with real coefficients such that all zeros of p are on the unit circle, then $p(z) = a_0 + a_1 z + a_2 z^2 + \cdots$ where $|a_1/a_0| \le n$ and $|a_2/a_0| \le n(n-1)/2$. Applying these observations to $P_m(z,q)$ we conclude from (40) that unless it has a zero inside the unit circle, we must have

$$\begin{aligned} &n + \frac{q^m(1-q)^m}{d_m(q)^2} - \frac{2}{d_m(q)} \le n, \\ &\frac{n(n-1)}{2} + (n-1)\frac{q^m(1-q)^m}{d_m(q)^2} - (n-1)\frac{2}{d_m(q)} \\ &\quad + \frac{2(2^m-1)}{d_m(q)} - \frac{q^m(2-q)^m - q^m(1-q)^m + (1-q^2)^m - 1}{d_m(q)^2} \le \frac{n(n-1)}{2}. \end{aligned}$$

If $m = 4k-1$, then it follows from Lemma 5 that $d_m(q) < 0$ and the first inequality above gives a contradiction. Hence we assume that $m = 4k+1$ so that $d_m(q) > 0$. The second inequality above implies that

$$d_m(q) \le \frac{q^m(2-q)^m - nq^m(1-q)^m + (1-q^2)^m - 1}{2(2^m - n)}. \tag{41}$$

A straightforward calculation shows that $q^3(2-q)^3 + (1-q^2)^3 < 1$ when $q \in (0,1)$, and since $q(2-q) \leq 1$ and $1-q^2 \leq 1$ it follows that $q^m(2-q)^m + (1-q^2)^m < 1$ for $m \geq 3$. But then it follows from (41) that $d_m(q) < 0$ and we have a contradiction. This completes the proof of the fact that, if $m > 1$ is odd, then there exists a point z_1 with $|z_1| < 1$ such that $P_m(z_1, q) = 0$.

Now we know that if there exists a periodic solution, then there exists a complex number z_0 with $|z_0| < 1$ such that $f(z_0) = 0$. To complete the proof we use the same argument as in [7] to show that the problem can be reduced to an equation of the type (4) with $f(z)/(1-z) = \sum_{n=0}^{\infty} z^n \alpha_n$, and then we invoke Proposition 4. Note that what is needed for the reduction to (4) is not necessarily that $y_*'(0-) < 0$ and $y_*'(T-) > 0$ but that u_* is strictly thermostat controlled by y_* and that each switching time depends continuously on the data. If this last condition is not satisfied, then the periodic solution is clearly not locally stable.

4. Proof of Proposition 4. We use the following notation: if $n \in \mathbf{Z}$ and if $\{\zeta_j\}$ is a sequence of elements in $\mathbf{R}^m$ defined at least for all indices $j \leq n$, then the sequence $H_n\zeta : \mathbf{N} \to \mathbf{R}^m$ is defined by $(H_n\zeta)_j = \zeta_{n-j}$, i.e., $H_n\zeta$ consists of the part of ζ "before" n.

Let us first assume that $\det[\alpha_0] \neq 0$. The resolvent kernel ρ associated with α is defined to be the solution of the equation

$$\sum_{j=0}^{n} \alpha_{n-j}\rho_j = \sum_{j=0}^{n} \rho_{n-j}\alpha_j = \delta_{0n}, \qquad n \geq 0.$$

(Here $\delta_{ij} = I$ if $i = j$ and zero otherwise.)

Since there exists a complex number z_0 such that $|z_0| < 1$, $\det[\sum_{n=0}^{\infty} z_0^n \alpha_n] = 0$, and $\det[\alpha_0] \neq 0$, we may assume that there exists real numbers $0 < r_0 < r_1 < 1$ so that

$$\left[\sum_{n=0}^{\infty} z^n \alpha_n\right]^{-1} = \sum_{l=1}^{L} \sum_{k=0}^{p_l-1} \left(\frac{1}{2} \frac{1}{(z_l - z)^{k+1}} A_{l,k} + \frac{1}{2} \frac{1}{(\overline{z_l} - z)^{k+1}} \overline{A_{l,k}} \right) + F(z), \tag{42}$$

where $L \geq 1$, $p_l \geq 1$, $|z_l| = r_0$, $l = 1, \cdots, L$, and F is (a matrix of functions) analytic in $\{ z \in \mathbf{C} \mid |z| \leq r_1 \}$.

It follows from (42) that the resolvent ρ can be expressed in the form

$$\rho_n = u_n + \varrho_n, \qquad n \geq 0, \tag{43}$$

where

$$u_n = \Re \left[\sum_{l=1}^{L} \sum_{k=0}^{p_l-1} \binom{n+k}{k} z_l^{-(n+k+1)} A_{l,k} \right]$$

and

$$\sum_{n=0}^{\infty} r_1^n |\varrho_n| \leq c_\varrho < \infty.$$

(Observe that here $\Re M$ is the matrix with each element equal to the real part of the corresponding element in M.)

Let $P = \max\{ p_l \mid l = 1, \cdots, L \}$ and define

$$b_n = \sum_{k=0}^{P-1} \left| \binom{k-n}{k} \right| r_0^{-(k+1)} \left(\frac{r_0}{r_1} \right)^n, \qquad n \geq 0.$$

We choose numbers B_∞ and B_1 such that

$$B_\infty \geq r_0^P \sup_{n\geq 0} b_n, \qquad B_1 \geq \sum_{n=1}^{\infty} b_n.$$

We shall later use the fact that B_∞ and B_1 can be chosen independently of r_0 when r_0 is close to zero.

We take the constant c_A to be such that

$$|A_{l,k}| \leq c_A.$$

We may, without loss of generality, assume that $p_1 = P$. It is easy to see that there exists a vector $v \in \mathbf{R}^m$ with $|v| = 1$ and a number $\gamma > 0$ such that

$$\left| \Re \left[\sum_{k=0}^{P-1} z_1^{-(k+1)} A_{1,k} v \right] \right| \geq r_0^{-P} \gamma B_\infty c_A. \tag{44}$$

Let $\Gamma = (2+\gamma)/\gamma$ and let ϵ be a positive number such that if $\|\zeta\|_{l^\infty(\mathbf{N})} \leq \Gamma\epsilon$, then

$$\|\Phi(\zeta)\|_{l^\infty(\mathbf{N})} \leq \|\zeta\|_{l^\infty(\mathbf{N})} \frac{1}{2\Gamma} \min\left\{ \frac{1}{c_\varrho}, \frac{1}{Lc_A B_1} \right\}. \tag{45}$$

If the conclusion of Proposition 4 does not hold, then there exists a number δ such that (5) implies (6). Pick a positive integer N such that

$$r_1^{N+1} \leq \frac{\delta}{\Gamma\epsilon}. \tag{46}$$

Define the set Ω by

$$\begin{aligned} \Omega = &\left\{ \{\vartheta_n\}_{n=0}^N \;\middle|\; \vartheta_n \in \mathbf{R}^m, \quad |\vartheta_n| \leq \Gamma\epsilon r_1^{N-n}, \quad 0 \leq n \leq N \right\} \\ &\times \left\{ \{\theta_n\}_{n=0}^N \;\middle|\; \theta_n \in \mathbf{R}^m, \quad |\theta_n| \leq \frac{\epsilon r_1^{N-n}}{Lc_A B_1}, \quad 0 \leq n \leq N \right\}. \end{aligned}$$

Next we consider a mapping $(\vartheta, \theta) \in \Omega \mapsto (G(\vartheta,\theta), g(\vartheta,\theta))$ defined as follows.

$$\begin{aligned} G_n(\vartheta,\theta) =& \frac{2\epsilon r_0^P}{\gamma c_A B_\infty} \Re \left[\sum_{k=0}^{P-1} \binom{n-N+k}{k} z_1^{N-n-k-1} A_{1,k} v \right] \\ &- \sum_{j=n+1}^{N} u_{n-j} \Phi_j(H_j\xi) + \sum_{j=0}^{n} \varrho_{n-j} \Phi_j(H_j\xi), \qquad 0 \leq n \leq N; \\ g_n(\vartheta,\theta) =& \Phi_n(H_n\xi), \qquad 0 \leq n \leq N, \end{aligned} \tag{47}$$

where

$$\xi_n = \begin{cases} \dfrac{2\epsilon r_0^P}{\gamma C_a B_\infty} \Re \left[\displaystyle\sum_{k=0}^{P-1} \binom{n-N+k}{k} z_1^{N-n-k-1} A_{1,k} v \right] - \displaystyle\sum_{j=0}^{N} u_{n-j}\theta_j, & n < 0; \\ \vartheta_n, & 0 \leq n \leq N. \end{cases} \tag{48}$$

It is obvious that the mapping $(\vartheta,\theta)\mapsto(G(\vartheta,\theta),g(\vartheta,\theta))$ is continuous (in the topology of $\mathbf{R}^{m\times(2N+2)}$), and we must show that it maps Ω into itself.

We clearly have for all $n\le N$

$$\left|\frac{2\epsilon r_0^P}{\gamma c_A B_\infty}\Re\left[\sum_{k=0}^{P-1}\binom{n-N+k}{k}z_1^{N-n-k-1}A_{1,k}v\right]\right|\le\frac{2\epsilon c_A r_0^P b_{N-n}r_1^{N-n}}{\gamma c_A B_\infty}\le\frac{2\epsilon r_1^{N-n}}{\gamma}.$$

Next we observe that, if $(\vartheta,\theta)\in\Omega$, then it follows that

$$|\xi_i|\le\frac{2\epsilon r_1^{N-i}}{\gamma}+\sum_{j=0}^{N}Lc_A r_1^{N-i}b_{j-i}\frac{\epsilon}{Lc_A B_1}\le\Gamma\epsilon r_1^{N-i},\qquad i<0.$$

Because $r_1<1$ and $N-i\ge N+1\ge N-n$ when $i<0\le n$, we therefore see that (46) and the definitions of ξ and Ω imply that

$$\sup_{n<0}|\xi_n|\le\delta\quad\text{and}\quad\|H_n\xi\|_{l^\infty(\mathbf{N})}\le\Gamma\epsilon r_1^{N-n},\quad 0\le n\le N. \tag{49}$$

An easy calculation shows that

$$\begin{aligned}\left|\sum_{j=n+1}^{N}u_{n-j}\Phi_j(H_j\xi)\right|&\le\sum_{j=n+1}^{N}Lc_A\sum_{k=0}^{P-1}\left|\binom{n-j+k}{k}\right|r_0^{j-n-k-1}\frac{\Gamma\epsilon r_1^{N-j}}{2\Gamma Lc_A B_1}\\&=\frac{\epsilon r_1^{N-n}}{2B_1}\sum_{j=n+1}^{N}b_{j-n}\le\frac{\epsilon r_1^{N-n}}{2}.\end{aligned} \tag{50}$$

In the same way we get

$$\left|\sum_{j=0}^{n}\varrho_{n-j}\Phi_j(H_j\xi)\right|\le\sum_{j=0}^{n}|\varrho_{n-j}|\Gamma\epsilon r_1^{N-j}\frac{1}{2\Gamma c_\varrho}\le\frac{\epsilon r_1^{N-n}}{2}. \tag{51}$$

It is immediately clear from (45) and (49) that

$$|g_n(\vartheta,\theta)|\le\frac{\epsilon r_1^{N-n}}{2Lc_A B_1},\qquad 0\le n\le N.$$

This inequality combined with (50) and (51) shows that Ω is mapped into itself.

Since Ω is convex, we can apply Schauder's fixed point theorem and find a point $(\vartheta^*,\theta^*)\in\Omega$ such that $G(\vartheta^*,\theta^*)=\vartheta^*$ and $g(\vartheta^*,\theta^*)=\theta^*$. Define ξ^* by replacing (ϑ,θ) in (48) by (ϑ^*,θ^*) and let

$$\psi_n=\frac{2\epsilon r_0^P}{\gamma C_a B_\infty}\Re\left[\sum_{k=0}^{P-1}\binom{n-N+k}{k}z_1^{N-n-k-1}A_{1,k}v\right]-\sum_{j=0}^{N}u_{n-j}\theta_j^*,\quad n\le N.$$

Since $\xi_n^*=G_n(\vartheta^*,\theta^*)$ and $\theta_n^*=\Phi_n(H_n(\xi^*))$ for $0\le n\le N$, we get

$$\xi_n^*=\psi_n+\sum_{j=0}^{n}(u_{n-j}+\varrho_{n-j})\Phi_j(H_j\xi^*),\qquad 0\le n\le N.$$

Let $\hat{\alpha}(z) = \sum_{n=0}^{\infty} z^n \alpha_n$. For every $l = 1, 2, \cdots, L$ we have

$$\begin{aligned} &\lim_{z \to z_l} \frac{\mathrm{d}^{p_l - 1}}{\mathrm{d}z^{p_l - 1}} \frac{(-1)^{p_l - 1}}{(p_l - 1)!} \sum_{k=0}^{p_l - 1} \hat{\alpha}(z) z^{-n-1} (z_l - z)^{p_l - k - 1} A_{l,k} \\ &= \lim_{z \to z_l} \sum_{k=0}^{p_l - 1} \frac{(-1)^k}{k!} \frac{\mathrm{d}^k}{\mathrm{d}z^k} \sum_{j=-\infty}^{n} \alpha_{n-j} z^{-j-1} A_{l,k} \\ &= \sum_{j=-\infty}^{n} \alpha_{n-j} \sum_{k=0}^{p_l - 1} \binom{j+k}{k} z_l^{-(j+k+1)} A_{l,k}. \end{aligned} \tag{52}$$

On the other hand, it follows from (42) that there exists a function G, analytic at z_l such that

$$\sum_{k=0}^{p_l - 1} \hat{\alpha}(z) z^{-n-1} (z_l - z)^{p_l - k - 1} A_{l,k} = (z_l - z)^{p_l} G(z).$$

We combine this result with (52) and note that the argument can be repeated with z_l and $A_{l,k}$ replaced by $\overline{z_l}$ and $\overline{A_{l,k}}$, respectively. This shows that for every $l = 1, 2, \cdots, L$ we have

$$\sum_{j=-\infty}^{n} \alpha_{n-j} \Re \left[\sum_{k=0}^{p_l - 1} \binom{j+k}{k} z_l^{-(j+k+1)} A_{l,k} \right] = 0, \qquad n \in \mathbf{Z}.$$

It follows that $\sum_{j=-\infty}^{n} \alpha_{n-j} \psi_j = 0$, and therefore we conclude from (43) that ξ^* satisfies the equation

$$\begin{aligned} &\sum_{j=-\infty}^{n} \alpha_{n-j} \xi_j = \Phi_n(H_n \xi), \qquad 0 \le n \le N, \\ &\xi_i = \psi_i, \qquad i < 0. \end{aligned}$$

In order to get the desired contradiction we note that by (44), (47), and (51),

$$|\xi_N^*| = |G_N(\vartheta^*, \theta^*)| \ge \frac{2\epsilon r_0^P r_0^{-P} \gamma c_A B_\infty}{\gamma c_A B_\infty} - \frac{\epsilon}{2} = \frac{3\epsilon}{2}.$$

This completes the proof in the case where $r_0 > 0$.

Finally we have to consider the case where $r_0 = 0$ (and therefore $L = 1$). We construct a continuous function $t \in [0,1] \mapsto \alpha(t) \in l^1(\mathbf{N}; \mathbf{R}^{m \times m})$ such that $\alpha(0) = \alpha$ and so that

$$\left[\sum_{n=0}^{\infty} z^n \alpha_n(t) \right]^{-1} = \sum_{k=0}^{P-1} \left(\frac{1}{2} \frac{1}{(z(t) - z)^{k+1}} A_k(t) + \frac{1}{2} \frac{1}{(\overline{z(t)} - z)^{k+1}} \overline{A_k(t)} \right) + F(z,t),$$

where $z \mapsto F(z,t)$ is analytic for $|z| \le r_1(t)$ and $z(t) \ne 0$ for $t > 0$. Moreover, we require that $z(t)$, $r_1(t)$, $A_k(t)$, and $F(z,t)$ are continuous functions of t. As a consequence we can in the above take the constants c_ϱ, c_A, γ, B_1, and B_∞ to be independent of t. These constants determine the number ϵ that we choose. We assume that $\delta > 0$ is such that (5) implies (6). Now we take a number $t > 0$ to be so small that

$$\sum_{n=0}^{\infty} |\alpha_n(t) - \alpha_n| < \frac{\delta}{\Gamma \epsilon}. \tag{53}$$

By the argument above, there exists a sequence $\{\psi_i\}_{i=-\infty}^{-1}$ with $\sup_{i\le -1}|\psi_i| \le \delta$ and a solution ξ of the equation

$$\sum_{j=-\infty}^{n} \alpha_{n-j}(t)\xi_j = \Phi(H_n\xi), \qquad 0 \le n \le N,$$
$$\xi_i = \psi_i, \qquad i < 0,$$

such that $|\xi_N| > \epsilon$. But then ξ is also a solution of the equation

$$\sum_{j=-\infty}^{n} \alpha_{n-j}\xi_j = \Phi(H_n\xi) + \varphi_n, \qquad 0 \le n \le N,$$
$$\xi_i = \psi_i, \qquad i < 0,$$

where $\varphi_n = \sum_{j=-\infty}^{n}(\alpha_{n-j} - \alpha_{n-j}(t))\xi_j$. Since $|\varphi_n| \le \delta$ by (53) and the definition of the set Ω, we have a contradiction. This completes the proof of Proposition 4.

Acknowledgment. The author thanks the referees for pointing out some errors in an earlier version of the paper.

REFERENCES

[1] H.W. ALT, *On the thermostat problem,* Control Cybernet., 14 (1985), pp. 171–193.
[2] M. BROKATE AND A. FRIEDMAN, *Optimal design for heat conduction problems with hysteresis,* SIAM J. Control Optim., 27 (1989), pp. 697–717.
[3] A. FRIEDMAN AND L.-S. JIANG, *Periodic solutions for a thermostat control problem,* Comm. Partial Differential Equations, 13 (1988), pp. 515–550.
[4] K. GLASHOFF AND J. SPREKELS, *An application of Glicksberg's theorem to set-valued integral equations arising in the theory of thermostats,* SIAM J. Math. Anal., 12 (1981), pp. 477–486.
[5] ———, *The regulation of temperature by thermostats and set-valued integral equations,* J. Integral Equations, 4 (1982), pp. 95–112.
[6] G. GRIPENBERG, *On periodic solutions of a thermostat equation,* SIAM J. Math. Anal., 18 (1987), pp. 694–702.
[7] ———, *Stability of periodic solutions in thermostat control,* SIAM J. Math. Anal., 20 (1989), pp. 1459–1471.
[8] N. KENMOCHI AND M. PRIMICERIO, *One-dimensional heat conduction with a class of automatic heat-source controls,* IMA J. Appl. Math., 40 (1988), pp. 205–216.
[9] J. PRÜSS, *Periodic solutions of the thermostat problem,* in Proc. Conference on "Differential Equations in Banach Spaces," Bologna, July 1985, Lecture Notes in Math. 1223, Springer-Verlag, Berlin, New York, 1987, pp. 216–226.
[10] T. I. SEIDMAN, *Optimal control for switching systems,* in Proc. Conference on Information Science and Systems, Johns Hopkins University, Baltimore, MD, 1987, pp. 485–489.
[11] ———, *Switching systems,* I, Math. Res. Report 86-78, University of Maryland, Baltimore, MD, 1986.
[12] ———, *Switching systems and periodicity,* Math. Res. Report 88-02, University of Maryland, Baltimore, MD, 1988.
[13] Q. ZHENG, *Über die Existenz von unendlich vielen periodischen Lösungen einer einfachen Temperaturregelung,* Numer. Math., 55 (1989), pp. 327–341.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1349-1353, September 1991

009

EXCHANGE OF STABILITIES FOR FLOW ALONG A CONCAVE WALL*

ISOM H. HERRON†

Abstract. The equations of hydrodynamical stability derived by Görtler for flow along a concave wall are proved to satisfy the principle of exchange of stabilities (PES). The case where a constant suction normal to the wall is present is also treated in the same formulation.

Key words. Görtler flow, suction, linear operators

AMS(MOS) subject classifications. 76E05, 47E05

1. Introduction. The first study of the instability of the boundary layer due to the curvature of the flow along a concave wall is attributed to Görtler [4]. Interest in this type of instability has only increased with time. The purpose here is to prove exchange of stabilities in the following sense. *Principle of exchange of stabilities* (PES): *The first unstable eigenvalue has imaginary part equal to zero.* By means of a technique due to Weinberger [5], our previous work on this problem [2] proved exchange of stabilities for flow near a free surface with small curvature, ignoring the effects of surface tension. As a sequel, the wall bounded case is now solved. Since most of the details are developed in the previous paper [2], this work will simply show how the wall bounded case may be transformed to the case already treated.

The governing disturbance equations are

$$(D^2 - v_0 D - a^2 - \sigma)(D^2 - a^2)v + a^2 \mu U u = 0, \tag{1.1a}$$

$$(D^2 - v_0 D - a^2 - \sigma)u - U'v = 0 \tag{1.1b}$$

for the components $u(\eta)$ along the wall and $v(\eta)$ perpendicular to the wall, with boundary conditions

$$v(0) = v'(0) = u(0) = 0, \tag{1.1c}$$

and decay at infinity

$$v, v', u \to 0 \quad \text{as } \eta \to \infty. \tag{1.1d}$$

This is the notation of Drazin and Reid [1], where $D = d/d\eta$, $U(\eta)$ is the velocity component of the basic flow along the wall, a is the wave number, μ is the small gap Taylor number (linearly proportional to curvature), σ is the eigenvalue, and $v_0 = 0$ (no suction) or $v_0 = -1$ (suction). The two cases $v_0 = 0$ and $v_0 = -1$ are handled in one formulation. As in the previous work [2], it is assumed that $U \geqq 0$ and $U' \geqq 0$ for $0 \leqq \eta < \infty$.

2. Abstract formulation and PES. In operator form the differential equations may be written

$$(S^* + \sigma)Mv + a^2 \mu U u = 0, \tag{2.1a}$$

$$U'v + (\tilde{S} + \sigma)u = 0, \tag{2.1b}$$

* Received by the editors July 2, 1990; accepted for publication October 30, 1990. This work was supported by the Office of Naval Research.

† Department of Mathematics, Howard University, Washington, DC 20059.

where

$$Mv = mv = (-D^2 + a^2)v, \qquad v \in \text{dmn } M,$$
$$(S^* + \sigma)Mv = (-D^2 + v_0 D + a^2 + \sigma)Mv, \qquad v \in \text{dmn } (S^* M),$$
$$(\tilde{S} + \sigma)u = (-D^2 + v_0 D + a^2 + \sigma)u, \qquad u \in \text{dmn } \tilde{S}.$$

Next define the operator M^* by

$$M^* v = mv = (-D^2 + a^2)v, \qquad v \in \text{dmn } M^*,$$

where

$$\text{dmn } M^* = \{v \in L_2[0, \infty) \mid v, v' \text{ abs. cont.}, mv \in L_2[0, \infty)\}. \tag{2.2a}$$

Thus M^* has no boundary conditions; it is the adjoint of M given above and

$$\text{dmn } M = \{v \in \text{dmn } M^* \mid v(0) = v'(0) = 0\}. \tag{2.2b}$$

Consequently, M is symmetric and positive definite, but not maximal and not invertible. Another operator needed in what follows is $\tilde{M}$, a positive-definite self-adjoint extension of M: $\tilde{M}v = mv$, $v \in \text{dmn } \tilde{M}$, where

$$\text{dmn } \tilde{M} = \{v \in \text{dmn } M^* \mid v(0) = 0\}. \tag{2.3}$$

A *generalized inverse* [3] to M is $M^\dagger$. The null space of M^*, nul M^* is spanned by $e^{-a\eta}$. The projection operator Q onto nul M^* is defined by

$$(Q\varphi)(\eta) = \int_0^\infty g_Q(\eta, \xi)\varphi(\xi)\, d\xi, \tag{2.4a}$$

where

$$g_Q(\eta, \xi) = e^{-a\eta}\left[\int_0^\infty e^{-2as}\, ds\right]^{-1} e^{-a\xi} = 2a\, e^{-a(\eta+\xi)}. \tag{2.4b}$$

Then $g^\dagger$, the kernel of the generalized inverse, satisfies

$$\left(-\frac{\partial^2}{\partial \eta^2} + a^2\right) g^\dagger(\eta, \xi) = \delta(\eta - \xi) - g_Q(\eta, \xi), \tag{2.5a}$$

$$g^\dagger(0, \xi) = \frac{\partial g^\dagger}{\partial \eta}(0, \xi) = 0, \quad g^\dagger, \frac{\partial g^\dagger}{\partial \eta} \to 0 \quad \text{as } \eta \to \infty, \tag{2.5b}$$

so that

$$(M^\dagger \varphi)(\eta) = \int_0^\infty g^\dagger(\eta, \xi)\varphi(\xi)\, d\xi. \tag{2.6}$$

Some properties of $M^\dagger$ are

$$MM^\dagger = I - Q, \tag{2.7a}$$

$$M^\dagger M = I, \tag{2.7b}$$

since nul M is trivial.

When (2.1a, b) are considered in more detail, S^* has no boundary conditions, while $\tilde{S}$ has boundary conditions and is maximal. The space of functions on which the operators act is weighted (if $v_0 \neq 0$) defined as

$$H_{v_0} = \left\{\varphi \;\middle|\; \int_0^\infty e^{-v_0\eta}|\varphi|^2\, d\eta < \infty\right\}, \tag{2.8a}$$

with inner product

$$\langle \varphi, \psi \rangle_{v_0} = \int_0^\infty \varphi(\eta)\bar{\psi}(\eta)\, e^{-v_0\eta}\, d\eta, \qquad \varphi, \psi \in H_{v_0}, \tag{2.8b}$$

and norm

$$\|\varphi\|_{v_0} = (\langle \varphi, \varphi \rangle_{v_0})^{1/2}. \tag{2.8c}$$

The projection Q onto the null space of M^* is employed. Define $\zeta = Mv \Rightarrow v = M^\dagger \zeta$, and let

$$Z(y) = \zeta(\eta) - T\zeta(\eta) = (I - T)\zeta, \tag{2.9a}$$

such that $Z \in \operatorname{dmn} \tilde{M}$. This is done by setting

$$T\zeta(\eta) = \zeta(0)\, e^{-a\eta}, \tag{2.9b}$$

which is an element of the null space of M^*. Thus, $I - T$ maps the range of M onto the domain of $\tilde{M}$. Then

$$QZ = Q\zeta - QT\zeta = -T\zeta. \tag{2.10}$$

From (2.7) we have

$$M^\dagger(Z - QZ) = M^\dagger Z, \tag{2.11}$$

where Q is given by (2.4). Then (2.1) may be written in terms of Z as

$$(\tilde{S} + \sigma)Z + a^2\mu Uu = (S^* + \sigma)QZ, \tag{2.12a}$$

$$U'M^\dagger Z + (\tilde{S} + \sigma)u = 0. \tag{2.12b}$$

It is not difficult to show that $\tilde{S}$ is maximal positive definite; $u \in \operatorname{dmn} \tilde{S}$, $\langle \tilde{S}u, u \rangle_{v_0} \geqq (v_0^2/4 + a^2)\langle u, u \rangle_{v_0}$. Thus $(\tilde{S} + \sigma)^{-1}$ exists for $\sigma \notin \Sigma = \{\sigma \in \mathbb{C} \,|\, \operatorname{Re}(\sigma) \leqq -(v_0^2/4 + a^2), \operatorname{Im}(\sigma) = 0\}$. The operation of $(\tilde{S} + \sigma)^{-1}$ is [2]

$$(\tilde{S} + \sigma)^{-1}\varphi = \int_0^\infty h(\eta, \xi; \sigma)\varphi(\xi)\, d\xi, \tag{2.13a}$$

where

$$h(\eta, \xi; \sigma) = \left[\frac{e^{-p|\eta - \xi|} - e^{-p(\eta + \xi)}}{2p}\right] e^{v_0(\eta - \xi)/2}, \tag{2.13b}$$

and

$$p = \sqrt{a^2 + v_0^2/4 + \sigma} \quad \text{is the positive square root.}$$

Application of $(\tilde{S} + \sigma)^{-1}$ to (2.12a) gives

$$Z + a^2\mu B(\sigma)Uu = B(\sigma)(S^* + \sigma)QZ, \tag{2.14a}$$

where

$$B(\sigma) = (\tilde{S} + \sigma)^{-1}. \tag{2.14b}$$

Integration by parts on the right side of (2.14a) leads to

$$\begin{aligned} Z(\eta) + (a^2\mu B(\sigma)Uu)(\eta) = \Bigg[h(\eta, \xi; \sigma)\left(-\frac{\partial}{\partial \xi}(QZ)(\xi) + v_0(QZ)(\xi)\right) \\ + \frac{\partial}{\partial \xi} h(\eta, \xi; \sigma)(QZ)(\xi)\Bigg]_{\xi=0}^{\infty} + (QZ)(\eta). \end{aligned} \tag{2.15}$$

From (2.13b) we have

$$Z(\eta)+(a^2\mu BUu)(\eta)=-\frac{\partial}{\partial\xi}h(\eta,0;\sigma)(QZ)(0)+(QZ)(\eta). \tag{2.16}$$

It is possible to rewrite (2.16) as

$$Z+(a^2\mu BU)u=(I-A(\sigma))QZ, \tag{2.17}$$

where $A(\sigma)$ is the projection onto nul $(S^*+\sigma)$. With (2.12b), (2.17) becomes

$$Z-a^2\mu BUBU'M^\dagger Z=(I-A)QZ. \tag{2.18}$$

Application of Q to both sides of (2.18) gives

$$a^2\mu QBUBU'M^\dagger Z=QAQZ. \tag{2.19}$$

The left side of (2.19) depends explicitly (linearly) on μ; the right side does not. Let $\mu\to 0$. Then $QAQZ\to 0$, which is not unexpected.

The condition $QAQZ=0$, interpreted in the light of (2.13) and (2.16), means

$$(QZ)(0)\int_0^\infty g_Q(\eta,\gamma)\frac{\partial h}{\partial\xi}(\gamma,0;\sigma)\,d\gamma=0. \tag{2.20}$$

When this calculation is carried out the result is

$$\frac{2a\,e^{-a\eta}}{a+p-v_0/2}(QZ)(0)=0. \tag{2.21}$$

This can only hold if $(QZ)(0)=0$, which from (2.10) means $QZ\equiv 0$, and this implies $Z\equiv 0$.

Next, by considering the norms of both sides of (2.19), the relation $QAQZ=0$ must hold for $\mu\neq 0$ as well. To see this, first set

$$J\equiv QBUBU'M^\dagger \tag{2.22}$$

so that from (2.19)

$$\|QAQZ\|_{v_0}=a^2\mu\|JZ\|_{v_0}$$

and the ratio

$$\frac{\|QAQZ\|_{v_0}}{\|JZ\|_{v_0}}=a^2\mu\geqq\frac{\|QAQZ\|_{v_0}}{\|J\|_{v_0}\|Z\|_{v_0}}. \tag{2.23}$$

Thus, as $\mu\to 0$, the right side of (2.23) must approach zero as $\|Z\|_{v_0}\to 0$, and this requires $\|QAQZ\|_{v_0}\to 0$ faster than $\|Z\|_{v_0}\to 0$, which for a linear operator cannot be true, unless $QAQZ\equiv 0$.

The calculations which led to (2.20) and (2.21) can be carried out again, giving the same result: $QZ\equiv 0\Rightarrow Z\equiv 0$, so the system (2.12) is really homogeneous. Furthermore, $\tilde{M}M^\dagger Z=(I-Q)Z=Z$, so $M^\dagger=\tilde{M}^{-1}$ when operating on Z.

The system (2.12) may be written as a single equation:

$$(\tilde{S}+\sigma)Z-a^2\mu UB(\sigma)U'\tilde{M}^{-1}Z=0, \tag{2.24}$$

or

$$Z-K(\sigma)Z=0, \tag{2.25a}$$

where

$$K(\sigma)=a^2\mu B(\sigma)UB(\sigma)U'B(0). \tag{2.25b}$$

The formulation (2.25) is a reduction to the form considered in the previous article [2]. It is true that $K(\sigma)$ will be a *compact* operator if the decay of U' to zero as $\eta \to \infty$ is sufficiently strong. By the technique of Weinberger [5], PES thereby follows. Now our earlier discussion (between equations (2.12) and (2.13)) has shown that the original system (2.1) and the transformed system (2.24) have spectra that agree, except on a set Σ, which is a subset of the negative real halfline. Consequently, PES will hold for the original system as well.

REFERENCES

[1] P. G. DRAZIN AND W. H. REID, *Hydrodynamic Stability*, Cambridge University Press, Cambridge, 1981.

[2] I. H. HERRON, *Exchange of stabilities for Görtler flow*, SIAM J. Appl. Math., 45 (1985), pp. 775–779.

[3] W. S. LOUD, *Some examples of generalized Green's functions and generalized Green's matrices*, SIAM Rev., 12 (1970), pp. 194–210.

[4] H. SCHLICHTING, *Boundary Layer Theory*, Seventh ed., McGraw-Hill, New York, 1979.

[5] H. F. WEINBERGER, *Exchange of stability in Couette flow*, in Bifurcation Theory and Nonlinear Eigenvalue Problems, J. B. Keller and S. Antman, eds., Benjamin, New York, 1969.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1354-1371, September 1991

010

GENEALOGY AND BIFURCATION SKELETON FOR CYCLES OF THE ITERATED TWO-EXTREMUM MAP OF THE INTERVAL*

J. RINGLAND†‡ AND M. SCHELL†

Abstract. It is shown how the skeleton of the bifurcation structure of iterated maps of the interval with two extrema may be constructed without reference to kneading sequences. This construction of the skeleton is based on local rules and is accomplished by considering not just the bones, i.e., the curves in the parameter plane that correspond to the existence of superstable cycles, but also the curves that correspond to the existence of itineraries from one turning point to the other: curves called ligaments here. A graphical genealogy of the cycles ensues.

Key words. genealogy, periodic orbits, skeleton, superstability, bimodal map

AMS(MOS) subject classifications. 58C25, 58F14, 58F22, 34C35

1. Introduction. Analysis of the dynamics of iterated maps of the interval with two extrema has focused on the parameter-plane curves that correspond to the existence of superstable cycles.[1] The curves have been termed the *bones* [1] of the regions in the parameter-plane where cycles are stable. The way in which the bifurcation structure hangs on these bones has been described [1], [2]. A more difficult problem is determining the arrangement of the *skeleton*—the collection of all the bones. Mackay and Tresser [2] have shown how it may be solved by applying the kneading theory [2]–[6]. Here we present an alternative solution that is much simpler and, due to its radically different character, quite complementary to the previous work. In our approach, no reference is made to the lexicology of kneading sequences. Impossible cycles that in the kneading theory approach must be excluded case by case, using a rather laborious admissibility test, simply do not arise. More importantly, the result of our construction is a graphical genealogy of the cycles that embodies their intricately knitted relationships and provides a locally causal explanation of the existence and arrangement of the bones.

Our construction of the skeleton is accomplished by considering not only the bones, but also the parameter-plane curves that correspond to the existence of itineraries from one turning point to the other: curves we call *ligaments*. As we intend the name to suggest, the ligaments serve to connect the bones. We stress that while each ligament is in fact the locus of occurrence of specific (finite) kneading sequence, at no point do we make use of the kneading sequences themselves. And parenthetically we note that while bones correspond to attractors of the map, ligaments do not.

The construction is founded on the local analysis of three classes of singular point of the parameter plane from which bones and ligaments can be considered to emanate. The three corresponding kinds of singular[2] condition of the map, which for brevity we denote (arbitrarily) by the letters α, χ, and ψ, are exemplified in Fig. 1. These

* Received by the editors November 27, 1989; accepted for publication (in revised form) September 10, 1990. This research was supported by the Petroleum Research Fund, administered by the American Chemical Society.

† Department of Chemistry and Center for Nonequilibrium Structures, Southern Methodist University, Dallas, Texas 75275.

‡ Department of Mathematics, State University of New York at Buffalo, Buffalo, New York 14214.

[1] These are cycles that include a turning point of the map, and hence have eigenvalue zero.

[2] The term does not imply that the map lacks smoothness (see Fig. 1).

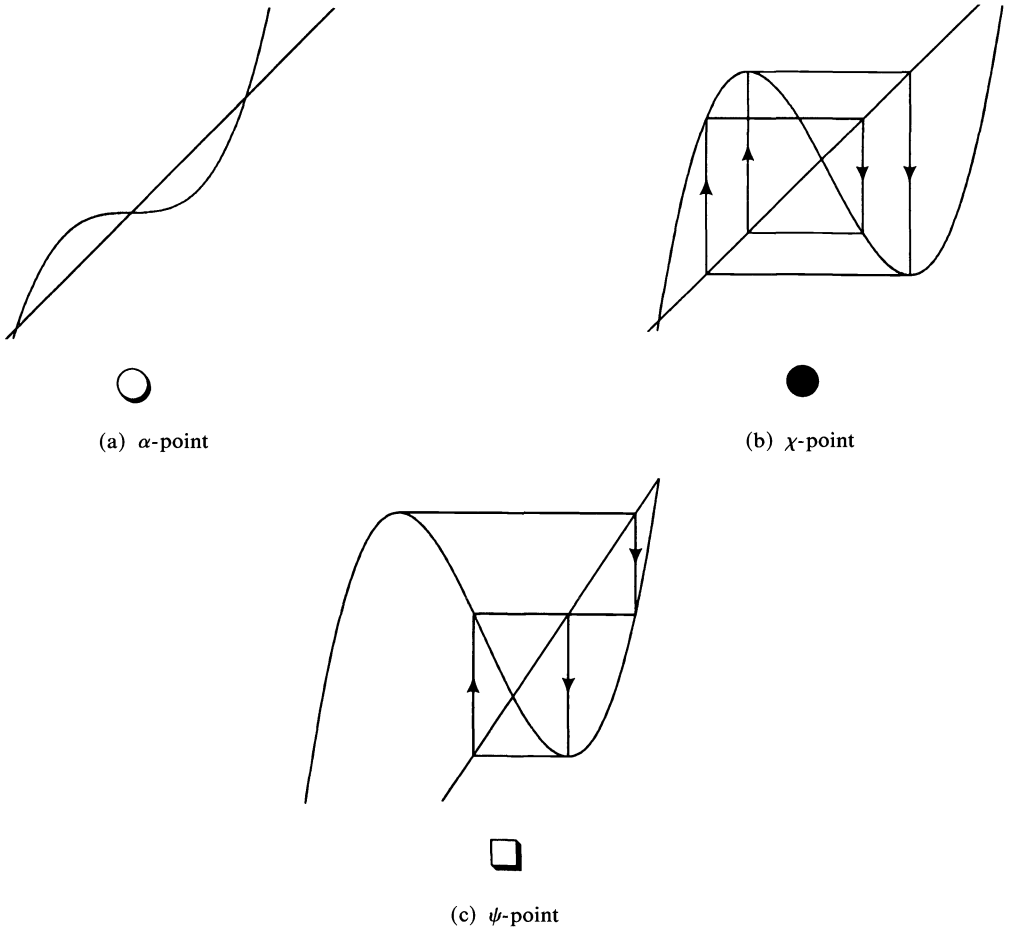


FIG. 1. *Maps (sketched) in the three kinds of condition that correspond to sources of bones and/or ligaments.* (a) *At an α-point, a point of degeneracy of two turning points.* (b) *At a χ-point, the location of a doubly superstable cycle, i.e., a cycle that contains both of the turning points of the map.* (c) *At a ψ-point, where an itinerary from one turning point to the other coexists with a cycle that contains the terminal turning point of that itinerary. The symbols appearing below the sketches are used in later figures to mark points in the parameter plane where the respective conditions arise.*

singular points will be discussed in detail in the next section, but first we mention briefly their attributes in order to sketch the logic of our approach. The α-point is a source of ligaments, the χ-point is a point of intersection of two ligaments and is the source of a pair of bones and of additional ligaments, and the ψ-point is a point of intersection of a ligament and a bone and is the source of additional ligaments. The idea of the construction is that the ligaments emanating from the α-point participate in intersections that constitute χ-points. The bones and ligaments that emanate from these χ-points participate in intersections that constitute additional χ-points and ψ-points. These points in turn are the sources of bones and ligaments which give rise to more χ-points and ψ-points, and so on. In this way the entire skeleton is generated, and in what follows we develop the specific prescription. Related methods have been applied in other dynamical systems contexts [7], [8].

In § 2 we describe the singular points, and derive the existence and arrangement of the bones and/or ligaments that emanate from them. In § 3 we investigate the

consequences of these results for map families that are *full* and *nice*: properties defined in that section that, in the spirit of the previous work [2], ensure, respectively, the completeness and maximal simplicity of the skeleton. Consideration of the possibility of other kinds of interaction of ligaments and bones then permits the completion of the set of rules for constructing the skeleton. We carry out the construction as far as to include all bones corresponding to cycles of length 6 or less.

2. Local analysis of the three points of origin of bones and/or ligaments. We consider analytic two-parameter families of maps $x \to \Phi(x, \lambda, \mu)$ of the interval into itself (coordinate x, parameters λ, μ), with $\partial\Phi/\partial x > 0$ except at two turning points (where $\partial\Phi/\partial x = 0$) and between them where $\partial\Phi/\partial x < 0$.

2.1. The α-point. In order to consider the starting point of the skeleton, we broaden slightly the class of maps considered and allow the turning points to merge.

DEFINITION. A point (λ_0, μ_0) of the parameter plane of the map $\Phi(x, \lambda, \mu)$ is an *α-point* if there exists x_0 such that $\Phi(x_0, \lambda_0, \mu_0) = x_0$, $\partial\Phi(x_0, \lambda_0, \mu_0)/\partial x = \partial^2\Phi(x_0, \lambda_0, \mu_0)/\partial x^2 = 0$, and $\partial^3\Phi(x_0, \lambda_0, \mu_0)/\partial x^3 > 0$.

The graph of a map at an α-point is pictured in Fig. 1(a).

THEOREM 1. *Generically, a sufficiently small parameter-plane circle centered at an α-point is intersected (transversally) by ligaments $\{\mathcal{L}_n\}$ corresponding to itineraries of lengths $n = 0, 1, 2, \cdots$ between turning points, and by a bone $\mathcal{B}$ corresponding to fixture of a turning point (equivalently, superstability of a fixed point), in the following (cyclic) order: $\mathcal{L}_0(L \to R)$, $\mathcal{B}(R)$, $\cdots$, $\mathcal{L}_3(L \to R)$, $\mathcal{L}_2(L \to R)$, $\mathcal{L}_1(L \to R)$, $\mathcal{L}_1(R \to L)$, $\mathcal{L}_2(R \to L)$, $\mathcal{L}_3(R \to L)$, $\cdots$, $\mathcal{B}(L)$, $\mathcal{L}_0(R \to L)$, where the arguments indicate the association with the left (L) and right (R) turning points. In other words, the generic unfolding of an α-point is qualitatively as depicted in Fig.* 2(a).

The proof follows a description of Fig. 2(a). The ligament corresponding to an inter-turning-point itinerary of length zero, i.e., the locus of emergence of the turning points, is labeled zero. The hatching indicates the side of this curve where the map is monotonic. On the other side is the bone (unbroken curve), on which exists a superstable fixed point of the map, and the bundle of ligaments, each labeled by the length of the corresponding itinerary. Ligaments corresponding to longer itineraries are closer to the bone. The asterisk in Fig. 2(a) corresponds to a situation like that depicted in Fig. 3(a).

In Fig. 2 (and Figs. 4 and 5) we use dashed and dotted curves to distinguish between the two *types* of ligament: those that correspond to itineraries from the left turning point to the right one, and those that correspond to itineraries from the right turning point to the left one.

The ligaments emanating from an α-point change type at the α-point: this fact is associated with the coincidence there of the two turning points. So on one "side" of the α-point the ligaments correspond to itineraries from the left turning point to the right one, and on the other side they correspond to itineraries from the right turning point to the left one. As discussed in detail in § 2.2, two ligaments of opposite type can intersect, giving rise to a χ-point: the location of a cycle that includes both turning points, and a source of bones and additional ligaments.

Proof of Theorem 1. Let us write the map as

$$\Phi(x, \lambda, \mu) = \Sigma a_{ijk} x^i \lambda^j \mu^k. \tag{2.1.1}$$

Here, as elsewhere below, the sum is on all indices over the nonnegative integers. An α-point at $x = \lambda = \mu = 0$ is obtained if we set $a_{000} = a_{100} = a_{200} = 0$. Generically, $a_{300} \neq 0$, and therefore $a_{300} > 0$ by the previously stated requirements on $\partial\Phi/\partial x$. Also generically

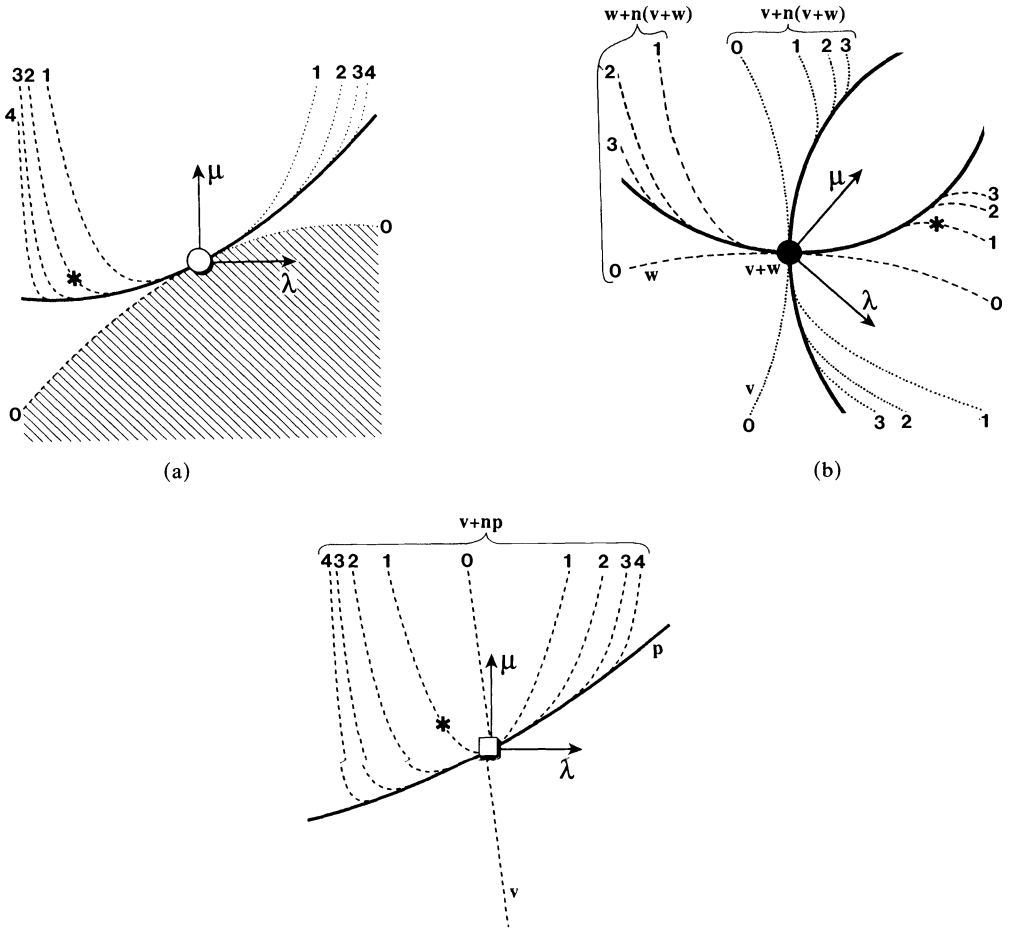


FIG. 2. *Sketches of the parameter-plane neighborhoods of each of the three kinds of condition depicted in Fig.* 1. *Unbroken curves represent bones (existence loci of superstable cycles). Broken curves represent ligaments (existence loci of itineraries from one turning point to the other); dashed and dotted are used to distinguish between the two types.* (a) *The α-point. Ligaments are labeled with the length of the corresponding itinerary. The ligament* 0 *is the locus of degeneracy of the two turning points; on the hatched side of this curve the map is monotonic. On the other side exist the (period-*1*) bone and bundle of ligaments which emanate from the α-point. The ligaments are ordered by length of corresponding itinerary, with the longer ones closer to the bone, as proved in* § 2.1. (b) *the χ-point. Two bones and two bundles of ligaments emanate from each χ-point. The ligaments are labeled with the value of n in the respective formulas for the length of the corresponding itinerary. That the local qualitative arrangement of the bones and ligaments is generically as pictured is proved in* § 2.2. (c) *The ψ-point, from which a bundle of ligaments emanates. The ligaments are labeled with the value of n in the formula v + np for the length of the corresponding itinerary. That this is the generic picture is proved in* § 2.3.

$a_{101} \neq 0$, so the implicit function theorem locally guarantees the existence of a unique analytic function

$$\mu^*(r, \lambda) = \Sigma b_{ij} r^i \lambda^j \tag{2.1.2}$$

($b_{00} = 0$) that satisfies

$$\frac{\partial \Phi}{\partial x}(r, \lambda, \mu^*(r, \lambda)) = 0, \tag{2.1.3}$$

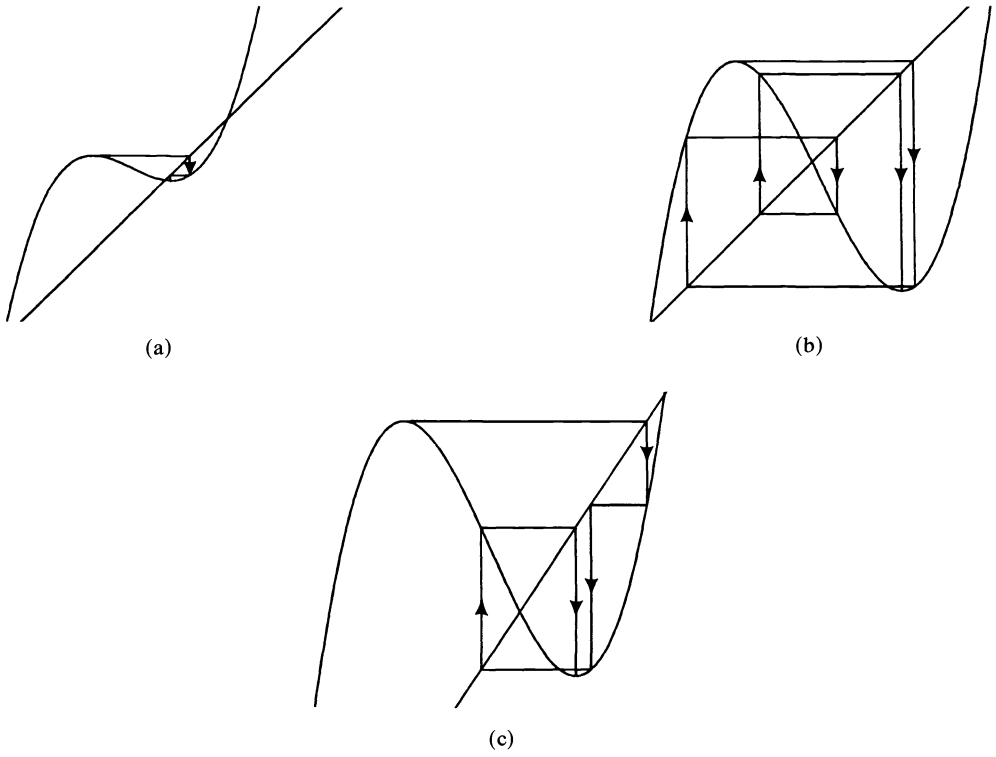


FIG. 3. *Sketches of maps near each of the three conditions depicted in Fig.* 1, *showing itineraries between turning points which exist on ligaments emanating from the singular points.* (a) *Near an* α*-point, an itinerary of length* 2 *from the left to the right turning point.* (b) *Near the* χ*-point of Fig.* 1(b), *an itinerary from the left to the right turning point of length* 5, *which at the* χ*-point is degenerate with itineraries of length* 1, 5, 9, 13, $\cdots$, *the increment being the period* (4) *of the doubly superstable cycle that exists at the* χ*-point.* (c) *Near the* ψ*-point of Fig.* 1(c), *an itinerary* (*length* 4) *that at the* ψ*-point consists of the direct itinerary* (*length* 2) *plus one loop around the superstable cycle* (*period* 2).

i.e., that makes r a turning point of Φ. Substitution of (2.1.1) and (2.1.2) into (2.1.3), and equating like terms yields $b_{10} = -2a_{200}/a_{101} = 0$.

There also exists another turning point given by an analytic function

$$s(r, \lambda) = \Sigma c_{ij} r^i \lambda^j \tag{2.1.4}$$

($c_{00} = 0$) satisfying

$$\frac{\partial \Phi}{\partial x}(s(r, \lambda), \lambda, \mu^*(r, \lambda)) = 0. \tag{2.1.5}$$

Since $a_{200} = 0$ and $a_{300} \neq 0$, $s(r, \lambda)$ as defined is double-valued. One value is simply the turning point r itself: $s(r, \lambda) = r$. We are interested in the other value, which by substitution of (2.1.1) and (2.1.4) into (2.1.5) has $c_{10} = -1$.

The implicit function theorem also guarantees the existence of a set of analytic functions $r_n(\lambda)$, $n = 0, 1, 2, \cdots$, with $r_n(0) = 0$, for all n, that satisfy

$$I_n(r_n(\lambda), \lambda) = 0, \tag{2.1.6}$$

where

$$I_n(r,\lambda)\equiv\Phi^n(r,\lambda,\mu^*(r,\lambda))-s(r,\lambda), \tag{2.1.7}$$

since $\partial I_0(0,0)/\partial r=1-c_{10}=2\ (\neq 0)$ and $\partial I_{n>0}(0,0)/\partial r=-c_{10}=1\ (\neq 0)$. (By Φ^n we mean the n-fold composition of Φ with itself.) Then the set of analytic functions $\{\mu_n(\lambda)\equiv\mu^*(r_n(\lambda),\lambda),\ n=0,1,2,\cdots\}$ represents a bundle of ligaments emanating from the α-point. We conclude that the itineraries of all lengths that degenerately exist at the α-point persist each on a unique curve in the unfolding.

Similarly, a bone is represented by the analytic function $\mu_{ss}(\lambda)\equiv\mu^*(r_{ss}(\lambda),\lambda)$ where

$$r_{ss}(\lambda)=\Sigma d_j\lambda^j \tag{2.1.8}$$

($d_0=0$) is defined by

$$I_{ss}(r_{ss}(\lambda),\lambda)=0, \tag{2.1.9}$$

where

$$I_{ss}(r,\lambda)\equiv\Phi(s(r,\lambda),\lambda,\mu^*(r,\lambda))-s(r,\lambda). \tag{2.1.10}$$

(Existence and uniqueness of $r_{ss}(\lambda)$ are guaranteed by $\partial I_{ss}(0,0)/\partial r=-c_{10}=1\neq 0$.) Note that $r_{ss}(\lambda)$ is the selection of the turning point such that the *other* turning point s is a fixed point of Φ.

To determine the *arrangement* of the ligaments and the bone we compute the deviations of the ligaments $\{\mu_n(\lambda)\}$ from the bone $\mu_{ss}(\lambda)$. That is, we compute the differences

$$\Delta\mu_n(\lambda)\equiv\mu^*(r_n(\lambda),\lambda)-\mu^*(r_{ss}(\lambda),\lambda). \tag{2.1.11}$$

If we define $\Delta r_n(\lambda)\equiv r_n(\lambda)-r_{ss}(\lambda)$, then

$$\Delta\mu_n(\lambda)=\mu^*(r_{ss}(\lambda)+\Delta r_n(\lambda),\lambda)-\mu^*(r_{ss}(\lambda),\lambda). \tag{2.1.12}$$

Substituting $r_{ss}(\lambda)+\Delta r_n(\lambda)$ for r in (2.1.2), using (2.1.8), and recalling that $b_{10}=d_0=0$, we obtain

$$\Delta\mu_n(\lambda)=s\lambda\Delta r_n(\lambda)+T\Delta r_n^2(\lambda)+\text{h.o.t.}, \tag{2.1.13}$$

where

$$S=2d_1b_{20}+b_{11},\qquad T=b_{20}, \tag{2.1.14}$$

and by h.o.t. we mean terms of higher order in λ than the leading term of the expressions to the left. Straightforward substitution of the series expansions of μ_{ss}, s, and r_{ss} into the definitions of these quantities ((2.1.2), (2.1.4), and (2.1.8), respectively), and equating like terms yields the following identities required for the evaluation of S and T:

$$\begin{aligned}
&d_1=c_{01}-b_{01}a_{001}-a_{010},\\
&c_{01}=(b_{11}a_{101}-2b_{01}a_{201}-2a_{210})/6a_{300},\\
&b_{01}=-a_{110}/a_{101},\\
&b_{11}=-2(b_{01}a_{201}+a_{210})/a_{101},\\
&b_{20}=-3a_{300}/a_{101}.
\end{aligned} \tag{2.1.15}$$

Now, from (2.1.6),

$$\begin{aligned}
0&=I_n(r_{ss}(\lambda)+\Delta r_n(\lambda),\lambda)\\
&=\Phi^n(r_{ss}(\lambda)+\Delta r_n(\lambda),\lambda,\mu(r_{ss}(\lambda)+\Delta r_n(\lambda),\lambda))-s(r_{ss}(\lambda)+\Delta r_n(\lambda),\lambda).
\end{aligned} \tag{2.1.16}$$

Substituting (2.1.8) into (2.1.4), and applying Φ, we get

$$0=\begin{cases} I_n(r_{ss}(\lambda),\lambda)+(1-c_{10})\Delta r_n(\lambda)+O(\Delta r_n(\lambda)^2), & n=0,\\ I_n(r_{ss}(\lambda),\lambda)+(-c_{10})\Delta r_n(\lambda)+O(\Delta r_n(\lambda)^2), & n>0. \end{cases} \tag{2.1.17}$$

Since $c_{10}=-1$, (2.1.17) implies that

$$\Delta r_n(\lambda)=\begin{cases} -\frac{1}{2}I_n(r_{ss}(\lambda),\lambda)+\text{h.o.t.}, & n=0,\\ -I_n(r_{ss}(\lambda),\lambda)+\text{h.o.t.}, & n>0. \end{cases} \tag{2.1.18}$$

By using (2.1.1), a recursion formula may be obtained for the $\{I_n(r_{ss}(\lambda),\lambda)\}$ as follows:

$$\begin{aligned} I_{n+1}(r_{ss}(\lambda),\lambda)&=\Phi(\Phi^n(r_{ss}(\lambda),\lambda,\mu(r_{ss}(\lambda),\lambda)),\lambda,\mu(r_{ss}(\lambda),\lambda))-s(r_{ss}(\lambda),\lambda),\\ &=\Phi(I_n(r_{ss}(\lambda),\lambda)+s(r_{ss}(\lambda),\lambda),\lambda,\mu(r_{ss}(\lambda),\lambda))-s(r_{ss}(\lambda),\lambda). \end{aligned} \tag{2.1.19}$$

Expanding, we then obtain

$$\begin{aligned} I_{n+1}(r_{ss}(\lambda),\lambda)&=\Phi(s(r_{ss}(\lambda),\lambda),\lambda,\mu(r_{ss}(\lambda),\lambda))-s(r_{ss}(\lambda),\lambda)\\ &\quad+Q\lambda I_n^2(r_{ss}(\lambda),\lambda)+RI_n^3(r_{ss}(\lambda),\lambda)+\text{h.o.t.}\\ &=Q\lambda I_n^2(r_{ss}(\lambda),\lambda)+RI_n^3(r_{ss}(\lambda),\lambda)+\text{h.o.t.}, \end{aligned} \tag{2.1.20}$$

where

$$Q=3a_{300}(c_{01}-d_1)+b_{01}a_{201}+a_{210},\qquad R=a_{300}. \tag{2.1.21}$$

(Recall that s is the location of a quadratic extremum of Φ except at $\lambda=\mu=0$ where Φ has a cubic inflection at s.)

The starting point for our recursive determination of the $\{\Delta\mu_n(\lambda)\}$ is, from (2.1.7) with $n=0$, (2.1.4) and (2.1.8),

$$I_0(r_{ss}(\lambda),\lambda)=U\lambda, \tag{2.1.22}$$

where

$$U=2d_1-c_{01}. \tag{2.1.23}$$

Combining the results (2.1.13), (2.1.18), (2.1.20), and (2.1.22), we have

$$\begin{aligned} &\Delta\mu_0(\lambda)=(-\tfrac{1}{2}SU+\tfrac{1}{4}TU^2)\lambda^2+\text{h.o.t.},\\ &\Delta\mu_1(\lambda)=-S(Q+RU)U^2\lambda^4+\text{h.o.t.},\\ &\Delta\mu_n(\lambda)=-SQQ^{(2^{n-1}-2)}(Q+RU)^{2^{n-1}}U^{2^n}\lambda^{2^{n+1}}+\text{h.o.t.},\qquad n>1. \end{aligned} \tag{2.1.24}$$

Thus the ligaments $\mu_n(\lambda)$ deviate from the bone $\mu_{ss}(\lambda)$ at orders λ^2, λ^4, λ^8, $\lambda^{16},\cdots$, for $n=0,1,2,3,\cdots$. The arrangement of the curves is ascertained by considering that for $n>1$, the sign of the leading order term of $\Delta\mu_n(\lambda)$ is always the same: the sign of $-SQ$. Moreover, the sign of $\Delta\mu_1(\lambda)$ is also the same, since the product of coefficients

$$[-S(Q+RU)][-SQ]=\frac{S^2V^2}{3a_{101}^2}, \tag{2.1.25}$$

where $V=3a_{300}(a_{101}a_{010}-a_{110}a_{001})+a_{210}a_{101}-a_{201}a_{110}$, is not less than zero. In contrast, $\mu_0(\lambda)$, the locus of turning point coalescence, deviates from the bone on the opposite side since the product

$$\left[-\frac{1}{2}SU+\frac{1}{4}TU^2\right][-SQ]=-\frac{2}{3}\frac{1}{a_{300}}\left[\frac{1}{a_{101}}\right]^6V^4 \tag{2.1.26}$$

is not *greater* than zero if $a_{300}>0$ as originally posited. Thus the qualitative arrangement of the ligaments and the bone is universal, and is as specified in Theorem 1. □

2.2. The χ-point. An intersection of two ligaments of opposite type, that is to say, of a ligament corresponding to an itinerary from the left turning point of the map to the right one and a ligament corresponding to an itinerary from the right turning point of the map to the left one, is a point of existence of a doubly superstable cycle (a cycle that includes both turning points). We show below that such a point is the source of a pair of bones, and of two bundles of additional ligaments. For the purposes of our construction of the skeleton we regard the intersection mentioned above as "causative" of the doubly superstable point and of the bones and other ligaments which emanate from it. But for purposes of analysis it is more convenient to define the point as follows.

DEFINITION. A point (λ_0, μ_0) of the parameter plane of the map $\Phi(x, \lambda, \mu)$ is a *χ-point* if there exist v, $w \in \mathbb{N}$ such that $\Phi^v(r_0, \lambda_0, \mu_0) = s_0$, and $\Phi^w(s_0, \lambda_0, \mu_0) = r_0$, where r_0, s_0 are (distinct) quadratic turning points of $\Phi(x, \lambda_0, \mu_0)$.

The existence of the two "causative" ligaments, as well as of the other ligaments and the bones, follows from the definition as we now demonstrate.

THEOREM 2. *Let (λ_0, μ_0) be a χ-point of $\Phi(x, \lambda_0, \mu_0)$ for which the turning points are r_0 and s_0, and let $v = \min(v' \in \mathbb{N} | \Phi^{v'}(r_0, \lambda_0, \mu_0) = s_0)$ and $w = \min(w' \in \mathbb{N} | \Phi^{w'}(s_0, \lambda_0, \mu_0) = r_0)$. Generically, a sufficiently small parameter-plane circle centered at (λ_0, μ_0) is intersected (transversally) by ligaments $\{\mathcal{L}_n^{r\to s}\}$ and $\{\mathcal{L}_n^{s\to r}\}$, $n = 0, 1, 2, \cdots$, corresponding to itineraries between the turning points (r and s) of lengths $v + n(v+w)$ and $w + n(v+w)$, respectively, and by bones $\mathcal{B}^r$ and $\mathcal{B}^s$ corresponding to (superstable) cycles containing r and s respectively, in the following (cyclic) order: $\mathcal{L}_0^{r\to s}$, $\mathcal{B}^s, \cdots, \mathcal{L}_3^{r\to s}$, $\mathcal{L}_2^{r\to s}$, $\mathcal{L}_1^{r\to s}$, $\mathcal{L}_0^{s\to r}$, $\mathcal{L}_1^{s\to r}$, $\mathcal{L}_2^{s\to r}, \cdots, \mathcal{B}^r$, $\mathcal{B}^s, \cdots, \mathcal{L}_2^{r\to s}$, $\mathcal{L}_1^{r\to s}$, $\mathcal{L}_0^{r\to s}$, $\mathcal{L}_1^{s\to r}$, $\mathcal{L}_2^{s\to r}$, $\mathcal{L}_3^{s\to r}, \cdots, \mathcal{B}^r$, $\mathcal{L}_0^{s\to r}$. In other words, the generic unfolding of a χ-point is qualitatively as depicted in the sketch of Fig. 2(b).*

The proof follows a brief description of Fig. 2(b). The pair of unbroken curves tangent at the χ-point to the causative ligaments (itinerary lengths v and w) are the bones, one associated with each turning point. The two bundles of emanating ligaments, one tangent to each bone at the χ-point, are represented by the first four members of each. In both bundles, ligaments corresponding to longer itineraries are closer to the respective bone: the ligaments are marked with the value of the index n of Theorem 1. The asterisk corresponds to a situation like that depicted in Fig. 3(b).

Proof of Theorem 2. Let there be a doubly superstable cycle of Φ, period $v+w$, at $\lambda = \mu = 0$ with

$$\Phi^v(r_0, 0, 0) = s_0, \qquad \Phi^w(s_0, 0, 0) = r_0 \tag{2.2.1}$$

(with no smaller v or w satisfying these equations), where

$$\frac{\partial \Phi}{\partial x}(r_0, 0, 0) = \frac{\partial \Phi}{\partial x}(s_0, 0, 0) = 0, \quad \frac{\partial^2 \Phi}{\partial x^2}(r_0, 0, 0) \neq 0, \quad \frac{\partial^2 \Phi}{\partial x^2}(s_0, 0, 0) \neq 0. \tag{2.2.2}$$

We write the v- and w-fold compositions of Φ as

$$\begin{aligned} f(x, \lambda, \mu) &= \Sigma a_{ijk} x^i \lambda^j \mu^k \equiv \Phi^v(r_0 + x, \lambda, \mu), \\ g(x, \lambda, \mu) &= \Sigma b_{ijk} x^i \lambda^j \mu^k \equiv \Phi^w(s_0 + x, \lambda, \mu), \end{aligned} \tag{2.2.3}$$

with $a_{000} = b_{000} = a_{100} = b_{100} = 0$ from (2.2.1) and (2.2.2). Since by stipulation (and generically) $a_{200} \neq 0$ and $b_{200} \neq 0$, there exist unique analytic functions $r(\lambda, \mu)$ and

$s(\lambda, \mu)$ for the turning points, defined, respectively, by

$$\frac{\partial f}{\partial x}(r(\lambda, \mu), \lambda, \mu) = 0, \qquad \frac{\partial g}{\partial x}(s(\lambda, \mu), \lambda, \mu) = 0. \tag{2.2.4}$$

We write them as

$$r(\lambda, \mu) = \Sigma c_{jk} \lambda^j \mu^k, \qquad s(\lambda, \mu) = \Sigma d_{jk} \lambda^j \mu^k \tag{2.2.5}$$

(with $c_{00} = d_{00} = 0$). We define the functions $\mu_{ss}(\lambda)$, $\mu_{rr}(\lambda)$, representing the bones (associated with s and r, respectively) as follows:

$$\begin{aligned} fg(s(\lambda, \mu_{ss}(\lambda)), \lambda, \mu_{ss}(\lambda)) - s(\lambda, \mu_{ss}(\lambda)) = 0, \\ gf(r(\lambda, \mu_{rr}(\lambda)), \lambda, \mu_{rr}(\lambda)) - r(\lambda, \mu_{rr}(\lambda)) = 0, \end{aligned} \tag{2.2.6}$$

and expand them as

$$\mu_{ss}(\lambda) = \Sigma e_m \lambda^m, \qquad \mu_{rr}(\lambda) = \Sigma \overline{e_m} \lambda^m. \tag{2.2.7}$$

Unique analytic functions $\mu_{ss}(\lambda)$ and $\mu_{rr}(\lambda)$ are guaranteed since $d_{01} = -b_{101}/2b_{200} \neq a_{001}$ generically, and $c_{01} = -a_{101}/2a_{200} \neq b_{001}$ generically. The fact that the two equations in (2.2.6) are related by interchanging the coefficients $\{a_{ijk}\} \leftrightarrow \{b_{ijk}\}$, i.e., by switching the identification of the turning points, means that the $\{e_m\}$ and $\{\overline{e_m}\}$ are related by the same interchange (once they have been evaluated in terms of those problem-defining coefficients). In what follows an overbar will denote the interchange operation.

We now define for the $n = 0, 1, 2, \cdots$,

$$\begin{aligned} I_n^{r \to s}(\lambda, \mu) \equiv (fg)^n f(r, \lambda, \mu) - s, \\ I_n^{s \to r}(\lambda, \mu) \equiv (gf)^n g(s, \lambda, \mu) - r. \end{aligned} \tag{2.2.8}$$

Since $\partial I_n^{r \to s}(0, 0)/\partial \mu = a_{001} - d_{01}$ (for all n) is generically nonzero, as is $\partial I_n^{s \to r}(0, 0)/\partial \mu = b_{001} - c_{01}$, ligaments represented by analytic functions $\mu_n^{r \to s}(\lambda)$, $\mu_n^{s \to r}(\lambda)$, $n = 0, 1, 2, \cdots$, that are the unique solutions of

$$I_n^{r \to s}(\lambda, \mu_n^{r \to s}(\lambda)) = 0, \qquad I_n^{s \to r}(\lambda, \mu_n^{s \to r}(\lambda)) = 0, \tag{2.2.9}$$

respectively, are guaranteed. We write the deviations of the ligaments from the respective bones as

$$\Delta \mu_n^{r \to s}(\lambda) \equiv \mu_n^{r \to s}(\lambda) - \mu_{ss}(\lambda), \qquad \Delta \mu_n^{s \to r}(\lambda) \equiv \mu_n^{s \to r}(\lambda) - \mu_{rr}(\lambda). \tag{2.2.10}$$

It is the leading terms of these deviations in which we are interested. The first step is to derive a relation between $\Delta \mu_n^{r \to s}(\lambda)$ and the ligament-defining function $I_n^{r \to s}(\lambda, \mu)$, *evaluated on the bone* $\mu_{ss}(\lambda)$, as follows using (2.2.8) and (2.2.9):

$$\begin{aligned} 0 \equiv I_n^{r \to s}(\lambda, \mu_n^{r \to s}(\lambda)) = I_n^{r \to s}(\lambda, \mu_{ss}(\lambda) + \Delta \mu_n^{r \to s}(\lambda)), \\ = I_n^{r \to s}(\lambda, \mu_{ss}(\lambda)) - P \Delta \mu_n^{r \to s}(\lambda) + O(\lambda) \Delta \mu_n^{r \to s}(\lambda) + O((\Delta \mu_n^{r \to s}(\lambda))^2), \end{aligned} \tag{2.2.11}$$

where

$$P = -(a_{001} - d_{01}). \tag{2.2.12}$$

Substituting (2.2.5) into (2.2.4) yields $d_{01} = -b_{101}/2b_{200}$. Thus

$$\Delta \mu_n^{r \to s}(\lambda) = \frac{1}{P} I_n^{r \to s}(\lambda, \mu_{ss}(\lambda)) + \text{h.o.t.}, \tag{2.2.13}$$

where, as before, by h.o.t. we mean terms of higher order than that of the leading term of the quantity to the left. Next we obtain a recursion relation for the $\{I_n^{r\to s}(\lambda, \mu_{ss})\}$ using (2.2.8):

$$\begin{aligned} I_{n+1}^{r\to s}(\lambda,\mu) &= (fg)^{n+1} f(r,\lambda,\mu) - s \\ &= fg(fg)^n f(r,\lambda,\mu) - s \\ &= fg(I_n^{r\to s}+s,\lambda,\mu) - s. \end{aligned} \tag{2.2.14}$$

Expanding, we obtain

$$\begin{aligned} I_{n+1}^{r\to s}(\lambda,\mu_{ss}) &= fg(s(\lambda,\mu_{ss}) + I_n^{r\to s}(\lambda,\mu_{ss}),\lambda,\mu_{ss}) - s(\lambda,\mu_{ss}) \\ &= Q\lambda(I_n^{r\to s}(\lambda,\mu_{ss}))^2 + O(\lambda^2)(I_n^{r\to s}(\lambda,\mu_{ss}))^2 + O((I_n^{r\to s}(\lambda,\mu_{ss}))^4) \end{aligned} \tag{2.2.15}$$

(recall that s is a turning point of fg that is quadratic except at $0=\lambda=\mu_{ss}(0)$ where it is quartic) where, by substituting into (2.2.3), it is found that

$$Q = b_{200}[e_1(2a_{200}b_{001} + a_{101}) + 2a_{200}b_{010} + a_{110}]. \tag{2.2.16}$$

From the definition of μ_{ss} (2.2.6a),

$$e_1 = -(a_{010} - d_{10})/(a_{001} - d_{01}), \tag{2.2.17}$$

where, from the definition of s (2.2.5), $d_{10} = -b_{110}/2b_{200}$. Combining the two results (2.2.13) and (2.2.15), we obtain a recursion relation for the deviations $\{\Delta\mu_n^{r\to s}(\lambda)\}$:

$$\begin{aligned} \Delta\mu_{n+1}^{r\to s}(\lambda) &= \frac{1}{P} I_{n+1}^{r\to s}(\lambda,\mu_{ss}) + \text{h.o.t.} \\ &= \frac{Q}{P}\lambda[I_n^{r\to s}(\lambda,\mu_{ss})]^2 + \text{h.o.t.}, \\ &= \frac{Q}{P}\lambda[P\Delta\mu_n^{r\to s}(\lambda)]^2 + \text{h.o.t.}, \\ &= PQ\lambda[\Delta\mu_n^{r\to s}(\lambda)]^2 + \text{h.o.t.} \end{aligned} \tag{2.2.18}$$

The first element of this sequence, $\Delta\mu_0^{r\to s}(\lambda)$, may be evaluated directly from (2.2.3), (2.2.9), and (2.2.13) as

$$\Delta\mu_0^{r\to s}(\lambda) = \frac{R^2}{S}\lambda^2 + O(\lambda^3), \tag{2.2.19}$$

where

$$\begin{aligned} R = {} & +4b_{200}b_{001}a_{200}a_{010} + 2b_{200}a_{101}a_{010} - 4a_{200}a_{001}b_{200}b_{010} - 2a_{200}b_{101}b_{010} \\ & + 2a_{200}b_{110}b_{001} + a_{101}b_{110} - 2b_{200}a_{110}a_{001} - b_{101}a_{110}, \end{aligned} \tag{2.2.20}$$

and

$$S \equiv 2\frac{a_{200}}{b_{200}}[a_{001}^2 b_{200}^2(12b_{101} + 8a_{001}b_{200}) + b_{101}^2(6a_{001}b_{200} + b_{101})]. \tag{2.2.21}$$

Noting that in fact $PQ = R/2$, we have finally that

$$\Delta\mu_n^{r\to s}(\lambda) = \frac{2}{R}\left[\frac{R^3}{2S}\right]^{2^n} \lambda^{3(2^n)-1} + \text{h.o.t.} \tag{2.2.22}$$

So we find that the leading orders of the deviations $\{\Delta\mu_n^{r\to s}(\lambda)\}$ are λ^2, λ^5, λ^{11}, $\lambda^{23}, \cdots$, for $n=0, 1, 2, 3, \cdots$: all odd powers for $n>0$. Moreover, the sign of the leading order coefficients for $n>0$ are all the same: the sign of R. This means that if we consider the quadrants into which the neighborhood of the χ-point is divided by the bones $\mu_{ss}(\lambda)$, $\mu_{rr}(\lambda)$ (note that $e_1 \neq \overline{e_1}$ generically), the $r\to s$ ligaments for $n>0$ run from one quadrant to the opposing one and, as n increases, accumulate on $\mu_{ss}(\lambda)$, just as depicted in Fig. 2(b). The $n=0$, or causative, ligament is quadratically tangent to the bone (thus occupying a pair of adjacent quadrants).

By symmetry, the results derived for the $r\to s$ ligaments and the s-bone apply equally to the $s\to r$ ligaments and the r-bone; the analogous quantitative relations are obtained by interchanging the coefficients $\{a_{ijk}\}\leftrightarrow\{b_{ijk}\}$.

To complete the analysis we first observe that for $n>0$, the sign of $\Delta\mu_n^{r\to s}(\lambda)\Delta\mu_n^{s\to r}(\lambda)=\Delta\mu_n^{r\to s}(\lambda)\overline{\Delta\mu_n^{r\to s}(\lambda)}$ is the sign of $R\bar{R}$. But R is, by inspection, antisymmetric under the interchange operation. Therefore

$$R\bar{R}\leqq 0. \tag{2.2.23}$$

This result implies that the bundles of ligaments given by $\mu_n^{r\to s}(\lambda)$ and $\mu_n^{s\to r}(\lambda)$, $n>0$, occupy the same pair of quadrants, leaving two quadrants unoccupied by ligaments with $n>0$.

To determine the relative placement of the $n=0$ (causative) ligaments, we consider the product of coefficients:

$$[\mu_0^{r\to s}(\lambda)-\mu_0^{s\to r}(\lambda)]\Delta\mu_1^{r\to s}(\lambda)\Delta\mu_0^{r\to s}(\lambda)\overline{\Delta\mu_0^{r\to s}(\lambda)}=T\lambda^{10}+O(\lambda^{11}), \tag{2.2.24}$$

where

$$\begin{aligned} T&=(e_1-\overline{e_1})\frac{R}{2}\left(\frac{R^2}{S}\right)^2\frac{R^2}{S}\frac{\bar{R}^2}{\bar{S}}\\ &=(e_1-\overline{e_1})\frac{R}{2}\frac{1}{S\bar{S}}\frac{R^4}{S^2}\bar{R}^2. \end{aligned} \tag{2.2.25}$$

But from (2.2.17),

$$(e_1-\overline{e_1})=-\frac{R}{U}, \tag{2.2.26}$$

where

$$U=[4a_{200}a_{001}b_{200}b_{001}+2a_{101}a_{001}b_{200}+a_{101}b_{101}+2b_{101}b_{001}a_{200}], \tag{2.2.27}$$

and from (2.2.21),

$$S\bar{S}=4U^3, \tag{2.2.28}$$

whence

$$T=-\frac{1}{8}\left[\frac{R^4\bar{R}}{SU^2}\right]^2\leqq 0. \tag{2.2.29}$$

Examination of Fig. 2(b) shows that any arrangement of the curves which is in accord with the result $R\bar{R}\leqq 0$, but which is qualitatively different from that shown (other than in orientation with respect to the λ, μ axes, and the sign of T is invariant under such orientation changes) will contradict $T\leqq 0$. Thus the qualitative arrangement of the ligaments and bones is universal, and is as specified in Theorem 2. □

2.3. The ψ-point. The intersection of a ligament with a bone that is associated with the terminal turning point of the ligament's itinerary, a situation such as the one

sketched in Fig. 1(c), constitutes the third kind of singular point of the parameter plane that we consider. We show below that such a point is the source of a bundle of additional ligaments. As in the case of the χ-point, for purposes of construction of the skeleton we consider the intersection of the ligament and bone to be "causative" of the bundle of additional ligaments, but for analytical convenience we define the point as follows.

DEFINITION. A point (λ_0, μ_0) of the parameter plane of the map $\Phi(x, \lambda, \mu)$ is a *ψ-point* if there exist v, $p \in \mathbb{N}$ such that $\Phi^v(r_0, \lambda_0, \mu_0) = s_0$, $\Phi^p(s_0, \lambda_0, \mu_0) = s_0$, where r_0, s_0 are (distinct) quadratic turning points of $\Phi(x, \lambda_0, \mu_0)$, and there does not exist $w(<p) \in \mathbb{N}$ such that $\Phi^w(s_0, \lambda_0, \mu_0) = r_0$.

The existence of the causative ligament and bone, as well as of the bundle of additional ligaments, follows from this definition as we demonstrate below. The additional ligaments correspond to itineraries that at the ψ-point are degenerate and consist of the direct inter-turning-point itinerary plus a some number of circuits around the superstable cycle. Figure 3(c) sketches the situation near a ψ-point on one such ligament.

THEOREM 3. *Let (λ_0, μ_0) be a ψ-point of $\Phi(x, \lambda, \mu)$ for which the turning points are r_0 and s_0, and let $v = \min(v' \in \mathbb{N} | \Phi^{v'}(r_0, \lambda_0, \mu_0) = s_0)$ and $p = \min(p' \in \mathbb{N} | \Phi^{p'}(s_0, \lambda_0, \mu_0) = s_0)$. Generically, a sufficiently small parameter-plane circle centered at (λ_0, μ_0) is intersected (transversally) by ligaments $\{\mathscr{L}_n^{r\to s}\}$, $n = 0, 1, 2, \cdots$, corresponding to itineraries of lengths $v + np$ from turning point r to turning point s, and by a bone $\mathscr{B}^s$ corresponding to a superstable cycle containing s, in the following (cyclic) order: $\mathscr{L}_0^{r\to s}$, $\mathscr{B}^s, \cdots, \mathscr{L}_3^{r\to s}$, $\mathscr{L}_2^{r\to s}$, $\mathscr{L}_1^{r\to s}$, $\mathscr{L}_0^{s\to r}$, $\mathscr{L}_1^{r\to s}$, $\mathscr{L}_2^{r\to s}$, $\mathscr{L}_3^{r\to s}, \cdots, \mathscr{B}^s$. In other words, the generic unfolding of a ψ-point is qualitatively as depicted in the sketch of Fig.* 2(c).

In Fig. 2(c) the unbroken curve represents the bone, and in the bundle of ligaments the ones corresponding to longer itineraries are closer to the bone. The asterisk corresponds to a situation like that depicted in Fig. 3(c).

Proof of Theorem 3. At $\lambda = \mu = 0$ let there exist an itinerary of length v between turning points of the map $\Phi(x, \lambda, \mu)$ and a superstable cycle of period p that includes the terminal turning point of that itinerary:

$$\Phi^v(r_0, 0, 0) = s_0, \qquad \Phi^p(s_0, 0, 0) = s_0 \tag{2.3.1}$$

(with v, p the smallest numbers satisfying these equations) where

$$\frac{\partial \Phi}{\partial x}(r_0, 0, 0) = \frac{\partial \Phi}{\partial x}(s_0, 0, 0) = 0, \quad \frac{\partial^2 \Phi}{\partial x^2}(r_0, 0, 0) \neq 0, \quad \frac{\partial^2 \Phi}{\partial x^2}(s_0, 0, 0) \neq 0. \tag{2.3.2}$$

In a fashion similar to that of § 2.2, we write the v- and p-fold composites of Φ as

$$\begin{aligned} f(x, \lambda, \mu) &= \Sigma a_{ijk} x^i \lambda^j \mu^k \equiv \Phi^v(r_0 + x, \lambda, \mu), \\ g(x, \lambda, \mu) &= \Sigma b_{ijk} x^i \lambda^j \mu^k \equiv \Phi^p(s_0 + x, \lambda, \mu), \end{aligned} \tag{2.3.3}$$

with $a_{000} = b_{000} = a_{100} = b_{100} = 0$ from (2.3.2), and the analytic functions $r(\lambda, \mu)$, $s(\lambda, \mu)$, representing the two turning points, defined, respectively, by

$$\frac{\partial f}{\partial x}(r(\lambda, \mu), \lambda, \mu) = 0, \qquad \frac{\partial g}{\partial x}(s(\lambda, \mu), \lambda, \mu) = 0 \tag{2.3.4}$$

(and guaranteed, respectively, by $a_{200} \neq 0$, $b_{200} \neq 0$) as

$$r(\lambda, \mu) = \Sigma c_{jk} \lambda^j \mu^k, \qquad s(\lambda, \mu) = \Sigma d_{jk} \lambda^j \mu^k \tag{2.3.5}$$

($c_{00} = d_{00} = 0$). The existence of a bone associated with s, represented by the function $\mu_{ss}(\lambda)$ that satisfies

$$g(s(\lambda, \mu_{ss}(\lambda)), \lambda, \mu_{ss}(\lambda)) - s(\lambda, \mu_{ss}(\lambda)) = 0, \tag{2.3.6}$$

is guaranteed by the generic inequality of b_{001} and d_{01}. We expand $\mu_{ss}(\lambda)$ as

$$\mu_{ss}(\lambda) = \Sigma e_j \lambda^j. \tag{2.3.7}$$

We then define for $n = 0, 1, 2, \cdots$

$$I_n^{r\to s}(\lambda, \mu) \equiv g^n f(r, \lambda, \mu) - s. \tag{2.3.8}$$

Since $\partial I_n^{r\to s}(0, 0)/\partial\mu = a_{001} - d_{01}$ (for all n) is generically nonzero, we are guaranteed, for each n, a ligament represented by the unique function $\mu_n^{r\to s}(\lambda)$ that satisfies

$$I_n^{r\to s}(\lambda, \mu_n^{r\to s}(\lambda)) = 0. \tag{2.3.9}$$

We write the deviations of the ligaments from the bone as

$$\Delta\mu_n^{r\to s}(\lambda) \equiv \mu_n^{r\to s}(\lambda) - \mu_{ss}(\lambda). \tag{2.3.10}$$

As before, we compute the leading terms of these deviations. Setting $n = 0$ in (2.3.10), substituting the series expansions for the various quantities, and collecting coefficients yields

$$\Delta\mu_0^{r\to s}(\lambda) = \frac{R}{S}\lambda + O(\lambda^2), \tag{2.3.11}$$

where

$$R = 2b_{200}[2b_{200}(a_{001}b_{010} - b_{001}a_{010}) + b_{110}(a_{001} - b_{010}) - b_{101}(a_{010} - b_{010})], \tag{2.3.12}$$

and

$$S = (2b_{200}a_{001} + b_{101})(2b_{200}b_{001} + b_{101}). \tag{2.3.13}$$

A recursion relation is obtained as follows:

$$\begin{aligned} 0 \equiv I_n^{r\to s}(\lambda, \mu_n^{r\to s}(\lambda)) &= I_n^{r\to s}(\lambda, \mu_{ss}(\lambda) + \Delta\mu_n^{r\to s}(\lambda)), \\ &= I_n^{r\to s}(\lambda, \mu_{ss}(\lambda)) - P_n\Delta\mu_n^{r\to s}(\lambda) + O(\lambda)\Delta\mu_n^{r\to s}(\lambda) + O((\Delta\mu_n^{r\to s}(\lambda))^2), \end{aligned} \tag{2.3.14}$$

where

$$P_n = \begin{cases} -(a_{001} - d_{01}), & n = 0, \\ -(b_{001} - d_{01}), & n > 0, \end{cases} \tag{2.3.15}$$

and $d_{01} = -b_{101}/2b_{200}$. Thus

$$\Delta\mu_n^{r\to s}(\lambda) = \frac{1}{P_n} I_n^{r\to s}(\lambda, \mu_{ss}(\lambda)) + \text{h.o.t.} \tag{2.3.16}$$

We have also that

$$\begin{aligned} I_{n+1}^{r\to s}(\lambda, \mu) &= g^{n+1} f(r, \lambda, \mu) - s \\ &= g g^n f(r, \lambda, \mu) - s \\ &= g(I_n^{r\to s} + s, \lambda, \mu) - s. \end{aligned} \tag{2.3.17}$$

Thus

$$\begin{aligned} I_{n+1}^{r\to s}(\lambda, \mu_{ss}) &= g(s(\lambda, \mu_{ss}) + I_n^{r\to s}(\lambda, \mu_{ss}), \lambda, \mu_{ss}) - s(\lambda, \mu_{ss}) \\ &= Q[I_n^{r\to s}(\lambda, \mu_{ss})]^2 + O(\lambda^2)[I_n^{r\to s}(\lambda, \mu_{ss})]^2 + O([I_n^{r\to s}(\lambda, \mu_{ss})]^4), \end{aligned} \tag{2.3.18}$$

where

$$Q = b_{200}. \tag{2.3.19}$$

(Recall that s is a turning point of g that is quadratic even at $\lambda = \mu = 0$.) We thus obtain

$$\begin{aligned}
\Delta\mu_{n+1}^{r\to s}(\lambda) &= \frac{1}{P_{n+1}} I_{n+1}^{r\to s}(\lambda, \mu_{ss}) + \text{h.o.t.} \\
&= \frac{Q}{P_{n+1}} \lambda [I_n^{r\to s}(\lambda, \mu_{ss})]^2 + \text{h.o.t.} \\
&= \frac{Q}{P} \lambda [P_n \Delta\mu_n^{r\to s}(\lambda)]^2 + \text{h.o.t.} \\
&= \begin{cases} \dfrac{Q}{P_1} P_0^2 \lambda [\Delta\mu_n^{r\to s}(\lambda)]^2 + \text{h.o.t.}, & n = 0, \\ P_1 Q \lambda [\Delta\mu_n^{r\to s}(\lambda)]^2 + \text{h.o.t.}, & n > 0. \end{cases}
\end{aligned} \tag{2.3.20}$$

And finally,

$$\Delta\mu_n^{r\to s}(\lambda) = \begin{cases} \dfrac{R}{S}\lambda + O(\lambda)^2, & n = 0, \\ \left[\dfrac{(P_1 Q)^{2^{n-1}}}{P_1 Q}\right]\left[\dfrac{Q}{P_1}\right]^{2^{n-1}}\left[\dfrac{P_0 R}{S}\right]^{2^n} \lambda^{2^n} + O(\lambda^{2^n+1}), & n > 0. \end{cases} \tag{2.3.21}$$

The $n = 0$ result indicates the generic transversal intersection of the ligament and bone which we think of as causing the ψ-point. The leading orders of the $\{\Delta\mu_n^{r\to s}(\lambda)\}$ representing the ligaments which emanate from the ψ-point are $\lambda^2, \lambda^4, \lambda^8, \lambda^{16}, \cdots$ for $n = 1, 2, 3, 4, \cdots$: all even powers. The sign of the leading order coefficients for $n > 0$ are all the same: the signs of $P_1 Q$. (Note that for $n = 1$, it is the sign of Q/P_1, which is of course the same.) Thus the qualitative arrangement of the ligaments and bones is universal, and is as specified in Theorem 3. □

3. Extending the bones and ligaments away from the neighborhood of their point of origin. Having determined the manner in which bones and ligaments originate, we now consider the rules which govern their behavior *away* from their points of origin, and examine the consequences.

In the present paper we restrict our attention to "full" two-parameter families of maps.

DEFINITION. A two-parameter family of two-extremum maps is *full* if there is complete transversal self-intersection of the ligament bundle emanating from an α-point.

The cubic family $\Phi(x, \lambda, \mu) = x^3 - \lambda x + \mu$, for example, with α-point at $\lambda = \mu = 0$, is full. The ligaments emanating from an α-point of a full family are sketched in Fig. 4(a). The χ-points caused by the intersections of these ligaments are marked by large dots. We can proceed with the derivation of the qualitative form of the rest of the skeleton by adding, according to the local rules derived in § 2, the bones and secondary ligaments that originate from each χ-point. The addition of the ligaments generates additional χ-points, and ψ-points are generated by the addition of the bones. Then at each of these points, the emanating bones and/or ligaments are again added. This procedure has a topologically unique result (independent of the order in which we choose to deal with the χ- and ψ-points) if, following [2], we insist that the family is "nice."

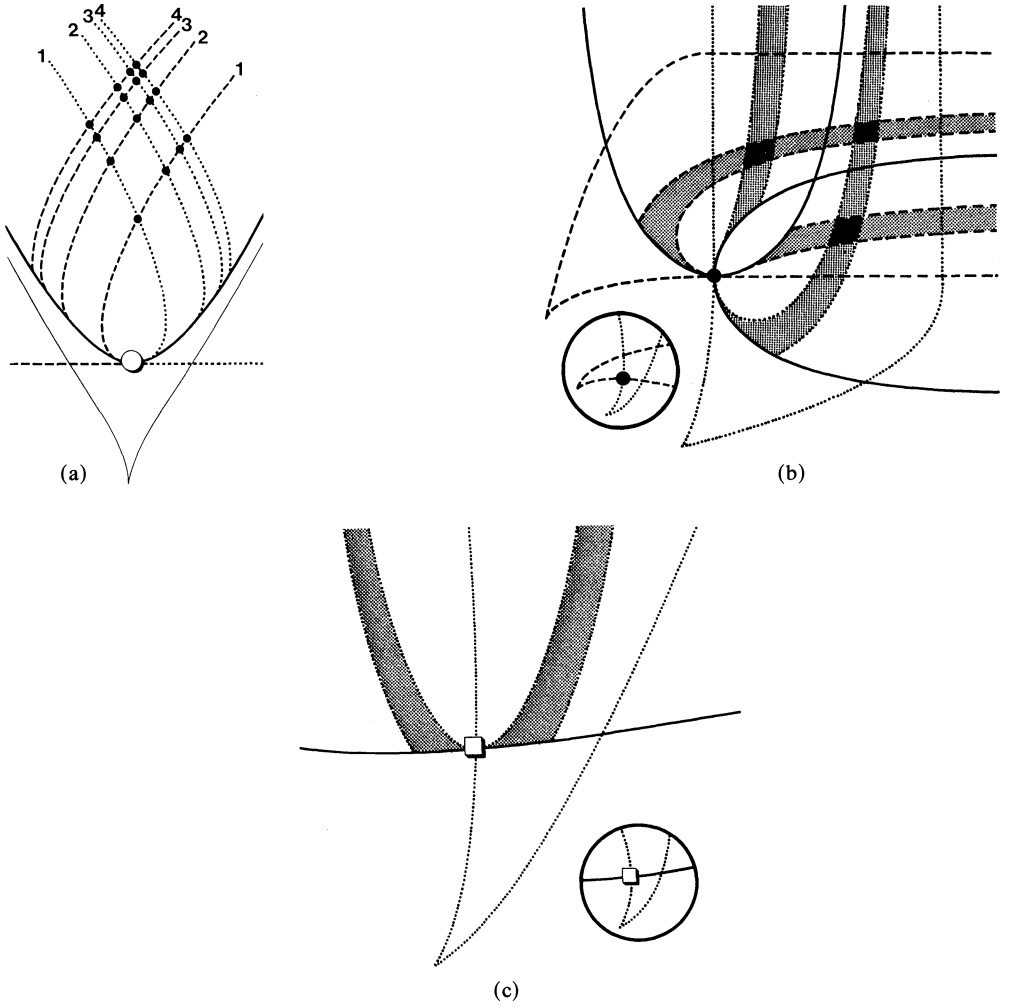


FIG. 4. (a) *A "full" two-parameter family of maps: the ligament bundle (broken curves) emanating from the α-point intersects itself completely (only first four ligaments shown) and transversally. This diagram is the starting point for the construction of the skeleton pursued in Fig. 5. The unbroken curve is the period-1 bone, as in Fig.* 2(a). *The thin* v*-shaped curve is a locus of tangent bifurcation which is commonly present (see* [1], *for example).* (b) *Inset: environment of every χ-point in Fig.* 4(a). *Main figure: necessary paths of ligaments and bones emanating from a χ-point situated as in the inset. The shaded bands represent the ligament bundles.* (c) *Inset: environment of every ψ-point in Fig.* 4(b). *Main figure: necessary paths of ligaments and bones emanating from a ψ-point situated as in the inset. The shaded band represents the ligament bundle.*

DEFINITION. A two-parameter family is *nice* if each bone and ligament has only one point of origin in the sense of § 2 (cf. [2]).

This property, which is imposed to guarantee that things are as simple as possible, implies the following:

(1) Two ligaments of the same type may not touch except at the point of origin of at least one of them, for it can easily be shown that a point of coincidence of two ligaments of the same type is necessarily an α-, χ-, or ψ-point; and

(2) Away from its point of origin, a bone may not touch a ligament whose itinerary begins at the turning point with which the bone is associated, for such a point of coincidence would necessarily be another χ-point of origin of the bone.

TABLE 1
Rules of contact and intersection of curves (r denotes one turning point, s the other).

	$s \to r$ ligament	$r \to s$ ligament	r-bone	s-bone
r-bone	ψ-point	forbidden[2]	forbidden	unrestricted[3]
$s \to r$ ligament	forbidden[1]	χ-point		

[1] By niceness, unless point of origin of at least one of the ligaments.
[2] Other than at χ-point of origin of bone.
[3] May be χ-point of origin of the bones, otherwise no interaction.

Additionally, it is a proximate consequence of the implicit function theorem that (without assumption) the coincidence of two bones associated with the same turning point is structurally unstable in two-parameter families. So generically no two bones associated with the same turning point intersect or touch. On the other hand, two bones can cross with no restriction and no interaction if they are associated with different turning points.

The rules of contact and interaction of bones and ligaments we have established are summarized in Table 1, and are sufficient to guide the construction of the skeleton unambiguously, as is now explained. We start with Fig. 4(a). Consider that the environment of each χ-point in Fig. 4(a) is as depicted in the circular inset of Fig. 4(b). That is, each causative ligament is intersected "downstream" by a ligament which has the same point of origin as the other causative ligament. We make three observations. First, of the two ways that the local picture of Fig. 2(b) can be oriented at the marked χ-point, only one does not lead to a conflict with the "crossability" rules described above when the emanating curves are extended; a ligament or bone cannot emanate into a sealed box of uncrossable walls. The unique orientation that is possible is as sketched in the main part of Fig. 4(b), where the shaded bands represent the bundles of emanating ligaments. Second, the way the emanating curves are extended in Fig. 4(b) away from the locality of the χ-point is the only way of reconciling the local picture with the crossability rules. Third, the environment of every additional χ-point caused by the emanating ligaments is exactly analogous to that of the marked χ-point in the inset; the same arguments thus apply to the curves emanating from these points.

Similarly, consider that the environment of each ψ-point generated by the introduction of the bones in Fig. 4(b) is as depicted in the circular inset of Fig. 4(c). That is, it has a neighboring ψ-point, with common bone, whose causative ligament has the same point of origin as its own. The ψ-point's ligament bundle cannot emanate "downwards" into the sealed trilateral box, but must emanate "upwards" as sketched in the main part of Fig. 4(c). (The shaded band represents the bundle of emanating ligaments.) All the χ-points and additional ψ-points caused by the ligaments emanating from the ψ-point again have environments analogous to those depicted in the insets of Figs. 4(b) and 4(c), respectively, and so the above arguments apply recursively to all the χ- and ψ-points generated. Thus if we start with Fig. 4(a), and follow the rules described, a topologically unique network of bones and ligaments will be generated.

In Fig. 5 we show this network constructed as far as to include all bones up to period 6. (A true sketch was drawn according to the rules: the figure as presented is a fair copy of that sketch with the spacing of the curves adjusted for better clarity.) A subset of this picture (bones up to period 5, no ligaments) was generated by Mackay and Tresser [2] following an entirely different procedure based on ordering of kneading

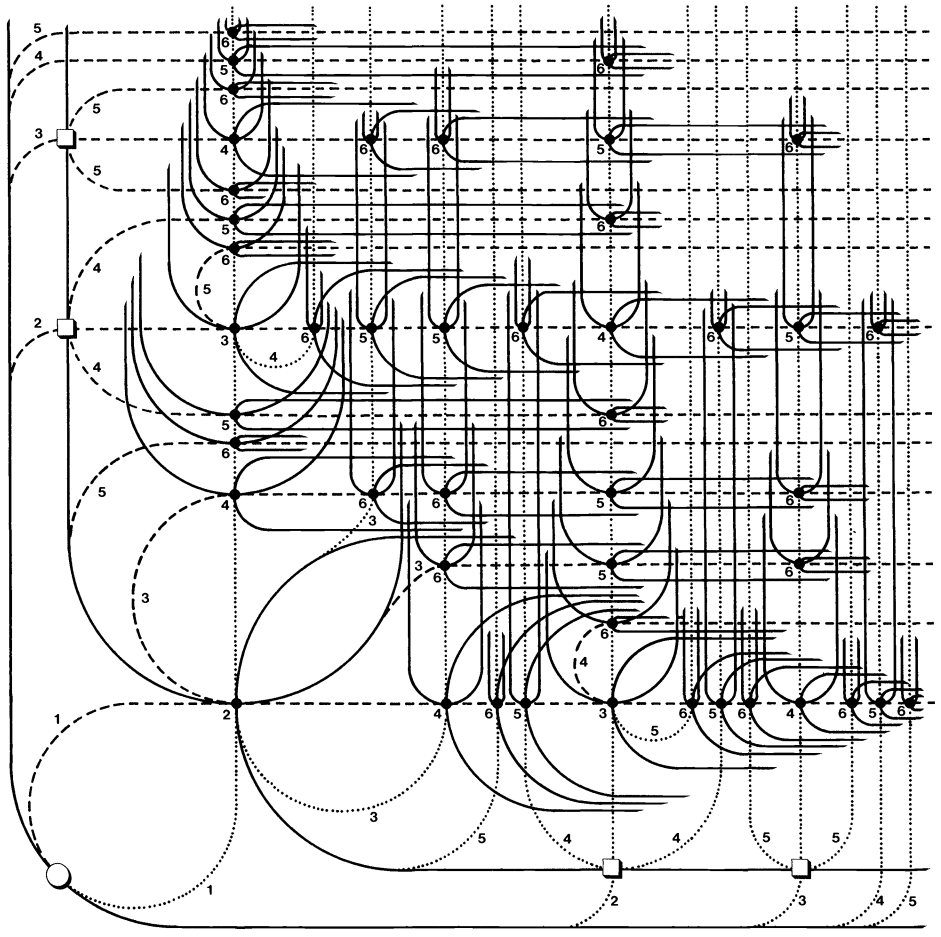


FIG. 5. *The skeleton of bones and ligaments constructed, according to the rules developed, to the extent of including every bone of period* 6 *and less. The figure constitutes a genealogy of the cycles. The length of the itinerary associated with each ligament, and the period of the cycle associated with each* χ*-point and its emanating bones, are marked. The reader is invited to check the arithmetic of these numbers. The same symbols as before are used for the* α*-,* χ*-, and* ψ*-points. (Only four* ψ*-points participate at this level; this number increases greatly if period-*7 *bones are to be included.) In order to avoid a dense mass of bones at the upper right, bones are produced only as far as is necessary to visually establish their positions relative to other bones.*

sequences. Bones of the inadmissible cycles of [2] do not arise in our construction simply because the loci (ligaments) of the would-be constituent kneading sequences do not get an opportunity to intersect, due to the relative positions of their (α-, χ-, or ψ-) points of origin.

Of the numerous interesting structural features of the skeleton of bones and ligaments, we mention two. One is the substructure responsible for the *U*-sequence [5] of doubly superstable (χ-) points which occurs along each causative ligament of every doubly superstable point. The lexical self-similarity of the U-sequence is reflected in structural self-similarity of the parameter-plane skeleton. This can be seen, for example, in Fig. 5 on either of the ligaments of itinerary length 1 that emanate from the α-point. Consider the U-sequence based on the period-2 χ-point and the one based

on the neighboring period-4 χ-point. The ligaments (itinerary lengths 3, 5, (7, 9, $\cdots$)) emanating from the period-2 χ-point can be seen to play the role for the latter analogous to that played for the former by the ligaments (itinerary lengths 1, 2, 3, 4, $\cdots$) emanating from the α-point. A more extensive development of the U-sequence substructure appears in Fig. 5 of [9], where the results of the present paper, Theorems 2 and 3, are used. A second self-similar substructure of the bone-and-ligament skeleton is the two-parameter Farey tree, as developed in [9] using Theorem 2 of the present paper.

The great ease of constructing the parameter-plane skeleton using our method is a facet of the method's value.[3] But more significant is that the genealogy of the cycles is realized and manifest in our skeleton of bones and ligaments, and that the emergence of the global structure upon the application of local rules offers a new perspective on the dynamics of two-extremum maps.

Acknowledgment. The authors thank C. Frenzen for useful discussions.

REFERENCES

[1] J. BELAIR AND L. GLASS, *Universality and self-similarity in the bifurcation of circle maps*, Phys. D, 16 (1985), pp. 143-154.

[2] R. S. MACKAY AND C. TRESSER, *Some flesh on the skeleton: the bifurcation structure of bimodal maps*, Phys. D, 27 (1987), pp. 412-422.

[3] P. COLLET AND J.-P. ECKMANN, *Iterated maps of the interval as dynamical systems*, Birkhäuser, Boston, 1980.

[4] J. GUCKENHEIMER, *Bifurcations of dynamical systems*, in Dynamical Systems, J. Guckenheimer, J. Moser, and S. Newhouse, eds., 1978 CIME Lectures, Birkhäuser, Boston, 1980.

[5] N. METROPOLIS, M. L. STEIN, AND P. R. STEIN, *On finite limit sets for transformations on the unit interval*, J. Combin. Theory, 15 (1973), pp. 25-44.

[6] J. MILNOR AND W. THURSTON, *On Iterated Maps of the Interval*, Lecture Notes in Math. 1342, Springer-Verlag, Berlin, New York, 1988.

[7] J. P. KEENER AND L. GLASS, *Global bifurcations of a periodically forced nonlinear oscillator*, J. Math. Biol., 21 (1984), pp. 175-190.

[8] A. VANDERBAUWHEDE, *Subharmonic branching in reversible systems*, SIAM J. Math. Anal., 21 (1990), pp. 954-979.

[9] J. RINGLAND AND M. SCHELL, *The Farey tree embodied—in bimodal maps of the interval*, Phys. Lett. A, 136 (1989), pp. 379-386.

[3] The ease comes at the expense of ignorance of the symbol sequences of the cycles, but for many purposes the periodicities alone (supplied by our method) will be the primary concern.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1372–1387, September 1991

011

CONSTRUCTING SYMMETRIC NONNEGATIVE MATRICES WITH PRESCRIBED EIGENVALUES BY DIFFERENTIAL EQUATIONS*

MOODY T. CHU† AND KENNETH R. DRIESSEL‡

Abstract. The inverse eigenvalue problem is solved for symmetric nonnegative matrices by means of a differential equation. If the given spectrum is feasible, then a symmetric nonnegative matrix can be constructed simply by following the solution curve of the differential system. The choice of the vector field is based on the idea of minimizing the distance between the cone of symmetric nonnegative matrices and the isospectral surface determined by the given spectrum. The projected gradient of the objective function is explicitly described. Using center manifold theory, it is also shown that the ω-limit set of any solution curve is a single point. Some numerical examples are presented.

Key words. nonnegative matrix, eigenvalue, projected gradient, stable manifold

AMS(MOS) subject classifications. 58F10, 49D10, 15A18, 65F15

1. Introduction.

A matrix $A \in R^{n\times n}$ is said to be nonnegative if no entry of A is negative. Nonnegative matrices arise frequently in various applied areas [3]. The Perron–Frobenius theorem concerning the spectrum of nonnegative matrices may be regarded as the central result in the theory of nonnegative matrices. It is, therefore, of great interest to study the following inverse eigenvalue problem.

PROBLEM 1. *Given a set $\sigma := \{\lambda_1, \cdots, \lambda_n\} \subset C$, find necessary and sufficient conditions for σ to be the spectrum of some nonnegative matrix.*

In practice, the prescribed eigenvalues $\lambda_1, \cdots, \lambda_n$ often are real numbers. Thus it is also interesting to ask the following.

PROBLEM 2. *Given a set $\sigma = \{\lambda_1, \cdots, \lambda_n\} \subset R$, find necessary and sufficient conditions for σ to be the spectrum of some symmetric nonnegative matrix.*

For decades researchers have been trying to answer these problems. A few necessary and a few sufficient conditions can be found, for example, in [2], [10]–[14], [16], [18], or more recently in [4]. To our knowledge, however, neither Problem 1 nor Problem 2 has been completely solved.

In this paper we want to address the problem of constructing a nonnegative matrix with prescribed spectrum. The problem is stated as follows.

PROBLEM 3. *Given a set σ of n real values that is known a priori to be the spectrum of some nonnegative matrix, numerically construct a symmetric nonnegative matrix whose spectrum is exactly σ.*

We have not found much discussion of Problem 3 in the literature. The most constructive result we have seen is the sufficient condition studied by Soules [18]. But Soules' condition is still limited because his construction depends on the specification of the Perron eigenvector—in particular, the components of the Perron eigenvector need to satisfy certain inequalities in order for his construction to work.

For our consideration, we shall need the following notation. Let $\mathcal{O}(n)$ denote the set of all orthogonal matrices in $R^{n\times n}$. Let Λ denote the diagonal matrix with

* Received by the editors December 4, 1989; accepted for publication (in revised form) October 5, 1990.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205. This research was supported in part by National Science Foundation grant DMS-9006135.

‡ Department of Mathematics, Idaho State University, Pocatello, Idaho 83209-0009.

diagonal entries $\lambda_1, \cdots, \lambda_n$; in symbols,

$$\Lambda := \operatorname{diag}\{\lambda_1, \cdots, \lambda_n\}. \tag{1}$$

The set

$$\mathcal{M}(\Lambda) := \{Q^T \Lambda Q | Q \in \mathcal{O}(n)\} \tag{2}$$

will be called the isospectral surface corresponding to Λ. Although the assumption is not required in our discussion, it can be shown that $\mathcal{M}(\Lambda)$ is indeed a smooth manifold with dimension $n(n-1)/2$ if all λ_i are distinct. The set of all symmetric nonnegative matrices in $R^{n\times n}$ is denoted by $\pi_s(R_+^n)$. We note that Problem 2 is equivalent to the following.

PROBLEM 2′. *Find necessary and sufficient conditions for the intersection of the isospectral surface $\mathcal{M}(\Lambda)$ and the cone $\pi_s(R_+^n)$ to be nonempty.*

Thus we are motivated to explore the idea of developing a way to systematically reduce the distance between $\pi_s(R_+^n)$ and $\mathcal{M}(\Lambda)$. If $\mathcal{M}(\Lambda)$ does intersect $\pi_s(R_+^n)$, then, of course, the distance is zero. Otherwise, our approach still finds a matrix from $\mathcal{M}(\Lambda)$ and a matrix from $\pi_s(R_+^n)$ such that their distance is a local minimum. In the latter case, the matrix from $\pi_s(R_+^n)$ is expected to be on a face of the cone $\pi_s(R_+^n)$, i.e., some of the entries of the nonnegative matrix are zero. We shall see that this property indeed shows up naturally in the development of our theory. Another fact, obvious from the geometry, is also worth mentioning: if $\mathcal{M}(\Lambda)$ intersects $\pi_s(R_+^n)$ at an interior point, then $\mathcal{M}(\Lambda)$ intersects $\pi_s(R_+^n)$ in a relative neighborhood of that point. In this case there are infinitely many symmetric nonnegative matrices corresponding to the given spectrum.

We can precisely formulate our idea as a constrained optimization problem. We first note that the set $\pi(R_+^n)$ of all nonnegative matrices in $R^{n\times n}$ can be formed as

$$\pi(R_+^n) = \{B * B | B \in R^{n\times n}\}, \tag{3}$$

where $X * Y$ denotes the Hadamard product of matrices X and Y. Let $\mathcal{S}(n)$ denote the set of all symmetric matrices in $R^{n\times n}$. Let

$$\langle A, B\rangle := \operatorname{trace}(AB^T) = \sum_{i,j} a_{ij} b_{ij} \tag{4}$$

denote the Frobenius inner product of two matrices $A, B \in R^{n\times n}$. We shall consider the following minimization problem.

PROBLEM 4. *Minimize*

$$F(Q, R) := \tfrac{1}{2}\|Q^T \Lambda Q - R * R\|^2, \tag{5}$$

subject to

$$(Q, R) \in \mathcal{O}(n) \times \mathcal{S}(n),$$

where $\|\cdot\|$ represents the Frobenius matrix norm. We shall show that the projection of the gradient vector of the objective function F onto the manifold $\mathcal{O}(n) \times \mathcal{S}(n)$ can be calculated explicitly. Consequently, we can introduce a steepest descent vector field on $\mathcal{O}(n) \times \mathcal{S}(n)$. This vector field can easily be transformed into a "flow" on the isospectral surface $\mathcal{M}(\Lambda)$ and a "flow" in the cone $\pi_s(R_+^n)$. Both flows are moving

in the steepest descent direction to minimize their distance until an equilibrium is reached. Our approach to Problem 4, *therefore, is a continuous realization process.*

In our earlier works, we have applied similar ideas to tackle the inverse Toeplitz eigenvalue problems [9] and other least squares matrix approximation problems subject to spectral constraint [8]. Our approach there proves to be quite successful. In this paper, we shall use some of our previously developed ideas. In §2 we develop the differential system; this is our main result. In §3 we use center manifold theory to study the stability properties of the resulting differential system. We argue that generically the ω-limit set of a solution flow contains only a single point. This proves the global convergence of our method. In §4, we study in detail the stability of equilibria for the case $n = 2$. Although this represents the simplest case, the stability analysis should shed some light on the behavior of our flow for higher-dimension cases. We present some numerical examples in the last section.

2. Projected gradient. In the product space $R^{n\times n} \times R^{n\times n}$, we shall use the induced Frobenius inner product:

$$\langle (A_1, A_2), (B_1, B_2) \rangle := \langle A_1, B_1 \rangle + \langle A_2, B_2 \rangle. \tag{6}$$

With this topology, the feasible set $\mathcal{O}(n) \times \mathcal{S}(n)$ of Problem 4 is clearly a smooth manifold. It is not difficult to show [8] that the space tangent to $\mathcal{O}(n) \times \mathcal{S}(n)$ at a point $(Q, R) \in \mathcal{O}(n) \times \mathcal{S}(n)$ is given by

$$\mathcal{T}_{(Q,R)}\mathcal{O}(n) \times \mathcal{S}(n) = \mathcal{T}_Q\mathcal{O}(n) \times \mathcal{T}_R\mathcal{S}(n) = Q\mathcal{S}(n)^\perp \times \mathcal{S}(n), \tag{7}$$

where $\mathcal{S}(n)^\perp$ denotes the orthogonal complement of $\mathcal{S}(n)$ and is composed of all skew-symmetric matrices in $R^{n\times n}$.

We first extend the definition of the function F in (5) in an obvious way to the entire space $R^{n\times n} \times R^{n\times n}$. A straightforward calculation shows that the Fréchet derivative of F at a general point $(A, B) \in R^{n\times n} \times R^{n\times n}$ acting on $(H, K) \in R^{n\times n} \times R^{n\times n}$ is

$$\begin{aligned} F'(A,B)(H,K) &= \langle A^T\Lambda A - B * B, H^T\Lambda A + A^T\Lambda H - K * B - B * K \rangle \\ &= \langle \Lambda A[(A^T\Lambda A - B * B)^T + (A^T\Lambda A - B * B)], H \rangle \\ &\quad + \langle -2(A^T\Lambda A - B * B) * B, K \rangle. \end{aligned} \tag{8}$$

The adjoint property $\langle A, BC \rangle = \langle AC^T, B \rangle = \langle B^T A, C \rangle$ has been used to rearrange terms in (8). It follows that, with respect to the inner product (6), the gradient of F at (A, B) is a pair of matrices; in fact, we have

$$\begin{aligned} \nabla F(A,B) &= \big(\Lambda A[(A^T\Lambda A - B * B)^T + (A^T\Lambda A - B * B)], \\ &\qquad - 2(A^T\Lambda A - B * B) * B\big). \end{aligned} \tag{9}$$

We are interested only in the case when $(A, B) = (Q, R) \in \mathcal{O}(n) \times \mathcal{S}(n)$. In this case, (9) is simplified to

$$\nabla F(Q,R) = (2\Lambda Q(Q^T\Lambda Q - R * R), -2(Q^T\Lambda Q - R * R) * R). \tag{10}$$

We now calculate the projection of $\nabla F(Q, R)$ on the manifold $\mathcal{O}(n) \times \mathcal{S}(n)$. Because we are using a product topology, the projection of $\nabla F(Q, R)$ on $\mathcal{O}(n) \times \mathcal{S}(n)$ is the direct product of the projections of the two components of $\nabla F(Q, R)$ on $\mathcal{O}(n)$

and $\mathcal{S}(n)$, respectively. Each of these projections can be calculated easily. In [8] we presented a simple way to do the projection on $\mathcal{O}(n)$: Since

$$R^{n\times n} = \mathcal{T}_Q\mathcal{O}(n) \oplus \mathcal{N}_Q\mathcal{O}(n) = Q\mathcal{S}(n)^{\perp} \oplus Q\mathcal{S}(n), \tag{11}$$

any matrix $A \in R^{n\times n}$ has a unique orthogonal splitting

$$A = Q\{\tfrac{1}{2}(Q^TA - A^TQ)\} + Q\{\tfrac{1}{2}(Q^TA + A^TQ)\} \tag{12}$$

as the sum of elements from $\mathcal{T}_Q\mathcal{O}(n)$ and $\mathcal{N}_Q\mathcal{O}(n)$. In particular, the projection of $2\Lambda Q(Q^T\Lambda Q - R*R)$ onto $\mathcal{T}_Q\mathcal{O}(n)$ is

$$\begin{aligned}&\tfrac{1}{2}Q\left\{Q^T[2\Lambda Q(Q^T\Lambda Q - R*R)] - [2\Lambda Q(Q^T\Lambda Q - R*R)]^TQ\right\}\\ &\quad = Q\left\{-Q^T\Lambda Q(R*R) + (R*R)Q^T\Lambda Q\right\}.\end{aligned} \tag{13}$$

On the other hand, $\mathcal{S}(n)$ is a vector space already, so the projection of $-2(Q^T\Lambda Q - R*R)*R$ onto $\mathcal{S}(n)$ is just itself. Thus we have found that the projection $g(Q,R)$ of $\nabla F(Q,R)$ onto the manifold $\mathcal{O}(n)\times\mathcal{S}(n)$ is given by the pair of matrices,

$$g(Q,R) = \left(Q\left\{-Q^T\Lambda Q(R*R) + (R*R)Q^T\Lambda Q\right\}, -2(Q^T\Lambda Q - R*R)*R\right). \tag{14}$$

The differential equation

$$\frac{d(Q,R)}{dt} = -g(Q,R), \tag{15}$$

therefore, defines a "steepest" descent vector field on $\mathcal{O}(n)\times\mathcal{S}(n)$ for the objective function $F(Q,R)$.

We now transport the flow (15) to the surface $\mathcal{M}(\Lambda)$ and the cone $\pi_s(R^n_+)$. For $Q(t)\in\mathcal{O}(n)$ and $R(t)\in\mathcal{S}(n)$, let

$$X(t) := Q(t)^T\Lambda Q(t), \tag{16}$$

$$Y(t) := R(t)*R(t). \tag{17}$$

Upon differentiating $X(t)$ and $Y(t)$ with respect to the variable t and using (15), we find that $X(t)$ and $Y(t)$ are governed by the differential system

$$\frac{dX}{dt} = [X,[X,Y]], \tag{18}$$

$$\frac{dY}{dt} = 4Y*(X-Y). \tag{19}$$

In (18) we have used the Lie bracket notation $[A,B] := AB - BA$. Together with an initial value $(X(0),Y(0))\in\mathcal{M}(\Lambda)\times\pi_s(R^n_+)$, we have reformulated Problem 4 as an initial value problem for (X,Y). The initial value problem is readily solved by available software.

The vector field on the right-hand sides of (18) and (19) is well defined for every $(X,Y)\in R^{n\times n}\times R^{n\times n}$. However, it is important to note that we intend to start the flow from an initial value $(X(0),Y(0))$ in $\mathcal{M}(\Lambda)\times\pi_s(R^n_+)$. Then $X(t)\in\mathcal{M}(\Lambda)$ and $Y(t)\in\pi_s(R^n_+)$ throughout the interval of existence. By the way these flows are

constructed, we know both $X(t)$ and $Y(t)$ are bounded and, hence, exist for $t \in [0, \infty)$. In fact, if we define

$$G(t) := F(X(t), Y(t)) = \tfrac{1}{2}\|X(t) - Y(t)\|^2 \geq 0, \tag{20}$$

then it is easy to calculate that

$$\frac{dG}{dt} = -\langle [X, Y], [X, Y] \rangle - 4\langle (X - Y), Y * (X - Y) \rangle \leq 0. \tag{21}$$

According to Lyapunov's second method [5, Thm. 5.5], the limit points of $(X(t), Y(t)) \in \mathcal{M}(\Lambda) \times \pi_s(R_+^n)$ must satisfy the equation $dG/dt = 0$. That is, $(\hat{X}, \hat{Y})$ will be a limit point only if

$$[\hat{X}, \hat{Y}] = 0 \tag{22}$$

and

$$\hat{Y} * (\hat{X} - \hat{Y}) = 0. \tag{23}$$

It is crucial to note from (18) and (19) that the conditions (22) and (23) are also sufficient for the condition that $(\hat{X}, \hat{Y})$ be an equilibrium point for the system. (In fact, if all eigenvalues in σ are distinct, then it can be shown that conditions (22) and (23) are also necessary.) Let

$$\mathcal{L} := \{(X, Y) \in \mathcal{M}(\Lambda) \times \pi_s(R_+^n) | [X, Y] = 0, Y * (X - Y) = 0\}. \tag{24}$$

We conclude that if we start with any $(X(0), Y(0)) \in \mathcal{M}(\Lambda) \times \pi_s(R_+^n)$, then the solution flow $(X(t), Y(t))$ approaches the set $\mathcal{L}$ as $t \longrightarrow \infty$ [5, Lemma 5.4]. That is, for every $\epsilon > 0$, there exists a $T > 0$ such that for every $t > T$ there exist a point $(\hat{X}, \hat{Y}) \in \mathcal{L}$ (possibly depending on t) such that $\|(X(t), Y(t)) - (\hat{X}, \hat{Y})\| < \epsilon$.

The above convergence result is not entirely satisfactory. For example, the flow $(X(t), Y(t))$ might oscillate around a nontrivial limit set. It will be interesting and important if we can show that the ω-limit set of any orbit $(X(t), Y(t))$ contains only a singleton. In the next section we shall use center manifold theory to prove that if the ω-limit set of an orbit $(X(t), Y(t))$ contains a point of the type $(\hat{X}, \hat{X})$, then $(X(t), Y(t))$ indeed converges to $(\hat{X}, \hat{X})$.

For computation, we obviously may choose $X(0) = \Lambda$. We note (using (19)) that if one component of $Y(t)$ is zero, then that component remains zero. For a feasible set of "generic" values, therefore, we should begin the flow $Y(t)$ with an interior point (i.e., a positive matrix) of the cone $\pi_s(R_+^n)$. Other than this restriction, the choice of $Y(0)$ is arbitrary. Different initial values of $Y(0)$ may lead to different limit points. We shall see some numerical examples in the last section.

Finally, we remark that a limit point $\hat{Y}$ of a flow $Y(t)$ could lie in one of the faces of $\pi_s(R_+^n)$ even if the flow starts from the interior of $\pi_s(R_+^n)$. We expect this situation when $\mathcal{M}(\Lambda) \cap \pi_s(R_+^n) = \emptyset$, i.e., when the given spectrum σ is not associated with any element of $\pi_s(R_+^n)$. But the most interesting case occurs when no component of the limit point $\hat{Y}$ is zero. Then, by (23), the symmetric nonnegative matrix $\hat{Y}$ must be the same as the isospectral matrix $\hat{X}$. In this case, we have numerically constructed a symmetric nonnegative matrix that has a prescribed spectrum.

3. Convergence. We have pointed out earlier that the ω-limit set of any orbit $(X(t), Y(t))$ is nonempty and invariant and that the orbit approaches its ω-limit set. In this section we shall take a closer look at the convergence behavior of the solution flow $(X(t), Y(t))$. We first use center manifold theory [6] to study the behavior of $(X(t), Y(t))$ near an equilibrium point. We then argue that the ω-limit set of any orbit $(X(t), Y(t))$ contains only a singleton.

Let $(\hat{X}, \hat{Y})$ be an equilibrium point of the system (18) and (19). If $\hat{X} \neq \hat{Y}$ or if $(\hat{X}, \hat{Y})$ does not belong to $\mathcal{M}(\Lambda) \times \pi_s(R_+^n)$, then we have not yet solved the inverse eigenvalue problem. We shall consider only the opposite case, namely, $\hat{X} = \hat{Y} \in \pi_s(R_+^n)$.

Our first approach is similar to the work done in [7]. For convenience, we first briefly review center manifold theory: Consider the system

$$\frac{dx}{dt} = Ax + f(x, y), \tag{25}$$

$$\frac{dy}{dt} = By + g(x, y), \tag{26}$$

where $x \in R^n$, $y \in R^m$, and A, B are constant matrices such that all eigenvalues of A have zero real parts while all those of B have negative real parts, the functions f and g are C^2 with $f(0,0) = 0, f'(0,0) = 0, g(0,0) = 0$, and $g'(0,0) = 0$. Then there exists an invariant manifold, called the center manifold, for the system (25) and (26). The center manifold is characterized by a C^2 function h from R^n to R^m with the property

$$y = h(x), \quad h(0) = 0, \quad h'(0) = 0. \tag{27}$$

Furthermore, the stability of $(0,0) \in R^n \times R^m$ for the system (25) and (26) is equivalent to the stability of $0 \in R^n$ for the system

$$\frac{dz}{dt} = Az + f(z, h(z)). \tag{28}$$

In addition, if $0 \in R^n$ is stable for (28), then with $(x(0), y(0))$ sufficiently small, there exists a solution $z(t)$ of (28) such that as $t \longrightarrow \infty$,

$$x(t) = z(t) + O(e^{-\mu t}), \tag{29}$$

$$y(t) = h(z(t)) + O(e^{-\mu t}) \tag{30}$$

for some constant $\mu > 0$.

We now apply these results to (18) and (19). Near an equilibrium point $(\hat{X}, \hat{X})$, we define

$$U(t) := X(t) - \hat{X}, \tag{31}$$

$$W(t) := X(t) - Y(t). \tag{32}$$

It is easy to see that (18) and (19) are equivalent to the following equations:

$$\frac{dU}{dt} = [\hat{X}, [W, \hat{X}]] + [U, [W, \hat{X}]] + [\hat{X}, [W, U]] + [U, [W, U]], \tag{33}$$

$$\frac{dW}{dt} = [\hat{X}, [W, \hat{X}]] - 4\hat{X} * W + [U, [W, \hat{X}]] + [\hat{X}, [W, U]] + [U, [W, U]] - 4W * (U - W). \tag{34}$$

Readers should distinguish between the linear and the nonlinear terms in each of the above expressions. We note that (33) is not quite in the same form as (25) since the linear term in (33) is in the variable W. But this discrepancy can easily be fixed through a simple linear transformation. Additionally, we are more interested in knowing whether $W(t)$ converges to zero than what $X(t)$ converges to. Thus we shall not be bothered to perform the transformation explicitly.

Since all underlying matrices are symmetric, it suffices to consider only the upper triangular parts of the matrices. Let Ω be the $n(n+1)/2 \times n(n+1)/2$ matrix representing the upper triangular part of the linear operator $[\hat{X},[W,\hat{X}]] - 4\hat{X} * W$. Applying center manifold theory, we first study the behavior of a solution flow near an equilibrium point.

LEMMA 3.1. *Suppose that all eigenvalues of Ω at an equilibrium point $(\hat{X},\hat{X})$ have negative real part. Then, starting with any matrix $(X(0),Y(0))$ sufficiently close to $(\hat{X},\hat{X})$, the solution flow $(X(t),Y(t))$ of (18) and (19) converges to a constant matrix of the form $(\hat{Z},\hat{Z})$. (Note that $\hat{Z}$ may not be the same as $\hat{X}$.)*

Proof. If all eigenvalues of Ω have negative real part, then obviously (see [6, Thm. 3, p. 5])

$$W \equiv h(U) \equiv 0 \tag{35}$$

is a center manifold for the system (33) and (34). It follows that the corresponding system (28) on the center manifold has constant solution. From (29) and (30), we conclude that $U(t)$ converges to a constant matrix while $W(t)$ converges to the zero matrix as $t \longrightarrow \infty$. We note that center manifold theory does not provide any information regarding which limit point $U(t)$ (and hence $X(t)$) is converging to, although it does guarantee that $X(t)$ and $Y(t)$ are converging to the same point.

The critical supposition that all eigenvalues of Ω have negative real part is difficult to justify in general. Even for the special case $n = 2$ to be discussed in the next section, the explicit expressions for eigenvalues of Ω are very complicated. Nonetheless, we have observed the following fact concerning this supposition.

LEMMA 3.2. *At any equilibrium point $(\hat{X},\hat{X}) \in \mathcal{M}(\Lambda) \times \pi_s(R_+^n)$, no eigenvalue of the corresponding Ω can have positive real part.*

Proof. We recall the definition $W(t) = X(t) - Y(t)$ and the fact that the differential equations (18) and (19) are designed to fulfill the specific purpose of reducing $\|X(t) - Y(t)\|$. Thus the Frobenius norm of the upper triangular part of $W(t)$ cannot grow as a function of t. Since $W(t)$ is related to its derivative by (34), the assertion follows.

In order that some eigenvalues of Ω have zero real parts, the components of $\hat{X}$ must satisfy certain algebraic equations. (Some examples are demonstrated in the next section.) The algebraic constraint, therefore, limits these special matrices, denoted by $(\tilde{X},\tilde{X})$, to a lower dimensional manifold in $\mathcal{M}(\Lambda) \times \pi_s(R_+^n)$. Forming a set of measure zero in the relative topology of $\mathcal{M}(\Lambda) \times \pi_s(R_+^n)$, points like $(\tilde{X},\tilde{X})$ should be regarded as nongeneric. Thus, for *almost all* equilibrium points of the kind $(\hat{X},\hat{X})$, all eigenvalues of the corresponding Ω have negative real part. Lemma 3.1, therefore, serves to explain the generic behavior of the dynamics of (18) and (19).

Near an equilibrium point of the kind $(\tilde{X},\tilde{X})$, the corresponding center manifold becomes much more complicated than (35). However, it can be proved that any flow, starting sufficiently close to $(\tilde{X},\tilde{X})$, still converges to a single point $(\tilde{Z},\tilde{Z})$. The proof is tedious but straightforward. We shall not give the full account of details here. But examples in the next section should illustrate our point.

It should be noted that Lemma 3.1 proves only a *local* convergence result. But we also know in the earlier discussion that the semiorbit of $(X(t), Y(t))$ approaches arbitrarily close to its ω-limit set which is a subset of all equilibrium points. These observations together imply that a solution flow $(X(t), Y(t))$ converges *globally* to a single point [1, Thm. 2.3]. Indeed, we have the following result.

LEMMA 3.3. *Let $(X(t), Y(t))$ be a solution flow of the differential system* (18) *and* (19). *Suppose $(\hat{X}, \hat{X})$ is an ω-limit point of the orbit $(X(t), Y(t))$ where all eigenvalues of the corresponding Ω have negative real parts. Then $(X(t), Y(t)) \longrightarrow (\hat{X}, \hat{X})$ as $t \longrightarrow \infty$.*

Proof. Since $(\hat{X}, \hat{X})$ is an ω-limit point of $(X(t), Y(t))$, there exists $T > 0$ such that $(X(T), Y(T))$ is sufficiently close to $(\hat{X}, \hat{X})$. By Lemma 3.1, the solution flow that begins at $(X(T), Y(T))$ converges to a single point $(\hat{Z}, \hat{Z})$. It follows that $(X(t), Y(t))$ converges to $(\hat{Z}, \hat{Z})$. Since $(\hat{X}, \hat{X})$ is an ω-limit point, it must be that $\hat{X} = \hat{Z}$.

In the above lemma, the assumption that all eigenvalues of Ω have negative real parts can be weakened. In fact, all we need in the proof of global convergence is the fact of local convergence to a single point. Thus, we restate the lemma as follows.

LEMMA 3.4. *Let $(X(t), Y(t)) \in \mathcal{M}(\Lambda) \times \pi_s(R_+^n)$ be a solution of the differential system* (18) *and* (19). *If $(\hat{X}, \hat{X})$ is an ω-limit point of this solution, then* $\lim_{t \to \infty}(X(t), Y(t)) = (\hat{X}, \hat{X})$.

We conclude this section with one final remark on Lemma 3.4. It is obvious that not every given set σ of n real values can be the spectrum of some nonnegative matrix. If a nonfeasible spectrum is given, we cannot expect the ω-limit set of any solution $(X(t), Y(t)) \in \mathcal{M}(\Lambda) \times \pi_s(R_+^n)$ to contain a point of the form $(\hat{X}, \hat{X})$. But even if σ is feasible, it is possible that an orbit $(X(t), Y(t))$ contains *no* limit point of the form $(\hat{X}, \hat{X})$. We have not analyzed this type of equilibrium points yet. In either case, however, our numerical experiment seems to suggest that the ω-limit set of $(X(t), Y(t))$ still contains a single point.

4. Stability analysis for $n = 2$. We shall now analyze the differential system (18) and (19) for the case $n = 2$ in detail. The answers to Problems 2 and 3 are obviously known for this simple case. But we hope the following study will provide some interesting insight into the understanding of the higher-dimensional case.

First we explain the geometry of $\mathcal{M}(\Lambda)$ and $\pi_s(R_+^n)$. Due to symmetry, it suffices to study the behavior of the six variables $(x_{11}, x_{12}, x_{22}; y_{11}, y_{12}, y_{22})$ only. We note that the set $\mathcal{O}(2)$ consists of two kinds of orthogonal matrices

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \qquad \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$

with $\theta \in [0, 2\pi)$. From (16), it follows that

$$\begin{aligned} x_{11} &= \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta, \\ x_{12} &= (\lambda_1 - \lambda_2)\cos\theta \sin\theta, \\ x_{22} &= \lambda_1 \sin^2\theta + \lambda_2 \cos^2\theta. \end{aligned} \tag{36}$$

These equations provide a parametric representation of the ellipse in R^3 formed by the intersection of the plane Π with equation $x_{11} + x_{22} = \lambda_1 + \lambda_2$ and the ellipsoid with equation $x_{11}^2 + 2x_{12}^2 + x_{22}^2 = \lambda_1^2 + \lambda_2^2$. If $\lambda_1 = \lambda_2$, then this ellipse is degenerate. Thus, the isospectral surface $\mathcal{M}(\Lambda)$ for $n = 2$ is represented by an ellipse. The distance from the plane Π to the origin is $\frac{1}{2}|\lambda_1 + \lambda_2|$. The cone $\pi_s(R_+^2)$ is the set of points in the

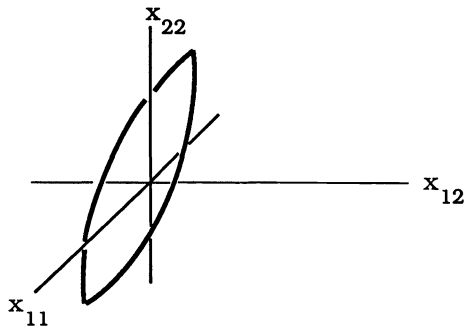


FIG. 1. *The set* $\mathcal{M}(\Lambda)$.

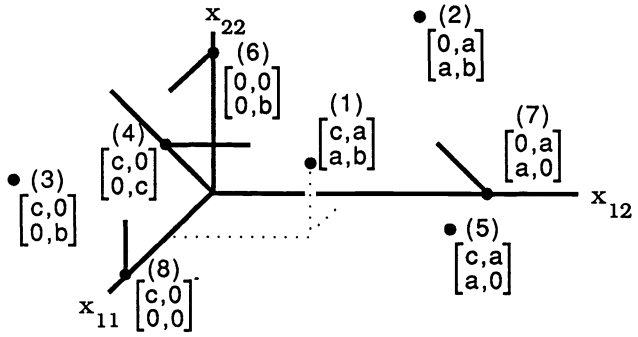


FIG. 2. *Representation of points in* $\mathcal{L}$.

first octant of R^3. The inverse eigenvalue problem will have a solution if and only if the ellipse intersects the first octant (see Fig. 1). It is clear from the geometry that this condition is equivalent to $\lambda_1 + \lambda_2 \geq 0$.

For $n = 2$, the set $\mathcal{L}$ defined in (24) contains points of the following eight types:

(1) $(y_{11}, y_{12}, y_{22}; y_{11}, y_{12}, y_{22})$; y_{ij} arbitrary but > 0;
(2) $(0, y_{12}, y_{22}; 0, y_{12}, y_{22})$; y_{ij} arbitrary but > 0;
(3) $(y_{11}, 0, y_{22}; y_{11}, 0, y_{22})$; y_{ij} arbitrary but > 0, $y_{11} \neq y_{22}$;
(4) $(y_{11}, x_{12}, y_{11}; y_{11}, 0, y_{11})$; x_{12}, y_{ij} arbitrary, but $y_{11} > 0$;
(5) $(y_{11}, y_{12}, 0; y_{11}, y_{12}, 0)$; y_{ij} arbitrary but > 0;
(6) $(x_{11}, 0, y_{22}; 0, 0, y_{22})$; x_{11}, y_{22} arbitrary, but $y_{22} > 0$;
(7) $(x_{11}, y_{12}, x_{11}; 0, y_{12}, 0)$; x_{11}, y_{12} arbitrary, but $y_{12} > 0$;
(8) $(y_{11}, 0, x_{22}; y_{11}, 0, 0)$; x_{22}, y_{11} arbitrary, but $y_{11} > 0$.

We note that the set $\mathcal{L}$ is the union [15] of one three-dimensional manifold (points of type (1)) and several two-dimensional manifolds (points of types (2)–(8)). The set $\mathcal{L}$ is represented in Fig. 2. For convenience, we have identified a representative for each type of point in Fig. 2. For example, the first open octant represents type (1) points; the first open quadrant in the yz-plane represents type (2) points, and so on. The extra lines sticking out from the coordinate axes or planes represent the freedom of variables for the matrix X. These are types (4), (6), (7), and (8) points, respectively.

For these types of limit points, note that the matrix Y is fixed to be the single point at the foot of these lines.

We consider the case when

$$\hat{X} = \hat{Y} = \begin{bmatrix} c & a \\ a & b \end{bmatrix}.$$

The corresponding matrix Ω in (19) is given by

$$\Omega = \begin{bmatrix} -4c - 2a^2 & 2ca - 2ba & 2a^2 \\ ca - ba & 2cb - c^2 - b^2 - 4a & -ca + ba \\ 2a^2 & -2ca + 2ba & -2a^2 - 4b \end{bmatrix}. \tag{37}$$

A general formula for eigenvalues of Ω is difficult to compute even with the help of a symbolic package. However, we already know from Lemma 3.2 that all eigenvalues of Ω have nonpositive real part. As an example, when $a = 2$, $b = 3$, and $c = 5$, we find the eigenvalues of Ω are approximately -35.53793480, -15.53700242, and -8.925062779. It can be seen easily from the characteristic polynomial that Ω will have two purely imaginary eigenvalues only if a, b, and c come from a very special two-dimensional hypersurface in R^3.

It turns out that some eigenvalues of Ω can be zero when $\hat{X}$ is on the faces or edges of the cone $\pi_s(R_+^n)$. In the following, we consider the local convergence for some of the following special cases.

Case 1. Type (8) limit point where

$$\tilde{X} = \tilde{Y} = \begin{bmatrix} c & 0 \\ 0 & 0 \end{bmatrix}.$$

In this case, $\Omega = \operatorname{diag}\{-4c, -c^2, 0\}$. So Lemma 3.1 cannot be applied. We give below a somewhat extended argument to demonstrate how the convergence of a flow near $(\tilde{X}, \tilde{X})$ can be reached. The differential system (33) and (34) becomes

(38)

$$\begin{aligned}
u_{11}' &= 2u_{12}w_{12}c + 2u_{12}w_{12}u_{11} + 2w_{22}u_{12}^2 - 2w_{11}u_{12}^2 - 2u_{12}w_{12}u_{22};\\
u_{12}' &= -w_{12}c^2 - 2w_{12}u_{11}c + 2u_{22}w_{12}c + cw_{11}u_{12} - cw_{22}u_{12} + u_{11}w_{11}u_{12}\\
&\quad +2u_{11}w_{12}u_{22} - w_{12}u_{11}^2 - u_{11}w_{22}u_{12} - w_{11}u_{12}u_{22} - w_{12}u_{22}^2 + w_{22}u_{12}u_{22};\\
u_{22}' &= -2u_{12}w_{12}c + 2w_{11}u_{12}^2 + 2u_{12}w_{12}u_{22} - 2u_{12}w_{12}u_{11} - 2w_{22}u_{12}^2;\\
w_{11}' &= -4w_{11}c + 2u_{12}w_{12}c + 2u_{12}w_{12}u_{11} + 2w_{22}u_{12}^2 - 2w_{11}u_{12}^2\\
&\quad -2u_{12}w_{12}u_{22} - 4w_{11}u_{11} + 4w_{11}^2;\\
w_{12}' &= -w_{12}c^2 - 2w_{12}u_{11}c + 2u_{22}w_{12}c + cw_{11}u_{12} - cw_{22}u_{12} + u_{11}w_{11}u_{12}\\
&\quad +2u_{11}w_{12}u_{22} - w_{12}u_{11}^2 - u_{11}w_{22}u_{12} - w_{11}u_{12}u_{22} - w_{12}u_{22}^2\\
&\quad +w_{22}u_{12}u_{22} - 4w_{12}u_{12} + 4w_{12}^2;\\
w_{22}' &= -2u_{12}w_{12}c + 2w_{11}u_{12}^2 + 2u_{12}w_{12}u_{22} - 2u_{12}w_{12}u_{11} - 2w_{22}u_{12}^2\\
&\quad -4w_{22}u_{22} + 4w_{22}^2.
\end{aligned}$$

We note there are two negative linear terms in w_{11}' and w_{12}'. So the convergence is contingent upon how the flow behaves on the center manifold. By center manifold theory, the center manifold of (38) is given by

$$(w_{11}, w_{12}) = h(u_{11}, u_{12}, u_{22}, w_{22}) \tag{39}$$

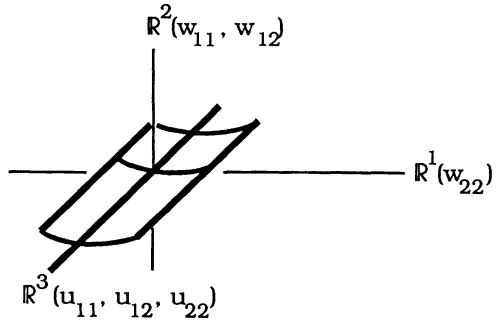


FIG. 3. *Four-dimensional center manifold.*

for some smooth function h. The geometry is illustrated in Fig. 3 where we use the x-axis to represent the three variables $(u_{11}, u_{12}, u_{22}) \in R^3$, the y-axis to represent the variable $w_{22} \in R$, and the z-axis to represent the two variables $(w_{11}, w_{12}) \in R^2$. We note that the x-axis where $W \equiv 0$ also represents the three-dimensional equilibrium points of type (1). Although h is difficult to compute explicitly, the following function can be shown to be an $O(\|(U, W)\|^3)$ approximation to h [6] near the origin:

$$
\begin{aligned}
w_{11} &:= \frac{u_{12}^2 w_{22}}{2c}; \\
w_{12} &:= \frac{-cu_{12}w_{22} - u_{11}u_{12}w_{22} + u_{12}u_{22}w_{22}}{c^2}.
\end{aligned} \tag{40}
$$

Upon substitution, we find the flow (28) on the center manifold is given by

$$
\begin{aligned}
u_{11}' &= -w_{22}u_{12}^2(-4u_{22}c + 4u_{11}c - 4u_{22}u_{11} + 2u_{11}^2 + u_{12}^2c + 2u_{22}^2)/c^2; \\
u_{12}' &= w_{22}u_{12}(-12u_{22}u_{11}c + 6u_{11}^2c + 4u_{11}c^2 + 6u_{22}^2c - 4u_{22}c^2 \\
&\qquad + u_{12}^2c^2 + u_{12}^2u_{11}c + 6u_{22}^2u_{11} - 6u_{11}^2u_{22} + 2u_{11}^3 \\
&\qquad - u_{12}^2u_{22}c - 2u_{22}^3)/(2c^2); \\
u_{22}' &= w_{22}u_{12}^2(-4u_{22}c + 4u_{11}c - 4u_{22}u_{11} + 2u_{11}^2 + u_{12}^2c + 2u_{22}^2)/c^2; \\
w_{22}' &= w_{22}(-4u_{12}^2u_{22}c + 4u_{12}^2u_{11}c + u_{12}^4c + 2u_{12}^2u_{22}^2 - 4u_{12}^2u_{22}u_{11} \\
&\qquad + 2u_{12}^2u_{11}^2 - 4u_{22}c^2 + 4w_{22}c^2)/c^2.
\end{aligned} \tag{41}
$$

The most dominant term in (41) is

$$
w_{22}' = 4w_{22}(w_{22} - u_{22}) + \text{ higher-order terms.} \tag{42}
$$

It is also obvious that

$$
(w_{22} - u_{22})' = 4w_{22}(w_{22} - u_{22}). \tag{43}
$$

We are interested only in the case where $Y(t) \in \pi_s(R_+^n)$. Therefore, $w_{22} = x_{22} - y_{22} = u_{22} - y_{22} \leq u_{22}$ since $y_{22} \geq 0$. By checking the signs of the right-hand sides of (42) and (43), we find that the projection of the vector field (41) at any point $(u_{11}, u_{12}, u_{22}, w_{22}) \in R^4$ onto the (u_{22}, w_{22})-plane must be within the shaded region as shown in Fig. 4. It is obvious from the geometry that $u_{22}(t)$ converges to a fixed

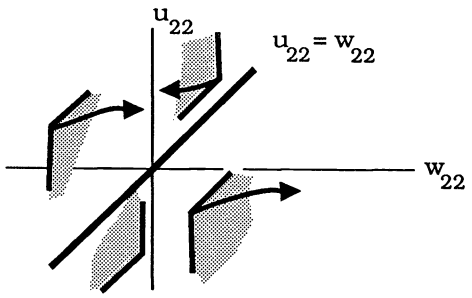


FIG. 4. *Projection of flows.*

point and $w_{22}(t)$ converges to zero as t converges to infinity. In other words, we have shown that near a equilibrium point $(\tilde{X}, \tilde{X})$ of type (8), the solution flow $(X(t), Y(t))$ with $Y(0) \in \pi_s(R_+^n)$ converges to a fixed point of the form $(\tilde{Z}, \tilde{Z})$. This should manifest our point made in the preceding section concerning the convergence when Ω has zero eigenvalues. In Fig. 4 we have also drawn the projection of vector field of (41) when $Y(0)$ is not in $\pi_s(R_+^n)$. This corresponds to the region below the diagonal $u_{22} = w_{22}$. It is interesting to note that $w_{22}(t)$ may diverge to infinity. This is because the differential system (18) and (19) has the descending property only if $Y(t) \in \pi_s(R_+^n)$.

Case 2. Type (3) limit points where

$$\hat{X} = \hat{Y} = \begin{bmatrix} c & 0 \\ 0 & b \end{bmatrix}$$

and $b \neq c$. We find $\Omega = \operatorname{diag}\{-4c, -(b-c)^2, -4b\}$. Thus Lemma 3.1 can be applied.

Case 3. Type (7) limit points where

$$\tilde{X} = \tilde{Y} = \begin{bmatrix} 0 & a \\ a & 0 \end{bmatrix}.$$

At such a limit point, the matrix Ω has eigenvalues $\{0, -4a, -4a^2\}$. An argument similar to the one given in Case 1 can be made. The center manifold should become more complicated because the eigenvectors, $[1,0,1]^T, [0,1,0]^T$, and $[1,0,-1]^T$, of Ω indicate that there are couplings between components. Instead of using center manifold theory, we now take a geometric viewpoint to study this limit point. Limit points of Type (7) have a unique feature that makes them special—that is, the ellipse $\mathcal{M}(\Lambda)$ containing $(\tilde{X}, \tilde{X})$ intersects the first octant only at $(\tilde{X}, \tilde{X})$. Therefore, the ellipse corresponding to a slightly perturbed spectrum, say $\sigma = \{a - \epsilon, -a\}$ with $\epsilon > 0$, will not intersect the first octant at all. This observation perhaps explains why we experience some numerical difficulty in constructing the second example in [18] by our method. We shall report this difficulty in the next section.

Case 4. Type (2) limit points where

$$\hat{X} = \hat{Y} = \begin{bmatrix} 0 & a \\ a & b \end{bmatrix}.$$

It can be checked that the characteristic polynomial of Ω is given by $p(\lambda) = \lambda^3 + (4a^2 + b^2 + 4a + 4b)\lambda^2 + (4b^3 + 16a^3 + 8a^2b + 16ab)\lambda + 32a^3b$. Clearly, $p(\lambda)$ must have

one negative real root. The other two roots can be purely imaginary numbers only if $(4a^2 + b^2 + 4a + 4b)(4b^3 + 16a^3 + 8a^2b + 16ab) = 32a^3b$. We therefore conclude that for almost all values of a and b, all three roots of $p(\lambda)$ have negative real part.

5. Numerical results. In this section we briefly report some of our numerical experiments with the differential equations (18) and (19).

We use the subroutine ODE in [17] as the integrator. Both local control parameters ABSERR and RELERR are set to be 10^{-12}. This criterion is used to control the accuracy in following the solution path. We examine the output values at time interval of 1. Normally, we should expect the loss of one or two digits in the global error. Thus, when the norm of the difference between two consecutive output points becomes less than 10^{-9}, we assume the path has converged to an equilibrium point. The execution is then terminated automatically. We always use $X(0) = \Lambda$ as the starting value for $X(t)$.

Example 1. We consider the spectrum $\sigma = \{5, 0, -2, -2\}$ which satisfies the so-called condition (K) in [10]. Let E denote the matrix whose components are all 1's. We report below various choices of $Y(0)$ and the corresponding approximate equilibrium points $\hat{Y}$. We also report the approximate length of t for convergence. These lengths may depend on the initial values and the integrators, but they should be independent of the computing machine.

(a) $Y(0) = E$, $\hat{Y} \approx Y(120) \approx$

$$\begin{bmatrix} .3035817600D{+}0 & .5849737957D{+}0 & .2062366873D{+}1 & .2062366873D{+}1 \\ .5849737957D{+}0 & .5568227490D{+}0 & .1257189377D{+}1 & .1257189377D{+}1 \\ .2062366873D{+}1 & .1257189377D{+}1 & .6979774550D{-}1 & .2069797746D{+}1 \\ .2062366873D{+}1 & .1257189377D{+}1 & .2069797746D{+}1 & .6979774550D{-}1 \end{bmatrix};$$

(b) $Y(0) = 2E$, $\hat{Y} \approx Y(120) \approx$

$$\begin{bmatrix} .2384681793D{+}0 & .5682102400D{+}0 & .2008259297D{+}1 & .2008259297D{+}1 \\ .5682102400D{+}0 & .6025576213D{+}0 & .1336108181D{+}1 & .1336108181D{+}1 \\ .2008259297D{+}1 & .1336108181D{+}1 & .7948709968D{-}1 & .2079487100D{+}1 \\ .2008259297D{+}1 & .1336108181D{+}1 & .2079487100D{+}1 & .7948709968D{-}1 \end{bmatrix};$$

(c) $Y(0) = 12E$, $\hat{Y} \approx Y(120) \approx$

$$\begin{bmatrix} .2054687518D{+}0 & .5725349916D{+}0 & .1995497108D{+}1 & .1995497108D{+}1 \\ .5725349916D{+}0 & .6259762947D{+}0 & .1349043940D{+}1 & .1349043940D{+}1 \\ .1995497108D{+}1 & .1349043940D{+}1 & .8427747670D{-}1 & .2084277477D{+}1 \\ .1995497108D{+}1 & .1349043940D{+}1 & .2084277477D{+}1 & .8427747670D{-}1 \end{bmatrix};$$

(d) $Y(0) = 200E$, $\hat{Y} \approx Y(120) \approx$

$$\begin{bmatrix} .2002571961D{+}0 & .5800168336D{+}0 & .2000122546D{+}1 & .2000122546D{+}1 \\ .5800168336D{+}0 & .6302391750D{+}0 & .1339883371D{+}1 & .1339883371D{+}1 \\ .2000122546D{+}1 & .1339883371D{+}1 & .8475181443D{-}1 & .2084751814D{+}1 \\ .2000122546D{+}1 & .1339883371D{+}1 & .2084751814D{+}1 & .8475181442D{-}1 \end{bmatrix};$$

(e) $Y(0) =$ a randomly generated symmetric positive matrix, $\hat{Y} \approx Y(240) \approx$

$$\begin{bmatrix} .5806553062D{+}0 & .8075689805D{+}0 & .2354051323D{+}1 & .2276884031D{+}1 \\ .8075689805D{+}0 & .5285565377D{-}1 & .1981177836D{+}0 & .4150231898D{+}0 \\ .2354051323D{+}1 & .1981177836D{+}0 & .3084574129D{+}0 & .2165951279D{+}1 \\ .2276884031D{+}1 & .4150231898D{+}0 & .2165951279D{+}1 & .5803162721D{-}1 \end{bmatrix}.$$

It is interesting to note that different choices of $Y(0)$ lead to different equilibrium points. It is also interesting to note that it takes about the same length of integration to reach convergence, although this observation is not conclusive.

Example 2. We consider the spectrum $\sigma = \{3 - t, 1 + t, -1, -1, -1, -1\}$ for $0 < t < 1$. It can be checked easily that the sufficient condition (K) in [10] is not satisfied. But in [18] it is proved that the set σ is indeed the spectrum of the nonnegative matrix

$$N := \begin{bmatrix} A & B \\ B & A \end{bmatrix},$$

where

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

and

$$B = \frac{(1-t)}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Considering $t = \frac{1}{2}$, we find the matrix $\Omega \in R^{21\times 21}$ at the point $(\tilde{X}, \tilde{X})$ with $\tilde{X} = N$ has one zero eigenvalue. Thus, this example is one of the exceptional cases we mentioned earlier. Furthermore, the sum of elements of σ is equal to zero for every value of t. A slight perturbation of σ, therefore, may make the spectrum unfeasible. We think this is a situation similar to Case 3 discussed in the previous section. Indeed, the matrix N has zeros on its diagonal, which indicates that N is at the intersection of n faces of $\pi_s(R^6_+)$.

(a) Suppose $Y(0) = N$. It can be calculated that $\|X(0) - Y(0)\| = 5$. We find $\hat{Y} \approx Y(130) \approx$

$$\begin{bmatrix} .0000D{+}0 & .9928D{+}0 & .9978D{+}0 & .2279D{+}0 & .2279D{+}0 & .2279D{+}0 \\ .9928D{+}0 & .0000D{+}0 & .9985D{+}0 & .1096D{+}0 & .1096D{+}0 & .1096D{+}0 \\ .9978D{+}0 & .9985D{+}0 & .0000D{+}0 & .1631D{+}0 & .1631D{+}0 & .1631D{+}0 \\ .2279D{+}0 & .1096D{+}0 & .1631D{+}0 & .0000D{+}0 & .1000D{+}1 & .1000D{+}1 \\ .2279D{+}0 & .1096D{+}0 & .1631D{+}0 & .1000D{+}1 & .0000D{+}0 & .1000D{+}1 \\ .2279D{+}0 & .1096D{+}0 & .1631D{+}0 & .1000D{+}1 & .1000D{+}1 & .0000D{+}0 \end{bmatrix},$$

where $\|X(130) - Y(130)\| \approx 7.9105 \times 10^{-11}$. We note this nonnegative matrix is different from the one constructed in [18] even though $Y(0)$ itself is already a solution to Problem 3.

(b) Suppose $Y(0) = E$. Then $\|X(0) - Y(0)\| \approx 6.9642$. We are surprised to find that our flow does not converge to an equilibrium point of the form $(\hat{X}, \hat{X})$, but rather $X(t)$ converges to $\hat{X} \approx$

$$\begin{bmatrix} -.3000D{+}0 & -.2041D{-}3 & .7000D{+}0 & .7000D{+}0 & .7000D{+}0 & .7000D{+}0 \\ -.2041D{-}3 & .1500D{+}1 & .5867D{-}3 & .5867D{-}3 & .5867D{-}3 & .5867D{-}3 \\ .7000D{+}0 & .5867D{-}3 & -.3000D{+}0 & .7000D{+}0 & .7000D{+}0 & .7000D{+}0 \\ .7000D{+}0 & .5867D{-}3 & .7000D{+}0 & -.3000D{+}0 & .7000D{+}0 & .7000D{+}0 \\ .7000D{+}0 & .5867D{-}3 & .7000D{+}0 & .7000D{+}0 & -.3000D{+}0 & .7000D{+}0 \\ .7000D{+}0 & .5867D{-}3 & .7000D{+}0 & .7000D{+}0 & .7000D{+}0 & -.3000D{+}0 \end{bmatrix},$$

while $Y(t)$ converges to $\hat{Y} \approx$

$$\begin{bmatrix} .5909-307 & .3325D-6 & .7000D+0 & .7000D+0 & .7000D+0 & .7000D+0 \\ .3325D-6 & .1500D+1 & .6963D-3 & .6963D-3 & .6963D-3 & .6963D-3 \\ .7000D+0 & .6963D-3 & .7348-307 & .7000D+0 & .7000D+0 & .7000D+0 \\ .7000D+0 & .6963D-3 & .7000D+0 & .7348-307 & .7000D+0 & .7000D+0 \\ .7000D+0 & .6963D-3 & .7000D+0 & .7000D+0 & .7348-307 & .7000D+0 \\ .7000D+0 & .6963D-3 & .7000D+0 & .7000D+0 & .7000D+0 & .7348-307 \end{bmatrix}$$

with $\|X(9120) - Y(9120)\| \approx .6708$. It is interesting to note that the eigenvalues of $\hat{Y}$ are $\{2.8, 1.5, -.7, -.7 - .7, -.7\}$. The true equilibrium point $(\hat{X}, \hat{Y})$ is where all the small components in the second row and the second column except the $(2,2)$-position of the above two matrices are zero. We have observed that all the significant components of $(\hat{X}, \hat{Y})$ are reached as early as $t \approx 500$. The overall slow convergence is due to the slow rate of change of components in the second row and the second column. To see this, we rerun the code by choosing

$$Y(0) = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

so that the small components become zero. Then the corresponding orbit converges to the limit point within $t \approx 80$. This example also illustrates that an orbit $(X(t), Y(t))$ may not necessarily have a limit point of the form $(\hat{X}, \hat{X})$.

(c) Suppose $Y(0) = 12E$. Then $\|X(0) - Y(0)\| \approx 72.0868$. Again, we find that $X(t)$ converges to

$$\begin{bmatrix} -.3000D+0 & -.2487D-3 & .7000D+0 & .7000D+0 & .7000D+0 & .7000D+0 \\ -.2487D-3 & .1500D+1 & .7158D-3 & .7158D-3 & .7158D-3 & .7158D-3 \\ .7000D+0 & .7158D-3 & -.3000D+0 & .7000D+0 & .7000D+0 & .7000D+0 \\ .7000D+0 & .7158D-3 & .7000D+0 & -.3000D+0 & .7000D+0 & .7000D+0 \\ .7000D+0 & .7158D-3 & .7000D+0 & .7000D+0 & -.3000D+0 & .7000D+0 \\ .7000D+0 & .7158D-3 & .7000D+0 & .7000D+0 & .7000D+0 & -.3000D+0 \end{bmatrix}$$

while $Y(t)$ converges to

$$\begin{bmatrix} .7206-307 & .7133D-6 & .7000D+0 & .7000D+0 & .7000D+0 & .7000D+0 \\ .7133D-6 & .1500D+1 & .8494D-3 & .8494D-3 & .8494D-3 & .8494D-3 \\ .7000D+0 & .8494D-3 & .7205-307 & .7000D+0 & .7000D+0 & .7000D+0 \\ .7000D+0 & .8494D-3 & .7000D+0 & .7205-307 & .7000D+0 & .7000D+0 \\ .7000D+0 & .8494D-3 & .7000D+0 & .7000D+0 & .7205-307 & .7000D+0 \\ .7000D+0 & .8494D-3 & .7000D+0 & .7000D+0 & .7000D+0 & .7205-307 \end{bmatrix}$$

with $\|X(7140) - Y(7140)\| \approx .6708$. We believe the true limit point is the same as the one in (b).

Example 3. We consider the spectrum $\sigma = \{11, -3, -2, -2, -1, -1\}$ which satisfies a sufficient condition in [10, Thm. 2.4]. Then,

(a) With $Y(0) = 2E$, then $\|X(0) - Y(0)\| \approx 16.6132$ and $\hat{Y} \approx Y(70) \approx$

$$\begin{bmatrix} .5514D+0 & .1920D+1 & .2194D+1 & .1620D+1 & .1551D+1 & .1920D+1 \\ .1920D+1 & .2243D+0 & .2550D+1 & .2265D+1 & .1920D+1 & .2224D+1 \\ .2194D+1 & .2550D+1 & .1158D+0 & .3058D+1 & .2194D+1 & .2550D+1 \\ .1620D+1 & .2265D+1 & .3058D+1 & .3328D+0 & .1620D+1 & .2265D+1 \\ .1551D+1 & .1920D+1 & .2194D+1 & .1620D+1 & .5514D+0 & .1920D+1 \\ .1920D+1 & .2224D+1 & .2550D+1 & .2265D+1 & .1920D+1 & .2243D+0 \end{bmatrix}$$

with $\|X(70) - Y(70)\| \approx 4.3653 \times 10^{-10}$.

(b) With $Y(0) = 12E$, then $\|X(0) - Y(0)\| \approx 72.6360$ and $\hat{Y} \approx Y(70) \approx$

$$\begin{bmatrix} .5741D{+}0 & .1946D{+}1 & .2206D{+}1 & .1604D{+}1 & .1574D{+}1 & .1946D{+}1 \\ .1946D{+}1 & .2292D{+}0 & .2550D{+}1 & .2230D{+}1 & .1946D{+}1 & .2229D{+}1 \\ .2206D{+}1 & .2550D{+}1 & .1100D{+}0 & .3025D{+}1 & .2206D{+}1 & .2550D{+}1 \\ .1604D{+}1 & .2230D{+}1 & .3025D{+}1 & .2833D{+}0 & .1604D{+}1 & .2230D{+}1 \\ .1574D{+}1 & .1946D{+}1 & .2206D{+}1 & .1604D{+}1 & .5741D{+}0 & .1946D{+}1 \\ .1946D{+}1 & .2229D{+}1 & .2550D{+}1 & .2230D{+}1 & .1946D{+}1 & .2292D{+}0 \end{bmatrix}$$

with $\|X(70) - Y(70)\| \approx 6.0561 \times 10^{-10}$.

REFERENCES

[1] B. AULBACH, *Continuous and Discrete Dynamics near Manifolds of Equilibria*, Lecture Notes in Mathematics, vol. 1058, Springer-Verlag, Berlin, 1984.

[2] W. W. BARRETT AND C. R. JOHNSON, *Possible spectra of totally positive matrices*, Linear Algebra Appl., 62 (1984), pp. 231–233.

[3] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[4] M. BOYLE AND D. HANDELMAN, *The spectra of nonnegative matrices via symbolic dynamics*, Ann. Math., 133 (1991), pp. 249–316.

[5] F. BRAUER, AND J. A. NOHEL, *Qualitative Theory of Ordinary Differential Equations*, Benjamin, New York, 1969.

[6] J. CARR, *Applications of Centre Manifold Theory*, Applied Mathematical Sciences, Vol. 35, Springer-Verlag, Berlin, 1981.

[7] M. T. CHU, *The generalized Toda flow, the* QR *algorithm and the center manifold theory*, SIAM J. Algebraic Discrete Methods, 5 (1984), pp. 187–201.

[8] M. T. CHU, AND K. R. DRIESSEL, *The projected gradient method for least squares matrix approximations with spectral constraints*, SIAM J. Numer. Anal., 27 (1990), pp. 1050–1060.

[9] K. R. DRIESSEL AND M. T. CHU, *Can real symmetric Toeplitz matrices have arbitrary spectra?*, SIAM J. Matrix Anal., submitted.

[10] M. FIEDLER, *Eigenvalues of nonnegative symmetric matrices*, Linear Algebra Appl., 9 (1974), pp. 119–142.

[11] S. FRIEDLAND, *On an inverse problem for nonnegative and eventually nonnegative matrices*, Israel J. Math., 29 (1978), pp. 43–60.

[12] S. FRIEDLAND AND A. A. MELKMAN, *On the eigenvalues of nonnegative Jacobi matrices*, Linear Algebra Appl., 25 (1979), pp. 239–254.

[13] D. HERSHKOWITS, *Existence of matrices with prescribed eigenvalues and entries*, Linear and Multilinear Algebra, 14 (1983), pp. 315–342.

[14] R. LOEWY AND D. LONDON, *A note on an inverse problems for nonnegative matrices*, Linear and Multilinear Algebra, 6 (1978), pp. 83–90.

[15] J. MILNOR, *Singular Points of Complex Hypersurfaces*, Ann. of Math. Stud., Vol. 61, Princeton University Press, Princeton, NJ, 1968.

[16] G. N. OLIVEIRA, *Nonnegative matrices with prescribed spectrum*, Linear Algebra Appl., 54 (1983), pp. 117–121.

[17] L. F. SHAMPINE AND M. K. GORDON, *Computer Solution of Ordinary Differential Equations: The Initial Value Problems*, Freeman, San Francisco, 1975.

[18] G. W. SOULES, *Constructing symmetric nonnegative matrices*, Linear and Multilinear Algebra, 13 (1983), pp. 241–251.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1388-1410, September 1991

012

TWO-SCALE DIFFERENCE EQUATIONS I. EXISTENCE AND GLOBAL REGULARITY OF SOLUTIONS*

INGRID DAUBECHIES†‡ AND JEFFREY C. LAGARIAS†

Abstract. A *two-scale difference equation* is a functional equation of the form $f(x)=\sum_{n=0}^{N} c_n f(\alpha x-\beta_n)$, where $\alpha>1$ and $\beta_0<\beta_1<\cdots<\beta_n$ are real constants, and c_n are complex constants. Solutions of such equations arise in spline theory, in interpolation schemes for constructing curves, in constructing wavelets of compact support, in constructing fractals, and in probability theory. This paper studies the existence and uniqueness of L^1-solutions to such equations. In particular, it characterizes L^1-solutions having compact support. A time-domain method is introduced for studying the special case of such equations where $\{\alpha, \beta_0, \cdots, \beta_n\}$ are integers, which are called *lattice two-scale difference equations.* It is shown that if a lattice two-scale difference equation has a compactly supported solution in $C^m(\mathbb{R})$, then $m<(\beta_n-\beta_0)/(\alpha-1)-1$.

Key words. wavelets, subdivision algorithms, fractals

AMS(MOS) subject classifications. 26A15, 26A18, 39A10, 42A05

1. Introduction.

A *two-scale difference equation* is a functional equation of the form

$$f(x)=\sum_{n=0}^{N} c_n f(\alpha x-\beta_n) \tag{1.1}$$

where $\alpha>1$ and $\beta_0<\beta_1<\cdots<\beta_n$ are real constants, and x takes real values, while the c_n are complex constants. The right side of (1.1) is typical for difference equations, and the name two-scale difference equation reflects the fact that (1.1) relates translates of scaled versions of the same function, involving two different scales. A *lattice two-scale difference equation* is the special case where α and all β_n are integers, i.e.,

$$f(x)=\sum_{n=0}^{N} c_n f(kx-n) \tag{1.2}$$

where $k\geqq 2$ is an integer. The apparently more general equation

$$f(x)=\sum_{n=-N_1}^{N_2} c_n f(kx-n) \tag{1.3}$$

can be reduced to the form (1.2) by the change of variable $y=x-N_1/(k-1)$.

This paper and its sequel (Daubechies and Lagarias (1988), hereafter called part II) study L^1-solutions of two-scale difference equations, and of lattice two-scale difference equations in particular. The basic questions concern the existence, uniqueness, and degree of regularity of solutions for a given equation. We treat in detail L^1-solutions having compact support. In fact, two-scale difference equations always have solutions in the sense of distributions and may also possess functions not in $L^1(\mathbb{R})$ as solutions, e.g., if $\sum_{n=0}^{N} c_n=1$, then the constant functions are solutions. However, only for special sets of $\{\alpha, \beta_n, c_n\}$ will (1.1) have any nonzero L^1-solutions.

Functions that satisfy lattice two-scale difference equations arise in several different contexts. G. de Rham is credited with an example of a continuous,

* Received by the editors November 30, 1988; accepted for publication (in revised form) September 5, 1990.

† AT&T Bell Laboratories, Murray Hill, New Jersey 07974.

‡ "Bevoegdverklaard Navorser" at the Belgian National Foundation for Scientific Research (on leave); on leave also from the Department of Theoretical Physics, Vrije Universiteit Brussel, Belgium.

nowhere-differentiable function which satisfies (1.2) with $k=3$ and $c_0=1$, $c_1=c_{-1}=\frac{1}{3}$, $c_2=c_{-2}=\frac{2}{3}$, and all other $c_n\equiv 0$. (This was communicated to us by Meyer (1987). We have not found a direct reference to this function in de Rham's papers but similar functions appear in de Rham (1947), (1956), (1957), (1959).) Such functions also arise as limits of "uniform subdivision schemes" for constructing curves and surfaces. As observed and generalized in the work of Dahmen and Micchelli (1984), (1988) and Micchelli and Prautzsch (1987a), (1987b), (1989), normalized B-splines and d-dimensional box splines each satisfy a lattice two-scale difference equation with $k=2$. They point out that this two-scale property is really the basic ingredient in a subdivision algorithm for numerically evaluating B-spline curves and surfaces, given by Lane and Riesenfeld (1980). This can be exploited to define and study other subdivision schemes for the design of curves and surfaces, also characterized by a lattice two-scale equation (Cavaretta and Micchelli (1989)). Dyn and Levin (1989) similarly link subdivision algorithms with the study of a lattice two-scale equation. Dubuc (1986) proposed a dyadic interpolation scheme where the "fundamental function" satisfies an equation of type (1.3) with $k=2$. For special values of the parameters, he proved smoothness results of this fundamental function. Dyn, Gregory, and Levin (1987) independently and by different techniques proved similar results for the same dyadic interpolation schemes. In Deslauriers and Dubuc (1987) this interpolation scheme was applied to the construction of fractal objects and functions with fractal properties. Deslauriers and Dubuc (1989) extend the dyadic interpolation scheme to other integer values of k; they use the properties of solutions of (1.3) corresponding to specific values of the c_n to study Lagrange iterative interpolation processes. In another field, Daubechies (1988) constructed orthonormal bases of compactly supported wavelets, i.e., orthonormal bases $\{h_{mn}(x)\}$ of $L^2(\mathbb{R})$ generated by translating and dilating a single compactly supported function h via

$$h_{mn}(x)=2^{-m/2}h(2^{-m}x-n).$$

The construction of such h requires an auxiliary function which is a solution of a lattice two-scale difference equation, and our interest in these equations arose from these functions. All of these examples actually involve L^1-solutions having compact support.

Solutions of general two-scale difference equations (1.1) arise in other areas of mathematics as well. Kershner and Wintner (1935) considered symmetric Bernoulli convolutions $d\lambda(x,\beta)$ whose Fourier transform $\Lambda(u,\beta)=\int_{-\infty}^{\infty} e^{iux}\,d\lambda(x,\beta)$ has

$$\Lambda(u,\beta)=\prod_{n=0}^{\infty}\cos(\beta^n u). \tag{1.4}$$

For certain values of β in $(0,1)$ the measure $d\lambda(x,\beta)=\lambda'(x,\beta)\,dx$ is absolutely continuous, and $\lambda'(x,\beta)$ then satisfies the two-scale difference equation

$$\lambda'(x)=\frac{\beta}{2}\left(\lambda'\left(\frac{1}{\beta}x-1\right)+\lambda'\left(\frac{1}{\beta}x+1\right)\right).$$

Smoothness properties of this and related Bernoulli convolutions were studied by Jessen and Wintner (1935), Erdös (1939), (1940), Garsia (1962), and Brown and Moran (1973). It remains a difficult open problem to characterize the set of β for which $d\lambda(x,\beta)$ is absolutely continuous. More recently, Barnsley and Demko (1985, Ex. 21) in studying iterated function systems construct a function $f_\beta(z)$ defined on $\mathbb{C}-A$ where

A is the Cantor set, satisfying the two-scale difference equation

$$f_\beta(z) = \frac{3^\beta}{2}\,(f_\beta(3z-2)+f_\beta(3z)).$$

Here $f_\beta(z) = \int_A d\mu(x)/(z-x)^\beta$, where $d\mu$ is uniform measure on the Cantor set.

We use two methods for studying these equations. They are a Fourier transform approach that applies to the general equation (1.1), and a time-domain approach described further below that applies only to lattice two-scale difference equations. Part I describes Fourier transform results on existence and uniqueness, introduces the time-domain approach, and uses it to establish bounds on the smoothness of L^1-solutions of compact support. Part II studies the time-domain construction in detail and gives sufficient conditions for the existence of nonzero continuous solutions of compact support, and determines their local and global regularity properties.

The Fourier transform provides a method for the study of L^1-solutions f of general two-scale difference equations (1.1). The convolution character of the right side of (1.1) leads to an infinite product expansion for the Fourier transform $\hat{f}(u)$ permitting detailed study. Section 2 uses this approach to obtain existence and uniqueness results for L^1-solutions to (1.1). These depend in a crucial way on the quantity

$$\Delta = \alpha^{-1} \sum_{m=0}^{N} c_n. \tag{1.5}$$

There are no nonzero L^1-solutions if $|\Delta| < 1$ or if $|\Delta| = 1$ and $\Delta \neq 1$. The case of most interest is $\Delta = 1$; it has at most one nonzero L^1-solution, up to a multiplicative scale factor. This solution, if it exists, is of compact support with support $(f) \subset [0, (\alpha-1)^{-1}N]$, and has $\int_{-\infty}^{\infty} f(x)\,dx \neq 0$. For $|\Delta| > 1$ it is possible to have zero, one, or infinitely many L^1-solutions, which need not have compact support, depending on the values $\{\alpha, \beta_n, c_n\}$.

Section 3 studies L^1-functions of compact support solving (1.1), and shows that they are all derived from solutions of the case $\Delta = 1$, in the following sense. If a two-scale difference equation (1.1) has a nonzero L^1-solution f of compact support, then it is unique (up to normalization), and necessarily,

(1) $\Delta = \alpha^m$ for some nonnegative integer m;

(2) The two-scale difference equation with $\Delta = 1$ obtained by replacing the coefficients $\{c_n\}$ with $\{\alpha^{-m}c_n\}$ has a nonzero L^1-solution g of compact support, and for a suitable choice of normalization,

$$\frac{d^m}{dx^m}\, g(x) \equiv f(x) \quad \text{a.e;}$$

The remainder of part I and part II use time-domain methods that apply only to the special case of lattice two-scale difference equations (1.2). This approach exploits the special feature that lattice two-scale difference equations make sense when restricted to the discrete domain $\mathbb{Z}$. Suppose we are given data $\{f(n);\, n \in \mathbb{Z}\}$ which satisfy

$$f(x) = \sum_{n=0}^{N} c_n f(kx-n) \tag{1.6}$$

for all $x \in \mathbb{Z}$. The functional equation then determines $f(x)$ for $x \in \mathbb{Z}/k$, and by iteration for $x \in \bigcup_{n=1}^{\infty} \mathbb{Z}/k^n$. In particular, such data $\{f(n);\, n \in \mathbb{Z}\}$ can be interpolated by at most one continuous solution of (1.6). This approach thus applies most naturally to the problem of finding *continuous* solutions of (1.6). Here we have the two subproblems

of finding solutions to (1.6) on $\mathbb{Z}$, and then determining conditions under which such solutions interpolate to solutions on $\mathbb{R}$.

The time-domain approach applies particularly well in the study of compactly supported continuous solutions, since we then know that $\{f(n);\ n \in \mathbb{Z}\}$ has $f(n) = 0$ off the finite set $0 \leqq n \leqq N/(k-1)$, and the set of solutions on $\mathbb{Z}$ to (1.6) is a finite-dimensional vector space. The iterative process of recursively solving (1.6) on $\{\mathbb{Z}/k^n;\ n = 1, 2, \cdots\}$ can be encoded using products of a finite set of matrices, as is explained in part II, and this provides a vehicle for studying convergence and smoothness of solutions. Such an approach was initiated by Micchelli and Prautzsch (1987a), as we discovered after completing this work.

In the rest of part I we apply the time-domain approach to obtain information about compactly supported solutions in the case where $\Delta = 1$, which by the results of § 3 is essentially the most general case.

Section 4 obtains results on two different iterative methods to find solutions of the lattice two-scale equation (1.6). A solution is a fixed point $f = Vf$ of the linear operator

$$Vf(x) = \sum_{n=0}^{N} c_n f(kx - n), \tag{1.7}$$

and a natural approach is to consider iterative schemes $f_j = Vf_{j-1}$ that converge to a fixed point f starting from suitable f_0. Given data $\{f(n);\ n \in \mathbb{Z}\}$ for an L^1-solution of such an equation with $\Delta = 1$, we can construct a piecewise linear spline f_0 that has $f_0(n) = f(n)$ for all $n \in \mathbb{Z}$. We show that if f is continuous, then the iterates $f_{j+1} = Vf_j$ are piecewise linear splines with successively finer knot sets (Theorem 4.1) and that $f_j \to f$ pointwise, with a rate of convergence depending on the smoothness of f. If f is L times continuously differentiable, then f_0 can be chosen to be a C^L piecewise polynomial spline of degree $2L+1$, with $f_0^{(l)}(n) = f^{(l)}(n)$ for all $n \in \mathbb{Z}$, $l = 0, \cdots, L$. Then the iterates $f_{j+1} = Vf_j$ are again C^L piecewise polynomial splines of degree $2L+1$ with successively finer knot sets, and we prove $f_j^{(l)} \to f^{(l)}$ pointwise, for all $l = 0, \cdots, L$. These results show in particular that convergence to a C^0-solution $f(x)$ occurs when one exists, if we start with correct initial conditions on $\mathbb{Z}$. However, they give no information concerning existence of such solutions.

The second iterative method for finding solutions to (1.6) discussed in § 4 is the "cascade algorithm." The successive approximations f_j in this scheme are again defined by $f_j = Vf_{j-1}$, but the starting point is now $f_0(x) = 1 - |x|$ for $-\frac{1}{2} \leqq x \leqq \frac{1}{2}$ and $f_0(x) = 0$ otherwise. These initial conditions are not usually a solution to (1.6) on $\mathbb{Z}$. The advantage of the cascade algorithm is that f_j can be computed via a "local" method: at every step j, the value of $f_j(x)$ can be determined using *only* the values obtained in the previous step in the region $\{y;\ |y - x| \leqq C2^{-j}\}$ (C independent of j), a neighborhood of x becoming exponentially small as j increases. This lends a "zoom-in" quality to the successive steps of the cascade algorithm (when it converges). It is known that the cascade algorithm does not always converge pointwise to a nonzero C^0-solution when one exists. This scheme has been studied by several authors (cf. Deslauriers and Dubuc (1989), Dyn, Gregory, and Levin (1989), (1990)), and various sufficient conditions for its convergence are known.

Section 5 obtains bounds on the global regularity of any nonzero L^1-solution to (1.6) when $\Delta = 1$. Theorem 5.1 shows that if such a solution is in $C^m(\mathbb{R})$, then $m < N/(k-1) - 1$. This result is best possible in the sense that there exist equations having C^m-solutions, for which $m \geqq N/(k-1) - 2$, for arbitrarily large m.

Finally, § 6 applies these results to three examples.

A number of authors have studied more general solutions to two-scale difference equations and related functional equations. In constructing fractals Barnsley and Demko (1985) study measures μ which are solutions of

$$\mu(S) = \sum_{n=0}^{N} c_n \mu(\alpha S - \beta_n), \qquad S \text{ a Borel set.}$$

The *adjoint operator* (in the L^2-sense) to (1.1) is

$$V^A g(x) = \sum_{n=0}^{N} c_n g(\alpha^{-1}(x+\beta_n)).$$

The stationary measures μ studied by Diaconis and Shashahani (1986) are fixed points of generalizations of such adjoint operators. It is also interesting to note that the mth Bernoulli polynomial $B_m(x)$ satisfies the equation $V^A B_m(x) = B_m(x)$ with

$$V^A g(x) = \sum_{n=0}^{m-1} n^{m-1} g\left(\frac{1}{m}(x+n)\right).$$

2. Existence and uniqueness of L^1-solutions. We are interested in L^1-solutions f to the two-scale difference equation

$$f(x) = \sum_{n=0}^{N} c_n f(\alpha x - \beta_n). \tag{2.1}$$

Since f is in $L^1(\mathbb{R})$, its Fourier transform $\hat{f}$,

$$\hat{f}(u) = \int_{-\infty}^{\infty} e^{ixu} f(x)\, dx,$$

is a bounded continuous function. By viewing (2.1) as a convolution equation, we see that $\hat{f}$ satisfies

$$\hat{f}(u) = P(\alpha^{-1}u)\hat{f}(\alpha^{-1}u) \tag{2.2}$$

where

$$P(u) = \frac{1}{\alpha} \sum_{n=0}^{N} c_n e^{i\beta_n u}. \tag{2.3}$$

The existence and uniqueness of L^1-solutions to (2.1) are governed by the value

$$\Delta = P(0) = \frac{1}{\alpha} \sum_{n=0}^{N} c_n,$$

as shown in the following result.

THEOREM 2.1. *Let Δ be defined as above. Then the following are true.*

(a) *If $|\Delta| \leqq 1$ and $\Delta \neq 1$, then the only L^1-solution of* (2.1) *is the trivial solution $f \equiv 0$.*

(b) *If $\Delta = 1$ then there exists, up to normalization, at most one nontrivial L^1-solution to* (2.1). *If it exists, then its Fourier transform is given by*

$$\hat{f}(u) = A \prod_{j=1}^{\infty} P(\alpha^{-j}u) \tag{2.4}$$

where $A = \hat{f}(0) = \int f(x)\, dx$ and the infinite product converges for all u. Conversely, if the right-hand side of (2.4) *is the inverse Fourier transform of an L^1-function f, then f is a nontrivial L^1-solution to* (2.1).

(c) *If* $|\Delta|>1$, *then the Fourier transform of any* L^1*-solution* f *is necessarily of the form*

$$\hat{f}(u)=\left[\prod_{j=1}^{\infty} p(\alpha^{-j}u)\right]\exp\left(\frac{\ln\Delta}{\ln\alpha}\ln|u|\right)g_{\operatorname{sgn}(u)}\left(\frac{\ln|u|}{\ln\alpha}\right) \tag{2.5}$$

where

$$p(u)=\Delta^{-1}P(u),$$

$g_{\pm}$ *are continuous periodic functions with period* 1, *and*

$$\ln\Delta=\ln|\Delta|+i\theta \quad \text{where } \Delta=|\Delta|\,e^{i\theta} \quad \text{and} \quad -\pi<\theta\leqq\pi.$$

Furthermore, the infinite product converges for all complex u. *Conversely, if* g_+, g_- *are continuous periodic functions of period* 1 *such that the inverse Fourier transform* f *of the right side of* (2.5) *is in* $L^1(\mathbb{R})$, *then* f *satisfies* (2.1).

Proof. (1) We have

$$\begin{aligned}|P(u)-\Delta| &\leqq \alpha^{-1}\sum_{n=0}^{N}|c_n||e^{i\beta_n u}-1| \\ &\leqq K\min(1,|u|)\exp[B|\operatorname{Im}(u)|]\end{aligned} \tag{2.6}$$

where $B=\max|\beta_i|$ and $K=2\alpha^{-1}\sum_{n=0}^{N}|c_n|(1+|\beta_n|)$.

(2) We first treat the case where $|\Delta|<1$. Since $f\in L^1(\mathbb{R})$, $\hat{f}(u)$ is continuous for $u\in\mathbb{R}$. From (2.6) we have (for real u)

$$|\hat{f}(u)|=|P(\alpha^{-1}u)||\hat{f}(\alpha^{-1}u)|\leqq(|\Delta|+K\alpha^{-1}|u|)|\hat{f}(\alpha^{-1}u)|.$$

It follows that, for all $j\in\mathbb{N}$,

$$|\hat{f}(u)|\leqq\|f\|_1\prod_{l=1}^{j}(|\Delta|+K\alpha^{-l}|u|). \tag{2.7}$$

For any real u we can make the product on the right side of (2.7) arbitrarily small by choosing j large enough, since $|\Delta|<1$. Hence $f\equiv 0$. This proves (a), except for $|\Delta|=1$, which we treat below.

(3) For $|\Delta|\geqq 1$ define $p(u)=\Delta^{-1}P(u)$. By (2.6) we have

$$|p(u)-1|\leqq K\,e^{B}\Delta^{-1}|u|:=K'|u|, \tag{2.8}$$

for complex $|u|<1$. Now we define

$$\hat{f}_0(u):=\prod_{j=1}^{\infty}p(\alpha^{-j}u), \tag{2.9}$$

and (2.8) shows that the infinite product converges absolutely and uniformly on compact subsets of $\mathbb{C}$ to an entire function. The bound (2.8) shows that for $|u|\leqq(2K')^{-1}$, $|p(u)|\geqq 1-K'|u|\geqq(1+2K'|u|)^{-1}$ and

$$\begin{aligned}|\hat{f}(\alpha^{-j}u)| &= |\hat{f}(u)|\prod_{l=1}^{j}|P(\alpha^{-l}u)|^{-1} \\ &\leqq |\Delta|^{-j}|\hat{f}(u)|\prod_{l=1}^{j}(1+2K'\alpha^{-l}|u|) \\ &\leqq |\Delta|^{-j}\|f\|_1\exp[2K'(\alpha-1)^{-1}|u|].\end{aligned}$$

This implies that

$$|\hat{f}(u)|\leqq\exp[(\alpha-1)^{-1}]\|f\|_1|\Delta|^{-j} \quad \text{when } |u|\leqq(2K')^{-1}\alpha^{-j}.$$

For any $|u| \leqq (2K')^{-1}$ we can find j so that $(2K')^{-1}\alpha^{-(j+1)} \leqq |u| \leqq (2K')^{-1}\alpha^{-j}$. It follows that

$$\begin{aligned} |\hat{f}(u)| &\leqq \exp[(\alpha-1)^{-1}]\|f\|_1 |\Delta|^{-j} \\ &\leqq \exp[(\alpha-1)^{-1}]\|f\|_1 \alpha^{-j \ln|\Delta|/\ln\alpha} \\ &\leqq C|u|^{\ln|\Delta|/\ln\alpha} \end{aligned} \tag{2.10}$$

for all $|u| \leqq (2K')^{-1}$. Since $|\hat{f}(u)|$ is bounded for real u ($f \in L^1$), (2.10) also extends (with possibly a different constant) to all $u \in \mathbb{R}$. Note that for $|\Delta| > 1$, (2.10) implies that

$$\int_{-\infty}^{\infty} f(x)\,dx = \hat{f}(0) = 0,$$

which can also be obtained directly from (2.1) by integration. Define $F(u)$ by

$$F(u) := \exp\left(-\frac{\ln|\Delta|}{\ln\alpha}\ln|u|\right)\hat{f}(u).$$

Since $\hat{f}$ is continuous, F is continuous as well, except possibly at $u = 0$, and by (2.10), F is bounded near $u = 0$. The function F satisfies the recursion

$$F(u) = e^{i\theta}p(\alpha^{-1}u)F(\alpha^{-1}u), \tag{2.11}$$

where $\Delta = |\Delta|\, e^{i\theta}$ and $-\pi < \theta \leqq \pi$. This yields

$$F(u) = \left[\prod_{j=1}^{J} p(\alpha^{-j}u)\right] e^{iJ\theta}F(\alpha^{-J}u).$$

The first factor has the limit $\hat{f}_0(u)$ as $J \to \infty$. When this limit is not zero, the second factor must also converge as $J \to \infty$, and we denote this limit by $\phi(u)$, so that

$$\phi(u) := [\hat{f}_0(u)]^{-1}F(u).$$

The function ϕ is continuous, except possibly at $u = 0$, where F may be discontinuous, and at the zeros of $\hat{f}_0(u)$. From (2.8) we find for complex $|u| \leqq 1$ that

$$|\hat{f}_0(u) - 1| \leqq K'|u| \exp[B(\alpha-1)^{-1}|u|],$$

which implies that $\hat{f}_0(u)$ is bounded away from zero in a neighborhood of $u = 0$. It follows that ϕ is continuous in a neighborhood of zero, except possibly at $u = 0$. On the other hand, the recursion (2.11) for F gives

$$\phi(u) = e^{i\theta}\phi(\alpha^{-1}u).$$

Define the two functions

$$g_{\pm}(t) := \phi(\pm\alpha^t)\exp(i\theta t), \qquad t \in \mathbb{R}.$$

Then

$$g_{\pm}(t+1) = g_{\pm}(t). \tag{2.12}$$

If the periodic functions $g_{\pm}$ had any singularity (including discontinuities), then ϕ would have infinitely many singularities in a neighborhood of $u = 0$. Since this is impossible, it follows that the $g_{\pm}$ are continuous. To prove the converse statement, it suffices to observe that, under the stated conditions, for $|\Delta| \geqq 1$, the right-hand side of (2.5) satisfies the functional equation $\hat{f}(u) = P(\alpha^{-1}u)\hat{f}(\alpha^{-1}u)$. This establishes (c).

(4) If $|\Delta| = 1$, then the above construction simplifies. We have

$$\hat{f}(u) = \hat{f}_0(u)\phi(u), \tag{2.13}$$

where ϕ satisfies

$$\phi(u) = \Delta\phi(\alpha^{-1}u) = e^{i\theta}\phi(\alpha^{-1}u).$$

Since $\hat{f}_0(u)$ is bounded away from zero for small $|u|$, it follows from (2.13) that ϕ is continuous at zero. In particular,

$$\phi(0) = e^{i\theta}\phi(0),$$

implying either $\Delta = e^{i\theta} = 1$ or $\phi(0) = 0$. However, we also know that

$$\phi(u) = g_{\mathrm{sgn}(u)}\left(\frac{\ln|u|}{\ln\alpha}\right)\exp\left(i\theta\frac{\ln|u|}{\ln\alpha}\right)$$

where $g_{\pm}$ are continuous periodic functions with period 1. If $\Delta \neq 1$ then $\phi(0) = 0$; hence $|g_{\pm}(t)| = |\phi(\pm\alpha^t)| \to 0$ as $t \to -\infty$. Since the functions $g_{\pm}$ are periodic, this forces $g_{\pm} \equiv 0$; hence $\phi \equiv 0$ and $f \equiv 0$ for $\Delta \neq 1$, which proves the rest of (a). If $\Delta = 1$ then $g_{\pm}(t) = \phi(\pm\alpha^t) \to \phi(0)$ as $t \to \infty$. By the periodicity of $g_{\pm}$ this implies that $g_{\pm} \equiv \phi(0)$ are both the same constant function; hence $\hat{f}(u) = \phi(0)\hat{f}_0(u)$. This proves (b). □

Remarks. (1) Theorem 2.1 also holds for "infinite" two-scale difference equations, i.e.,

$$f(x) = \sum_{n=-\infty}^{\infty} c_n f(\alpha x - \beta_n)$$

provided that $\sum_{n=-\infty}^{\infty} |c_n||\beta_n|^{\delta} < \infty$ for some $\delta > 0$. The estimate (2.6) then becomes

$$|P(u) - \Delta| \leqq \alpha^{-1} \sum_{n=-\infty}^{\infty} |c_n| \min\left(2, |\beta_n u|^{\min(1,\delta)}\right) e^{|B||\mathrm{Im}\, u|},$$

and the other estimates can be adjusted similarly.

(2) Note that there may exist distributional solutions even if $|\Delta| < 1$. One example is the equation

$$f(x) = \tfrac{1}{12}f(2x) + \tfrac{1}{6}[f(2x+1) + f(2x-1)] - \tfrac{1}{12}[f(2x+2) + f(2x-2)],$$

which admits $f(x) = x^2$ as a solution. The Fourier transform of this solution is a distribution supported at the origin, so that the continuity argument used in the proof of Theorem 2.1 does not apply.

(3) There are no distributional solutions with Fourier transform continuous at zero if $|\Delta| < 1$. For $|\Delta| \geqq 1$, (2.4) and (2.5) always give distributional solutions to the two-scale equation (2.1), but there may exist other distributional solutions with discontinuous Fourier transforms.

Theorem 2.1 has the following corollary, proved in the lattice case by Deslauriers and Dubuc (1987).

COROLLARY 2.2. *If the two-scale difference equation* (2.1) *with* $\Delta = 1$ *possesses a nontrivial* L^1*-solution* f*, then* $\int_{-\infty}^{\infty} f(x)\,dx \neq 0$ *and* f *has compact support, with*

$$\mathrm{supp}\,(f) \subset [\beta_0(\alpha-1)^{-1}, \beta_N(\alpha-1)^{-1}]. \tag{2.14}$$

Proof. By Theorem 2.1

$$\hat{f}(u) = \hat{f}(0)\prod_{j=1}^{\infty} P(\alpha^{-j}u) = \hat{f}(0)\hat{f}_0(u).$$

Hence $\int_{-\infty}^{\infty} f(x)\,dx = \hat{f}(0) \neq 0$. We can without loss of generality reduce to the case that $\beta_0 = -\beta_N$ by considering

$$f_1(x) = f\left(x - (\alpha-1)^{-1}\frac{\beta_N - \beta_0}{2}\right),$$

which is easily checked to satisfy

$$f_1(x) = \sum_{n=0}^{N} c_n f_1\left(\alpha x - \beta_j + \frac{\beta_0 - \beta_N}{2}\right).$$

Now suppose $\beta_0 = -\beta_N$ and (2.6) becomes

$$|P(u) - 1| \leqq K \min(1, |u|) \exp(B|\operatorname{Im}(u)|)$$

with $B = \beta_N$. On the annulus $\alpha^k \leqq |u| \leqq \alpha^{k+1}$ this gives

$$\begin{aligned} |\hat{f}_0(u)| &\leqq \prod_{j=1}^{k+1} [1 + K\Delta^{-1} \exp(B\alpha^{-j}|\operatorname{Im}(u)|)] \prod_{j=k+2}^{\infty} [1 + K'\alpha^{-j} \exp(B)] \\ &\leqq C(1+|u|)^M \exp[B(\alpha-1)^{-1}|\operatorname{Im}(u)|], \end{aligned} \tag{2.15}$$

where $M = [|\ln((K\Delta^{-1}+1)/\alpha)| + 1]$ and C is a constant. By the Paley–Wiener theorem for distributions (see, e.g., Reed and Simon (1975, Thm. IX.12)) $\hat{f}_0(u)$ is the Fourier transform of a distribution f_0 in $\mathcal{S}'(\mathbb{R})$ having compact support in the interval $|x| \leqq B(\alpha-1)^{-1}$. By hypothesis this distribution is the L^1-function $[\hat{f}(0)]^{-1} f$; hence f has compact support in $|x| \leqq (\beta_n - \beta_0)/(2(\alpha-1))$ and (2.14) follows. □

This proof shows that *all* two-scale difference equations with $\Delta = 1$ possess a distributional solution f in $\mathcal{S}'(\mathbb{R})$ having compact support in $[\beta_0(\alpha-1)^{-1}, \beta_N(\alpha-1)^{-1}]$ which has Fourier transform (2.4). The arguments of § 3 will show that up to a scale factor this is the unique distribution in $\mathcal{S}'(\mathbb{R})$ which satisfies (3.1) and has compact support. The following examples illustrate a few cases, for different values of Δ.

Examples. (1) Consider the lattice two-scale difference equation

$$f(x) = \tfrac{1}{2}\Delta[f(2x) + 2f(2x-1) + f(2x-2)].$$

Depending on the value of Δ, there will be one, infinitely many, or no nontrivial L^1-solutions. If $\Delta = 1$, then any candidate L^1-solution satisfies

$$\begin{aligned} \hat{f}(u) &= \hat{f}(0) \prod_{j=1}^{\infty} \left\{\frac{1}{4}[1 + \exp(i2^{-j}u)]^2\right\} \\ &= \hat{f}(0)\, e^{iu}\left(\frac{\sin(u/2)}{u/2}\right)^2. \end{aligned}$$

It follows that f is a multiple of the function g,

$$g(x) = \begin{cases} x, & 0 \leqq x \leqq 1, \\ 2-x, & 1 \leqq x \leqq 2, \\ 0 & \text{otherwise.} \end{cases}$$

Up to normalization, we therefore have a unique L^1-solution in this case. For $\Delta = 2$, we find

$$\hat{f}(u) = |u|\, e^{iu}\left(\frac{\sin(u/2)}{u/2}\right)^2 g_{\operatorname{sgn}(u)}\left(\frac{\ln|u|}{\ln 2}\right),$$

where $g_{\pm}$ are periodic, continuous functions of period 1. Clearly, $\hat{f} \in L^2(\mathbb{R})$. If $g_{\pm}$ are C^1, then we easily check that also $(\hat{f})' \in L^2$, which implies $f \in L^1$. There is therefore

clearly an infinity of different possible L^1-solutions in this case. Only one of these L^1-solutions is compactly supported (see § 3). For $\Delta = 4$, however, we have

$$\hat{f}(u) = 4\, e^{iu}[\sin(u/2)]^2 g_{\mathrm{sgn}(u)}\left(\frac{\ln|u|}{\ln 2}\right).$$

This tends to zero for $|u| \to \infty$ only if both functions $g_{\pm} \equiv 0$. The only L^1-solution is therefore the trivial solution $f \equiv 0$.

(2) The following example shows that $\Delta = 1$ does not imply the existence of a nontrivial L^1-solution. Take the lattice two-scale difference equation

$$f(x) = 2f(2x-1).$$

Every candidate L^1-solution satisfies

$$\hat{f}(u) = \hat{f}(0) \prod_{j=1}^{\infty} [\exp(i2^{-j}u)] = \hat{f}(0)\, e^{iu}.$$

Since e^{iu} is the Fourier transform of the Dirac δ-measure at $x = 1$, there are no nontrivial L^1-solutions.

(3) Consider the family of two-scale difference equations

$$f_\alpha(x) = \frac{\alpha}{2}\{f_\alpha(\alpha x - 1) + f_\alpha(\alpha x + 1)\},$$

which all have $\Delta = 1$. This equation always has a distributional solution with Fourier transform

$$\hat{f}_\alpha(u) = \prod_{n=1}^{\infty} \cos(\alpha^{-n}u),$$

which has compact support in $[-(\alpha-1)^{-1}, (\alpha-1)^{-1}]$ by the same argument as in the proof of Corollary 2.2. The smoothness of this distribution as a function of $\beta = \alpha^{-1}$ for $0 < \beta < 1$ was studied by Kershner and Wintner (1935), Erdös (1939), (1940), and Garsia (1962). It is known that for $\alpha = 2^{1/k}$, with k sufficiently large, the function f_α is continuous (hence in $L^1(\mathbb{R})$) and arbitrarily smooth. Erdös (1940) showed that for any k there is a constant $c(k)$ such that for almost all $\beta = \alpha^{-1}$ in the interval $[c(k), 1]$ the distribution f_α is a function in $C^{(k)}(\mathbb{R})$.

3. L^1-solutions having compact support. We consider the general two-scale difference equation

$$f(x) = \sum_{n=0}^{N} c_n f(\alpha x - \beta_n), \tag{3.1}$$

and derive necessary conditions for the existence of nonzero L^1-function of compact support.

THEOREM 3.1. *Suppose that the two-scale difference equation* (3.1) *possesses a nonzero L^1-solution f having compact support. Then*:

(a) $\Delta = \alpha^m$ *for a nonnegative integer* m.

(b) f *is unique up to a scale factor and has Fourier transform*

$$\hat{f}(u) = Au^m \prod_{k=1}^{\infty} p(\alpha^{-k}u) \tag{3.2}$$

where $p(u) = \Delta^{-1} \sum_{n=0}^{n} c_n\, e^{i\beta_n u}$.

(c) *The two-scale equation with* $\Delta = 1$ *obtained by replacing* $\{c_n\}$ *by* $\{\alpha^{-m}c_n\}$ *has a nonzero* L^1*-solution* g *unique up to scale, and with a proper choice of scale* $(d^m/dx^m)g(x) \equiv f(x)$.

The main ingredient in the proof is the following result.

LEMMA 3.2. *Suppose that the function* M_0, *defined on* $\mathbb{R} - \{0\}$ *by*

$$M_0(u) := \exp(\gamma_0 \ln |u|) g_{\mathrm{sgn}(u)}(\gamma_1 \ln |u|), \tag{3.3}$$

is such that

(1) $g_{\pm}(t)$ *are periodic functions of period* 1.

(2) γ_1 *is real.*

(3) $M_0(u)$ *possesses an analytic continuation to* $\mathbb{C}$ *which is an entire function of exponential type.*

Then $M_0(u) = Au^m$ *where* m *is a nonnegative integer.*

Proof. (1) The function $h(u) = \exp(\gamma_0 \ln u)$ can be continued analytically to the simply-connected two-sheeted region $\mathbf{R} = \{z = re^{i\theta};\ r > 0 \text{ and } -\frac{3}{2}\pi \leqq \theta \leqq \frac{3}{2}\pi\}$ where $re^{i\theta}$ and $re^{i(\theta+2\pi)}$ are viewed as distinct points and

$$h(re^{i(\theta+2\pi)}) = \exp(2\gamma_0 \pi i) h(re^{i\theta}), \tag{3.4}$$

whenever both sides are in $\mathbf{R}$.

(2) Since $M_0(u)$ is entire and $h(z)$ is nonzero on $\mathbf{R}$, it follows from (3.3) that $g_+(\gamma_1 \ln u)$ has an analytic continuation to $\mathbf{R}$. Therefore, in terms of the variable $t = \gamma_1 \ln u$, $g_+(t)$ has an analytic continuation to the horizontal strip $\mathbf{T} = \{t;\ -\frac{3}{2}\pi\gamma_1 < \mathrm{Im}\,(t) < \frac{3}{2}\pi\gamma_1\}$. The periodicity

$$g_+(t+1) = g_+(t) \tag{3.5}$$

on the real axis extends to this strip by analytic continuation. Now the single-valuedness of $M_0(u)$ on $\mathbb{C}$ means that

$$M_0(re^{i(\theta+2\pi)}) = M_0(re^{i\theta}), \qquad i = 1, 2,$$

on the region $\mathbf{R}$, and combined with (3.3) and (3.4) this implies that

$$g_+(t + 2\gamma_1 \pi i) = \exp(-2\gamma_0 \pi i) g_+(t) \tag{3.6}$$

is valid when both t, $t + 2\gamma_1 \pi i \in \mathbf{T}$, i.e., for $\{t;\ -\frac{3}{2}\pi\gamma_1 < \mathrm{Im}\,(t) < \frac{1}{2}\pi\gamma_1\}$. This relation allows us to continue g_+ analytically to the entire plane, and (3.5), (3.6) then hold for all complex t.

(3) We claim next that $g_+(t)$ is an entire function of exponential type. To see this, note that it is bounded by a constant, say C, on the rectangle $0 \leqq \mathrm{Re}\,(t) \leqq 1$, $0 \leqq \mathrm{Im}\,(t) \leqq 2\pi\gamma_1$, and the periodicity relations (3.5) and (3.6) then give

$$|g_+(t)| \leqq C \exp(2\pi |\gamma_0| |\mathrm{Im}\,(t)|),$$

proving the claim.

Next we show that $g_+(t)$ has no zeros. For if it had a zero at t_0, the periodicity relations (3.5) and (3.6) would give it zeros at $t_0 + m + (2\gamma_1 \pi i)n$ for m, $n \in \mathbb{Z}$, which contradicts the property that an entire function of exponential type has $O(R)$ zeros in the disc $\{t: |t| \leqq R\}$ as $R \to \infty$.

Since $g_+(t)$ is an entire function of exponential type having no zeros, $g_+(t) = A_+ \exp(c_+ z)$ for some constant c_+. The periodicity (3.5) forces $c_+ = 2k\pi i$ for some $k \in \mathbb{Z}$.

(4) A similar argument applied to $g_-(\gamma_1 \ln(-u))$ shows that $g_-(t) = A_- \exp(c_- z)$ where $c_- = 2l\pi i$ for some $l \in \mathbb{Z}$.

(5) Now we have for real u that

$$M_0(u) = \begin{cases} A_+ \exp[(\gamma_0 + 2k\pi i)\ln u], & u > 0, \\ A_- \exp[(\gamma_0 + 2l\pi i)\ln(-u)], & u < 0. \end{cases}$$

The first expression has a singularity at $u = 0$ unless $\gamma_0 + 2k\pi i = m_+$ is an integer. Similarly, we conclude that $\gamma_0 + 2l\pi i = m_-$ is an integer. Analyticity of $M_0(u)$ at $u = 0$ yields $A_+ = A_-$ and $m_+ = m_- = m \geqq 0$, and $M_0(u) = Au^m$ for a nonnegative integer m. □

We now proceed to prove Theorem 3.1.

Proof of Theorem 3.1. (1) By the Paley-Wiener theorem the Fourier-Laplace transform $\hat{f}$ of f is necessarily an entire function of exponential type, satisfying

$$|\hat{f}(u)| \leqq C \exp(B|\operatorname{Im} u|) \tag{3.7}$$

for some constants B, C. Now set

$$\hat{f}_0(u) := \prod_{j=1}^{\infty} p(\alpha^{-j}u),$$

which by the argument in the proof of Corollary 2.2 is also entirely of exponential type. We claim that

$$M(u) := \frac{\hat{f}(u)}{\hat{f}_0(u)} \tag{3.8}$$

is also an entire function of exponential type, i.e., any zero of $\hat{f}(u)$ has at least the multiplicity of $\hat{f}_0(u)$. To see this, note that for any zero u_0 of $\hat{f}_0(u)$ of multiplicity m there is a finite product $\prod_{j=1}^{J} p(\alpha^{-j}u)$ having a zero of the same multiplicity. Iterating the basic recursion (2.2) yields

$$\hat{f}(u) = \Delta^J \left(\prod_{j=1}^{J} p(\alpha^{-j}u) \right) \hat{f}(\alpha^{-J}u).$$

Since all terms on the right side of this expression are analytic at u_0 and the product has a zero of multiplicity m, $\hat{f}(u)$ has a zero there of at least that multiplicity, and the claim follows.

(2) Since $f \in L^1$, it satisfies the formula of Theorem 2.1(c). Thus the hypotheses of Lemma 3.2 are satisfied for $M(u)$ given by (3.8). Consequently, $M = Au^m$ and

$$\hat{f}(u) = Au^m \prod_{k=1}^{\infty} p(\alpha^{-k}u). \tag{3.9}$$

This proves claim (b). On the other hand, the two-scale equation (3.1) implies

$$\hat{f}(u) = \Delta p(\alpha^{-1}u)\hat{f}(\alpha^{-1}u).$$

Substituting (3.9) gives $\Delta = \alpha^m$, which proves (a).

(3) For (c), observe that if $m \geqq 1$ then $\hat{f}(0) = 0$; hence

$$\int_{-\infty}^{\infty} f(x)\,dx = 0. \tag{3.10}$$

Define $f_1(x) = \int_{-\infty}^{x} f(w)\,dw$ and observe that since f has compact support (3.10) shows that $f_1(x)$ is in $L^1(\mathbb{R})$ with compact support. Furthermore, f_1 satisfies the two-scale equation (3.1) with $\{c_n\}$ replaced by $\{\alpha^{-1}c_n\}$, by integrating (3.1). Of course $(d/dx)(f_1(x)) = f(x)$. By integrating m times, (c) follows. □

Remarks. (1) Under the weaker hypothesis that (3.1) possesses a solution which is a tempered distribution of compact support, the conclusions (a) and (b) of the theorem still hold.

(2) L^1-solutions satisfying (a) can exist for arbitrarily large values of m, for suitable values of α and of the c_j, β_j. By (c), this is equivalent to saying that there exist L^1-solutions with arbitrary high regularity for two-scale difference equations with $\Delta = 1$. Examples are given in § 6.

4. Lattice two-scale difference equations: iterative approximations. In the remainder of this paper we study compactly supported continuous solutions of lattice two-scale difference equations,

$$f(x) = \sum_{n=0}^{N} c_n f(kx - n), \tag{4.1}$$

where k is an integer $\geqq 2$. We will suppose that $\Delta = (1/k)\sum_{n=0}^{N} c_n = 1$, which involves essentially no loss of generality by Theorem 3.1.

A continuous solution of such an equation is a fixed point $Vf = f$ of the linear operator

$$Vf(x) = \sum_{n=0}^{N} c_n f(kx - n) \tag{4.2}$$

acting on a function space, e.g., $C^0(\mathbb{R})$. A natural method to construct a solution of (4.1) is as a limit of the iterative approximation scheme $f_{j+1} = Vf_j$, where f_0 is a suitable initial function. In this section we discuss the convergence of two such approximation schemes.

We first suppose that a compactly supported continuous solution $f(x)$ exists, and that the data $\{f(n): n \in \mathbb{Z}\}$ are known. We consider initial functions f_0 which are piecewise linear splines interpolating these data with knots at the integers $\mathbb{Z}$. That is, f_0 is defined by

$$f_0(x) = f(n)(n+1-x) + f(n+1)(x-n) \quad \text{for } n \leqq x \leqq n+1.$$

Since $f(n) = 0$ for $n \notin [0, (k-1)^{-1}N]$, f_0 has compact support in $[0, \lceil (k-1)^{-1}N \rceil]$, so we may regard it as being defined on the finite knot set $\mathbb{Z} \cap [0, \lceil (k-1)^{-1}N \rceil]$. (As usual $\lceil a \rceil$ stands for "smallest integer larger than or equal to a.") It immediately follows that $f_j = V^j f_0$ is a piecewise linear spline with knots at the $k^{-j}n$, $0 \leqq n \leqq \lceil k^j(k-1)^{-1}N \rceil$, which agrees with f at these knots. Consequently we have Theorem 4.1.

THEOREM 4.1. *Suppose that the lattice two-scale equation* (4.1) *with* $\Delta = 1$ *has a nonzero continuous solution* f *of compact support. Let* f_0 *be the spline of degree* 1 *with knot set* $\{n;\ n \in \mathbb{Z} \cap [0, \lceil (k-1)^{-1}N \rceil]$, *and with* $f_0(n) = f(n)$. *Define* $f_j = V^j f_0$, *with* V *as in* (4.2). *Then*

(1) f_j *is an interpolating spline of degree* 1 *with knot set* $k^{-j}(\mathbb{Z} \cap [0, \lceil k^j(k-1)^{-1}N \rceil])$.

(2) f_j *agrees with* f *at its knots,* $f_j(k^{-j}n) = f(k^{-j}n)$.

(3) $\|f - f_j\|_{L^\infty} \to 0$ *as* $j \to \infty$.

(4) *If* $f \in \mathrm{Lip}^\alpha$ *for* $0 < \alpha \leqq 1$, *i.e.,* $|f(x) - f(y)| \leqq C|x-y|^\alpha$, *then* $\|f - f_j\|_{L^\infty} \leqq Ck^{-j\alpha}$.

Proof. (1) and (2) were derived above; (3) and (4) are standard spline convergence results; see, e.g., Schumaker (1981, Thm. 6.15) or Theorem 4.2 below. □

If the compactly supported solution f has more regularity, e.g., if $f \in \mathrm{Lip}^{L,\alpha}$ (which means $f \in C^L$ and $d^L f/dx^L \in \mathrm{Lip}^\alpha$), then the same piecewise linear f_j converge even faster to f (Schumaker (1981, Thm. 6.15)). In order to obtain convergence of the derivatives as well, we need to use an initial function f_0 that is more regular. This can

be achieved by choosing for f_0 a C^L piecewise polynomial spline of degree $2L+1$ that agrees with f and its first L derivatives on the knot set $\mathbb{Z}$. (Similar fast convergence of the f_j and their derivatives can be achieved by other C^L choices for f_0, for which the derivatives on $\mathbb{Z}$ do not necessarily agree with those of f. In our present case, however, we can determine the $f^{(l)}(n)$ easily, and we can therefore afford to pick this particular f_0. The associated f_j will play a role in part II as well.)

THEOREM 4.2. *Suppose that the lattice two-scale equation* (4.1) *with* $\Delta=1$ *has a nonzero solution* f *of compact support which is* L *times continuously differentiable. Let* f_0 *be the* C^L *interpolating spline of degree* $2L+1$ *with knot set* $\{n;\ n\in\mathbb{Z}\cap[0,\ \lceil(k-1)^{-1}N\rceil]\}$ *and such that* $f_0^{(l)}(n)=f^{(l)}(n)$, $l=0,\cdots,L$. *Define* $f_j=V^jf_0$, *with* V *as in* (4.2). *Then*

(1) f_j *is a* C^L *interpolating spline of degree* $2L+1$ *with knot set* $k^{-j}(\mathbb{Z}\cap[0,\ \lceil k^j(k-1)^{-1}N\rceil])$.

(2) $f_j^{(l)}(k^{-j}n)=f^{(l)}(k^{-j}n)$ *for* $n\in\mathbb{Z}$ *and,* $l=0,\cdots,L$.

(3) *For all* l, $0\leqq l\leqq L$, $\|f^{(l)}-f_j^{(l)}\|_{L^\infty}\leqq Ck^{-j(L-l)}$.

(4) *If* $f\in \mathrm{Lip}^{L,\alpha}$, *then* $\|f^{(l)}-f_j^{(l)}\|_{L^\infty}\leqq Ck^{-j(L-l+\alpha)}$ *for* $l=0,\cdots,L$.

Proof. Note first that f_0 exists and is uniquely determined by the constraints imposed: on every interval $[n, n+1]$, the $2L+2$ coefficients of f_0 are linear functions of the $2L+2$ boundary values $f_0^{(l)}(n), f_0^{(l)}(n+1)$, $l=0,\cdots,L$. It is obvious that $f_0\in C^L$. It then immediately follows that f_j is also C^L, that f_j is piecewise polynomial of degree $2L+1$, with knots at $k^{-j}\mathbb{Z}$, and that $f_j^{(l)}(k^{-j}n)=f^{(l)}(k^{-j}n)$, for $l=0,\cdots,L$. Points 3 and 4 are again standard results in spline approximation theory (they can, e.g., easily be proved by methods similar to those used in the proof of Theorem 6.15 in Schumaker (1981)); for the sake of convenience we also give an explicit and simple proof in the Appendix. □

Theorems 4.1 and 4.2 guarantee convergence of spline interpolants, provided we start from the right data $\{f(n);\ n\in\mathbb{Z}\}$ or $\{f^{(l)}(n);\ n\in\mathbb{Z}, l=0,\cdots,L\}$. In the latter case, we obtain very fast convergence of f *and* its derivatives. However, the theorems do not show how to determine these data or how to estimate smoothness of f given the data $\{k, c_1,\cdots,c_n\}$ specifying (4.1).

In the next section we shall see that the $f(n)$, $n=0$ to $\lceil(k-1)^{-1}N\rceil$ can be related to the eigenvector of eigenvalue 1 of a particular matrix constructed from the coefficients c_n. If this eigenvalue is nondegenerate, then this provides a way to determine the $f(n)$. Similarly, nondegenerate eigenvectors of this matrix, corresponding to the eigenvalue k^{-l}, are linked to the $f^{(l)}(n)$. We shall also see how this matrix provides an upper bound for the regularity of f; more subtle matrix techniques in part II will lead to more precise regularity estimates.

There exists another iterative scheme that is often used for the construction of f. The jth approximation function f_j in this scheme is also a spline function with knot set $2^{-j}\mathbb{Z}$, and $f_{j+1}=Vf_j$, but the initial function f_0 is different. It is a continuous, piecewise linear spline, with $f_0(0)=1$, $f_0(n)=0$ for $n\neq 0$. The advantage of this choice for f_0 is that it results in a "local" algorithm called the *cascade algorithm*. We check (see, e.g., Daubechies (1988)) that, for $0\leqq l<k$,

$$f_j(k^{-j}(km+l))=\sum_n c_{l+kn}f_{j-1}(k^{-j+1}(m-n)).$$

This means that the $f_j(k^{-j}n)$ can be computed by using only the values of f_{j-1} in a small neighborhood of $k^{-j}n$; more precisely, $f_j(k^{-j}n)$ is determined by the $f_{j-1}(k^{-j+1}l)$ with $k^{-j}(n-N)\leqq k^{-(j-1)}l\leqq k^{-j}n$. This is quite unlike the previous scheme, where $f_j(k^{-j}n)$ was computed from the $f_{j-1}(k^{-j-1}n-m)$, $0\leqq m\leqq N$. We remark that in general

$f_0(n) = \delta_{n0}$ does not satisfy the two-scale difference equation (4.1) restricted to $\mathbb{Z}$. It does so only if all but one of the coefficients c_{kn} (with index of a multiple of k) vanish, $c_{kn} = \delta_{n0}$. (We suppose here that $c_n = 0$ for $n < N_1$, or $n > N_2$, where N_1 need not be equal to zero. We can shift this to the standard situation $c_n = 0$ for $n < 0$ or $n > N$; in this case we would have $c_{kn+n_0} = \delta_{n0}$ for some n_0, and we would choose $f_0(n) = \delta_{nn_0}$ correspondingly.) In this case the cascade algorithm corresponds to an *interpolating subdivision scheme* (Chaiken (1974), Dyn, Gregory, and Levin (1987), (1989), (1990), Micchelli (1986), Micchelli and Prautzsch (1987a), (1987b), (1989)): at every level j, the function f_j coincides with f_{j-1} at the knots of f_{j-1}, i.e.,

$$f_j(k^{-j+1}n) = f_{j-1}(k^{-j+1}n);$$

the intermediate values $f_j(k^{-j}(kn+l))$, $0 < l < k$, are computed by an appropriate interpolation procedure (determined by the c_n). The "local" aspect of the cascade algorithm makes subdivision schemes of interest for the construction of curves and surfaces. In Daubechies (1988) the same scheme was called the "graphical" construction algorithm.

A drawback of the cascade algorithm is that it does not always converge, even when a continuous solution to the two-scale difference equation exists. An example is

$$f(x) = \tfrac{1}{2}f(2x-3) + f(2x) + \tfrac{1}{2}f(2x+3).$$

This equation corresponds to a subdivision scheme. It has a continuous solution with support $[-3, 3]$, namely,

$$f(x) = \begin{cases} 1 - |x|/3, & |x| \leqq 3, \\ 0 & \text{otherwise.} \end{cases}$$

The cascade algorithm converges to this solution in the sense of distributions, but not in $C^0(\mathbb{R})$: indeed $f_n(1) = 0$ for all n. In the special case of interpolating subdivision schemes ($c_{km} = \delta_{m0}$), Dyn and Levin (1989) give necessary and sufficient conditions to ensure convergence of the cascade algorithm. Daubechies (1988) lists a different set of sufficient conditions for convergence of the cascade algorithm.

5. Lattice two-scale difference equations: global regularity of compactly supported solutions. Assume that the lattice two-scale difference equation

$$f(x) = \sum_{n=0}^{N} c_n f(kx - n), \tag{5.1}$$

with $\Delta = (1/k)\sum_{n=0}^{N} c_n = 1$, has a nontrivial L^1-solution, necessarily of compact support. The regularity of this function can be bounded above purely in terms of its support width.

THEOREM 5.1. *Given a lattice two-scale difference equation, $f(x) = \sum_{n=0}^{N} c_n f(kx-n)$ with $\Delta = 1$. Let N_0 be the largest integer strictly smaller than $N/(k-1)$, and define* M *to be the $N_0 \times N_0$ matrix*

$$\mathsf{M}_{ij} = c_{ki-j}, \qquad i, j = 1, \cdots, N_0. \tag{5.2}$$

If there exists a nontrivial L^1-solution f which is in $C^m(\mathbb{R})$, then $\{1, k^{-1}, \cdots, k^{-m}\} \subset$ spectrum (M). *In particular,*

$$m < \frac{N}{k-1} - 1. \tag{5.3}$$

Proof. (1) Since f is continuous, and support $(f) \subset [0, N/(k-1)]$, it follows that $f(l) = 0$ for $l \leqq 0$ and $l > N_0$. Define $\mathsf{v}^0 \in \mathbb{R}^{N_0}$ by

$$\mathsf{v}_j^0 = f(j), \qquad j = 1, \cdots, N_0.$$

Substituting $x = j$, with $j = 1, \cdots, N_0$, into equation (5.1) leads to

$$\mathsf{v}^0 = \mathsf{M}\mathsf{v}^0,$$

where M is defined by (5.2).

(2) On the other hand, it is clear that $\mathsf{v}^0 \neq 0$. Indeed, if $\mathsf{v}^0 = \mathbf{0}$, i.e. $f(j) = 0$ for all $j \in \mathbb{Z}$, then $f(k^{-l}m) = 0$ would follow, for all $l \in \mathbb{N}$, $m \in \mathbb{Z}$, by applying (5.1). By continuity this would imply $f \equiv 0$. Since $f \neq 0$, we have $\mathsf{v}^0 \neq 0$. Consequently, 1 is an eigenvalue of M.

(3) Similarly we define, for $l \leqq m$, $\mathsf{v}^l \in \mathbb{R}^{N_0}$ by

$$\mathsf{v}_j^l = f^{(l)}(j), \qquad j = 1, \cdots, N_0,$$

where $f^{(l)}$ denotes the lth derivative of f. Differentiating (5.1) l times, and substituting $x = 1, \cdots, N_0$ leads to

$$\mathsf{v}^l = k^l \mathsf{M}\mathsf{v}^l.$$

Again $\mathsf{v}^l = 0$ would imply $f^{(l)} \equiv 0$, hence $f^{(l-1)} \equiv$ constant. Since $f^{(l-1)}$ has compact support, $f^{(l-1)} \equiv 0$ would follow. By induction this would imply that $f \equiv 0$. Since $f \neq 0$, $\mathsf{v}^l \neq 0$, and k^{-l} is an eigenvalue of M for $0 \leqq l \leqq m$.

(4) $f \in C^m(\mathbb{R})$ implies that the $N_0 \times N_0$ matrix M has $m+1$ eigenvalues. Hence $m \leqq N_0 - 1 < N/(k-1) - 1$. □

Remarks. (1) The bound (5.3) of Theorem 5.1 cannot be improved. For $N = (k-1)L$ there exist $\{c_n;\ n = 0, 1, \cdots, N\}$ such that the $(L-1) \times (L-1)$ matrix M has exactly the eigenvalues $1, k^{-1}, \cdots, k^{-L+2}$, and such that the corresponding f is in c^{L-2}. One such example is given by

$$p(\xi) = \sum_{n=0}^{(k-1)L} c_n e^{in\xi} = [(1 + e^{i\xi} + \cdots + e^{i(k-1)\xi})/k]^L,$$

leading to $\hat{f}(\xi) = [(1 + e^{i\xi})/\xi]^L$. The function f is a B-spline of degree $L-1$; it is in C^{L-2}. The fact that any C^{n-1}-spline with knot set $\mathbb{Z}$ must have support width greater than or equal to $n+2$ has long been known (cf. Schoenberg (1973, p. 13)).

(2) The condition $\{1, k^{-1}, \cdots, k^{-m}\} \subset$ spectrum (M) is not sufficient to ensure that $f \in C^m$. For $k = 2$, $N = 3$, e.g., all the choices $c_0 = \frac{1}{4} + \lambda$, $c_1 = \frac{3}{4} + \lambda$, $c_2 = \frac{3}{4} - \lambda$, $c_3 = \frac{1}{4} - \lambda$, all other $c_n = 0$, where $\lambda \in \mathbb{R}$ is arbitrary, lead to 2×2 matrices M with the same spectrum, namely, $\{1, \frac{1}{2}\}$. Nevertheless, the regularity of f depends on λ. Using the techniques of part II, we can check that f is continuous if and only if $|\lambda| < \frac{3}{4}$, and that $f \in C^1$ if and only if $|\lambda| < \frac{1}{4}$. For $\lambda = \frac{1}{4}$, e.g., we find $p(\xi) = (1 + e^{i\xi})^2/4$; hence $f(x) = x$ for $0 \leqq x \leqq 1$, $2 - x$ for $1 \leqq x \leqq 2$, zero otherwise, which is clearly not in C^1.

(3) It follows from the proof that, provided that they are nondegenerate, the eigenvectors of M with eigenvalue k^{-l} determine the $f^{(l)}(n)$, up to normalization. It is not a priori obvious how to choose these $m+1$ different normalizations (one for each l) in a coherent way. In part II we shall see how this can be done, modulo some restrictions on the c_n.

6. Examples.

6.1. The de Rham function. The de Rham function is a classical example of a continuous nowhere-differentiable function. Like many such examples, it is defined as the limit of successive approximations.

Define

$$f_0(x)=\begin{cases}1+x, & -1\leqq x\leqq 0,\\ 1-x, & 0\leqq x\leqq 1,\\ 0, & |x|\geqq 1.\end{cases}$$

Clearly f_0 is piecewise linear; its restriction to the intervals $[m, m+1]$, $m\in\mathbb{Z}$, is linear. The next function f_1 in the approximation scheme is constructed as follows: f_1 is again piecewise linear, with its restriction to the intervals $[m/3, (m+1)/3]$ linear, for all $m\in\mathbb{Z}$. The nodes of f_1 are given by $f_1(m)=f_0(m)$, $f_1(m+\frac{1}{3})=f_0(m+\frac{2}{3})$, $f_1(m+\frac{2}{3})=f_0(m+\frac{1}{3})$, for all $m\in\mathbb{Z}$. Graphically, this corresponds to splitting every interval on which f_0 is linear into three equal parts, exchanging the values at f_0 at the two interior points, and linearly interpolating between the nodes obtained in this way (see Fig. 1). Exactly the same procedure is then repeated to obtain f_{j+1} from f_j, for all $j\in\mathbb{N}$. The resulting f_j are piecewise linear, with linear restrictions to the intervals $[m3^{-j}, (m+1)3^{-j}]$, for all $m\in\mathbb{Z}$. Geometrically it is clear that this process converges pointwise to a continuous limit function f.

It can be checked fairly easily that $f_{n+1}=Vf_n$ where

$$Vf(x)=f(3x)+\tfrac{1}{3}[f(3x+1)+f(3x-1)]+\tfrac{2}{3}[f(3x+2)+f(3x-2)].$$

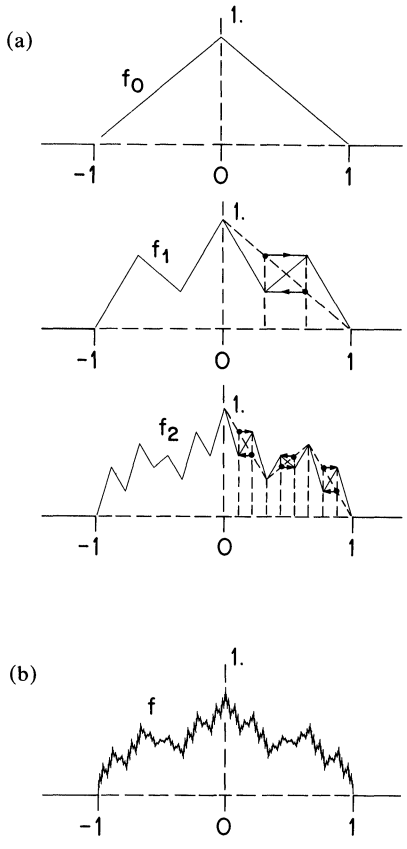


FIG. 1. (a) *The first three approximations f_0, f_1, f_2 to the de Rham function.* (b) *The de Rham function.* (*Note. We have plotted f_8 rather than f. At the scale of the figure, they are indistinguishable.*)

As the pointwise limit of the f_j, the de Rham function f satisfies the two-scale difference equation

$$f(x)=f(3x)+\tfrac{1}{3}[f(3x+1)+f(3x-1)]+\tfrac{2}{3}[f(3x+2)+f(3x-2)]. \tag{6.1}$$

The corresponding trigonometric polynomial is given by

$$p(\xi)=e^{-2i\xi}[(1+e^{i\xi}+e^{2i\xi})/3][(2-e^{i\xi}+2e^{2i\xi})/3];$$

it is not clear how to deduce the continuity of f from this expression for p! In part II we shall use a time-domain method to prove that f is Hölder continuous, with exponent $\gamma=1-\ln 2/\ln 3=.36907\cdots$ but is nowhere differentiable. The method of part II also allows us to analyze local properties of f, to show that there exist fractal sets with nonzero Hausdorff dimension, but zero Lebesgue measure, on which f is "almost" differentiable, in the sense that the local Hölder exponent can be chosen arbitrarily close to 1. (The choice of the fractal set depends on the desired Hölder exponent.)

A variant on the de Rham function is obtained by choosing $k=3$, $c_0=1$, $c_1=c_{-1}=\frac{1}{2}-\alpha$, $c_2=c_{-2}=\frac{1}{2}+\alpha$, all other $c_n\equiv 0$. The corresponding f_j^α and f^α are plotted in Fig. 2; for $\alpha=\frac{1}{6}$ we obviously revert to the de Rham case. The analysis of part II will show

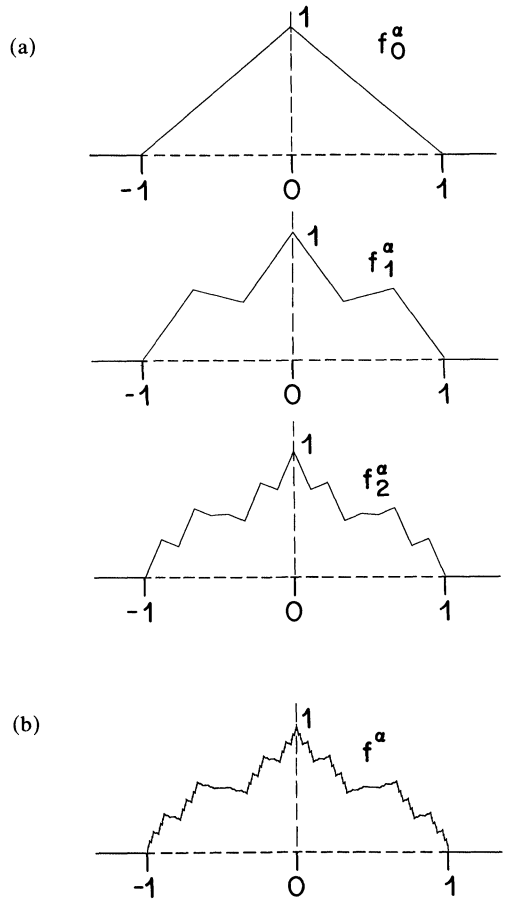


FIG. 2. (a) *The first three approximations* f_j^α, $j=0,1,2$ *for the generalized de Rham function corresponding to* $\alpha=\frac{1}{12}$. (b) *The generalized de Rham function* f^α *itself.*

that for sufficiently small α, $\alpha(1+2\alpha)^2<\frac{2}{27}$, or $\alpha<.0492\cdots$, the resulting function f^α is Lipschitz almost everywhere.

Note that for any α we have $N=4$, hence $N_0=1$, so that the matrix M reduces to the scalar 1. It therefore follows immediately from Theorem 5.1 that f and f^α cannot possibly be C^1, since then M would have at least the two eigenvalues 1 and $\frac{1}{2}$.

6.2. The Lagrange interpolation functions of Deslauriers and Dubuc. These functions are obtained by choosing an integer, $k>1$, called the "base" of the interpolation scheme, and an even integer M, $M\geqq 2$, called the "number of nodes," in the language of Deslauriers and Dubuc (1989). The interpolation function is then defined by the recursive process

$$f[k^{-j}(km+n)]=\sum_{m'=-M/2+1}^{M/2}\beta_{m',n}f[k^{-(j-1)}(m+m')],$$

where

$$\sum_{m'=-M/2+1}^{M/2}\beta_{m',n}=1 \quad \text{for all } n=1,\cdots,k-1. \tag{6.2}$$

This corresponds to a two-scale equation of the type

$$f(x)=f(kx)+\sum_{m=-M/2}^{M/2-1}\sum_{m=1}^{k-1}\alpha_{m,n}f(kx-km-n)$$

where $\alpha_{mn}=\beta_{-m,n}$. The $\beta_{m,n}$, or α_{mn}, are determined by (6.2), by the requirement that $p(\xi)$ be divisible by as many factors $[1+e^{i\xi}+\cdots+e^{i(k-1)\xi}]$ as possible, and by the symmetry condition $\beta_{-l,n}=\beta_{l+1,k-n}$, $0\leqq l\leqq M/2$, $n=1,\cdots,k-1$. For base 2, with four nodes ($k=2$, $M=4$), this leads to the two-scale difference equation

$$f(x)=f(2x)+\tfrac{9}{16}[f(2x+1)+f(2x-1)]-\tfrac{1}{16}[f(2x+3)+f(2x-3)], \tag{6.3}$$

corresponding to

$$\begin{aligned}p(\xi)&=e^{-3i\xi}[(1+e^{i\xi})/2]^4[-\tfrac{1}{2}+2e^{i\xi}-\tfrac{1}{2}e^{2i\xi}]\\&=(\cos\xi/2)^4(2-\cos\xi).\end{aligned}$$

Using $\sup_\xi|2-\cos\xi|=3$, we obtain $|\hat{f}(\xi)|\leqq C|\xi|^{-4+\log_2 3}$ (see, e.g., Lemma 3.2 in Daubechies (1988)), from which it follows that $f\in C^1$. We can bound the regularity of $f(x)$ by Theorem 5.1. We have $N=6$, $k=2$, so that $N_0=5$. The spectrum of the 5×5 matrix M is in this case $\{1,\frac{1}{2},\frac{1}{4},\frac{1}{8}\}$ where the eigenvalue $\frac{1}{4}$ has multiplicity 2. By Theorem 5.1 we know therefore that f can be at most C^3. In fact, however, f is not even C^2, as is shown in Deslauriers and Dubuc (1989), using the infinite product formula for $\hat{f}$ together with the special property that $p(\xi)\geqq 0$. They show that f is "almost" C^2, in the sense that f' is Hölder continuous with exponent $1-\varepsilon$ (ε arbitrarily small), but not C^1. In Dubuc (1986) the sharper estimate $|f(x)-f(x+t)|\leqq C|t|\log(1/|t|)$ is proved, for small enough t. This same example was also studied by Dyn, Gregory, and Levin (1987), with more general weights in (6.3) ($\frac{1}{2}+w$, w instead of $\frac{9}{16}$, $\frac{1}{16}$). For the parameters fixed as in (6.3), their results are slightly weaker than Dubuc's. For a thorough and detailed analysis of this example we refer to Dubuc (1986), Dyn, Gregory, and Levin (1987), or to part II.

6.3. Orthonormal bases of compactly supported wavelets. A family of wavelets is generated by translating and dilating one single function,

$$\psi_{jk}(x)=2^{-j/2}\psi(2^{-j}x-k), \qquad j,k\in\mathbb{Z}.$$

For some choices of ψ, the family ψ_{jk} constitutes an orthonormal basis of $L^2(\mathbb{R})$. One such choice is

$$\psi(x) = \begin{cases} 1, & 0 \leqq x < \frac{1}{2}, \\ -1, & \frac{1}{2} \leqq x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding orthonormal basis is well known; it is called the Haar basis, and provides an unconditional basis for all L^p-spaces, $1 < p < \infty$. Recently some other, more interesting choices for ψ have been found. The first one was constructed by Stromberg (1982); later Meyer (1985/86) constructed independently another wavelet basis, which was extended to higher dimensions by Lemarié and Meyer (1986). In the Meyer construction $\hat{\psi}$ is C^∞ and compactly supported; the basis $\{\psi_{jk}\}$ is not only an orthonormal basis for $L^2(\mathbb{R})$, but also an unconditional basis for all the L^p-spaces $(1 < p < \infty)$, the Sobolev spaces, the Besov spaces, etc. Later Battle (1987) and Lemarié (1987) constructed other orthonormal bases of wavelets, based on ψ which have faster (exponential) decay; their examples are K times continuously differentiable (K arbitrarily large, but finite). Mallat (1989) and Meyer (1986), (1990) devised a scheme into which all these constructions fit naturally, which they call *multiresolution analysis.* Finally, Daubechies (1988) constructed orthonormal bases of wavelets generated by *compactly supported* ψ which are K times differentiable.

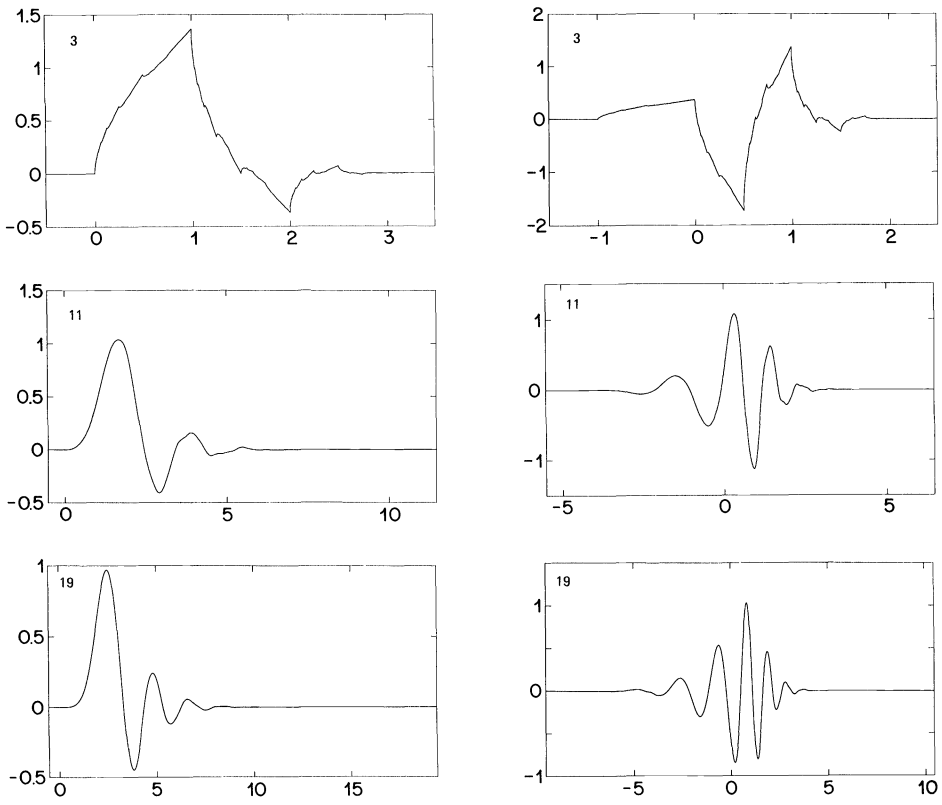


FIG. 3. *Some examples of orthonormal wavelet bases with compact support constructed in Daubechies* (1988). *In every case both* ϕ *and* ψ *are plotted. The number of nonvanishing* c_n *is, respectively,* 4, 12, *and* 20, *corresponding to support widths of, respectively,* 3, 11, *and* 19.

A typical construction of an orthonormal basis of wavelets uses an auxiliary function ϕ such that

$$\phi(x)=\sum c_n\phi(2x-n). \tag{6.4}$$

Provided the $\phi(x-n)$ are an orthonormal set, the function ψ is then given by

$$\psi(x)=\sum_n (-1)^n c_{n+1}\phi(2x+n).$$

(If the functions $\phi(x-n)$ are not orthonormal, we first construct $\tilde{\phi}$ by $\hat{\tilde{\phi}}(\xi)=\hat{\phi}(\xi)\times(\sum_{m\in\mathbb{Z}}|\hat{\phi}(\xi+2\pi m)|^2)^{-1/2}$; the $\tilde{\phi}(x-n)$ are orthonormal, and satisfy an equation of type (6.4), with different $\tilde{c}_n$; for more details see Daubechies (1988) or Mallat (1989)).

A construction of ϕ using only finitely many c_n results in a compactly supported ϕ (see § 2), and therefore a compactly supported ψ. As a finite linear combination of translated and dilated versions of ϕ, ψ has the same regularity as ϕ. It follows that a good understanding of the regularity of solutions of finite two-scale difference equations is important in the construction of orthonormal bases of compactly supported wavelets. The examples constructed in Daubechies (1988) have the property that their support width increases linearly with their regularity. This is illustrated by Fig. 3, which shows the pairs ϕ, ψ for support widths 3, 11, and 19, respectively. It is clear that ϕ, ψ become more regular as their width increases; Daubechies (1988) showed that there exists $\mu>0$ such that

$$\phi_N,\ \psi_N\in C^{\mu N} \quad \text{where } |\operatorname{supp}\phi_N|=|\operatorname{supp}\psi_N|=N.$$

The question then arose whether this linear increase of the support width was necessary. This question is now answered affirmatively by Theorem 5.1: if $\phi\in C^K$, then $|\operatorname{supp}\phi|\geqq K+2$. This also provides a simple proof for the (known) fact that it is impossible to construct wavelet bases generated by a compactly supported C^∞-function ϕ.

Appendix.

PROPOSITION. *Suppose f is a compactly supported function in $\operatorname{Lip}^{L,\alpha}$. Define functions f_j by:*

(1) *On every interval $[n2^{-j},(n+1)2^{-j}]$, f_j is a polynomial of degree $2L+1$.*

(2) *f_j is in C^L and*

$$f_j^{(l)}(n2^{-j})=f^{(l)}(n2^{-j}) \quad \text{for } l=0,\cdots,L, \quad n\in\mathbb{Z}.$$

Then $\|f^{(l)}-f_j^{(l)}\|_{L^\infty}\leqq C2^{-j(L-l+\alpha)}$ for $l=0,\cdots,L$, and for some C independent of j and l.

Proof. (1) Choose $x\in\operatorname{support}(f)$, $j\in\mathbb{N}$ arbitrary. Find n so that $2^{-j}n\leqq x\leqq 2^{-j}(n+1)$. Then

$$\begin{aligned}|f^{(l)}(x)-f_j^{(l)}(x)|\leqq{}&\left|f^{(l)}(x)-\sum_{k=l}^{L}\frac{1}{(k-l)!}f^{(k)}(2^{-j}n)(x-2^{-j}n)^{(k-l)}\right|\\&+\left|f_j^{(l)}(x)-\sum_{k=l}^{L}\frac{1}{(k-l)!}f^{(k)}(2^{-j}n)(x-2^{-j}n)^{(k-l)}\right|.\end{aligned} \tag{A.1}$$

Since $f\in\operatorname{Lip}^{L,\alpha}$ and $\operatorname{support}(f)$ is finite, the first term is bounded by $C|x-2^{-j}n|^{\alpha+(L-l)}\leqq C2^{-j(\alpha+L-l)}$ with C independent of x or j. It therefore suffices to bound the second term.

(2) On $[2^{-j}n,2^{-j}(n+1)]$ we have

$$f_j(y)=\sum_{l=0}^{L}\frac{1}{l!}f^{(l)}(2^{-j}n)(y-2^{-j}n)^l+\sum_{l=0}^{L}\frac{1}{(L+1+l)!}a_{n,l}^j(y-2^{-j}n)^{L+1+l}$$

with the $a^j_{n,l}$ determined by the $L+1$ equations, $0 \leqq m \leqq L$,

$$\sum_{l=0}^{L} \frac{1}{(L+1+l-m)!} a^j_{n,l} 2^{-j(L+1+l-m)} = f^{(m)}(2^{-j}(n+1)) - \sum_{l=m}^{L} \frac{1}{(l-m)!} f^{(l)}(2^{-j}n) 2^{-j(l-m)},$$

or

$$\sum_{l=0}^{L} \frac{1}{(L+1+l-m)!} a^j_{n,l} 2^{-j(l+1)} = b^j_{n,m} \tag{A.2}$$

where

$$b^j_{n,m} = 2^{-j(L-m)} \left[f^{(m)}(2^{-j}(n+1)) - \sum_{l=m}^{L} \frac{1}{(l-m)!} f^{(l)}(2^{-j}n) 2^{-j(l-m)} \right]$$

is bounded, uniformly in n, by $C2^{-j\alpha}$ because $f \in \mathrm{Lip}^{L,\alpha}$ and f is compactly supported. It follows, by inverting the system (A.2), that

$$|a^j_{n,l} 2^{-j(l+1)}| \leqq C 2^{-j\alpha}.$$

(3) Consequently,

$$\begin{aligned} &\left| f_j^{(l)}(x) - \sum_{k=1}^{L} \frac{1}{(k-l)!} f^{(k)}(2^{-n}j)(x-2^{-j}n)^{(k-l)} \right| \\ &\quad = \left| \sum_{k=0}^{L} \frac{1}{(L+1+k-l)!} a^j_{n,k} (x-2^{-j}n)^{L+1+k-l} \right| \\ &\quad \leqq \sum_{k=0}^{L} \frac{1}{(L+1+k-l)!} C 2^{j(k+1)} 2^{-j\alpha} 2^{-j(L+1+k-l)} \\ &\quad \leqq C 2^{-j(L-l+\alpha)}. \end{aligned}$$

Hence $(A.1) \leqq C2^{-j(L-l+\alpha)}$ for all x, with C independent of j, l, or x, and the proposition is proved. □

Acknowledgments. The authors are indebted to C. A. Micchelli and to anonymous referees for supplying many references, and for suggesting the discussion on approximation by splines given in § 4. A helpful conversation with A. M. Odlyzko led to the proof of Lemma 3.2. After completing this work, we learned that some results have also been derived by P. Auscher (1989).

REFERENCES

P. AUSCHER (1989), *Ondelettes fractales et applications*, Ph.D. thesis, Université de Paris, Dauphine, France.

M. F. BARNSLEY AND S. DEMKO (1985), *Iterated function systems and the global construction of fractals*, Proc. Roy. Soc. London Ser. A, 399, pp. 243-275.

G. BATTLE (1987), *A block spin construction of ondelettes* I: *Lemarie functions*, Comm. Math. Phys., 110, pp. 601-615.

G. BROWN AND W. MORAN (1973), *A dichotomy for infinite products of discrete measures*, Proc. Cambridge Philos. Soc., 73, pp. 307-316.

A. S. CAVARETTA AND C. A. MICCHELLI (1989), *The design of curves and surfaces by subdivision algorithms*, in Mathematical Methods in Computer-Aided Geometric Design, T. Lyche and L. Schumaker, eds., Academic Press, New York, pp. 113-153.

G. M. CHAIKIN (1974), *An algorithm for high speed curve generation*, Comput. Graphics Image Process., 3, pp. 346-349.

W. DAHMEN AND C. A. MICCHELLI (1984), *Subdivision algorithms for the generation of box spline surfaces*, Comput. Aided Geom. Des., 1, pp. 115-129.

W. DAHMEN AND C. A. MICCHELLI (1988), *Local spline interpolation schemes in one and several variables*, in Approximation and Optimization, Lecture Notes in Math. 1354, Springer-Verlag, New York, pp. 11-24.

I. DAUBECHIES (1988), *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41, pp. 909-996.

I. DAUBECHIES AND J. C. LAGARIAS (1988), *Two scale difference equations* II. *Local regularity, infinite products of matrices and fractals*, AT&T Bell Laboratories, preprint.

G. DE RHAM (1947) *Un peu de mathématiques à propos d'une courbe plane*, Elem. Math., 2, pp. 73-76, 89-97.

——— (1956), *Sur une courbe plane*, J. Math. Pures Appl. (9), 35, pp. 25-42.

——— (1957), *Sur un exemple de fonction continue sans dérivée*, Enseign. Math. (2), 3, pp. 71-72.

——— (1959), *Sur les courbes limites de polygones obtenus par trisection*, Enseign. Math. (2), 5, pp. 29-43.

G. DESLAURIERS AND S. DUBUC (1987), *Interpolation dyadique*, in Fractals: Dimensions Non Entières et Applications, G. Cherbit, ed., Masson, Paris, pp. 44-55.

——— (1989), *Symmetric iterative interpolation*, Constr. Approx., 5, pp. 49-68.

P. DIACONIS AND M. SHASHAHANI (1986), *Products of random matrices and computer image generation*, Contemp. Math., 50, pp. 173-182.

S. DUBUC (1986), *Interpolation through an iterative scheme*, J. Math. Anal. Appl., 114, pp. 185-204.

N. DYN, J. A. GREGORY, AND D. LEVIN (1987), *A 4-point interpolatory subdivision scheme for curve design*, Comput. Aided Geom. Des., 4, pp. 257-268.

——— (1989), *Analysis of uniform binary subdivision schemes for curve design*, in Algorithms for the Approximation of Functions, J. C. Cox and J. C. Mason, ed.

——— (1990), *Uniform subdivision algorithms for curves and surfaces*, Constr. Approx., to appear.

N. DYN AND D. LEVIN (1989), *Interpolating subdivision schemes for the generation of curves and surfaces*, to appear.

P. ERDÖS (1939), *On a family of symmetric Bernoulli convolutions*, Amer. J. Math., 61, pp. 974-976.

——— (1940), *On the smoothness properties of a family of symmetric Bernoulli convolutions*, Amer. J. Math., 62, pp. 180-186.

A. GARSIA (1962), *Arithmetic properties of Bernoulli convolutions*, Trans. Amer. Math. Soc., 102, pp. 409-432.

G. JESSEN AND A. WINTNER (1935), *Distribution functions and the Riemann zeta function*, Trans. Amer. Math. Soc., 38, pp. 48-88.

R. KERSHNER AND A. WINTNER (1935), *On symmetric Bernoulli convolutions*, Amer. J. Math., 57, pp. 541-548.

J. M. LANE AND R. F. RIESENFELD (1980), *A theoretical development for the computer generation of piecewise polynomial surfaces*, IEEE Trans. Pattern Anal. Machine Intelligence, 2, pp. 35-46.

P. G. LEMARIÉ (1987), *Ondelettes à localisation exponentielle*, J. Math. Pures Appl., 67, pp. 227-236.

P. G. LEMARIÉ AND Y. MEYER (1986), *Ondelettes et bases Hilbertiennes*, Rev. Mat. Iberoamericana, 2, pp. 1-18.

S. MALLAT (1989), *Multiresolution approximation and wavelets*, Trans. Amer. Math. Soc., 315, pp. 69-88.

Y. MEYER (1985/86), *Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs*, Seminar Bourbaki No. 662.

——— (1986), *Ondelettes et fonctions splines*, Séminaire EDP, Ecole Polytechnique, Paris, France.

——— (1987), Private communication.

——— (1990), *Ondelettes et Operateurs* I, II, Hermann, Paris.

C. A. MICCHELLI (1986), *Subdivision algorithms for curves and surfaces*, in Proc. SIGGRAPH '86, September 4, 1986.

C. MICCHELLI AND H. PRAUTZSCH (1987a), *Refinement and subdivision for spaces of integer translates of compactly supported functions*, in Numerical Analysis, D. F. Griffith and G. A. Watson, eds., Academic Press, New York, pp. 192-222.

——— (1987b), *Computing curves invariant under halving*, Comput. Aided Geom. Des., 4, pp. 133-140.

——— (1989), *Uniform refinement of curves*, Linear Algebra Appl., 114/115, pp. 841-870.

M. REED AND B. SIMON (1975), *Methods of Mathematical Physics*, Vol. II. *Fourier Analysis, Self-Adjointness*, Academic Press, New York.

I. J. SCHOENBERG (1973) *Cardinal Spline Interpolation*, CBMS-NSF Regional Conference Series in Applied Mathematics 12, Society for Industrial and Applied Mathematics, Philadelphia, PA.

L. L. SCHUMAKER (1981), *Spline Functions: Basic Theory*, John Wiley, New York.

J. O. STROMBERG (1982), *A modified Haar system and higher order spline systems*, in Proc. Conference in Harmonic Analysis in Honor of A. Zygmund, Vol. II, W. Beckner et al., eds., Wadsworth Mathematical Series, pp. 475-493.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1411-1429, September 1991

013

SPECIAL FUNCTIONS AND SINGULAR QUASILINEAR PARTIAL DIFFERENTIAL EQUATIONS*

VICTOR L. SHAPIRO†

Abstract. For Qu a second-order singular quasilinear elliptic operator, the notion of a first eigenvalue is introduced. The terminology singular arises from the fact that the coefficients of Q may be zero on part or all of the boundary or the region of definition may be unbounded. A one-sided nonlinear result for Q is established below the first eigenvalue, and a theorem and corollary (called results at resonance) are established at the first eigenvalue. The corollary presents a condition which is both necessary and sufficient. Examples are given involving Hermite polynomials, Bessel functions, and other special functions.

Key words. quasilinear partial differential equations, special functions, resonance, first eigenvalue

AMS(MOS) subject classifications. 35J60, 33A65

1. Introduction. Let $\Omega \subset \mathbb{R}^N$, $N \geqq 1$, be an open (possibly unbounded) connected set and let p and ρ be two positive functions in Ω with $p \in C^1(\bar{\Omega})$ and $\rho \in C^0(\Omega)$. We set

$$Qu(x) = -D_i[p(x)A_i(x, u, Du)] + \rho(x)B_0(x, u, Du)u \tag{1.0}$$

where $D_i = \partial/\partial x_i$ and $D = (D_1, \cdots, D_N)$. We shall study the singular quasi-linear equation

$$Qu(x) = \rho(x)f(x, u) + G. \tag{1.1}$$

The term singular arises from the fact that Ω may be unbounded or that p may equal zero on part or all of the boundary of Ω. A similar situation occurs for the singular Sturm-Liouville system studied in elementary differential equations (see [1, p. 57], [2, p. 554], or [3, p. 560]).

In (1.0) and for the rest of the paper the summation convention is used for $i = 1, \cdots, N$. Throughout we shall suppose

$$p \in C^0(\bar{\Omega}), \quad p > 0 \quad \text{in } \Omega, \quad \int_\Omega p < \infty, \quad \text{and } p = 0 \quad \text{on } \Gamma_1 \subset \partial\Omega \tag{1.2}$$

where Γ_1 may be the empty set, i.e., $\Gamma_1 = \{x \in \partial\Omega: p(x) = 0\}$.

$$\rho \in C^0(\Omega), \quad \rho > 0 \quad \text{in } \Omega, \quad \text{and} \int_\Omega \rho < \infty. \tag{1.3}$$

In order to explain the type of problem we intend to study for (1.1), we have to introduce some Hilbert spaces. In particular, we let $C^0_\rho(\Omega)$ and $C^1_{p,\rho}(\Omega)$ be the following two pre-Hilbert spaces:

$$C^0_\rho(\Omega) = \left\{ u \in C^0(\Omega): \int_\Omega u^2\rho < \infty \right\} \tag{1.4}$$

* Received by the editors December 27, 1989; accepted for publication (in revised form) September 27, 1990.

† Department of Mathematics, University of California, Riverside, California 92521.

with $\langle u, v\rangle_\rho = \int_\Omega uv\rho$;

$$(1.5)\quad C^1_{p,\rho}(\Omega) = \left\{ u \in C^0(\bar{\Omega}) \cap C^1(\Omega) \colon u = 0 \text{ on } \partial\Omega \backslash \Gamma_1 \text{ and } \int_\Omega |Du|^2 p + u^2 \rho < \infty \right\}$$

with $\langle u, v\rangle_{p,\rho} = \int_\Omega [Du \cdot Dvp + uv\rho]$.

$L^2_\rho(\Omega)$ will designate the Hilbert space that we obtain by completing $C^0_\rho(\Omega)$ using the method of Cauchy sequences with respect to the norm $\|u\|_\rho = \langle u, u\rangle^{1/2}_\rho$, and $H^1_{p,\rho}(\Omega)$ will designate the Hilbert space that we obtain by completing $C^1_{p,\rho}(\Omega)$ with respect to the norm $\|u\|_{p,\rho} = \langle u, u\rangle^{1/2}_{p,\rho}$.

We define $L^q_\rho(\Omega)$ in a similar manner for $1 \leqq q < \infty$ and we also observe that $L^2_p(\Omega)$ is a well-defined space.

Continuing, we make the following assumptions for the functions $A_i(x, t, \xi)$, $i = 1, \cdots, N$, and $B_0(x, t, \xi)$ where $x \in \Omega$, $t \in \mathbb{R}$, and $\xi \in \mathbb{R}^N$.

(Q1) $A_i(x, t, \xi) : \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ and $B_0(x, t, \xi) : \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$, and both satisfy the Caratheodory conditions (i.e., $A_i(x, t, \xi)$ is measurable for x in Ω for every fixed $(t, \xi) \in \mathbb{R} \times \mathbb{R}^N$ and is continuous in (t, ξ) for almost every fixed $x \in \Omega$, and similarly, for $B_0(x, t, \xi)$).

(Q2) There exist $h \geqq 0$ and $c_1 \geqq 0$ with $h \in L^2_p(\Omega)$ and c_1 a constant such that, respectively,

$$|A_i(x, t, \xi)| \leqq h(x) + c_1|\xi| \quad \text{for a.e. } x \in \Omega \quad \text{or}$$

$$|A_i(x, t, \xi)| \leqq h(x) + c_1(t^2 + |\xi|^2)^{1/2} \quad \text{for a.e. } x \in \Omega$$

accordingly as $L^2_\rho(\Omega)$ is not or is continuously imbedded in $L^2_p(\Omega)$ (i.e., the latter means $L^2_\rho(\Omega) \subset L^2_p$ and $\|u\|_p \leqq K\|u\|_\rho$ for all $u \in L_\rho$).

(Q3) There exists a positive constant c_2 such that $|B_0(x, t, \xi)| \leqq c_2$ for almost every $x \in \Omega$ and for all $(t, \xi) \in \mathbb{R} \times \mathbb{R}^N$.

(Q4) There exists a positive constant c_3 and a nonnegative function $Z \in L^1_p(\Omega)$ subject to $A_i(x, t, \xi)\xi_i \geqq c_3|\xi|^2 - Z(x)$ for almost every $x \in \Omega$ and for all $(t, \xi) \in \mathbb{R} \times \mathbb{R}^N$.

(Q5) $[A_i(x, t, \xi) - A_i(x, t, \xi')](\xi_i - \xi'_i) > 0$ for almost every $x \in \Omega$, for all $t \in \mathbb{R}^N$, and for all $\xi, \xi' \in \mathbb{R}^N$ with $\xi \neq \xi'$.

Next, we introduce the semilinear form

$$(1.6)\quad \mathcal{Q}(u, v) = \int_\Omega pA_i(x, u, Du)D_i v + \rho B_0(x, u, Du)uv$$

and we see from (Q1)-(Q3) that $\mathcal{Q}(u, v)$ is well defined for $u, v \in H^1_{p,\rho}$. Also, we define

$$(1.7)\quad \lambda_1^* = \liminf_{\|u\|_\rho \to \infty} \mathcal{Q}(u, u)/\|u\|^2_\rho, \qquad u \in H^1_{p,\rho}.$$

It is clear that if $\mathcal{Q}$ were a linear elliptic operator and p and ρ were equal to one, λ_1^* would be the first eigenvalue associated with the Dirichlet problem (see [7, p. 213]). An easy computation using (Q2)-(Q4) shows that λ_1^* is a finite-valued real number.

The first theorem that we shall present here will involve the interplay between the left-hand side and the right-hand side of (1.1), i.e., between Q and f, at values strictly below λ_1^*. In particular, we shall assume the following:

(f1) $f(x, t)$ meets the usual Caratheodory condition.

(f2) There exist $\tilde{h}_1 \geqq 0$ and $\varepsilon_0 > 0$ with $\tilde{h}_1 \in L^2_\rho(\Omega)$ subject to

$$tf(x, t) \leqq (\lambda_1^* - \varepsilon_0)t^2 + \tilde{h}_1(x)|t| \quad \forall t \in \mathbb{R} \text{ and a.e. } x \in \Omega.$$

(f3) There exist $K > 0$ and $\tilde{h}_2 \geqq 0$ with $\tilde{h}_2 \in L^2_\rho(\Omega)$ subject to

$$|f(x, t)| \leqq \tilde{h}_2(x) + K|t| \quad \forall t \in \mathbb{R} \text{ and a.e. } x \in \Omega.$$

We shall also assume that $G \in [H^1_{p,\rho}(\Omega)]^*$, i.e., $G: H^1_{p,\rho}(\Omega) \to \mathbb{R}$, G is linear, and there exist K_1 subject to

$$|G(u)| \leqq K_1 \|u\|_{p,\rho} \quad \forall u \in H^1_{p,\rho}. \tag{1.8}$$

Next, we state the first theorem we shall prove in this paper. This statement will involve an assumption labeled $O_{p,\rho}(\Omega)$ which will be explained and illustrated using special functions in the paragraphs following the theorem.

THEOREM 1. *Let $\Omega \subset \mathbb{R}^N$, $N \geqq 1$, be an open connected set and let p and ρ satisfy* (1.2) *and* (1.3). *Assume* (Q1)-(Q5), (f1)-(f3), $O_{p,\rho}(\Omega)$, *and that $G \in [H^1_{p,\rho}(\Omega)]^*$. Then there exist $u^* \in H^1_{p,\rho}(\Omega)$ such that*

$$\mathcal{Q}(u^*, v) = \langle f(\cdot, u^*), v\rangle_\rho + G(v) \quad \forall v \in H^1_{p,\rho}(\Omega) \tag{1.9}$$

where $\mathcal{Q}(u^*, v)$ is defined by (1.6).

It is easy to give examples to show that the above theorem is a best possible result, i.e., (1.9) is false if we allow ε_0 to be zero in (f2). We shall do this in the concluding paragraphs of this section.

By assumption $O_{p,\rho}(\Omega)$ we shall mean that the following three conditions (O1)-(O3) hold

(O1) There exists a complete orthonormal sequence of functions $\{\psi_n\}_{n=1}^\infty$ in $L^2_\rho(\Omega)$. Also $\psi_n \in H^1_{p,\rho}(\Omega)$ for all n.

(O2) There exist $\{\lambda_n\}_{n=1}^\infty$ with $1 \leqq \lambda_1 < \lambda_2 \leqq \lambda_3 \leqq \cdots \leqq \lambda_n \leqq \cdots \to \infty$ subject to

$$\langle \psi_n, v\rangle_{p,\rho} = \lambda_n \langle \psi_n, v\rangle_\rho \quad \forall v \in H^1_{p,\rho}(\Omega).$$

(O3) $\psi_1 > 0$ almost everywhere in Ω.

Now there are many unbounded (as well as bounded) open connected sets Ω that satisfy assumption $O_{p,\rho}(\Omega)$. We illustrate this fact using the theory of special functions.

First and first foremost, we take $\Omega = \mathbb{R}^2$ (with a similar situation prevailing in $\mathbb{R}^N$, $N \geqq 1$), set $\rho(x) = p(x) = e^{-(x_1^2 + x_2^2)}$, and take

$$\phi_{mn}(x_1, x_2) = H_m(x_1)H_n(x_2)/(2^{-m}2^{-n}m!\,n!\,\pi)^{1/2} \tag{1.10}$$

where $H_n(t) = (-1)^n e^{t^2} d^n e^{-t^2}/dt^n$ is the familiar Hermite polynomial for $n = 0, 1, 2, \cdots$. As is well known, $\{H_n(t)/(2^n n!\,\pi^{1/2})^{1/2}\}_{n=0}^\infty$ forms a complete orthonormal system over $\mathbb{R}^1$ with respect to the weight e^{-t^2}. Hence it follows from (1.10) that $\{\phi_{mn}\}_{m=0,n=0}^{\infty,\infty}$ forms a CONS for $L^2_\rho(\mathbb{R}^2)$ where $\rho = e^{-(x_1^2+x_2^2)}$ and therefore (O1) holds. Also, $[e^{-t^2}H_n'(t)]' = -2n\,e^{-t^2}H_n(t)$. An easy computation, using this last fact, shows that if $v \in C^1_{p,\rho}(\Omega)$ as defined by (1.5), then with $D_1 = \partial/\partial x_1$,

$$\int_{\mathbb{R}^2} e^{-(x_1^2+x_2^2)} H_n(x_2) D_1 H_m(x_1) D_1 v(x_1, x_2) = 2m \int_{\mathbb{R}^2} e^{-(x_1^2+x_2^2)} H_n(x_2) H_m(x_1) v(x_1, x_2).$$

From this fact, it follows that

$$\langle \phi_{mn}, v\rangle_{p,\rho} = [2(m+n)+1]\langle \phi_{mn}, v\rangle_\rho \quad \forall v \in H^1_{p,\rho}.$$

Consequently, we see that (O2) holds with $\lambda_1 = 1$, $\lambda_2 = 3$, $\lambda_3 = 3$, $\lambda_4 = 5$, etc. Also, we see from (1.10) that $\psi_1 = \pi^{-1/2}$; hence (O3) holds and $O_{p,\rho}(\Omega)$ is established for this case.

For the second example, we work in $\mathbb{R}^3$, take $\tilde{\Omega} \subset \mathbb{R}^2$ to be a bounded open connected set, and $\Omega = \tilde{\Omega} \times \mathbb{R}^1$. So Ω is an infinite cylinder whose axis is parallel to the x_3-axis. We take $p(x) = \rho(x) = e^{-x_3^2}$ and

$$\Phi_{mn}(x) = \phi_m(x_1, x_2) H_n(x_3)/(2^n n! \pi^{1/2})^{1/2}, \qquad m = 1, 2, \cdots, n = 0, 1, \cdots, \tag{1.11}$$

where $\{\phi_m(x_1, x_2)\}_{m=1}^{\infty}$ and $\{\eta_m\}_{m=1}^{\infty}$ are the familiar eigenfunctions and eigenvalues of the Laplace operator in $W_0^{1,2}(\tilde{\Omega})$. Since $\{\phi_m(x_1, x_2)\}_{m=1}^{\infty}$ is a CONS for $L^2(\tilde{\Omega})$ and $\{H_n(x_3)/(2^n n! \pi^{1/2})^{1/2}\}_{n=0}^{\infty}$ is a CONS over $\mathbb{R}^1$ with respect to the weight $e^{-x_3^2}$, it follows that $\{\Phi_{mn}\}_{m=1,n=0}^{\infty,\infty}$ is a CONS for $L_\rho^2(\Omega)$. Therefore (O1) holds.

Using the properties alluded to in the first example, an easy computation shows that

$$\begin{aligned} &\int_\Omega e^{-x_3^2} \phi_m(x_1, x_2) D_3 H_n(x_3) D_3 v(x) \\ &\qquad = 2n \int_\Omega e^{-x_3^2} \phi_m(x_1, x_2) H_n(x_3) v(x) \quad \forall v \in C_{p,\rho}^1(\Omega). \end{aligned} \tag{1.12}$$

Hence, it follows from (1.11) that

$$\langle \Phi_{mn}, v \rangle_{p,\rho} = (\eta_m + 2n + 1) \langle \Phi_{mn}, v \rangle_\rho$$

for all $v \in H_{p,\rho}^1$. Consequently, we see (O2) holds with $\lambda_1 = \eta_1 + 1$ and $\lambda_2 = \eta_2 + 1$ or $\eta_1 + 3$, depending on which is smaller. As is well known [7, p. 214], $\phi_1(x_1, x_2) > 0$ almost everywhere in $\tilde{\Omega}$. Hence $\psi_1(x_1, x_2, x_3) = \phi_1(x_1, x_2) \pi^{-1/2}$ and we see that (O3) holds. Consequently, $O_{p,\rho}(\Omega)$ is established.

For our third example, we work in $\mathbb{R}^2$ and take $\Omega = (0, 1) \times \mathbb{R}^1$ with $p(x_1, x_2) = x_1 e^{-x_2^2} = \rho(x_1, x_2)$. Also we set

$$\phi_{mn}(x_1, x_2) = J_0(\eta_m x_1) H_n(x_2)/[2^{-1} J_1(\eta_m) 2^n n! \pi^{1/2}]^{1/2} \tag{1.13}$$

for $m = 1, 2, \cdots, n = 0, 1, 2, \cdots$, where $J_0(t)$ and $J_1(t)$ are the familier Bessel functions of the first kind and η_m is the mth positive zero of $J_0(t)$. As is well known [8, p. 264], $\{J_0(\eta_m t)/[2^{-1} J_1(\eta_m)]^{1/2}\}_{m=1}^{\infty}$ is CONS with respect to weight t on $[0, 1]$. Hence, it follows that $\{\phi_{mn}\}_{m=1,n=0}^{\infty,\infty}$ is a CONS for $L_\rho^2(\Omega)$. Observing that

$$\int_\Omega e^{-x_2^2} H_n(x_2) x_1 D_1 J_0(\eta_m x_1) D_1 v(x) = \eta_m^2 \int_\Omega e^{-x_2^2} x_1 H_n(x_2) J_0(\eta_m x_1) v(x)$$

for all $v \in C_{p,\rho}^1(\Omega)$ and using the analogue of (1.12) for the current situation, we see from (1.13) that

$$\langle \phi_{mn}, v \rangle_{p,\rho} = (\eta_m^2 + 2n + 1) \langle \phi_{mn}, v \rangle_\rho$$

for all $v \in H_{p,\rho}^1$. Consequently, we see that (O2) holds with $\lambda_1 = \eta_1^2 + 1$, $\lambda_2 = \eta_1^2 + 3$, $\lambda_3 = \eta_1^2 + 5$, etc. Since $J_0(\eta_1 t) > 0$ in $[0, 1)$, we see that $\psi_1(x_1, x_2) = J_0(\eta_1 x_1) \pi^{-1/2} > 0$ in Ω and (O3) holds. Therefore, $O_{p,\rho}(\Omega)$ is established.

For our fourth example, we continue to work in $\mathbb{R}^2$ and take $\Omega = (0, 1) \times (0, 1)$ with $p(x) = \rho(x) = x_1 x_2$. We set

$$\phi_{mn}(x_1, x_2) = J_0(\eta_m x_1) J_0(\eta_n x_2)/[2^{-2} J_1(\eta_m) J_1(\eta_n)]^{1/2}, \tag{1.14}$$

$m, n = 1, 2, \cdots$, where we use the notation of the previous example. Observing that

$$\int_\Omega x_1 x_2 J_0(\eta_n x_2) D_1 J_0(\eta_m x_1) D_1 v(x) = \eta_m^2 \int_\Omega x_1 x_2 J_0(\eta_m x_1) J_0(\eta_n x_2) v(x)$$

for all $v \in C^1_{p,\rho}(\Omega)$, we see, as in the previous example, that (01)-(03) hold here also with

$$\lambda_1 = 2\eta_1^2 + 1, \quad \lambda_2 = \eta_1^2 + \eta_2^2 + 1, \quad \lambda_3 = \eta_1^2 + \eta_2^2 + 1,$$

etc., and

$$\psi_1(x_1, x_2) = J_0(\eta_1 x_1) J_0(\eta_1 x_2) / 2^{-1} J_1(\eta_1).$$

Hence, $O_{p,\rho}(\Omega)$ is established for this case.

In $\mathbb{R}^1$ we see that $O_{p,\rho}(\Omega)$ holds for the following situations:

$$\Omega = (0, \infty), \quad p(x) = x^{\alpha+1} e^{-x}, \quad \rho(x) = x^{\alpha} e^{-x}, \quad \alpha > -1, \tag{1.15}$$

$$\Omega = (-1, 1), \quad p(x) = (1-x)^{\alpha+1}(1+x)^{\beta+1}, \quad \rho = (1-x)^{\alpha}(1+x)^{\beta}, \quad \alpha, \beta > -1. \tag{1.16}$$

For (1.15) we use the Laguerre polynomials and for (1.16) we use the Jacobi polynomials. For the details to be used to show that (O1)-(O3) hold in these cases, we refer the reader to [14].

The terminology special functions usually refers to the Jacobi polynomials, Laguerre polynomials, Hermite polynomials, and Bessel functions. The above examples illustrate why we have used this terminology in the title of the paper.

In closing this section, we shall use the first example above to show that ε_0 cannot be taken equal to zero in (f2); i.e., the theorem as stated is in general false if $\varepsilon_0 = 0$ in (f2).

To see this fact, we take $\Omega = \mathbb{R}^2$ and use the first example from above with $p(x) = \rho(x) = e^{-(x_1^2 + x_2^2)}$. Then as we pointed out, $\lambda_1 = 1$ and $\psi_1 = \pi^{-1/2}$. We take $A_1(x, t, \xi) = \xi_1$, $A_2(x, t, \xi) = \xi_2$, and $B_0(x, t, \xi) = 1$. Therefore,

$$Qu = -[D_1 p D_1 u + D_2 p D_2 u] + pu$$

and we see that (Q1)-(Q5) holds. Also we see that

$$\mathcal{Q}(u, v) = \int_{\mathbb{R}^2} p(x)[D_1 u D_1 v + D_2 u D_2 v + uv] \tag{1.17}$$

for all $u, v \in H^1_{p,\rho}$. Since $p = \rho$ we obtain from (1.17) that $\mathcal{Q}(u, u)/\|u\|^2_\rho \geqq 1$ for $u \neq 0$. Hence it follows from (1.7) that $\lambda_1^* \geqq 1$. But $\psi_1 \in H^1_{p,\rho}$ and $\mathcal{Q}(\psi_1, \psi_1)/\|\psi_1\|^2_\rho = 1$. Therefore, $\lambda_1^* = 1 = \lambda_1$.

Next, we take $f(x, t) = t$ and observe that $f(x, t)$ meets (f1), (f3), and (f2) with $\varepsilon_0 = 0$. Also, we take $G(v) = \langle \psi_1, v \rangle_\rho$ for all $v \in H^1_{p,\rho}$. Suppose the theorem held under such circumstances. Then from (1.9) we see that there exist $u^* \in H^1_{p,\rho}(\mathbb{R}^2)$ such that

$$\mathcal{Q}(u^*, v) = \langle u^*, v \rangle_\rho + \langle \psi_1, v \rangle_\rho \quad \forall v \in H^1_{p,\rho}. \tag{1.18}$$

We take $v = \psi_1$ in (1.18) and observe from (1.17) and the fact $p = \rho$ that

$$\langle u^*, \psi_1 \rangle_\rho = \langle u^*, \psi_1 \rangle_\rho + \langle \psi_1, \psi_1 \rangle_\rho.$$

Therefore, $\langle \psi_1, \psi_1 \rangle_\rho = 0$. But $\langle \psi_1, \psi_1 \rangle_\rho = 1$. We have arrived at a contradiction. No such u^* as in (1.18) exists. Hence our theorem is in a certain sense best possible, as asserted.

The monotonicity assumption (Q5) above enables us to deal with the nonlinear aspect of Du in $A_i(x, u, Du)$ and this part of the paper was motivated by the classical 1970 article of Browder [5] and the 1965 result of Leray and Lions [11].

Theorem 2 of this paper is stated and proved in § 4.

2. Fundamental lemmas. The first lemma we prove is the following.

LEMMA 1. *Assume* (O1)-(O3) *and that* $g \in L^2_\rho(\Omega)$. *Set*

$$\hat{g}(n) = \langle \psi_n, g \rangle_\rho. \tag{2.1}$$

Then $g \in H^1_{p,\rho}$ *if and only if* $\sum_{n=1}^{\infty} \lambda_n |\hat{g}(n)|^2 < \infty$. *Furthermore, if* $g \in H^1_{p,\rho}$, $\|g\|^2_{p,\rho} = \sum_{n=1}^{\infty} \lambda_n |\hat{g}(n)|^2$.

To prove Lemma 1, we observe from (O1) and (O2) that $\langle \psi_n, \psi_n \rangle_{p,\rho} = \lambda_n > 0$. Also, we see that if $v \in H^1_{p,\rho}$ is such that $\langle \psi_n, v \rangle_{p,\rho} = 0$ for all n, then $\langle \psi_n, v \rangle_{\rho} = 0$ for all n. Since $\{\psi_n\}_{n=1}^{\infty}$ is CONS in L^2_ρ, it follows that $v = 0$ almost everywhere in Ω. Hence we conclude that

$$\{\psi_n / \lambda_n^{1/2}\}_{n=1}^{\infty} \quad \text{is CONS in } H^1_{p,\rho}. \tag{2.2}$$

It follows from (2.2) that if $g \in H^1_{p,\rho}$, then

$$\|g\|^2_{p,\rho} = \sum_{n=1}^{\infty} \langle g, \psi_n \rangle^2_{p,\rho} \lambda_n^{-1} < \infty. \tag{2.3}$$

But from (O2) and (2.1), we have $\langle g, \psi_n \rangle_{p,\rho} = \lambda_n \hat{g}(n)$. Consequently, we conclude from (2.3) that

$$\|g\|^2_{p,\rho} = \sum_{n=1}^{\infty} \lambda_n |\hat{g}(n)|^2 < \infty \tag{2.4}$$

and the only if part of the above lemma is established. Also the last line in the conclusion of the lemma is established.

To establish the if part, we assume

$$\sum_{n=1}^{\infty} \lambda_n |\hat{g}(n)|^2 < \infty, \tag{2.5}$$

and set $h_n = \sum_{k=1}^{n} \hat{g}(k)\psi_k = \sum_{k=1}^{n} \sqrt{\lambda_k}\, \hat{g}(k)\psi_k / \sqrt{\lambda_k}$. Then for $m > n$, $\|h_m - h_n\|^2_{p,\rho} = \sum_{n+1}^{m} \lambda_k |\hat{g}(k)|^2$. Hence it follows from (2.5) that $\{h_n\}_{n=1}^{\infty}$ is a Cauchy sequence in the Hilbert space $H^1_{p,\rho}$. Therefore there exist $h \in H^1_{p,\rho}$ such that $\|h_n - h\|_{p,\rho} \to 0$. But this implies that $\|h_n - h\|_{\rho} \to 0$. Since it follows from the definition of h_n and the fact that $g \in L^2_\rho$ that $\|h_n - g\|_{\rho} \to 0$. We conclude that $g = h$ and the lemma is established.

LEMMA 2. *Assume* (O1)-(O3). *The* $H^1_{p,\rho}(\Omega)$ *is compactly imbedded in* $L^2_\rho(\Omega)$.

To establish this lemma, assume that there exists a sequence $\{v_n\}_{n=1}^{\infty}$ in $H^1_{p,\rho}$ and a constant K such that $\|v_n\|^2_{p,\rho} \leqq K$ for all n. Then it follows from Lemma 1 and (2.4) that

$$\sum_{k=1}^{\infty} \lambda_k |\hat{v}_n(k)|^2 \leqq K. \tag{2.6}$$

To complete the proof of the lemma, we have to show that there exist $v \in H^1_{p,\rho}$ and a subsequence $\{v_{n_m}\}_{m=1}^{\infty}$ such that

$$\|v_{n_m} - v\|^2_{\rho} \to 0. \tag{2.7}$$

To establish (2.7), using (2.6), we choose a subsequence $\hat{v}_{(n,1)}(1)$ such that $\hat{v}_{(n,1)}(1) \to a_1$ as $n \to \infty$. Next we choose a subsequence $(n, 2)$ of $(n, 1)$ such that $\hat{v}_{(n,2)}(2) \to a_2$. Continuing in this manner, we obtain subsequences (n, k) and finite number a_k such that

$$\lim_{n \to \infty} \hat{v}_{(n,k)}(k) = a_k \quad \forall k, \tag{2.8}$$

where $(n, k+1)$ is a subsequence of (n, k). Now let j be a fixed positive integer. Then it follows from (2.6) that $\sum_{k=1}^{j} \lambda_k |\hat{v}_{(n,j)}(k)|^2 \leqq K$. Consequently, it follows from (2.8) that $\sum_{k=1}^{j} \lambda_k |a_k|^2 \leqq K$. But then letting $j \to \infty$, we see that

$$\sum_{k=1}^{\infty} \lambda_k |a_k|^2 \leqq K. \tag{2.9}$$

It therefore follows from Lemma 1 that

$$\exists v \in H^1_{p,\rho} \quad \text{subject to } \hat{v}(k) = a_k. \tag{2.10}$$

Next we define $v_{n_m} = v_{(m,m)}$, and we observe from (2.4) and (2.10) that

$$\|v_{n_m} - v\|_\rho^2 = \sum_{k=1}^{\infty} |\hat{v}_{(m,m)}(k) - \hat{v}(k)|^2. \tag{2.11}$$

Once again we let j be a fixed positive integer. Then we see from (2.10) and (2.11) that

$$\|v_{n_m} - v\|_\rho^2 \leqq \sum_{k=1}^{j} |\hat{v}_{(m,m)}(k) - a_k|^2 + \lambda_j^{-1} \sum_{k=1}^{\infty} \lambda_k |\hat{v}_{(m,m)}(k) - a_k|^2.$$

Consequently, it follows from (2.6), (2.8), and (2.9) that

$$\limsup_{m\to\infty} \|v_{n_m} - v\|_\rho^2 \leqq 4K/\lambda_j.$$

But by (O2), $\lambda_j \to \infty$. Hence $\limsup_{m\to\infty} \|v_{n_m} - v\|_\rho^2 = 0$. Therefore (2.7) is established, and the proof of the lemma is complete.

Next, we establish the following lemma.

LEMMA 3. *Assume that the conditions in the hypothesis of the theorem hold. Let n be a fixed positive integer, and let S_n be the subspace of $H^1_{p,\rho}$ spanned by $\psi_1, \cdots, \psi_n$. Then there exist $u_n \in S_n$ such that*

$$\mathcal{Q}(u_n, v) = \langle f(\cdot, u_n), v\rangle_\rho + G(v) \quad \forall v \in S_n. \tag{2.12}$$

To prove this lemma, we let $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n$ and set

$$F_k(\alpha) = \mathcal{Q}(\alpha_j\psi_j, \psi_k) - \langle f(\cdot, \alpha_j\psi_j), \psi_k\rangle_\rho - G(\psi_k) \tag{2.13}$$

for $k = 1, \cdots, n$ where we have used the summation convention for $j = 1, \cdots, n$. Then with $F(\alpha) = (F_1(\alpha), \cdots, F_n(\alpha))$, we obtain that

$$F(\alpha)\cdot\alpha = \mathcal{Q}(\alpha_j\psi_j, \alpha_k\psi_k) - \langle f(\cdot, \alpha_j\psi_j), \alpha_k\psi_k\rangle_\rho - G(\alpha_k\psi_k).$$

Consequently, we obtain from (f2) and (O1) that

$$F(\alpha)\cdot\alpha \geqq \mathcal{Q}(\alpha_j\psi_j, \alpha_k\psi_k) - (\lambda_1^* - \varepsilon_0)|\alpha|^2 - \|\tilde{h}\|_\rho|\alpha| - G(\alpha_k\psi_k) \quad \forall \alpha \in \mathbb{R}^n. \tag{2.14}$$

Since $G \in [H^1_{p,\rho}]^*$, there exist K_2 such that

$$|G(v)| \leqq K_2\|v\|_{p,\rho}. \tag{2.15}$$

Also, it follows from Lemma 1 that there exist $K_{3,n}$ such that $\|v\|_{p,\rho} \leqq K_{3,n}\|v\|_\rho$ for all $v \in S_n$. Consequently, we conclude from (2.14) and (2.15) that

$$F(\alpha)\cdot\alpha \geqq \mathcal{Q}(\alpha_j\psi_j, \alpha_k\psi_k) - (\lambda_1^* - \varepsilon_0)|\alpha|^2 - [\|\tilde{h}\|_\rho + K_2K_{3,n}]|\alpha| \quad \forall \alpha \in \mathbb{R}^n. \tag{2.16}$$

From (1.7) we see that there exist $R_0 > 0$ such that $\mathcal{Q}(\alpha_j\psi_j, \alpha_k\psi_k) \geqq |\alpha|^2(\lambda_1^* - \varepsilon_0/2)$ for $|\alpha| \geqq R_0$. Hence we obtain from (2.16) that $|\alpha| \geqq R_0$,

$$F(\alpha)\cdot\alpha \geqq 2^{-1}\varepsilon_0|\alpha|^2 - [\|\tilde{h}\|_\rho + K_2K_{3,n}]|\alpha|. \tag{2.17}$$

From this last inequality, it follows that there exists $R_1 > 0$ such that

$$F(\alpha)\cdot\alpha > 0 \quad \forall |\alpha| \geqq R_1.$$

From (Q1)-(Q3) in conjunction with (1.6) and from (f1) and (f3), it is an easy matter to see from (2.13) that $F_k(\alpha) \in C^0(\mathbb{R}^n)$ for $k = 1, \cdots, n$. Hence it follows from [12, p. 18] that there exist $\alpha^* \in \mathbb{R}^n$ with $|\alpha| < R_1$ such that $F_k(\alpha^*) = 0$ for $k = 1, \cdots, n$. Setting $u_n = \alpha_1^*\psi_1 + \cdots + \alpha_n^*\psi_n$, we consequently see from (2.13) that

$$\mathcal{Q}(u_n, \psi_k) = \langle f(\cdot, u_n), \psi_k\rangle_\rho - G(\psi_k) \quad \text{for } k = 1, \cdots, n. \tag{2.18}$$

But if $v \in S_n$, $v = \hat{v}(1)\psi_1 + \cdots + \hat{v}(n)\psi_n$ and the conclusion of the lemma follows immediately from (2.18).

3. Proof of Theorem 1. To prove Theorem 1 we invoke Lemma 3 and obtain a sequence $\{u_n\}_{n=1}^{\infty}$ where

$$u_n = \alpha_1^n \psi_1 + \cdots + \alpha_n^n \psi_n \tag{3.1}$$

with the property that (2.12) holds.We claim there exists a constant K_4 such that

$$\|u_n\|_{p,\rho} \leqq K_4 \quad \forall n. \tag{3.2}$$

Suppose to the contrary that (3.2) does not hold. For ease of notation and without loss in generality we assume

$$\lim_{n\to\infty} \|u_n\|_{p,\rho} = \infty. \tag{3.3}$$

We will show this assumption gives a contradiction.

Now one of the following two cases present themselves:

$$\exists K_5 \quad \text{subject to } \|u_n\|_\rho \leqq K_5 \quad \forall n, \tag{3.4}$$

$$\exists \{u_{n_j}\}_{j=1}^{\infty} \quad \text{subject to } \lim_{j\to\infty} \|u_{n_j}\|_\rho = \infty. \tag{3.5}$$

Suppose (3.4) holds. Then taking $v = u_n$ in (2.12) and using (f2) and (1.8), we obtain

$$\mathcal{Q}(u_n, u_n) \leqq (\lambda_1^* - \varepsilon_0)\|u_n\|_\rho^2 + \|\tilde{h}\|_\rho \|u_n\|_\rho + K_1 \|u_n\|_{p,\rho}. \tag{3.6}$$

From (Q3), (Q4), and (3.4), we infer from this last inequality that

$$c_3\|u_n\|_{p,\rho}^2 - K_1\|u_n\|_{p,\rho} \leqq (\lambda_1^* - \varepsilon_0)K_5^2 + \|\tilde{h}\|_\rho K_5 + c_3 K_5^2 + \int_\Omega Zp + c_2 K_5^2 \quad \forall n.$$

Since $Z \in L_p^1(\Omega)$, this last inequality is incompatible with (3.3). Hence (3.2) holds for the first case (3.4).

We now suppose that (3.3) and the second case (3.5) hold. Then it follows from (3.6), (1.7), and (1.8) that there exist j_0 such that for $j \geqq j_0$,

$$(\lambda_1^* - \varepsilon_0 2^{-1})\|u_{n_j}\|_\rho^2 \leqq (\lambda_1^* - \varepsilon_0)\|u_{n_j}\|_\rho^2 + \|\tilde{h}\|_\rho \|u_{n_j}\|_\rho + K_1\|u_{n_j}\|_{p,\rho}.$$

Hence we have, for $j \geqq j_0$,

$$2^{-1}\varepsilon_0\|u_{n_j}\|_\rho^2 - \|\tilde{h}\|_\rho \|u_{n_j}\|_\rho \leqq K_1\|u_{n_j}\|_{p,\rho}.$$

We conclude from this inequality and (3.5) that there exist K_6 and j_1 such that

$$\|u_{n_j}\|_\rho^2 \leqq K_6\|u_{n_j}\|_{p,\rho} \quad \text{for } j \geqq j_1. \tag{3.7}$$

From this last inequality, we obtain from (3.6) that there exist K_7 and $j_2 \geqq j_1$ such that

$$\mathcal{Q}(u_{n_j}, u_{n_j}) \leqq K_7\|u_{n_j}\|_{p,\rho} \quad \text{for } j \geqq j_2. \tag{3.8}$$

From (Q3), (Q4), (3.7), and (3.8), we in turn obtain that

$$c_3\|u_{n_j}\|_{p,\rho}^2 \leqq (K_7 + c_2 K_6)\|u_{n_j}\|_{p,\rho} + \int_\Omega Zp$$

for $j \geqq j_2$. Since $Z \in L_p^1(\Omega)$ and c_2 and c_3 are positive constants, this last inequality is incompatible with (3.3), and we have arrived at a contradiction for the second case (3.5) also. Hence (3.2) is fully established.

Since $H^1_{p,\rho}(\Omega)$ is a separable Hilbert space, we see from Lemma 2 that there exists a subsequence (which for ease of notation we take to be the full sequence) and a function $u^* \in H^1_{p,\rho}$ with the following properties:

$$\lim_{n\to\infty} \|u_n - u^*\|_\rho = 0, \tag{3.9}$$

$$\lim_{n\to\infty} u_n(x) = u^*(x) \quad \text{for a.e. } x \text{ in } \Omega, \tag{3.10}$$

$$\lim_{n\to\infty} \langle D_i u_n, v\rangle_p = \langle D_i u^*, v\rangle_p, \quad \forall v \in L^2_p \text{ and } i = 1, \cdots, N, \tag{3.11}$$

$$\lim_{n\to\infty} G(u_n) = G(u^*). \tag{3.12}$$

Next, we propose to show there exists a subsequence $\{u_{n_j}\}_{j=1}^\infty$ subject to

$$\lim_{j\to\infty} Du_{n_j}(x) = Du^*(x) \quad \text{for a.e. } x \text{ in } \Omega. \tag{3.13}$$

Once (3.13) is established it will be an easy matter to complete the proof of the theorem.

To establish (3.13) it is sufficient to establish the following two facts. (1) There exists a subsequence $\{u_{n_j}\}_{j=1}^\infty$ subject to

$$\lim_{j\to\infty} [A_i(x, u_{n_j}, Du_{n_j}) - A_i(x, u_{n_j}, Du^*)][D_i u_{n_j}(x) - D_i u^*(x)] = 0 \tag{3.14}$$

for a.e. x in Ω.

(2) With $\{u_{n_j}\}_{j=1}^\infty$ designating the same subsequence as in (3.14),

$$\{|Du_{n_j}(x)|\}_{j=1}^\infty \text{ is pointwise bounded for almost every } x \text{ in } \Omega. \tag{3.15}$$

To see that (3.14) and (3.15) together imply (3.13), let Ω_1 be the subset of Ω for which (Q5), (3.10), (3.14), and (3.15) hold simultaneously. Consequently,

$$\text{meas}\,(\Omega - \Omega_1) = 0. \tag{3.16}$$

Suppose there exist $x_0 \in \Omega_1$ for which (3.13) does not hold. Hence by (3.15) there exist a further sequence $\{Du_{n_{j_k}}(x_0)\}_{k=1}^\infty$ and a $\xi^* \in \mathbb{R}^N$ with

$$Du^*(x_0) \neq \xi^* \tag{3.17}$$

such that $\lim_{k\to\infty} Du_{n_{j_k}}(x_0) = \xi^*$. Therefore from (3.10)

$$\begin{aligned}\lim_{k\to\infty} &[A_i(x_0, u_{n_{j_k}}, Du_{n_{j_k}}) - A_i(x_0, u_{n_{j_k}}, Du^*)][D_i u_{n_{j_k}}(x_0) - D_i u^*(x_0)] \\ &= [A_i(x_0, u^*(x_0), \xi^*) - A_i(x_0, u^*(x_0), Du^*(x_0)][\xi_i^* - D_i u^*(x_0)].\end{aligned} \tag{3.18}$$

From (Q5) and (3.17) we see that the right-hand side of (3.18) is strictly positive. Hence the limit on the left-hand side of (3.18) is strictly positive. But $x_0 \in \Omega_1$ and by the definition of Ω_1 and (3.14), this limit equals zero. We have arrived at a contradiction. Hence the equality in (3.13) holds for every point in Ω_1 and therefore by (3.16), for almost every x in Ω, and statement (3.13) is fully established.

It remains to show that (3.14) and (3.15) hold. To establish (3.14) we observe from (Q2), (3.9), (3.10), and Egoroff's theorem [13, p. 75] that

$$\lim_{n\to\infty} \int_\Omega |A_i(x, u_n, Du^*) - A_i(x, u^*, Du^*)|^2 p(x) = 0 \tag{3.19}$$

for $i = 1, \cdots, N$. Also, we have from (3.11) that

$$\lim_{n\to\infty} \int_\Omega A_i(x, u^*, Du^*)[D_i u_n - D_i u^*] p(x) = 0. \tag{3.20}$$

From (3.19) and (3.20) in conjunction with (3.2) and Schwarz's inequality, it is not difficult to see that

$$\lim_{n\to\infty}\int_\Omega [A_i(x, u_n, Du_n) - A_i(x, u_n, Du^*)][D_iu_n - D_iu^*]p(x) = 0, \tag{3.21}$$

provided we show

$$\lim_{n\to\infty}\int_\Omega A_i(x, u_n, Du_n)(D_iu_n - D_iu^*)p(x) = 0. \tag{3.22}$$

By (Q5), the integrand in (3.21) is nonnegative for almost every $x\in\Omega$. Hence the integrand converges in L_p^1-norm to zero and (3.14) follows from [13, p. 70]. Therefore, to complete the proof of (3.14), it remains to establish (3.22). From (Q3) and (3.9) we see that (3.22) will follow once we show

$$\lim_{n\to\infty} \mathcal{Q}(u_n, u_n - u^*) = 0. \tag{3.23}$$

Now $u^*\in H_{p,\rho}^1$ and

$$P_nu^* \equiv \sum_{j=1}^{n} \hat{u}^*(j)\psi_j \quad \text{is in } S_n \tag{3.24}$$

where S_n is defined in Lemma 3. Hence it follows from Lemma 1 that

$$\lim_{n\to\infty} \|P_nu^* - u^*\|_{p,\rho} = 0. \tag{3.25}$$

From (Q2), (Q3), (1.6), (3.2), and (3.25), it also follows that $\lim_{n\to\infty} \mathcal{Q}(u_n, P_nu^* - u^*) = 0$. Consequently, we see that (3.23) will follow once we show that

$$\lim_{n\to\infty} \mathcal{Q}(u_n, u_n - P_nu^*) = 0. \tag{3.26}$$

To establish (3.26), we observe from (3.24) that (2.12) holds for u_n with v replaced by $u_n - P_nu^*$. It is easy to see from (1.8), (3.12), and (3.25) that $\lim_{n\to\infty} G(u_n - P_nu^*) = 0$. Consequently, we obtain from (2.12) that (3.26) will follow once we show

$$\lim_{n\to\infty}\int_\Omega f(x, u_n)[u_n - P_nu^*]\rho = 0. \tag{3.27}$$

But from (3.9) and (f3), we see that $\{\|f(\cdot, u_n)\|_\rho^2\}_{n=1}^\infty$ is a uniformly bounded sequence. Also, we see from (3.9) and (3.25) that $\lim_{n\to\infty} \|u_n - P_nu^*\|_\rho = 0$. We conclude from these last two facts and Schwarz's inequality that (3.25) does indeed hold. Hence (3.21) holds and the proof of (3.14) is complete.

To establish (3.15), we let Ω_2 be the set where simultaneously $u^*(x)$, $|Du^*(x)|$, $Z(x)$, $h(x)$, $u_{n_j}(x)$, $A_i(x, u_{n_j}(x), Du_{n_j}(x))$, and $A_i(x, u_{n_j}(x), u^*(x))$ are finite-valued for $i = 1, \cdots, N$ and $j = 1, 2, \cdots$, where (Q2) and (Q4) hold and also where the limits in (3.10) and (3.14) exist. Then

$$\text{meas}\,(\Omega - \Omega_2) = 0. \tag{3.28}$$

Suppose there exist $x_0\in\Omega_2$ such that

$$\lim_{k\to\infty} |Du_{n_{j_k}}(x_0)| = \infty. \tag{3.29}$$

Now,

$$c_3|Du_{n_{j_k}}(x_0)|^2 \leqq A_i(x_0, u_{n_{j_k}}, Du_{n_{j_k}})D_iu_{n_{j_k}}(x_0) + Z(x_0), \tag{3.30}$$

and also

$$
\begin{aligned}
&A_i(x_0, u_{n_{j_k}}, Du_{n_{j_k}})D_iu_{n_{j_k}}(x_0)\\
&\quad = A_i(x_0, u_{n_{j_k}}, Du_{n_{j_k}})D_iu^*(x_0)\\
&\quad\quad + A_i(x_0, u_{n_{j_k}}, Du^*)[D_iu_{n_{j_k}}(x_0) - D_iu^*(x_0)]\\
&\quad\quad + [A_i(x_0, u_{n_{j_k}}, Du_{n_{j_k}}) - A_i(x_0, u_{n_{j_k}}, Du^*)][D_iu_{n_{j_k}}(x_0) - D_iu^*(x_0)].
\end{aligned}
\tag{3.31}
$$

We divide both sides of (3.30) by $|Du_{n_{j_k}}(x_0)|^{3/2}$ and conclude from (Q2), (3.31), (3.14), and the definition of Ω_2 that $\lim_{k\to\infty} |Du_{n_{j_k}}(x_0)|^{1/2} = 0$. This is a contradiction to (3.29). Hence $\{|Du_{n_j}(x)|\}_{j=1}^{\infty}$ is a pointwise bounded sequence for every x in Ω_2. But then (3.15) follows immediately from this fact and (3.28). As we have already shown, (3.14) and (3.15) imply (3.13). Hence (3.13) is completely established.

Next, we let $v_J \in S_J$ where J is a fixed, but arbitrary positive integer. Then it follows from (Q3), (3.9), (3.10), and (3.13) that

$$
\lim_{j\to\infty} \langle B_0(\cdot, u_{n_j}, Du_{n_j})u_{n_j}, v_J\rangle_\rho = \langle B_0(\cdot, u^*, Du^*)u^*, v_J\rangle_\rho. \tag{3.32}
$$

We see next from (f3) and (3.9) that $\{\|f(\cdot, u_{n_j}) - f(\cdot, u^*)\|_\rho\}_{j=1}^{\infty}$ is a uniformly bounded sequence and from (3.10), Egoroff's theorem, and the fact $v_J \in L^2_\rho$ that

$$
\lim_{j\to\infty} \langle f(\cdot, u_{n_j}), v_J\rangle_\rho = \langle f(\cdot, u^*), v_J\rangle_\rho. \tag{3.33}
$$

Likewise, we see from (Q2) and (3.2) that $\{\|A_i(\cdot, u_{n_j}, Du_{n_j}) - A_i(\cdot, u^*, Du^*)\|_p\}_{j=1}^{\infty}$ is a uniformly bounded sequence for $i = 1, \cdots, N$. Hence, it follows from (3.10), (3.13), Egoroff's theorem, and the fact that $|Dv_J| \in L^2_p$ that

$$
\lim_{j\to\infty} \langle A_i(\cdot, u_{n_j}, Du_{n_j}), D_iv_J\rangle_p = \langle A_i(\cdot, u^*, Du^*), D_iv_J\rangle_p. \tag{3.34}
$$

Since (2.12) holds for u_{n_j} and $v_J \in S_J$ when $n_j \geqq J$, we conclude from (1.6), (3.32), (3.33), and (3.34) that

$$
\mathcal{Q}(u^*, v_J) = \langle f(\cdot, u^*), v_J\rangle_\rho + G(v_J). \tag{3.35}
$$

Next given $v \in H^1_{p,\rho}$, we define $P_Jv \in S_J$ as in (3.24) and observe that $\lim_{J\to\infty} \|P_Jv - v\|_{p,\rho} = 0$ as in (3.25). Hence it follows from (Q2) and (Q3) that $\mathcal{Q}(u^*, P_Jv) \to \mathcal{Q}(u^*, v)$, $\langle f(\cdot, u^*), P_Jv\rangle_\rho \to \langle f(\cdot, u^*), v\rangle_\rho$, and $G(P_Jv) \to G(v)$ as $J \to \infty$. We conclude from (3.35) that

$$
\mathcal{Q}(u^*, v) = \langle f(\cdot, u^*), v\rangle_\rho + G(v),
$$

and the proof of Theorem 1 is complete.

4. Resonance and singular quasilinearity. Let

(L1) $a_{ij}(x)$ and $b_0(x)$ be in $L^\infty(\Omega)$, $i, j = 1, \cdots, N$ with $a_{ij}(x) = a_{ji}(x)$ almost everywhere in Ω; and

(L2) $a_{ij}(x)\xi_i\xi_j \geqq c_4|\xi|^2$ for almost every $x \in \Omega$ where $c_4 > 0$

and where the summation convention is used. Set

$$
Lu = -D_i[p(x)a_{ij}(x)D_ju] + b_0(x)\rho(x)u \tag{4.0}
$$

and

$$
\mathcal{L}(u, v) = \int_\Omega pa_{ij}D_iuD_jv + b_0uv\rho \tag{4.1}
$$

for all $u, v \in H^1_{p,\rho}$ where once again we have used the summation convention for $i, j = 1, \cdots, N$.

From Lemma 2 we know that $H^1_{p,\rho}$ is compactly imbedded in L^2_ρ. Hence, using the reasoning on [7, p. 213], it follows that there exists

$$\sigma_1 \leqq \sigma_2 \leqq \cdots \leqq \sigma_n \leqq \cdots \to +\infty \tag{4.2}$$

and a sequence $\{\phi_n\}_{n=1}^\infty$ with $\phi_n \in H^1_{p,\rho}$ enjoying the following two properties:

$$\mathscr{L}(\phi_n, v) = \sigma_n \langle \phi_n, v \rangle_\rho \quad \forall n, \tag{4.3}$$

$$\{\phi_n\}_{n=1}^\infty \text{ is a CONS in } L^2_\rho. \tag{4.4}$$

Furthermore, applying the reasoning of [7, p. 214] to compact subsets of Ω, we see that $\sigma_1 < \sigma_2$ and we can take $\phi_1 > 0$ almost everywhere in Ω. We record this observation as

$$\sigma_1 < \sigma_2 \quad \text{and} \quad \phi_1(x) > 0 \quad \text{for a.e. } x \in \Omega. \tag{4.5}$$

Next with Q as in § 1, we shall say Q is $*$-related to L if the following two facts hold:

$$\lambda_1^* = \sigma_1, \tag{4.6}$$

$$\liminf_{\|u\|_{p,\rho} \to \infty} [\mathscr{Q}(u, u) - \mathscr{L}(u, u)] / \|u\|_{p,\rho} \geqq 0. \tag{4.7}$$

In particular, we see that L is $*$-related to itself. Also, it is easy to see that $Qu = -D_i p\{1 + [1 + |Du|^2]^{-1/2}\} D_i u + \rho u$ meets (Q1)-(Q5) and is $*$-related to $Lu = -D_i p D_i u + \rho u$. (See [9, p. 96]. In the next section the details will be supplied and other examples with be given.) We intend to establish a result similar to Theorem 1, except that in (f2) we shall take $\varepsilon_0 = 0$. As we have shown at the end of § 1, a result of this nature is in general false and therefore to obtain such a result, another condition, usually referred to as a Landesman-Lazer condition, is required. (See [4, p. 284] and [10].) To this end, we define the following:

(f4) There exist $\tilde{h}_1 \geqq 0$ with $\tilde{h}_1 \in L^2_\rho$ subject to $tf(x, t) \leqq \lambda_1^* t^2 + \tilde{h}_1(x)|t|$ for all $t \in \mathbb{R}$ and almost every $x \in \Omega$.

The theorem we intend to prove is the following.

THEOREM 2. *Let the hypotheses of Theorem* 1 *hold with* (f4) *replacing* (f2). *Suppose also that Q is $*$-related to L where L satisfies* (L1) *and* (L2). *Set* $g_+(x) = \limsup_{t \to \infty} [f(x, t) - \lambda_1^* t]$ *and* $g_-(x) = \liminf_{t \to -\infty} [f(x, t) - \lambda_1^* t]$ *and suppose furthermore that*

$$\int_\Omega g_+ \phi_1 \rho < -G(\phi_1) < \int_\Omega g_- \phi_1 \rho. \tag{4.8}$$

Then there exist $u^* \in H^1_{p,\rho}$ *such that*

$$\mathscr{Q}(u^*, u) = \langle f(\cdot, u^*), u \rangle_\rho + G(u) \quad \forall u \in H^1_{p,\rho}. \tag{4.9}$$

From the start, by adding a constant positive multiple of $\rho(x)u$ to Qu in (1.0) and Lu in (4.0), and this same constant multiple to u to $f(x, u)$, we see from (1.7), (4.6), and (4.9) that, with no loss in generality, to prove the theorem we can assume that

$$\lambda_1^* = \sigma_1 > 0 \quad \text{and} \quad b_0(x) > 0 \quad \text{a.e. in } \Omega. \tag{4.10}$$

Also, we set

$$g(x, t) = f(x, t) - \lambda_1^* t \quad \text{for } (x, t) \in \Omega \times \mathbb{R} \tag{4.11}$$

and observe from (f4) and Theorem 1 that there exist $\{u_n\}_{n=1}^{\infty}$ with $u_n \in H^1_{p,\rho}$ such that

$$\mathscr{Q}(u_n, u) = [\lambda_1^* - n^{-1}]\langle u_n, u\rangle_\rho + \langle g(\cdot, u_n), u\rangle_\rho + G(u) \quad \forall u \in H^1_{p,\rho}. \tag{4.12}$$

Next we set

$$\check{w}(n) = \langle \phi_n, w\rangle_\rho \quad \text{for } w \in L^2_\rho \tag{4.13}$$

and also observe from (4.2)-(4.4) and (4.10) that

$$u \in H^1_{p,\rho} \quad \text{and} \quad \mathscr{L}(u, \phi_n) = 0 \quad \forall n \Rightarrow u = 0. \tag{4.14}$$

Furthermore, from (L2), (4.1), and (4.10), we see that if $u \in H^1_{p,\rho}$ and $u \neq 0$, then $\mathscr{L}(u, u) > 0$. Consequently, $\mathscr{L}(\cdot, \cdot)$ can be viewed as an inner product on $H^1_{p,\rho}$ and it follows from (4.3), (4.4), (4.13), and (4.14) that

$$\begin{aligned} &\text{(a)} \ \langle w, w\rangle_\rho = \sum_{n=1}^{\infty} |\check{w}(n)|^2 \qquad \text{for } w \in L^2_\rho, \\ &\text{(b)} \ \mathscr{L}(u, u) = \sum_{n=1}^{\infty} \sigma_n |\check{u}(n)|^2 \text{ for } u \in H^1_{p,\rho}. \end{aligned} \tag{4.15}$$

Continuing with the proof, the first fact we claim is that with $\{u_n\}_{n=1}^{\infty}$ the sequence given in (4.12),

$$\text{There exist } K_8 \text{ such that } \|u_n\|_{p,\rho} \leqq K_8 \text{ for all } n. \tag{4.16}$$

Suppose that (4.16) does not hold. Then there exists a subsequence (which for ease of notation we take to be the full sequence) such that

$$\lim_{n\to\infty} \|u_n\|_{p,\rho} = \infty. \tag{4.17}$$

We shall arrive at a contradiction by showing that this fact implies that one of the inequalities in (4.8) does not hold.

To accomplish this we replace u by u_n in (4.12) and obtain from (f4) and (1.8) that

$$\mathscr{Q}(u_n, u_n) \leqq (\lambda_1^* - n^{-1})\|u_n\|_\rho^2 + \|\tilde{h}_1\|_\rho \|u_n\|_\rho + K_1 \|u_n\|_{p,\rho}. \tag{4.18}$$

Consequently, we obtain from (L1), (L2), (4.7), (4.10), (4.17), and this last inequality that there is a positive constant K_9 such that

$$2^{-1} c_4 \|u_n\|_{p,\rho}^2 - K_1 \|u_n\|_{p,\rho} \leqq K_9 \|u_n\|_\rho^2 + \|\tilde{h}_1\|_\rho \|u_n\|_\rho$$

for n sufficiently large. Since c_4 is a positive constant, it follows from (4.17) that there is a positive constant K_{10} and a positive integer n_1 such that

$$\|u_n\|_{p,\rho} \leqq K_{10} \|u_n\|_\rho \quad \text{for } n \geqq n_1. \tag{4.19}$$

In particular, from (4.17) and (4.19) we see that $\lim_{n\to\infty} \|u_n\|_\rho = \infty$. Next we set

$$v_n = u_n / \|u_n\|_\rho \tag{4.20}$$

and

$$v_{1n} = \check{v}_n(1)\phi_1 \quad \text{and} \quad v_{2n} = v_n - v_{1n}. \tag{4.21}$$

From (4.7), (4.10), (4.15), (4.18), and (4.19), we see that there exist n_2 such that

$$\sum_{k=2}^{\infty} (\sigma_k - \sigma_1)|\check{v}_n(k)|^2 \leqq -n^{-1}\|v_n\|_\rho^2 + [\|\tilde{h}_1\|_\rho + (K_1 + 1)K_{10}]\|v_n\|_\rho \|u_n\|_\rho^{-1}$$

for $n \geqq \max(n_1, n_2)$. From (4.20), we see that $\|v_n\|_\rho = 1$, and from (4.5) and (4.15)(a) that the left-hand side of this last inequality majorizes $(\sigma_2 - \sigma_1)\|v_{n2}\|_\rho^2$. We, consequently, conclude from (4.17) and (4.19) that

$$\lim_{n\to\infty} \|v_{n2}\|_\rho^2 = 0. \tag{4.22}$$

But then it follows from (4.5) and (4.21) that $\lim_{n\to\infty} \|v_{n1}\|_\rho = 1$, and hence

$$\lim_{n\to\infty} |\check{v}_n(1)|^2 = 1. \tag{4.23}$$

From (4.19) and (4.20) we see that $\|v_n\|_{p,\rho} \leqq K_{10}$ for $n \geqq n_1$. Therefore we have from Lemma 2 that there exist $v \in H^1_{p,\rho}$ and a subsequence (which for ease of notation we take to be the full sequence) such that

$$\begin{aligned}
&\text{(a)} \quad \lim_{n\to\infty} \|v_n - v\|_\rho = 0, \\
&\text{(b)} \quad \lim_{n\to\infty} v_n(x) = v(x) \quad \text{for a.e. } x \text{ in } \Omega, \\
&\text{(c)} \quad \lim_{n\to\infty} \int_\Omega D_i v_n w p = \int_\Omega D_i v w p \quad \forall w \in L^2_p, \\
&\text{(d)} \quad \lim_{n\to\infty} G(v_n) = G(v).
\end{aligned} \tag{4.24}$$

From (4.13), (4.21), (4.22), and (4.24)(a), we see that $\check{v}(k) = 0$ for $k \geqq 2$. Consequently, $v = \check{v}(1)\phi_1$ and since $\lim_{n\to\infty} \check{v}_n(1) = \check{v}(1)$, we conclude from (4.23) that $\check{v}(1) = 1$ or -1. We shall assume

$$\check{v}(1) = 1 \quad \text{and} \quad v = \phi_1 \tag{4.25}$$

and arrive at a contradiction. Similar reasoning will lead to a contradiction for the assumption $\check{v}(1) = -1$.

To arrive at this contradiction, we replace u by u_n in (4.12) and use (4.6), (4.7), and (4.15) to obtain the following. Given $\varepsilon > 0$, there exist n_3 such that for $n \geqq n_3$

$$\sum_{k=2}^{\infty} (\sigma_k - \sigma_1)|\check{u}_n(k)|^2 \leqq \langle g(\cdot, u_n), u_n\rangle_\rho + G(u_n) + \varepsilon\|u_n\|_{p,\rho}.$$

Consequently, from (4.19) and (4.20) we see that

$$\int_\Omega [\tilde{h}_1(x)|v_n| - g(x, u_n)v_n|\rho - \int_\Omega \tilde{h}_1(x)|v_n|\rho \leqq G(v_n) + \varepsilon K_{10} \tag{4.26}$$

for $n \geqq n_3$. (f4) in conjunction with (4.11) tells us that the integrand in the first integral on the left in this last inequality is nonnegative. Since $\tilde{h}_1 \in L^2_\rho$, we see from (4.24)(a) that the limit of the second integral on the left is $\langle \tilde{h}_1, |v|\rangle_\rho$. We consequently conclude from (4.24)(d), (4.26), and Fatou's lemma [13, p. 24] that

$$-\int_\Omega \limsup_{n\to\infty} [g(x, u_n)v_n]\rho \leqq G(v) + \varepsilon K_{10}.$$

But $\varepsilon > 0$ is arbitrary. Consequently, we conclude from this last inequality that

$$-\int_\Omega \limsup_{n\to\infty} [g(x, u_n)v_n]\rho \leqq G(v). \tag{4.27}$$

From (4.24)(b) and (4.25), we see that $\lim_{n\to\infty} v_n(x) = \phi_1(x)$ almost everywhere in Ω. Since $u_n = \|u_n\|_\rho v_n$, we conclude from (4.5), (4.17), and (4.19) that $\lim_{n\to\infty} u_n(x) = \infty$ for almost every $x \in \Omega$. Hence we see from (4.11), (4.27), and the definition of $g_+(x)$ that

$$\int_\Omega g_+(x)\phi_1\rho \geqq -G(\phi_1).$$

But this is a direct contradiction of the first inequality in (4.8). Hence, (4.16) does indeed hold.

From (4.16), we immediately obtain from Lemma 2 that there exist $u^* \in H^1_{p,\rho}$ with properties (3.9)-(3.12). We would like to show that (3.13) holds. Equation (3.13) follows from (3.14) and (3.15) exactly in the same manner as before. Hence, it remains to establish (3.14) and (3.15). To establish (3.14), we see that (3.19)-(3.21) follow once we show that (3.22) holds. Since (3.21) implies (3.14) via [13, p. 70], to establish (3.14) it suffices to establish (3.22). We now do this. In particular, we have to show

$$\lim_{n\to\infty} \int_\Omega A_i(x, u_n, Du_n)(D_i u_n - D_i u^*)p(x) = 0 \tag{4.28}$$

under the assumptions that (4.12) and (4.16) hold and also that (3.9)-(3.12) hold. From (3.9) and (Q3), we see that (4.28) will follow once we show

$$\lim_{n\to\infty} \mathcal{Q}(u_n, u_n - u^*) = 0. \tag{4.29}$$

Replacing u by $u_n - u^*$ in (4.12), we see from (3.9) and (3.12) that (4.29) will follow once we show

$$\lim_{n\to\infty} \langle g(\cdot, u_n), u_n - u^*\rangle_\rho = 0. \tag{4.30}$$

Now it follows from (4.11) and (f3) that

$$|g(x, u_n)| \leqq \tilde{h}_2(x) + (K + \lambda_1^*)|u_n|. \tag{4.31}$$

Hence it follows from (3.9) that there exist K_{11} such that $\int_\Omega |g(x, u_n)|^2\rho \leqq K_{11}$ for all n. Using (3.9) once again, in conjunction with this last fact and Schwarz's inequality, gives (4.30). Hence (3.14) is established. Equation (3.15) follows exactly as before using (3.30). Since (3.14) and (3.15) hold, as we have already shown, (3.13) holds.

To complete the proof of the theorem, it remains to show that (3.9)-(3.13) along with (4.12) and (4.16) imply (4.9). Let u be given in $H^1_{p,\rho}$. Then it follows (Q3), (3.9), (3.10), and (3.13) that

$$\lim_{j\to\infty} \langle B_0(\cdot, u_{n_j}, Du_{n_j})u_{n_j}, u\rangle_\rho = \langle B_0(\cdot, u^*, Du^*)u^*, u\rangle_\rho. \tag{4.32}$$

Next, we see from (4.31) and (3.9) that

$$\{\|g(\cdot, u_{n_j}) - g(\cdot, u^*)\|_\rho\}_{n=1}^\infty$$

is a uniformly bounded sequence. Hence, it follows from (3.10) and Egoroff's theorem that

$$\lim_{j\to\infty} \langle g(\cdot, u_{n_j}), u\rangle_\rho = \langle g(\cdot, u^*), u\rangle_\rho. \tag{4.33}$$

Likewise, we see from [Q2] and (4.16) that

$$\{\|A_i(\cdot, u_{n_j}, Du_{n_j}) - A_i(\cdot, u^*, Du^*)\|_p\}_{j=1}^\infty$$

is a uniformly bounded sequence for $i = 1, \cdots, N$. Hence, it follows from (3.10), (3.13), Egoroff's theorem, and the fact $D_i u \in L_p^2$ that

$$\lim_{j\to\infty} \langle A_i(\cdot, u_{n_j}, Du_{n_j}), D_i u\rangle_p = \langle A_i(\cdot, u^*, Du^*), D_i u\rangle_p. \tag{4.34}$$

From (4.12) and (3.9) in conjunction with (4.32)-(4.34), we conclude that

$$\langle \mathcal{Q}(u^*, u)\rangle = \lambda_1^* \langle u^*, u\rangle_p + \langle g(\cdot, u^*), u\rangle_\rho + G(u).$$

From (4.11) we see this is the same as (4.9). The proof of Theorem 2 is therefore complete.

5. A corollary with some examples. Using the techniques presented in this paper, we can also obtain the singular quasilinear analogue of Theorem 3 of [6]. The details are left to the interested reader.

We close with the following corollary and some examples illustrating said corollary. We shall need the following asumption in the sequel. $g(t)$ is a bounded continuous function with

$$g_+ = \limsup_{t\to\infty} g(t), \quad g_- = \liminf_{t\to-\infty} g(t), \quad \text{and } g_+ < g(t) < g_- \quad \text{for } -\infty < t < t\infty. \tag{5.1}$$

COROLLARY. *Let $\Omega \subset \mathbb{R}^N$, $N \geqq 1$, be an open connected set and let p and ρ satisfy* (1.2) *and* (1.3). *Assume that Q given by* (1.0) *satisfies* (Q1)-(Q5), *that $O_{p,\rho}(\Omega)$ holds, and that $G \in [H_{p,\rho}^1(\Omega)]^*$. Assume also that L given by* (4.0) *satisfies* (L1) *and* (L2), *that Q is $*$-related to L, and that*

$$\mathcal{Q}(u, \phi_1) = \lambda_1^* \langle u, \phi_1\rangle_\rho \quad \forall u \in H_{p,\rho}^1(\Omega) \tag{5.2}$$

where ϕ_1 is given in (4.5). *Assume furthermore that g is a bounded continuous function satisfying* (5.1). *Then a necessary and sufficient condition that there exist $u^* \in H_{p,\rho}^1(\Omega)$ with the property that*

$$\mathcal{Q}(u^*, u) = \lambda_1^* \langle u^*, u\rangle_\rho + \langle g(u^*), u\rangle_\rho + G(u) \tag{5.3}$$

for all $u \in H_{p,\rho}^1(\Omega)$ is that

$$g_+ \int_\Omega \phi_1 \rho < -G(\phi_1) < g_- \int_\Omega \phi_1 \rho. \tag{5.4}$$

It is clear that $f(x, t) = g(t) + \lambda_1^* t$ satisfies (f1), (f3), and (f4). Hence the sufficiency condition of the above corollary follows immediately from Theorem 2. To prove the necessary condition, suppose there exist $u^* \in H_{p,\rho}^1$ such that (5.3) holds. Take $u = \phi_1$ in (5.3). Then it follows from (5.2) that

$$-G(\phi_1) = \int_\Omega g(u^*)\phi_1 \rho. \tag{5.5}$$

But (5.1) implies that

$$g_+ < g(u^*(x)) < g_- \quad \text{for a.e. } x \text{ in } \Omega,$$

and (5.4) follows from (1.2), (4.5), and (5.5). Hence the necessary condition of the above corollary is established.

We now give some examples for which the corollary holds. In particular, in the following examples we shall have to show that Q satisfies (Q1)-(Q5), that Q is $*$-related to L, and that (5.2) holds.

For our first example, we consider the case where $\Gamma_1 = \partial\Omega$ in (1.2) and take

$$\begin{aligned} Lu &= -D_i[pD_iu] + \rho u, \\ Qu &= -D_i p\{1+[1+|Du|^2]^{-1/2}\}D_i u + \rho u. \end{aligned} \tag{5.6}$$

We see that $A_i(x, t, \xi) = \{1+[1+|\xi|^2]^{-1/2}\}\xi_i$ and $B_0 = 1$. It is clear from [9, p. 96] that Q meets (Q1)-(Q5) and that

$$\mathcal{Q}(u, v) - \mathcal{L}(u, v) = \int_\Omega p[1+|Du|^2]^{-1/2} D_i u D_i v. \tag{5.6'}$$

Since $\Gamma_1 = \partial\Omega$, constants are in $C^1_{p,\rho}(\Omega)$. It therefore follows from (4.1) and (4.3) that $\phi_1 = \psi_1 = c_5$ where c_5 is a positive constant and $\lambda_1 = \sigma_1 = 1$. Also, we have from (5.6') that $\mathcal{Q}(u, u) - \mathcal{L}(u, u) \geqq 0$. Consequently, (4.7) follows immediately, and also we see from (1.7), (4.15), and this last inequality that $\lambda_1^* \geqq \sigma_1$. But from (5.6'), we have that $\mathcal{Q}(n\phi_1, n\phi_1) = n^2\mathcal{L}(\phi_1, \phi_1) = n^2\sigma_1$. Therefore, $\lim_{n\to\infty} \mathcal{Q}(n\phi_1, n\phi_1)/n^2\langle\phi_1, \phi_1\rangle_\rho = \sigma_1$; hence $\lambda_1^* \leqq \sigma_1$. Thus $\lambda_1^* = \sigma_1$, and (4.6) is established. Since we have already established (4.7), we have that Q is $*$-related to L. From (5.6') we also see that $\mathcal{Q}(u, \phi_1) = \mathcal{L}(u, \phi_1) = \sigma_1\langle u, \phi_1\rangle_\rho$. Consequently, (5.2) holds and our example is fully established.

We observe that $\Gamma_1 = \partial\Omega$ in the cases (1.10), (1.15), and (1.16) discussed in § 1. Hence the above example covers these cases.

For a more sophisticated example, we consider the case where $F(t)$ is a continuous nondecreasing function on $[0, \infty)$ with

$$F(0) \geqq 0 \quad \text{and} \quad \lim_{t\to\infty} t[1-F(t)] = 0, \tag{5.7}$$

e.g., $F(t) = t/\sqrt{1+t^2}$, $t^2/(1+t^2)$, $[(1+t^2)/(2+t^2)]^{3/2}$, etc. For such an F, the first example we consider is where L has the eigenfunctions that are given in (1.11). For this case $\Omega = \tilde{\Omega} \times \mathbb{R}^1$ where $\tilde{\Omega} \subset \mathbb{R}^2$ and $p(x) = \rho(x) = e^{-x_3^2}$. We then have that

$$Lu = -e^{-x_3^2}D_1^2 u - e^{-x_3^2}D_2^2 u - D_3[e^{-x_3^2}D_3 u] + e^{-x_3^2}u \tag{5.8}$$

and we take

$$Qu = -e^{-x_3^2}[D_1^2 u + D_2^2 u] - 2^{-1}D_3\, e^{-x_3^2}[1+F(|D_3 u|)]D_3 u + e^{-x_3^2}u. \tag{5.9}$$

Now it follows from the discussion in § 1 that $\sigma_1 = \lambda_1 = \eta_1 + 1$ and $\phi_1 = \psi_1 = \tilde{\phi}_1(x_1, x_2)H_0(x_3)/\pi^{1/4}$.

From the fact that $H_0(x_3) = 1$, it follows from (1.6) that

$$\begin{aligned} \mathcal{Q}(u, \phi_1) &= \pi^{-1/4}\int_\Omega e^{-x_3^2}\{D_1 u D_1\tilde{\phi}_1 + D_2 u D_2\tilde{\phi}_1 + 2^{-1}[1+F(|D_3 u|)]D_3 u D_3\tilde{\phi}_1\} + \langle u, \phi_1\rangle_\rho \\ &= (\eta_1+1)\langle u, \phi_1\rangle_\rho \quad \forall u \in C^1_{p,\rho}(\Omega). \end{aligned}$$

Hence, it follows that

$$\mathcal{Q}(u, \phi_1) = \sigma_1\langle u, \phi_1\rangle_\rho \quad \forall u \in H^1_{p,\rho} \tag{5.10}$$

and (5.2) is established provided we can show $\sigma_1 = \lambda_1^*$, i.e., that Q satisfies (4.6). We also have to show that Q satisfies (4.7) and (Q1)-(Q5).

To establish these facts, we observe from (5.9) that $B_0=1$, $A_1(x,t,\xi)=\xi_1$, $A_2(x,t,\xi)=\xi_2$, and $A_3(x,t,\xi)=2^{-1}[1+F(|\xi_3|)]\xi_3$. It is clear from (5.7) and the fact that $F(t)$ is nondecreasing on $[0,\infty)$ that Q meets (Q1)-(Q5). Also it follows from (5.8) and (5.9) that

$$\mathcal{Q}(u,u)-\mathcal{L}(u,u)=-2^{-1}\int_\Omega e^{-x_3^2}[1-F(|D_3u|)]|D_3u|^2. \tag{5.11}$$

Consequently, $\mathcal{Q}(u,u)\leqq\mathcal{L}(u,u)$ for all $u\in H^1_{p,\rho}$ and we conclude from (1.7) that

$$\lambda_1^*\leqq\liminf_{\|u\|_\rho^2\to\infty}\mathcal{L}(u,u)/\|u\|_\rho^2,\qquad u\in H^1_{p,\rho}. \tag{5.12}$$

But for $u\neq 0$, $\mathcal{L}(u,u)/\|u\|_\rho^2=\mathcal{L}(u/\|u\|_\rho, u/\|u\|_\rho)$ and from (4.15)(a), (b) we see that

$$\inf_{\|u\|_\rho^2=1}\mathcal{L}(u,u)=\sigma_1,\qquad u\in H^1_{p,\rho}.$$

Hence, it follows from (5.12) that

$$\lambda_1^*\leqq\sigma_1. \tag{5.13}$$

On the other hand, from (Q3) and (Q4) if $\lim_{n\to\infty}\langle u_n,u_n\rangle_\rho=\infty$ and

$$\lim_{n\to\infty}\mathcal{Q}(u_n,u_n)/\langle u_n,u_n\rangle_\rho=\lambda_1^*, \tag{5.13}'$$

we see from (4.7) that $\lambda_1^*\geqq\sigma_1$. This fact coupled with (5.13) gives (4.6).

It remains to establish (4.7). It is clear from (5.11) that (4.7) will follow if we can show the following. Given $\{u_n\}_{n=1}^\infty$ an arbitrary sequence in $C^1_{p,\rho}$ such that

$$\lim_{n\to\infty}\|u_n\|_{p,\rho}=\infty, \tag{5.14}$$

then

$$\lim_{n\to\infty}\int_\Omega e^{-x_3^2}[1-F(|D_3u_n|)]|D_3u_n|^2/\|u_n\|_{p,\rho}=0. \tag{5.15}$$

Now we recall once again that $\Omega=\tilde{\Omega}\times\mathbb{R}^1$ where $\tilde{\Omega}\subset\mathbb{R}^2$ is a bounded open connected set. Let $\varepsilon>0$ be given. Using (5.7), choose $T>0$ such that $t|1-F(t)|<\varepsilon$ for $t\geqq T$. Let $\Omega_{1n}=\{x\in\Omega\colon|D_3u_n(x)|>T\}$ and $\Omega_{2n}=\Omega\backslash\Omega_{1n}$. Since $\Omega=\Omega_{1n}\cup\Omega_{2n}$, it follows that

$$\int_\Omega e^{-x_3^2}|1-F(|D_3u_2|)||D_3u_n|^2\leqq T^2(\text{meas }\tilde{\Omega})\pi^{1/2}+\varepsilon\|u_n\|_{p,\rho}(\text{meas }\tilde{\Omega})^{1/2}\pi^{1/4}.$$

Dividing both sides of this last inequality by $\|u_n\|_{p,\rho}$, we see that

$$\limsup_{n\to\infty}\int_\Omega e^{-x_3^2}|1-F(|D_3u_n|)||D_3u_n|^2/\|u_n\|_{p,\rho}\leqq\varepsilon(\text{meas }\tilde{\Omega})^{1/2}\pi^{1/4}.$$

Since ε is arbitrary, (5.15) is true. Hence (4.7) holds, and our example is completely established.

For our final illustration we choose $\Omega=(0,1)\times\mathbb{R}^1$ and $p(x_1,x_2)=\rho(x_1,x_2)=x_1e^{-x_2^2}$. We take

$$Lu=-D_1[x_1\,e^{-x_2^2}D_1u]-D_2[x_1\,e^{-x_2^2}D_2u]+x_1\,e^{-x_2^2}u \tag{5.16}$$

and

$$Qu=-D_1[x_1\,e^{-x_2^2}D_1u]-D_2\{x_1\,e^{-x_2^2}2^{-1}[1+F(|D_2u|)]D_2u\}+x_1\,e^{-x_2^2}u. \tag{5.17}$$

Then it follows from (1.6) and (4.1) that

$$\mathcal{Q}(u, v) - \mathcal{L}(u, v) = -\int_\Omega x_1 e^{-x_2^2} 2^{-1}[1 - F(|D_2 u|)] D_2 u D_2 v. \tag{5.18}$$

Since $\phi_1(x_1, x_2) =$ constant multiple of $J_0(\eta_1 x_1)$ where η_1 is the first positive zero of J_0, it also follows that the integral on the right-hand side of (5.18) is zero when $v = \phi_1$. Hence, we see from (5.18) and (4.3) that

$$\mathcal{Q}(u, \phi_1) = \mathcal{L}(u, \phi_1) = \sigma_1 \langle u, \phi_1 \rangle_\rho \quad \forall u \in H^1_{p,\rho}. \tag{5.19}$$

Therefore, (5.2) will follow once we show that (4.6) holds. We also have to show that (4.7) holds and that Q satisfies (Q1)-(Q5).

We see from (1.0) and (5.17) that $B_0 = 1$, $A_1(x, t, \xi) = \xi_1$, and $A_2(x, t, \xi) = 2^{-1}[1 + F(|\xi_2|)]\xi_2$. It is an easy matter to see from (5.7) and the fact that $F(t)$ is nondecreasing on $[0, \infty)$ that (Q1)-(Q5) hold. Also it follows from (5.18) that $\mathcal{Q}(u, u) \leqq \mathcal{L}(u, u)$ and hence as we did in (5.12) and (5.13), we obtain that $\lambda_1^* \leqq \sigma_1$. But from (5.13'), we obtain as before using (Q3)-(Q4) and (4.7) that $\lambda_1^* \geqq \sigma_1$. Consequently, (4.6) holds. To establish (4.7) we assume (5.14) holds and show with the identical proof used before for (5.15) that

$$\lim_{n \to \infty} \int_\Omega x_1 e^{-x_2^2}[1 - F(|D_2 u_n|)] D_2 u_n / \|u_n\|_{p,\rho} = 0.$$

Consequently, it follows from (5.18) that (4.7) is true. Therefore, Lu and Qu given, respectively, by (5.16) and (5.17) satisfy the conditions in the hypothesis of the corollary and our last example is completely established.

REFERENCES

[1] L. C. ANDREWS, *Elementary Partial Differential Equations with Boundary Value Problems*, Academic Press, New York, 1986.

[2] R. L. BORRELLI AND C. S. COLEMAN, *Differential Equations, a Modeling Approach*, Prentice Hall, Englewood Cliffs, NJ, 1987.

[3] W. E. BOYCE AND R. C. DI PRIMA, *Elementary Differential Equations and Boundary Value Problems*, Third ed., John Wiley, New York, 1977.

[4] H. BREZIS AND L. NIRENBERG, *Characterization of the ranges of some nonlinear operators and applications to boundary value problems*, Ann. Sci. Norm Sup. Pisa, 5 (1978), pp. 225-236.

[5] F. E. BROWDER, *Existence theorems in partial differential equations*, in Proc. Symposia in Pure Mathematics, Vol. XVI, American Mathematical Society, Providence, RI, 1970, pp. 1-60.

[6] D. G. DE FIGUEIREDO AND J. P. GOSSEZ, *Nonlinear perturbations of a linear elliptic problem near its first eigenvalue*, J. Differential Equations, 30 (1978), pp. 1-19.

[7] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Second ed., Springer-Verlag, Berlin, 1983.

[8] H. HOCHSTADT, *The Functions of Mathematical Physics*, Wiley-Interscience, New York, 1971.

[9] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[10] E. M. LANDESMAN AND A. C. LAZER, *Nonlinear perturbations of linear elliptic boundary value problems*, J. Math. Mech., 7 (1970), pp. 609-623.

[11] J. LERAY AND J. L. LIONS, *Quelques résultats de Višik sur les problèmes elliptiques non linéaires par les méthodes de Minty-Browder*, Bull. Soc. Math. France, 93 (1965), pp. 97-107.

[12] L. NIRENBERG, *Topics in Nonlinear Functional Analysis*, Lecture Notes, Courant Institute of Mathematical Sciences, New York University, New York, 1974.

[13] W. RUDIN, *Real and Complex Analysis*, Second ed., McGraw-Hill, New York, 1974.

[14] G. SZEGO, *Orthogonal Polynomials*, American Mathematical Society Colloquium Publications, Fourth ed., American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1430–1441, September 1991

014

BIORTHOGONALITY OF A SYSTEM OF RATIONAL FUNCTIONS WITH RESPECT TO A POSITIVE MEASURE ON [−1,1]*

MIZAN RAHMAN†

Abstract. By using the summation and transformation formulas for balanced very well poised basic hypergeometric series an extension of Askey and Wilson's q-beta integral involving five parameters is evaluated, where one of the parameters has modulus greater than 1. The result is then applied to obtain the biorthogonality relation for a system of real-valued rational functions on $[-1,1]$ representable as terminating balanced very well poised ${}_{10}\phi_9$ series.

Key words. balanced and very well poised basic hypergeometric series, q-beta integrals, biorthogonality of rational functions

AMS(MOS) subject classification. 33A65

1. Introduction.

Euler's beta integral

$$\int_{-1}^{1}(1-x)^{\alpha}(1+x)^{\beta}\,dx = 2^{\alpha+\beta+1}\frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}, \tag{1.1}$$

$\mathrm{Re}(\alpha,\beta) > -1$, has been extended in many different ways (see [1]–[10], [17]), but the extension that has received particular attention in recent years is Askey and Wilson's [10] q-beta integral given by

$$\frac{1}{2\pi}\int_{-1}^{1} w(x;a,b,c,d)\,dx = \frac{(abcd;q)_\infty}{(q,ab,ac,ad,bc,bd,cd;q)_\infty}, \tag{1.2}$$

where

$$w(x;a,b,c,d) = \frac{h(x;1,-1,q^{1/2},-q^{1/2})}{h(x;a,b,c,d)}(1-x^2)^{-1/2}, \tag{1.3}$$

with

$$h(x;a_1,a_2,\cdots,a_k) = \prod_{j=1}^{k} h(x;a_j), \tag{1.4}$$

$$\begin{aligned} h(x;a) &= \prod_{n=0}^{\infty}(1-2axq^n+a^2q^{2n}) \\ &= (ae^{i\theta},ae^{-i\theta};q)_\infty \qquad \text{when } x=\cos\theta, \end{aligned} \tag{1.5}$$

and the q-shifted factorials are defined by

$$(a;q)_n = \begin{cases} 1 & \text{if } n=0, \\ (1-a)(1-aq)\cdots(1-aq^n) & \text{if } n=1,2,\cdots, \end{cases} \tag{1.6}$$

*Received by the editors September 8, 1989; accepted for publication September 10, 1990.

†Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6, Canada. This research was supported by the Natural Sciences and Engineering Research Council of Canada under grant A6197.

$$(a,q)_\infty = \lim_{n\to\infty}(a;q)_n \quad \text{when } |q|<1, \tag{1.7}$$

$$(a_1,a_2,\cdots,a_k;q)_n = \prod_{j=1}^{k}(a_j;q)_n. \tag{1.8}$$

It will be assumed throughout the paper that $|q|<1$. Formula (1.2) is valid provided $\max(|a|,|b|,|c|,|d|)<1$. If one of the parameters, say a, is greater than 1 in modulus but is such that $|aq^{m+1}|<1<|aq^m|$ for some nonnegative integer m then, as Askey and Wilson [10] showed by a contour integration argument, and as Gasper and Rahman [13] proved by real integration, (1.2) must be modified by the formula

$$\frac{1}{2\pi}\int_{-1}^{1} w(x;a,b,c,d)\,dx + \sum_{k=0}^{m} w_k = \frac{(abcd;q)_\infty}{(q,ab,ac,ad,bc,cd;q)_\infty}, \tag{1.9}$$

where

(1.10)

$$w_k = \frac{(a^{-2};q)_\infty}{(q,ab,ac,ad,a/b,a/c,a/d;q)_\infty}\cdot\frac{(1-a^2q^{2k})(a^2,ab,ac,ad;q)_k}{(1-a^2)(q,aq/b,aq/c,aq/d;q)_k}\left(\frac{q}{abcd}\right)^k.$$

Nasrallah and Rahman [15] considered an extension of the integral in (1.2) by introducing a fifth parameter and in [16] Rahman gave the following extension:

$$\frac{1}{2\pi}\int_{-1}^{1} v(x;a,b,c,d,f)\,dx = \kappa(a,b,c,d,f), \tag{1.11}$$

where

$$v(x;a,b,c,d,f) = \frac{h(x;1,-1,q^{1/2},-q^{1/2},abcdf)}{h(x;a,b,c,d,f)}(1-x^2)^{-1/2}, \tag{1.12}$$

and

$$\kappa(a,b,c,d,f) = \frac{(abcd,abcf,abdf,acdf,bcdf;q)_\infty}{(q,ab,ac,ad,af,bc,bd,bf,cd,cf,df;q)_\infty}, \tag{1.13}$$

provided

$$\max(|a|,|b|,|c|,|d|,|f|)<1. \tag{1.14}$$

Askey [5] gave a direct proof of (1.11) by a functional equation method. Our first objective in this paper is to give an evaluation of the integral in (1.11) when the modulus of one of the parameters exceeds 1 while the moduli of the other four parameters remain less than 1. Suppose m is a nonnegative integer such that

$$|fq^{m+1}|<1<|fq^m|. \tag{1.15}$$

We will prove in §3 that

$$\begin{aligned}\frac{1}{2\pi}\int_{-1}^{1} v(x;a,b,c,d,f)\,dx + \sum_{k=0}^{m} v_k(a,b,c,d,f)\\ = \kappa(a,b,c,d,f),\end{aligned} \tag{1.16}$$

where

$$v_k(a,b,c,d,f) = \frac{(abcd, abcdf^2, f^{-2}; q)_\infty}{(q, af, bf, cf, df, a/f, b/f, c/f, d/f; q)_\infty} \cdot \frac{(1-f^2q^{2k})(f^2, af, bf, cf, df, q/abcd; q)_k}{(1-f^2)(q, fq/a, fq/b, fq/c, fq/d, abcdf^2; q)_k} q^k. \tag{1.17}$$

It is clear that (1.16) reduces to (1.9) when any one of the parameters a, b, c, d approaches zero.

The main objective of this paper, however, is to prove the following theorems.

THEOREM 1. *Let* $\max(|a|,|b|,|c|,|d|) < 1$, $|f| \le \min(|a|,|b|,|c|,|d|)$ *and let* N *be a nonnegative integer such that*

$$|fq^{-N}| < 1 < |fq^{-N-1}|. \tag{1.18}$$

Let

$$\begin{aligned} R_m(x) &= R_m(x; a, b, c, d, f) \\ &= \frac{(a^2bcdf, bcdf, de^{i\theta}, de^{-i\theta}; q)_m}{(ad, d/a, abcdfe^{i\theta}, abcdfe^{-i\theta}; q)_m} \\ &\quad \cdot M_m(x; a, b, c, d, f), \end{aligned} \tag{1.19}$$

$$\begin{aligned} S_n(x) &= S_n(x; a, b, c, d, f) \\ &= \frac{(aq/f, q/af, de^{i\theta}, de^{-i\theta}; q)_n}{(ad, d/a, qe^{i\theta}/f, qe^{-i\theta}/f; q)_n} \\ &\quad \cdot M_n(x; a, b, c, d, fq^{-1}), \end{aligned} \tag{1.20}$$

where $x = \cos\theta$, $0 \le \theta \le \pi$, *and*

$$\begin{aligned} &M_k(x; a, b, c, d, f) \\ &\quad = {}_{10}W_9(aq^{-k}/d; q^{1-k}/bd, q^{1-k}/cd, 1/df, abcf, ae^{i\theta}, ae^{-i\theta}, q^{-k}; q, q), \end{aligned} \tag{1.21}$$

where the symbol ${}_{10}W_9$ *represents a very well poised* ${}_{10}\phi_9$ *series defined in* §2. *Let*

$$L_{m,n} := \int_{-1}^{1} v(x; a, b, c, d, f) R_m(x) S_n(x)\, dx. \tag{1.22}$$

If $m = 0, 1, \cdots$, *and* $n = 0, 1, \cdots, N$, *then*

$$L_{m,n} = \kappa(a, b, c, d, f)\delta_{m,n}/h_n, \tag{1.23}$$

where

$$h_n = \frac{(1-abcdq^{2n-1})(abcdq^{-1}, ab, ac, ad, 1/af, bcdf/q; q)_n}{(1-abcdq^{-1})(q, cd, bd, bc, a^2bcdf, aq/f; q)_n} q^n. \tag{1.24}$$

THEOREM 2. *Let* a, b, c, d, f *satisfy the same conditions as in Theorem* 1. *If* $n \ge N+1$, *then*

$$\begin{aligned} &L_{m,n} + \sum_{k=0}^{n-N-1} v_k(a, b, c, dq^n, fq^{-n}) R_m(x_{n-k}; a, b, c, d, f) \\ &\quad \cdot M_n(x_{n-k}; a, b, c, d, fq^{-1}) \frac{(afq^{-n}, fq^{-n}/a; q)_n}{(ad, d/a; q)_n} \\ &\quad = \kappa(a, b, c, d, g)\delta_{m,n}/h_n, \end{aligned} \tag{1.25}$$

where

$$x_{n-k} = \tfrac{1}{2}(fq^{k-n} + f^{-1}f^{n-k}). \tag{1.26}$$

In §4 we will give the proofs of Theorems 1 and 2. In §2 we will give the definition and some properties of the very well poised basic hypergeometric series that will be needed to complete the proofs in §3 and §4.

In [16] Rahman found a biorthogonality relation for essentially the same rational functions, but with respect to a complex measure on the unit circle. This paper deals with the problems that arise when we try to convert that integral to a real one on the interval $[-1, 1]$.

2. Very well poised basic hypergeometric series. A basic hypergeometric series ${}_{r+1}\phi_r$ in base q and with $r+1$ numerator parameters and r denominator parameters is defined by

$${}_{r+1}\phi_r\left[\begin{matrix}a_1, a_2, \cdots, a_{r+1} \\ b_1, \cdots, b_r\end{matrix}; q, z\right] = \sum_{n=0}^{\infty} \frac{(a_1, a_2, \cdots, a_{r+1}; q)_n}{(q, b_1, \cdots, b_r; q)_n} z^n, \tag{2.1}$$

which terminates and hence becomes a polynomial of degree k in z when one of the numerator parameters is of the form $q^{-k}, k = 0, 1, \cdots$. If the series does not terminate, then we will naturally assume that the series converges, which is guaranteed by the inequality $|z| < 1$. Whether or not the series terminates it will always be assumed that the denominator parameters are such that no zero factors appear in any term of the series in (2.1). When $z = q$ and $qa_1a_2 \cdots a_{r+1} = b_1 \cdots b_r$, we call the series balanced. If $qa_1 = a_2b_1 = a_3b_2 = \cdots = a_{r+1}b_r$, the series is said to be well poised. If, in addition, $a_2 = qa_1^{1/2}$, $a_3 = -qa_1^{1/2}$, the series is said to be very well poised. Because of the frequent use of very well poised series in this paper we shall use the following abbreviated notation:

$$\begin{aligned} &{}_{r+1}W_r(a_1; a_4, a_5, \cdots, a_{r+1}; q, z) \\ &\quad := {}_{r+1}\phi_r\left[\begin{matrix}a_1, qa_1^{1/2}, -qa_1^{1/2}, a_4, \cdots, a_{r+1} \\ a_1^{1/2}, -a_1^{1/2}, a_1q/a_4, \cdots, a_1q/a_{r+1}\end{matrix}; q, z\right]. \end{aligned} \tag{2.2}$$

One of the most useful summation formulas in the theory of basic hypergeometric series is Jackson's [14] formula:

$${}_8W_7(a; b, c, d, e, q^{-n}; q, q) = \frac{(aq, aq/bc, aq/bd, aq/cd; q)_n}{(aq/b, aq/c, aq/d, aq/bcd, ; q)_n}, \tag{2.3}$$

where $n = 0, 1, \cdots$, and

$$a^2q^{n+1} = bcde, \tag{2.4}$$

which expresses that the terminating ${}_8W_7$ series in (2.3) is balanced. Almost as important as (2.3) is Bailey's [11] nonterminating extension of it:

$$\begin{aligned} &{}_8W_7(a; b, c, d, e, f; q, q) \\ &\quad + \frac{(aq, b/a, c, d, e, f, bq/c, bq/d, bq/e, bq/f; q)_\infty}{(qb^2/a, a/b, bc/a, bd/a, be/a, bf/a, aq/c, aq/e, aq/f; q)_\infty} \\ &\quad \cdot {}_8W_7(b^2/a; b, bc/a, bd/a, be/a, bf/a; q, q) \\ &\quad = \frac{(aq, b/a, aq/cd, aq/ce, aq/cf, aq/de, aq/ef; q)_\infty}{(aq/c, aq/d, aq/e, aq/f, bc/a, bd/a, be/a, bf/a; q)_\infty} \end{aligned} \tag{2.5}$$

where the balance condition (2.4) is replaced by

$$qa^2 = bcdef. \tag{2.6}$$

We shall also need Bailey's [12] ${}_{10}\phi_9$ transformation formula see [13, eqn. (2.12.9)], which can be written in a somewhat compact form as follows:

$$\begin{aligned} &V(a;b,c,d,e,f,g,h;q,q) \\ &\quad = \frac{(aq,b/a,\lambda q/f,\lambda q/g,\lambda q/h,bf/\lambda,bg/\lambda,bh/\lambda;q)_\infty}{(\lambda q,b/\lambda,aq/f,aq/g,aq/h,bf/a,bg/a,bh/a;q)_\infty} \\ &\qquad \cdot V(\lambda;b,\lambda c/a,\lambda d/a,\lambda e/a,fg,h;q,q), \end{aligned} \tag{2.7}$$

where

$$q^2a^3 = bcdefgh, \qquad \lambda = qa^2/cde, \tag{2.8}$$

and

(2.9)

$$\begin{aligned} &V(a;b,c,d,e,f,g,h;q,q) \\ &= {}_{10}W_9(a;b,c,d,e,f,g,h;q,q) \\ &\quad + \frac{(aq,b/a,c,d,e,f,g,h,bq/c,bq/d,bq/f,bq/g,bq/h;q)_\infty}{(qb^2/a,a/b,bc/a,bd/a,be/a,bf/a,bg/a,bh/a,aq/c,aq/d,aq/e,aq/f,aq/g,aq/h;q)_\infty} \\ &\quad \cdot {}_{10}W_9(b^2/a;b,bc/a,bd/a,be/a,bf/a,bg/a,bh/a;q,q). \end{aligned}$$

If $h = q^{-n}, n = 0, 1, \cdots$, then the coefficients of the second ${}_{10}W_9$ series in both V-functions in (2.7) vanish and we get the terminating form of (2.7), namely,

(2.10)

$$\begin{aligned} &{}_{10}W_9(a;b,c,d,e,f,g,q^{-n},q,q) \\ &\quad = \frac{(aq,aq/bf,aq/bg,aq/fg;q)_n}{(aq/b,aq/f,aq/b,aq/bfg;q)_n} {}_{10}W_9(\lambda;b,c\lambda/a,d\lambda/a,e\lambda/a,f,g,q^{-n};q,q). \end{aligned}$$

In [16] the author found a representation of a V-function in terms of an integral of the type (1.2) and (1.11). By an iteration of the transformation formula (2.7) that representation can be expressed in the following form:

$$\begin{aligned} &\frac{1}{2\pi}\int_{-1}^{1} \frac{h(x;1,-1.q^{1/2},-q^{1/2},c,d)(1-x^2)^{-1/2}}{h(x;a_1,a_2,a_3,a_4,a_5,a_6)}\,dx \\ &\quad = \frac{\prod_{j=1}^{6}(ca_j,d/a_j;q)_\infty}{(q,c^2,d/c;q)_\infty \prod_{1\le j<k\le 6}(a_ja_k;q)_\infty} \\ &\qquad \cdot V(c^2/q;cd/q,c/a_1,c/a_2,c/a_3,c/a_4,c/a_5,c/a_6;q,q), \end{aligned} \tag{2.11}$$

where the balance condition of the V-function as well as the integrand in (2.11) is expressed by the relation

$$cd = a_1a_2a_3a_4a_5a_6. \tag{2.12}$$

The modulus of each of the denominator parameters $a_j, j = 1, \cdots, 6$, is, of course, less than 1. Situations where this restriction may not hold are no interest to us in this paper. The form of the expression on the right side of (2.11) may not be the most useful one but it is the easiest one to remember because of its symmetry.

3. Proof of (1.16). If $f = \pm 1$ in (1.12) then the functions $h(x; \pm 1)$ and $h(x; f)$ cancel, and hence it follows by continuity that

$$\frac{1}{2\pi}\int_{-1}^{1} v(x; a, b, c, d, \pm 1)\, dx = \kappa(a, b, c, d, \pm 1), \tag{3.1}$$

provided, of course, that

$$\max(|a|, |b|, |c|, |d|) < 1. \tag{3.2}$$

If $f \neq \pm 1$ but $|f| = 1$, then $h(x; f) = 0$ for some x in the interval $(-1, 1)$, and so the integral in (1.11) does not converge. Similary, this integral does not converge if $|fq^n| = 1$ and $fq^n \neq \pm 1$, for some positive integer n. If there is a nonnegative integer m such that

$$|fq^{m+1}| < 1 < |fq^m| \tag{3.3}$$

and if $fa^{\pm 1}, fb^{\pm 1}, fc^{\pm 1}, fd^{\pm 1}$ are not of the form q^{-n} for any nonnegative integer n, then the integral in (1.11) converges and it can be evaluated in the following way. Since

$$\begin{aligned} h(x; f) &= (fe^{i\theta}, fe^{-i\theta}; q)_{m+1} h(x; fq^{m+1}) \\ &= f^{2m+2} q^{m+m^2} \frac{h(x; fq^{m+1}, f^{-m}/f)}{h(x; q/f)}, \end{aligned} \tag{3.4}$$

we get by denoting the integral on the left side of (1.11) by $J(a, b, c, d, f)$,

$$\begin{aligned} &J(a, b, c, d, f) \\ &\quad = f^{-2m-2} q^{-m-m^2} \int_{-1}^{1} \frac{h(x; 1, -1, q^{1/2}, -q^{1/2}, abcdf, q/f)(1 - x^2)^{-1/2}}{h(x; a, b, c, d, fq^{m+1}, q^{-m}/f)}\, dx, \end{aligned} \tag{3.5}$$

where it is assumed that the inequality (3.2) is also satisfied. Since the parameters in the integrand of (3.5) satisfy (2.12), we may apply (2.11) to obtain the following relation:

$$\begin{aligned} J(a, b, c, d, f) =& K_m(a, b, c, d, f) \\ &\cdot V((abcdf)^2/q; abcd, abcf, abdf, acdf, bcdf, abcdq^{-m-1}, abcdf^2 q^m; q, q), \end{aligned} \tag{3.6}$$

where

$$\begin{aligned} K_m(a, b, c, d, f) = &- \frac{(a^2bcdf, ab^2cdf, abc^2df, abcd^2f, abcdf^2; q)_\infty}{(q, ab, ac, ad, bc, bd, cd, af, bf, cf, df; q)_\infty} \\ &\cdot \frac{(abcd, f^{-2}; q)_\infty}{(a/f, b/f, c/f, d/f, (abcdf)^2, 1/abcdf^2; q)_\infty} \\ &\cdot \frac{(qf^2, q/abcd, afq, bfq, cfq, dfq; q)_m}{(q, qabcdf^2, fq/a, fq/b, fq/c, fq/d; q)_m}. \end{aligned} \tag{3.7}$$

We now apply (2.7) on the V-function on the right side of (3.6) to get

$$
\begin{aligned}
&V((abcdf)^2/q; abcd, abcf, abdf, acdf, bcdf, abcdq^{-m-1}, abcdf^2q^m; q, q)\\
&\quad = \frac{((abcdf)^2, q/abcdf^2, abcfq, abc/f, dfq, d/f; q)_\infty}{(a^2b^2c^2df, q/abcf, abcd^2f, q/df, qabcdf^2, f^{-2}; q)_\infty}\\
&\qquad \cdot (qabcdf^2, q, fq/d, fq/abc; q)_m/(abcfq, dfq, qf^2, q/abcd; q)_m\\
&\qquad \cdot V(a^2b^2c^2df/q; abcd, ab, ac, bc, abcf, abcdq^{-m-1}, abcdf^2q^m; q, q).
\end{aligned}
\tag{3.8}
$$

Using the definition (2.9) of the V-functions, we have

$$
\begin{aligned}
&V(a^2b^2c^2df/q; qbcd, 2b, 2c, bc, abcf, abcdq^{-m-1}, abcdf^2q^m; q, q)\\
&\quad = {}_{10}W_9(a^2b^2c^2df/q; abcd, ab, ac, bc, abcf, abcdq^{-m-1}, abcdf^2q^m; q, q)\\
&\qquad + \frac{(a^2b^2c^2df, q/abcf, ab, ac, bc, abcf, abcdq^{-m-1}, abcdf^2q^m; q)_\infty}{(dq^2/f, abcf/q, q/cf, q/bf, q/af, q, dq^{-m}/fdfq^{m+1}; q)_\infty}\\
&\qquad \cdot \frac{(q/cd, q/bd, q/ad, dq/f, q^{m+2}, q^{1-m}/f^2; q)_\infty}{(abc^2df, ab^2cdf, a^2bcdf, abcd, abcfq^{m+1}, abcq^{-m}/f; q)_\infty}\\
&\qquad \cdot {}_{10}W_9(dq/f; abcd, q/af, q/bf, q/cf, q, dq^{-m}/f, dfq^{m+1}; q, q).
\end{aligned}
\tag{3.9}
$$

Because of the cancellation of a numerator and a denominator parameter in the first ${}_{10}W_9$ series on the right side of (3.9), it reduces to

$$
{}_8W_7(a^2b^2c^2df/q; ab, ac, bc, abcdq^{-m-1}, abcdf^2q^m; q, q),
$$

which is very well poised and balanced and so Bailey's summation formula (2.5) is applicable. Thus

$$
\begin{aligned}
&{}_{10}W_9(a^2b^2c^2df/q; abcd, ab, ac, bc, abcf, abcdq^{-m-1}, abcdf^2q^m; q, q)\\
&\quad = \frac{(a^2b^2c^2df, abdf, acdf, bcdf, aq^{-m}/f, bq^{-m}/f, cq^{-m}/f, q^{-m}/abcf; q)_\infty}{(a^2bcdf, ab^2cdf, abc^2df, df, q^{-m}/af, q^{-m}/bf, q^{-m}/cf, abcq^{-m}/f; q)_\infty}\\
&\qquad - \frac{(a^2b^2c^2df, ab, ac, bc, abcdf^2q^m, adq^{-m}, bdq^{-m}, cdq^{-m}; q)_\infty}{(a^2bcdf, ab^2cdf, abc^2df, df, abcfq^m, q^{-m}/af, q^{-m}/bf, q^{-m}/cf; q)_\infty}\\
&\qquad \cdot \frac{(q^{-m}/abcf, q^{-2m}/f^2; q)_\infty}{(abcq^{-m}/f, dq^{-2m}/f; q)_\infty}\\
&\qquad \cdot {}_8W_7(dq^{-2m-1}/f; abcdq^{-m-1}, q^{-m}/af, q^{-m}/bf, q^{-m}/cf, df; q, q).
\end{aligned}
\tag{3.10}
$$

By using cancellation this last ${}_8W_7$ series can be written as a very well poised nonterminating balanced ${}_{10}W_9$ series in many ways, but the one of particular interest to us is the following form:

$$
\begin{aligned}
&{}_8W_7(dq^{-2m-1}/f; abcdq^{-m-1}, q^{-m}/af, q^{-m}/bf, q^{-m}/cf, df; q, q)\\
&\quad = {}_{10}W_9(dq^{-2m-1}/f; dq^{-m}/f, q^{-m}/af, q^{-m}/bf,\\
&\qquad q^{-m}/cf, abcdq^{-m-1}, df, q^{-m}; q, q).
\end{aligned}
\tag{3.11}
$$

The appearance of q^{-m} in the numerator may indicate that it is a terminating series, but in fact it is not because of the cancellation with the q^{-m} term that occurs in the

denominator below dq^{-m}/f when we write out the ${}_{10}W_9$ series in (3.11) as a very well poised ${}_{10}W_9$ series. The whole purpose of this particular choice is that the ${}_{10}W_9$ in (3.11) can now be matched with the second ${}_{10}W_9$ series on the right side of (3.9) and their coefficients are such that they combine into a single V-function.

Using (3.7)–(3.10) in (3.6) and simplifying the coefficients we obtain

(3.12)
$$\begin{aligned} &J(a,b,c,d,f)\\ &\quad=\kappa(a,b,c,d,f)-\frac{(abcd,abcfq^{m+1},abcdf^2q^m,f^{-2};q)_\infty(1-abcf)}{(q,df,abcfq^m,afq,bfq,cfq,a/f,b/f,c/f,d/f;q)_\infty}\\ &\qquad\cdot\frac{(q/ad,q/bd,q/cd,abcfq;q)_m(qf^2;q)_{2m}(df)^m}{(fq/a,fq/b,fq/c;q)_m(fq/d;q)_{2m}}\\ &\qquad\cdot V(dq^{-2m-1}/f;dq^{-m}/f,q^{-m}/af,q^{-m}/bf,q^{-m}/cf,df,abcdq^{-m-1},q^{-m};q,q). \end{aligned}$$

We now apply (2.7) on the V-function on the right side of (3.12) and find that because $(q^{-m};q)_\infty=0$ for $m=0,1,\cdots$, it transforms to a multiple of a single terminating balanced ${}_{10}W_9$ series. Thus we find that

(3.13)
$$\begin{aligned} &J(a,b,c,d,f)\\ &\quad=\kappa(a,b,c,d,f)-\frac{(abcd,abcdf^2,f^{-2};q)_\infty}{(q,af,bf,cf,df,a/f,b/f,c/f,d/f;q)_\infty}\\ &\qquad\cdot\frac{(q/ad,q/bd,q/cd,q/abcd,qf^2,dfq;q)_m}{(q,fq/a,fq/b,fq/c,fq/d,q/abcd^2f;q)_m}\\ &\qquad\cdot{}_{10}W_9(abcd^2fq^{-m-1};dq^{-m}/f,abdf,acdf,bcdf,df,abcdq^{-m-1},q^{-m};q,q). \end{aligned}$$

Using an iteration of (2.10) (see also [13, Ex. (2.19)]), we may now transform this ${}_{10}W_9$ series as follows:

(3.14)
$$\begin{aligned} &{}_{10}W_9(abcd^2fq^{-m-1};dq^{-m}/f,abdf,acdf,bcdf,df,abcdq^{-m-1},q^{-m};q,q)\\ &\quad=\frac{(q/abcd^2f,fq/a,fq/b,fq/c,fq/d,q;q)_m}{(q/ad,q/bd,q/cd,q/abcd,qf^2,dfq;q)_m}\\ &\qquad\cdot{}_{10}W_9(f^2;af,bf,cf,df,q/abcd,f^2q^{m+1},q^{-m};q,q)\\ &\quad=\frac{(q/abcd^2f,fq/a,fq/b,fq/c,fq/d,q;q)_m}{(q/ad,q/bd,q/cd,q/abcd,qf^2,dfq;q)_m}\\ &\qquad\cdot\sum_{k=0}^{m}\frac{(1-f^2q^{2k})(f^2,af,bf,cf,df,q/abcd;q)_k}{(1-f^2)(q,fq/a,fq/b,fq/c,fq/d,abcdf^2;q)_k}q^k. \end{aligned}$$

Substitution of (3.14) into (3.13) then yields (1.16).

It may be remarked that the same method can be applied to evaluate the integral in (1.11) when two or more parameters satisfy inequalities of the form (1.15) and other restrictions so that no zero factors appear in the denominators.

4. Proofs of Theorems 1 and 2. By successive application of (2.10) it can be shown that

$$
\begin{aligned}
R_m(x) &= {}_{10}W_9(a^2bcdfq^{-1}; ae^{i\theta}, ae^{-i\theta}, abcf, abdf, acdf, abcdq^{m-1}, q^{-m}; q, q) \\
&= \frac{(bc, bd, 1/bf, a^2bcdf; q)_m}{(ac, ad, 1/af, ab^2cdf; q)_m} \\
&\quad \cdot {}_{10}W_9(ab^2cdfq^{-1}; be^{i\theta}, be^{-i\theta}, bcdf, bdaf, bcaf, abcdq^{-m-1}, q^{-m}; q, q) \\
&= \frac{(bc, bd, 1/bf, a^2bcdf; q)_m}{(ac, ad, 1/af, ab^2cdf; q)_m} \sum_{j=0}^{m} \frac{(1 - ab^2cdfq^{2j-1})(ab^2cdfq^{-1}, abcf; q)_j}{(1 - ab^2cdfq^{-1})(q, bd; q)_j} \\
&\quad \cdot \frac{(abdf, bcdf, abcdq^{m-1}, q^{-m}; q)_j}{(bc, ab, bfq^{1-m}, ab^2cdfq^m; q)_j} q^j \frac{h(x; b, abcdfq^j)}{h(x; bq^j, abcdf)}.
\end{aligned} \tag{4.1}
$$

Also, (1.20) can be written in the form

$$
\begin{aligned}
&S_n(x) \\
&= \frac{(fq^{-n}/a, afq^{-n}; q)_n}{(ad, d/a; q)_n} \sum_{k=0}^{n} \frac{(1 - aq^{2k-n}/d)(aq^{-n}/d, q^{1-n}/bd, q^{1-n}/cd; q)_k}{(1 - aq^{-n}/d)(q, ab, ac; q)_k} \\
&\quad \cdot \frac{(q/df, abcf/q, q^{-n}; q)_k}{(afq^{-n}, q^{2-n}/bcdf, aq/d; q)_k} d^{2k} q^{2nk-k^2} \frac{h(x; a, d, f)}{h(x; aq^k, dq^{n-k}, fq^{-n})}.
\end{aligned} \tag{4.2}
$$

In (4.1) and (4.2) $x = \cos\theta$, and the h functions are as defined in (1.4) and (1.5). Substituting (4.1) and (4.2) into (1.22) we find that

$$
\begin{aligned}
L_{m,n} &= \frac{(bc, bd, 1/bf, a^2bcdf; q)_m (fq^{-n}/a, afq^{-n}; q)_n}{(ac, ad, 1/af, ab^2cdf; q)_m (ad, d/a; q)_n} \\
&\quad \cdot \sum_{j=0}^{m} \frac{(1 - ab^2cdfq^{2j-1})(ab^2cdfq^{-1}, abcf, abdf, bcdf; q)_j}{(1 - ab^2cdfq^{-1})(q, bd, bc, ab; q)_j} \\
&\quad \cdot \frac{(abcdq^{m-1}, q^{-m}; q)_j}{(bfq^{1-m}, ab^2cdfq^m; q)_j} q^j \sum_{k=0}^{n} \frac{(1 - aq^{2k-n}/d)(aq^{-n}/d, q^{1-n}/bd; q)_k}{(1 - aq^{-n}/d)(q, ab; q)_k} \\
&\quad \cdot \frac{(q^{1-n}/cd, q/df, abcf/q, q^{-n}; q)_k}{(ac, afq^{-n}, q^{2-n}/bcdf, aq/d; q)_k} d^{2k} q^{2nk-k^2} I_{m,n,j,k},
\end{aligned} \tag{4.3}
$$

where

$$
I_{m,n,j,k} = \frac{1}{2\pi} \int_{-1}^{1} v(x; aq^k, bq^j, c, dq^{n-k}, fq^{-n}) dx. \tag{4.4}
$$

For $j \geq 0$, $n \geq k \geq 0$, where $n = 0, 1, \cdots, N$ with $|fq^{-N}| < 1$, the parameters in the integrand of (4.4) are all less than 1 in the modulus and so, by (1.11),

$$
\begin{aligned}
I_{m,n,j,k} &= \kappa(aq^k, bq^j, c, dq^{n-k}, fq^{-n}) \\
&= \kappa(a, b, c, d, f) \frac{(ad, bd, cd, abcfq^{-n}; q)_n}{(afq^{-n}, bfq^{-n}, cfq^{-n}, abcd; q)_n} \\
&\quad \cdot \frac{(ab, bc, bfq^{-n}, bdq^n; q)_j}{(abcdq^n, abcfq^{-n}, abdf, bcdf; q)_j} \\
&\quad \cdot \frac{(afq^{-n}, abq^j, ac, q^{1-j}/bcdf; q)_k}{(q/df, q^{1-n-j}/bd, q^{1-n}/cd, abcfq^{j-n}; q)_k} d^{-2k} q^{k+k^2-2nk}.
\end{aligned} \tag{4.5}
$$

When we substitute this into (4.3) the series over k reduces to

$$ {}_8W_7(aq^{-n}/d; q^{1-n}/bd, abq^j, q^{1-j}/bcdf, abcf/q, q^{-n}; q, q), $$

which is balanced and terminating and hence can be summed by (2.3). The sum is

$$ \frac{(aq^{1-n}/d, cfq^{-n}, ab^2cdfq^{j-1}, q^{-j}; q)_n}{(ab, q^{1-j-n}/bd, abcfq^{j-n}, bcdf/q; q)_n}, $$

which vanishes unless $j \geq n$. Thus, $L_{m,n} = 0$ if $m < n$. Hence, for $m \geq n$ we find after some simplifications that

(4.6)

$$ \begin{aligned} L_{m,n} &= \kappa(a, b, c, d, f) \\ &\quad \cdot \frac{(bc, bd, 1/bf, a^2bcdf; q)_m (aq/f, cd; q)_n (ab^2cdf; q)_{2n}}{(ac, ad, 1/af, ab^2cdf; q)_m (ab, bcdf/q; q)_n (abcd; q)_{2n}} (bf)^n \\ &\quad \cdot \frac{(abcdq^{m-1}, q^{-m}; q)_n}{(bfq^{1-m}, ab^2cdfq^m; q)_n} {}_6W_5(ab^2cdfq^{2n-1}; bf, abcdq^{m+n-1}, q^{n-m}; q, q). \end{aligned} $$

The ${}_6W_5$ series in (4.6) is a special case of the ${}_8W_7$ series in (2.3) and it has the sum

$$ \frac{(ab^2cdfq^{2n}, q^{1+n-m}; q)_{m-n}}{(abcdq^{2n}, bfq^{1-m+n}; q)_{m-n}}, $$

which vanishes unless $m = n$. Hence

$$ \begin{aligned} L_{m,n} &= \kappa(a, b, c, d, f) \\ &\quad \cdot \frac{(bc, bd, cd, 1/bf, aq/f, a^2bcdf; q)_n (abcdq^{n-1}, q^{-n}; q)_n}{(ad, ac, ab, bcdf/q, 1/af; q)_n (bfq^{1-n}; q)_n (abcd; q)_{2n}} \\ &\quad \cdot (bf)^n \delta_{m,n} \qquad (4.7) \\ &= \kappa(a, b, c, d, f) \\ &\quad \cdot \frac{(1 - abcdq^{-1})(q, bc, bd, cd, aq/f, a^2bcdf; q)_n}{(1 - abcdq^{2n-1})(abcdq^{-1}, ad, ac, ab, bcdf/q, 1/af; q)_n} q^{-n} \delta_{m,n}. \end{aligned} $$

This completes the proof of Theorem 1.

Thus, the system of rational functions $S_n(x)$, $n = 0, 1, \cdots, N$, is biorthogonal to $R_m(x)$, $m = 0, 1, \cdots$, when $|fq^{-N}| < 1 < |fq^{-N-1}|$ with respect to the positive measure $v(x; a, b, c, d, f)dx$ on $(-1, 1)$. Orthogonality of finitely many polynomials with respect to positive measures has been a topic of current interest; see Askey [7], [8].

To prove Theorem 2 we observe that when $n = N + 1, N + 2, \cdots, L_{m,n}$ is again given by (4.3) and (4.4), but because of the fact that $|fq^{-n}| > 1$ with $|fq^{-N}| < 1 < |fq^{-N-1}|$, $I_{m,n,j,k}$ has an additional contribution which, via (1.16), is

$$ \begin{aligned} &- \frac{(abcdq^{n+j}, abcdf^2q^{j-n}, f^{-2}q^{2n}; q)_\infty}{(q, afq^{k-n}, bfq^{j-n}, cfq^{-n}, dfq^{-k}, aq^{k+n}/f, bq^n/f, cq^n/f, dq^{2n-k}/f; q)_\infty} \\ &\quad \cdot \sum_{l=0}^{n-N-1} \frac{(1 - f^2q^{2l-2n})(f^2q^{-2n}, afq^{k-n}, bfq^{j-n}, cfq^{-n}; q)_l}{(1 - f^2q^{-2n})(q, fq^{1-k-n}/a, fq^{1-j-n}/b, fq^{1-n}/c; q)_l} \qquad (4.8) \\ &\quad \cdot \frac{(dfq^{-k}, q^{1-n-j}/abcd; q)_l q^l}{(fq^{1-2n+k}/d, abcdf^2q^{j-n}; q)_l}. \end{aligned} $$

As a consequence $L_{m,n}$ has an additional term given by

(4.9)
$$
\begin{aligned}
&- \sum_{l=0}^{n-N-1} v_l(a,b,c,dq^n,fq^{-n}) \\
&\quad \cdot \frac{(bc,bd,1/bf,a^2bcdf;q)_m (fq^{-n}/a,afq^{-n};q)_n}{(ac,ad,1/af,ab^2cdf;q)_m (ad,d/a;q)_n} \\
&\quad \cdot {}_{10}W_9(ab^2cdfq^{-1};abcf,abdf,bcdf,bfq^{l-n},bq^{n-l}/f,abcdq^{m-1},q^{-m};q,q) \\
&\quad \cdot {}_{10}W_9(aq^{-n}/d;q^{1-n}/bd,q^{1-n}/cd,q/df,abcf/q,afq^{l-n},aq^{n-l}/f,q^{-n};q,q).
\end{aligned}
$$

By (1.21),

$$
\begin{aligned}
&{}_{10}W_9(aq^{-n}/d;q^{1-n}/bd,q^{1-n}/cd,q/df,abcf/q,afq^{l-n},aq^{n-l}/f,q^{-n};q,q) \\
&\qquad = M_n(x_{n-l};a,b,c,d,fq^{-1}),
\end{aligned} \tag{4.10}
$$

where x_{n-l} is defined in (1.26).

Also, by (4.1),

$$
\begin{aligned}
&{}_{10}W_9(ab^2cdfq^{-1};abcf,abdf,bcdf,bfq^{l-n},bq^{n-l}/f,abcdq^{m-1},q^{-m};q,q) \\
&\qquad = \frac{(ac,ad,1/af,ab^2cdf;q)_m}{(bc,bd,1/bf,a^2bcdf;q)_m} R_m(x_{n-l};a,b,c,d,f).
\end{aligned} \tag{4.11}
$$

Substitution of (4.10) and (4.11) in (4.9) completes the proof of Theorem 2.

REFERENCES

[1] G. E. Andrews and R. Askey, *Another q-extension of the beta function*, Proc. Amer. Math. Soc., 81 (1981), pp. 97–100.

[2] R. Askey, *The q-gamma and q-beta functions*, Appl. Anal., 8 (1978), pp. 125–141.

[3] ———, *Ramanujan's extensions of the gamma and beta functions*, Amer. Math. Monthly, 87 (1980), pp. 346–359.

[4] ———, *A q-extension of Cauchy's form of the beta integral*, Quart. J. Math. Oxford Ser. (2), 32 (1981), pp. 255–266.

[5] ———, *Beta integrals in Ramanujan's papers, his unpublished work and further examples*, in Ramanujan Revisited, G. E. Andrews, eds., Academic Press, New York, 1988, pp. 561–590.

[6] ———, *Beta integrals and q-extensions*, Papers of the Ramanujan Centennial International Conference, Annamalainagar, December 15–18, 1987, pp. 85–102.

[7] ———, *Continuous q-Hermite polynomials when $q > 1$*, in Workshop on q-Series and Partitions D. Stanton, ed., IMA Volumes in Mathematics and its Applications, 18, Springer-Verlag, Berlin, New York, 1989.

[8] ———, *Beta integrals and the associated orthogonal polynomials*, in Number Theory, K. Alladi, ed., Lecture Notes in Math., 1395, Springer-Verlag, New York, 1989, pp. 84–121.

[9] R. Askey and R. Roy, *More q-beta integrals*, Rocky Mountain J. Math., 16 (1986), pp. 365–372.

[10] R. Askey and J. A. Wilson, *Some basic hypergeometric polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319 (1985).

[11] W. N. Bailey, *Series of hypergeometric type which are infinite in both directions*, Quart. J. Math. Oxford, 7 (1936), pp. 105–115.

[12] ———, *Well-poised basic hypergeometric series*, Quart. J. Math. Oxford, 18 (1947), pp. 157–166.

[13] G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, 1990.

[14] F. H. JACKSON, *Summation of q-hypergeometric series*, Messenger of Math., 50 (1921), pp. 101–112.

[15] B. NASSRALLAH AND M. RAHMAN, *Projection formulas, a reproducing kernel and a generating function for q-Wilson polynomials*, SIAM J. Math. Anal., 16 (1985), pp. 186–197.

[16] M. RAHMAN, *An integral representation of a* ${}_{10}\phi_9$ *and continuous biorthogonal* ${}_{10}\phi_9$ *rational functions*, Canad. J. Math., 38 (1986), pp. 605–618.

[17] ———, *Some extensions of the beta integral and the hypergeometric function*, Proc. NATO-ASI on Orthogonal Polynomials and Their Applications, Columbus, OH, 1989.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1442-1449, September 1991

015

ON ASYMPTOTICS OF JACOBI POLYNOMIALS*

LI-CHEN CHEN†‡ AND MOURAD E. H. ISMAIL†

This paper is dedicated to Frank Olver on the occasion of his 65th birthday.

Abstract. The asymptotic behavior of Jacobi polynomials $P_n^{\alpha+an,\beta+bn}(x)$ and Laguerre polynomials $L_n^{\alpha+an}(b+nx)$ is found when $n\to\infty$ and a, b, α, β are real constants, $a>-1$, $b>-1$, and x is a real variable.

Key words. Jacobi polynomials, Darboux's method, strong asymptotics

AMS(MOS) subject classifications. 33A65, 30E15

1. Introduction. The Jacobi polynomials $\{P_n^{\alpha,\beta}(x)\}$ are orthogonal on $[-1, 1]$ with respect to $(1-x)^\alpha(1+x)^\beta$ when $\alpha>-1$, $\beta>-1$. We shall refer to α and β as the parameters of the Jacobi polynomial. The asymptotics of the Jacobi polynomials, as $n\to\infty$ for fixed x, α, β follow from Darboux's asymptotic method and generating functions (Fields [7], Olver [12], Rainville [14], Szegö [17]). The polynomials oscillate when $x\in(-1, 1)$ and grow exponentially in the complex plane cut along $[-1, 1]$.

In this work we study the asymptotics of $P_n^{\alpha+an,\beta+bn}(x)$ and $L_n^{\alpha+an}(b+nx)$ as $n\to\infty$, and a, b, α, β, x remain fixed. The asymptotics of Jacobi polynomials lay the groundwork for a proposed study of the asymptotics of the Racah coefficients [1], [2], [4], [5], [13]. The connection between the two problems is that the Racah coefficients are integrals of products of Jacobi polynomials.

Askey and Wilson [3] introduced a q-analogue of the 6-j symbols and the Racah coefficients. Ismail and Wilson [9] derived the main term in the asymptotic expansion of these q analogues. Later Ismail [8] derived the complete asymptotic expansions of the q analogue of the 6-j symbols and the Racah coefficients.

The approach we used in this work is to apply the method of Darboux to the generating function

$$\sum_{n=0}^{\infty} P_n^{\alpha+an,\beta+bn}(x)t^n=(1+\xi)^{\alpha+1}(1+\eta)^{\beta+1}[1-a\xi-b\eta-(1+a+b)\xi\eta]^{-1}, \tag{1.1}$$

where ξ and η depend on x and t in the following fashion:

$$\xi=\tfrac{1}{2}(x+1)t(1+\xi)^{1+a}(1+\eta)^{1+b} \quad\text{and}\quad \eta=\tfrac{1}{2}(x-1)t(1+\xi)^{1+a}(1+\eta)^{1+b}. \tag{1.2}$$

The generating function (1.1)-(1.2) is due to Srivastava and Singhal [16].

The singularities of the right-hand side of (1.1) are at $\xi=-1$, $\eta=-1$, or when

$$1-a\xi-b\eta-(1+a+b)\xi\eta=0.$$

It is easy to see that, if $a>-1$ and $b>-1$, then the t-singularities of smallest absolute value make

$$1-a\xi-b\eta-(1+a+b)\xi\eta=0.$$

This occurs if and only if

$$(x+1)-[a(x+1)+b(x-1)]\xi+(1+a+b)(1-x)\xi^2=0, \tag{1.3}$$

* Received by the editors May 17, 1988; accepted for publication September 4, 1990. This research was partially supported by a grant from the National Science Foundation. The support of the Graduate College at Arizona State University in the form of a research assistantship is acknowledged. Part of this work was done at Arizona State University.

† Department of Mathematics, University of South Florida, Tampa, Florida 33620.

‡ Present address, Department of Business Mathematics, Soochow University, 56, Kuei-Yang St., Sec 1 Taipei, 10001, Taiwan, Republic of China.

since

$$\eta = (x-1)\xi/(x+1). \tag{1.4}$$

The roots of (1.3) are

$$\xi_{\pm} = \frac{b(x-1)+a(1+x)\pm\sqrt{\Delta}}{-2(1+a+b)(x-1)}, \tag{1.5}$$

where

$$\Delta = [a(x+1)+b(x-1)]^2 - 4(1+a+b)(1-x^2). \tag{1.6}$$

The corresponding η's will be denoted by $\eta_{\pm}$.

Section 2 contains some preliminary calculations that will be used to determine the strong asymptotics of the Jacobi polynomials under consideration. Our main results on the strong asymptotics of Jacobi polynomials are stated and proved in § 3. In § 4 we derive the corresponding asymptotic results involving Laguerre polynomials. This work concludes with § 5 where we describe the connection between our results and the results of Mhaskar and Saff [10], Moak, Saff, and Varga [11], and Saff and Varga [15] on nth root asymptotics of Jacobi and Laguerre polynomials. The work on nth root asymptotics was motivated by a theorem of Lorentz on approximation by incomplete polynomials.

2. Preliminaries. Throughout this section we will assume that a, b, and x are real. It follows from (1.2) and (1.5) that the t-roots of (1.3) are

$$t_{\pm} = \frac{b(x-1)+a(1+x)\pm\sqrt{\Delta}}{(a+b+1)(1-x^2)}[1+\xi_{\pm}]^{-a-1}[1+\eta_{\pm}]^{-b-1}. \tag{2.1}$$

We need to know how to expand ξ as a function of t near the singularities $t_{\pm}$. Differentiating the first relation in (1.2) and using (1.4), we find

$$\xi(1+\xi)(1+\eta)\frac{dt}{d\xi} = t[1-a\xi-b\eta-(a+b+1)\xi\eta], \tag{2.2}$$

which shows that $dt/d\xi$ vanishes at $\xi=\xi_{\pm}$. We then proceed to compute $d^2t/d\xi^2$. A calculation gives

$$\left.\frac{d^2t}{d\xi^2}\right|_{\xi_{\pm}} = \pm t_{\pm}\sqrt{\Delta}\,\{\xi_{\pm}(1+x)(1+\xi_{\pm})(1+\eta_{\pm})\}^{-1}.$$

We set

$$A_{\pm} = \pm\tfrac{1}{2}t_{\pm}\sqrt{\Delta}\,\{\xi_{\pm}(1+x)(1+\xi_{\pm})(1+\eta_{\pm})\}^{-1}. \tag{2.3}$$

Therefore as $t\to t_{\pm}$ and $\xi\to\xi_{\pm}$ we must have

$$t_{\pm}-t \approx -A_{\pm}(\xi_{\pm}-\xi)^2. \tag{2.4}$$

This shows that $\xi_{\pm}-\xi \approx \{(t_{\pm}-t)/(-A_{\pm})\}^{1/2}$ as $t\to t_{\pm}$, respectively. We then choose the comparison function

$$g(t) = B_+(t_+-t)^{-1/2} + B_-(t_--t)^{-1/2} \tag{2.5}$$

with constants B_+ and B_- to be determined later. Here $(t_+-t)^{1/2}$ is the branch of the square root which is continuous in the t plane cut along the outward ray through $t_+ = |t_+|\exp(i\tau_+)$ and which satisfies $(t_+-t)^{-1/2}\to|t_+|^{-1/2}\exp(-i\tau_+/2)$ as $t\to 0$. The branch of $(t_--t)^{-1/2}$ is similarly defined.

At this stage we have to consider three separate cases:

$$\text{Case 1: } \Delta<0, \quad \text{Case 2: } \Delta>0, \quad \text{Case 3: } \Delta=0. \tag{2.6}$$

Case 1 is the oscillatory case, while in Cases 2 and 3 the polynomials under consideration grow exponentially.

We first treat Case 1. It follows from (1.6) and $\Delta<0$ that

$$(1+a+b)(1-x^2)>0. \tag{2.7}$$

In this case both t_+ and t_- have the same absolute value. If $f(t)$ denotes the right-hand side of (1.1) then the limit of $(t_\pm - t)^{1/2} f(t)$ as $t \to t_\pm$ from the left is $B_\pm$. In order to compute the comparison function in (2.5) explicitly and derive the desired asymptotic formula we will need the following identities. In this case simple calculations give

$$|3a+b+2-(a+b+2)x \pm \sqrt{\Delta}|^2 = 8(1-x)(a+1)(a+b+1), \tag{2.8}$$

and

$$|-(a+b+2)x-(a+3b+2) \pm \sqrt{\Delta}|^2 = 8(1+x)(b+1)(a+b+1). \tag{2.9}$$

Therefore in Case 1

$$(1-x)(a+1)(a+b+1)>0 \quad \text{and} \quad (1+x)(b+1)(a+b+1)>0 \tag{2.10}$$

hold. We set

$$\frac{a(x+1)+b(x-1)+\sqrt{\Delta}}{(1+a+b)(1-x^2)} = 2\{(1+a+b)(1-x^2)\}^{-1/2} e^{i\rho}, \qquad -\pi<\rho\leqq\pi, \tag{2.11}$$

$$\frac{(a+b+2)x-(3a+b+2)-\sqrt{\Delta}}{2(x-1)(a+b+1)} = \left[\frac{2(a+1)}{(1-x)(a+b+1)}\right]^{1/2} e^{i\theta}, \qquad -\pi<\theta\leqq\pi, \tag{2.12}$$

$$\frac{(a+b+2)x+a+3b+2-\sqrt{\Delta}}{2(x+1)(a+b+1)} = \left[\frac{2(b+1)}{(x+1)(a+b+1)}\right]^{1/2} e^{i\gamma}, \qquad -\pi<\gamma\leqq\pi. \tag{2.13}$$

We then use (1.1), (1.2), (2.1), and (2.3)-(2.5) to establish

$$\begin{aligned} B_+ &= \lim_{t\to t_+} (1+\xi)^{\alpha+1}(1+\eta)^{\beta+1}[1-a\xi-b\eta-(a+b+1)\xi\eta]^{-1}(t_+-t)^{1/2} \\ &= -i(\Delta)^{-1/4}\left[\frac{(a+b+2)x-(3a+b+2)-\sqrt{\Delta}}{2(a+b+1)(x-1)}\right]^{\alpha-a/2} \\ &\quad \cdot\left[\frac{(a+b+2)x+a+3b+2-\sqrt{\Delta}}{2(a+b+1)(x+1)}\right]^{\beta-b/2}. \end{aligned} \tag{2.14}$$

Similarly,

$$\begin{aligned} B_- &= i(-\sqrt{\Delta})^{-1/2}\left[\frac{(a+b+2)x-(3a+b+2)+\sqrt{\Delta}}{2(a+b+1)(x-1)}\right]^{\alpha-a/2} \\ &\quad \cdot\left[\frac{(a+b+2)x+a+3b+2+\sqrt{\Delta}}{2(a+b+1)(x+1)}\right]^{\beta-b/2}. \end{aligned} \tag{2.15}$$

The coefficient of t^n in the comparison function $g(t)$ of (2.5) is

$$(-1)^n \binom{-1/2}{n} [B_+(t_+)^{-n-1/2} + B_-(t_-)^{-n-1/2}],$$

which can be simplified using (2.11)-(2.13), and

$$(-1)^n\binom{-1/2}{n}=\frac{\Gamma(n+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(n+1)}\approx\frac{n^{-1/2}}{\sqrt{\pi}},$$

to establish

$$\begin{aligned}P_n^{\alpha+an,\beta+bn}(x)\approx&\left(\frac{4\sqrt{(-\Delta)}}{\pi n}\right)^{1/2}\left[\frac{(1-x)(a+b+1)}{2(a+1)}\right]^{n(-a-1)/2-\alpha/2-1/4}\\&\cdot\left[\frac{(1+x)(a+b+1)}{2(b+1)}\right]^{n(-b-1)/2-\beta/2-1/4}\left[\frac{(1-x^2)(a+b+1)}{4}\right]^{n/2+1/4}\\&\cdot\cos\left[\left\{(1+a)n+\alpha+\frac{1}{2}\right\}\frac{\theta}{2}+\left\{-(b+1)n+\beta+\frac{1}{2}\right\}\frac{\gamma}{2}-(2n+1)\frac{\rho}{4}-\frac{\pi}{4}\right].\end{aligned}\tag{2.16}$$

We now treat Case 2. It is obvious that $\xi_+\neq\xi_-$ when $\Delta>0$. From (2.2) we see that $dt/d\xi$ does not vanish between ξ_+ and ξ_-, hence t takes different values at ξ_+, ξ_-. Set

$$\begin{aligned}&t_0=t_+\quad\text{if } |t_-|>|t_+|,\quad t_0=t_-\quad\text{if } |t_+|>|t_-|,\\&t_1=t_+\quad\text{if } |t_-|<|t_+|,\quad t_0=t_-\quad\text{if } |t_+|<|t_-|.\end{aligned}\tag{2.17}$$

We also denote the B's that correspond to t_0 and t_1 by B_0 and B_1, respectively. The function $g(t)$ of (2.5) remains a comparison function, but the term that contributes to the dominant term in the asymptotic development of the polynomials under consideration is $B_0(t_0-t)^{-1/2}$.

In Case 3, $\Delta=0$, hence $\xi_+=\xi_-=\xi_0$ and $t_+=t_-$. Set $t_0=t_+=t_-$ and $\xi_0=\xi_+=\xi_-$. In this case both $dt/d\xi$ and $d^2t/d\xi^2$ vanish but $d^3t/d\xi^3$ does not vanish. Indeed upon differentiating (2.2) twice with respect to ξ we get

$$\left.\frac{d^3t}{d\xi^3}\right|_{\xi_0}=\frac{2t_0(a+b+1)}{\xi_0(1+\eta_0)(1+\xi_0)}\cdot\frac{1-x}{x+1}=:c.\tag{2.18}$$

Therefore $\xi_0-\xi\approx\sqrt[3]{6/c}\,(t_0-t)^{1/3}$ as $t\to t_0$. If, as before, we denote the right-hand side of (1.1) by $f(t)$, then we replace the comparison function (2.5) by

$$g(t)=D(t_0-t)^{-2/3},\tag{2.19}$$

where

$$D=\lim_{t\to t_0}\frac{(1+\xi)^{1+\alpha}(1+\eta)^{1+\beta}(t_0-t)^{2/3}}{1-a\xi-b\eta-(a+b+1)\xi\eta}.\tag{2.20}$$

We will evaluate D, using (2.19) and (2.20), and find the asymptotic behavior of the corresponding P_n's in § 3.

3. Asymptotics of Jacobi polynomials. Recall that x is assumed to be real and a and b belong to $(-1,\infty)$. We first state and prove a lemma that characterizes the case $\Delta<0$.

LEMMA 3.1. *In order for* $\Delta<0$ *to hold it is necessary and sufficient that*

(i) *Formulas* (2.7) *and* (2.10) *hold, and so* $a+b+1>0$.

(ii) $(a+b+2)^2x\in(b^2-a^2-4[(a+1)(1+b)(a+b+1)]^{1/2},$
$b^2-a^2+4[(a+1)(1+b)(a+b+1)]^{1/2})$, *when* $a+b\neq-2$.

Proof. Observe that

$$\Delta=(a+b+2)^2x^2-2(b^2-a^2)x+[(b-a)^2-4(a+b+1)]\tag{3.1}$$

follows from (1.6). In § 2 we already demonstrated the necessity of (2.7) and (2.10). It is easy to see that (2.7) and (2.10) imply

$$a+b+1>0. \tag{3.2}$$

Then $\Delta<0$ implies the restrictions on x stated in (ii). This proves the necessity of the conditions in this lemma. We can prove the sufficiency of the assumptions by first establishing (3.2), and then analyze Δ as a quadratic function of x. We will omit the details.

Observe that if $a+b=-1$ the x-interval in (ii) becomes empty.

LEMMA 3.2. *The roots $\xi_{\pm}$ of* (1.3) *are distinct and have equal absolute values if and only if $a+b\neq -1$ and $\Delta<0$.*

Proof. This follows from the well-known fact that in the complex z plane the quantities $z\pm(z^2-1)^{1/2}$ have equal absolute values if and only if z is real and $-1\leqq z\leqq 1$.

THEOREM 3.3. *Assume $a>-1$, $b>-1$, and x satisfies condition* (ii) *of Lemma* 3.1; *then the asymptotic relationship* (2.16) *holds, as $n\to\infty$, where ρ, γ and θ are as in* (2.11)-(2.13).

THEOREM 3.4. *Assume that $a>-1$, $b>-1$, and $\Delta>0$. Then*

$$P_n^{\alpha+an,\beta+bn}(x)\approx\frac{B_0}{\sqrt{n\pi}}\,t_0^{-n-1/2}. \tag{3.3}$$

Proof. The coefficient of t^n in the power series expansion of $(t_0-t)^{-1/2}$ is

$$\frac{\Gamma(n+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(n+1)}\,t_0^{-n-1/2},$$

which is asymptotic to $(n\pi)^{-1/2}t_0^{-n-1/2}$. This proves (3.3).

Finally, we discuss the case $\Delta=0$. This case can be thought of as a transition between Case 1 and Case 2. In Case 3 both $dt/d\xi$ and $d^2t/d\xi^2$ vanish but $d^3t/d\xi^3$ does not vanish. This case occurs if and only if

$$x^2(a+b+2)^2+2x(a^2-b^2)+(a-b)^2-4(a+b+1)=0 \tag{3.4}$$

holds (see (3.1)). There is no choice for a and b that will make the left-hand side of (3.4) identically zero. Thus $\Delta=0$ has at most two solutions. They are

$$x=\frac{b^2-a^2\pm 4\sqrt{(1+a)(1+b)(1+a+b)}}{(2+a+b)^2}\quad\text{if } a+b>-1, \tag{3.5}$$

and at $x=2a+1$ if $a+b=-1$. A calculation shows that the constant D of (2.19) and (2.20) is given by

$$D=9^{-1/3}(1+\xi)^{\alpha+1/3}(1+\eta)^{\beta+1/3}, \tag{3.6}$$

where

$$1+\xi=-a^{-1}\left[-a-1\pm\sqrt{\frac{(1+a)(1+b)}{1+a+b}}\right], \tag{3.7}$$

$$1+\eta=-b^{-1}\left[-b-1\pm\sqrt{\frac{(1+a)(1+b)}{1+a+b}}\right]. \tag{3.8}$$

The calculation above is particularly simplified by the observation

$$b^2\frac{x-1}{x+1}=-2(a+b+1)-ab\pm 2\sqrt{(1+a)(1+b)(1+a+b)}.$$

Note that the difference between $1+\xi$ and $1+\eta$ is that a and b are interchanged and the sign of the square root is reversed. The coefficient of t^n in $g(t)$ is $(t_0)^{-n-2/3}D(2/3)_n/n!$. Therefore the P_n's behave asymptotically like the aforementioned coefficient of t^n. Thus we established the following asymptotic result:

$$P_n^{\alpha+an,\beta+bn}(x) \approx \frac{[(1+x)/(2\xi)]^{n+2/3}}{\sqrt[3]{9n}\,\Gamma(\frac{2}{3})}(1+\xi)^{n(-a-1)+\alpha+1+2a/3}(1+\eta)^{n(-b-1)+\beta+1+2b/3}. \tag{3.9}$$

4. Laguerre polynomials. In this section we briefly analyze the strong asymptotics of the Laguerre polynomials $L_n^{\alpha+an}(b+nx)$ as $n\to\infty$ for real b, α, β, x, and positive a. We will consider only the oscillatory range.

The generating function

$$\sum_{n=0}^{\infty} L_n^{\alpha+an}(b+nx)t^n = (1-\xi)^{1-\alpha}[(1-\xi)^2 - a\xi(1-\xi)+x\xi]^{-1}\exp[-b\xi/(1-\xi)], \tag{4.1}$$

where

$$t = \xi(1-\xi)^a \exp[x\xi/(1-\xi)] \tag{4.2}$$

was proved by Carlitz [6] using Lagrange inversion. It is a limiting case of the Srivastava-Singhal generating function (1.1)-(1.2), which was also proved by Lagrange inversion.

The ξ-singularities of (4.1) are at $\xi=1$, $\xi=\xi_\pm$, where $\xi_\pm$ are the roots of

$$(1-\xi)^2 - a\xi(1-\xi)+x\xi = 0. \tag{4.3}$$

Therefore

$$\xi_\pm = \frac{2+a-x\pm\sqrt{(a-x)^2-4x}}{2(1+a)}. \tag{4.4}$$

In the oscillatory case we must have

$$x-2\sqrt{x} < a < x+2\sqrt{x}. \tag{4.5}$$

It follows that $|\xi_\pm| = 1/(1+a) < 1$, since $a>0$. Thus $\xi=\xi_\pm$ are the singularities of the generating function that are closest to the origin. Let

$$t_\pm = \xi_\pm(1-\xi_\pm)^a \exp[x\xi_\pm/(1-\xi_\pm)]. \tag{4.6}$$

A calculation shows that

$$\frac{dt}{d\xi} = 0, \qquad \frac{d^2t}{d^2\xi} = (1-\xi)^a(1-\xi)^{-3}[-a(1-\xi)+x(1+\xi)]\exp\{x[1-(1-\xi)^{-1}]\} \quad \text{at } \xi=\xi_\pm. \tag{4.7}$$

We can prove

$$L_n^{\alpha+an}(b+nx) \approx \sqrt{\frac{-2}{n\pi\Delta}}\left(\frac{-x\Delta}{1+a}\right)^{1/4}\left(\frac{\sqrt{1+a}}{x}\right)^{an+a+1/2}(1+a)^{n/2+1/4} \cdot \exp\left\{\left(\frac{2b-x}{2\sqrt{x}}\right)\cos\phi\right\}\cos\left\{\left[-\left(\alpha+\frac{1}{2}\right)+\frac{a}{2}\right]\theta-\left(n+\frac{1}{2}\right)(\omega+a\theta)+\frac{\gamma}{2} + \exp\left\{\left[\frac{2b-x}{2\sqrt{x}}\right]\sin\phi\right\}\right\}, \tag{4.8}$$

where

$$\Delta=(a-x)^2-4x, \qquad 1-(1-\xi_{\pm})^{-1}=(x)^{-1/2}\,e^{\pm i\phi},$$
$$-a(1-\xi_{\pm})+x(1+\xi_{\pm})=(-x\Delta/(1+a))^{1/4}\,e^{\pm i\gamma}, \qquad \omega=\arg(\xi_+). \tag{4.9}$$

5. Remarks. In their interesting work on the sharpness of the Lorentz theorem on incomplete polynomials, Saff and Varga [15] were led to two problems involving asymptotics and distribution of zeros of the Jacobi polynomials $\{P_n^{\alpha_n,\beta_n}(x)\}$ as $n\to\infty$. Let $\alpha_n>-1$ and $\beta_n>-1$. The second problem was solved in [11], where Moak, Saff, and Varga proved the following result.

THEOREM 5.1. *Let s_n and l_n be, respectively, the smallest and largest zeros of the Jacobi polynomial $P_n^{\alpha_n,\beta_n}(x)$ and assume that the limits*

$$\lim_{n\to\infty}\frac{\alpha_n}{2n+\alpha_n+\beta_n}=A \quad \textit{and} \quad \lim_{n\to\infty}\frac{\beta_n}{2n+\alpha_n+\beta_n}=B$$

exist. Then as $n\to\infty$, $s_n\to S$, and $l_n\to L$, where

$$S=B^2-A^2-[(A^2+B^2-1)^2-4A^2B^2]^{1/2}$$

and

$$L=B^2-A^2+[(A^2+B^2-1)^2-4A^2B^2]^{1/2}.$$

Furthermore, the zeros of the sequence $\{P_n^{\alpha_n,\beta_n}(x)\}$ are dense in $[S, L]$.

The special case $\alpha_n=\alpha+an$, $\beta_n=\beta+bn$, $a>-1$, $b>-1$, of Theorem 5.1 follows from (2.16). Indeed for $\alpha>-1$ and $\beta>-1$ the zeros of $P_n^{\alpha,\beta}(x)$ belong to $(-1, 1)$. The zeros increase with β and decrease with α. Thus $\{s_n\}$ and $\{l_n\}$ are monotone sequences. The oscillatory nature of the asymptotic formula (2.16) shows that the zeros of the Jacobi polynomials $\{P_n^{\alpha+an,\beta+bn}(x)\}$ are dense in the interval characterized by $\Delta<0$. The interval $\Delta<0$ is precisely the interval (S, L).

The first problem of Saff and Varga [15] involving Jacobi polynomials was to determine the asymptotic behavior of $P_n^{\alpha,\beta+2\theta n/(1-\theta)}(x)$, as $n\to\infty$, for real x, $x\neq-1$. Saff and Varga used the method of steepest descent to prove their results. The asymptotic estimates (2.16) and (3.3) generalize the Saff-Varga estimates. Saff and Varga used their results to prove that a theorem of Lorentz on incomplete polynomials is sharp.

Mhaskar and Saff [10] developed nth root asymptotics for $\{L_n^{\alpha+an}(b+nx)\}$ as a concrete example to illustrate their general results and demonstrate their sharpness. Formula (4.8) gives the main term in the strong (pointwise) asymptotics of the aforementioned Laguerre polynomials and implies the nth root asymptotics. Furthermore, (4.8) shows that the zeros of the Laguerre polynomials under consideration are dense in the interval $(2+a-2\sqrt{1+a},\ 2+a+2\sqrt{1+a})$.

If a sequence of polynomials $\{p_n(x)\}$ is orthogonal with respect to a positive measure $d\mu$ and μ' is supported on a set E, then the zeros of the p_n's are dense in E. The zeros of the polynomial sequences $\{P_n^{\alpha+an,\beta+bn}(x)\}$ and $\{L_n^{\alpha+an}(b+nx)\}$ are dense in certain intervals but neither set is orthogonal unless $a=b=0$, and $a=x=0$, respectively.

Acknowledgment. We thank Ed Saff for the information about [10] and for suggesting that we do the strong asymptotics of § 4.

REFERENCES

[1] G. E. ANDREWS AND R. A. ASKEY, *Classical orthogonal polynomials*, in Polynômes Orthogonaux et Applications, C. Brezinski, A. Draux, A. P. Magnus, P. Maroni, and A. Ronveaux, eds., Lecture Notes in Math. 1171, Springer-Verlag, Berlin, 1985, pp. 36–62.

[2] R. A. ASKEY AND J. A. WILSON, *A set of orthogonal polynomials that generalize the Racah coefficients or 6-j symbols*, SIAM J. Math. Anal., 10 (1979), pp. 1008–1016.

[3] ———, *Some basic hypergeometric polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319 (1985), pp. 1–54.

[4] L. C. BIEDENHARN AND J. D. LOUCK, *Angular Momentum in Quantum Physics: Theory and Application*, Cambridge University Press, Cambridge, 1984.

[5] ———, *The Racah–Wigner Algebra in Quantum Theory*, Cambridge University Press, Cambridge, 1984.

[6] L. CARLITZ, *A class of generating functions*, SIAM J. Math. Anal., 8 (1977), pp. 518–532.

[7] J. L. FIELDS, *A uniform treatment of Darboux's method*, Arch. Rational Mech. Anal., 27 (1968), pp. 289–305.

[8] M. E. H. ISMAIL, *Asymptotics of the Askey–Wilson and q-Jacobi polynomials*, SIAM J. Math. Anal., 17 (1986), pp. 1,475–1,482.

[9] M. E. H. ISMAIL AND J. A. WILSON, *Asymptotic and generating relations for the q-Jacobi and* ${}_4\phi_3$ *polynomials*, J. Approx. Theory Appl., 36 (1982), pp. 43–54.

[10] H. N. MHASKAR AND E. B. SAFF, *Polynomials with Laguerre weights in* L^p, in Rational Approximation and Interpolation, P. R. Graves-Morris and E. B. Saff, eds., Lecture Notes in Math. 1105, Springer-Verlag, Berlin, 1984, pp. 511–523.

[11] D. S. MOAK, E. B. SAFF, AND R. S. VARGA, *On the zeros of Jacobi polynomials* $P_n^{(\alpha_n,\beta_n)}(x)$, Trans. Amer. Math. Soc., 249 (1979), pp. 159–163.

[12] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[13] G. PONZONO AND T. REGGE, *Semiclassical limit of Racah coefficients*, in Spectroscopic and Group Theoretical Methods in Physics, North-Holland, Amsterdam, 1968, pp. 1–58.

[14] E. D. RAINVILLE, *Special Functions*, Macmillan, New York, 1960.

[15] E. B. SAFF AND R. S. VARGA, *The sharpness of Lorentz's theorem on incomplete polynomials*, Trans. Amer. Math. Soc., 249 (1979), pp. 163–186.

[16] H. M. SRIVASTAVA AND J. P. SINGHAL, *New generating functions for Jacobi and related polynomials*, J. Math. Anal. Appl., 41 (1972), pp. 748–752.

[17] G. SZEGÖ, *Orthogonal Polynomials*, Colloquium Publications 23, Fourth ed., American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1450–1459, September 1991

016

HYPERGEOMETRIC EXPANSIONS OF HEUN POLYNOMIALS*

E.G. KALNINS† AND W. MILLER, JR.‡

Abstract. The product of two Heun polynomials is expanded in terms of products of two Jacobi polynomials. This is done by making crucial use of group theory and the knowledge of separable coordinate systems on the n-sphere. The expansion presented includes as a special case hypergeometric function expansions of single Heun polynomials that have been derived previously by other methods.

Key words. multivariable orthogonal polynomials, the n-sphere, Heun functions

AMS(MOS) subject classifications. 22E70, 33A65, 33A75

1. Introduction.

Any Fuchsian equation of second order with four singularities can be reduced to the form

$$\frac{d^2w}{dx^2}+\left[\frac{\gamma}{x-e_1}+\frac{\delta}{x-e_2}+\frac{\epsilon}{x-e_3}\right]\frac{dw}{dx}+\frac{\alpha\beta x-q}{(x-e_1)(x-e_2)(x-e_3)}w=0, \tag{1.1}$$

where $\alpha+\beta-\gamma-\delta-\epsilon+1=0$.

The singularities are located at $x=e_1,e_2,e_3$, and ∞ and have indices depending upon $\alpha,\cdots,\epsilon$. The constant q is known as the accessory parameter. This is Heun's equation [1] and solutions may be characterised by the P symbol [2].

$$P\left\{\begin{array}{cccccc} e_1 & e_2 & e_3 & \infty & \\ 0 & 0 & 0 & \alpha & x \\ 1-\gamma & 1-\delta & 1-\epsilon & \beta & \end{array}\right\}. \tag{1.2}$$

Power series expansions for the solutions of Heun's equation have been studied by Heun for various arguments [1], [3]. There turn out to be 96 distinct types of power series. Alternatively, solutions of Heun's equations can be expanded in series of hypergeometric functions. Such expansions were studied by Svartholm [4] and Erdélyi [5]. Typically such expansions have the form

$$P\left\{\begin{array}{cccccc} e_1 & e_2 & e_3 & \infty & \\ 0 & 0 & 0 & \alpha & x \\ 1-\gamma & 1-\delta & 1-\epsilon & \beta & \end{array}\right\}=\sum_{m=0}^{\infty}A_mP\left\{\begin{array}{ccccc} 0 & 1 & \alpha & \\ 0 & 0 & \lambda+m & x \\ 1-\gamma & 1-\delta & \mu-m & \end{array}\right\}, \tag{1.3}$$

where $\lambda+\mu=\gamma+\delta-1=\alpha+\beta-\epsilon$. Two types of expansion were given:

(i) Series of type I for which $\lambda=\alpha,\mu=\beta-\epsilon$. These series converge outside an ellipse with foci at e_1,e_2 and which passes through e_3. There are three distinct expansions of this type.

(ii) Series of type II for which $\mu=0$, $\gamma-1$, $\delta-1$, or $\gamma+\delta-2$.

In all these expansions the coefficients A_m satisfy three-term recurrence relations

$$b_0A_0+c_1A_1=0, \tag{1.4}$$

$$a_rA_{r-1}+b_rA_r+c_{r+1}A_{r+1}=0, \qquad r=1,2,\cdots,$$

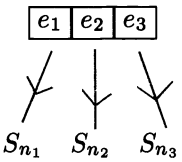
* Received by the editors June 6, 1990; accepted for publication October 3, 1990. This research was supported in part by the National Science Foundation under grant DMS 88-23054.

†Department of Mathematics and Statistics, University of Waikato, Hamilton, New Zealand.

‡School of Mathematics, 127 Vincent Hall, University of Minnesota, Minneapolis, Minnesota 55455.

where a_r, b_r, c_r are known expressions in r and $c_r \neq 0$. If q is chosen from a number of characteristic values then expansions of this type converge. In this article we derive some of these expansions for the case of Heun polynomials from considerations based on group theory and its connection with separation of variables solutions of the Laplace–Beltrami eigenvalue equation on the n-sphere. The method used makes a judicious choice of coordinates on the n-sphere. The expansions that are first derived are for products of Heun polynomials as sums of products of Jacobi polynomials. The coefficients in the expansions obey three-term recurrence relations. The corresponding single variable expansions are then obtained by allowing one of the variables to take a fixed value. This paper is an extension of [8], in which its motivation and background can be found.

2. Derivation of the expansion formula. The graphical calculus of separable coordinates for the Laplace–Beltrami eigenvalue equation on the n-sphere has been completely worked out by Kalnins and Miller [6], [7]. To derive an expansion for Heun polynomials we consider coordinate systems corresponding to graphs of the type



on the n sphere, $n = n_1 + n_2 + n_3 + 2$. A suitable choice of coordinates is

$$
\begin{aligned}
s_i &= u_1 w_i, && i = 1, \cdots, n_1 + 1, \\
s_{j+n_1+1} &= u_2 t_j, && j = 1, \cdots, n_2 + 1, \\
s_{k+n_1+n_2+2} &= u_3 z_k, && k = 1, \cdots, n_3 + 1,
\end{aligned} \tag{2.1}
$$

where

$$
\sum_{i=1}^{n_1+1} w_i^2 = 1, \quad \sum_{j=1}^{n_2+1} t_j^2 = 1, \quad \sum_{k=1}^{n_3+1} z_k^2 = 1,
$$

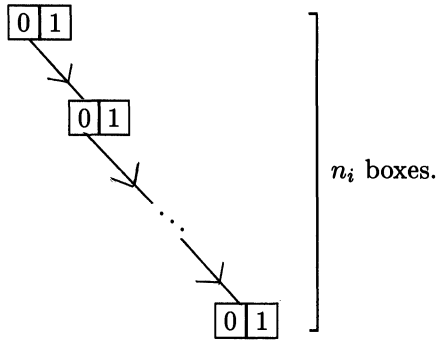
and

$$
u_i^2 = \frac{(x-e_i)(y-e_i)}{(e_j-e_i)(e_k-e_i)}, \qquad i = 1,2,3, \quad i,j,k \text{ pairwise distinct.} \tag{2.2}
$$

The metric on the n sphere is

$$
\begin{aligned}
ds^2 = &-\frac{(x-y)}{4}\left[\frac{dx^2}{(x-e_1)(x-e_2)(x-e_3)} - \frac{dy^2}{(y-e_1)(y-e_2)(y-e_3)}\right] \\
&+ \frac{(x-e_1)(y-e_1)}{(e_2-e_1)(e_3-e_1)} \sum_{i=1}^{n_1+1} dw_i^2 + \frac{(x-e_2)(y-e_2)}{(e_3-e_2)(e_1-e_2)} \sum_{j=1}^{n_2+1} dt_j^2 \\
&+ \frac{(x-e_3)(y-e_3)}{(e_2-e_3)(e_1-e_3)} \sum_{k=1}^{n_3+1} dz_k^2.
\end{aligned} \tag{2.3}
$$

The coordinate systems chosen for w_i, t_j, z_k can be taken to be, say, spherical coordinates in each case, corresponding to the graph [6].



We then seek eigenfunctions ψ of the Laplacian satisfying

$$\Delta\psi = -J(J+n_1+n_2+n_3+1)\psi, \tag{2.4}$$

where J is a nonnegative integer. In the coordinates we have chosen, this equation has the form

(2.5)

$$\begin{aligned}\Delta\psi = -\frac{4}{(x-y)}\Bigg[&(x-e_1)(x-e_2)(x-e_3)\\ &\cdot\left[\frac{\partial^2}{\partial x^2}+\frac{1}{2}\left[\frac{n_1+1}{x-e_1}+\frac{n_2+1}{x-e_2}+\frac{n_3+1}{x-e_3}\right]\frac{\partial}{\partial x}\right]\psi\\ &-(y-e_1)(y-e_2)(y-e_3)\left[\frac{\partial^2}{\partial y^2}+\frac{1}{2}\left[\frac{n_1+1}{y-e_1}+\frac{n_2+1}{y-e_2}+\frac{n_3+1}{y-e_3}\right]\frac{\partial}{\partial y}\right]\psi\Bigg]\\ &+\left[\frac{(e_1-e_2)(e_1-e_3)}{(x-e_1)(y-e_1)}\Delta_1+\frac{(e_2-e_1)(e_2-e_3)}{(x-e_2)(y-e_2)}\Delta_2+\frac{(e_3-e_1)(e_3-e_2)}{(x-e_3)(y-e_3)}\Delta_3\right]\psi\\ &=-J(J+n_1+n_2+n_3+1)\psi,\end{aligned}$$

where Δ_k is the Laplacian on the sphere S_{n_k}.

If we seek eigenfunctions such that

$$\Delta_i\psi = -\ell_i(\ell_i+n_i-1)\psi, \qquad i=1,2,3, \tag{2.6}$$

where the ℓ_i are nonnegative integers, then writing

$$\psi = \prod_{i=1}^{3}[(x-e_i)(y-e_i)]^{\ell_i/2}\phi, \tag{2.7}$$

we find (2.5) has the form

(2.8)

$$\begin{aligned}-\frac{4}{(x-y)}\Bigg\{&(x-e_1)(x-e_2)(x-e_3)\left(\frac{\partial^2}{\partial x^2}+\left[\frac{\ell_1+\frac{1}{2}(n_1+1)}{x-e_1}+\frac{\ell_2+\frac{1}{2}(n_2+1)}{x-e_2}\right.\right.\\ &\left.\left.+\frac{\ell_3+\frac{1}{2}(n_3+1)}{x-e_3}\right]\frac{\partial}{\partial x}\right)\phi+Ax\phi\\ &-(y-e_1)(y-e_2)(y-e_3)\left(\frac{\partial^2}{\partial y^2}+\left[\frac{\ell_1+\frac{1}{2}(n_1+1)}{y-e_1}+\frac{\ell_2+\frac{1}{2}(n_2+1)}{y-e_2}\right.\right.\\ &\left.\left.+\frac{\ell_3+\frac{1}{2}(n_3+1)}{y-e_3}\right]\frac{\partial}{\partial y}\right)\phi-Ay\phi\Bigg\}=-J(J+N+1)\phi,\end{aligned} \tag{2.8}$$

where

$$A = \tfrac{1}{4}(L+N+1)L, \quad L = \ell_1+\ell_2+\ell_3, \quad N = n_1+n_2+n_3.$$

The corresponding separable solutions have the form

$$\psi = u_1^{\ell_1}u_2^{\ell_2}u_3^{\ell_3}\Phi^1_{J\ell_1\ell_2\ell_3 q}(x)\Phi^2_{J\ell_1\ell_2\ell_3 q}(y)\Theta_{\ell_1\ell_2\ell_3}(\mathbf{w},\mathbf{t},\mathbf{z}), \tag{2.9}$$

where a complete set of functions $\Theta_{\ell_1\ell_2\ell_3}(\mathbf{w},\mathbf{t},\mathbf{z})$ can be taken as

$$\Theta_{\ell_1\ell_2\ell_3}(\mathbf{w},\mathbf{t},\mathbf{z}) = \Theta_{\ell_1}(\mathbf{w})\Theta_{\ell_2}(\mathbf{t})\Theta_{\ell_3}(\mathbf{z}) \tag{2.10}$$

and typically,

$$\Theta_{\ell_1}(\mathbf{w}) = \prod_{j=0}^{n_1-2} C^{\frac{1}{2}(n_1-j-1)+K_{j+1}}_{K_j-K_{j+1}}(\cos(\theta_{n_1-j}))(\sin\theta_{n_1-j})^{K_j+1}e^{\pm iK_{n_1-1}\theta_1} \tag{2.11}$$

for $\ell_1 = K_0 \geq K_1 \geq \cdots \geq K_{n_1-1} \geq 0$, and

$$\Delta_{(k)}\Theta_{\ell_1}(\mathbf{w}) = -K_k(K_k+n_1-k-1)\Theta_{\ell_1}(\mathbf{w}), \tag{2.12}$$

where $C^{\nu}_{\mu}(z)$ is a Gegenbauer polynomial. The coordinates on S_{n_1} are

$$\begin{aligned} w_1 &= \sin\theta_{n_1}\cdots\sin\theta_2\sin\theta_1 \\ w_2 &= \sin\theta_{n_1}\cdots\sin\theta_2\cos\theta_1 \\ &\vdots \\ w_{n_1} &= \sin\theta_{n_1}\cos\theta_{n_1-1} \\ w_{n_1+1} &= \cos\theta_{n_1} \end{aligned} \tag{2.13}$$

and the operator $\Delta_{(k)}$ is given by

$$\Delta_{(k)} = \sum_{r<\ell\leq n_1+1-k} I^2_{r\ell}, \quad I_{r\ell} = w_r\frac{\partial}{\partial w_\ell} - w_\ell\frac{\partial}{\partial w_r}, \quad k = 0,\cdots,n_1-1. \tag{2.14}$$

(The $\Delta_{(k)}$ are the second-order symmetry operators for Δ_1 whose eigenvalue equations (2.12) characterize the separable coordinates (2.13); see [6], [7].) The corresponding separation equations are

$$\begin{aligned} &\Big[-4(\lambda-e_1)(\lambda-e_2)(\lambda-e_3) \\ &\cdot\Big[\frac{d^2}{d\lambda^2}+\Big(\frac{\ell_1+\frac{1}{2}(n_1+1)}{\lambda-e_1}+\frac{\ell_2+\frac{1}{2}(n_2+1)}{\lambda-e_2}+\frac{\ell_3+\frac{1}{2}(n_3+1)}{\lambda-e_3}\Big)\frac{d}{d\lambda}\Big] \\ &\quad+(J-L)(J+L+N+1)\lambda+4q\Big]\Phi^{\epsilon}_{J\ell_1\ell_2\ell_3 q}(\lambda) = 0, \end{aligned} \tag{2.15}$$

where $\lambda = x, y$ according as $\epsilon = 1, 2$, respectively. This is Heun's equation of the form (1.1) with $\gamma = \ell_1+\frac{1}{2}(n_1+1)$, $\delta = \ell_2+\frac{1}{2}(n_2+1)$, $\epsilon = \ell_3+\frac{1}{2}(n_3+1)$, $\alpha = \frac{1}{2}(L-J)$, $\beta = \frac{1}{2}(L+J+N+1)$. The solutions for the functions $\Phi^{\epsilon}_{J\ell_1\ell_2\ell_3 q}(\lambda)$ are Heun polynomials which for fixed J will form a complete set of basis functions once the eigenvalues q have been calculated. To calculate the eigenvalues it is convenient to observe that in the coordinate system (2.1) the operator $\mathscr{M}$ whose eigenvalue χ is

$$\begin{aligned} \chi &= (e_1+e_2+e_3)[\ell_1^2+\ell_2^2+\ell_3^2+\ell_1n_1+\ell_2n_2+\ell_3n_3-J(J+N+1)] \\ &\quad+2\ell_1\ell_2e_3+2\ell_1\ell_3e_2+2\ell_2\ell_3e_1-\ell_1e_1-\ell_2e_2-\ell_3e_3 \\ &\quad+\ell_1n_2e_3+\ell_1n_3e_2+\ell_2n_1e_3+\ell_2n_3e_1+\ell_3n_1e_2+\ell_3n_2e_1-4q \end{aligned} \tag{2.16}$$

is given by [6], [7]

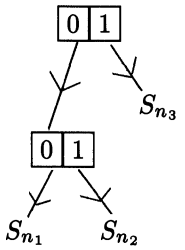
$$\mathscr{M} = (e_1+e_2)\sum_{p\in P}\sum_{q\in Q} I_{pq}^2 + (e_2+e_3)\sum_{q\in Q}\sum_{r\in R} I_{rq}^2 + (e_1+e_3)\sum_{p\in P}\sum_{r\in R} I_{pr}^2, \tag{2.17}$$

$$P = \{1,\cdots,n_1+1\}, Q = \{n_1+2,\cdots,n_1+n_2+2\},$$

$$R = \{n_1+n_2+3,\cdots,n_1+n_2+n_3+3\}.$$

That is, $\mathscr{M}$ is the second-order symmetry operator for the Laplacian ($[\mathscr{M},\Delta]=0$), which corresponds to the separable coordinates x, y. (The separable solutions (2.9) are eigenfunctions of $\mathscr{M}$ with eigenvalues χ.) Expression (2.16) gives the relationship between the eigenvalues χ and q. (The terms involving the ℓ_j result from consideration of the factor $u_1^{\ell_1}u_2^{\ell_2}u_3^{\ell_3}$.)

The basis functions on the sphere S_n corresponding to coordinates of the graph can also be expanded in terms of the basis functions of the coordinate system corresponding to the graph [6],



i.e., the coordinates (2.1) with

$$u_1 = \sin\theta\cos\phi, \quad u_2 = \sin\theta\sin\phi, \quad u_3 = \cos\theta \tag{2.18}$$

and the infinitesimal distance

$$ds^2 = d\theta^2 + \sin^2\theta d\phi^2 + \sin^2\theta\cos^2\phi\sum_{i=1}^{n_1+1} dw_i^2 + \sin^2\theta\sin^2\phi\sum_{j=1}^{n_2+1} dt_j^2 + \cos^2\theta\sum_{k=1}^{n_3+1} dz_k^2. \tag{2.19}$$

Eigenfunction solutions of (2.4) in these coordinates are

$$\begin{aligned}\psi &= (\sin\theta)^M(\cos\theta)^{\ell_3}(\sin\phi)^{\ell_2}(\cos\phi)^{\ell_1} \\ &\quad\times P_{(J-M-\ell_3)/2}^{M+\frac{1}{2}(n_1+n_2),\ell_3+\frac{1}{2}(n_3-1)}(\cos 2\theta) \\ &\quad\times P_{(M-\ell_1-\ell_2)/2}^{\ell_2+\frac{1}{2}(n_2-1),\ \ell_1+\frac{1}{2}(n_1-1)}(\cos 2\phi)\ \Theta_{\ell_1\ell_2\ell_3}(\mathbf{w},\mathbf{t},\mathbf{z}) \\ &= \psi_{JM}\Theta_{\ell_1\ell_2\ell_3},\end{aligned} \tag{2.20}$$

where $P_n^{\alpha,\beta}(z)$ are Jacobi polynomials. Here $J = L+2j$ and $M = L+2m$, where $j = 0,1,\cdots$, $m = 0,1,\cdots j-1, j$. The eigenfunctions satisfy

$$\Delta'\psi = -M(M+n_1+n_2)\psi, \tag{2.21}$$

where

$$\Delta' = \sum_{i>j} I_{ij}^2 \tag{2.22}$$

and i, j range from 1 to $n_1 + n_2 + 2$.

Note that in terms of the Cartesian coordinates u_1, u_2, u_3 on the 2-sphere ($u_1^2 + u_2^2 + u_3^2 = 1$) these eigenfunctions take the form

$$\begin{aligned}\psi_{JM} &= u_1^{\ell_1} u_2^{\ell_2} u_3^{\ell_3} (u_1^2 + u_2^2)^{(M-\ell_1-\ell_2)/2} \\ &\quad \times P_{(J-M-\ell_3)/2}^{M+\frac{1}{2}(n_1+n_2),\ \ell_3+\frac{1}{2}(n_3-1)}(1 - 2u_1^2 - 2u_2^2) \\ &\quad \times P_{(M-\ell_1-\ell_2)/2}^{\ell_2+\frac{1}{2}(n_2-1),\ \ell_1+\frac{1}{2}(n_1-1)}\left(\frac{2u_1^2}{u_1^2+u_2^2} - 1\right) \\ &= u_1^{\ell_1} u_2^{\ell_2} u_3^{\ell_3} \Phi_{jm},\end{aligned} \tag{2.23}$$

i.e., the form $u_1^{\ell} u_2^{\ell_2} u_3^{\ell_3} \Phi(u_1^2, u_2^2)$, where Φ is a polynomial.

This remark leads to another way of viewing the Heun and Jacobi bases. In the equation $\Delta\psi = -J(J + N + 1)\psi$ with $\Delta\psi$ given by (2.5) and Δ_k replaced by the values $-\ell_k(\ell_k + n_k - 1)$, $k = 1, 2, 3$, we set $\psi = u_1^{\ell} u_2^{\ell_2} u_3^{\ell_3} \Phi(x_1, x_2)$ and introduce the new coordinates $x_1 = u_1^2, x_2 = u_2^2$. The eigenvalue equation for Φ reads

$$H\Phi = -j(j + G - 1)\Phi, \tag{2.24}$$

where

$$H = \sum_{i,j=1}^{2} (x_i\delta_{ij} - x_i x_j)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{2} (\gamma_i - G x_i)\frac{\partial}{\partial x_i}. \tag{2.25}$$

Here $G = \gamma_1 + \gamma_2 + \gamma_3$ and in this particular case

$$\begin{aligned}\gamma_i &= \ell_i + \tfrac{1}{2}(n_i + 1), \qquad i = 1, 2, 3, \\ j &= \tfrac{1}{2}(J - L) = 0, 1, 2, \cdots.\end{aligned} \tag{2.26}$$

This coincides with equation (1.4) in [8]. In particular H maps polynomials of maximum degree m_i in x_i to polynomials of the same type. Furthermore, it is easy to see that the polynomial eigenfunctions of H form a basis for the space of all polynomials $f(x_1, x_2)$ and that the spectrum of H acting on this space is exactly $\{-j(j + G - 1) : j = 0, 1, \cdots\}$. It is also shown in [8] that $H = \Delta_2 + \Lambda_2$ where Δ_2 is the Laplace–Beltrami operator on S_2 and

$$\Lambda_2 = \sum_{i=1}^{2} \left[\gamma_i - \frac{1}{2} + \left(\frac{3}{2} - G\right) x_i\right] \frac{\partial}{\partial x_i}. \tag{2.27}$$

Moreover, H is self-adjoint with respect to the inner product

$$(f_1, f_2) = \iint_{x_1, x_2 > 0, 1 - x_1 - x_2 > 0} f_1(\mathbf{x})\overline{f_2(\mathbf{x})}\, d\omega, \tag{2.28}$$

where

$$\begin{aligned}dw &= x_1^{\gamma_1 - 1} x_2^{\gamma_2 - 1}(1 - x_1 - x_2)^{\gamma_3 - 1} dx_1 dx_2 : \\ &(Hf_1, f_2) = (f_1, Hf_2).\end{aligned} \tag{2.29}$$

Here f_1, f_2 are polynomials in $\mathbf{x} = (x_1, x_2)$. For fixed j the polynomials

(2.30)
$$\Phi_{jm}(x_1, x_2) = (x_1 + x_2)^m P_{j-m}^{\gamma_1+\gamma_2+2m-1,\gamma_3-1}(2x_1 + 2x_2 - 1) \times P_m^{\gamma_2-1,\ \gamma_1-1}\left(\frac{2x_1}{x_1 + x_2} - 1\right), \qquad m = 0, 1, \cdots, j$$

form an orthogonal basis for the eigenspace corresponding to eigenvalue $-j(j + G - 1)$. (This is the orthogonal basis of Proriol [9] and of Karlin and McGregor [10].) Similarly the Heun polynomials $\Phi^1_{J\ell_1\ell_2\ell_3 q}(x)\Phi^2_{J\ell_1\ell_2\ell_3 q}(y)$ where q runs over the possible eigenvalues, form an alternate orthogonal basis for this same space. Moreover as pointed out in [11] these bases correspond to spherical and ellipsoidal coordinates on the 2-sphere and are the only coordinates in which Δ_2 separates.

With this point of view we are operating on the sphere S_2 rather than S_n and our two distinguished orthogonal bases are the only ones possible rather than two out of a multiplicity of separable systems on S_n for large n. The principal advantage of this new point of view is that the eigenfunctions are obviously polynomials in x_1, x_2 and that the only requirement on the constants $\gamma_1, \gamma_2, \gamma_3$ to ensure orthogonality is that they be strictly positive. Thus the ℓ_i and n_i need not be integers; it is only required that $2\ell_i + n_i + 1 > 0$.

In the following our expansion formulas are valid for all real $\gamma_i > 0$. In the special case $\gamma_1 = \gamma_2 = \gamma_3 = \frac{1}{2}$ we have $H = \Delta_2$, the Laplace–Beltrami operator on S_2. In this case the eigenvalue equation $\Delta_2\Phi = -j(j + \frac{1}{2})\Phi$ admits the Lie algebra $so(3)$ as a symmetry algebra. A basis for $so(3)$ is $\{u_1\partial_{u_2} - u_2\partial_{u_1}, u_3\partial_{u_1}, u_3\partial_{u_2}\}$ where $u_3 = \pm(1 - u_1^2 - u_2^2)^{1/2}$. This extra symmetry is associated with the fact that there are additional polynomial solutions of the eigenvalue equation (see §3 of reference [8]). In particular the equation admits polynomial solutions of the form $f(u_1, u_2)$ and the spectrum of Δ_2 acting on the space of all such polynomials is $-j(j + \frac{1}{2})$ where now $2j = 0, 1, 2\cdots$. Furthermore, there exist solutions of the form $u_3 g(u_1, u_2)$ with g a polynomial and with the same eigenvalues. The dimension of each eigenspace is $2j + 1$ rather than $j + 1$ for the general case. In this special case the eigenfunctions corresponding to spherical coordinates are just the spherical harmonics whereas those corresponding to ellipsoidal coordinates are products of Lamé polynomials. For the solution of the problem of expanding the Lamé basis in terms of a spherical harmonic basis, see [11]–[13].

Returning to the case of general ℓ_i, n_i we consider the problem of expanding the Heun basis (2.9) in terms of the Jacobi polynomial basis (2.20), (2.23), (2.30):

(2.31)
$$\psi = u_1^{\ell_1} u_2^{\ell_2} u_3^{\ell_3} \Phi^1_{J\ell_1\ell_2\ell_3 q}(x)\Phi^2_{J\ell_1\ell_2\ell_3 q}(y) = \sum_{m=0}^{j} \xi_m \psi_{J\ell_1\ell_2\ell_3 M}(\theta, \phi).$$

Three-term recurrence relations for the expansion coefficients ξ_m (where $M = \ell_1 + \ell_2 + 2m$) can be deduced by requiring that

(2.32)
$$\mathscr{M}\psi = \chi\psi.$$

To obtain the recurrence relations we need the action of the various pieces of $\mathscr{M}$ on the Jacobi bases $\psi_{J\ell_1\ell_2\ell_3 M}(\theta, \phi)$. Since $\mathscr{M}$ commutes with H there must exist an

expansion of the form $\mathscr{M}\psi_{jm} = \sum_r X_r \psi_{j,m+r}$. Indeed, we have

$$\mathscr{M}\psi_{J\ell_1\ell_2\ell_3 M}(\theta,\phi) = \sum_{r=-1}^{+1} X_r \psi_{J\ell_1\ell_2\ell_3,M+2r}(\theta,\phi), \tag{2.33}$$

where

(2.34)

$$X_1(m,j) = \frac{4(e_1-e_2)(\gamma_1+\gamma_2+\gamma_3+m+j-1)(\gamma_3-m+j-1)(m+1)(\gamma_1+\gamma_2+m-1)}{(\gamma_1+\gamma_2+2m-1)(\gamma_1+\gamma_2+2m)},$$

$$X_{-1}(m,j) = \frac{4(e_1-e_2)(\gamma_1+\gamma_2+m+j-1)(-m+j+1)(\gamma_2-1)(\gamma_1-1)}{(\gamma_1+\gamma_2+2m-1)(\gamma_1+\gamma_2+2m-2)},$$

$$\begin{aligned} X_0(m,j)-\chi &= \frac{2(e_1-e_2)[m^2+m(\gamma_1+\gamma_2-1)-j^2-j(\gamma_1+\gamma_2+\gamma_3--1)](\gamma_1+\gamma_2-2)(\gamma_1-\gamma_2)}{(\gamma_1+\gamma_2+2m-2)(\gamma_1+\gamma_2+2m)} \\ &\quad + 4\frac{(e_1-e_2)m\gamma_3(\gamma_1-\gamma_2)(m+\gamma_2)}{(\gamma_1+\gamma_2+2m-2)(\gamma_1+\gamma_2+2m)} \\ &\quad + 2(e_1+e_2)[-m^2-m(\gamma_1+\gamma_2-1)+j^2+j(\gamma_1+\gamma_2+\gamma_3-1)] \\ &\quad + 4e_3[m^2+m(\gamma_1+\gamma_2-1)]+4q. \end{aligned}$$

Keys to deriving this result are the following recurrence formulas for Jacobi polynomials $P_n^{\alpha,\beta}(x)$:

$$xP_n^{\alpha,\beta} = \mathscr{A}P_{n-1}^{\alpha,\beta} + \mathscr{B}P_n^{\alpha,\beta} + \mathscr{C}P_{n+1}^{\alpha,\beta}, \tag{2.35}$$

$$\mathscr{A} = \frac{2(n+\alpha)(n+\beta)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta)}, \qquad \mathscr{B} = \frac{(\beta-\alpha)(\beta+\alpha)}{(2n+\alpha+\beta+2)(2n+\alpha+\beta)},$$

$$\mathscr{C} = \frac{2(n+1)(n+\alpha+\beta+1)}{(2n+\alpha+\beta+2)(2n+\alpha+\beta+1)},$$

$$(1-x^2)\frac{d}{dx}P_n^{\alpha,\beta} = AP_{n-1}^{\alpha,\beta} + BP_n^{\alpha,\beta} + CP_{n+1}^{\alpha,\beta},$$

$$A = \frac{2(n+\alpha)(n+\beta)(n+\alpha+\beta+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta)}, \qquad B = \frac{2n(\alpha-\beta)(n+\alpha+1)}{(2n+\alpha+\beta+2)(2n+\alpha+\beta)},$$

$$C = -\frac{2n(n+1)(n+\alpha+\beta+1)}{(2n+\alpha+\beta+2)(2n+\alpha+\beta+1)}.$$

(To prove (2.33) it is enough to use relations (2.35) to evaluate both sides of (2.33) for a fixed choice of the variable θ. Thus a one-variable expansion in ϕ leads to a two-variable expansion in θ and ϕ.) Now substituting the expansion (2.31) into the eigenvalue equation $\mathscr{M}\psi = \chi\psi$ and using (2.33) we find the three-term recurrence relation

$$X_1(m-1,j)\xi_{m-1} + (X_0(m,j)-\chi)\xi_m + X_{-1}(m+1,j)\xi_{m+1} = 0, \tag{2.36}$$

where $m = 0, 1, \cdots, j$. Consequently, the $j+1$ independent eigenvalues q are calculated from the determinant

(2.37)
$$\begin{vmatrix} X_0(j,j)-\chi & X_1(j-1,j) & & & \\ X_{-1}(j,j) & X_0(j-1,j)-\chi & X_1(j-2,j) & & \\ \ddots & \ddots & \ddots & & \\ & & & & \\ & & & X_{-1}(1,j) & X_0(0,j)-\chi \end{vmatrix} = 0.$$

To obtain the expansions in terms of one variable from (2.31) we proceed as follows. For the two choices of u_i, $i = 1, 2, 3$ given by (2.2) and (2.18), take $y = e_3$, $\theta = \pi/2$. Then the expression has the form

$$\Phi^1_{jq}(x) \equiv \Phi^1_{J\ell_1\ell_2\ell_3 q}(x) = \sum_{m=0}^{j} \hat{\gamma}_m P_m^{\gamma_2-1,\gamma_1-1}(\cos 2\phi), \tag{2.38}$$

where

$$\cos 2\phi = 2\frac{(x-e_1)}{(e_2-e_1)} - 1.$$

This is an expansion of type 2 with $\mu = 0$. A different type of expansion can be obtained by taking $\phi = \pi/2$ and $y = e_1$. The resulting expression has the form

$$\Phi^1_{jq}(x) = \sum_{m=0}^{j} \tilde{\gamma}_m (\sin\theta)^{2m} P_{j-m}^{2m+\gamma_1+\gamma_2-1,\ \gamma_3-1}(\cos 2\theta), \tag{2.39}$$

where

$$\cos 2\theta = -2\frac{(x-e_2)}{(e_2-e_3)} - 1.$$

In both these examples the dependence of the $\hat{\gamma}_m$ and $\tilde{\gamma}_m$ coefficients on the indices $\ell_1, \ell_2, \ell_3, q$ has been suppressed.

This second type of expansion of a Heun polynomial appears to be new. Nothing that was done in the derivation of expansions (except the limits of summation on r) could not be extended to the representation of Heun functions when $J, \ell_1, \ell_2, \ell_3$ are complex. Consequently, representations of such functions in terms of expansions whose coefficients obey three-term recurrence relations can be derived. The convergence of series of this type will be discussed elsewhere.

REFERENCES

[1] K. HEUN, *Zur Theorie der Riemannischen Functionen zweiter Ordung mit vier Verzweignungspunkten*, Math. Ann., 33 (1989), pp. 161–179.

[2] E. L. INCE, *Ordinary Differential Equations*, reprint, Dover, New York, 1956.

[3] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F.G. TRICOMI, *Higher Transcendental Functions, Vol.* 3, McGraw–Hill, 1955.

[4] N. SVARTHOLM, *Die Lösing der Fuchischen Differential Gleichung zweiter Ordnung durch hypergeometrische Polynome*, Math. Ann., 116 (1939), pp. 413–421.

[5] A. ERDÉLYI, *Certain expansions of solutions of the Heun equation*, Quart J. Math. Oxford Ser. (2), 15 (1944), pp. 62–69.

[6] E. G. KALNINS AND W. MILLER, *Separation of variables on n dimensional Riemannian manifolds. The n sphere S_n and Euclidean n space E_n*, J. Math. Phys, 27 (1986), pp. 1721–1736.

[7] E. G. KALNINS, *Separation of Variables for Spaces of Constant Riemannian Curvature*, Pitman, Harlow, U.K., 1986.

[8] E. G. KALNINS, W. MILLER, AND M. V. TRATNIK, *Families of orthogonal and biorthogonal polynomials on the n-sphere*, SIAM J. Math. Anal., 22 (1991), pp. 272–294.

[9] J. PRORIOL, *Sur une familie de polynomes a deux variables orthogonaux dans un triangle*, C.R. Acad. Sci. Paris, 245 (1957), pp. 2459–2461.

[10] S. KARLIN AND J. MCGREGOR, *Some stochastic models in genetics*, Stochastic Models in Medicine and Biology, J. Gurland, ed., University of Wisconsin Press, Madison, WI, 1964.

[11] J. PATERA AND P. WINTERNITZ, *A new basis for the representations of the rotation group. Lame and Heun polynomials*, J. Math. Phys., 14 (1973), pp. 1130–1139.

[12] E. G. KALNINS AND W. MILLER, *Lie theory and separation of variables 4. The groups $SO(2,1)$ and $SO(3)$*, J. Math. Phys., 15 (1974), pp. 1263–1274.

[13] W. MILLER, *Symmetry and Separation of Variables*, Addison–Wesley, Reading, MA, 1977.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1460-1474, September 1991

017

UNIFORM, EXPONENTIALLY IMPROVED, ASYMPTOTIC EXPANSIONS FOR THE GENERALIZED EXPONENTIAL INTEGRAL*

F. W. J. OLVER†

Abstract. By allowing the number of terms in an asymptotic expansion to depend on the asymptotic variable, it is possible to obtain an error term that is exponentially small as the asymptotic variable tends to its limit. This procedure is called "exponential improvement." It is shown how to improve exponentially the well-known Poincaré expansions for the generalized exponential integral (or incomplete Gamma function) of large argument. New uniform expansions are derived in terms of elementary functions, and also in terms of the error function.

Inter alia, the results supply a rigorous foundation for some of the recent work of M. V. Berry on a smooth interpretation of the Stokes phenomenon.

Key words. coalescing critical points, converging factors, Dingle's terminants, error function, incomplete Gamma function, Stokes' phenomenon

AMS(MOS) subject classifications. primary 41A60; secondary 33A70

1. Introduction.

Suppose that a function $f(z)$ has an asymptotic expansion of the form

$$f(z)\sim f_0+\frac{f_1}{z}+\frac{f_2}{z^2}+\cdots,$$

as $z\to\infty$ in a certain region $\mathbf{R}$, say, of the complex plane. By definition, if the series is truncated at the term f_{n-1}/z^{n-1}, where n is an arbitrary fixed integer, then the error in representing $f(z)$ by the truncated series is $O(z^{-n})$ as $z\to\infty$ in $\mathbf{R}$. For some time, however, it has been known that if we permit the number of terms in the truncated series to depend on $|z|$, then it is possible to make the truncation error *exponentially* small as $z\to\infty$ in $\mathbf{R}$.

Consider, for example, the exponential integral defined by

$$E_1(z)=e^{-z}\int_0^\infty\frac{e^{-zt}}{1+t}\,dt$$

when $|\text{ph } z|<\pi/2$, and by analytic continuation elsewhere. This has the well-known expansion.

$$e^zE_1(z)\sim\sum_{s=0}^{\infty}(-)^s\frac{s!}{z^{s+1}},\qquad z\to\infty\quad\text{in }|\text{ph } z|\leqq\frac{3}{2}\pi-\delta, \tag{1.1}$$

where δ is an arbitrary positive constant. Let $R_n(z)$ denote the nth remainder term in this expansion, given by

$$e^zE_1(z)=\sum_{s=0}^{n-1}(-)^s\frac{s!}{z^{s+1}}+R_n(z). \tag{1.2}$$

In [9, Chap. 14, § 3.1] it was shown that if ζ is a real or complex variable and θ is real, then

$$R_n\{(n+\zeta)\,e^{i\theta}\}=O(n^{-1/2}\,e^{-n-\zeta}), \tag{1.3}$$

* Received by the editors February 9, 1990; accepted for publication September 19, 1990.

† Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.
This research was supported by the National Science Foundation under grant DMS 87-23039.

as $n \to \infty$, uniformly with respect to $\theta \in [-\pi+\delta, \pi-\delta]$ and bounded values of $|\zeta|$. Let us now restrict ζ to be real, set $z=(n+\zeta)\,e^{i\theta}$, and regard z as the asymptotic variable instead of n. Then we have

$$R_n(z) = O(z^{-1/2}\,e^{-|z|}), \qquad |z| \to \infty, \tag{1.4}$$

uniformly with respect to ph $z \in [-\pi+\delta, \pi-\delta]$ and bounded values of $\|z|-n|$. For example, (1.4) would apply if $n = \text{int}\,[|z|]$, the integer part of $|z|$. When the approximation given by (1.2) is modified in this way, that is, by taking

$$n = |z| - \zeta \tag{1.5}$$

with $|\zeta|$ bounded, we shall say that the resulting partial sum is *uniformly exponentially accurate* in the sector $|\text{ph}\, z| \leqq \pi - \delta$.

The results in [9, Chap. 14, § 3] actually go much further. The object in this reference was to provide a rigorous basis for the theory of "converging factors." In the case of the asymptotic expansion (1.1) of $E_1(z)$ it was shown that

$$R_n(z) = (-)^n \frac{n!}{z^{n+1}}\, C_n(z), \tag{1.6}$$

where the converging factor $C_n(z)$ possesses an asymptotic expansion of the form

$$C_n\{(n+\zeta)\,e^{i\theta}\} \sim \sum_{s=0}^{\infty} \frac{P_s(\alpha, \zeta)}{n^s}, \qquad n \to \infty, \tag{1.7}$$

in which the $P_s(\alpha, \zeta)$ are polynomials in $\alpha \equiv (1+e^{i\theta})^{-1}$ and ζ. Again, (1.7) is uniformly valid for $\theta \in [-\pi+\delta, \pi-\delta]$ and bounded $|\zeta|$. As before, we restrict ζ to be real and set $z=(n+\zeta)\,e^{i\theta}$. Then (1.5) applies and the combination of (1.6) and (1.7) yields

$$R_n(z) \sim (-)^n \frac{\Gamma(|z|-\zeta+1)}{z^{|z|-\zeta+1}} \sum_{s=0}^{\infty} \frac{P_s(\alpha, \zeta)}{(|z|-\zeta)^s}, \qquad |z| \to \infty.$$

On replacing the Gamma function by Stirling's series and setting $|z| = z\,e^{-i\theta}$, we see that this expansion can be rearranged as an asymptotic expansion in descending powers of z; thus

$$R_n(z) \sim (-)^n\, e^{-|z|} z^{-1/2}\, e^{(\zeta-|z|-1/2)i\theta} \sum_{s=0}^{\infty} \frac{Q_s(\theta, \zeta)}{z^s}, \qquad |z| \to \infty, \tag{1.8}$$

where the coefficients $Q_s(\theta, \zeta)$ are rational functions of $e^{i\theta}$ and polynomials in ζ. Again, this expansion is uniformly valid for ph $z \in [-\pi+\delta, \pi-\delta]$ and bounded $|\zeta|$. We call the combination of (1.2) and (1.8) a *uniform, exponentially improved* ("UEI"), or more precisely, a *uniform, $e^{-|z|}$-improved, asymptotic expansion.*

Recently, Ursell [11] has considered functions $f(z)$ representable as Laplace transforms

$$f(z) = \int_0^{\infty} e^{-zt}\, \psi(t)\, dt,$$

in which $\psi(t)$ is analytic in a neighborhood of the origin. He has shown that the asymptotic expansion of $f(z)$ obtained by straightforward application of Watson's lemma can always be rendered exponentially accurate, in the sense described above. Ursell's analysis applies only to positive real values of z, but it could be extended easily to the complex plane.

The purpose of the present investigation is to show how to construct UEI expansions for a wide class of functions. The new theory goes beyond that presented in [9, Chap. 14] in three ways: (i) the number of functions that can be treated successfully is greatly increased; (ii) nonelementary functions may be used in constructing the expansions; (iii) as a consequence of (ii), regions of validity are extended considerably.

One of the theoretical and practical consequences of this investigation is to provide a rigorous basis for a recent powerful interpretation of the Stokes phenomenon that has been developed by Berry [1]. Indeed, it was this aspect that motivated this work. This is explained fully in [10], and inter alia we shall supply proofs, and significant extensions, of all results that were stated in [10].

In the present paper we confine our attention to the construction of UEI expansions for the generalized exponential integral $E_p(z)$, when p is fixed and $|z|$ is large, both p and z being real or complex. This is because the generalized exponential integral plays a fundamental role in other cases—an observation first made by Dingle [4].

Another investigation of the generalized exponential integral (or incomplete Gamma function) has been provided by Jones. Jones' paper [6] was presented independently of and at the same time as [10]. There is substantial overlap of [6] with the present investigation, but there are significant differences in the method of proof and results achieved.

2. UEI expansions in terms of elementary functions.

2.1. General properties of the generalized exponential integral. For real or complex values of p and z, other than $z=0$, the generalized exponential integral $E_p(z)$ is defined by

$$E_p(z)=z^{p-1}\Gamma(1-p,z)=z^{p-1}\int_z^\infty \frac{e^{-t}}{t^p}\,dt. \tag{2.1}$$

The only restriction on the integration path is that it must not intersect the origin. Accordingly, unless p is a nonpositive integer, $E_p(z)$ is a multiple-valued function of z. On the other hand, $E_p(z)$ is a single-valued function of p; indeed from (2.1) it is clear that for fixed nonzero values of z each branch of $E_p(z)$ is an entire function of p.[1]

On replacing t by $z(1+t)$ we find that $E_p(z)$ is also given by

$$E_p(z)=e^{-z}\int_0^\infty \frac{e^{-zt}}{(1+t)^p}\,dt, \qquad |\mathrm{ph}\, z|<\frac{1}{2}\pi, \tag{2.2}$$

where the integration path now runs along the real axis, and $(1+t)^p$ has its principal value.

An integral representation that we shall find particularly useful is given by

$$E_p(z)=\frac{z^{p-1}e^{-z}}{\Gamma(p)}\int_0^\infty \frac{e^{-zt}t^{p-1}}{1+t}\,dt, \qquad \mathrm{Re}\, p>0, \quad |\mathrm{ph}\, z|<\frac{1}{2}\pi. \tag{2.3}$$

Again the integration path is along the real axis, and t^{p-1} and z^{p-1} assume their principal values. This result may be derived from (2.2) with the aid of the convolution formula for the Laplace transform. The integral on the right-hand side can be regarded as either a Laplace transform or a Stieltjes transform.

Other branches of $E_p(z)$ can be related to the principal branch (for which $-\pi<\mathrm{ph}\, z\leqq\pi$) by rotation of the integration path in (2.3), and analytic continuation with

[1] For the case in which p is a positive integer, various properties and some numerical tables of $E_p(z)$ will be found in [7] and [8].

respect to z. Each time the path crosses the simple pole at $t=-1$ a contribution is made by the residue. Thus we find that

$$E_p(z)=\frac{2\pi i\, e^{-kp\pi i}}{\Gamma(p)}\,\frac{\sin(kp\pi)}{\sin(p\pi)}\,z^{p-1}+E_p(z\,e^{-2k\pi i}), \tag{2.4}$$

where k is any integer. Analytic continuation removes all restrictions on p and z, other than $z=0$, from this result.

2.2. Poincaré expansions for large $|z|$. Let p be fixed and $z\to\infty$. Then by applying the extended form of Watson's lemma, as given by Theorem 3.3 of [9, Chap. 4], to the integral representation (2.2), we find that

$$E_p(z)\sim\frac{e^{-z}}{z}\sum_{s=0}^{\infty}(-)^s\frac{(p)_s}{z^s},\qquad |\mathrm{ph}\,z|\leqq\frac{3}{2}\pi-\delta, \tag{2.5}$$

where $(p)_s$ is Pochhammer's symbol for the ascending factorial $p(p+1)\cdots(p+s-1)$, and δ continues to denote an arbitrary positive constant. Corresponding expansions in other phase ranges are obtainable by combination of (2.4) and (2.5). For example, by taking $k=1$ in (2.4) we derive

$$E_p(z)\sim\frac{2\pi i\, e^{-p\pi i}}{\Gamma(p)}\,z^{p-1}+\frac{e^{-z}}{z}\sum_{s=0}^{\infty}(-)^s\frac{(p)_s}{z^s},\qquad \frac{1}{2}\pi+\delta\leqq\mathrm{ph}\,z\leqq\frac{7}{2}\pi-\delta. \tag{2.6}$$

In the common sector $\pi/2+\delta\leqq\mathrm{ph}\,z\leqq\frac{3}{2}\pi-\delta$, both (2.5) and (2.6) apply. Their apparent discrepancy is a term that is exponentially small compared with the main contribution; hence there is no anomaly. This is, of course, simply a manifestation of Stokes' phenomenon. On the ray $\mathrm{ph}\,z=\pi$, the dominance of one contribution in (2.6) over the other is maximal, and for this reason we call $\mathrm{ph}\,z=\pi$ a *Stokes line.*[2] The other Stokes lines for $E_p(z)$ for large $|z|$ occur at $\mathrm{ph}\,z=k\pi$, where k is any integer.

2.3. UEI expansions for the sector $|\mathrm{ph}\,z|\leqq\pi-\delta$. If n is an arbitrary nonnegative integer, then we have identically

$$\frac{1}{1+t}=1-t+t^2-\cdots+(-)^{n-1}t^{n-1}+(-)^n\frac{t^n}{1+t},\qquad t\neq-1. \tag{2.7}$$

On substituting this expansion in (2.3) and integrating term by term, we obtain the following explicit representation of the remainder term in (2.5):

$$E_p(z)=\frac{e^{-z}}{z}\sum_{s=0}^{n-1}(-)^s\frac{(p)_s}{z^s}+(-)^n\frac{2\pi}{\Gamma(p)}\,z^{p-1}F_{n+p}(z), \tag{2.8}$$

where

$$F_{n+p}(z)=\frac{e^{-z}}{2\pi}\int_0^{\infty}\frac{e^{-zt}t^{n+p-1}}{1+t}\,dt=\frac{\Gamma(n+p)}{2\pi}\,\frac{E_{n+p}(z)}{z^{n+p-1}}. \tag{2.9}$$

In deriving this result we have assumed that $\mathrm{Re}\,p>0$ and $|\mathrm{ph}\,z|<\pi/2$. However, by analytic continuation with respect to p the first condition is replaceable by $\mathrm{Re}\,p>-n$. To ease the second restriction, we rotate the path of integration through an angle $-\gamma$,

[2] Sometimes the name anti-Stokes line is used instead; see the discussion on p. 518 of [9].

where $-\pi<\gamma<\pi$, and set $t=\tau e^{-i\gamma}$. Then by analytic continuation with respect to z we have

$$F_{n+p}(z)=\frac{e^{-i(n+p)\gamma}\,e^{-z}}{2\pi}\int_0^\infty \frac{\exp(-z\tau e^{-i\gamma})\tau^{n+p-1}}{1+\tau e^{-i\gamma}}\,d\tau, \tag{2.10}$$

valid when $\operatorname{Re} p>-n$ and $|\operatorname{ph}(z e^{-i\gamma})|<\pi/2$.

For future reference we also record here the continuation formula for $F_{n+p}(z)$ corresponding to (2.4). This is given by

$$F_{n+p}(z)=(-)^n i\,e^{-kp\pi i}\frac{\sin(kp\pi)}{\sin(p\pi)}+e^{-2kp\pi i}F_{n+p}(z e^{-2k\pi i}), \tag{2.11}$$

$k=0,\pm1,+2,\cdots$.

The ratio of the $(s+1)$st term to the sth term in (2.5) is $(1-p-s)/z$. In consequence, when $|z|$ is large compared with $|p|$, the optimum number of terms (that is, the number of terms preceding the numerically smallest term) is $\operatorname{int}[\hat{n}]+1$, where $\hat{n}$ is the positive number satisfying $|p+\hat{n}|=|z|$. Accordingly, we set

$$z=\rho e^{i\theta},\qquad n=\rho-p+\alpha, \tag{2.12}$$

where ρ is a large positive parameter, θ is real and $|\alpha|$ is bounded.[3] With $\gamma=\theta$ in (2.10), we have

$$F_{n+p}(z)=\frac{e^{-i(\rho+\alpha)\theta}\,e^{-z}}{2\pi}\int_0^\infty \frac{e^{-\rho\tau}\tau^{\rho+\alpha-1}}{1+\tau e^{-i\theta}}\,d\tau, \tag{2.13}$$

valid when $-\pi<\theta<\pi$ and $\operatorname{Re}(\rho+\alpha)>0$.

For large ρ the integrand has a saddlepoint at $\tau=1$. By application of Theorem 7.1 of [9, Chap. 4] (Laplace's method) and use of the second of (2.12) we find that

$$F_{n+p}(z)\sim\frac{e^{-i(\rho+\alpha)\theta}}{1+e^{-i\theta}}\frac{e^{-\rho-z}}{(2\pi\rho)^{1/2}}\sum_{s=0}^{\infty}\frac{a_{2s}(\theta,\alpha)}{\rho^s},\qquad \rho\to\infty, \tag{2.14}$$

where the coefficients $a_{2s}(\theta,\alpha)$ are continuous functions of θ and α. Moreover, by means of a straightforward extension of the proof of the cited theorem, we can prove that the expansion (2.14) holds uniformly with respect to θ and α, when $\theta\in[-\pi+\delta,\pi-\delta]$ and $|\alpha|$ is bounded.

The combination of (2.8), (2.12), and (2.14) is the required expansion. It is a UEI expansion throughout its sector of validity $|\operatorname{ph} z|\leqq\pi-\delta$, because if the series in (2.14) is truncated at its mth term, where m is arbitrary, but fixed, then the relative error in the resulting approximation to $E_p(z)$ is uniformly $O(z^{p-m-1/2}e^{-|z|})$.

The coefficients $a_{2s}(\theta,\alpha)$ are calculable in the usual manner (see, for example, [9, Chap. 3, § 8.1]). One way to express them is found to be

$$a_{2s}(\theta,\alpha)=1\cdot3\cdot5\cdots(2s-1)\sum_{j=0}^{2s}\lambda_{2s-1,j-1}A_j(\theta,\alpha), \tag{2.15}$$

where

$$A_s(\theta,\alpha)=\sum_{j=0}^{s}(-)^j\binom{\alpha}{s-j}\frac{1}{(1+e^{i\theta})^j}, \tag{2.16}$$

[3] This α is unrelated to the α of § 1.

and $\lambda_{s,j}$ is the coefficient of x^s in the expansion of t^j, t being related to x via

$$\tfrac{1}{2}x^2 = t - \ln(1+t) = \tfrac{1}{2}t^2 - \tfrac{1}{3}t^3 + \tfrac{1}{4}t^4 - \cdots,$$

with $x \sim t$ as $t \to 0$. Reversion of the last expansion yields

$$t = x + \tfrac{1}{3}x^2 + \tfrac{1}{36}x^3 - \tfrac{1}{270}x^4 + \tfrac{1}{4320}x^5 + \cdots, \tag{2.17}$$

and with the aid of this result we arrive at the following explicit expressions for the first three coefficients:

$$\begin{aligned} &a_0(\theta, \alpha) = 1, \qquad a_2(\theta, \alpha) = \tfrac{1}{12} + A_2(\theta, \alpha), \\ &a_4(\theta, \alpha) = \tfrac{1}{288} + \tfrac{1}{12}A_2(\theta, \alpha) + 2A_3(\theta, \alpha) + 3A_4(\theta, \alpha). \end{aligned} \tag{2.18}$$

2.4. Remarks. (i) From (2.15), (2.16), and (2.18) we observe that the coefficients $a_{2s}(\theta, \alpha)$ become unbounded as $\theta \to \pm\pi$; in consequence, the interval of validity $|\theta| \leqq \pi - \delta$ of (2.14) is maximal.

(ii) The expansion (2.14) can be rearranged as an expansion for the converging factor in descending powers of n, as indicated in § 1. The writer has verified that in the case where $p = 1$ the first three coefficients in the resulting expansion agree with those found in [9, Chap. 14, § 3.2] by quite different methods. (They are also simpler in form than their counterparts in (2.14).)

(iii) In [10] the function

$$T_{n+p}(z) = e^{(n+p)\pi i}\frac{\Gamma(n+p)}{2\pi i}\frac{E_{n+p}(z)}{z^{n+p-1}}$$

played the role of $F_{n+p}(z)$. An advantage of using $F_{n+p}(z)$ in the present paper is that it leads to a symmetric form for the main result; compare (5.10) and (5.11) below.

3. Nonelementary UEI expansions: formal derivation.

3.1. Approach. The region of validity of the UEI expansion of § 2 falls short of the Stokes lines at ph $z = \pm\pi$. We now seek UEI expansions that hold uniformly in sectors containing these lines. To begin with, we concentrate on the sector $0 < \text{ph } z < 2\pi$.

The essential difficulty is that the integrand in (2.13) has a saddlepoint at $\tau = 1$ and a simple pole at $\tau = -e^{i\theta}$, and these points coincide as $\theta \to \pi$. To overcome this difficulty we use a modification of Laplace's method introduced by van der Waerden [12] and developed subsequently by Jones [5] and Bleistein [2]. This yields an expansion involving a nonelementary function, namely, the complementary error function.

3.2. Transformation of integration variable. We employ a quadratic change of integration variable, given by

$$\tfrac{1}{2}w^2 = \tau - \ln\tau - 1. \tag{3.1}$$

The saddlepoint at $\tau = 1$ then corresponds to $w = 0$. We resolve the ambiguity in the choice of branch of $(2\tau - 2\ln\tau - 2)^{1/2}$ by requiring

$$w \sim \tau - 1 \quad \text{as } \tau \to 1. \tag{3.2}$$

Let $\mathbf{\Delta}$ denote the sector $|\text{ph } \tau| < \pi$ and $\mathbf{W}$ be its map in the w-plane. The mapping is easily determined by passage through the intermediate variable v, given by

$$v = \tau - \ln\tau - 1, \qquad w = \sqrt{2v}.$$

Corresponding points in the τ and w planes are indicated in Figs. 1 and 2. $\mathbf{W}$ is the unshaded region shown in Fig. 2; its boundaries are the curves C_1D_1 and C_2D_2 defined parametrically by

$$\tfrac{1}{2}w^2 = -\sigma - \ln\sigma \mp i\pi - 1, \qquad 0 \leqq \sigma < \infty.$$

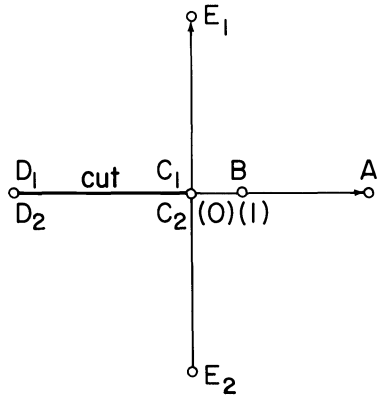


FIG. 1. *τ-plane*: *domain* **Δ**.

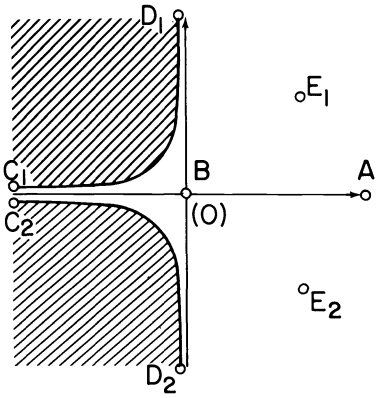


FIG. 2. *w-plane*: *domain* **W**.

From these diagrams and the corresponding map in the v-plane (not illustrated), it is clear that the mapping between the domains **Δ** and **W** is one-to-one, and that τ and w are analytic functions of each other when $\tau \in \mathbf{\Delta}$ and $w \in \mathbf{W}$.

In the w-plane, the pole at $\tau = e^{i(\theta-\pi)}$ corresponds to $w = -ic(\theta)$, say, where

$$\tfrac{1}{2}\{c(\theta)\}^2 = -e^{i(\theta-\pi)} + i(\theta-\pi) + 1. \tag{3.3}$$

In consequence of (3.2) we have

$$c(\theta) \sim -(\theta-\pi) \quad \text{as } \theta \to \pi; \tag{3.4}$$

in fact, the Taylor series expansion of $c(\theta)$ at $\theta = \pi$ begins

$$c(\theta) = -(\theta-\pi) - \frac{i}{6}(\theta-\pi)^2 + \frac{1}{36}(\theta-\pi)^3 + \cdots. \tag{3.5}$$

Graphs of $\frac{1}{2}\{c(\theta)\}^2$ and $c(\theta)$ are indicated in Figs. 3 and 4.[4]

Now consider formula (2.13). This was obtained by rotation of the path of integration in (2.9) and analytic continuation with respect to z, and is valid when

[4] Although in this section we consider only values of θ in the interval $(0, 2\pi)$, the graphs of $\frac{1}{2}\{c(\theta)\}^2$ and $c(\theta)$ are shown for the more extensive range $[-\pi, 3\pi]$ for subsequent reference. The reader should also note that the scales of Figs. 2, 3, and 4 are not the same.

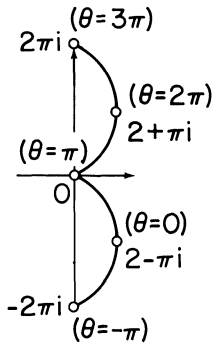


FIG. 3. $\frac{1}{2}\{c(\theta)\}^2$-*plane.*

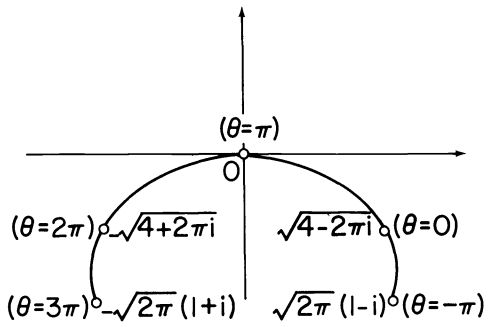


FIG. 4. $c(\theta)$-*plane.*

$-\pi < \theta < \pi$. However, an extension of this process shows that (2.13) continues to be valid when $\pi \leqq \theta < 2\pi$, as long as the integration path in (2.13) is deformed to pass above the pole at $\tau = e^{i(\theta-\pi)}$. In the w-plane this pole is mapped onto the point $w = -ic(\theta)$. In consequence, on transforming variables from τ to w in (2.13), we obtain

$$F_{n+p}(z) = \frac{e^{-i(\rho+\alpha)\theta}\, e^{-\rho-z}}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\rho w^2\right) \frac{f(\theta, \alpha, w)}{w + ic(\theta)}\, dw, \qquad 0 < \theta < 2\pi, \tag{3.6}$$

where[5]

$$f(\theta, \alpha, w) = \frac{w + ic(\theta)}{1 + \tau e^{-i\theta}}\, \tau^{\alpha-1} \frac{d\tau}{dw}, \tag{3.7}$$

and the path is indented to pass above the point $w = -ic(\theta)$ when $\theta \geqq \pi$.

3.3. Formal expansion of the transformed integral. The function $f(\theta, \alpha, w)$ is analytic at $w = -ic(\theta)$, and we may therefore set

$$f(\theta, \alpha, w) = f\{\theta, \alpha, -ic(\theta)\}[1 - \{w + ic(\theta)\}g(\theta, \alpha, w)], \tag{3.8}$$

where $g(\theta, \alpha, w)$ is analytic throughout **W**, including $w = -ic(\theta)$. On substituting this

[5] The reader is cautioned that several notational changes have been made from the earlier paper [10]. For example, c has become $-c(\theta)$, and to obtain the new $f(\theta, \alpha, w)$ from that of [10] it is necessary to replace θ and w by $\theta - \pi$ and $w + ic(\theta)$, respectively, and then multiply the result by $e^{i\alpha(\theta-\pi)}$.

expression for $f(\theta, \alpha, w)$ into (3.6), we obtain

$$F_{n+p}(z) = \frac{e^{-i(\rho+\alpha)\theta}\, e^{-\rho-z}}{2\pi} f\{\theta, \alpha, -ic(\theta)\} \cdot \left\{ \int_{-\infty}^{\infty} \frac{\exp\left(-\frac{1}{2}\rho w^2\right)}{w + ic(\theta)}\, dw - \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\rho w^2\right) g(\theta, \alpha, w)\, dw \right\}. \tag{3.9}$$

The first integral in the right-hand side of the last equation is evaluable in terms of the complementary error function:

$$\int_{-\infty}^{\infty} \frac{\exp\left(-\frac{1}{2}\rho w^2\right)}{w + ic(\theta)}\, dw = -\pi i \exp\left(\frac{1}{2}\rho\{c(\theta)\}^2\right) \operatorname{erfc}\left\{ c(\theta) \sqrt{\frac{1}{2}\rho} \right\}. \tag{3.10}$$

In the second integral the path may be taken to be a straight line, since the integrand is analytic in $\mathbf{W}$. We may expand $g(\theta, \alpha, w)$ in a Taylor series centered on the transformed saddlepoint at the origin, in the form

$$g(\theta, \alpha, w) = \sum_{s=0}^{\infty} g_s(\theta, \alpha) w^s, \tag{3.11}$$

and on multiplying by $\exp\left(-\frac{1}{2}\rho w^2\right)$ and integrating formally term by term, we arrive at

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\rho w^2\right) g(\theta, \alpha, w)\, dw \sim \sum_{s=0}^{\infty} \Gamma\left(s + \frac{1}{2}\right) g_{2s}(\theta, \alpha) \left(\frac{2}{\rho}\right)^{s+1/2}. \tag{3.12}$$

Lastly, the value of $f\{\theta, \alpha, -ic(\theta)\}$ is found by letting $\tau \to e^{i(\theta-\pi)}$ in (3.7); thus

$$f\{\theta, \alpha, -ic(\theta)\} = -e^{i\alpha(\theta-\pi)}. \tag{3.13}$$

Combination of (3.9), (3.10), (3.12), and (3.13), followed by use of (3.3) and (2.12) yields the desired expansion in the form

$$F_{n+p}(z) \sim (-)^n i\, e^{-p\pi i} \left[\frac{1}{2} \operatorname{erfc}\left\{ c(\theta) \sqrt{\frac{1}{2}\rho} \right\} - i \frac{\exp\left(-\frac{1}{2}\rho\{c(\theta)\}^2\right)}{(2\pi\rho)^{1/2}} \sum_{s=0}^{\infty} \left(\frac{1}{2}\right)_s g_{2s}(\theta, \alpha) \left(\frac{2}{\rho}\right)^s \right]. \tag{3.14}$$

To evaluate the coefficients $g_{2s}(\theta, \alpha)$, we find by reversion of (3.1)

$$\tau = 1 + w + \frac{1}{3} w^2 + \frac{1}{36} w^3 - \frac{1}{270} w^4 + \frac{1}{4320} w^5 + \cdots;$$

compare (2.17). We now substitute this Maclaurin expansion into the formula

$$g(\theta, \alpha, w) = e^{i\alpha(\pi-\theta)} \frac{\tau^{\alpha-1}}{1 + \tau e^{-i\theta}} \frac{d\tau}{dw} + \frac{1}{w + ic(\theta)} \tag{3.15}$$

obtained from (3.7), (3.8), and (3.13), and identify the coefficient of w^{2s} on the right-hand side of (3.11). For example, the first one is found to be

$$g_0(\theta, \alpha) = \frac{e^{i\alpha(\pi-\theta)}}{1 + e^{-i\theta}} - \frac{i}{c(\theta)}, \tag{3.16}$$

the right-hand side being replaced by its limiting value

$$g_0(\pi, \alpha) = \tfrac{2}{3} - \alpha, \tag{3.17}$$

when $\theta = \pi$.

In the next two sections we shall establish the asymptotic nature of the formal series (3.14), and determine its maximum region of validity in the z-plane. At the same time, we shall supply another way of calculating the coefficients $g_{2s}(\theta, \alpha)$.

4. Proof of the validity of the expansion (3.14) when $\delta \leqq \text{ph } z \leqq 2\pi - \delta$.

4.1. Objective. In this section we shall show that if the infinite series in the expansion (3.14) is truncated at the term $s = m-1$, say, where m is any fixed integer, then the difference between the resulting approximation and $F_{n+p}(z)$ is $\exp[-\frac{1}{2}\rho\{c(\theta)\}^2]O(\rho^{-m-1/2})$ as $\rho \to \infty$, uniformly with respect to θ and α. The method of proof is similar to that used in [9, Chap. 9, §§ 9-13].

4.2. Bounds for the coefficients $g_s(\theta, \alpha)$. A preliminary step is to prove that as s increases, the rate of growth of the coefficients $g_s(\theta, \alpha)$ is uniformly algebraic.

Referring to (3.11) and applying Cauchy's integral formula for the derivatives of an analytic function, we have

$$g_s(\theta, \alpha) = \frac{1}{s!}\left[\frac{d^s g(\theta, \alpha, w)}{dw^s}\right]_{w=0} = \frac{1}{2\pi i}\int_{\mathcal{W}} \frac{g(\theta, \alpha, w)}{w^{s+1}}\, dw,$$

where $\mathcal{W}$ is any simple closed contour that encloses the origin and lies within $\mathbf{W}$ (Fig. 2). Let us also require $\mathcal{W}$ to enclose the point $-ic(\theta)$. Then we may substitute in the last integral by means of (3.15) to obtain

$$g_s(\theta, \alpha) = \frac{e^{i\alpha(\pi-\theta)}}{2\pi i}\int_{\mathcal{T}} \frac{1}{w^{s+1}} \frac{\tau^{\alpha-1}}{1+\tau e^{-i\theta}}\, d\tau + \frac{1}{2\pi i}\int_{\mathcal{W}} \frac{dw}{w^{s+1}\{w+ic(\theta)\}}, \tag{4.1}$$

where $\mathcal{T}$ is the map of $\mathcal{W}$ in the τ-plane.

By hypothesis, $\theta \in [\delta, 2\pi - \delta]$. In consequence, we can construct a simple closed contour that (i) lies in the domain $\mathbf{\Delta}$ (Fig. 1); (ii) is independent of θ; (iii) has the arc $\tau = e^{i(\theta-\pi)}$, $\delta \leqq \theta \leqq 2\pi - \delta$, in its interior. We take $\mathcal{T}$ to be this contour and $\mathcal{W}$ its w-map. Let (i) $2\pi l_1$ and $2\pi l_2$ denote the lengths of $\mathcal{T}$ and $\mathcal{W}$, respectively; (ii) M be the maximum value of $|\tau^{\alpha-1}/(1+\tau e^{-i\theta})|$ when $\tau \in \mathcal{T}$, $\theta \in [\delta, 2\pi - \delta]$ and α ranges over its (bounded) set of values; (iii) d be the minimum value of $|w+ic(\theta)|$ when $w \in \mathcal{W}$ and $\theta \in [\delta, 2\pi - \delta]$. Clearly, l_1, l_2, M are finite, and $d > 0$. Since $c(\pi) = 0$, it is also clear that $|w| \geqq d$ when $w \in \mathcal{W}$. Majorizing the two integrals on the right-hand side of (4.1) by use of these bounds, we find that

$$|g_s(\theta, \alpha)| \leqq \frac{l_1 M e^{(\pi-\delta)|\operatorname{Im}\alpha|}}{d^{s+1}} + \frac{l_2}{d^{s+2}} \leqq \frac{A}{d^s}, \tag{4.2}$$

where A is assignable independently of θ and α.

4.3. Bounding the error term. Let m be an arbitrary nonnegative integer, and define $\phi_{2m}(\theta, \alpha, w)$ by the equation

$$g(\theta, \alpha, w) = \sum_{s=0}^{2m-1} g_s(\theta, \alpha) w^s + w^{2m}\phi_{2m}(\theta, \alpha, w); \tag{4.3}$$

compare (3.11). Correspondingly, in (3.14) we have

$$F_{n+p}(z) = (-)^n i\, e^{-p\pi i}\left[\frac{1}{2}\operatorname{erfc}\left\{c(\theta)\sqrt{\frac{1}{2}\rho}\right\} - i\frac{\exp(-\frac{1}{2}\rho\{c(\theta)\}^2)}{(2\pi\rho)^{1/2}} \cdot \left\{\sum_{s=0}^{m-1}\left(\frac{1}{2}\right)_s g_{2s}(\theta, \alpha)\left(\frac{2}{\rho}\right)^s + \Phi_m(\theta, \alpha, \rho)\right\}\right], \tag{4.4}$$

where

$$\Phi_m(\theta, \alpha, \rho) = \sqrt{\frac{\rho}{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\rho w^2\right) w^{2m} \phi_{2m}(\theta, \alpha, w)\, dw. \tag{4.5}$$

Suppose first that $-d/2 \leqq w \leqq d/2$, where d is defined in § 4.2. Then using (4.2) we see that

$$|\phi_{2m}(\theta, \alpha, w)| = \left| \sum_{s=2m}^{\infty} g_s(\theta, \alpha) w^{s-2m} \right| \leqq \frac{2A}{d^{2m}}.$$

Hence

$$\begin{aligned} &\left| \int_{-d/2}^{d/2} \exp\left(-\frac{1}{2}\rho w^2\right) w^{2m} \phi_{2m}(\theta, \alpha, w)\, dw \right| \\ &\quad < \frac{2A}{d^{2m}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\rho w^2\right) w^{2m}\, dw = O\left(\frac{1}{\rho^{m+1/2}}\right). \end{aligned} \tag{4.6}$$

Second, suppose that $d/2 \leqq w < \infty$. From (3.1) we see that as $w \to +\infty$ we have $\tau \sim w^2/2$ and $d\tau/dw \sim w$. Hence

$$\frac{\tau^{\alpha-1}}{1+\tau e^{-i\theta}} \frac{d\tau}{dw} \sim 2^{2-\alpha} e^{i\theta} w^{2\alpha-3} \quad \text{uniformly as } w \to +\infty.$$

We now refer to (3.15) and bear in mind that $|1+\tau e^{-i\theta}|$ and $|w+ic(\theta)|$ each have positive lower bounds when $w \in [d/2, \infty)$ and $\theta \in [\delta, 2\pi-\delta]$; compare Figs. 1, 2, and 4. In this way we may verify that

$$|g(\theta, \alpha, w)| \leqq A w^a, \qquad \tfrac{1}{2}d \leqq w < \infty,$$

where

$$a = \max_{\forall \alpha} (2 \operatorname{Re} \alpha - 3, -1), \tag{4.7}$$

and we are now using the symbol A generically to denote a finite constant that is independent of θ and α. Substituting into (4.3) by means of this result and (4.2), we conclude that when $m \geqq a/2$ we have

$$|\phi_{2m}(\theta, \alpha, w)| \leqq A, \qquad \tfrac{1}{2}d \leqq w < \infty.$$

Hence

$$\begin{aligned} &\left| \int_{d/2}^{\infty} \exp\left(-\frac{1}{2}\rho w^2\right) w^{2m} \phi_{2m}(\theta, \alpha, w)\, dw \right| \\ &\quad \leqq A \int_{0}^{\infty} \exp\left(-\frac{1}{2}\rho w^2\right) w^{2m}\, dw = O\left(\frac{1}{\rho^{m+1/2}}\right). \end{aligned} \tag{4.8}$$

Lastly, suppose that $-\infty < w \leqq -d/2$. From (3.1) we see that if $w \to -\infty$, then $\tau \sim \exp(-w^2/2 - 1)$ and $d\tau/dw \sim -w\tau$. Hence

$$\frac{\tau^{\alpha-1}}{1+\tau e^{-i\theta}} \frac{d\tau}{dw} \sim -w\tau^{\alpha} \sim -w \exp\left\{-\alpha\left(\frac{1}{2}w^2 + 1\right)\right\},$$

uniformly, and therefore from (3.15) we have

$$|g(\theta, \alpha, w)| \leqq A|w|\, e^{bw^2}, \qquad -\infty < w \leqq -\tfrac{1}{2}d,$$

where

$$b = \max_{\forall \alpha} (-\tfrac{1}{2} \operatorname{Re} \alpha, 0); \tag{4.9}$$

compare (4.7). Substituting in (4.3) by means of this result and (4.2), we conclude that when $m \geqq 1$ we have

$$|\phi_{2m}(\theta, \alpha, w)| \leqq A\, e^{bw^2}, \qquad -\infty < w \leqq -\tfrac{1}{2} d.$$

Hence if, also, $\rho > 2b$ we have

$$\left| \int_{-\infty}^{-d/2} \exp\left(-\frac{1}{2}\rho w^2\right) w^{2m} \phi_{2m}(\theta, \alpha, w)\, dw \right| \leqq A \int_{-\infty}^{0} \exp\left\{-\left(\frac{1}{2}\rho - b\right) w^2\right\} w^{2m}\, dw = O\left(\frac{1}{\rho^{m+1/2}}\right). \tag{4.10}$$

On combining (4.5), (4.6), (4.8), and (4.10), we conclude that the error term in the expansion (4.4) satisfies

$$\Phi_m(\theta, \alpha, \rho) = O(\rho^{-m}), \qquad \rho \to \infty, \tag{4.11}$$

uniformly with respect to $\theta \in [\delta, 2\pi - \delta]$ and bounded values of $|\alpha|$. This is the desired result.

4.4. Remark. The significance of the results established in this section apropos the Stokes phenomenon is as follows. From Fig. 4 it is seen that $\xi \equiv c(\theta)\sqrt{\rho/2}$ lies in the sector $-\pi/4 \leqq \text{ph}\, \xi \leqq 0$ when $-\pi \leqq \theta \leqq \pi$, and in the sector $0 \leqq \text{ph}\,(-\xi) \leqq \pi/4$ when $\pi \leqq \theta \leqq 3\pi$. It is well known that $\text{erfc}\,(\xi) = O(e^{-\xi^2})$ uniformly throughout the first sector, and $\text{erfc}\,(\xi) = 2 + O(e^{-\xi^2})$ uniformly throughout the second sector (see, for example, [9, Chaps. 2-4]). In consequence, if ρ is large and fixed and θ increases continuously from values just below π to values just above π, then $\frac{1}{2}\,\text{erfc}\,\{c(\theta)\sqrt{\rho/2}\}$ changes rapidly, but smoothly, from being exponentially small to being exponentially close to 1. From (4.2) and (4.11) it is easily seen that the same conclusion also applies to the whole content of the square brackets in (4.4). On substituting into (2.8) by means of this result we perceive that if $|z|$ is large and the asymptotic expansion (2.5) is truncated at its optimum number of terms, then there is a rapid, *but smooth*, transition to the form (2.6) in the neighborhood of $\text{ph}\, z = \pi$. Berry described the Stokes phenomenon in this illuminating manner in [1], but his analysis was purely formal. The present investigation serves to place Berry's conclusions on a rigorous basis—in the case of the generalized exponential integral. For further details concerning this aspect see [10].

5. Main results and conclusions.

5.1. Comparison of the expansions of §§ 2 and 4; formulae for the coefficients $g_{2s}(\theta, \alpha)$. Let us summarize the results obtained so far. Setting $z = \rho\, e^{i\theta}$, $n = \rho - p + \alpha$ and defining $F_{n+p}(z)$ by (2.8), we proved in § 2 that

$$F_{n+p}(z) \sim \frac{e^{-i(\rho+\alpha)\theta}}{1 + e^{-i\theta}} \frac{e^{-\rho - z}}{(2\pi\rho)^{1/2}} \sum_{s=0}^{\infty} \frac{a_{2s}(\theta, \alpha)}{\rho^s}, \qquad \rho \to \infty, \tag{5.1}$$

uniformly with respect to bounded $|\alpha|$ and $\theta \in [-\pi + \delta, \pi - \delta]$, δ being an arbitrary positive constant. In §§ 3 and 4, we proved that

$$F_{n+p}(z) \sim (-)^n i\, e^{-p\pi i}\left[\frac{1}{2}\,\text{erfc}\left\{c(\theta)\sqrt{\frac{1}{2}\rho}\right\} - i\frac{\exp\left(-\frac{1}{2}\rho\{c(\theta)\}^2\right)}{(2\pi\rho)^{1/2}} \sum_{s=0}^{\infty} \left(\frac{1}{2}\right)_s g_{2s}(\theta, \alpha)\left(\frac{2}{\rho}\right)^s\right] \tag{5.2}$$

as $\rho \to \infty$, uniformly with respect to bounded $|\alpha|$ and $\theta \in [\delta, 2\pi - \delta]$; here $c(\theta)$ is defined as in § 3.2.

Now consider the common sector of validity, given by $\delta \leqq \theta \leqq \pi - \delta$. From Fig. 4 we see that $c(\theta)$ lies in the fourth quadrant of $\mathbb{C}$ and is bounded away from the origin. We may therefore replace $\frac{1}{2}\operatorname{erfc}\{c(\theta)\sqrt{\rho/2}\}$ by an asymptotic expansion in descending powers of $c(\theta)\sqrt{\rho/2}$ that is appropriate for this region. This is given by[6]

$$\frac{1}{2}\operatorname{erfc}\left\{c(\theta)\sqrt{\frac{1}{2}\rho}\right\} \sim \frac{\exp\left(-\frac{1}{2}\rho\{c(\theta)\}^2\right)}{c(\theta)(2\pi\rho)^{1/2}} \sum_{s=0}^{\infty} (-)^s \frac{(\frac{1}{2})_s}{\{c(\theta)\}^{2s}} \left(\frac{2}{\rho}\right)^s, \tag{5.3}$$

this result being uniformly valid with respect to θ in an interval that includes $[\delta, \pi - \delta]$. Substitution into (5.2) yields

$$\begin{aligned} F_{n+p}(z) \sim (-)^n i\, e^{-p\pi i} & \frac{\exp\left(-\frac{1}{2}\rho\{c(\theta)\}^2\right)}{(2\pi\rho)^{1/2}} \\ & \cdot \left\{ \sum_{s=0}^{\infty} (-)^s \frac{(\frac{1}{2})_s}{\{c(\theta)\}^{2s+1}} \left(\frac{2}{\rho}\right)^s - i \sum_{s=0}^{\infty} \left(\frac{1}{2}\right)_s g_{2s}(\theta, \alpha) \left(\frac{2}{\rho}\right)^s \right\}. \end{aligned} \tag{5.4}$$

We also observe from (2.12) and (3.3) that (5.1) may be recast in the form

$$F_{n+p}(z) \sim (-)^n\, e^{-p\pi i} \frac{e^{i\alpha(\pi-\theta)}}{1+e^{-i\theta}} \frac{\exp\left(-\frac{1}{2}\rho\{c(\theta)\}^2\right)}{(2\pi\rho)^{1/2}} \sum_{s=0}^{\infty} \frac{a_{2s}(\theta, \alpha)}{\rho^s}. \tag{5.5}$$

We now have two uniform asymptotic expansions of the same function in descending powers of ρ. They must therefore be identical. Equating coefficients, we conclude that

$$g_{2s}(\theta, \alpha) = \frac{e^{i\alpha(\pi-\theta)}}{1+e^{-i\theta}} \frac{a_{2s}(\theta, \alpha)}{1 \cdot 3 \cdots (2s-1)} + \frac{(-)^{s-1} i}{\{c(\theta)\}^{2s+1}}, \qquad s = 0, 1, 2, \cdots. \tag{5.6}$$

Analytic continuation extends this formula to all values of θ, other than odd multiples of π; furthermore, at $\theta = \pi$ we may replace the right-hand side by its limiting value.

With the aid of the formula $a_0(\theta, \alpha) = 1$ given at the end of § 2.3, we immediately perceive that (5.6) agrees with (3.16) in the case $s = 0$. For other values of s, formula (5.6) provides a convenient way of evaluating $g_{2s}(\theta, \alpha)$.

5.2. Extending the sector of validity. In expanding the error function that appears in (5.2) in the form (5.3), we considered only values of θ in the interval $[\delta, \pi - \delta]$. However, from Figs. 3 and 4 it is clear that (5.3) continues to be uniformly valid in an interval that includes $[-\pi + \delta, \pi - \delta]$. Since this again leads to an expansion, namely (5.5), that is known to be valid in this larger interval, it follows that (5.2) itself must be uniformly valid in this interval.

By symmetry, we also expect (5.2) to be valid in the interval $[\pi + \delta, 3\pi - \delta]$. And this can be verified by expanding the complementary error function in (5.2) in a form appropriate for this sector, and comparing the result with the expansion

$$F_{n+p}(z) \sim (-)^n i\, e^{-p\pi i} + \frac{e^{-i(\rho+\alpha)\theta}}{1+e^{-i\theta}} \frac{e^{-\rho-z}}{(2\pi\rho)^{1/2}} \sum_{s=0}^{\infty} \frac{a_{2s}(\theta, \alpha)}{\rho^s}, \qquad \pi + \delta \leqq \theta \leqq 3\pi - \delta, \tag{5.7}$$

obtained from (2.11), with $k = 1$, and (5.1).[7]

5.3. Main theorem. Another expansion for $F_{n+p}(z)$ can be obtained from (5.2) by replacing p and z by their complex conjugates, and then taking the conjugate of the whole result. On combining all the expansions obtained so far we arrive at the following theorem.

---

[6] See, for example, [9, pp. 67, 112].

[7] In deriving (5.7), we also need to use (2.12) and the fact that $a_{2s}(\theta, \alpha)$ is periodic in θ, with period 2π.

THEOREM 1. *Let*

$$z=\rho e^{i\theta}, \qquad n=\rho-p+\alpha, \tag{5.8}$$

and define $c(\theta)$ *by* (3.3), (3.4), *and* $g_{2s}(\theta,\alpha)$ *by* (5.6), (2.15), (2.16). *Then*

$$E_p(z)=\frac{e^{-z}}{z}\sum_{s=0}^{n-1}(-)^s\frac{(p)_s}{z^s}+(-)^n\frac{2\pi}{\Gamma(p)}z^{p-1}F_{n+p}(z), \tag{5.9}$$

where for fixed p *and large* $|z|$

$$F_{n+p}(z)\sim(-)^n i\,e^{-p\pi i}\left[\frac{1}{2}\operatorname{erfc}\left\{c(\theta)\sqrt{\frac{1}{2}\rho}\right\}-i\frac{\exp\left(-\frac{1}{2}\rho\{c(\theta)\}^2\right)}{(2\pi\rho)^{1/2}}\sum_{s=0}^{\infty}\left(\frac{1}{2}\right)_s g_{2s}(\theta,\alpha)\left(\frac{2}{\rho}\right)^s\right], \tag{5.10}$$

valid when $-\pi+\delta\leqq\theta\leqq3\pi-\delta$, *and*

$$F_{n+p}(z)\sim(-)^{n-1}i\,e^{p\pi i}\left[\frac{1}{2}\operatorname{erfc}\left\{\overline{c(-\theta)}\sqrt{\frac{1}{2}\rho}\right\}+i\frac{\exp\left(-\frac{1}{2}\rho\{\overline{c(-\theta)}\}^2\right)}{(2\pi\rho)^{1/2}}\sum_{s=0}^{\infty}\left(\frac{1}{2}\right)_s \overline{g_{2s}(-\theta,\bar{\alpha})}\left(\frac{2}{\rho}\right)^s\right], \tag{5.11}$$

valid when $-3\pi+\delta\leqq\theta\leqq\pi-\delta$. *These expansions are uniform with respect to* θ *and bounded values of* $|\alpha|$; *furthermore, the* θ*-intervals of validity are maximal.*

The only part of this theorem we have not established already is the maximality of the θ-intervals of validity. However, this is an immediate consequence of the fact that the coefficients $g_{2s}(\theta,\alpha)$ become infinite as $\theta\to-\pi$ or $\theta\to3\pi$; compare (5.6) and (2.15).

5.4. Remarks. (i) If we truncate the infinite series in (5.10) at the term $s=m-1$, where m is an arbitrary fixed nonnegative integer, then the error in the resulting approximation to $F_{n+p}(z)$ is $\exp\left[-\frac{1}{2}\rho\{c(\theta)\}^2\right]z^{-m-1/2}O(1)$, that is, $e^{-z-|z|}z^{-m-1/2}O(1)$; compare (3.3) (again). On referring to the expansions (2.5) and (2.6) we see that the combination of (5.9) and (5.10) is a uniform, $e^{-|z|}$-improved, expansion (in the sense of § 1) throughout the region of validity of (5.10). Similarly, the combination of (5.9) and (5.11) is a uniform, $e^{-|z|}$-improved, expansion throughout the region of validity of (5.11).

(ii) In applications, the use of (5.10) and (5.11) can be confined to the sectors $0\leqq\operatorname{ph}z\leqq\pi$ and $-\pi\leqq\operatorname{ph}z\leqq0$, respectively, and the continuation formula (2.11) (or equivalently (2.4)) used elsewhere. Indeed, this procedure is consistent with maximum accuracy; compare the observations made in [9, Chap. 14, § 1]. However, there are two reasons we sought the maximal regions of validity of these expansions. First, these extensions show that, for example, the combination of (5.9) and (5.10) includes both (2.5) and (2.6) within most of their regions of validity. Second, the extensions complete the proof of results stated in [10, § 3] in connection with Berry's interpretation of the Stokes phenomenon.

(iii) The expansions (5.10) and (5.11) have the common interval of validity $|\operatorname{ph}z|\leqq\pi-\delta$. The two expansions are not the same here, but their discrepancy is absorbable in the implied error terms. Thus there is also a Stokes phenomenon associated with these expansions in the neighborhood of $\operatorname{ph}z=0$, and it would appear to be possible to extend the whole process to construct UEI expansions of the UEI expansions.[8]

[8] For further comments on this possibility see [3, § 4].

(iv) The most serious deficiency in the present results is a set of realistic error bounds for the expansions (5.10) and (5.11). Unfortunately, at this time this remains a general drawback to asymptotic expansions obtained from integral representations with coalescing critical points.

Acknowledgements. The author is pleased to acknowledge helpful suggestions by the referees.

REFERENCES

[1] M. V. BERRY, *Uniform asymptotic smoothing of Stokes's discontinuities*, Proc. Roy. Soc. London Ser. A, 422 (1989), pp. 7–21.

[2] N. BLEISTEIN, *Uniform asymptotic expansions of integrals with stationary point near algebraic singularity.* Comm. Pure Appl. Math., 19 (1966), pp. 353–370.

[3] W. G. C. BOYD, *Stieltjes transforms and the Stokes phenomenon*, Proc. Roy. Soc. London Ser. A, 429 (1990), pp. 227–246.

[4] R. B. DINGLE, *Asymptotic expansions and converging factors* I. *General theory and basic converging factors*, Proc. Roy. Soc. London Ser. A, 244 (1958), pp. 456–475.

[5] D. S. JONES, *Asymptotic behavior of integrals*, SIAM Rev., 14 (1972), pp. 286–317.

[6] ———, *Uniform asymptotic remainders*, in Asymptotic and Computational Analysis, R. Wong, ed., Marcel Dekker, New York, 1990, pp. 241–264.

[7] G. F. MILLER, *Tables of Generalized Exponential Integrals*, N. P. L. Mathematical Tables, Vol. 3, H.M. Stationery Office, London, 1960.

[8] National Bureau of Standards, *Handbook of Mathematical Functions*, Applied Mathematical Series No. 55, M. Abramowitz and I. A. Stegun, eds., U.S. Government Printing Office, Washington, DC, 1964.

[9] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic, Press, New York, 1974.

[10] ———, *On Stokes' phenomenon and converging factors*, in Asymptotic and Computational Analysis, R. Wong, ed., Marcel Dekker, New York, 1990, pp. 329–355.

[11] F. URSELL, *Integrals with a large parameter. A strong form of Watson's lemma*, in Elasticity: Mathematical Methods and Applications, R. W. Ogden and G. Eason, eds., Ellis Horwood,Chichester, 1990, pp. 391–395.

[12] B. L. VAN DER WAERDEN, *On the method of saddle points*, Appl. Sci. Res. Ser. B, 2 (1951), pp. 33–45.

SIAM J. MATH. ANAL.
Vol. 22, No. 5, pp. 1475-1489, September 1991

018

UNIFORM, EXPONENTIALLY IMPROVED, ASYMPTOTIC EXPANSIONS FOR THE CONFLUENT HYPERGEOMETRIC FUNCTION AND OTHER INTEGRAL TRANSFORMS*

F. W. J. OLVER†

Abstract. A new generalized asymptotic expansion is constructed for the confluent hypergeometric function $U(a, a-b+1, z)$ in which the parameters a and b are real or complex constants, and z is a large complex variable. The expansion is expressed in terms of generalized exponential integrals (or, equivalently, incomplete Gamma functions). It has a larger region of validity and greater accuracy than the conventional expansions of Poincaré type; moreover, it provides insight into the manner in which the Poincaré expansions change smoothly, albeit rapidly, from one to the other in the vicinity of the so-called Stokes lines. The expansion is accompanied by strict error bounds in the most important part of its region of validity.

The method used is quite general and can be applied to other functions that are representable as transforms of Laplace or Stieltjes type.

Key words. coalescing critical points, error bounds, generalized exponential integral, incomplete Gamma function, Laplace transform, Stieltjes transform, Stokes' phenomenon

AMS(MOS) subject classifications. primary 41A60; secondary 33A30

1. Introduction and summary.

In the preceding paper [10] we investigated the generalized exponential integral $E_p(z)$, in which p and z are unrestricted real or complex variables. We began with the identity

$$E_p(z)=\frac{e^{-z}}{z}\sum_{s=0}^{n-1}(-)^s\frac{(p)_s}{z^s}+(-)^n\frac{2\pi}{\Gamma(p)}z^{p-1}F_{n+p}(z), \tag{1.1}$$

in which n is an arbitrary nonnegative integer, $(p)_s$ denotes the ascending factorial $p(p+1)\cdots(p+s-1)$, and

$$F_p(z)=\frac{\Gamma(p)}{2\pi}\frac{E_p(z)}{z^{p-1}}. \tag{1.2}$$

Writing z in polar form $z=\rho e^{i\theta}$, we were able to prove that if $\rho\to\infty$, with p fixed, then

$$\begin{aligned}F_{n+p}(z)\sim(-)^n i e^{-p\pi i}\Bigg[&\frac{1}{2}\operatorname{erfc}\left\{c(\theta)\sqrt{\frac{1}{2}\rho}\right\}\\&-i\frac{\exp(-\frac{1}{2}\rho\{c(\theta)\}^2)}{(2\pi\rho)^{1/2}}\sum_{s=0}^{\infty}\left(\frac{1}{2}\right)_s g_{2s}(\theta,\alpha)\left(\frac{2}{\rho}\right)^s\Bigg],\end{aligned} \tag{1.3}$$

uniformly with respect to $\theta\in[-\pi+\delta, 3\pi-\delta]$ and bounded values of $|\alpha|$, where $\alpha=n-\rho+p$. Here (and elsewhere) δ denotes an arbitrarily small positive constant; furthermore,

$$c(\theta)=\sqrt{2\{e^{i\theta}+i(\theta-\pi)+1\}},$$

with an appropriate choice of branch of the square root, and the coefficients $g_{2s}(\theta,\alpha)$ are continuous functions of θ and α. We also supplied a similar expansion for $F_{n+p}(z)$ when $\theta\in[-3\pi+\delta, \pi-\delta]$.

The importance of the expansion (1.3) is threefold.

* Received by the editors February 9, 1990; accepted for publication September 19, 1990.

† Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.
This research was supported by the National Science Foundation under grant DMS 87-23039.

First, the combination of (1.1) and (1.3) unifies the Poincaré-type expansion

$$E_p(z)\sim\frac{e^{-z}}{z}\sum_{s=0}^{\infty}(-)^s\frac{(p)_s}{z^s},\qquad \theta\in\left[-\frac{3}{2}\pi+\delta,\frac{3}{2}\pi-\delta\right], \tag{1.4}$$

and the compound expansion

$$E_p(z)\sim\frac{2\pi i\,e^{-p\pi i}}{\Gamma(p)}z^{p-1}+\frac{e^{-z}}{z}\sum_{s=0}^{\infty}(-)^s\frac{(p)_s}{z^s},\qquad \theta\in\left[\frac{1}{2}\pi+\delta,\frac{7}{2}\pi-\delta\right], \tag{1.5}$$

since these expansions can be derived from (1.1) and (1.3) by appropriate re-expansion in the common regions of validity.

Second, as a consequence of this unification, the combination of (1.1) and (1.3) accurately quantifies the Stokes phenomenon, that is, the rapid but smooth change in form from (1.4) to (1.5) as θ passes through the common interval of validity $[\pi/2+\delta, \frac{3}{2}\pi-\delta]$.

Third, in much of its region of validity the combination of (1.1) and (1.3) is more accurate than either (1.4) or (1.5), owing to the fact that it contains an extra factor $e^{-|z|}$ in its uniform error estimates. For this reason, we called the combination of (1.1) and (1.3) a *uniform, exponentially improved* (UEI), *asymptotic expansion.*

The purpose of the present paper is to furnish a general method for constructing generalized asymptotic expansions that enjoy similar properties, especially uniform exponential improvement. The functions we shall consider are representable as Laplace or Stieltjes transforms. We illustrate the method by detailed treatment of the confluent hypergeometric function defined by

$$U(a,a-b+1,z)=\frac{1}{\Gamma(a)}\int_0^{\infty}e^{-zt}t^{a-1}(1+t)^{-b}\,dt, \tag{1.6}$$

when $\operatorname{Re} a>0$ and $|\operatorname{ph} z|<\pi/2$, and by analytic continuation elsewhere. Special cases of this function include Bessel and modified Bessel functions, parabolic cylinder functions, and Airy functions. However, our method applies to a broader class of functions than those of confluent hypergeometric type.

Instead of expressing the generalized asymptotic expansions in terms of the complementary error function, as was done for the function $E_p(z)$, we find it more convenient, and elegant, to express these expansions in terms of the F_p functions. Our main result is as follows:

THEOREM 1. *Define* $R_n(a,b,z)$ *by*

$$U(a,a-b+1,z)=z^{-a}\sum_{s=0}^{n-1}(-)^s\frac{(a)_s(b)_s}{s!\,z^s}+R_n(a,b,z), \tag{1.7}$$

where

$$n=|z|-a-b+1+\alpha, \tag{1.8}$$

$|z|$ *being large,* a *and* b *being fixed real or complex parameters, and* $|\alpha|$ *being bounded. Then*

$$R_n(a,b,z)=(-)^n2\pi\frac{z^{b-1}\,e^z}{\Gamma(a)\Gamma(b)}\left\{\sum_{s=0}^{m-1}(-)^s\frac{(1-a)_s(1-b)_s}{s!}\frac{F_{n-s+a+b-1}(z)}{z^s}\right.$$
$$\left.+(1-a)_m(1-b)_mR_{m,n}(a,b,z)\right\}, \tag{1.9}$$

where m *is an arbitrary fixed integer, and*

$$R_{m,n}(a,b,z)=O(e^{-z-|z|}z^{-m}),\qquad |\operatorname{ph} z|\leqq\pi, \tag{1.10}$$

$$R_{m,n}(a,b,z)=O(z^{-m}),\qquad \pi\leqq|\operatorname{ph} z|\leqq\tfrac{5}{2}\pi-\delta. \tag{1.11}$$

Furthermore, these sectors of validity are maximal.

Theorem 1 is proved in the next four sections. In § 6 we derive a strict and realistic bound for $R_{m,n}(a, b, z)$ in the sector $|\text{ph } z| \leqq \pi$. Some general conclusions are drawn in § 7.

To conclude this Introduction we make the following observations:

(i) From (1.10) and (1.11) it is seen that the combination of (1.7) and (1.9) furnishes a uniform exponentially improved expansion in a sector that includes $|\text{ph } z| \leqq \pi$. For other phase ranges improved expansions can be constructed immediately by combining (1.7) and (1.9) with the connection formulae [4, p. 27]

$$\begin{aligned} & e^{k(a-b)\pi i} U(a, a-b+1, z e^{2k\pi i}) \\ & = \pm \frac{e^{\mp(a+b)\pi i} \sin k(a-b)\pi - \sin (k \mp 1)(a-b)\pi}{\sin (a-b)\pi} U(a, a-b+1, z) \\ & \quad + \frac{2\pi i \sin k(a-b)\pi}{\sin (a-b)\pi} \frac{e^{\mp b\pi i}}{\Gamma(a)\Gamma(b)} e^z U(1-b, a-b+1, z e^{\pm \pi i}), \end{aligned} \tag{1.12}$$

in which k is an arbitrary integer and either upper signs, or lower signs, are taken throughout.

(ii) The well-known Poincaré-type expansion

$$U(a, a-b+1, z) \sim z^{-a} \sum_{s=0}^{\infty} (-)^s \frac{(a)_s (b)_s}{s! z^s}, \qquad |\text{ph } z| \leqq \tfrac{3}{2}\pi - \delta, \tag{1.13}$$

and the two compound expansions

$$\begin{aligned} U(a, a-b+1, z) & \sim z^{-a} \sum_{s=0}^{\infty} (-)^s \frac{(a)_s (b)_s}{s! z^s} \\ & \quad \mp 2\pi i \frac{e^{\mp(a+b)\pi i}}{\Gamma(a)\Gamma(b)} z^{b-1} e^z \sum_{s=0}^{\infty} \frac{(1-a)_s (1-b)_s}{s! z^s}, \\ & |\text{ph } \{z \exp (\mp \tfrac{3}{2}\pi i)\}| \leqq \pi - \delta, \end{aligned} \tag{1.14}$$

obtained by substitution of (1.13) into (1.12), with $k = \pm 1$, can be recovered from (1.7) and (1.9) by appropriate re-expansion.

(iii) Owing to the presence of the factors $(1-a)_m$ and $(1-b)_m$ in (1.9), $R_n(a, b, z)$ can be expressed exactly as a finite combination of F_p functions when a or b is a positive integer, by taking m to be sufficiently large. Furthermore, the estimates (1.10) and (1.11) enjoy certain uniformity properties with respect to the parameters a and b. In the case of the sector $|\text{ph } z| \leqq \pi$ the last statement is verifiable by inspection of the error bound for $R_{m,n}(a, b, z)$ that is supplied in § 6 below.

(iv) The original purpose of the present paper was to supply proofs of results that were stated in [9, § 4]. In turn, the main purpose of [9] was to provide a rigorous basis for the formal researches of Berry on Stokes' phenomenon [2]. These objectives are achieved, but we go well beyond them, especially with respect to regions of validity and exponential improvement.

(v) A formal expansion of $R_n(a, b, z)$ in terms of the function $z^{n+b} e^z F_{n+b}(z)$ and its derivatives was found by Dingle [5]. This expansion is not equivalent to (1.9) and Dingle was concerned only with the sector $|\text{ph } z| < \pi$. Whether his expansion is valid in a larger sector has not been investigated, but perhaps this question is moot since our new expansion is of a simpler and more revealing form.

(vi) For work on the modified Bessel function $K_\nu(z)$ that is closely related to [9, § 4] and the present paper, see [3] and [6].

2. Representation of the remainder term as a Stieltjes transform. If, in (1.6), we rotate the path of integration through an angle $-\gamma$, γ being an arbitrary number in the interval $(-\pi, \pi)$, then we obtain the analytic continuation of $U(a, a-b+1, z)$ in the sector $|\mathrm{ph}\,(z\,e^{-i\gamma})| < \pi/2$, given by

$$U(a, a-b+1, z) = \frac{1}{\Gamma(a)} \int_0^{\infty e^{-i\gamma}} e^{-zt} t^{a-1} (1+t)^{-b}\, dt. \tag{2.1}$$

We now expand the factor $(1+t)^{-b}$ by Taylor's theorem, in the form

$$(1+t)^{-b} = \sum_{s=0}^{n-1} (-)^s \frac{(b)_s}{s!} t^s + \chi_n(b, t), \tag{2.2}$$

where n is an arbitrary nonnegative integer,

$$\chi_n(b, t) = \frac{t^n}{2\pi i} \int_{\mathscr{L}} \frac{(1+w)^{-b}}{w^n (w-t)}\, dw, \tag{2.3}$$

and $\mathscr{L}$ is a simple closed contour that encloses $w=0$ and $w=t$, but not $w=-1$ (see, for example, [1, Chap. 4, § 3.1]). On substituting into (2.1) by means of (2.2) we obtain (1.7) with the following representation of the remainder term:

$$R_n(a, b, z) = \frac{1}{\Gamma(a)} \int_0^{\infty e^{-i\gamma}} e^{-zt} t^{a-1} \chi_n(b, t)\, dt. \tag{2.4}$$

Now suppose that $\mathrm{Re}\, b + n > 0$. Then the integral of $(1+w)^{-b} w^{-n} (w-t)^{-1}$ around any large circle in the w-plane vanishes as the radius of the circle tends to infinity. Accordingly, the contour $\mathscr{L}$ in (2.3) may be deformed into the loop contour depicted in Fig. 1, giving

$$\chi_n(b, t) = \frac{t^n}{2\pi i} \int_{-\infty}^{(-1-)} \frac{(1+w)^{-b}}{w^n (w-t)}\, dw.$$

If, also, $\mathrm{Re}\, b < 1$, then the loop may be collapsed on to the interval $(-\infty, -1]$ in the usual way. Subsequent replacement of w by $-w$ yields

$$\chi_n(b, t) = (-)^n \frac{\sin(\pi b)}{\pi} t^n \int_1^{\infty} \frac{(w-1)^{-b}}{w^n (w+t)}\, dw, \tag{2.5}$$

where $(w-1)^{-b}$ has its principal value.

We now substitute in (2.4) by means of (2.5), to obtain

$$R_n(a, b, z) = (-)^n \frac{\sin(\pi b)}{\pi \Gamma(a)} \int_0^{\infty e^{-i\gamma}} e^{-zt} t^{n+a-1}\, dt \int_1^{\infty} \frac{(w-1)^{-b}}{w^n (w+t)}\, dw, \tag{2.6}$$

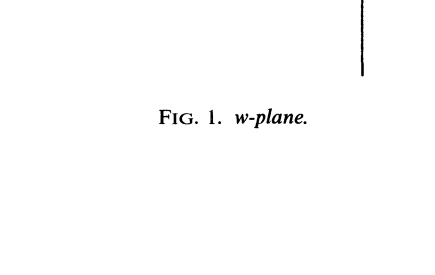


FIG. 1. *w-plane.*

where the fractional powers continue to assume their principal values. The assumptions made so far are given by $0<\operatorname{Re} a$, $-n<\operatorname{Re} b<1$, $|\operatorname{ph}(z\,e^{-i\gamma})|<\pi/2$. However, by analytic continuation with respect to a, we see that (2.4) and (2.6) remain valid when the first condition is relaxed to $-n<\operatorname{Re} a$. By absolute convergence of the repeated integrals we may interchange the order of integration (see, for example, [7, Thm. 1]). Then by replacing t by $w\tau$, we arrive at

$$R_n(a,b,z)=\frac{(-)^n\sin(\pi b)}{\pi\Gamma(a)}\int_1^\infty (w-1)^{-b}w^{a-1}\,dw\int_0^{\infty e^{-i\gamma}}\frac{e^{-zw\tau}\tau^{n+a-1}}{1+\tau}\,d\tau.$$

Provided that $\operatorname{Re} a<\operatorname{Re} b$ the order of integration may again be reversed, and on replacing w by $1+v$, we obtain the desired integral representation for the remainder term, given by

$$R_n(a,b,z)=(-)^n\frac{\sin(\pi b)}{\pi\Gamma(a)}\int_0^{\infty e^{-i\gamma}}\frac{e^{-z\tau}\tau^{n+a-1}}{1+\tau}\,d\tau\int_0^\infty e^{-z\tau v}v^{-b}(1+v)^{a-1}\,dv, \tag{2.7}$$

valid with the conditions

$$-n<\operatorname{Re} a<\operatorname{Re} b<1,\qquad |\operatorname{ph}(z\,e^{-i\gamma})|<\tfrac{1}{2}\pi. \tag{2.8}$$

Remark. A key step in the foregoing analysis is the use of Cauchy's integral formula for the remainder term in a Taylor-series expansion. Since the objective is to represent the remainder term $R_n(a,b,z)$ in the expansion (1.7) as a Stieltjes transform, applicable methods based on properties of Laplace or Stieltjes transforms, given, for example, in [11] and [12], can be used instead. This approach was used by Boyd [3] and Jones [6].

3. Proof of Theorem 1 when $|\operatorname{ph} z|\leqq\pi-\delta$. Throughout this section we continue to assume that all fractional powers take their principal values and that γ is an arbitrary constant in the interval $(-\pi,\pi)$.

Let m be an arbitrary nonnegative integer and $v\in[0,\infty)$. By Taylor's theorem we have

$$(1+v)^{a-1}=\sum_{s=0}^{m-1}(-)^s\frac{(1-a)_s}{s!}v^s+(1-a)_m v^m\phi_m(a,v), \tag{3.1}$$

where $\phi_0(a,v)=(1+v)^{a-1}$ and

$$\phi_m(a,v)=\frac{(-)^m}{(m-1)!}\int_0^1(1-t)^{m-1}(1+vt)^{a-m-1}\,dt,\qquad m\geqq 1. \tag{3.2}$$

Next, if we refer to the definition of $F_p(z)$, given by [10, (2.9)] and then apply analytic continuation with respect to z, we see that

$$F_p(z)=\frac{e^{-z}}{2\pi}\int_0^{\infty e^{-i\gamma}}\frac{e^{-zt}t^{p-1}}{1+t}\,dt,\quad \operatorname{Re} p>0,\quad |\operatorname{ph}(z\,e^{-i\gamma})|<\tfrac{1}{2}\pi. \tag{3.3}$$

Provided that $m<n+\operatorname{Re} a+\operatorname{Re} b$ we may substitute for $(1+v)^{a-1}$ by means of (3.1) in the inner integral in (2.7) and integrate term by term. On referring to (3.3) and using the reflection and recurrence formulae for the Gamma function, we arrive at (1.9) with

$$R_{m,n}(a,b,z)=\frac{z^{1-b}e^{-z}}{2\pi\Gamma(m+1-b)}\int_0^{\infty e^{-i\gamma}}\frac{e^{-z\tau}\tau^{n+a-1}}{1+\tau}\,d\tau\int_0^\infty e^{-z\tau v}v^{m-b}\phi_m(a,v)\,dv. \tag{3.4}$$

The last result has been established with the conditions $\operatorname{Re} a+\operatorname{Re} b>m-n$ and (2.8). From its definition we see that $\phi_m(a,v)$ is a continuous function of a and v and

an entire function of a. Also, if $m > \operatorname{Re} a - 1$, then $|(1+vt)^{a-m-1}| \leqq 1$, and in consequence

$$|\phi_m(a, v)| \leqq 1/m!, \qquad v \in [0, \infty), \quad m > \operatorname{Re} a - 1. \tag{3.5}$$

Hence by analytic continuation with respect to a and b we see that if $|\operatorname{ph}(z e^{-i\gamma})| < \pi/2$ and $n > m$, then (1.9) and (3.4) hold in the region defined by

$$\operatorname{Re} a < m+1, \quad \operatorname{Re} b < m+1, \quad \operatorname{Re} a + \operatorname{Re} b > m+1-n. \tag{3.6}$$

We now seek to majorize $|R_{m,n}(a, b, z)|$ for fixed values of a, b, and m, bounded values of $|\alpha|$, and large values of $\rho \equiv |z|$. Here α is defined (1.8) and m exceeds both $\operatorname{Re} a - 1$ and $\operatorname{Re} b - 1$. (The other two conditions, $n > m$ and $\operatorname{Re} a + \operatorname{Re} b > m+1-n$, are satisfied automatically when $\rho \to \infty$ with $|\alpha|$ bounded.) Also, since we are assuming throughout this section that $\operatorname{ph} z \in [-\pi+\delta, \pi-\delta]$, we may set $\gamma = \operatorname{ph} z = \theta$ without violating the condition $|\operatorname{ph}(z e^{-i\gamma})| < \pi/2$.

With $\gamma = \theta$ we have $\tau = t e^{-i\theta}$, where $0 \leqq t < \infty$. Hence (3.4) may be recast as

$$R_{m,n}(a, b, z) = \frac{e^{-(n+a)i\theta} z^{1-b} e^{-z}}{2\pi\Gamma(m+1-b)} \int_0^\infty \frac{e^{-\rho t} t^{n+a-1}}{1+t e^{-i\theta}}\, dt \int_0^\infty e^{-\rho t v} v^{m-b} \phi_m(a, v)\, dv. \tag{3.7}$$

By use of (3.5) we see that the inner integral satisfies

$$\left| \int_0^\infty e^{-\rho t v} v^{m-b} \phi_m(a, v)\, dv \right| \leqq \frac{\Gamma(m+1-\operatorname{Re} b)}{m!(\rho t)^{m+1-\operatorname{Re} b}}. \tag{3.8}$$

For the outer integral we have

$$|1+t e^{-i\theta}|^{-1} \leqq \operatorname{cosec} \delta. \tag{3.9}$$

Hence

$$|R_{m,n}(a, b, z)| \leqq \frac{e^{\pi(|\operatorname{Im} a| + |\operatorname{Im} b|)} \Gamma(m+1-\operatorname{Re} b)}{2\pi \sin\delta\, m!\, |\Gamma(m+1-b)|} \frac{\Gamma(n-m+\operatorname{Re} a+\operatorname{Re} b-1)|e^{-z}|}{\rho^{n+\operatorname{Re} a+\operatorname{Re} b-1}}. \tag{3.10}$$

Substituting for n by means of (the real part of) (1.8) and using Stirling's formula, we conclude that as $\rho \to \infty$ the right member of this inequality is $O\{e^{-z-|z|}/z^{m+1/2}\}$ uniformly for bounded $|\alpha|$; compare (1.10).

Lastly, the conditions $m > \operatorname{Re} a - 1$ and $m > \operatorname{Re} b - 1$ may be removed in the usual way by replacing m in (1.9) by a sufficiently large integer and referring to [10, (2.14)].[1] This completes the proof of Theorem 1 in the case where $|\operatorname{ph} z| \leqq \pi - \delta$; indeed, we have a slightly stronger asymptotic estimate for $R_{m,n}(a, b, z)$—by a factor $|z|^{-1/2}$—for this sector.

4. Proof of Theorem 1 when $\delta \leqq \operatorname{ph} z \leqq 2\pi - \delta$. We return to the representation of $R_n(a, b, z)$ furnished by (1.9) and (3.4), and valid with the conditions $|\gamma| < \pi$, $|\operatorname{ph}(z e^{-i\gamma})| < \pi/2$, $n > m$, and (3.6). The natural way to try to extend the proof of § 3 to larger values of $\operatorname{ph} z$ would be to rotate further the path of the outer integral in (3.4), that is, to take values of γ equaling or exceeding π. The difficulty is that as $\gamma \to \pi$, the integration path approaches the singularity of the integrand at $\tau = -1$. To overcome this difficulty we adopt, temporarily, the extra conditions $0 < \gamma < \pi$, $\pi/2 < \operatorname{ph} z < \frac{3}{2}\pi$, and decompose $R_{m,n}(a, b, z)$ as follows:

$$R_{m,n}(a, b, z) = R^{(1)}_{m,n}(a, b, z) + R^{(2)}_{m,n}(a, b, z), \tag{4.1}$$

[1] Compare also (4.5) below.

where

$$R^{(1)}_{m,n}(a,b,z)=\frac{z^{1-b}e^{-z}}{2\pi\Gamma(m+1-b)}\int_0^{\infty e^{-i\gamma}}\frac{e^{-z\tau}\tau^{n+a-1}}{1+\tau}\,d\tau \times\int_0^{\infty}e^{zv}v^{m-b}\phi_m(a,v)\,dv, \tag{4.2}$$

$$R^{(2)}_{m,n}(a,b,z)=\frac{z^{1-b}e^{-z}}{2\pi\Gamma(m+1-b)}\int_0^{\infty e^{-i\gamma}}e^{-z\tau}\tau^{n+a-1}\,d\tau \times\int_0^{\infty}\frac{e^{-z\tau v}-e^{zv}}{1+\tau}v^{m-b}\phi_m(a,v)\,dv. \tag{4.3}$$

The right member of (4.2) may be regarded as the product of two single integrals, instead of as a double integral. The first of these is identifiable as an F_p function; the second may be continued analytically by rotating the path of integration through an arbitrary angle $-\hat{\gamma}$, say, where $\hat{\gamma}\in(-\pi,\pi)$. With the aid of (3.3) we obtain

$$R^{(1)}_{m,n}(a,b,z)=\frac{z^{1-b}}{\Gamma(m+1-b)}F_{n+a}(z)\int_0^{\infty e^{-i\hat{\gamma}}}e^{zv}v^{m-b}\phi_m(a,v)\,dv, \tag{4.4}$$

valid with the conditions $\pi/2<|\mathrm{ph}\,(z\,e^{-i\hat{\gamma}})|<\frac{3}{2}\pi$ and (3.6). (The conditions $|\mathrm{ph}\,(z\,e^{-i\gamma})|<\pi/2$ and $\pi/2<\mathrm{ph}\,z<\frac{3}{2}\pi$ disappear.) From Theorem 1 of [10], we obtain[2]

$$F_{n+a}(z)=O(e^{-z-|z|}),\qquad |\mathrm{ph}\,z|\leqq\pi, \tag{4.5}$$

$$F_{n+a}(z)=O(1),\qquad \pi\leqq|\mathrm{ph}\,z|\leqq3\pi-\delta. \tag{4.6}$$

Next, in this section we consider only the sector $\delta\leqq\mathrm{ph}\,z\leqq2\pi-\delta$; hence we may set $\hat{\gamma}=\theta-\pi$. On replacing v by $v\,e^{i(\pi-\theta)}$ we derive

$$\int_0^{\infty e^{-i\hat{\gamma}}}e^{zv}v^{m-b}\phi_m(a,v)\,dv=e^{i(\pi-\theta)(m-b+1)}\int_0^{\infty}e^{-|z|v}v^{m-b}\phi_m(a,v\,e^{i(\pi-\theta)})\,dv. \tag{4.7}$$

One of our assumptions is that $\mathrm{Re}\,a<m+1$; accordingly, from (3.2) we perceive that

$$|\phi_m(a,v\,e^{i(\pi-\theta)})|\leqq A_m, \tag{4.8}$$

where A_m is assignable independently of $\theta\in[\delta,2\pi-\delta]$ and $v\in[0,\infty)$. In consequence, also, of the second condition of (3.6) it follows that

$$\int_0^{\infty e^{-i\hat{\gamma}}}e^{zv}v^{m-b}\phi_m(a,v)\,dv=O(z^{b-m-1}).$$

Then combining this result with (4.4)-(4.6), we arrive at

$$R^{(1)}_{m,n}(a,b,z)=\begin{cases}O(e^{-z-|z|}z^{-m}), & \delta\leqq\mathrm{ph}\,z\leqq\pi,\\ O(z^{-m}), & \pi\leqq\mathrm{ph}\,z\leqq2\pi-\delta.\end{cases} \tag{4.9}$$

For the contribution of $R^{(2)}_{m,n}(a,b,z)$, we first rotate the path for the inner integral in (4.3) through an angle $(\pi-\gamma)/2$. This yields

$$R^{(2)}_{m,n}(a,b,z)=\frac{z^{1-b}e^{-z}}{2\pi\Gamma(m+1-b)}\int_0^{\infty e^{-i\gamma}}e^{-z\tau}\tau^{n+a-1}\,d\tau \times\int_0^{\infty e^{i(\pi-\gamma)/2}}\frac{e^{-z\tau v}-e^{zv}}{1+\tau}v^{m-b}\phi_m(a,v)\,dv, \tag{4.10}$$

[2] In deriving (4.5) and (4.6) from the cited theorem, we have used the following facts: $\mathrm{erfc}\,(\xi)=O(e^{-\xi^2})$ throughout the sector $|\mathrm{ph}\,\xi|\leqq\pi/4$; $\mathrm{erfc}\,(\xi)=O(1)$ throughout the sector $|\mathrm{ph}\,(-\xi)|\leqq\pi/4$ (compare [10, § 4.4]).

valid when $0<\gamma<\pi$, $|\mathrm{ph}\,(z\,e^{-i\gamma})|<\pi/2$, $|\mathrm{ph}\,(z\,e^{i(\pi-3\gamma)/2})|<\pi/2$, and $|\mathrm{ph}\,(z\,e^{-i(\pi+\gamma)/2})|<\pi/2$, that is, when

$$0<\gamma<\pi \quad \text{and} \quad \tfrac{1}{2}\gamma<\mathrm{ph}\,z<\tfrac{3}{2}\gamma. \tag{4.11}$$

Furthermore, because the integrand is now free from singularity at $\tau=-1$ we may extend the rotation of each integration path by increasing γ. Thus (4.10) is also valid with the conditions

$$\pi\leqq\gamma<2\pi \quad \text{and} \quad \tfrac{3}{2}\gamma-\pi<\mathrm{ph}\,z<\pi+\tfrac{1}{2}\gamma. \tag{4.12}$$

In accordance with (4.11) we may set $\gamma=\theta$ $(=\mathrm{ph}\,z)$ in (4.10) when $\delta\leqq\mathrm{ph}\,z<\pi$. Alternatively, if $\pi\leqq\mathrm{ph}\,z\leqq2\pi-\delta$, then we may set $\gamma=\theta$ in (4.10) in accordance with (4.12). In either event we find that

$$\begin{aligned} R^{(2)}_{m,n}(a,b,z)=&\frac{z^{1-b}\,e^{-z}}{2\pi\Gamma(m+1-b)}\int_0^{\infty e^{-i\theta}} e^{-z\tau}\tau^{n+a-1}\,d\tau \\ &\times\int_0^{\infty e^{i(\pi-\theta)/2}}\frac{e^{-z\tau v}-e^{zv}}{1+\tau}\,v^{m-b}\phi_m(a,v)\,dv, \end{aligned} \tag{4.13}$$

valid when $\delta\leqq\mathrm{ph}\,z\leqq2\pi-\delta$.

Once again, subject to (3.6) we seek to estimate $R^{(2)}_{m,n}(a,b,z)$ for fixed values of a, b, and m, bounded values of $|\alpha|$, and large values of $\rho\equiv|z|$.

We first observe that the integration variable of the inner integral in (4.13) lies within the sector $|\mathrm{ph}\,v|<\pi/2$. Hence

$$|\phi_m(a,v)|\leqq e^{\pi|\mathrm{Im}\,a|/2}/m!; \tag{4.14}$$

compare (3.5) and (4.8).

Next, we have identically

$$\frac{e^{-z\tau v}-e^{zv}}{1+\tau}=zv\,e^{-z\tau v}\frac{1-e^{zv(1+\tau)}}{zv(1+\tau)}=zv\,e^{zv}\frac{e^{-zv(1+\tau)}-1}{zv(1+\tau)}.$$

Furthermore, by application of the maximum-modulus theorem it is easily verified that

$$\left|\frac{1-e^{-\zeta}}{\zeta}\right|\leqq1, \qquad \mathrm{Re}\,\zeta\geqq0. \tag{4.15}$$

Combination of the last two results yields

$$\left|\frac{e^{-z\tau v}-e^{zv}}{1+\tau}\right|\leqq\begin{cases}|zv\,e^{-z\tau v}|, & \mathrm{Re}\,\{zv(1+\tau)\}\leqq0,\\ |zv\,e^{zv}|, & \mathrm{Re}\,\{zv(1+\tau)\}\geqq0.\end{cases}$$

With $\mathrm{ph}\,\tau=-\theta$ and $\mathrm{ph}\,v=\pi/2-\theta/2$, we have

$$\mathrm{Re}\,\{zv(1+\tau)\}=\rho|v|(|\tau|-1)\sin\tfrac{1}{2}\theta.$$

Accordingly,

$$\mathrm{Re}\,\{zv(1+\tau)\}\gtrless0 \quad \text{according as} \quad |\tau|\gtrless1.$$

Therefore when $0<|\tau|\leqq1$ we have

$$\begin{aligned} &\left|\int_0^{\infty e^{i(\pi-\theta)/2}}\frac{e^{-z\tau v}-e^{zv}}{1+\tau}\,v^{m-b}\phi_m(a,v)\,dv\right| \\ &\leqq\frac{e^{\pi|\mathrm{Im}\,a|/2}\,e^{(\pi-\theta)\mathrm{Im}\,b/2}\rho}{m!}\int_0^{\infty}e^{-\rho|\tau v|\sin\theta/2}|v|^{m-\mathrm{Re}\,b+1}\,d|v| \\ &\leqq\frac{e^{\pi(|\mathrm{Im}\,a|+|\mathrm{Im}\,b|)/2}\rho}{m!}\,\frac{\Gamma(m+2-\mathrm{Re}\,b)}{(\rho|\tau|\sin\frac{1}{2}\delta)^{m+2-\mathrm{Re}\,b}}. \end{aligned} \tag{4.16}$$

Similarly, when $|\tau| \geqq 1$ we find that

$$\left| \int_0^{\infty e^{i(\pi-\theta)/2}} \frac{e^{-z\tau v} - e^{zv}}{1+\tau} v^{m-b} \phi_m(a, v)\, dv \right| \leqq \frac{e^{\pi(|\operatorname{Im} a|+|\operatorname{Im} b|)/2} \rho}{m!} \frac{\Gamma(m+2-\operatorname{Re} b)}{(\rho \sin \frac{1}{2}\delta)^{m+2-\operatorname{Re} b}}. \tag{4.17}$$

We now prepare to substitute in (4.13) by means of the inequalities just obtained. At this stage it is convenient to use the symbol A_m generically to denote a quantity that may depend on m, a, and b, but is independent of n, z and (bounded values of) α; compare, also, (4.8). In (4.13) we have

$$|\tau^{n+a-1}| \leqq A_m |\tau|^{\rho+\operatorname{Re}\alpha-\operatorname{Re} b};$$

compare (1.8). Hence

$$|R^{(2)}_{m,n}(a, b, z)| \leqq A_m \frac{|e^{-z}|}{\rho^m} \left\{ \int_0^1 e^{-\rho|\tau|} |\tau|^{\rho+\operatorname{Re}\alpha-m-2}\, d|\tau| + \int_1^{\infty} e^{-\rho|\tau|} |\tau|^{\rho+\operatorname{Re}\alpha-\operatorname{Re} b}\, d|\tau| \right\}.$$

Replacing both integration ranges by $(0, \infty)$ and subsequently using Stirling's formula, we derive

$$|R^{(2)}_{m,n}(a, b, z)| \leqq A_m \frac{|e^{-z}|}{\rho^m} \left\{ \frac{\Gamma(\rho+\operatorname{Re}\alpha-m-1)}{\rho^{\rho+\operatorname{Re}\alpha-m-1}} + \frac{\Gamma(\rho+\operatorname{Re}\alpha+1-\operatorname{Re} b)}{\rho^{\rho+\operatorname{Re}\alpha+1-\operatorname{Re} b}} \right\} = O\left(\frac{e^{-z-|z|}}{z^{m+1/2}} \right). \tag{4.18}$$

This result is valid when $\delta \leqq \operatorname{ph} z \leqq 2\pi - \delta$, and on combining it with (4.1) and (4.9) we arrive at

$$R_{m,n}(a, b, z) = \begin{cases} O(e^{-z-|z|} z^{-m}), & \delta \leqq \operatorname{ph} z \leqq \pi, \\ O(z^{-m}), & \pi \leqq \operatorname{ph} z \leqq 2\pi - \delta. \end{cases}$$

After removing the restrictions (3.6) on m in the same manner as that used at the end of § 3 (compare (4.5) and (4.6)), we see that the proof of Theorem 1 is complete for the sector $\delta \leqq \operatorname{ph} z \leqq 2\pi - \delta$.

5. Completion of the proof of Theorem 1. The next extension is to the sector $\frac{3}{2}\pi + \delta \leqq \operatorname{ph} z \leqq \frac{5}{2}\pi - \delta$. Rather than attempting to extend the method of proof introduced in § 4, we appeal to the connection formulae (1.12). An advantage of this approach is that it also determines the maximum regions of validity.

From (1.12) with $k = 1$, z replaced by $z e^{-2\pi i}$ and upper signs taken, we have

$$U(a, a-b+1, z) = e^{-2a\pi i} U(a, a-b+1, z e^{-2\pi i}) + 2\pi i \frac{e^{-a\pi i}}{\Gamma(a)\Gamma(b)} e^{z} U(1-b, a-b+1, z e^{-\pi i}). \tag{5.1}$$

Substituting in each side of this equation by means of (1.7) and (1.9), then using the identity

$$F_{n-s+a+b-1}(z e^{-2\pi i}) = (-)^{n+s} i\, e^{(a+b)\pi i} + e^{2(a+b)\pi i} F_{n-s+a+b-1}(z) \tag{5.2}$$

obtained from [10], (2.11), and carrying out some reduction, we eventually arrive at

$$R_{m,n}(a,b,z)=e^{-2(a+b)\pi i}R_{m,n}(a,b,z\,e^{-2\pi i})+R^{(3)}_{m,n}(a,b,z)+R^{(4)}_{m,n}(a,b,z), \tag{5.3}$$

where

$$R^{(3)}_{m,n}(a,b,z)=(-)^{n-1}i\,e^{-(a+b)\pi i}\sum_{s=m}^{n-1}\frac{(m+1-a)_{s-m}(m+1-b)_{s-m}}{s!\,z^s}, \tag{5.4}$$

and

$$\begin{aligned}R^{(4)}_{m,n}(a,b,z)=2\pi i&\frac{z^{1-a-b}\,e^{-z}}{\Gamma(m+1-a)\Gamma(m+1-b)}\\ &\cdot\left\{\sum_{s=0}^{m-1}\frac{(a)_s(b)_s}{s!}\frac{F_{n-s-a-b+1}(z\,e^{-\pi i})}{z^s}\right.\\ &\left.+(a)_m(b)_m R_{m,n}(1-b,1-a,z\,e^{-\pi i})\right\}.\end{aligned} \tag{5.5}$$

Since $\frac{3}{2}\pi+\delta\leqq\operatorname{ph}z\leqq\frac{5}{2}\pi-\delta$, we have $|\operatorname{ph}(z\,e^{-2\pi i})|\leqq\pi/2-\delta$; hence from the results of § 3 it follows that

$$R_{m,n}(a,b,z\,e^{-2\pi i})=O(e^{-z-|z|}z^{-m-1/2}). \tag{5.6}$$

Next, although the number of terms in the sum in (5.4) increases with $|z|$, it is routine to verify that

$$R^{(3)}_{m,n}(a,b,z)=(-)^{n-1}\frac{i\,e^{-(a+b)\pi i}}{m!\,z^m}+O\left(\frac{1}{z^{m+1}}\right). \tag{5.7}$$

For $R^{(4)}_{m,n}(a,b,z)$ we have $\frac{1}{2}\pi+\delta\leqq\operatorname{ph}(z\,e^{-\pi i})\leqq\frac{3}{2}\pi-\delta$; accordingly, from the results of § 4 it follows that

$$R_{m,n}(1-b,1-a,z\,e^{-\pi i})=O(z^{-m}),$$

and

$$F_{n-s-a-b+1}(z\,e^{-\pi i})=O(1),$$

for each s. Hence

$$R^{(4)}_{m,n}(a,b,z)=O(z^{1-a-b}\,e^{-z}). \tag{5.8}$$

On substituting in (5.3) by means of (5.6), (5.7), and (5.8) we conclude that

$$R_{m,n}(a,b,z)=O(z^{-m}), \tag{5.9}$$

as required. Then combining this estimate with the results of §§ 3 and 4 we establish (1.10) in the sector $-\pi+\delta\leqq\operatorname{ph}z\leqq\pi$, and (1.11) in the sector $\pi\leqq\operatorname{ph}z\leqq\frac{5}{2}\pi-\delta$. A similar proof holds for the conjugate sector, or we can appeal to symmetry.

To complete the proof of Theorem 1, we have only to demonstrate that the estimate (1.10) breaks down after we cross the rays $\operatorname{ph}z=\pm\pi$, and the estimate (1.11) breaks down after we cross $\operatorname{ph}z=\pm\frac{5}{2}\pi$.

In the case where $\operatorname{ph}z=\pi$, we observe from (4.4) and (4.7) that as $z\to\infty$ in the sector $\delta\leqq\operatorname{ph}z\leqq2\pi-\sigma$, we have

$$R^{(1)}_{m,n}(a,b,z)\sim(-)^{m-1}\,e^{-b\pi i}\phi_m(a,0)\frac{F_{n+a}(z)}{z^m}=-\frac{e^{-b\pi i}F_{n+a}(z)}{m!\,z^m};$$

compare also (3.2). And when $\pi+\delta\leqq\operatorname{ph}z\leqq2\pi-\delta$, $(-)^nF_{n+a}(z)$ tends to the constant $i\,e^{-a\pi i}$ (compare Theorem 1 of [10]), giving $R^{(1)}_{m,n}(a,b,z)\sim(\text{constant})z^{-m}$. Obviously from (4.18) this algebraic behavior cannot be cancelled in (4.1) by the contribution of $R^{(2)}_{m,n}(a,b,z)$. A similar argument applies to the ray $\operatorname{ph}z=-\pi$.

Lastly, after we cross $\operatorname{ph}z=\frac{5}{2}\pi$, the contribution of $R^{(4)}_{m,n}(a,b,z)$ in (5.3) is no longer absorbable in the estimate $O(z^{-m})$ because it becomes exponentially large. On

the other hand, the term $e^{-2(a+b)\pi i}R_{m,n}(a, b, z\,e^{-2\pi i})$ remains exponentially small and $R^{(3)}_{m,n}(a, b, z)$ remains $O(z^{-m})$. Similarly for the ray ph $z = -\frac{5}{2}\pi$.

6. Error bounds. For computational purposes the use of (1.7) and (1.9) should be confined to the sector $|\text{ph } z| \leqq \pi$ in which the error term has undergone maximum exponential improvement; elsewhere continuation formulae are to be applied. As always, it is highly desirable to supply a rigorous and realistic bound for the error term in any asymptotic approximation. The analyses supplied in §§ 3-5 lend themselves only in part to achieving this objective.

In § 3, for example, the bound (3.10) breaks down as $\delta \to 0$. However, a slight modification of the analysis leading to this result yields a satisfactory bound when $|\text{ph } z| \leqq \pi/2$: in this sector the right-hand side of (3.9) is replaceable by unity, thus eliminating the factor $(\sin \delta)^{-1}$ from (3.10). Similarly, if we restrict ph $z \in [\pi/2, \pi]$, then the bounds derived in § 4 can be tightened; however, the factor $(\text{cosec } \delta/2)^{m+2-\text{Re}\,b}$ that appears in (4.16) and (4.17) can be replaced only by $(\sqrt{2})^{m+2-\text{Re}\,b}$, which is a drawback.

In this section, we modify the analysis of § 4 in a more substantial manner in order to arrive at a sharper bound for $R_{m,n}(a, b, z)$ in the sector $|\text{ph } z| \leqq \pi$. This modification is based partly on the analysis of Boyd [3] for constructing error bounds for a similar expansion for the modified Bessel function $K_\nu(z)$. The bounds we shall derive are similar to, but not the same as, those of Boyd when the parameters a and b (and ipso facto ν) are real, and they are more readily computable when a, b (and ν) are complex.

As before, we set $z = \rho\, e^{i\theta}$. Throughout the following analysis we restrict $|\theta| < \pi$, relaxing this condition to $|\theta| \leqq \pi$ at the close by appeal to continuity.

In (3.4) we take $\gamma = \theta$, and decompose $R_{m,n}(a, b, z)$ in a manner slightly different from that given by (4.1)-(4.3). We write

$$R_{m,n}(a, b, z) = S^{(1)}_{m,n}(a, b, z) + S^{(2)}_{m,n}(a, b, z), \tag{6.1}$$

where

$$S^{(1)}_{m,n}(a, b, z) = \frac{z^{1-b}\, e^{-z}}{2\pi\Gamma(m+1-b)} \int_0^{\infty e^{-i\theta}} \frac{e^{-z\tau}\tau^{n+a-1}}{1+\tau}\, d\tau \int_0^\infty e^{-\rho v} v^{m-b} \phi_m(a, v)\, dv, \tag{6.2}$$

$$\begin{aligned} S^{(2)}_{m,n}(a, b, z) = {} & \frac{z^{1-b}\, e^{-z}}{2\pi\Gamma(m+1-b)} \int_0^{\infty e^{-i\theta}} e^{-z\tau}\tau^{n+a-1}\, d\tau \\ & \cdot \int_0^\infty \frac{e^{-z\tau v} - e^{-\rho v}}{1+\tau} v^{m-b} \phi_m(a, v)\, dv, \end{aligned} \tag{6.3}$$

these equations being valid with the conditions $|\theta| < \pi$ and (3.6).

As in § 4, the first integral in (6.2) may be identified in terms of $F_{n+a}(z)$. For the second integral, we utilize the bound (3.5) for $\phi_m(a, v)$. Thus we derive

$$|S^{(1)}_{m,n}(a, b, z)| \leqq \frac{e^{\theta\,\text{Im}\,b}\Gamma(m+1-\text{Re}\,b)}{m!|\Gamma(m+1-b)|} \frac{|F_{n+a}(z)|}{\rho^m}. \tag{6.4}$$

In (6.3) we set $\tau = t\, e^{-i\theta}$; thus

$$\begin{aligned} S^{(2)}_{m,n}(a, b, z) = {} & \frac{e^{-(n+a)i\theta}}{2\pi\Gamma(m+1-b)} z^{1-b}\, e^{-z} \int_0^\infty e^{-\rho t} t^{n+a-1}\, dt \\ & \cdot \int_0^\infty \frac{e^{-\rho t v} - e^{-\rho v}}{1+t\, e^{-i\theta}} v^{m-b} \phi_m(a, v)\, dv. \end{aligned} \tag{6.5}$$

To bound the inner integral, the inequalities

$$0 \leqq \frac{e^{-\rho t v} - e^{-\rho v}}{1-t} \leqq \begin{cases} \rho v\, e^{-\rho t v}, & 0 \leqq t \leqq 1, \\ \rho v\, e^{-\rho v}, & t \geqq 1, \end{cases} \tag{6.6}$$

and

$$\left| \frac{1-t}{1+t\,e^{-i\theta}} \right| \leqq 1 \tag{6.7}$$

are applied: (6.6) is easily verified with the aid of (4.15), and (6.7) is verifiable by considering the bilinear transformation $T = (1-t)/(1+t\,e^{-i\theta})$; compare [3, § 4]. Referring again to (3.5), we conclude that

$$\left| \int_0^\infty \frac{e^{-\rho t v} - e^{-\rho v}}{1+t\,e^{-i\theta}}\, v^{m-b} \phi_m(a, v)\, dv \right| \leqq \frac{\rho}{m!} \frac{\Gamma(m+2-\operatorname{Re} b)}{(\rho t_1)^{m+2-\operatorname{Re} b}},$$

where $t_1 = \min(t, 1)$. If we further restrict $\operatorname{Re} a + \operatorname{Re} b > m+2-n$ (compare (3.6)), then we may substitute this bound in (6.5). Referring once again to (1.8), we find that

$$\begin{aligned} |S^{(2)}_{m,n}(a, b, z)| \leqq{}& \frac{e^{\theta(\operatorname{Im} a + \operatorname{Im} b)} \Gamma(m+2-\operatorname{Re} b)}{2\pi\, m!\, |\Gamma(m+1-b)|} \frac{|e^{-z}|}{\rho^m} \\ & \cdot \left\{ \int_0^1 e^{-\rho t} t^{\rho - m + \operatorname{Re}\alpha - 2}\, dt + \int_1^\infty e^{-\rho t} t^{\rho - \operatorname{Re} b + \operatorname{Re}\alpha}\, dt \right\}. \end{aligned} \tag{6.8}$$

Inequalities (6.4) and (6.8) furnish the desired upper bound on $|R_{m,n}(a, b, z)|$ via (6.1). The condition $|\theta| < \pi$ may now be eased by continuity considerations. Thus the aggregate conditions on (6.4) and (6.8) are given by $|\theta| \leqq \pi$ and

$$\operatorname{Re} a < m+1, \quad \operatorname{Re} b < m+1, \quad \operatorname{Re} a + \operatorname{Re} b > m+2-n. \tag{6.9}$$

It remains to bound the integrals that appear in (6.8). A simple approach is to extend each integration range to $[0, \infty)$. This immediately yields

$$\int_0^1 e^{-\rho t} t^{\rho - m + \operatorname{Re}\alpha - 2}\, dt < \frac{\Gamma(\rho - m + \operatorname{Re}\alpha - 1)}{\rho^{\rho - m + \operatorname{Re}\alpha - 1}},$$

$$\int_1^\infty e^{-\rho t} t^{\rho - \operatorname{Re} b + \operatorname{Re}\alpha}\, dt < \frac{\Gamma(\rho - \operatorname{Re} b + \operatorname{Re}\alpha + 1)}{\rho^{\rho - \operatorname{Re} b + \operatorname{Re}\alpha + 1}}.$$

However, each of these bounds overestimates the actual value of the corresponding left-hand side by a factor that approaches two as $\rho \to \infty$. More sophisticated bounds that are free from this defect are supplied in the Appendix.

7. Conclusions. We have constructed a uniform generalized asymptotic expansion for the confluent hypergeometric function $U(a, a-b+1, z)$, valid when the parameters a and b are fixed and z tends to infinity in the sector $|\operatorname{ph} z| \leqq \frac{5}{2}\pi - \delta$, where δ is an arbitrarily small positive constant. This expansion is in terms of generalized exponential integrals (or equivalently incomplete Gamma functions) and it includes as special cases three well-known expansions of Poincaré type valid in the ph z intervals $[-\frac{5}{2}\pi + \delta, -\frac{1}{2}\pi - \delta]$, $[-\frac{3}{2}\pi + \delta, \frac{3}{2}\pi - \delta]$, $[\frac{1}{2}\pi + \delta, \frac{5}{2}\pi - \delta]$. It also quantifies accurately the rapid, but smooth, transitions between these expansions in the neighborhoods of the Stokes lines ph $z = \pm\pi$. These smooth transitions were first described by M. V. Berry using formal analysis and error functions as approximants. The present results serve to place

Berry's conclusions on a rigorous foundation in the case of the confluent hypergeometric function. However, it should be noted that Berry's analysis applied only in the neighborhoods of the Stokes lines, whereas our result is uniformly valid in a sector of total angle $5\pi - 2\delta$.

Our new expansion enjoys the important property that it is a uniform, exponentially improved expansion throughout the sector $|\text{ph } z| \leqq \pi$, in the sense described in [10, § 1]. Furthermore, we have derived strict and realistic bounds for the error term in this sector. Uniform, exponentially improved expansions in other sectors can be easily constructed with the aid of standard connection formulae, and normally we need not compute the new expansion outside the sector $|\text{ph } z| \leqq \pi$.

Although we have treated only the confluent hypergeometric function in the present paper, our method is applicable to a large class of functions that can be represented by Laplace transforms.[3] The basic steps are as follows: (i) The remainder term of the Maclaurin expansion of the kernel of the transform is represented as a contour integral by means of Cauchy's integral formula. (ii) The integral obtained in step (i) is used to construct a Stieltjes transform in the form of a double integral for the remainder term associated with the desired asymptotic expansion. (iii) The double integral is expanded in a series of generalized exponential integrals. (iv) The region of validity is extended by suitable rotations of integration paths and use of continuation formulae. The only step in which we made use of special properties of the confluent hypergeometric function was (iv), where we relied on continuation formulae. However, this was largely a matter of convenience, and in any case this step is the least important.

In Berry's theory the conditions adopted ensure that the function whose asymptotic expansions are under investigation is representable as a Laplace transform of the type admitted above. The scope of the present method therefore appears to be essentially the same as that of Berry.

Appendix. Bounds for some incomplete Gamma functions. In this appendix we construct bounds for the incomplete Gamma functions

$$I_\lambda^{(1)}(\rho) = \int_1^\infty e^{-\rho t} t^{\rho+\lambda-1}\, dt, \qquad I_\lambda^{(2)}(\rho) = \int_0^1 e^{-\rho t} t^{\rho+\lambda-1}\, dt, \tag{A1}$$

that have the same asymptotic forms as the functions themselves when $\rho \to \infty$ with λ positive and fixed. These results were referred to at the end of § 6.

LEMMA. *When λ and ρ are positive*

$$I_\lambda^{(1)}(\rho) \leqq \sqrt{\frac{\pi}{2\rho}}\, e^{-\rho} \exp\left(\frac{\kappa_1^2}{2\rho}\right)\left\{1 + \operatorname{erf}\left(\frac{\kappa_1}{\sqrt{2\rho}}\right)\right\}, \tag{A2}$$

$$I_\lambda^{(2)}(\rho) \leqq \sqrt{\frac{\pi}{2\rho}}\, e^{-\rho} \exp\left(\frac{\kappa_2^2}{2\rho}\right)\left\{1 + \operatorname{erf}\left(\frac{\kappa_2}{\sqrt{2\rho}}\right)\right\}, \tag{A3}$$

where

$$\kappa_1 = \sup_{\tau \in (0,\infty)} \omega(\tau), \qquad \kappa_2 = \sup_{\tau \in (-1,0)} \omega(\tau), \tag{A4}$$

[3] This includes, for example, the integral $\int_0^\infty e^{-zt} t^{d-1} F(a, b; c; -t)\, dt$, in which a, b, c, d are parameters. This is a generalized hypergeometric function of ${}_2F_2$ type.

and[4]

$$\omega(\tau)=\frac{1}{\sqrt{2\tau-2\ln(1+\tau)}}\ln\left\{\frac{(1+\tau)^{\lambda}}{|\tau|}\sqrt{2\tau-2\ln(1+\tau)}\right\}. \tag{A5}$$

To establish this result, we first observe that

$$\omega(\tau)\to 0,\quad \tau\to+\infty;\qquad \omega(\tau)\to\lambda-\tfrac{1}{3},\quad \tau\to 0+;$$
$$\omega(\tau)\to-\infty,\quad \tau\to-1+;\qquad \omega(\tau)\to\tfrac{1}{3}-\lambda,\quad \tau\to 0-.$$

Hence κ_1 and κ_2 are both finite; furthermore, $\kappa_1\geqq\max(\lambda-\frac{1}{3},0)$ and $\kappa_2\geqq\frac{1}{3}-\lambda$.

In the definition (A1) of $I_{\lambda}^{(1)}(\rho)$ let us replace t by $1+\tau$, and then set

$$\tfrac{1}{2}v^2=\tau-\ln(1+\tau), \tag{A6}$$

with $v\sim\tau$ when $\tau\to 0+$. We find that

$$I_{\lambda}^{(1)}(\rho)=e^{-\rho}\int_0^{\infty}\exp\left(-\frac{1}{2}\rho v^2\right)f(v)\,dv, \tag{A7}$$

where

$$f(v)=(1+\tau)^{\lambda-1}\frac{d\tau}{dv}=(1+\tau)^{\lambda}\frac{v}{\tau}.$$

Clearly, $f(v)$ is positive when $v\in(0,\infty)$, and tends to unity when $v\to 0+$. Furthermore, from (A4), (A5), and (A6) we see that

$$f(v)\leqq e^{\kappa_1 v},\qquad 0<v<\infty. \tag{A8}$$

Substitution of this bound into (A7) leads to (A2).

The proof of (A3) is similar.

We note that by applying Laplace's method [8, Chap. 3], to the integrals in the definitions (A1) we derive

$$I_{\lambda}^{(1)}(\rho),\ I_{\lambda}^{(2)}(\rho)\sim\sqrt{\frac{\pi}{2\rho}}\,e^{-\rho},\quad \rho\to\infty.$$

Obviously, the right-hand sides of (A2) and (A3) also have this asymptotic form.

We also observe that for computational purposes the bounds (A2) and (A3) are simplifiable, with an insignificant loss of sharpness when ρ is large. This is achieved via the inequality $\operatorname{erf}(t)\leqq 2t/\sqrt{\pi}$, which is valid for all nonnegative values of t. In this way we derive

$$\begin{aligned} I_{\lambda}^{(1)}(\rho)&\leqq\sqrt{\frac{\pi}{2\rho}}\,e^{-\rho}\exp\left(\frac{\kappa_1^2}{2\rho}\right)\left(1+\sqrt{\frac{2}{\pi\rho}}\,\kappa_1\right)\\ &<\sqrt{\frac{\pi}{2\rho}}\,e^{-\rho}\exp\left(\frac{2\kappa_1^2}{\pi\rho}\right)\left(1+\sqrt{\frac{2}{\pi\rho}}\,\kappa_1\right)<\frac{\pi\,e^{-\rho}}{\sqrt{2\pi\rho}-2\kappa_1}, \end{aligned} \tag{A9}$$

provided that $\rho>2\kappa_1^2/\pi$. (The last step is a consequence of the inequality $(1+t)\,e^{t^2}<(1-t)^{-1}$, $0<t<1$.) The corresponding simplification for $I_{\lambda}^{(2)}(\rho)$ is found to be

$$I_{\lambda}^{(2)}(\rho)<\frac{\pi\,e^{-\rho}}{\sqrt{2\pi\rho}-2\kappa_2},\qquad \rho>\frac{2\kappa_2^2}{\pi}, \tag{A10}$$

[4] In (A5) the positive square root must be taken in both instances. Thus $\sqrt{2\tau-2\ln(1+\tau)}\sim|\tau|$ as $\tau\to 0$ from above or below.

provided that $\kappa_2 > 0$. If $\kappa_2 \leqq 0$, then we have

$$I_\lambda^{(2)}(\rho) \leqq \sqrt{\frac{\pi}{2\rho}}\, e^{-\rho}; \tag{A11}$$

compare (A7) and (A8).

Remark. The method that we have used in this Appendix may also be applied to the general problem of bounding error terms that arise with Watson's lemma and the method of steepest descents. It enjoys some advantages over that described in [8, Chap. 3, § 9; Chap. 4, § 10].

Acknowledgment. The writer is pleased to acknowledge helpful discussions with W. G. C. Boyd.

REFERENCES

[1] L. V. AHLFORS, *Complex Analysis*, Second ed., McGraw-Hill, New York, 1966.
[2] M. V. BERRY, *Uniform asymptotic smoothing of Stokes' discontinuities*, Proc. Roy. Soc. London Ser. A, 422 (1989), pp. 7-21.
[3] W. G. C. BOYD, *Stieltjes transforms and the Stokes phenomenon*, Proc. Roy. Soc. London Ser. A, 429 (1990), pp. 227-246.
[4] H. BUCHHOLZ, *The Confluent Hypergeometric Function* (translated by H. Lichtblau and K. Wetzel from 1953 German ed.), Springer Tracts in Natural Philosophy, Vol. 15, Springer-Verlag, Berlin, New York, 1969.
[5] R. B. DINGLE, *Asymptotic expansions and converging factors* IV. *Confluent hypergeometric, parabolic cylinder, modified Bessel, and ordinary Bessel functions*, Proc. Roy. Soc. London Ser. A, 249 (1958), pp. 270-283.
[6] D. S. JONES, *Uniform asymptotic remainders*, in Asymptotic and Computational Analysis, R. Wong, ed., Marcel Dekker, New York, 1990, pp. 241-264.
[7] E. R. LOVE, *Changing the order of integration*, J. Austral. Math. Soc. 9 (1970), pp. 421-432.
[8] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[9] ———, *On Stokes' phenomenon and converging factors*, in Asymptotic and Computational Analysis, R. Wong, ed., Marcel Dekker, New York, 1990, pp. 329-355.
[10] ———, *Uniform, exponentially improved, asymptotic expansions for the generalized exponential integral*, SIAM J. Math. Anal., this issue (1991), pp. 1460-1474.
[11] E. C. TITCHMARSH, *Introduction to the Theory of Fourier Integrals*, Second ed., Oxford University Press, London, New York, 1948.
[12] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1941.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1491-1499, November 1991

$C^{\alpha}(\bar{\Omega})$ SOLUTIONS OF A CLASS OF NONLINEAR DEGENERATE ELLIPTIC SYSTEMS ARISING IN THE THERMISTOR PROBLEM*

HONG XIE† AND W. ALLEGRETTO†

Abstract. Under realistic assumptions on the electrical and thermal conductivities, the existence is proven, for some $0<\alpha<1$, of positive $C^{\alpha}(\bar{\Omega})$ solutions for a system of degenerate elliptic equations which model a thermistor. A priori bounds are established for the solutions, and then the conductivities are truncated, so that a uniformly elliptic system is obtained. Next, $L^{2,\mu}(\Omega)$ estimates are used to obtain $C^{\alpha}(\bar{\Omega})$ estimates. Finally, the desired results follow from fixed-point argument.

Key words. degenerate, elliptic system, thermistor, estimates, truncation

AMS(MOS) subject classifications. 35J55, 35J60, 35Q20

1. Introduction. In this paper, we study the Boundary Value Problem for a class of nonlinear elliptic systems which models a temperature dependent electrical resistor. The electrical potential and the temperature distributions are denoted by φ and u, respectively. These are functions defined on a smooth domain (open, bounded set) Ω in R^N. The cases of practical interest are for $N\leqq 3$ but our proofs also hold for $N>3$. The relevant equations are as follows:

$$\nabla(\sigma(u)\nabla\varphi)=0 \quad \text{in } \Omega \quad \text{and} \quad \varphi=\varphi_0 \quad \text{on } \partial\Omega; \tag{1.1}$$

$$\nabla(k(u)\nabla u)=-\sigma(u)|\nabla\varphi|^2 \quad \text{in } \Omega \quad \text{and} \quad u=u_0 \quad \text{on } \partial\Omega. \tag{1.2}$$

Equations (1.1) and (1.2) express, respectively, the conservation of current and the energy balance including the electrical heating due to the Joule effect. The functions u_0, φ_0 are the distributions on the boundary, while σ, k denote the electrical and thermal conductivity. In many cases of practical interest, σ and k have the following forms:

$$\sigma(u)=Au^{\gamma}\exp(-C/Bu), \tag{1.3}$$

$$k(u)=(D+Eu+Fu^2)^{-1}, \tag{1.4}$$

where u is the temperature, A, B, C, D, E, and F are physical positive constants, and γ is a small positive number or a nonpositive number. Further details on the physical background of this system can be found in [1], [2], and [5].

We observe that system (1.1), (1.2) is degenerate, highly coupled, and that the right-hand side has quadratic growth in the gradient of one of the unknowns. These are features not often found in the nonlinear systems commonly considered in the literature. In particular, it does not appear that arguments based on variational theory and/or Sobolev's Embedding Theorems can be easily applied to system (1.1), (1.2).

Throughout this paper, we make the following assumptions:

(H1) *$\Omega\subset R^N$ is a bounded domain with $\partial\Omega\in C^1$.*

(H2) *$\varphi_0(x)$ and $u_0(x)$ are in $C^{1,\beta}(\bar{\Omega})(\beta>0)$, i.e., there exist positive constants φ_M, u_m and u_M such that*

$$0\leqq\varphi_0(x)\leqq\varphi_M \quad \text{and} \quad u_m\leqq u_0(x)\leqq u_M \quad \text{on } \bar{\Omega}. \tag{1.5}$$

* Received by the editors September 4, 1990; accepted for publication February 20, 1991. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

(H3) *$\sigma(t)$ and $k(t)$ are continuous for $t \geqq 0$, positive and smooth for $t > 0$, and there exists a smooth function f such that*

$$\frac{k(t)}{\sigma(t)} \geqq f(t) > 0 \quad \text{for } t \geqq u_M, \tag{1.6}$$

and: (a) *Either $\int_{u_M}^{\infty} f(t)\,dt = \infty$ or, if $\int_{u_M}^{\infty} f(t)\,dt < \infty$, then*

$$\int_{u_M}^{\infty} f(t)\,dt > \frac{\varphi_M^2}{2}; \tag{1.7}$$

(b) *There exists a sequence $\{t_n\}_1^{\infty}$, tending to infinity, such that $f(t_n) \geqq f(t)$ for $t \in [t_n, \infty)$.*

Remark 1.1. The functions (1.3), (1.4) satisfy (H3) if $F > 0$ and $\gamma \leqq -1$ in (1.3), (1.4). We need only take

$$f(t) = \frac{a_M}{t} \quad \text{with } a_M = \min_{t \geqq u_M} [t\, e^{C/Bt}/(D + Et + Ft^2)t^{\gamma}] > 0.$$

For another example, let $F > 0$ and $\gamma \leqq 0$, and we now choose $f(t) = a_M/t^2$ with $a_M = \min_{t \geqq u_M} [t^2\, e^{C/Bt}/(D + Et + Ft^2)t^{\gamma}] > 0$. Equation (1.7) now implies the restriction $u_M \cdot \varphi_M^2 < 2a_M$.

Remark 1.2. We observe that, unlike the temperature, the potential need not be positive. We can deal with this more general case as follows: Let the arbitrary constants C_m and C_M be such that

$$C_m \leqq \varphi_0(x) \leqq C_M.$$

Note that (1.1), (1.2) are invariant under shifts of φ by an arbitrary constant. Thus by shifting by C_m and letting $\varphi_M = C_M - C_m$ we have

$$0 \leqq \varphi_0(x) - C_m \leqq \varphi_M.$$

Everything remains unchanged except that (1.7) now becomes

$$\int_{u_M}^{\infty} f(t)\,dt > \frac{(C_M - C_m)^2}{2}. \tag{1.7}'$$

This restriction is essential in our method of proof, but we conjecture that the results may be true without it.

We will prove that there exists at least one positive $C^{\alpha}(\bar{\Omega})$ (some $0 < \alpha < 1$) solution of (1.1), (1.2) under (H1)-(H3). The same problem (1.1), (1.2) has been studied in [3] by Cimatti and Prodi. They obtained existence results in $H^1(\Omega)$ when $k(t)$ is constant, but no boundedness of the solution. Both physical and numerical results indicate that the solution should be bounded. Later in [2], Cimatti proved that all solutions of (1.1), (1.2) are bounded under the following assumption:

(H3)′ *There exist three positive constants σ_m, σ_M and k_m such that $\sigma_m \leqq \sigma(t) \leqq \sigma_M$ and $k(t) \geqq k_m$ for all $t \geqq u_m$.*

The assumption (H3)′ is very restrictive and it excludes important practical cases as in (1.3), (1.4).

We would like to indicate that the results in [2] and [3] furnished the motivation for this work.

2. Estimates. We define the weak solutions of the system (1.1), (1.2) as couples $(u(x), \varphi(x)) \in H^1(\Omega)$, which satisfy

$$\varphi - \varphi_0 \in H_0^1(\Omega), \quad \int_\Omega \sigma(u)\nabla\varphi\nabla\chi = 0 \quad \forall \chi \in H_0^1(\Omega); \tag{2.1}$$

$$u - u_0 \in H_0^1(\Omega), \quad \int_\Omega k(u)\nabla u\nabla v = \int_\Omega \sigma(u)|\nabla\varphi|^2 v \quad \forall v \in H_0^1(\Omega). \tag{2.2}$$

A solution (u, φ) is a $C^\alpha(\bar{\Omega})(0 \leqq \alpha \leqq 1)$ solution of (1.1), (1.2), or equivalently, of (2.1), (2.2) if and only if u and φ are in $C^\alpha(\bar{\Omega})$ and satisfy (2.1), (2.2).

In what follows, C denotes a generic constant which may differ from proof to proof or even within the same proof.

THEOREM 2.1 (Boundedness estimate). *Let (u, φ) be a positive $C^\alpha(\bar{\Omega})$ solution of* (2.1), (2.2), *then*

$$0 \leqq \varphi(x) \leqq \varphi_M \quad \text{on } \bar{\Omega}, \tag{2.3}$$

$$u_m \leqq u(x) \leqq M \quad \text{on } \bar{\Omega}, \tag{2.4}$$

where M is a positive constant independent of u.

Proof. Since $u \in C^\alpha(\bar{\Omega})(0 < \alpha < 1)$ and is positive, $0 \leqq \min_\Omega u \leqq u(x) \leqq \max_\Omega u < \infty$. So $0 < \min_\Omega \sigma(u) \leqq \max_\Omega \sigma(u) < \infty$ and $0 < \min_\Omega k(u) \leqq \max_\Omega k(u) < \infty$ by (H3). This implies that the system

$$\varphi - \varphi_0 \in H_0^1(\Omega), \quad \int_\Omega \sigma(u)\nabla\varphi\nabla\chi = 0 \quad \forall \chi \in H_0^1(\Omega), \tag{2.5}$$

$$u - u_0 \in H_0^1(\Omega), \quad \int_\Omega k(u)\nabla u\nabla v = \int_\Omega \sigma(u)|\nabla\varphi|^2 v \quad \forall v \in H_0^1(\Omega) \tag{2.6}$$

is uniformly elliptic. Furthermore, $\varphi \in C^1(\bar{\Omega})$ by (2.5) and (H2), (H3) and standard elliptic estimates [4]. We apply the weak maximum principle [4] to both equations. We have that $0 \leqq \min_{\partial\Omega} \varphi_0 \leqq \varphi(x) \leqq \max_{\partial\Omega} \varphi_0 \leqq \varphi_M$ and $u(x) \geqq \min_{\partial\Omega}(u_0) \geqq u_m$ on $\bar{\Omega}$. In order to get the upper bound on u, we employ the same transformation that was used in [2].

Let $\chi = \varphi v \in H_0^1(\Omega)$ for $v \in H_0^1(\Omega)$, and substitute it into (2.1). We have

$$\int_\Omega \sigma(u)|\nabla\varphi|^2 v = -\int_\Omega \sigma(u)\varphi\nabla\varphi\nabla v \quad \forall v \in H_0^1(\Omega). \tag{2.7}$$

Replace the right-hand side of the equation in (2.2) by (2.7). We see that u satisfies

$$u - u_0 \in H_0^1(\Omega), \quad \int_\Omega k(u)\nabla u\nabla v = -\int_\Omega \sigma(u)\varphi\nabla\varphi\nabla v \quad \forall v \in H_0^1(\Omega). \tag{2.8}$$

Furthermore, we let $\xi(x) = \frac{1}{2}\varphi^2(x) + \int_{u_m}^{u(x)} (k(t)/\sigma(t))\,dt$ and $\xi_0(x) = \frac{1}{2}\varphi_0^2(x) + \int_{u_m}^{u_0(x)} (k(t)/\sigma(t))\,dt$. It is easy to see that $\xi(x)$ and $\xi_0(x)$ are in $H^1(\Omega)$. From (2.8), we have that $\xi(x)$ satisfies

$$\xi - \xi_0 \in H_0^1(\Omega), \quad \int_\Omega \sigma(u)\nabla\xi\nabla v = 0 \quad \forall v \in H_0^1(\Omega).$$

By the weak maximum principle, $\min_{\partial\Omega} \xi_0 \leqq \xi(x) \leqq \max_{\partial\Omega} \xi_0$ on $\bar{\Omega}$, i.e., we have

$$\int_{u_m}^{u(x)} \frac{k(t)}{\sigma(t)}\,dt \leqq \frac{1}{2}\varphi^2 + \int_{u_m}^{u(x)} \frac{k(t)}{\sigma(t)}\,dt \leqq \frac{1}{2}\varphi_M^2 + \int_{u_m}^{u_M} \frac{k(t)}{\sigma(t)}\,dt.$$

It follows that

$$\int_{u_M}^{u(x)} \frac{k(t)}{\sigma(t)}\, dt \leqq \frac{1}{2}\varphi_M^2.$$

Now if $u(x) \leqq u_M$ for $x \in \bar{\Omega}$, we take $M = u_M$. Otherwise, if $u(x) \geqq u_M$ for some $x \in \bar{\Omega}$, then by (H3),

$$\frac{k(t)}{\sigma(t)} \geqq f(t) > 0 \quad \text{for } t \geqq u_M$$

implies

$$\int_{u_M}^{u(x)} f(t)\, dt \leqq \frac{1}{2}\varphi_M^2.$$

Assume first that $\int_{u_M}^{\infty} f(t)\, dt = +\infty$. The inequality above implies that there exists a bound $\tau > 0$ (which is independent of $u(x)$) such that $u(x) \leqq \tau$. Now select t_n such that $t_n \geqq \max\{\tau, u_M\}$ and note that $u(x) \leqq t_n$ for $x \in \bar{\Omega}$. We define $M = t_n$. Next, if $\int_{u_M}^{\infty} f(t)\, dt < \infty$ and $\int_{u_M}^{\infty} f(t)\, dt > \frac{1}{2}\varphi_M^2$, then there exists a small positive constant $\varepsilon > 0$ such that

$$\int_{u_M}^{\infty} f(t)\, dt \geqq \frac{1}{2}\varphi_M^2 + \varepsilon.$$

Thus

$$\int_{u_M}^{\infty} f(t)\, dt - \varepsilon \geqq \frac{1}{2}\varphi_M^2 \geqq \int_{u_M}^{u(x)} f(t)\, dt,$$

and so

$$\int_{u(x)}^{\infty} f(t)\, dt \geqq \varepsilon > 0.$$

This implies that there exists a $\tau > 0$ such that $u(x) \leqq \tau$. As before, we let $t_n \geqq \max\{u_M, \tau\}$ and let $M = t_n$. Estimate (2.4) is proved.

Now truncate $\sigma(t)$ and $k(t)$ as follows:

$$\sigma^*(t) = \begin{cases} \sigma(u_m), & 0 \leqq t \leqq u_m, \\ \sigma(t), & u_m \leqq t \leqq M, \\ \sigma(M), & M \leqq t < \infty \end{cases} \tag{2.9}$$

and

$$k^*(t) = \begin{cases} k(u_m), & 0 \leqq t \leqq u_m, \\ k(t), & u_m \leqq t \leqq M, \\ k(M), & M \leqq t < \infty. \end{cases} \tag{2.10}$$

Consider the following truncated system:

$$\varphi - \varphi_0 \in H_0^1(\Omega), \quad \int_\Omega \sigma^*(u)\nabla\varphi\nabla\chi = 0 \quad \forall \chi \in H_0^1(\Omega); \tag{2.11}$$

$$u - u_0 \in H_0^1(\Omega), \quad \int_\Omega k^*(u)\nabla u\nabla v = \int_\Omega \sigma^*(u)|\nabla\varphi|^2 v \quad \forall v \in H_0^1(\Omega). \tag{2.12}$$

We see that the truncated problem (2.11), (2.12) is a uniformly elliptic system. In fact, there exist two positive constants $\lambda_m \geqq \lambda_M$, such that

$$\lambda_m \leqq \sigma^*(t), \quad k^*(t) \leqq \lambda_M \quad \text{for all } t \geqq 0. \tag{2.13}$$

Moreover, $\sigma^*(t)$, $k^*(t)$ are still Lipschitz for $t \geqq 0$ and satisfy (1.6). Let us check in detail that $\sigma^*(t)$, $k^*(t)$ satisfy (1.6). By definition of $\sigma^*(t)$ and $k^*(t)$, we have

$$\frac{k^*(t)}{\sigma^*(t)} = \begin{cases} k(u_m)/\sigma(u_m), & 0 \leqq t \leqq u_m, \\ k(t)/\sigma(t), & u_m \leqq t \leqq M, \\ k(M)/\sigma(M), & M \leqq t < \infty. \end{cases}$$

Thus

$$\frac{k^*(t)}{\sigma^*(t)} = \frac{k(t)}{\sigma(t)} \geqq f(t) > 0 \quad \text{for } u_m \leqq u_M \leqq t \leqq M$$

$$\frac{k^*(t)}{\sigma^*(t)} = \frac{k(M)}{\sigma(M)} \geqq f(M) = f(t_n) \geqq f(t) \quad \text{for } t \in [t_n, \infty),$$

by (b) of H(3), i.e., (1.6) holds for $\sigma^*(t)$ and $k^*(t)$ with the same f.

LEMMA 2.1. *Any solution* (u, φ) *of* (2.11), (2.12) *satisfies the bound* (2.3), (2.4).

Proof. Since (2.11), (2.12) is uniformly elliptic, by the weak maximum principle [4], it follows that $0 \leqq \varphi(x) \leqq \varphi_M$ and $u(x) \geqq u_m$ on $\bar{\Omega}$. Let $\xi^*(x) = \frac{1}{2}\varphi^2(x) + \int_{u_m}^{u(x)} (k^*(t)/\sigma^*(t))\, dt$ and $\xi_0(x) = \frac{1}{2}\varphi_0^2(x) + \int_{u_m}^{u_0(x)} (k^*(t)/\sigma^*(t))\, dt = \frac{1}{2}\varphi_0^2(x) + \int_{u_m}^{u_0(x)} (k(t)/\sigma(t))\, dt$ since $u_0(x) \leqq M$. The result then follows by the same argument as in the proof of Theorem 2.1, if we can verify that ξ^* and ξ_0^* are in $H^1(\Omega)$. In fact, φ_0 and φ are bounded and in $H^1(\Omega)$, and so are $\frac{1}{2}\varphi^2$ and $\frac{1}{2}\varphi_0^2$. Also, since u and u_0 are in $H^1(\Omega)$ and $\lambda_m/\lambda_M \leqq k^*(t)/\sigma^*(t) \leqq \lambda_M/\lambda_m$ by (2.13), then $\int_{u_m}^{u(x)} (k^*(t)/\sigma^*(t))\, dt$ and $\int_{u_m}^{u_0(x)} (k^*(t)/\sigma^*(t))\, dt = \int_{u_m}^{u_0(x)} (k(t)/\sigma(t))\, dt$ are in $H^1(\Omega)$. The lemma is proved.

COROLLARY 2.1. *Any positive* $C^\alpha(\bar{\Omega})$ *solution of* (2.1), (2.2) *is a* $C^\alpha(\bar{\Omega})$ *solution of* (2.11), (2.12) *and any* $C^\alpha(\bar{\Omega})$ *solution of* (2.11), (2.12) *is a positive* $C^\alpha(\bar{\Omega})$ *solution of* (2.1), (2.2).

Hence, the existence of a positive $C^\alpha(\bar{\Omega})$ solution of (2.1), (2.2) is equivalent to the existence of a $C^\alpha(\bar{\Omega})$ solution of the uniformly elliptic system (2.11), (2.12). The rest of this paper is devoted to the proof of existence of a $C^\alpha(\bar{\Omega})$ solution of (2.11), (2.12).

THEOREM 2.2 ($H^1(\Omega)$ estimate). *Let* (u, φ) *be a* $C^\alpha(\bar{\Omega})$ *solution of* (2.11), (2.12). *We have the following estimates*:

$$\|\varphi\|_{H^1(\Omega)} \leqq C\|\varphi_0\|_{H^1(\Omega)}; \tag{2.14}$$

$$\|u\|_{H^1(\Omega)} \leqq C(\|\varphi_0\|_{H^1(\Omega)} + \|u_0\|_{H^1(\Omega)}), \tag{2.15}$$

where C *only depends on* u_m, M, λ_m, *and* λ_M.

Proof. Estimate (2.14) is obtained by letting $\chi = \varphi - \varphi_0$ in (2.11), i.e.,

$$\varphi - \varphi_0 \in H_0^1(\Omega), \qquad \int_\Omega \sigma^*(u)\nabla\varphi \cdot \nabla(\varphi - \varphi_0)\, dx = 0.$$

Thus

$$\int_\Omega \sigma^*(u)|\nabla(\varphi - \varphi_0)|^2\, dx = -\int_\Omega \sigma^*(u)\nabla\varphi_0 \cdot \nabla(\varphi - \varphi_0)\, dx.$$

Using (2.13) and the Schwarz inequality, we obtain

$$\|\nabla(\varphi - \varphi_0)\|_{L^2(\Omega)} \leqq C\|\nabla\varphi_0\|_{L^2(\Omega)} \leqq C\|\varphi_0\|_{H^1(\Omega)}.$$

Note that $\|\varphi-\varphi_0\|_{H^1(\Omega)} \leqq C\|\nabla(\varphi-\varphi_0)\|_{L^2(\Omega)}$ since $\varphi-\varphi_0 \in H_0^1(\Omega)$, and hence

$$\|\varphi\|_{H^1(\Omega)} \leqq \|\varphi-\varphi_0\|_{H^1(\Omega)} + \|\varphi_0\|_{H^1(\Omega)} \leqq C\|\varphi_0\|_{H^1(\Omega)}$$

by (2.7). We rewrite (2.12) as

$$u-u_0 \in H_0^1(\Omega), \qquad \int_\Omega k^*(u)\nabla u \nabla v\, dx = -\int_\Omega \sigma^*(u)\varphi\nabla\varphi\nabla v\, dx.$$

Since $\varphi \in L^\infty(\Omega)$ and $\nabla\varphi$ can be estimated by (2.14), the same argument as above can be used to obtain (2.15).

Next, we want to establish $C^\alpha(\bar{\Omega})(0<\alpha<1)$ estimates for the solutions of (2.11), (2.12). This is done by employing $L^{2,\mu}(\Omega)$ estimates for equations in divergence form. For convenience to the readers, we first state the definition and some related results about $L^{2,\mu}(\Omega)$ spaces.

DEFINITION 2.1. We say that an $L^2(\Omega)$-function $u \in L^{2,\mu}(\Omega)$ if u satisfies

$$[u]_{2,\mu,\Omega} = \left(\sup_{\substack{x_0\in\bar{\Omega} \\ 0<\rho<\infty}} \rho^{-\mu} \int_{\Omega\cap B(x_0,\rho)} |u-u_{x_0,\rho}|^2\, dx \right)^{1/2} < \infty, \tag{2.16}$$

where $u_{x_0,\rho} = (1/\text{mes}\,(B(x_0,\rho)\cap\Omega)) \int_{B(x_0,\rho)\cap\Omega} u\, dx$, $x_0 \in \bar{\Omega}$.

We recall that $\partial\Omega \in C^1$ by (H1).

LEMMA 2.2. *For $0 \leqq \mu \leqq N+2$, $L^{2,\mu}(\Omega)$ is a Banach space under the norm $\|u\|_{L^{2,\mu}(\Omega)} = \|u\|_{L^2(\Omega)} + [u]_{2,\mu,\Omega}$.*

LEMMA 2.3. (i) *If $0 \leqq \mu \leqq N$, $L^\infty(\Omega) \subset L^{2,\mu}(\Omega)$ and $L^\infty(\Omega)$ is a space of multipliers for $L^{2,\mu}(\Omega)$.*

(ii) *If $N < \mu \leqq N+2$, $L^{2,\mu}(\Omega)$ is isomorphic to $C^\alpha(\bar{\Omega})$ with $\alpha = (\mu-N)/2$.*

LEMMA 2.4. *Let $u \in H^1(\Omega)$ and $\nabla u \in L^{2,\mu}(\Omega)$ with $0 \leqq \mu < N$; then $u \in L^{2,\mu+2}(\Omega)$ with*

$$\|u\|_{L^{2,\mu+2}(\Omega)} \leqq C(\|u\|_{L^2(\Omega)} + \|\nabla u\|_{L^{2,\mu}(\Omega)}),$$

where C is independent of u.

The proofs of the above three lemmas can be found in [7, pp. 28–63]. Our $L^{2,\mu}(\Omega)$ estimate is based on the next lemma, which is a special case of Theorem 2.19 in [7, p. 124]. We note here that Theorem 2.19 in [7] is established explicitly only for $N \geqq 3$, but our $L^{2,\mu}(\Omega)$ estimates (Lemma 2.5) are true not only for $N \geqq 3$ but also for $N = 2$ (see [6, § 5]). For the case where $N = 1$, we can get $C^\alpha(\bar{\Omega})$ estimates directly by Sobolev's Embedding Theorem.

LEMMA 2.5. *Let $u \in H_0^1(\Omega)$ satisfy*

$$\int_\Omega a^{ij} u_{x_i} v_{x_j} = \int_\Omega f^i v_{x_i} \quad \forall v \in H_0^1(\Omega), \tag{2.17}$$

with $\lambda\|\xi\|^2 \leqq a^{ij}\xi_i\xi_j \leqq \Lambda\|\xi\|^2$, $\Lambda \geqq \lambda > 0$, $\xi \in R^N \setminus 0$, and $f^i \in L^{2,\mu}(\Omega)$. Then $\nabla u \in L^{2,\mu}(\Omega)$ and, moreover, we have

$$\|\nabla u\|_{L^{2,\mu}(\Omega)} \leqq C\,(\textstyle\sum \|f^i\|_{L^{2,\mu}(\Omega)} + \|u\|_{H^1(\Omega)}),$$

where $0 < \mu < \mu_0 = N-2+2\delta_0$, $0 < \delta_0 < 1$, and δ_0, C only depend on λ, Λ, and Ω.

THEOREM 2.3 ($L^{2,\mu}(\Omega)$ estimates). *Let $(u,\varphi) \in H^1(\Omega)$ be a $C^\alpha(\bar{\Omega})$ solution of* (2.11), (2.12). *Then we have that*

$$\|\nabla\varphi\|_{L^{2,\mu}(\Omega)} \leqq C(\|\varphi_0\|_{H^1(\Omega)} + \|\nabla\varphi_0\|_{L^{2,\mu}(\Omega)}), \tag{2.18}$$

$$\|\nabla u\|_{L^{2,\mu}(\Omega)} \leqq C(\|\varphi_0\|_{H^1(\Omega)} + \|u_0\|_{H^1(\Omega)} + \|\nabla\varphi_0\|_{L^{2,\mu}(\Omega)} + \|\nabla u_0\|_{L^{2,\mu}(\Omega)}), \tag{2.19}$$

where C only depends on λ_m, M, Λ_M, u_m, φ_M, and Ω, and $0<\mu<\mu_0=N-2+2\delta_0$, $0<\delta_0<1$ depends on λ_m and λ_M only.

Proof. Rewrite (2.11), (2.12) in the following forms:

$$\psi\in H_0^1(\Omega),\quad \int_\Omega \sigma^*(u)\nabla\psi\nabla\chi=-\int_\Omega \sigma^*(u)\nabla\varphi_0\nabla\chi \quad \forall\chi\in H_0^1(\Omega), \tag{2.20}$$

$$\begin{aligned} &W\in H_0^1(\Omega),\quad \int_\eta k^*(u)\nabla W\nabla v \\ &\qquad =\int_\Omega \sigma^*(u)|\nabla\varphi|^2 v-\int_\Omega k^*(u)\nabla u_0\nabla v \quad \forall v\in H_0^1(\Omega); \end{aligned} \tag{2.21}$$

and

$$\varphi(x)=\psi(x)+\varphi_0(x) \quad\text{and}\quad u(x)=W(x)+u_0(x) \qquad \text{on } \bar\Omega. \tag{2.22}$$

Since (u,φ) solves (2.11), (2.12), (W,ψ) solves (2.20), (2.21). By assumption (H2) and Lemmas 2.3 and 2.5, we have that

$$\|\nabla\psi\|_{L^{2,\mu}(\Omega)}\leqq C(\|\sigma^*(U)\nabla\varphi_0\|_{L^{2,\mu}(\Omega)}+\|\psi\|_{H^1(\Omega)}), \tag{2.23}$$

where $0<\mu<\mu_0=N-2+2\delta_0$, $0<\delta_0<1$ only depend on λ_m and λ_M. Since $\int_\Omega \sigma^*(u)|\nabla\varphi|^2 v\,dx=-\int_\Omega \sigma^*(u)\varphi\nabla\varphi\nabla v\,dx$ for all $v\in H_0^1(\Omega)$, we replace $\int_\Omega \sigma^*(u)|\nabla\varphi|^2 v\,dx$ by $-\int_\Omega \sigma^*(u)\varphi\nabla\varphi\nabla v\,dx$ on the right-hand side of the equation in (2.21). We apply Lemma 2.5 again to obtain

$$\|\nabla W\|_{L^{2,\mu}(\Omega)}\leqq C(\|\sigma^*(u)\varphi\nabla\varphi\|_{L^{2,\mu}(\Omega)}+\|k^*(u)\nabla u_0\|_{L^{2,\mu}(\Omega)}+\|W\|_{H^1(\Omega)}). \tag{2.24}$$

Now, by Lemma 2.3, since $\mu<\mu_0=N-2+2\delta_0<N$, it follows that

$$\|\sigma^*(u)\nabla\varphi_0\|_{L^{2,\mu}(\Omega)}\leqq\|\sigma^*(u)\|_{L^\infty(\Omega)}\|\nabla\varphi_0\|_{L^{2,\mu}(\Omega)}\leqq\lambda_M\|\nabla\varphi_0\|_{L^{2,\mu}(\Omega)}, \tag{2.25}$$

$$\|\sigma^*(u)\varphi\nabla\varphi\|_{L^{2,\mu}(\Omega)}\leqq\|\sigma^*(u)\|_{L^\infty(\Omega)}\|\varphi\|_{L^\infty(\Omega)}\|\nabla\varphi\|_{L^{2,\mu}(\Omega)}\leqq\lambda_M\varphi_M\|\nabla\varphi\|_{L^{2,\mu}(\Omega)} \tag{2.26}$$

$$\|k^*(u)\nabla u_0\|_{L^{2,\mu}(\Omega)}\leqq\|k^*(u)\|_{L^\infty(\Omega)}\|\nabla u_0\|_{L^{2,\mu}(\eta)}\leqq\lambda_M\|\nabla u_0\|_{L^{2,\mu}(\Omega)}. \tag{2.27}$$

By Theorem 2.2 and (2.22), we conclude

$$\|\psi\|_{H^1(\Omega)}\leqq\|\varphi\|_{H^1(\Omega)}+\|\varphi_0\|_{H^1(\Omega)}\leqq C\|\varphi_0\|_{H^1(\Omega)}, \tag{2.28}$$

$$\|W\|_{H^1(\Omega)}\leqq\|u\|_{H^1(\Omega)}+\|u_0\|_{H^1(\Omega)}\leqq C(\|u_0\|_{H^1(\Omega)}+\|\varphi_0\|_{H^1(\Omega)}). \tag{2.29}$$

By combining (2.23), (2.25), and (2.28), we get

$$\|\nabla\varphi\|_{L^{2,\mu}(\Omega)}\leqq\|\nabla\varphi\|_{L^{2,\mu}(\Omega)}+\|\nabla\varphi_0\|_{L^{2,\mu}(\Omega)}\leqq C(\|\nabla\varphi_0\|_{L^{2,\mu}(\Omega)}+\|\varphi_0\|_{H^1(\Omega)}). \tag{2.30}$$

And, by (2.24), (2.26), (2.27), (2.29), and (2.30), it follows that

$$\begin{aligned} \|\nabla u\|_{L^{2,\mu}(\Omega)}&\leqq\|\nabla W\|_{L^{2,\mu}(\Omega)}+\|\nabla u_0\|_{L^{2,\mu}(\Omega)} \\ &\leqq C(\|\nabla\varphi_0\|_{L^{2,\mu}(\Omega)}+\|\nabla u_0\|_{L^{2,\mu}(\Omega)}+\|\varphi_0\|_{H^1(\Omega)}+\|u_0\|_{H^1(\Omega)}). \end{aligned} \tag{2.31}$$

Since (2.30), (2.31) hold for all μ with $0<\mu<\mu_0=N-2+2\delta_0$, $0<\delta_0<1$, the theorem follows.

COROLLARY 2.2. *There exist positive constants K, δ_0 such that for any $C^\alpha(\bar\Omega)$ solution (u,φ) of* (2.11), (2.12), *we have*

$$\|u\|_{C^\gamma(\bar\Omega)}\leqq K \quad\text{and}\quad \|\varphi\|_{C^\gamma(\bar\Omega)}\leqq K \tag{2.32}$$

for all γ with $0<\gamma<\delta_0$.

Proof. We need only employ Lemmas 2.3 and 2.4 and Theorem 2.3.

In conclusion we remark that the $H^1(\Omega)$, $L^{2,\mu}(\Omega)$, and C^γ bounds are a consequence of the uniform coefficient bounds and not of the specific nonlinear structure of the equations.

3. Main theorem. We fix α_0 in $(0, \delta_0)$ and consider the existence of a $C^{\alpha_0}(\bar{\Omega})$ solution of (2.11), (2.12)).

THEOREM 3.1 (Existence Theorem). *There exists at least one $C^{\alpha_0}(\bar{\Omega})$ solution of* (2.11), (2.12).

Proof. Let $S = \{u \in C^{\alpha_0}(\bar{\Omega}) \mid \|u\|_{C^{\alpha_0}(\bar{\Omega})} \leqq K \text{ and } u(x) \geqq u_m \text{ on } \bar{\Omega}\}$, where K is determined in Corollary 2.2. Obviously, S is a closed convex set in $C^{\alpha_0}(\bar{\Omega})$. We prove the result by constructing a completely continuous map T from S into S and then applying Schauder's Fixed-Point Theorem. For $u \in S$, we define $W = Tu$ by the following:

$$\varphi - \varphi_0 \in H_0^1(\Omega), \quad \int_\Omega \sigma^*(u)\nabla\varphi\nabla\chi = 0 \quad \forall \chi \in H_0^1(\Omega); \tag{3.1}$$

$$W - u_0 \in H_0^1(\Omega), \quad \int_\Omega k^*(u)\nabla W\nabla v = \int_\Omega \sigma^*(u)|\nabla\varphi|^2 v \quad \forall v \in H_0^1(\Omega). \tag{3.2}$$

We first note that T is well defined from S into S. In fact, since (3.1) is a linear uniformly elliptic Dirichlet problem, we conclude by standard linear theory [4] that there exists a unique solution φ in $C^1(\bar{\Omega})$. By Lemma 2.5, $\nabla\varphi \in L^{2,\mu}(\Omega)$ with $0 < \mu < \mu_0 = N - 2 + 2\delta_0$. If we substitute the solution φ of (3.1) into (3.2), we get another linear uniformly elliptic Dirichlet problem. Since $\int_\Omega \sigma^*(u)|\nabla\varphi|^2 v\,dx = -\int_\Omega \sigma^*(u)\varphi\nabla\varphi\nabla v\,dx$, the right-hand side of the equation in (3.2) defines a bounded linear functional on $H_0^1(\Omega)$. By the Lax–Milgram theorem [4], there exists a unique solution $W \in H^1(\Omega)$ and also by Lemma 2.5, $\nabla W \in L^{2,\mu}(\Omega)$. Now (φ, W) solves (3.1), (3.2). By the same arguments as in the proof of Theorem 2.3 and the corollary, we have $\|W\|_{C^\alpha(\bar{\Omega})} \leqq K$, with α in $(0, \delta_0)$. In particular, let $\alpha = \alpha_0$. and note that we get $\|W\|_{C^{\alpha_0}(\bar{\Omega})} \leqq K$ and $W \geqq u_m$ by the maximum principle, i.e., there exists a unique $W = Tu$ in S. It follows that T is well defined.

Next we prove that (i) T is continuous and (ii) T is compact. Let $\{u_n\}_{n=1}^\infty$ and u in S be such that $u_n \to u$ in $C^{\alpha_0}(\bar{\Omega})$ as $n \to \infty$. Let (φ_n, W_n) and (φ, W) be the solutions of (3.1), (3.2) corresponding to u_n and u, $n = 1, 2, \cdots$; we show that $W_n \to W$ in $C^{\alpha_0}(\bar{\Omega})$. Since φ_n, φ solve (3.1) with respect to u_n and u, we have

$$\varphi_n - \varphi \in H_0^1(\Omega), \quad \int_\Omega \sigma^*(u_n)\nabla(\varphi_n - \varphi)\nabla\chi = \int_\Omega (\sigma^*(u) - \sigma^*(u_n))\nabla\varphi\nabla\chi \tag{3.3}$$

for any $\chi \in H_0^1(\Omega)$, i.e., $\psi = \varphi_n - \varphi$ solves the following problem:

$$\psi \in H_0^1(\Omega), \quad \int_\Omega \sigma^*(u_n)\nabla\psi\nabla\chi = \int_\Omega (\sigma^*(u) - \sigma^*(u_n))\nabla\varphi\nabla\chi \quad \forall \chi \in H_0^1(\Omega). \tag{3.4}$$

By the estimate of Lemma 2.5 we have

$$\|\nabla(\varphi_n - \varphi)\|_{L^{2,\mu}(\Omega)} \leqq C(\|(\sigma^*(u) - \sigma^*(u_n))\nabla\varphi\|_{L^{2,\mu}(\Omega)} + \|\varphi_n - \varphi\|_{H^1(\Omega)}). \tag{3.5}$$

From Lemma 2.3, $\|(\sigma^*(u) - \sigma^*(u_n))\nabla\varphi\|_{L^{2,\mu}(\Omega)} \leqq \max_\Omega \|\sigma^*(u) - \sigma^*(u_n)\| \cdot \|\nabla\varphi\|_{L^{2,\mu}(\Omega)} \leqq CK \max \|\sigma^*(u) - \sigma^*(u_n)\|$ and by letting $\chi = \varphi_n - \varphi$ in (3.3), we get $\|\varphi_n - \varphi\|_{H^1(\Omega)} \leqq C \max_\Omega \|\sigma^*(u) - \sigma^*(u_n)\| \|\varphi\|_{H^1(\Omega)} \leqq KC \max \|\sigma^*(u) - \sigma^*(u_n)\|$. Note

that $u_n \to u$ in $C^{\alpha_0}(\bar{\Omega})$ implies $u_n \to u$ uniformly on $\bar{\Omega}$. By the continuity of $\sigma^*(t)$. It follows that

$$\max_{\Omega} \|\sigma^*(u) - \sigma^*(u_n)\| \to 0 \quad \text{as } n \to \infty,$$

i.e.,

$$\|\nabla(\varphi_n - \varphi)\|_{L^{2,\mu}(\Omega)} \to 0 \quad \text{as } n \to \infty, \quad \text{for } 0 < \mu < \mu_0. \tag{3.6}$$

By Lemmas 2.3 and 2.4, $\varphi_n \to \varphi$ in $C^{\alpha_0}(\bar{\Omega})$. Moreover, for $0 < \mu < \mu_0$, we have

$$\|\varphi_n \nabla \varphi_n - \varphi \nabla \varphi\|_{L^{2,\mu}(\Omega)} \leqq \|\varphi_n - \varphi\|_{C(\bar{\Omega})} \|\nabla \varphi_n\|_{L^{2,\mu}(\Omega)} + \|\varphi\|_{C(\Omega)} \|\varphi_n - \nabla \varphi\|_{L^{2,\mu}(\Omega)},$$

i.e.

$$\|\varphi_n \nabla \varphi_n - \varphi \nabla \varphi\|_{L^{2,\mu}(\Omega)} \to 0 \quad \text{as } n \to \infty.$$

Next note that W_n and W satisfy the following:

$$\int_{\Omega} k^*(u_n) \nabla W_n \nabla v = \int_{\Omega} \sigma^*(u_n) |\nabla \varphi_n|^2 v \quad \forall v \in H_0^1(\Omega),$$

$$\int_{\Omega} k^*(u) \nabla W \nabla v = \int_{\Omega} \sigma^*(u) |\nabla \varphi|^2 v \quad \forall v \in H_0^1(\Omega).$$

Applying (2.7) to both the equations and subtracting yields

$$\int_{\Omega} k^*(u_n) \nabla (W_n - W) \nabla v = \int_{\Omega} [(\sigma^*(u) - \sigma^*(u_n)) \varphi_n \nabla \varphi_n + \sigma^*(u)(\varphi \nabla \varphi - \varphi_n \nabla \varphi_n)$$
$$+ (k^*(u) - k^*(u_n)) \nabla W] \nabla v \quad \forall v \in H_0^1(\Omega).$$

Now, since the same argument applies to the above equation, we have $W_n \to W$ in $C^{\alpha_0}(\bar{\Omega})$, i.e., T is continuous. Finally, that T is compact follows by noting that actually $W = Tu$ is in $C^{\alpha}(\bar{\Omega})$ for $\alpha_0 < \alpha < \delta_0$ and $C^{\alpha}(\bar{\Omega}) \hookrightarrow C^{\alpha_0}(\bar{\Omega})$ is a compact embedding for $\alpha > \alpha_0$. By the Schauder Fixed-Point Theorem T has at least one fixed point in S, which we denote by u^*. Let φ^* be the solution by (3.1) for $u = u^*$. We see that (u^*, φ^*) is a $C^{\alpha_0}(\bar{\Omega})$ solution of (2.11), (2.12), and the proof is complete.

REFERENCES

[1] W. ALLEGRETTO, A. NATHAN, K. CHAU, AND H. P. BALTES, *Two dimensional numerical simulations of electrothermal behavior in very large scale integrated contacts and vias*, Canad. J. Phys., 67 (1989), pp. 212-217.

[2] G. CIMATTI, *A bound for the temperature in the thermistor problem*, IMA J. Appl. Math., 40 (1988), pp. 15-22.

[3] G. CIMATTI AND G. PRODI, *Existence results for a nonlinear elliptic system modeling a temperature dependent electrical resistor*, Ann. Mat. Pura Appl., Ser. 4 (1988), pp. 227-236.

[4] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1983.

[5] F. J. HYDE, *Thermistors*, Iliffe Book, London, 1971.

[6] J. KADLEC AND J. NEČAS, *Sulla regolarita delle soluzioni di equazioni ellittiche negli spazi $H^{k,d}$*, Ann. Scuola Norm. Sup. Pisa Sci. Fis. Mat., 21 (1967), pp. 527-545.

[7] G. M. TROIANIELLO, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, London, 1987.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1500-1515, November 1991

002

EXISTENCE AND MULTIPLICITY OF POSITIVE RADIAL SOLUTIONS FOR SEMILINEAR ELLIPTIC EQUATIONS IN ANNULAR DOMAINS*

SONG-SUN LIN†‡ AND FENG-MING PAI†

Abstract. The existence and multiplicity of positive radial solutions of equation $\Delta u + f(u) = 0$ is studied in annular domains in $\mathbb{R}^n$, $n \geqq 2$. It is proven that if $f(0) \geqq 0$, f is somewhere negative in $(0, \infty)$ and superlinear at ∞, then there is a large positive radial solutions on all annuli. If $f(0) < 0$ and satisfies certain conditions, then the equation has no solution if the annuli are too wide. Multiplicity results are also obtained when f has many humps with positive areas.

Key words. elliptic, semilinear, positive radial solution, annular domain

AMS(MOS) subject classifications. 35B32, 35JB65, 35P30

1. Introduction. In this paper we consider the existence and multiplicity of positive radial solutions of the semilinear elliptic equation

$$\Delta u(x) + f(u(x)) = 0 \quad \text{in } a < |x| < b, \tag{1.1}$$

$$u(x) = 0 \quad \text{on } |x| = a \quad \text{and} \quad |x| = b, \tag{1.2}$$

$x \in \mathbb{R}^n$, $n \geqq 2$ and $f \in C^1((0, \infty)) \cap C^0([0, \infty))$ satisfying the following hypotheses:

(H1) f is negative somewhere in $(0, \infty)$;

(H2) f is superlinear at $u = \infty$, i.e., $\lim_{u \to \infty} f(u)/u = \infty$.

One of the problems for semilinear elliptic equations in annular domains which have been studied quite extensively in recent years is:

(P) Does (1.1), (1.2) possess a positive radial solution in every annulus?

The answer to (P) was proved affirmative by Nehari [20], assuming that f is positive in $(0, \infty)$ and satisfies the condition: $\exists \delta > 0$ such that $f(u)/u^{1+\delta}$ is monotone increasing in $(0, \infty)$.

Later, (P) was studied by Kazdan and Warner [15], Ni and Nussbaum [21], Bandle, Coffman, and Marcus [2], Garaizar [13], and Lin [17].

In [2], Bandle, Coffman, and Marcus showed that the answer to (P) is affirmative, provided that f is positive in $(0, \infty)$ and satisfies the following conditions:

(A1) f is nondecreasing in $(0, \infty)$;

(A2) $\lim_{u \to 0} f(u)/u = 0$;

(A3) $\lim_{u \to \infty} f(u)/u = \infty$.

In [2], it is remarked that (A1) is not a necessary condition for existence. This have been confirmed by Coffman and Marcus [8] and Lin [17] independently.

With a suitable change of independent variable, (1.1), (1.2) become equations of the form

$$u''(t) + G(t, u) = 0, \qquad t_0 < t < t_1, \tag{1.3}$$

$$u(t_0) = 0 = u(t_1). \tag{1.4}$$

* Received by the editors January 22, 1990; accepted for publication January 7, 1991.

† Department of Applied Mathematics, National Chiao Tung University, Hsin-chu, Taiwan, Republic of China.

‡ The work of this author was partially supported by the National Science Council of the Republic of China.

In [3], Bandle and Kwong showed that the answer to (P) is affirmative if G satisfies the following conditions:

(G1) G is C^0 in its first variable and C^1 in its second;

(G2) $\lim_{u\to\infty} G(t, u)/u = \infty$ uniformly on every $[t_0, t_1]$;

(G3) $\lim_{u\to 0} G(t, u)/u = 0$ uniformly on every $[t_0, t_1]$.

$G(t, u)$ is now allowed to be negative for small positive value and $G(t, 0) = 0$ is assumed implicitly with the limit involved exists and finite in (G3).

In this paper, we first generalize the results of Bandle and Kwong [3], showing that (P) is affirmative if (H1), (H2), and (H3) are satisfied.

(H3) $$f(0) \geqq 0.$$

Moreover, solutions obtained are "large" in the following sense: By (H1), there exists $(u_*, u^*) \subset (0, \infty)$ such that

(1.5) $$f(u) \geqq 0 \quad \text{in } (u^*, \infty), \qquad f(u) < 0 \quad \text{in } (u_*, u^*), \qquad f(u_*) = f(u^*) = 0.$$

Let $\gamma > u^*$ be the smallest number such that

(1.6) $$\int_{u_*}^{\gamma} f(u)\, du = 0.$$

The solution u of (1.1), (1.2) is called large if

(1.7) $$\|u\| \equiv \max\{u(x): a \leqq |x| \leqq b\} \geqq \gamma.$$

On the contrary, Garaizar [13] showed that (P) is negative, i.e., (1.1), (1.2) has no positive radial solution if $b - a$ is too large, if f satisfies the following conditions:

(i) $f(0) < 0$;

(ii) There exists $\bar{u} > 0$ such that $F(u) \leqq 0$ in $(0, \bar{u})$ and $f(u) > 0$ in $(\bar{u}, \infty)$;

(iii) There exists $k > 1$ and $d_2 \geqq d_1 > 0$ such that $d_1 u^k \leqq f(u) \leqq d_2 u^k$ for u large, where

$$F(u) = \int_0^u f(t)\, dt.$$

We can also obtain a similar nonexistence result without assuming condition (iii), i.e., if (H2) and the following hold true:

(H3)′(i) $f(0) < 0$;

(ii) There exists $\bar{u} > 0$ such that $F(u) < 0$ in $(0, \bar{u}]$ and $f(u) > 0$ in $(\bar{u}, \infty)$.

On the other hand, when f changes signs, the existence of multiple positive solutions of the equation

(1.8) $$\Delta u + \lambda f(u) = 0 \quad \text{in } \Omega,$$

(1.9) $$u = 0 \quad \text{on } \partial\Omega,$$

$\lambda \geqq 0$ and Ω is a bounded smooth domain in $\mathbb{R}^n$, $n \geqq 2$, has been studied by many authors (see, e.g., Brown and Budin [5], Hess [14], de Figueiredo [12], Clement and Sweers [6], and Wang and Kazarinoff [24]).

In [14], Hess showed that if f satisfies the following conditions:

(f_1) $f(0) > 0$;

(f_2) There exist m numbers $\bar{u}_m > \bar{u}_{m-1} > \cdots > \bar{u}_1 > 0$ such that $f(\bar{u}_k) = 0$ for $k = 1, \cdots, m$;

(f_3) $0 < \max\{F(s): 0 \leqq s \leqq \bar{u}_{k-1}\} < F(\bar{u}_k)$, $k = 2, \cdots, m$,

then there exists a number $\bar{\lambda} > 0$ such that for all $\lambda > \bar{\lambda}$, (1.7), (1.8) have at least $2m - 1$ positive solutions $\hat{u}_1, u_2, \hat{u}_2, \cdots, u_m, \hat{u}_m$ such that $\|\hat{u}_1\| < \bar{u}_1$ and $\bar{u}_{k-1} < \|u_k\|$, $\|\hat{u}_k\| < \bar{u}_k$ for $k = 2, \cdots, m$, and $\hat{u}_{k-1} < \hat{u}_k$ and $u_k < \hat{u}_k$ for $k = 2, \cdots, m$.

Later, de Figueiredo [12] obtained the existence of $2m-1$ ordered positive solutions under slightly different assumptions.

In this paper, we show that if f satisfies (f_2) and (f_3), then there exists $b_m > a$ such that for any $b > b_m$, (1.1), (1.2) has at least m ordered positive radial solutions u_k with $\|u_k\| \in (\bar{u}_{k-1}, \bar{u}_k)$ for $k = 2, \cdots, m$. Moreover, if $f(0) \geqq 0$, then (1.1), (1.2) has at least $2m-1$ positive radial solutions for $b > b_m$.

For the other related problems, note the following:

(i) Uniqueness of positive radial solution, when $f(u) > 0$ for $u \in (0, \infty)$, has been studied by Ni and Nussbaum [21], Bandle, Coffman, and Marcus [2], Bandle and Kwong [3], and Coffman and Marcus [8].

(ii) Symmetry breaking for positive radial solutions has been studied by Brezis and Nirenberg [4], Coffman [7], Suzuki and Nagasaki [22], [23], and Lin [16], [18], [19].

The methods used in this paper are shooting techniques, the phase-plane method, and variational methods. All results obtained in this paper can also be generalized to $f(r, u)$ which satisfies certain uniformity assumptions in r as in (G2) and (G3).

The paper is organized as follows. In § 2, we obtain some preliminary results which are useful. In § 3, we prove that (P) is affirmative when (H1) ~ (H3) are satisfied. In § 4, we prove (P) is negative when (H1) ~ (H3)′ are satisfied. In § 5, we obtain the multiplicity results for wide annuli.

2. Preliminaries. Since we are interested in positive radial solutions of (1.1), we write (1.1), (1.2) in the form

$$u''(r) + \frac{n-1}{r} u'(r) + f(u(r)) = 0 \quad \text{in } (a, b), \tag{2.1}$$

$$u(a) = 0 = u(b). \tag{2.2}$$

For fixed $a > 0$, we consider the family of solutions $u(\cdot) \equiv u(\cdot, \alpha)$ of the initial value problem

$$u''(r) + \frac{n-1}{r} u'(r) + f(u(r)) = 0 \quad \text{for } r > a, \tag{2.3}$$

$$u(a) = 0 \quad \text{and} \quad u'(a) = \alpha, \tag{2.4}$$

where $\alpha \geqq 0$ is the shooting parameter. Furthermore, (2.3), (2.4) can also be written as a dynamical system

$$u' = v, \tag{2.5}$$

$$v' = -\frac{n-1}{r} v - f(u), \tag{2.6}$$

with initial data

$$u(a) = 0 \quad \text{and} \quad v(a) = \alpha. \tag{2.7}$$

We define an energy function $H(\cdot) \equiv H(u(\cdot, \alpha))$ by

$$H(r) = \tfrac{1}{2} v^2(r) + F(u(r)). \tag{2.8}$$

Then, along each trajectory of solution of (2.5), (2.6), H is decreasing; in fact,

$$H'(r) = -\frac{n-1}{r} u'^2(r) \leqq 0. \tag{2.9}$$

Furthermore, $H(r)$ is strictly decreasing in r if $\alpha > 0$.

We first classify the solutions $u(\cdot, \alpha)$.

DEFINITION 2.1. For any $\alpha \geqq 0$, α belongs to one of the following three disjoint sets:

(i) $\alpha \in P$ (or $u(\cdot, \alpha)$ is a P-solution) if $u(r, \alpha) > 0$ for all $r > a$,

(ii) $\alpha \in N$ (or $u(\cdot, \alpha)$ is an N-solution) if there exists $b(\alpha) > 0$ such that $u(r, \alpha) > 0$ in $(a, b(\alpha))$, $u(b(\alpha), \alpha) = 0$ and $u'(b(\alpha), \alpha) < 0$,

(iii) $\alpha \in T$ (or $u(\cdot, \alpha)$ is a T-solution) if there exists $b(\alpha) > 0$ such that $u(r, \alpha) > 0$ in $(a, b(\alpha))$, $u(b(\alpha), \alpha) = 0$ and $u'(b(\alpha), \alpha) = 0$.

We then state some simple but basic properties of solutions $u(\cdot, \alpha)$.

LEMMA 2.2. (i) *If $\alpha \in N$, then $u(\cdot, \alpha)$ has only one local maximum.*

(ii) *If $\alpha \in N \cup T$, then $H(u(r, \alpha)) > 0$ for $r \in (a, b(\alpha))$.*

(iii) *If* (H2) *is satisfied and $u(r, \alpha) > 0$ for all $r > r_0 \geqq a$, then $u(r, \alpha)$ is bounded.*

(iv) *N is an open set.*

Proof. (i) The proof of (i), in the general case, was given by Garaizar [13]. The main idea is using energy $H(r)$, which decreases along the trajectory, and then obtain the following two facts:

(a) the trajectory cannot cross (intersect) itself;

(b) the trajectory cannot be tangent to the u-axis.

Therefore, (i) can be proved. For the details, see [13, Lemma 1].

(ii) Since $H(u(b(\alpha), \alpha)) \geqq 0$, (ii) follows.

(iii) Since

$$H(u(r, \alpha)) = \tfrac{1}{2}u'^2(r) + F(u(r, \alpha)) \leqq H(u(a, \alpha)) = \alpha^2/2,$$

we have $F(u(r, \alpha)) \leqq \alpha^2/2$ for all $r \geqq r_0$. Therefore, (H2) implies $u(r, \alpha)$ is bounded.

(iv) By the Implicit Function Theorem, $b(\alpha)$ is continuously differentiable in N and N is an open set. □

The following lemma indicates there is a great difference between cases $f(0) \geqq 0$ and $f(0) < 0$.

LEMMA 2.3. *If $f(0) \geqq 0$, then $T = \phi$. Furthermore, if $\alpha \in T$ then $u(r, \alpha) > 0$ for all $r > b(\alpha)$.*

Proof. If $f(0) = 0$, then $(u, v) = (0, 0)$ is an equilibrium. Hence, $T = \phi$. If $f(0) > 0$ and there were $\alpha \in T$, then $u''(b(\alpha), \alpha) = -f(0) < 0$. Therefore, $u(r, \alpha) < 0$ for $r < b(\alpha)$ and sufficiently close to it, a contradiction. This proves $T = \phi$.

If $\alpha \in T$, then it is necessary that $f(0) < 0$. Since $H(u(b(\alpha), \alpha)) = 0$ implies that $H(u(r, \alpha)) < 0$ for all $r > b(\alpha)$, then $u''(b(\alpha), \alpha) = -f(0) > 0$ implies that $u(r, \alpha) > 0$ for all $r > b(\alpha)$. □

The following lemma plays a crucial role in the study of problem (P).

LEMMA 2.4. *If there is a sequence $\{\alpha_k\} \subset N \cup T$ such that*

$$\alpha_k \to \bar{\alpha} > 0 \quad \textit{and} \quad b(\alpha_k) \to \infty$$

as $k \to \infty$, then $\bar{\alpha} \in P$ and $u(\cdot, \bar{\alpha})$ satisfies the following monotonicity property:

(M) (i) *$u(r, \bar{\alpha})$ is either strictly increasing in (a, ∞) or there exists $a_1 > a$ such that $u(r, \bar{\alpha})$ is strictly increasing in (a, a_1) and strictly decreasing in (a_1, ∞).*

(ii) *$u(r, \bar{\alpha}) \to \tilde{u}$ as $r \to \infty$ where $f(\tilde{u}) = 0$.*

Proof. First, we observe that $u(\cdot, \bar{\alpha})$ cannot have a local maximum followed by a local minimum. Otherwise, by continuous dependence of ordinary differential equations (o.d.e.), for k sufficiently large, $u(r, \alpha_k)$ will have at least two local maxima in $(a, b(\alpha_k))$, a contradiction to Lemma 2.2(i). It is also clear that $u(r, \bar{\alpha})$ cannot be constant on any finite interval of (a, ∞). Hence, $u(\cdot, \bar{\alpha})$ satisfies (M)(i). Condition (M)(ii) follows by Lemma 2.6 which will be proved later.

As in [2], [3], and [17], it is sometimes convenient to study the existence problem in the form of (1.3), (1.4).

For $n \geqq 3$, in terms of variables

$$s = r^{2-n} \quad \text{and} \quad w(s) = u(r), \tag{2.10}$$

equations (2.1), (2.2) can be rewritten as

$$w''(s) + \rho(s) f(w(s)) = 0 \quad \text{in } (s_0, s_1), \tag{2.11}$$

$$w(s_0) = 0 = w(s_1), \tag{2.12}$$

where $\rho(s) = (n-2)^{-2} s^{-k}$, $k = (2n-2)/(n-2)$, $s_0 = b^{2-n}$, and $s_1 = a^{2-n}$. For $n = 2$, in terms of variables

$$s = \tfrac{1}{2} - \log a + \log r \quad \text{and} \quad w(s) = u(r),$$

equation (2.1) can also be written as (2.11) with $\rho(s) = a^2 e^{2s-1}$, $s_0 = \frac{1}{2}$ and $s_1 = -\frac{1}{2} - \log a + \log b$. In the remaining part of the section, we only treat the case $n \geqq 3$; the case $n = 2$ can also be treated analogously.

The associated initial value problem, now backward shooting in an s-variable, is

$$w''(s) + \rho(s) f(w(s)) = 0 \quad \text{for } s < s_1, \tag{2.13}$$

$$w(s_1) = 0 \quad \text{and} \quad w'(s_1) = -\beta, \tag{2.14}$$

where $\beta \geqq 0$ is the shooting parameter and $s_1 = a^{2-n}$ is a fixed number.

It is easy to check that (2.13), (2.14) is equivalent to

$$w(s) = \beta(s_1 - s) - \int_s^{s_1} (t-s)\rho(t) f(w(t))\, dt \quad \text{for } s < s_1, \tag{2.15}$$

and the solution $w(\cdot, \beta)$ also satisfies the following equation:

$$w(s) = w(\bar{s}) + w'(\bar{s})(s - \bar{s}) + \int_{\bar{s}}^{s} (t-s)\rho(t) f(w(t))\, dt \quad \text{for } 0 < s, \bar{s} < s_1. \tag{2.16}$$

The associated energy function V is defined by

$$V(s) \equiv V(w(s, \beta)) = \tfrac{1}{2} w'^2(s) + \rho(s) F(w(s)). \tag{2.17}$$

It is clear that

$$V'(s) = \rho'(s) F(w(s)),$$

and so

$$V(s) = V(\bar{s}) + \int_{\bar{s}}^{s} \rho'(t) F(w(t))\, dt \tag{2.18}$$

for $0 < s, \bar{s} < s_1$.

If w has a zero in $(0, s_1)$, denote

$$s_0(\beta) = \inf \{ s_0 : w(s, \beta) > 0 \text{ in } (s_0, s_1) \},$$

and $\nu(\beta) \in (s_0(\beta), s_1)$ satisfies

$$w(\nu(\beta), \beta) = \max \{ w(s, \beta) : s \in (s_0(\beta), s_1) \}.$$

With a modification of the argument used in Lin [17], we can prove that $s_0(\beta)$ and $\nu(\beta)$ are well defined for sufficiently large β and tend to s_1 as $\beta \to \infty$. For completeness, we also give a full proof here.

LEMMA 2.5. *If condition* (H2) *is satisfied, then* $s_0(\beta)$ *and* $\nu(\beta)$ *are well defined for sufficiently large* β. *Moreover,*

$$\lim_{\beta\to\infty} \nu(\beta) = s_1, \tag{2.19}$$

$$\lim_{\beta\to\infty} s_0(\beta) = s_1, \tag{2.20}$$

and

$$\lim_{\beta\to\infty} w(\nu(\beta), \beta) = \infty. \tag{2.21}$$

Proof. We first prove (2.19). If (2.19) were false, then there would be a point $\nu_0 \in (0, s_1)$ and a sequence $\beta_k \to \infty$ with

$$w_k(s) > 0 \quad \text{and} \quad w_k'(s) \leqq 0 \quad \text{in } (\nu_0, s_1), \tag{2.22}$$

where $w_k(s) = w(s, \beta_k)$.

Letting $\bar{s} = (\nu_0 + s_1)/2$, we claim that

$$\limsup_{k\to\infty} w_k(\bar{s}) = \infty. \tag{2.23}$$

Suppose this is not the case; then there exists a constant $M > 0$ such that

$$w_k(\bar{s}) \leqq M \quad \text{for all } k. \tag{2.24}$$

Now, by (2.16) and (2.24), we have

$$w_k(\bar{s}) = \frac{1}{2}\beta_k(s_1 - \nu_0) - \int_{\bar{s}}^{s} (t - \bar{s})\rho(t) f(w_k(t))\, dt \geqq \frac{1}{2}\beta_k(s_1 - \nu_0) - C,$$

for some constant $C \geqq 0$. But, by (2.24), this is impossible. Therefore, (2.23) holds.

By choosing a subsequence of β_k if necessary, we may assume that

$$\lim_{k\to\infty} w_k(\bar{s}) = \infty. \tag{2.25}$$

Denote

$$h_k(s) = f(w_k(s))/w_k(s)$$

in $(\nu_0, \bar{s})$ and

$$m_k = \inf\{h_k(s): s \in [\nu_0, \bar{s}]\}.$$

By (2.25) and (H2),

$$\lim_{k\to\infty} m_k = \infty. \tag{2.26}$$

Now $w_k''(s) + \rho(s)h_k(s)w_k(s) = 0$ in $(\nu_0, \bar{s})$ with $\rho(s)h_k(s) \geqq \rho(\bar{s})m_k$ in $(\nu_0, \bar{s})$. By (2.26) and the Sturm Comparison Theorem, w_k has zeros in $(\nu_0, \bar{s})$ for sufficiently large k. But by (2.22) this is impossible. This proves (2.19).

Next, we prove (2.21). By (2.18), we have

$$\frac{1}{2}\beta_k^2 = \rho(\nu_k)F(u(\nu_k)) + \int_{\nu_k}^{s_1} \rho'(t)F(w_k(t))\, dt,$$

where $\nu_k = \nu(\beta_k)$, which implies that $F(w_k(\nu_k)) \to \infty$ as $k \to \infty$. By (H2), (2.21) follows.

Finally, we prove (2.20). If (2.20) were false, then there would be a point $s_0 \in (0, s_1)$ and a sequence $\beta_k \to \infty$ with

$$w_k(s) > 0 \quad \text{in } (s_0, \nu_k). \tag{2.27}$$

Denote $\bar{s} = \frac{1}{2}(s_0 + s_1)$. By (2.19), we may assume that $\bar{s} < \nu_k$ for all k. We first claim that

$$\limsup_{k\to\infty} w_k(\bar{s}) < \infty. \tag{2.28}$$

Let

$$L_k = \min\{w_k(s): s \in [\bar{s}, \nu_k]\}.$$

Then, there exists $L > 0$ such that

$$L_k \leqq L \quad \text{for all } k. \tag{2.29}$$

Otherwise, by the Sturm Comparison Theorem again, w_k has a zero in $(\bar{s}, \nu_k)$, a contradiction to (2.27).

If $w_k(\bar{s}) = L_k$, then (2.28) holds.

If $w_k(\bar{s}) > L_k$, let $s_k \in (\bar{s}, \nu_k)$ such that $w_k(s_k) = L_k$. Denote $r_k = s_k^{1/(2-n)}$, $\bar{r} = \bar{s}^{1/(2-n)}$, and $u_k = w_k$. Then $u_k'(r_k) = 0$, and we have

$$H(u_k(r_k)) = F(L_k) \geqq H(u_k(\bar{r})) \geqq F(u_k(\bar{r})).$$

By (H2) and (2.29),

$$u_k(\bar{r}) \leqq M$$

for some constant $M > 0$. This proves (2.28).

By (H2), there exists $u^* > 0$ such that $f(u) > 0$ for all $u > u^*$. Denote

$$A_k = \{s \in (0, \tau_k): w_k(s) \leqq u^*\}.$$

Then by (2.16), we have

$$\begin{aligned} w_k(\nu_k) &= w_k(\bar{s}) + w_k'(\bar{s})(\nu_k - \bar{s}) + \int_{\bar{s}}^{\nu_k} (t - \nu_k)\rho(t) f(w_k(t))\, dt \\ &\leqq w_k(\bar{s}) + w_k'(\bar{s})(\nu_k - \bar{s}) + \int_{A_k} (t - \nu_k)\rho(t) f(w_k(t))\, dt \\ &\leqq w_k(\bar{s}) + w_k'(\bar{s})(\nu_k - \bar{s}) + C \end{aligned}$$

for some constant $C \geqq 0$. Hence, by (2.20), we have

$$\lim_{k\to\infty} w_k'(\bar{s}) = \infty. \tag{2.30}$$

On the other hand, by (2.16) again, we have

$$\begin{aligned} w_k(s_0) &= w_k(\bar{s}) + w_k'(\bar{s})(s_0 - \bar{s}) + \int_{\bar{s}}^{s_0} (t - s_0)\rho(t) f(w_k(t))\, dt \\ &\leqq w_k(\bar{s}) - \frac{1}{2} w_k'(\bar{s})(s_1 - s_0) - \int_{A_k} (t - s_0)\rho(t) f(w_k(t))\, dt \\ &\leqq w_k(\bar{s}) - \frac{1}{2}(s_1 - s_0) w_k'(\bar{s}) + C_1 \end{aligned}$$

for some constant $C_1 \geqq 0$. By (2.28) and (2.30), $w_k(s_0) \to -\infty$ as $k \to \infty$, a contradiction to (2.27). This proves (2.20). □

LEMMA 2.6. *If $u(r, \alpha) > 0$ for $r > r_0 \geqq a$, then*

$$\liminf_{r \to \infty} |u(r, \alpha) - Z| = 0, \tag{2.31}$$

where $Z = \{\tilde{u} \geqq 0 : f(\tilde{u}) = 0\}$ and $|u - Z| = \inf\{|u - \tilde{u}| : \tilde{u} \in Z\}$. In particular, if $\lim_{r \to \infty} u(r, \alpha) = \tilde{u} \geqq 0$, then $f(\tilde{u}) = 0$.

Proof. If (2.31) were false, then there would be an $\varepsilon > 0$ such that

$$|f(u(r, \alpha))| \geqq \varepsilon.$$

Denote $w(s, \beta) = u(r, \alpha)$. Then by (2.15)

$$|w(s, \beta)| \to \infty \quad \text{as } s \to 0^+.$$

This is impossible in viewing $w(s, \beta) > 0$ and Lemma 2.2(iii). □

3. Existence of large solutions when $f(0) \geqq 0$. *In this section we shall prove that if $(H1) \sim (H3)$ are satisfied, the answer to (P) is affirmative. We first prove the following lemma.*

LEMMA 3.1. *If $\alpha \in N$ and $\|u\| \geqq \gamma$ then*

$$\|u\| > \gamma, \tag{3.1}$$

where γ is in (1.6). *A similar result holds for $\alpha \in T$ with*

$$\max\{u(r, \alpha) : r \in [a, b(\alpha)]\} \geqq \gamma. \tag{3.2}$$

Proof. If $\alpha \in N$, by Lemma 2.2(i), there exists a unique $\tau(\alpha) \in (a, b(\alpha))$, such that

$$u(\tau(\alpha), \alpha) = \|u\|. \tag{3.3}$$

Let $r_1(\alpha) \in (\tau(\alpha), b(\alpha))$ such that

$$u(r_1(\alpha), \alpha) = u_*,$$

where u_* is in (1.5). Then

$$H(u(\tau(\alpha))) = F(u(\tau(\alpha))) > H(u(r_1(\alpha))) > F(u_*),$$

which implies

$$\int_{u_*}^{u(\tau(\alpha))} f(u)\, du > 0.$$

Hence $u(\tau(\alpha)) > \gamma$. This proves (3.1). By the same argument, we can obtain (3.1) if $\alpha \in T$ and (3.2) holds. □

LEMMA 3.2. *Assume conditions* (H1) ~ (H3) *are satisfied. Then*

$$N_1 = \{\alpha \in N : \|u(\cdot, \alpha)\| > \gamma\} \tag{3.4}$$

is a nonempty open set.

Proof. By Lemmas 2.3 and 2.5, N_1 is nonempty. By Lemma 3.1 and continuous dependence of o.d.e., N_1 is an open set. □

We can now prove the main result of this section.

THEOREM 3.3. *Assume conditions* (H1) ~ (H3) *are satisfied. Then for any $b > a > 0$, there exists a positive radial solution $u(r)$ of* (2.1), (2.2) *with $\|u\| > \gamma$.*

Proof. By Lemma 3.2, there exists $\alpha^* \geqq 0$ such that $N_1 \supset (\alpha^*, \infty)$ with $\alpha^* \notin N_1$.

By Lemma 2.5, it suffices to show that

$$\lim_{\alpha \to (\alpha^*)^+} b(\alpha) = \infty. \tag{3.5}$$

We shall prove the theorem according to $f(0) > 0$ and $f(0) = 0$.

If $f(0) > 0$, we claim that $\alpha^* > 0$. In fact, $u''(a, 0) = -f(0) < 0$. Hence, there is an $\varepsilon > 0$ such that $u(r, 0) < 0$ for $r \in (a, a+\varepsilon)$. This implies $\alpha^* > 0$. We claim that $\alpha^* \in P$. If $\alpha^* \notin P$, then $(0, \infty) = N \cup P$ implies $\alpha^* \in N$. Since $\alpha^* \in N_1$, we have $u(\tau(\alpha^*), \alpha^*) \geqq \gamma$. By Lemma 3.1, $\alpha^* \in N_1$, a contradiction. Therefore, $\alpha^* \in P$ and (3.5) follows.

If $f(0) = 0$, then either $\alpha^* > 0$ or $\alpha^* = 0$. If $\alpha^* > 0$, then the previous argument also works and then (3.5) holds. If $\alpha^* = 0$ and (3.5) are false, then there are $b_0 > a$ and $\delta > 0$ such that $b(\alpha) \leqq b_0$ for all $\alpha \in (0, \delta)$. Since $\tau(\alpha) \in (0, b_0)$ for all $\alpha \in (0, \delta)$, there exists a sequence $\alpha_k \to 0$ such that $\tau(\alpha_k) \to \tau_0 \in [0, b_0]$. Since $u(\tau(\alpha_k), \alpha_k) > \gamma$, we have $u(\tau_0, 0) \geqq \gamma$, a contradiction to $u(r, 0) \equiv 0$. Hence, (3.5) holds. □

COROLLARY 3.4. *Assume conditions* (H1)~(H3) *are satisfied and* $f(0) > 0$. *Then for any* $a > 0$, *the equation*

$$u''(r) + \frac{n-1}{r} u'(r) + f(u(r)) = 0 \quad \text{in } (a, \infty), \tag{3.6}$$

$$u(a) = 0 \quad \text{and} \quad u(r) > 0 \quad \text{for } r > a, \tag{3.7}$$

has a solution u *which satisfies* (M).

Proof. In the proof of the previous theorem, we have $N_1 \supset (\alpha^*, \infty)$ with $\alpha^* > 0$ and $\alpha^* \in P$. By (3.5) and Lemma 2.4, $u(\cdot, \alpha^*)$ satisfies (M). □

4. Nonexistence on wide annuli when f(0)<0. In this section we shall prove that if (H2) and (H3)′ are satisfied, then (2.1), (2.2) has no positive solution when $b - a$ is too large.

We first show that $P \neq \phi$ when $f(0) < 0$.

LEMMA 4.1. *If* $f(0) < 0$, *then there exists* $\alpha_* > 0$ *such that* $[0, \alpha_*) \subset P$.

Proof. We first prove $0 \in P$. In fact, $u''(a, 0) = -f(0) > 0$ and $H(u(a, 0)) = 0 > H(u(r, 0))$ for $r > a$ implies $u(r, 0) > 0$ for $r > a$. Hence $0 \in P$.

Next, let $u_0 > 0$ such that $f(u) < 0$ in $[-u_0, u_0]$. Then there exist $\varepsilon > 0$ and $\alpha_0 > 0$ such that for all $\alpha \in [0, \alpha_0]$, we have $|u(r, \alpha)| \leqq u_0$ in $[a, a+\varepsilon]$. Therefore, for all $\alpha \in [0, \alpha_0]$,

$$u''(r) + \frac{n-1}{r} u'(r) > 0 \quad \text{in } [a, a+\varepsilon],$$

which implies $u(r, \alpha) > 0$ in $(a, a+\varepsilon]$.

On the other hand, if $H(u(a+\varepsilon, 0)) < 0$, by continuous dependence of o.d.e., there exists $\alpha_* \in (0, \alpha_0)$ such that $H(u(a+\varepsilon, \alpha)) < 0$ for all $\alpha \in (0, \alpha_*)$. Therefore, for any $\alpha \in (0, \alpha_*)$, $H(u(r, \alpha)) < 0$ for $r > a+\varepsilon$, which implies $u(r, \alpha) > 0$ for $r > a+\varepsilon$. This proves $(0, \alpha_*) \subset P$. □

We now prove the main result of this section.

THEOREM 4.2. *Assume conditions* (H2) *and* (H3)′ *are satisfied. Then there exists* $b^* > a$ *such that for any* $b > b^*$, (2.1), (2.2) *has no positive solution.*

Proof. By Lemma 2.5, there exist $\alpha^* > 0$ and $b_0 > a$ such that $(\alpha^*, \infty) \subset N \cup T$ and $b(\alpha) \leqq b_0$ for all $\alpha \in (\alpha^*, \infty)$. On the other hand, by Lemma 4.1, there exists $\alpha_* > 0$ such that $[0, \alpha_*) \subset P$. Therefore, it suffices to show that there exists $\tilde{b}_0 > a$ such that

$$b(\alpha) \leqq \tilde{b}_0 \quad \text{for any } \alpha \in [\alpha_*, \alpha^*] \cap (N \cup T). \tag{4.1}$$

If (4.1) were false, then there would be a sequence $\alpha_k \in [\alpha_*, \alpha^*] \cap (N \cup T)$ and $\bar{\alpha} \in [\alpha_*, \alpha^*]$ such that $\alpha_k \to \bar{\alpha}$ and $b(\alpha_k) \to \infty$ as $k \to \infty$. By Lemma 2.4, $\bar{\alpha} \in P$ and $u(\cdot, \bar{\alpha})$ satisfies (M). Now, (H3)'(ii) and $u(r, \bar{\alpha}) \to \tilde{u}$ as $r \to \infty$ with $f(\tilde{u}) = 0$ implies $H(u(r, \bar{\alpha})) \to F(\tilde{u}) < 0$ as $r \to \infty$. Therefore, there exists $r_0 > a$ such that $H(u(r_0, \bar{\alpha})) < 0$. By continuous dependence of o.d.e., we have $H(u(r_0, \alpha_k)) < 0$ for k sufficiently large, a contradiction to Lemma 2.2(ii). This proves (4.1). □

5. Multiplicity results on wide annuli. In the previous sections we studied the existence of large (and nonexistence of) positive solutions for (2.1), (2.2) under the various assumptions of f. In this section we shall study the existence of "intermediate size" solutions of (2.1), (2.2) when f may change signs several times and satisfies condition (f_3) in § 1, i.e., f satisfies the following hypothesis:

(H4) there exist m successive numbers $\bar{u}_m > \bar{u}_{m-1} > \cdots > \bar{u}_1 > 0$, which satisfy

(i) $f(\bar{u}_k) = 0$ for $k = 1, \cdots, m$; and

(ii) $0 < \max\{F(s) : 0 \leqq s \leqq \bar{u}_{k-1}\} < F(\bar{u}_k)$ for $k = 2, \cdots, m$.

Let γ_k be the least number in $(\bar{u}_{k-1}, \bar{u}_k)$ such that

$$\int_{\bar{u}_{k-1}}^{\gamma_k} f(u)\, du = 0, \tag{5.1}$$

for $k = 1, \cdots, m$, where $\bar{u}_0 \equiv 0$. We first prove the following lemma.

LEMMA 5.1. *Assume there exists an $\bar{u} > 0$ such that*

$$f(u) = 0 \quad \text{for } u \geqq \bar{u}. \tag{5.2}$$

Then we have the following conditions:

(i) *if $u(r_1, \alpha) = \bar{u}$ and $u'(r_1, \alpha) \geqq 0$ for some $r_1 > a$, then for $r > r_1$,*

$$u(r, \alpha) = \bar{u} + \frac{1}{n-2} r_1 u'(r_1, \alpha) - \frac{1}{n-2} r_1^{n-1} u'(r_1, \alpha) r^{2-n}; \tag{5.3}$$

(ii) *let $U = \{\alpha \in (0, \infty) : u(r_1, \alpha) = \bar{u}$ for some $r_1 > a\}$; then there exists $\alpha^* > 0$ such that $U \supseteq (\alpha^*, \infty)$.*

Proof. (i) By (5.2), we have

$$u''(r) + \frac{n-1}{r} u'(r) = 0 \tag{5.4}$$

as long as $u(r, \alpha) \geqq \bar{u}$. Therefore, by solving (5.4) with initial condition $u(r_1, \alpha) = \bar{u}$ and $u'(r_1, \alpha) \geqq 0$, (5.3) follows.

(ii) We shall prove (ii) by using the method of backward shooting. If (ii) were false, there would be a sequence $\beta_k \to \infty$ such that

$$w_k(s) < \bar{u} \quad \text{as long as } w_k \text{ remain positive}, \tag{5.5}$$

where $w_k(s) \equiv w(s, \beta_k)$. Let

$$\nu_k = \inf\{\tilde{s} \in (0, s_1) : w_k(s) > 0 \text{ and } w_k'(s) \leqq 0 \text{ in } (\tilde{s}, s_1)\}.$$

We claim that

$$\lim_{k \to \infty} \nu_k = s_1. \tag{5.6}$$

If (5.6) were false, there would be $\nu_0 < s_1$ and a subsequence of ν_k (for simplicity,

rename it ν_k), such that $\nu_k \leqq \nu_0$ for all k. Therefore, by (2.15) and (5.5), we have

$$w_k(\nu_0) = \beta_k(s_1 - \nu_0) - \int_{\nu_0}^{s_1} (t - \nu_0)\rho(t)f(w_k(t))$$
$$\geqq \beta_k(s_1 - \nu_0) - C,$$

for some constant $C \geqq 0$, a contradiction to (5.5). Hence (5.6) holds.

On the other hand, by (5.5) again, we have

$$w'(\tau_k) = -\beta_k + \int_{\tau_k}^{s_1} \rho(t)f(w_k(t))\, dt$$
$$\leqq -\beta_k + C$$

for some $C \geqq 0$. Therefore, $w'(\tau_k) < 0$ if k is large enough, a contradiction to the definition of ν_k. This proves (ii). □

The (energy) functional we want to minimize is

$$J(u) = \int_a^b r^{n-1}\left\{\frac{1}{2}u'^2(r) - F(u(r))\right\} dr$$

in $H_0^1((a, b))$, where $H_0^1((a, b)) = \{u : u$ is absolutely continuous in $[a, b]$ with $u(a) = 0 = u(b)$ and $u, u' \in L^2(a, b)\}$.

Since f may change signs, the minimizer u_b of J is not necessarily positive. However, for fixed a, if b is sufficiently large and (H4) is satisfied, then we can prove u_b is positive. To make the proof more transparent, we begin with the study of two simple cases.

LEMMA 5.2. *If f satisfies the following*:

(H5)(i) $f(0) \geqq 0$;

(ii) *There exists $\bar{u} > 0$ such that $f(\bar{u}) = 0$ and $f(u) > 0$ in $(0, \bar{u})$, then, we have the following results*:

(i) *There exists $b^* > a$ such that for any $b > b^*$, (2.1), (2.2) has a positive solution u_b that is also a local minimizer of $J(u)$ over $H_0^1((a, b))$. Moreover,*

$$0 < u_b < \bar{u} \quad \text{in } (a, b), \tag{5.7}$$

and

$$-\frac{b^n}{n}F(\bar{u}) \leqq J(u_b) \leqq -\frac{b^n}{n}F(\bar{u}) + C(b^{n-1} + 1) \tag{5.8}$$

for some positive constant C which is independent of b.

(ii) *If $f(0) > 0$, then there exists a positive solution $\tilde{u}$ of* (3.6), (3.7) *and $\tilde{u}$ is strictly increasing in (a, ∞) and $\tilde{u}(r) \to \bar{u}$ as $r \to \infty$.*

Proof. We first modify the function outside $[0, \bar{u}]$ as Clément and Sweers did in [6]. Denote

$$f_1(u) = \begin{cases} 0 & u \geqq \bar{u}, \\ f(u) & u \in [0, \bar{u}], \\ 2f(0) - f(-u) & \text{if } u < 0, \end{cases} \tag{5.9}$$

$$F_1(u) = \int_0^u f_1(t)\, dt \quad \text{and} \quad J_1(u) = \int_a^b r^{n-1}\left\{\frac{1}{2}u'^2(r) - F_1(u(r))\right\} dr.$$

It is easy to verify that

$$F_1(|u|) = F_1(u) + 2f(0)|u| \quad \text{for } u < 0. \tag{5.10}$$

Hence, for any $u \in H_0^1((a, b))$,

$$J_1(|u|) \leqq J_1(u). \tag{5.11}$$

Since $f(u)$ is bounded, the minimizer of $J_1(u)$ over $H_0^1((a, b))$ is achieved, say u_b, which is a solution of (2.1), (2.2). By (5.11), u_b can be chosen to be nonnegative. If $u_b \not\equiv 0$ in (a, b), then by Lemma 2.3, $u_b > 0$ in (a, b).

If $f(0) > 0$, then $u_b > 0$ in (a, b). If $f(0) = 0$, we want to prove that $u_b > 0$ in (a, b) if b is large enough. This can be done by choosing appropriate test functions $u_b^* \in H_0^1((a, b))$ as follows:

$$u_b^*(r) = \begin{cases} (r-a)\bar{u} & \text{for } r \in [a, a+1], \\ \bar{u} & \text{for } r \in [a+1, b-1], \\ (b-r)\bar{u} & \text{for } r \in [b-1, b]. \end{cases} \tag{5.12}$$

Then

$$J_1(u_b^*) \leqq -\frac{b^n}{n} F(\bar{u}) + C(b^{n-1}+1) \tag{5.13}$$

for some constant C which is independent of b. Therefore, if b is large enough, then $u_b > 0$ in (a, b), and by Lemma 5.1(i), $u_b < \bar{u}$ in (a, b). By (5.13), (5.8) follows. This proves (i).

To prove (ii), we first note that $f(0) > 0$ implies there exists $\alpha_* > 0$ such that $(0, \alpha_*) \subset N$ and

$$\lim_{\alpha \to 0^+} b(\alpha) = a \tag{5.14}$$

(see Lin [18]). On the other hand, by Lemma 5.1(ii), there exists $\alpha^* > \alpha_*$ such that $U \supset (\alpha^*, \infty)$. Therefore, for any $b > b^*$, there exists $\alpha(b) \in (0, \alpha^*]$ such that $u(\cdot, \alpha(b))$ is a minimizer of $J_1(u)$. Therefore, by (5.14) there exists $\bar{\alpha} > 0$ and a sequence $b_k \to \infty$ such that $\alpha(b_k) \to \bar{\alpha}$ as $k \to \infty$. By Lemma 2.4, $\bar{\alpha} \in P$ and $u(\cdot, \bar{\alpha})$ satisfies (M). By Lemma 5.1(i) and (H5)(ii), $u(\cdot, \bar{\alpha})$ is strictly increasing in (a, ∞). This proves (ii). □

Next, we treat the case $f(0) < 0$.

LEMMA 5.3. *If f satisfies the following conditions*:

(H5)′(i) $f(0) \leqq 0$;

(ii) *There exist* $\bar{u} > \underline{u} > 0$ *such that* $f(\underline{u}) = f(\bar{u}) = 0$ *and* $f(u) < 0$ *in* $(0, \underline{u})$ *and* $f(u) > 0$ *in* $(\underline{u}, \bar{u})$;

(iii) $\int_0^{\bar{u}} f(u)\, du > 0$,

then there exists $b^* > a$ *such that for any* $b > b^*$ *there exists a positive solution* u_b *of* (2.1), (2.2) *with* $\|u_b\| \in (\gamma, \bar{u})$, *where*

$$\int_0^{\gamma} f(u)\, du = 0. \tag{5.15}$$

Proof. If $f(0) = 0$, then the arguments in Lemma 5.2 also work and give the result as in Lemma 5.2(i). Note that (H5)′(iii) implies $J_1(u_b^*) < 0$ in (5.13) when b is large enough.

If $f(0) < 0$, then (5.10) implies the extension f_1 in (5.9) is no longer suitable to the minimization problem. Therefore, we want to modify f in a different way and use super- and subsolution methods to obtain solutions for (2.1), (2.2).

Since $f(0) < 0$, we can extend f into $(u_0, 0)$ such that

$$f(u_0) = 0, f(u) < 0 \quad \text{in } (u_0, 0) \tag{5.16}$$

and

$$\int_{u_0}^{\bar{u}} f(u)\,du > 0. \tag{5.17}$$

Let $v = u - u_0$ and denote $v_1 = \underline{u} - u_0$ and $v_2 = \bar{u} - u_0$. Let

$$g(v) = f(u) \quad \text{in } [0, v_2].$$

Then $g(0) = g(v_1) = g(v_2) = 0$,

$$g(v) < 0 \quad \text{in } (0, v_1) \quad \text{and} \quad g(v) > 0 \quad \text{in } (v_1, v_2).$$

We then extend g outside $[0, v_2]$ by making

$$g(v) = 0 \quad \text{for } v > v_2$$

and

$$g(v) = -g(-v) \quad \text{for } v < 0.$$

Denote

$$G(v) = \int_0^v g(t)\,dt.$$

Then, as in (5.10),

$$G(|v|) = G(v) \quad \text{for all } v < 0$$

and (5.17) can be rewritten as

$$G(v_2) = \int_0^{v_2} g(t)\,dt = \int_{u_0}^{\bar{u}} f(u)\,du > 0. \tag{5.18}$$

Define

$$\tilde{J}(v) = \int_a^b r^{n-1}\left\{\frac{1}{2}v'^2(r) - G(v(r))\right\} dr \tag{5.19}$$

in $H_0^1((a, b))$.

Let v_b^* be defined as u_b^* in (5.12), but replace $\bar{u}$ by v_2. Then

$$\tilde{J}(v_b^*) \leqq -\frac{b^n}{n} G(v_2) + C(b^{n-1} + 1) \tag{5.20}$$

for some positive constant C which is independent of b. Therefore, there exists $b^* > a$ such that for any $b > b^*$, the equations

$$v''(r) + \frac{n-1}{r} v'(r) + g(v(r)) = 0 \quad \text{in } (a, b), \tag{5.21}$$

$$v(a) = 0 = v(b) \tag{5.22}$$

have a positive solution v_b which is also a minimizer of $\tilde{J}$ with

$$\tilde{J}(v_b) \leqq -\frac{b^n}{n} G(v_2) + C(b^{n-1} + 1)$$

and

$$\|v_b\| \in (\gamma - u_0, v_2). \tag{5.23}$$

Let $\tilde{u}_b = u_0 + v_b$; then

$$\Delta\tilde{u}_b + f(\tilde{u}_b) = \Delta v_b + g(v_b) = 0 \quad \text{in } a < |x| < b,$$

and

$$\tilde{u}_b = u_0 < 0 \quad \text{on } |x| = a \quad \text{and} \quad |x| = b.$$

Hence, $\tilde{u}_b$ is a subsolution of (2.1), (2.2). Since $\bar{u}$ is a supersolution and $\bar{u} > \tilde{u}_b$, by monotone iteration scheme (see, e.g., [11], [18]), there is a positive solution u_b of (2.1), (2.2) and u_b satisfies $\tilde{u}_b < u_b < \bar{u}$, which also implies $\|u_b\| \in (\gamma, \bar{u})$. □

Now, we can prove the following multiplicity result for general case.

THEOREM 5.4. *Assume condition* (H4) *is satisfied. Then, we have the following results*:

(i) *If* $f(0) \geqq 0$, *then there exists* $b^* < a$ *such that for any* $b > b^*$, (2.1), (2.2) *has at least* $2m-1$ *positive solutions* $\tilde{u}_1, \underset{\sim}{u}_2, \tilde{u}_2, \cdots, \underset{\sim}{u}_m, \tilde{u}_m$ *with* $\tilde{u}_{k-1} < \tilde{u}_k$ *and* $\underset{\sim}{u}_k < \tilde{u}_k$ *in* (a, b), *and* $\|\underset{\sim}{u}_k\|, \|\tilde{u}_k\| \in (\gamma_k, \bar{u}_k)$ *for* $k = 2, \cdots, m$.

(ii) *If* $f(0) < 0$, *then there exists* $b^* > a$ *such that for any* $b > b^*$, (2.1), (2.2) *has at least* m *positive solutions* $\tilde{u}_1 < \cdots < \tilde{u}_m$ *with* $\|\tilde{u}_k\| \in (\gamma_k, \bar{u}_k)$ *for* $k = 1, \cdots, m$.

Proof. We shall prove the theorem by induction on m.

For $m = 1$, the results were proved in Lemmas 5.2 and 5.3 under the conditions (H5) and (H5)′, respectively. The arguments used in the last two lemmas are also valid for general cases; thus the details are omitted.

We first study the case $f(0) \geqq 0$. For $j = 2, \cdots, m$, denote

$$f_j(u) = \begin{cases} f(u) & \text{for } u \in [0, \bar{u}_j], \\ 0 & \text{for } u \in [\bar{u}_j, \infty), \\ 2f(0) - f(-u) & \text{for } u < 0, \end{cases}$$

$$F_j(u) = \int_0^u f_j(t)\, dt,$$

and

$$J_j(u) = \int_a^b r^{n-1} \left\{ \frac{1}{2} u'^2(r) - F_j(u(r)) \right\} dr.$$

It is clear that f_{j+1} is an extension of f_j and

$$\text{if } \|u\| < \bar{u}_j, \quad \text{then } \bar{J}_{j+1}(u) = J_j(u) \tag{5.24}$$

for $j = 1, \cdots, m-1$.

Assume $m = j (\geqq 1)$ is true. Then there exists a $b_j^* > a$ such that for any $b > b_j^*$, $J_j(u)$ has a minimizer $\tilde{u}_{j,b}$ which is a positive solution of (2.1), (2.2) and satisfies

$$\|\tilde{u}_{j,b}\| \in (\gamma_j, \bar{u}_j) \tag{5.25}$$

and

$$-\frac{b^n}{n} F_j(\bar{u}_j) \leqq J_j(\tilde{u}_{j,b}) \leqq -\frac{b^n}{n} F_j(\bar{u}_j) + C_j(b^{n-1} + 1) \tag{5.26}$$

for some positive constant C_j that is independent of b.

Let $u_{j+1,b}^*$ be as in (5.12) with $\bar{u}$ replaced by $\bar{u}_{j+1}$. Then

$$J_{j+1}(u_{j+1,b}^*) \leqq -\frac{b^n}{n} F_{j+1}(\bar{u}_{j+1}) + C_{j+1}(b^{n-1} + 1) \tag{5.27}$$

for some positive constant C_{j+1} independent of b. Therefore, by (5.24) ~ (5.27), there exists $\bar{b}^*_{j+1} \geqq b^*_j$, such that for any $b > \bar{b}^*_{j+1}$, minimizer $\tilde{u}_{j+1,b}$ of J_{j+1} satisfies

$$J_{j+1}(\tilde{u}_{j+1,b}) < J_j(\tilde{u}_{j,b}).$$

Hence, $\tilde{u}_{j+1,b} \neq \tilde{u}_{j,b}$ and $\|\tilde{u}_{j+1,b}\| \in (\gamma_{j+1}, \bar{u}_{j+1})$.

Let

$$N_{j+1} = \{\alpha \in N: \gamma_{j+1} < \|u\| < \bar{u}_{j+1}\}.$$

Then N_{j+1} is an open set and nonempty according to the last paragraph. Therefore, by Lemma 5.1, there exist two positive numbers $\bar{\alpha}_{j+1} > \underline{\alpha}_{j+1}$, such that $(\underline{\alpha}_{j+1}, \bar{\alpha}_{j+1}) \subset N_{j+1}$, $\underline{\alpha}_{j+1} \notin N_{j+1}$ and $\bar{\alpha}_{j+1} \notin N_{j+1}$. By Lemma 2.4, $\underline{\alpha}_{j+1}$ and $\bar{\alpha}_{j+1}$ belong to P. Hence,

$$\lim_{\alpha \to (\underline{\alpha}_{j+1})^+} b(\alpha) = \infty = \lim_{\alpha \to (\bar{\alpha}_{j+1})^-} b(\alpha),$$

and then both $u(\cdot, \underline{\alpha}_{j+1})$ and $u(\cdot, \bar{\alpha}_{j+1})$ satisfy (M). Therefore, there exists $b^*_{j+1} \geqq \bar{b}^*_{j+1}$, such that for any $b > b^*_{j+1}$, (2.1), (2.2) has at least two positive solutions having maximum value in $(\gamma_{j+1}, \bar{u}_{j+1})$. Since $\bar{u}_{j+1}$ is a supersolution, there exists the maximum positive solution $\tilde{u}_{j+1}$ of (2.1), (2.2) having maximum value in $(\gamma_{j+1}, \bar{u}_{j+1})$. This proves (i).

Condition (ii) can be proved by using the arguments used in (i) and Lemma 5.3; thus details are omitted. □

In the proof of last theorem, we obtain the following results for (3.6), (3.7).

COROLLARY 5.5. *Assume condition* (H4) *is satisfied. Then, we have the following results*:

(i) *If* $f(0) > 0$, *then there exist at least* $2m-1$ *positive solutions* $\tilde{u}_1, \underline{u}_2, \tilde{u}_2, \cdots, \tilde{u}_m$ *of* (3.6), (3.7) *and* $\tilde{u}_j(r) \to \bar{u}_j$ *as* $r \to \infty$ *for* $j = 1, \cdots, m$.

(ii) *If* $f(0) \leqq 0$, *then there exist at least* $2m-2$ *positive solutions* $\underline{u}_2, \tilde{u}_2, \cdots, \tilde{u}_m$ *of* (3.6), (3.7) *and* $\tilde{u}_j(r) \to \bar{u}_j$ *as* $r \to \infty$ *for* $j = 2, \cdots, m$.

REFERENCES

[1] A. AMBROSETTI AND P. HESS, *Positive solutions of asymptotically linear elliptic eigenvalue problems*, J. Math. Anal. Appl., 73 (1980), pp. 411-422.

[2] C. BANDLE, C. V. COFFMAN, AND M. MARCUS, *Nonlinear elliptic problems in annular domains*, J. Differential Equations, 69 (1987), pp. 322-345.

[3] C. BANDLE AND M. K. KWONG, *Semilinear elliptic problems in annular domains*, J. Appl. Math. Phys., 40 (1989), pp. 245-257.

[4] H. BREZIS AND L. NIRENBERG, *Positive solutions of nonlinear elliptic equations involving critical Sobolev exponents*, Comm. Pure Appl. Math., 36 (1983), pp. 437-477.

[5] K. J. BROWN AND H. BUDIN, *On the existence of positive solutions for a class of semilinear elliptic boundary value problems*, SIAM J. Math. Anal., 10 (1979), pp. 875-883.

[6] P. CLEMENT AND G. SWEERS, *Existence and multiplicity results for a semilinear elliptic equation*, Ann. Scuola Norm. Sup. Pisa, 14 (1987), pp. 97-121.

[7] C. V. COFFMAN, *A nonlinear boundary value problem with many positive solutions*, J. Differential Equations, 54 (1984), pp. 429-437.

[8] C. V. COFFMAN AND M. MARCUS, *Existence and uniqueness results for semilinear Dirichlet problems in annuli*, preprint.

[9] C. COSNER AND K. SCHMITT, A priori *bounds for positive solutions of a semilinear elliptic equation*, Proc. Amer. Math. Soc., 95 (1985), pp. 47-50.

[10] E. N. DANCER, *Multiple fixed points of positive mappings*, J. Reine Angew. Math., 352 (1986), pp. 46-66.

[11] E. N. DANCER AND K. SCHMITT, *On positive solutions of semilinear elliptic equations*, Proc. Amer. Math. Soc., 101 (1987), pp. 445-452.

[12] D. G. DE FIGUEIREDO, *On the existence of multiple ordered solutions of nonlinear eigenvalue problems*, Nonlinear Anal., T. M. 7 A, 11 (1987), pp. 481-492.

[13] X. GARAIZAR, *Existence of positive radial solutions for semilinear elliptic equations in the annulus*, J. Differential Equations, 70 (1987), pp. 69-92.

[14] P. HESS, *On multiple positive solutions of nonlinear elliptic equations*, Comm. Partial Differential Equations, 6 (1981), pp. 951-961.

[15] J. L. KAZDAN AND F. W. WARNER, *Remarks on some quasilinear elliptic equations*, Comm. Pure Appl. Math., 28 (1975), pp. 567-597.

[16] S. S. LIN, *On non-radially symmetric bifurcation in the annulus*, J. Differential Equations, 80 (1989), pp. 251-279.

[17] ———, *On the existence of positive radial solutions for nonlinear elliptic equations in annular domains*, J. Differential Equations, 81 (1989), pp. 221-233.

[18] ———, *Positive radial solutions and non-radial bifurcation for semilinear elliptic equations in annular domains*, J. Differential Equations, 86 (1990), pp. 367-391.

[19] ———, *Existence of positive non-radial solutions for nonlinear elliptic equations in annular domains*, Trans. Amer. Math. Soc., to appear.

[20] Z. NEHARI, *On a class of nonlinear second order differential equations*, Trans. Amer. Math. Soc., 95 (1960), pp. 101-123.

[21] W.-M. NI AND NUSSBAUM, *Uniqueness and non-uniqueness for positive radial solution of* $\Delta u + f(u, r) = 0$, Comm. Pure Appl. Math., 38 (1985), pp. 67-108.

[22] T. SUZUKI AND K. NAGASSAKI, *Lifting of local subdifferentiations and elliptic boundary value problems on symmetric domains*, I, Proc. Japan Acad., 64 (1988), pp. 1-4.

[23] ———, *On the nonlinear eigenvalue problem* $\Delta u + \lambda e^u = 0$, Trans. Amer. Math. Soc., 309 (1988), pp. 591-608.

[24] S. H. WANG AND N. D. KAZARINOFF, *On positive solutions of some nonlinear eigenvalue problems*, preprint.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1516-1551, November 1991

003

PARTIAL REGULARITY IN PROBLEMS MOTIVATED BY NONLINEAR ELASTICITY*

NICOLA FUSCO† AND J. HUTCHINSON‡

Abstract. Regularity is proven almost everywhere for minimisers of problems motivated by nonlinear elasticity. Model problems treated include

$$\int_\Omega |Du|^2 + |\det Du|^2,$$

where $u : \Omega (\subset \mathbb{R}^2) \to \mathbb{R}^2$, and

$$\int_\Omega |Du|^2 + |Du|^s + |AdDu|^s + |\det Du|^s,$$

where $u : \Omega (\subset \mathbb{R}^3) \to \mathbb{R}^3$ with $s > 2$.

In particular, continuity of minimisers is not assumed a priori. "Degenerate" convexity of the integrand in the higher minors M of Du is also allowed, in the sense that second derivatives in M may approach zero as $M \to 0$.

Key words. nonlinear elasticity, partial regularity, elliptic systems

AMS(MOS) subject classifications. 35J50, 35J60, 73C50

1. Introduction.

Partial regularity for minimisers of functionals of the form

$$I[u] = \int_\Omega F(u) \quad \left(\text{or} \int_\Omega F(x, u, Du) \right), \tag{1.1}$$

where $\Omega \subset \mathbb{R}^n$ and $u : \Omega \to \mathbb{R}^N$, and F is convex in Du, has been extensively studied. See [G] and the references therein.

Many natural problems, however, only require that F be *quasi-convex* (see [B], [E], and [G]). Existence theory in such cases is now well understood; see [G] for references and [AF] for a nearly optimal result. Partial regularity results were recently obtained by Evans [E] and extended in [FH] and [GM].

All known natural examples of quasi-convex functionals (such as in nonlinear elasticity theory; see [B]) are, in fact, *polyconvex.* Recall that a polyconvex functional in Du is a convex function of the various minors of the matrix [Du]. Polyconvexity implies quasiconvexity [M]. While the converse is not true, the known counter examples are quite pathological. It is natural, then, to investigate the regularity theory for polyconvex functionals.

In this paper we prove new partial regularity results for various polyconvex functionals, rather than for general quasi-convex functionals. Some model problems that are treated include

$$I[u] = \int_\Omega |Du|^2 + |\text{deT } Du|^2 \tag{1.2}$$

* Received by the editors April 6, 1990; accepted for publication (in revised form) November 26, 1990.

† Dipartimento di Matematica, Via Mezzocannone 8, 80134 Napoli, Italy.

‡ Department of Mathematics, Australian National University, GPO Box 4, Canberra ACT 2601, Australia. The work of this author was largely carried out at the University of Salerno, Italy, under the auspices of the "Comitato Nazionale per le Scienze Matematiche."

in the case where $n = N = 2$, and

$$I[u] = \int_\Omega |Du|^2 + |Du|^s + |\text{Ad } Du|^s + |\det Du|^s \tag{1.3}$$

for $s > 2$, in the case where $n = N = 3$.

Since $|\text{Ad } Du|$ is the Euclidean norm of the map $|Du \wedge Du|$ which sends two-forms to two-forms, and $|\det Du|$ is the corresponding norm in (1.3) for maps on three-forms and in (1.2) for maps on two-forms, the preceding functionals are natural geometric generalisations of the Dirichlet energy. More generally, the linear map $\Lambda_k Du(x)$ sends k-vectors in $\mathbb{R}^n$ to k-vectors in $\mathbb{R}^N$, and the partial regularity result we prove includes the case of the functional

$$I[u] = \int_\Omega |Du|^2 + |Du|^s + \sum_{i=2}^{k} |\Lambda_i Du|^s, \tag{1.4}$$

where $s \geqq 2$ if $n = 2$, and $s > n - 1$ if $n \geqq 3$.

In order to simplify the exposition, which is already quite long, we restrict ourselves to problems of the form

$$I[u] = \int_\Omega F^1(Du) + \sum_{i=2}^{k} F^i(\Lambda_i Du), \tag{1.5}$$

where each $F^i(q)$ has the same growth rate for large q. It is clear, however, that the methods apply to other problems with differing growth rates.

The main result (Theorem 4.1) is that minimisers of $I[u]$ are $C^{1,\alpha}$ except on a closed set of measure zero, for each $0 < \alpha < 1$. Higher regularity follows from the Euler–Lagrange equations by standard bootstrapping arguments.

We now comment on some aspects of the functionals that we consider, as well as on the proof of the theorem.

The set of those competing functions for which the integrand is finite is not a linear space, even for (1.2). This corresponds to the fact that, for 2×2 matrices A and B, we have

$$\det (A + B) = \det A + \det B + A_{11}B_{22} + A_{22}B_{11} - A_{12}B_{21} - A_{21}B_{12},$$

and so we cannot improve on the powers in the estimate

$$\int_\Omega (\det (Du + Dv))^2 \leqq c \int_\Omega ((\det Du)^2 + (\det Dv)^2 + |Du|^4 + |Dv|^4).$$

It follows that if ϕ is a function obtained by smoothly interpolating between u and v, then we cannot normally estimate $I[\phi]$ in terms of $I[u]$ and $I[v]$. In (5.42) we use a new comparison function construction that does enable us to obtain estimates of this form.

In order to include the model problem (1.2) in the theory developed by Evans [E], it would be necessary to include a term of the form $\varepsilon |Du|^4$, for some $\varepsilon > 0$, in the integrand in order to control the determinant term. Similar remarks apply to the more general class of integrands covered in Theorem 4.1.

Note that minimisers are not a priori continuous. Also, the structure conditions allow (in fact require, but that is not necessary, as noted in § 7.2) degenerate ellipticity in the higher-order minors.

The proof of the main theorem involves a blowup argument. As usual, we obtain a sequence of "bad" balls $B(x_m, r_m)$ and a corresponding sequence of normalised

functions v_m defined on the unit ball. We then show that each v_m minimises a certain normalised functional obtained not, as would normally be the case, by linearising about $(Du)_{x_m,r_m}$, but (in the case of (1.3), for example) by linearising the three terms in the integrand about $(Du)_{x_m,r_m}$, $(\mathrm{Ad}\, Du)_{x_m,r_m}$, and $(\det Du)_{x_m,r_m}$, respectively. The proof of this makes essential use of the fact that each minor can be written in divergence form.

The blowup argument is used to prove a decay estimate on the quantity

$$U(x,r)=\fint_{B(x,r)}\Big(|Du-(Du)_{x,r}|^2+|Du-(Du)_{x,r}|^s + \sum_{i=2}^{k}|\Lambda_i(Du)_{x,r}|^{s-2}|\Lambda_i(Du-(Du)_{x,r})|^2+\sum_{i=2}^{k}|\Lambda_i(Du-(Du)_{x,r})|^s\Big), \tag{1.6}$$

where s is the common growth rate of the $F^i(q)$ in (1.5). The first two terms in (1.6) are as in [E], the fourth term is accounted for by the growth rate of the F^i, and the third term is necessary to handle the degenerate ellipticity corresponding to the F^i term. This particular choice of third term was motivated by the quantity $U(x, r)$ shown to decay in [FH2], where an analogous degeneracy problem occurred.

To prove the decay lemma, we first establish that the normalised functions v_m converge to a limit function v in various weak ways (cf. (5.11)), and that v satisfies the linear equation (5.29). In order to obtain the required decay estimate on $U(x, r)$ we establish various strong convergences of the v_m in §§ 5C and 5D. The key point is the construction of suitable comparison functions in (5.42), using the method mentioned previously. Detailed estimates using the minimising properties of the v_m then establish the required forms of strong convergence in (5.64)–(5.67).

We remark that the proof avoids any use of Caccioppoli inequalities, and strongly uses the fact that u is a minimiser of $I[\cdot]$ rather than merely a stationary point. The guiding principle behind the proof is that weak convergence of minimisers should imply strong convergence.

For completeness, we also prove some results of Ball. Although he only considers the cases where $n = N = 2, 3$, the extensions are straightforward. In § 2.4 we show how the elementary symmetric functions of the singular values of Du give rise to polyconvex functions (cf. [B, Thm. 5.1]). The question of regularity of minimisers of such functionals has been raised recently by De Giorgi. In § 3 we establish the existence of minimisers for the class of problems considered in this paper (cf. [B, Thm. 7.7]). Finally, in § 7 we indicate various extensions that can be made to our results.

Under the assumptions with which we work, the weak and pointwise determinants agree (see Lemma 2.2.1), and so our spaces include the ones introduced in [GMS]. On the other hand, we do not know if the two spaces coincide (see also Remarks 2.3.2). However, if we restrict to the closure in $\Lambda_k W^{1,s}$ of C^1 functions, by modifying the proofs somewhat, we are still able to establish the main result, Theorem 4.1. The essential point is to show that the comparison functions stay in the same class. The arguments are somewhat involved.

1.1. Notation. We use the following standard notation:

$$B_t=\{y\in R^n:|y|<t\},$$

$$B(x,r)=\{y\in R^n:|x-y|<r\},$$

$$\fint_E f=|E|^{-1}\int_E f,$$

$$(f)_t = \fint_{B_t} f,$$

$$(f)_{x,r} = \fint_{B(x,r)} f,$$

"$f_m \rightharpoonup f$ in L^k" denotes weak convergence,

"$f_m \to f$ in L^k" denotes strong convergence.

2. Preliminary notions.

2.1. Algebraic preliminaries. Suppose that

$$L: \mathbb{R}^n \to \mathbb{R}^N \tag{2.1}$$

is a linear map. The *standard bases* for $\mathbb{R}^n$ and $\mathbb{R}^N$ will be denoted by $e_1, \cdots, e_n$ and $\varepsilon_1 \cdots, \varepsilon_N$, respectively. *Inner products* are denoted by $(\cdot, \cdot)$ and the associated *inner product norm* is denoted by $|\cdot|$. The *coordinates* of L are given by

$$L_{\alpha i} = (Le_i, \varepsilon_\alpha). \tag{2.2}$$

The *inner product norm* of L is given by

$$|L|^2 = \sum_{i,\alpha} L_{\alpha i}^2. \tag{2.3}$$

For each $k > 1$ the space of k-vectors from $\mathbb{R}^n$ is denoted by $\Lambda_k \mathbb{R}^n$. The standard basis elements for $\Lambda_k \mathbb{R}^n$ are denoted by

$$e_\lambda = e_{\lambda_1} \wedge \cdots \wedge e_{\lambda_k}, \qquad 1 \leqq \lambda_1 < \cdots < \lambda_k \leqq n. \tag{2.4}$$

We define $\Lambda_0 \mathbb{R}^n = \mathbb{R}$, $\Lambda_1 \mathbb{R}^n = \mathbb{R}^n$. If $p > n$, then $\Lambda_p \mathbb{R}^n$ is trivial.

The standard inner product on $\mathbb{R}^n$ induces an inner product on $\Lambda_k \mathbb{R}^n$ where the e_λ defined above form an orthonormal basis.

Similarly, the elements

$$\varepsilon_\mu = \varepsilon_{\mu_1} \wedge \cdots \wedge \varepsilon_{\mu_k}, \qquad 1 \leqq \mu_1 < \cdots < \mu_k \leqq N \tag{2.5}$$

form an orthonormal basis for $\Lambda_k \mathbb{R}^N$.

Recall the identity

$$(u_1 \wedge \cdots \wedge u_k, v_1 \wedge \cdots \wedge v_k) = \det [(u_i, v_j)], \tag{2.6}$$

where u_i and v_j are elements of $\mathbb{R}^n$ (or $\mathbb{R}^N$). In particular,

$$|u_1 \wedge \cdots \wedge u_k|^2 = \det [(u_i, u_j)]. \tag{2.7}$$

For each $k > 1$ there is an induced map $\Lambda_k L$ on k-vectors,

$$\Lambda_k L: \Lambda_k \mathbb{R}^n \to \Lambda_k \mathbb{R}^N, \tag{2.8}$$

where

$$\Lambda_k L(v_1 \wedge \cdots \wedge v_k) = Lv_1 \wedge \cdots \wedge Lv_k. \tag{2.9}$$

We often write

$$\langle \Lambda_k L, v_1 \wedge \cdots \wedge v_k \rangle \tag{2.10}$$

for $\Lambda_k L(v_1 \wedge \cdots \wedge v_k)$.

Analogously to (2.2), $\Lambda_k L$ has coordinates

$$\begin{aligned}(\Lambda_k L)_{\mu\lambda} &= (\Lambda_k L(e_\lambda), \varepsilon_\mu) \\ &= (Le_{\lambda_1} \wedge \cdots \wedge Le_{\lambda_k}, \varepsilon_{\mu_1} \wedge \cdots \wedge \varepsilon_{\mu_k}) \\ &= \det [L_{\mu_j \lambda_i}]_{i,j=1}^k,\end{aligned} \tag{2.11}$$

where we have used (2.6) and (2.2).

For example,

$$\begin{aligned}(\Lambda_2 L)_{(\mu_1,\mu_2)(\lambda_1,\lambda_2)} &= \begin{vmatrix} L_{\mu_1\lambda_1} & L_{\mu_1\lambda_2} \\ L_{\mu_2\lambda_1} & L_{\mu_2\lambda_2} \end{vmatrix} \\ &= L_{\mu_1\lambda_1} L_{\mu_2\lambda_2} - L_{\mu_1\lambda_2} L_{\mu_2\lambda_1}.\end{aligned} \tag{2.12}$$

Thus the coordinates of $\Lambda_k L$ are precisely the $k \times k$ minors of the matrix $[L_{\alpha i}]$.

The inner product norm of $\Lambda_k L$ is given by

$$|\Lambda_k L|^2 = \sum_{\lambda,\mu} (\Lambda_k L)_{\mu\lambda}^2. \tag{2.13}$$

If $A = [a_{\alpha i}]$ and $B = [b_{\alpha i}]$ are $k \times k$ matrices, then

$$\det (A+B) = \det A + \sum_{i=1}^{k-1} \sum A^{(k-i)} B^{(i)} + \det B. \tag{2.14}$$

Here $A^{(k-i)}$ ranges over all $(k-i)\times(k-i)$ minors from A, and $B^{(i)}$ is always the complementary $i \times i$ minor from B.

Suppose that both L, $M : \mathbb{R}^n \to \mathbb{R}^N$ are linear maps. Following (2.14) we write

$$\Lambda_k(L+M) = \Lambda_k L + \sum_{i=1}^{k-1} \Lambda_{k-i} L \odot \Lambda_i M + \Lambda_k M. \tag{2.15}$$

In other words, each component of $\Lambda_k(L+M)$ is a sum of terms of which the first is the corresponding component of $\Lambda_k L$ and the last is the corresponding component of $\Lambda_k M$. The expression $\Lambda_{k-i} L \odot \Lambda_i M$ denotes a sum of terms of the form lm, where l and m are coordinates of $\Lambda_{k-i} L$ and $\Lambda_i M$, respectively.

For example, as we see from (2.12)

$$\begin{aligned}(\Lambda_2(L+M))_{(\mu_1,\mu_2)(\lambda_1,\lambda_2)} &= \begin{vmatrix} L_{\mu_1\lambda_1} + M_{\mu_1\lambda_1} & L_{\mu_1\lambda_2} + M_{\mu_1\lambda_2} \\ L_{\mu_2\lambda_1} + M_{\mu_2\lambda_1} & L_{\mu_2\lambda_2} + M_{\mu_2\lambda_2} \end{vmatrix} \\ &= L_{\mu_1\lambda_1} L_{\mu_2\lambda_2} - L_{\mu_1\lambda_2} L_{\mu_2\lambda_1} \\ &\quad + (L_{\mu_1\lambda_1} M_{\mu_2\lambda_2} + L_{\mu_2\lambda_2} M_{\mu_1\lambda_1} - L_{\mu_1\lambda_1} M_{\mu_2\lambda_1} \\ &\quad - L_{\mu_2\lambda_1} M_{\mu_1\lambda_2}) \\ &\quad + (M_{\mu_1\lambda_1} M_{\mu_2\lambda_2} - M_{\mu_1\lambda_2} M_{\mu_2\lambda_1}).\end{aligned} \tag{2.16}$$

We sometimes write

$$\Lambda_k(L+M) = \sum_{i=0}^{k} \Lambda_{k-i} L \odot \Lambda_i M, \tag{2.17}$$

with the understanding that $\Lambda_0 L = 1$, $\Lambda_0 M = 1$.

2.2. Properties of $\Lambda_k Du$. First suppose

$$u : \Omega (\subset \mathbb{R}^k) \to \mathbb{R}^k, \tag{2.18}$$

where Ω is a smooth bounded domain and u is a smooth map. We will denote by

$$[Du / D_i u^\alpha] \tag{2.19}$$

the matrix obtained from the matrix $[Du]$ by deleting the αth row and ith column (i.e., the row and column containing $D_i u^\alpha$).

Then

$$\sum_{i=1}^{k} (-1)^i D_i \det [Du/D_i u^\alpha] = 0, \qquad \alpha = 1, \cdots, k. \tag{2.20}$$

This follows from the rule for differentiating a determinant together with the alternating character of a determinant (see [M, Lemma 4.4.6, p. 122]).

It follows that

$$\det Du = \sum_{i=1}^{k} (-1)^{i+\alpha} D_i (u^\alpha \det [Du/D_i u^\alpha]), \qquad \alpha = 1, \cdots, k \tag{2.21}$$

$$0 = \sum_{i=1}^{k} (-1)^i D_i (u^\alpha \det [Du/D_i u^\beta]), \qquad \alpha \neq \beta. \tag{2.22}$$

We next consider more general u.

First assume that $k = 2$ and $u \in W^{1,2}(\Omega)$. It is trivial that $\det Du \in L^1(\Omega)$. Moreover, (2.20)-(2.22) hold in the distributional sense. This is easily seen by first writing each of (2.20)-(2.22) in weak, i.e., integral form, and noting that the corresponding integrals are continuous in $W^{1,2}$. Since $C^\infty(\Omega)$ is dense in $W^{1,2}(\Omega)$, this establishes the claim.

Now assume $k > 2$ and $u \in W^{1,k-1}(\Omega)$. Then the determinant of any $j \times j$ minor of Du belongs to $L^{(k-1)/j}(\Omega)$ for $1 \leqq j \leqq k-1$, and in particular $\det [Du/D_i u^\alpha] \in L^1(\Omega)$ for each i and α. Moreover, (2.20) holds for such u by a similar approximation argument as before.

Now assume moreover that $\det [Du/D_i u^\alpha] \in L^{(k-1)/(k-2)}(\Omega)$ for each i and α. It follows immediately by the rule for expanding a determinant that $\det Du \in L^1(\Omega)$. We also claim that (2.21) and (2.22) are now true in the distributional sense.

To see this, let $\{u_m\} \subset C^\infty(\Omega)$ and $u_m \to u$ in $W^{1,k-1}(\Omega)$. Also (as in [B, Lemma 6.1]), let $(\det [Du/D_i u^\alpha])_m$ be obtained by mollifying $\det [Du/D_i u^\alpha]$ with ρ_{ε_m}, where $\rho_{\varepsilon_m}(x) = \varepsilon_m^{-k} \rho(\varepsilon_m^{-1} x)$, $\rho \in C_c^\infty(\Omega)$, $\rho \geqq 0$, $\int \rho = 1$, and $\varepsilon_m \to 0$. Then $(\det [Du/D_i u^\alpha])_m \to \det [Du/D_i u^\alpha]$ in $L^{(k-1)/(k-2)}_{\text{loc}}$. Moreover, for $d(x, \partial\Omega) > \varepsilon_m$, it follows from (2.20) that

$$\sum_{i=1}^{k} (-1)^i D_i (\det [Du/D_i u^\alpha])_m(x) = \sum_{i=1}^{k} (-1)^i \int_\Omega D_i \rho_{\varepsilon_m}(x-y) \det [Du/D_i u^\alpha](y)\, dy$$
$$= 0.$$

It follows for any $\varphi \in C_c^\infty(\Omega)$ and all sufficiently large m, that

$$\int \sum_{i=1}^{k} (-1)^{i+\alpha} D_i u_m^\alpha (\det [Du/D_i u^\alpha])_m \varphi = \int \sum_{i=1}^{k} (-1)^{i+\alpha} D_i (u_m^\alpha \det [Du/D_i u^\alpha])_m \varphi$$
$$= -\int \sum_{i=1}^{k} (-1)^{i+\alpha} u_m^\alpha (\det [Du/D_i u^\alpha])_m D_i \varphi.$$

Letting $m \to \infty$ we obtain (2.21).

(We remark that the key point in the preceding argument was to mollify $\det [Du/D_i u^\alpha]$, rather than to consider $\det [Du_m/D_i u_m^\alpha]$.)

Similarly, if $\alpha \neq \beta$,

$$\int \sum_{i=1}^{k} (-1)^i D_i u_m^\alpha (\det [Du/D_i u^\beta])_m \varphi = -\int \sum_{i=1}^{k} (-1)^{i+\alpha} u_m^\alpha (\det [Du/D_i u^\beta])_m D_i \varphi.$$

Letting $m \to \infty$ and using the fact that $0 = \sum_{i=1}^{k} (-1)^i A_i^\alpha \det [A/A_i^\beta]$ for any $k \times k$ determinant $[A]$ and for $\alpha \neq \beta$, we see (2.22) is also true in the distributional sense.

Equation (2.21) is often used to give a definition of $\det Du$ in the weak, or distributional, sense. Thus we have, in particular, established the following lemma.

LEMMA 2.2.1. *Suppose* $u:\Omega(\subset\mathbb{R}^k)\to\mathbb{R}^k$. *Assume that either* $k=2$ *and* $u\in W^{1,2}(\Omega)$ *or that* $k>2$, $u\in W^{1,k-1}(\Omega)$, *and* $\det[Du/D_iu^\alpha]\in L^{(k-1)/(k-2)}(\Omega)$ *for each* i *and* α. *Then the determinant of any* $j\times j$ *minor of* Du *exists both pointwise and in the weak sense, and both notions agree.*

Next assume

$$u:\Omega(\subset\mathbb{R}^n)\to\mathbb{R}^N, \tag{2.23}$$

where Ω is a smooth bounded domain and u is a smooth map. Then it follows from (2.11) and (2.21) that

$$\begin{aligned}(\Lambda_kDu)_{\mu\lambda}&=\det[D_{\lambda_i}u^{\mu_j}]_{i,j=1}^k\\&=\sum_{i=1}^k(-1)^{i+j}D_{\lambda_i}(u^{\mu_j}(\Lambda_{k-1}Du)_{(\mu_1\cdots\hat{\mu}_j\cdots\mu_k)(\lambda_1\cdots\hat{\lambda}_i\cdots\lambda_k)})\end{aligned} \tag{2.24}$$

for each $j=1,\cdots,k$.

In particular,

$$\begin{aligned}(\Lambda_2Du)_{(\mu_1,\mu_2)(\lambda_1,\lambda_2)}&=\begin{vmatrix}D_{\lambda_1}u^{\mu_1}&D_{\lambda_2}u^{\mu_1}\\D_{\lambda_1}u^{\mu_2}&D_{\lambda_2}u^{\mu_2}\end{vmatrix}\\&=D_{\lambda_1}(u^{\mu_1}D_{\lambda_2}u^{\mu_2})-D_{\lambda_2}(u^{\mu_1}D_{\lambda_1}u^{\mu_2}).\end{aligned} \tag{2.25}$$

A similar expression is obtained on expanding across the second row.

Similarly,

$$\begin{aligned}(\Lambda_3Du)_{(\mu_1,\mu_2,\mu_3)(\lambda_1,\lambda_2,\lambda_3)}&=\begin{vmatrix}D_{\lambda_1}u^{\mu_1}&D_{\lambda2}u^{\mu_1}&D_{\lambda_3}u^{\mu_1}\\D_{\lambda_1}u^{\mu_2}&D_{\lambda_2}u^{\mu_2}&D_{\lambda_3}u^{\mu_2}\\D_{\lambda_1}u^{\mu_3}&D_{\lambda_2}u^{\mu_3}&D_{\lambda_3}u^{\mu_3}\end{vmatrix}\\&=D_{\lambda_1}(u^{\mu_1}(\Lambda_2Du)_{(\mu_2,\mu_3)(\lambda_2,\lambda_3)})\\&\quad-D_{\lambda_2}(u^{\mu_1}(\Lambda_2Du)_{(\mu_2,\mu_3)(\lambda_1,\lambda_3)})\\&\quad+D_{\lambda_3}(u^{\mu_1}(\Lambda_2Du)_{(\mu_2,\mu_3)(\lambda_1,\lambda_2)}).\end{aligned} \tag{2.26}$$

A similar expression is again obtained by expanding across the second or third row.

We usually abbreviate (2.24) to

$$\Lambda_kDu=\sum D(u\odot\Lambda_{k-1}Du), \tag{2.27}$$

or even

$$\Lambda_kDu=D(u\Lambda_{k-1}Du). \tag{2.28}$$

The essential point to observe is that each component function of Λ_kDu is a linear combination of partial derivatives of terms of the form fg, where f is a component function of u and g is a component of $\Lambda_{k-1}Du$. The more precise form in (2.24) will not usually be relevant.

2.3. The class $\Lambda_kW^{1,s}(\Omega)$.

DEFINITION 2.3.1. Assume $\Omega(\subset\mathbb{R}^n)$ is a smooth bounded domain and

$$u:\Omega\to\mathbb{R}^N.$$

Assume $2\leqq k\leqq\min\{n,N\}$ and assume

$$s\geqq 2 \quad\text{if } n=2, \qquad s\geqq n-1 \quad\text{if } n>2.$$

Then

$$\Lambda_k W^{1,s}(\Omega)$$

is the class of functions $u \in W^{1,s}(\Omega)$ such that $\Lambda_i Du \in L^s$ for $i = 1, \cdots, k$.

Remarks 2.3.2. (i) It follows from Lemma 2.2.1 that (each component of) $\Lambda_i Du$ exists also in the weak sense and that the weak and pointwise notions agree. In other words, (2.24) is valid in the distributional sense.

(ii) We can define $\Lambda_k W^{1,s}(\Omega)$ in the weak sense for more general s and develop a corresponding theory, but the weak and pointwise notions will not necessarily agree. Although this is an appropriate setting for variational problems, the regularity theory in § 5 only applies if $s \geqq 2$ (for $n = 2$) or $s > n - 1$ (for $n > 2$).

For this reason we will in future maintain the restrictions on s given in the definition above.

(iii) More general classes of functions could be clearly defined by requiring $\Lambda_i Du \in L^{s_i}$ for different s_i. It is also possible to generalise the regularity results of § 5 in an analogous way.

(iv) It is clear that $\Lambda_k W^{1,s}(\Omega)$ is *not* generally a linear space. This is a basic reason for some of the difficulties we encounter in § 5.

Since $\Lambda_k Du$ is in particular a vector-valued function, we take the usual L^s norm

$$\|\Lambda_k Du\|_{L^s} = \left(\int_\Omega \sum_{\lambda,\mu} |(\Lambda_k Du)_{\mu\lambda}|^s \right)^{1/s}. \tag{2.29}$$

DEFINITION 2.3.3. A sequence $\{u_j\} \subset \Lambda_k W^{1,s}(\Omega)$ converges in the weak $\Lambda_k W^{1,s}$ sense to $u \in W^{1,s}(\Omega)$ if

(a) $u_j \rightharpoonup u$ in $W^{1,s}(\Omega)$,

(b) $\|\Lambda_i Du_j\|_{L^s} \leqq M$ for $1 = 2, \cdots, k$ and some $M < \infty$, independent of j.

We write

$$u_j \rightharpoonup u \quad \text{in} \quad \Lambda_k W^{1,s}(\Omega) \text{ (or } \Lambda_k W^{1,s}).$$

PROPOSITION 2.3.4. *If $u_j \rightharpoonup u$ in $\Lambda_k W^{1,s}(\Omega)$, then, moreover,*

(i) $u \in \Lambda_k W^{1,s}(\Omega)$,

(ii) $\Lambda_i Du_j \rightharpoonup \Lambda_i Du$ *in* $L^s(\Omega)$, $i = 2, \cdots, k$,

(iii) $\|\Lambda_i Du\|_{L^s} \leqq \liminf_{j\to\infty} \|\Lambda_i Du\|$, $i = 2, \cdots, k$.

Proof. We show by induction on i that $u \in \Lambda_i W^{1,s}$ for $i = 2, \cdots, k$ and that (ii) and (iii) are true.

In case $i = 2$ we observe that $u_j Du_j \rightharpoonup uDu$ in L^1, since $u_j \to u$ in L^s and $Du_j \rightharpoonup Du$ in L^s (and $s \geqq 2$). To see that (iii) is true we observe from (2.25) that

$$\|\Lambda_2 Du\|_{L^s} = \sup \left\{ \int_\Omega uDuD\varphi : \int_\Omega |\varphi|^{s/s-1} \leqq 1 \right\},$$

where we are using the same abuse of notation as in (2.28). Standard arguments now imply (iii).

It follows that (i) and (ii) are also true, since weak L^s convergence is equivalent to distributional convergence in the presence of uniform L^s bounds.

Similar arguments apply for $i > 2$ after writing $\Lambda_i Du = D(u\Lambda_{i-1} Du)$ in integral form. □

Although $\Lambda_k W^{1,s}(\Omega)$ is not usually a linear space, we do have from (2.15) that if $u \in \Lambda_k W^{1,s}(\Omega)$ and $\chi \in C^\infty(\Omega)$ then $u + \chi \in \Lambda_k W^{1,s}(\Omega)$ and

$$\Lambda_k D(u+\chi) = \Lambda_k Du + \sum_{i=1}^{k-1} \Lambda_{k-i} Du \odot \Lambda_i D\chi + \Lambda_k D\chi. \tag{2.30}$$

In particular,

$$|\Lambda_k D(u+\chi)| \leqq c(\chi)\left(1+\sum_{i=1}^{k}|\Lambda_i Du|\right).$$

Also, if $\psi:\Omega(\subset\mathbb{R}^n)\to\mathbb{R}^n$ is bi-Lipschitz and $u\in\Lambda_k W^{1,s}(\psi(\Omega))$, then $u\circ\psi\in\Lambda_k W^{1,s}(\Omega)$, and we calculate almost everywhere that

$$\begin{aligned}
&(\Lambda_k D(u\circ\psi))_{(\mu_1,\cdots,\mu_k)(\lambda_1,\cdots,\lambda_k)}\\
&\quad=((\Lambda_k D(u\circ\psi))(e_{\lambda_1}\wedge\cdots\wedge e_{\lambda_k}),\ \varepsilon_{\mu_1}\wedge\cdots\wedge\varepsilon_{\mu_k})\\
&\quad=(D(u\circ\psi)(e_{\lambda_1})\wedge\cdots\wedge D(u\circ\psi)(e_{\lambda_k}),\ \varepsilon_{\mu_1}\wedge\cdots\wedge\varepsilon_{\mu_k})\\
&\quad=(Du(D\psi(e_{\lambda_1}))\wedge\cdots\wedge Du(D\psi(e_{\lambda_k})),\ \varepsilon_{\mu_1}\wedge\cdots\wedge\varepsilon_{\mu_k})\\
&\quad=\sum_{\alpha_1,\cdots,\alpha_k=1}^{n}\frac{\partial\psi^{\alpha_1}}{\partial x_{\lambda_1}}\cdot\cdots\cdot\frac{\partial\psi^{\alpha_k}}{\partial x_{\lambda_k}}(Du(e_{\alpha_1})\wedge\cdots\wedge Du(e_{\alpha_k}),\ \varepsilon_{\mu_1}\wedge\cdots\wedge\varepsilon_{\mu_k})\\
&\quad=\sum_{1\leqq\alpha_1<\cdots<\alpha_k\leqq n}\frac{\partial(\psi^{\alpha_1},\cdots,\psi^{\alpha_k})}{\partial(x_{\lambda_1},\cdots,x_{\lambda_k})}(\Lambda_k Du)_{(\mu_1,\cdots,\mu_k)(\alpha_1,\cdots,\alpha_k)}.
\end{aligned}\tag{2.31}$$

2.4. Singular values and polyconvex functions. Recall that the singular values of a linear map $L:\mathbb{R}^n\to\mathbb{R}^N$ are the eigenvalues $\sigma_1,\cdots,\sigma_N$ of the positive semidefinite $N\times N$ symmetric matrix $\sqrt{L\circ L^*}$, where L^* is the transpose map.

Moreover, we have the following lemma.

LEMMA 2.4.1. *If $\sigma_1,\cdots,\sigma_N$ are the singular values of $L:\mathbb{R}^n\to\mathbb{R}^N$, then $\sigma_{i_1}\cdot\cdots\cdot\sigma_{i_k}$ for $1\leqq i_1<\cdots<i_k\leqq N$ are the singular values of $\Lambda_k L$.*

Proof. First note that if $T:\mathbb{R}^N\to\mathbb{R}^p$ is a linear map, then

$$\Lambda_k(T\circ L)=\Lambda_k T\circ\Lambda_k L.\tag{2.32}$$

This follows immediately by applying each side to an arbitrary basis element $e_{i_1}\wedge\cdots\wedge e_{i_k}$ and using (2.9).

Next note that

$$(\Lambda_k L)^*=\Lambda_k L^*,\tag{2.33}$$

where $(\Lambda_k L)^*$ is the transpose map of $\Lambda_k L$. This follows by considering basis elements $e_\lambda=e_{\lambda_1}\wedge\cdots\wedge e_{\lambda_k}$ in $\Lambda_k\mathbb{R}^n$ and $\varepsilon_\mu=\varepsilon_{\mu_1}\wedge\cdots\wedge\varepsilon_{\mu_k}$ in $\Lambda_k\mathbb{R}^N$. We have

$$\begin{aligned}
((\Lambda_k L^*)\varepsilon_\mu,\ e_\lambda)&=(L^*\varepsilon_{\mu_1}\wedge\cdots\wedge L^*\varepsilon_{\mu_k},\ e_{\lambda_1}\wedge\cdots\wedge e_{\lambda_k})\\
&=\det[(L^*\varepsilon_{\mu_i},\ e_{\lambda_j})]_{i,j=1}^{k}\quad\text{(from (2.6))}\\
&=\det[(\varepsilon_{\mu_i},\ Le_{\lambda_j})]_{i,j=1}^{k}\\
&=(\varepsilon_{\mu_1}\wedge\cdots\wedge\varepsilon_{\mu_k},\ Le_{\lambda_1}\wedge\cdots\wedge Le_{\lambda_k})\quad\text{(from (2.6))}\\
&=(\varepsilon_\mu,\ (\Lambda_k L)e_\lambda)\\
&=((\Lambda_k L)^*\varepsilon_\mu,\ e_\lambda).
\end{aligned}$$

This implies (2.33).

Let $v_1,\cdots,v_N$ be eigenvectors of $L\circ L^*$ with eigenvalues $\sigma_1^2,\cdots,\sigma_N^2$, respectively. We may assume the v_i form an orthonormal basis of $\mathbb{R}^N$.

It follows that

$$\begin{aligned}
\langle\Lambda_k L\circ(\Lambda_k L)^*,\ v_{i_1}\wedge\cdots\wedge v_{i_k}\rangle&=\langle\Lambda_k(L\circ L^*),\ v_{i_1}\wedge\cdots\wedge v_{i_k}\rangle\\
&=(L\circ L^*)v_{i_1}\wedge\cdots\wedge(L\circ L^*)v_{i_k}\\
&=\sigma_{i_1}^2\cdot\cdots\cdot\sigma_{i_k}^2 v_{i_1}\wedge\cdots\wedge v_{i_k}.
\end{aligned}$$

Thus the $v_{i_1} \wedge \cdots \wedge v_{i_k}$ form a basis of eigenvectors for $\Lambda_k L \circ (\Lambda_k L)^*$, with eigenvalues $\sigma_{i_1}^2 \cdot \cdots \cdot \sigma_{i_k}^2$. In particular, the lemma follows.

LEMMA 2.4.2. *Let*

$$F_j(\Lambda_j L) = g_j(\sigma_1 \cdot \cdots \cdot \sigma_j, \cdots, \sigma_{i_1} \cdot \cdots \cdot \sigma_{i_j}, \cdots),$$

where g is convex and increasing in each variable. Then F is polyconvex.

Proof. This follows immediately from the case where $j = 1$ and Lemma 2.4.1. The case where $j = 1$ is established in [B, Thm. 5.1(ii)]. □

Remark. The previous lemma gives a large class of polyconvex functionals, as shown in [B, § 5] in the case where $n = N = 2, 3$. Namely, if

$$F(Du, \Lambda_2 Du, \cdots, \Lambda_k Du) = \sum_{j=1}^{k} g_j(\sigma_1 \cdot \cdots \cdot \sigma_j, \cdots, \sigma_{i_1} \cdot \cdots \cdot \sigma_{i_j}, \cdots),$$

where g_j are as in the previous lemma, then F is polyconvex.

In particular,

$$\begin{aligned} |\Lambda_j Du|^2 &= \text{Trace } (\Lambda_j Du \circ (\Lambda_j Du)^*) \\ &= \text{Trace } (\Lambda_j (Du \circ Du^*)) \\ &= \sum_{i_1 < \cdots < i_j} \sigma_{i_1}^2 \cdot \cdots \cdot \sigma_{i_j}^2, \end{aligned}$$

where $\sigma_1, \cdots, \sigma_N$ are the singular values of Du.

3. Existence of minimisers. We prove the following existence theorem due to Ball [B, Thm. 7.7]. In fact, the theorem is true (with essentially the same proof) provided $s \geqq 2n/(n+1)$, but under this more general condition the weak and pointwise notions of $\Lambda_i Du$ will no longer necessarily agree. Existence results under even more general conditions have recently been established in [GMS, § 6.A] using techniques from geometric measure theory (they also consider the problem of invertibility), and by Müller [Mu].

Suppose $n = 2$ and $s \geqq 2$, or $n > 2$ and $s \geqq n - 1$.

Let

$$I[u] = \int_\Omega F(x, u, Du, \Lambda_2 Du, \cdots, \Lambda_k Du), \tag{3.1}$$

where

$$u : \Omega (\subset \mathbb{R}^n) \to \mathbb{R}^N. \tag{3.2}$$

We assume

$$F = F(x, z, a_1, \cdots, a_k) : \Omega \times \mathbb{R}^N \times \mathbb{R}^{q_1} \times \cdots \times \mathbb{R}^{q_k} \to \mathbb{R} \tag{3.3}$$

for the appropriate q_i. We write $a = (a_1, \cdots, a_k)$. The structure conditions we assume on F are as follows:

(3.4)
- (a) $F \geqq 0$,
- (b) F is measurable in x for all (z, a),
- (c) F is continuous in z for all (y, a),
- (d) F is convex in a for all (x, z),
- (e) $F(x, z, a_1, \cdots, a_k) \geqq \gamma \sum_{i=1}^{k} |a_i|^s$ for some $\gamma > 0$.

THEOREM 3.1 (Ball). *Suppose I and F are as in* (3.1) *and* (3.4), *and* $v \in W^{1,s}(\Omega)$. *Let*

$$\mathscr{C} = \{u \in \Lambda_k W^{1,s}(\Omega) : u - v \in W_0^{1,s}(\Omega)\}.$$

Then if $I[u] < \infty$ *for some* $u \in \mathscr{C}$, *there will exist* $u \in \mathscr{C}$ *which minimises* $I[u]$ *amongst all members of* $\mathscr{C}$.

Proof. Let $\{u_j\} \subset \mathscr{C}$ be a minimising sequence.

On passing to a subsequence we can assume by Poincaré's inequality that $u_j \rightharpoonup u$ in $W^{1,s}(\Omega)$ for some u. Moreover, from (3.4e) we have uniform bounds on $\|\Lambda_i Du_j\|_{L^s(\Omega)}$ for $i = 2, \cdots, k$. It follows from Definition 2.3.3 and Proposition 2.3.4 that $u_j \rightharpoonup u$ in $\Lambda_k W^{1,s}$ for some u, and moreover that $u \in \mathscr{C}$.

Since $\Lambda_i Du_j \rightharpoonup \Lambda_i Du$ in $L^s(\Omega)$ for each i, by Proposition 2.3.4 (ii), and $u_j \to u$ in L^s, a standard lower semicontinuity result [G, Thm. 2.3, p. 18] implies $I[u] \leqq \lim_{j\to\infty} I[u_j]$.

It follows that u is the required minimiser. □

4. Partial regularity theorem. Suppose

$$u : \Omega (\subset \mathbb{R}^n) \to \mathbb{R}^N \tag{4.1}$$

and $u \in \Lambda_k W^{1,s}$. Consider functionals of the form

$$I[u] = \int_\Omega F^1(Du) + \sum_{i=2}^{k} F^i(\Lambda_i Du). \tag{4.2}$$

Assume the following hypotheses:

(H1) F^i is C^2 for $i = 1, \cdots, k$;

(H2) $s \geqq 2$ if $n = 2$, $s > n - 1$ if $n > 2$;

(H3) (a) $|D^2 F^1(p)| \leqq c(1 + |p|^{s-2})$,

(b) $D_{p_i^\alpha p_j^\beta} F^1(p) \xi_i^\alpha \xi_j^\beta \geqq \gamma(1 + |p|^{s-2})|\xi|^2$

for all $\xi \in \mathrm{Hom}\,(\mathbb{R}^n; \mathbb{R}^N)$, some $\gamma > 0$; For $i = 2, \cdots, k$,

(H4) (a) $|D^2 F^i(q)| \leqq c|q|^{s-2}$,

(b) $D_{q_\alpha q_\beta} F^i(q) \xi_\alpha \xi_\beta \geqq \gamma |q|^{s-2} |\xi|^2$

for all $\xi \in \mathbb{R}^{qi}$, some $\gamma > 0$.

Model problems included under these hypotheses are

$$I[u] = \int_\Omega |Du|^2 + |Du|^s + \sum_{i=2}^{k} |\Lambda_i Du|^s, \tag{4.3}$$

where $s \geqq 2$ if $n = 2$, and $s > n - 1$ if $n > 2$.

The main theorem established in this paper is the following.

MAIN THEOREM 4.1. *Suppose* $u \in \Lambda_k W^{1,s}(\Omega)$ *is a minimiser of* $I[\cdot]$, *where* I *satisfies the hypotheses* (H1)-(H4). *Then there exists an open set* $\Omega_0 \subset \Omega$ *such that*

$$\mathscr{L}^n(\Omega \sim \Omega_0) = 0, \qquad u \in C^{1,\alpha}(\Omega_0)$$

for all $0 < \alpha < 1$.

Remark. In § 7 we indicate various extensions and generalisations.

5. Proof of decay estimate. In this section we continue to use the notation of § 4. In particular, $u \in \Lambda_k W^{1,s}(\Omega)$ is a minimiser of $I[\cdot]$, where $I[\cdot]$ satisfies the hypotheses (H1)-(H4).

Consider the following quantity:

$$U(x, r) := \fint_{B_r(x)} \Big[|Du - (Du)_{x,r}|^2 + |Du - (Du)_{x,r}|^s + \sum_{i=2}^{k} \big(|\Lambda_i (Du)_{x,r}|^{s-2} |\Lambda_i (Du - (Du)_{x,r})|^2 + |\Lambda_i (Du - (Du)_{x,r})|^s \big) \Big] \tag{5.1}$$

(if $s = 2$ we drop the second and third terms on the right side).

We will prove the following decay estimate.

THEOREM 5.1. *Suppose $M > 0$. Then for each $0 < \tau < \frac{1}{2}$ there exists $\varepsilon_0 = \varepsilon_0(\tau, M)$ such that for every $B_r(x) \subset \Omega$, if*

$$|(Du)_{x,r}| \leqq M,$$

$$U(x, r) \leqq \varepsilon_0,$$

then

$$U(x, \tau r) \leqq c_1 \tau^2 U(x, r)$$

for some $c_1 = c_1(M)$.

Proof. We specify $c_1(M)$ later (see (5.86) and (5.38)). Assume the theorem is not true for some M and τ, which we fix.

In the following we will allow c to be a constant which may depend on M, but *not* on τ or m, and which may change from line to line.

By assumption there exist balls $B(x_m, r_m) \subset \Omega$ such that $|(Du)_{x_m, r_m}| \leqq M$, $U(x_m, r_m) \to 0$ as $m \to \infty$, and

$$U(x_m, \tau r_m) > c_1 \tau^2 U(x_m, r_m). \tag{5.2}$$

We will establish a contradiction to (5.2) for some sufficiently large $c_1 = c_1(M)$. Let

$$\lambda_m^2 = U(x_m, r_m). \tag{5.3}$$

Then $\lambda_m \neq 0$ (as otherwise Du is constant on $B_{r_m}(x_m)$ and so both sides of (5.2) equal zero, contradicting (5.2)). Thus we take

$$0 < \lambda_m \to 0 \quad \text{as } m \to \infty. \tag{5.4}$$

Define

$$a_m = (u)_{x_m, r_m}, \tag{5.5}$$

$$A_m = (Du)_{x_m, r_m}. \tag{5.6}$$

Then by assumption,

$$|A_m| \leqq M. \tag{5.7}$$

Define the normalised function

$$v_m(z) = \frac{u(x_m + r_m z) - a_m - r_m A_m z}{\lambda_m r_m} \tag{5.8}$$

for $z \in B_1$.

Note that $v_m \in \Lambda_k W^{1,s}(B_1)$ as in (2.30). We also have

$$
\begin{aligned}
&D_{v_m}(z) = \lambda_m^{-1}(Du(x_m + r_m z) - A_m),\\
&(Dv_m)_\tau = \lambda_m^{-1}((Du)_{x_m,\tau r_m} - A_m),\\
&Dv_m(z) - (Dv_m)_\tau = \lambda_m^{-1}(Du(x_m + r_m z) - (Du)_{x_m,\tau r_m}),\\
&(v_m)_1 = 0,\\
&(Dv_m)_1 = 0.
\end{aligned} \tag{5.9}
$$

It follows from (5.3) that

$$
1 = \fint_{B_1} |Dv_m|^2 + \lambda_m^{s-2}|Dv_m|^s + \sum_{i=2}^{k} (|\Lambda_i A_m|^{s-2}\lambda_m^{2i-2}|\Lambda_i Dv_m|^2 + \lambda_m^{si-2}|\Lambda_i Dv_m|^s). \tag{5.10}
$$

On passing to a subsequence we claim that for some $A \in \mathrm{Hom}(\mathbb{R}^n; \mathbb{R}^N)$ and some $v \in L^2(B_1)$ that

$$
\begin{aligned}
&\text{(i)} && A_m \to A \quad \text{pointwise},\\
&\text{(ii)} && v_m \rightharpoonup v \quad \text{in } W^{1,2},\\
&\text{(iii)} && \lambda_m^{1-2/s} v_m \rightharpoonup 0 \quad \text{in } W^{1,s} \quad (\text{if } s > 2),\\
&\text{(iv)} && |\Lambda_i A_m|^{(s/2)-1}\lambda_m^{i-1}\Lambda_i Dv_m \rightharpoonup 0 \quad \text{in } L^2 \cdots 2 \leqq i \leqq k,\\
&\text{(v)} && \lambda_m^{i-(2/s)}\Lambda_i Dv_m \rightharpoonup 0 \quad \text{in } L^s \cdots 2 \leqq i \leqq k,\\
&\text{(vi)} && \lambda_m^{\delta} Dv_m \to 0 \quad \text{a.e. for any } \delta > 0.
\end{aligned} \tag{5.11}
$$

The first two claims are standard from the facts $|A_m| \leqq M$ and $f_{B_1}|Dv_m|^2 \leqq 1$.

To see (iii) we note that $v_m \rightharpoonup v$ in the distributional sense, and hence $\lambda_m^{1-2/s} v_m \rightharpoonup 0$ in the distributional sense. Since $f_{B_1} \lambda_m^{s-2}|Dv_m|^s$ is bounded uniformly in m and since $(v_m)_1 = 0$, it follows from Poincaré's inequality that $\lambda_m^{1-2/s} v_m$ is uniformly bounded in the $W^{1,s}(B_1)$ norm, and so (iii) follows.

To see (iv) in the case where $i = 2$ we note that $v_m \to v$ in L^2, $Dv_m \rightharpoonup Dv$ in L^2, and hence $v_m \odot Dv_m \rightharpoonup v \odot Dv$ in the distributional sense. But then $\Lambda_2 Dv_m \rightharpoonup \Lambda_2 Dv$ in the distributional sense, and hence $\lambda_m|\Lambda_2 A_m|^{s/2-1}\Lambda_2 Dv_m \rightharpoonup 0$ in the distributional sense. Since $\lambda_m|\Lambda_2 A_m|^{s/2-1}\Lambda_2 Dv_m$ is uniformly bounded in $L^2(B_1)$ by (5.10), result (iv) follows.

The proof of (v) in the case where $i = 2$ is similar.

Assume (iv) and (v) are true with i replaced by $i-1$ and $3 \leqq i \leqq k$. Since $v_m \to v$ in L^2 and $\lambda_m^{i-1-2/s}\Lambda_{i-1} Dv_m \rightharpoonup 0$ in L^s (by (v), recall $s \geqq 2$), it follows that $\lambda_m^{i-1-2/s} v_m \odot \Lambda_{i-1} Dv_m \rightharpoonup 0$ in the distributional sense. But then $\lambda_m^{i-1-2/s}\Lambda_i Dv_m \rightharpoonup 0$ in the distributional sense, and hence $\lambda_m^{i-2/s}\Lambda_i Dv_m \rightharpoonup 0$ in the distributional sense. Since $\lambda_m^{i-2/s}\Lambda_i Dv_m$ is uniformly bounded in L^s by (5.10), it follows that $\lambda_m^{i-2/s}\Lambda_i Dv_m \rightharpoonup 0$ in L^s. This establishes (v) for $i = 1, \cdots, k$, by induction.

We also have from the fact $\lambda_m^{i-1-2/s}\Lambda_i Dv_m \rightharpoonup 0$ that $|\Lambda_i A_m|^{s/2-1}\lambda_m^{i-1}\Lambda_i Dv_m \rightharpoonup 0$ in the distributional sense (recall that $|\Lambda_i A_m|$ is uniformly bounded from (5.7)). Since $|\Lambda_i A_m|^{s/2-1}\lambda_m^{i-1}\Lambda_i Dv_m$ is uniformly bounded in L^2 from (5.10), it follows that $|\Lambda_i A_m|^{s/2-1}\lambda_m^{i-1}\Lambda_i Dv_m \rightharpoonup 0$ in the L^2 sense. This establishes (iv) for $i = 1, \cdots, k$, by induction.

Finally (vi) follows immediately from the fact Dv_m is uniformly bounded in L^2 and so $\lambda_m^{\delta} Dv_m \to 0$ in L^2.

The proof of the theorem is rather long, and so we break it into a number of parts.

A. v_m minimises a normalised functional. Define the normalised functions

$$F^1_m(P) = \lambda_m^{-2}[F^1(A_m + \lambda_m P) - F^1(A_m) - DF^1(A_m)\lambda_m P], \tag{5.12}$$

$$\begin{aligned} F^i_m(P) = \lambda_m^{-2}[&F^i(\Lambda_i(A_m + \lambda_m P)) - F^i(\Lambda_i A_m) \\ &- DF^i(\Lambda_i A_m)(\Lambda_i(A_m + \lambda_m P) - \Lambda_i A_m)] \end{aligned} \tag{5.13}$$

for $i = 2, \cdots, k$.

The corresponding normalised functional is defined to be

$$I_m[w] = \fint_{B_1} F^1_m(Dw) + \sum_{i=2}^{k} F^i_m(Dw). \tag{5.14}$$

For $0 < t < 1$ we similarly define

$$I^t_m[w] = \fint_{B_t} \sum_{i=1}^{k} F^i_m(Dw). \tag{5.15}$$

Example 5.2. If $n = N = 2$ and $i = 2$, then

$$\begin{aligned} F^2_m(P) = \lambda_m^{-2}[&F^2(\det(A_m + \lambda_m P) - F^2(\det A_m) \\ &- DF^2(\det A_m)(\det(A_m + \lambda_m P) - \det A_m)]. \end{aligned} \tag{5.16}$$

If, moreover, $F^2(\det P) = (\det P)^2$, then

$$\begin{aligned} F^2_m(P) &= \lambda_m^{-2}[(\det(A_m + \lambda_m P))^2 - (\det A_m)^2 \\ &\qquad - 2\det A_m(\det(A_m + \lambda_m P) - \det A_m)] \\ &= \lambda_m^{-2}(\det(A_m + \lambda_m P) - \det A_m)^2 \\ &= (A_m \tilde{\odot} P + \lambda_m \det P)^2, \end{aligned} \tag{5.17}$$

where we use $\tilde{\odot}$ for the particular linear combination of terms above.

Thus

$$F^2_m(Dw) = (A_m \tilde{\odot} Dw + \lambda_m \det Dw)^2. \tag{5.18}$$

Remark 5.3. In [EG] the authors define a normalised functional which in the present framework would be obtained by setting for $i = 1, \cdots, k$,

$$\tilde{F}^i(P) = F^i(\Lambda_i P), \tag{5.19}$$

$$\tilde{F}^i_m(P) = \lambda_m^{-2}[\tilde{F}^i(A_m + \lambda_m P) - \tilde{F}^i(A_m) - D\tilde{F}^i(A_m)\lambda_m P], \tag{5.20}$$

$$I_m[w] = \fint_{B_1} \sum_{i=1}^{k} \tilde{F}^i_m(Dw). \tag{5.21}$$

Thus in [EG] the function $P \mapsto F^i(\Lambda_i P)$ is normalised about $P = A_m$, whereas here we normalise the function $\Lambda_i P \mapsto F^i(\Lambda_i P)$ about $\Lambda_i P = \Lambda_i A_m$.

If $i = 1$, then $\tilde{F}^i_m(P) = F^i_m(P)$. But if $i > 1$ this is, of course, not true.

In Example 5.2, where $F^2(\det P) = (\det P)^2$, we obtain from the rule for differentiating a determinant,

$$\begin{aligned} \tilde{F}^2_m(P) &= \lambda_m^{-2}[(\det(A_m + \lambda_m P))^2 - (\det A_m)^2 - 2\det A_m(A_m \tilde{\odot} \lambda_m P)] \\ &= \lambda_m^{-2}[\det(A_m + \lambda_m P))^2 - (\det A_m)^2 \\ &\qquad - 2\det A_m(\det(A_m + \lambda_m P) - \det A_m) + 2\lambda_m^2 \det A_m \det P] \\ &= F^2_m(\det P) + 2\det A_m \det P. \end{aligned} \tag{5.22}$$

Returning to the general case, it is easy to check, as noted in [EG], that v_m is a minimiser of $\tilde{I}_m$.

We show here that v_m is a minimiser of I_m, amongst functions in $\Lambda_k W^{1,s}(B_1)$ which agree with v_m outside some compact subset B_1. The essential point is that the extra term introduced as in (5.22) gives rise to an integrand which can be written in divergence form.

LEMMA 5.4. *Suppose* $w \in \Lambda_k W^{1,s}(B_1)$ *and* $w = v_m$ *outside some compact* $K \subset B_1$. *Then*

$$I_m[v_m] \leqq I_m[w].$$

Proof. First note that

$$\begin{aligned} \fint_{B_1} F^i(\Lambda_i(A_m + \lambda_m Dw)) &= \fint_{B_1} F^i(\Lambda_i D(A_m z + \lambda_m w)) \\ &= \fint_{B_{(x_m, r_m)}} F^i(\Lambda_i Dw^*), \end{aligned} \tag{5.23}$$

where

$$w^*(x_m + r_m z) = a_m + r_m A_m z + \lambda_m r_m w(z).$$

In particular, from (5.23) and the definition of v_m,

$$\fint_{B_1} F^i(\Lambda_i(A_m + \lambda_m Dv_m)) = \fint_{B_{(x_m, r_m)}} F^i(\Lambda_i Du).$$

Since u is a minimiser of $I[\cdot]$, it follows that

$$\fint_{B_1} \sum_{i=1}^{k} F^i(\Lambda_i(A_m + \lambda_m Dv_m)) \leqq \fint_{B_1} \sum_{i=1}^{k} F^i(\Lambda_i(A_m + \lambda_m Dw)). \tag{5.24}$$

But we also have that

$$\fint_{B_1} \Lambda_i(A_m + \lambda_m Dv_m) = \fint_{B_1} \Lambda_i(A_m + \lambda_m Dw). \tag{5.25}$$

To see this it is sufficient to show, in view of (2.24), that if $X \in L^1(B_1)$ is a vector-valued function, $X = 0$ outside some compact $K \subset B_1$, and $\operatorname{div} X \in L^1$ (where $\operatorname{div} X$ is defined in the distributional sense), that $\int_{B_1} \operatorname{div} X = 0$.

But if X_ε is obtained by mollifying X in the usual way, then for all sufficiently small ε, $\int \operatorname{div} X_\varepsilon = 0$. Since $\operatorname{div} X_\varepsilon = (\operatorname{div} X)_\varepsilon$ converges to $\operatorname{div} X$ in L^1, the result follows. □

B. v satisfies a linear equation. We need to calculate the Euler–Lagrange equation for v_m by calculating

$$\left.\frac{d}{dt}\right|_{t=0} \Lambda_i(A_m + \lambda_m(Dv_m + tD\varphi)).$$

Let us write

$$\Lambda_i(A+B) = \Lambda_i A + \Lambda_{i-1} A \tilde{\odot} B + \sum_{j=2}^{i-1} \Lambda_{i-j} A \odot \Lambda_j B + \Lambda_i B, \tag{5.26}$$

so that

$$\Lambda_{i-1} A \tilde{\odot} B$$

will always denote this *particular* linear combinations of terms $a_\lambda b_\lambda$, where a_λ and b_λ are components of $\Lambda_{i-1}A$ and B, respectively.

Then for $\varphi \in C_c^\infty(B_1)$,

$$\frac{d}{dt}\bigg|_{t=0} \Lambda_i(A_m + \lambda_m(Dv_m + tD\varphi)) = \Lambda_{i-1}(A_m + \lambda_m Dv_m) \tilde{\odot} \lambda_m D\varphi. \tag{5.27}$$

It follows from (5.12), (5.13), and Lemma 5.4 that v_m satisfies the Euler-Lagrange equation given by

$$\begin{aligned}
0 &= \lambda_m^{-1} \fint_{B_1} [DF^1(A_m + \lambda_m Dv_m) - DF^1(A_m)]\, D\varphi \\
&\quad + \lambda_m^{-1} \fint_{B_1} \sum_{i=2}^{k} [DF^i(\Lambda_i(A_m + \lambda_m Dv_m)) - DF^i(\Lambda_i A_m)] \\
&\qquad \cdot [\Lambda_{i-1}(A_m + \lambda_m Dv_m) \tilde{\odot} D\varphi] \\
&= \fint_{B_1} \left[\int_0^1 D^2F^1(A_m + t\lambda_m Dv_m)\, dt \right] Dv_m\, D\varphi \\
&\quad + \lambda_m^{-1} \fint_{B_1} \sum_{i=2}^{k} \left[\int_0^1 D^2F^i(\Lambda_i A_m + t(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m))\, dt \right] \\
&\qquad \cdot [\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m][\Lambda_{i-1}(A_m + \lambda_m Dv_m) \tilde{\odot} D\varphi] \\
&= \fint_{B_1} \left[\int_0^1 D^2F^1(A_m + t\lambda_m Dv_m)\, dt \right] Dv_m\, D\varphi \\
&\quad + \fint_{B_1} \sum_{i=2}^{k} \left[\int_0^1 D^2F^i(\Lambda_i A_m + t(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m))\, dt \right] \\
&\qquad \cdot \left[\Lambda_{i-1} A_m \tilde{\odot} Dv_m + \sum_{j=2}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv_m \right] \\
&\qquad \cdot \left[\Lambda_{i-1} A_m \tilde{\odot} D\varphi + \sum_{j=1}^{i-1} \lambda_m^j \Lambda_{i-1-j} A_m \odot \Lambda_j Dv_m \tilde{\odot} D\varphi \right].
\end{aligned} \tag{5.28}$$

We aim to show that v satisfies the linear equation

$$0 = \fint_{B_1} \left[D^2F^1(A)\, Dv\, D\varphi + \sum_{i=2}^{k} D^2F^i(\Lambda_i A)(\Lambda_{i-1} A \tilde{\odot} Dv)(\Lambda_{i-1} A \tilde{\odot} D\varphi) \right] \tag{5.29}$$

for all $\varphi \in C_c^\infty(B_1)$.

This is an elliptic system with ellipticity bounds given by

$$\gamma|\xi|^2 \leqq D^2F^1(A)\xi\xi + \sum_{i=2}^{k} D^2F^i(\Lambda_i A)(\Lambda_{i-1} A \tilde{\odot} \xi)(\Lambda_{i-1} A \tilde{\odot} \xi) \leqq c|\xi|^2 \tag{5.30}$$

for all $\xi \in \mathrm{Hom}\,(\mathbb{R}^n; \mathbb{R}^N)$.

The necessary calculations are simplified by the following lemmas.

LEMMA 5.5. *Suppose* $\{f_m\} \subset L^1(B_1)$, $\{g_m\} \subset L^{1+\eta}(B_1)$ (*some* $\eta > 0$), *and* α *is constant. Suppose*

$$f_m \to \alpha \quad \text{a.e.},$$

$$g_m \rightharpoonup g \quad \text{in } L^{1+\eta},$$

$$\int_\Omega |f_m g_m|^{1+\delta} \leqq A \text{ for some } \delta > 0, \quad \text{some } A < \infty, \quad \text{and all } m.$$

Then

$$f_m g_m \rightharpoonup \alpha g \quad \text{in } L^{1+\delta}.$$

Proof. For any $\varphi \in C^\infty(B_1)$ we have

$$\int_{B_1} (f_m g_m - \alpha g)\varphi = \int_{B_1} (f_m - \alpha) g_m \varphi + \alpha \int_{B_1} (g_m - g)\varphi. \tag{5.31}$$

The second integral on the right side converges to zero as $m \to \infty$.

To see that the same is true of the first integral we note that for each $\varepsilon > 0$ there exists $E \subset B_1$ with $|E| < \varepsilon$ such that $f_m \to \alpha$ uniformly on $B_1 \sim E$. Then

$$\int_{B_1 \sim E} |f_m - \alpha||g_m||\varphi| \leqq c \sup_{B_1 \sim E} |f_m - \alpha| \left(\int_{B_1} |g_m|^{1+\eta} \right)^{1/(1+\eta)}$$

$$\to 0 \quad \text{as } m \to \infty.$$

Moreover,

$$\int_E |f_m - \alpha||g_m||\varphi| \leqq c \left(\int_E |f_m g_m|^{1+\delta} \right)^{1/(1+\delta)} |E|^{\delta/(1+\delta)}$$

$$+ c \left(\int_E |g_m|^{1+\eta} \right)^{1/(1+\eta)} |E|^{\eta/(1+\eta)}$$

$$\leqq c(\varepsilon),$$

where $c(\varepsilon) \to 0$ as $\varepsilon \to 0$. Since $\varepsilon > 0$ is arbitrary, it follows that the first integral on the right side of (5.31) also converges to zero as $m \to \infty$.

This establishes the lemma. □

Remark. The case $g_m \equiv 1$ is also of interest.

LEMMA 5.6. *Suppose* $p, q \geqq 0$ *and all* $a_i \geqq 0$. *Then*

$$(\Sigma a_i^p)(\Sigma a_i^q) \leqq c \Sigma a_i^{p+q}.$$

Proof. Multiply out the left side and use Young's inequality. □

LEMMA 5.7. $|\Lambda_i A_m| \leqq c|\Lambda_j A_m|$ *if* $1 \leqq j \leqq i$, *for some* c *independent of* m.

Proof. If $j = i - 1$ this follows by expanding the determinant corresponding to each component of $\Lambda_i A_m$ along an arbitrary row (or column) and recalling $|A_m|$ is uniformly bounded in m. The result follows for general j by repeating the argument. □

LEMMA 5.8.

$$\Lambda_i(A_m + \lambda_m D v_m) \to \Lambda_i A \quad \text{a.e. as } m \to \infty.$$

Proof. This follows from (5.11)(vi) and the fact $A \mapsto \Lambda_i A$ is clearly a continuous function. □

We now proceed with the proof of (5.29).

Apply Lemma 5.5 to the first integral on the right side of (5.28), with

$$f_m = \int_0^1 D^2 F^1(A_m + t\lambda_m D v_m)\, dt,$$

$$g_m = D v_m, \qquad g = Dv.$$

Then $f_m \to D^2F^1(A)$ almost everywhere from (5.11)(vi), and $g_m \rightharpoonup g$ in L^2 from (5.11)(ii). Also,

$$\begin{aligned}|f_m g_m| &\leqq c(1+\lambda_m^{s-2}|Dv_m|^{s-2})|Dv_m| \\ &\leqq c(|Dv_m|+\lambda_m^{1-2/s}(\lambda_m^{s-2}|Dv_m|^s)^{(s-1)/s}),\end{aligned}$$

and so $f_m g_m \in L^{s/(s-1)}$ uniformly in m. Lemma 5.5 now implies

$$\fint_{B_1}\left[\int_0^1 D^2F^1(A_m+t\lambda_m Dv_m)\,dt\right]Dv_m\,D\varphi \to \fint_{B_1} D^2F^1(A)\,Dv\,D\varphi. \tag{5.32}$$

We next consider the second integral on the right side of (5.28). It is convenient to write this integral in the following form:

$$\begin{aligned}&\sum_{i=2}^{k}\left\{\fint_{B_1}\left[\int_0^1 D^2F^i(\Lambda_i A_m + t(\Lambda_i(A_m+\lambda_m Dv_m)-\Lambda_i A_m))\,dt\right]\right. \\ &\qquad\cdot\Bigg[(\Lambda_{i-1}A_m \tilde{\odot} Dv_m)(\Lambda_{i-1}A_m \tilde{\odot} D\varphi) \\ &\qquad\quad+\left(\sum_{j=2}^{i}\lambda_m^{j-1}\Lambda_{i-j}A_m \odot \Lambda_j Dv_m\right)(\Lambda_{i-1}A_m \tilde{\odot} D\varphi) \\ &\qquad\quad+\left(\sum_{j=1}^{i}\lambda_m^{j-1}\Lambda_{i-j}A_m \odot \Lambda_j Dv_m\right) \\ &\qquad\quad\left.\cdot\left(\sum_{j=1}^{i-1}\lambda_m^{j}\Lambda_{i-1-j}A_m \odot \Lambda_j Dv_m \tilde{\odot} D\varphi\right)\Bigg]\right\} \\ &\quad=\sum_{i=2}^{k}(A_m^i+B_m^i+C_m^i).\end{aligned} \tag{5.33}$$

We first claim

$$A_m^i \to \fint_{B_1} D^2F^i(\Lambda_i A)(\Lambda_{i-1}A \tilde{\odot} Dv)(\Lambda_{i-1}A \tilde{\odot} D\varphi). \tag{5.34}$$

To see this, apply Lemma 5.5 with

$$\begin{aligned}f_m &= \int_0^1 D^2F^i(\Lambda_i A_m + t(\Lambda_i(A_m+\lambda_m Dv_m)-\Lambda_i A_m))\,dt, \\ g_m &= (\Lambda_{i-1}A_m \tilde{\odot} Dv_m)\Lambda_{i-1}A_m.\end{aligned}$$

Then

$$\begin{aligned}&f_m \to D^2F^i(\Lambda_i A) \quad \text{a.e.}, \\ &g_m \rightharpoonup (\Lambda_{i-1}A \tilde{\odot} Dv)\Lambda_{i-1}A \quad \text{in } L^2.\end{aligned}$$

Moreover,

$$\begin{aligned}|f_m g_m| &\leqq c(1+|\Lambda_i(A_m+\lambda_m Dv_m)|^{s-2})|Dv_m| \\ &\leqq c\left(1+\sum_{j=1}^{i}\lambda_m^{j(s-2)}|\Lambda_j Dv_m|^{s-2}\right)|Dv_m| \\ &\leqq c\left(|Dv_m|+\lambda_m^{1-2/s}\sum_{j=1}^{i}(\lambda_m^{j-2/s}|\Lambda_j Dv_m|)^{s-2}(\lambda_m^{1-2/s}|Dv_m|)\right) \\ &\leqq c\Bigg(|Dv_m|+\lambda_m^{1-2/s}\sum_{j=1}^{i}(\lambda_m^{j-2/s}|\Lambda_j Dv_m|)^{s-1} \\ &\qquad\qquad+\lambda_m^{1-2/s}\sum_{j=1}^{i}(\lambda_m^{1-2/s}|Dv_m|)^{s-1}\Bigg).\end{aligned}$$

It follows from (5.11) that $f_m g_m \in L^{s/(s-1)}$, uniformly in m. Hence (5.34) is valid by Lemma 5.5.

We next claim

$$B^i_m \to 0 \quad \text{as } m \to \infty. \tag{5.35}$$

To establish this, apply Lemma 5.5 with

$$f_m = \left[\int_0^1 D^2 F^i(\Lambda_i A_m + t(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m))\, dt \right] \cdot \left(\sum_{j=2}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv_m \right) \Lambda_{i-1} A_m,$$

$$g_m = 1.$$

Then $f_m \to 0$ almost everywhere (using (5.11)(vi)). Also,

$$|f_m| \leqq c \left(|\Lambda_i A_m| + \sum_{j=1}^{i} \lambda_m^j |\Lambda_j Dv_m| \right)^{s-2} \left(\sum_{j=2}^{i} \lambda_m^{j-1} |\Lambda_j Dv_m| \right)$$

(using (H4) and (5.7))

$$\leqq c \sum_{j=2}^{i} |\Lambda_i A_m|^{s-2} \lambda_m^{j-1} |\Lambda_j Dv_m| + c\lambda_m^{1-2/s} \left(\sum_{j=1}^{i} \lambda_m^{j-2/s} |\Lambda_j Dv_m| \right)^{s-2} \left(\sum_{j=2}^{i} \lambda_m^{j-2/s} |\Lambda_j Dv_m| \right)$$

$$\leqq c \sum_{j=2}^{i} |\Lambda_j A_m|^{s-2} \lambda_m^{j-1} |\Lambda_j Dv_m| + c\lambda_m^{1-2/s} \left(\sum_{j=1}^{i} \lambda_m^{j-2/s} |\Lambda_j Dv_m| \right)^{s-1},$$

using Lemma 5.7.

It follows from (5.11) that $f_m \in L^{s/(s-1)}$ uniformly in m. Thus (5.35) is established from Lemma 5.5.

Finally, we claim that

$$C^i_m \to 0 \quad \text{as } m \to \infty. \tag{5.36}$$

To see this we calculate

$$|C^i_m| \leqq c \fint_{B_1} \left(|\Lambda_i A_m| + \sum_{j=1}^{i} \lambda_m^j |\Lambda_j Dv_m| \right)^{s-2} \cdot \left(\sum_{j=1}^{i} \lambda_m^{j-1} |\Lambda_j Dv_m| \right) \left(\sum_{j=1}^{i-1} \lambda_m^j |\Lambda_j Dv_m| \right)$$

$$\leqq c \fint_{B_1} |\Lambda_i A_m|^{s-2} \sum_{j=1}^{i} \lambda_m^{2j-1} |\Lambda_j Dv_m|^2 + c \sum_{j=1}^{i} \lambda_m^{js-1} |\Lambda_j Dv_m|^s$$

$$\to 0 \quad \text{as } m \to \infty$$

from (5.11). This proves (5.36).

Thus (5.29) is established, using (5.28), and (5.32)-(5.36).

From standard regularity theory it now follows that

$$v \in C^\infty(B_1), \tag{5.37}$$

and, moreover, that for any $K \subset\subset B_1$ and any $i \geqq 1$, we have from (5.29), (5.30), and (5.10) that

$$\sup_K |D^i v| \leqq c \fint_{B_1} |Dv|^2 \tag{5.38}$$

where $c = c(K, M, i)$.

C. An estimate of $I^t_m(v_m) - I^t_m(v)$ from above. We will show that for almost everywhere $t \in (\frac{1}{2}, 1)$ there exists a subsequence (depending on t) for which

$$\lim_{m\to\infty} (I^t_m(v_m) - I^t_m(v)) \leqq 0. \tag{5.39}$$

This, together with the estimate from below in Part D, are the main estimates.

In order to establish (5.39) we need to define comparison functions w_m that "connect" v to v_m.

More precisely, for each $t \in (\frac{1}{2}, 1)$ and each $\delta \in (0, \frac{1}{4})$ we define

$$w_m = w^{t,\delta}_m \in \Lambda_k W^{1,s}(B_1) \tag{5.40}$$

as follows.

For $x \in B_1$ write

$$x = r\omega, \quad \text{where } r = |x| \quad \text{and} \quad \omega = x/|x|. \tag{5.41}$$

Let

$$w_m(r\omega) = \begin{cases} v(r\omega) & \cdots\ r \leqq t-\delta, \\ v([t-\delta+2(r-(t-\delta))]\omega] & \cdots\ t-\delta \leqq r \leqq t-\delta/2, \\ \dfrac{t-r}{\delta/2}\, v(t\omega) + \dfrac{r-(t-\delta/2)}{\delta/2}\, v_m(t\omega) & \cdots\ t-\dfrac{\delta}{2} \leqq r \leqq t, \\ v_m(r\omega) & \cdots\ t \leqq r \leqq 1. \end{cases} \tag{5.42}$$

Thus w_m equals v on $B_{t-\delta}$. On the annulus $B_{t-\delta/2} \sim B_{t-\delta}$ we obtain w_m by taking v restricted to $B_t \sim B_{t-\delta}$ and then composing with a diffeomorphism from $B_{t-\delta/2} \sim B_{t-\delta}$ to $B_t \sim B_{t-\delta}$. On $B_t \sim B_{t-\delta/2}$ we obtain $w_m(r\omega)$ for $t - \delta/2 \leqq r \leqq t$ by linearly interpolating between $v(t\omega)$ and $v_m(t\omega)$. Finally, w_m agrees with v_m on $B_1 \sim B_t$.

We claim

$$w_m \in \Lambda_k W^{1,s}(B_1) \quad \text{for a.e. } t \in (\tfrac{1}{2}, 1), \tag{5.43}$$

and establish some estimates.

First note that

$$\sup_{B_{t-\delta/2}} |Dw_m| \leqq c \sup_{B_t} |Dv| \leqq c \tag{5.44}$$

from (5.38). It follows that

$$\sup_{B_{t-\delta/2}} |\Lambda_i Dw_m| \leqq c \tag{5.45}$$

for $i = 1, \cdots, k$.

Next consider the annulus $B_t \sim B_{t-\delta/2}$. Let $\tau_1, \cdots, \tau_{n-1}, \nu$ be an orthonormal basis for $\mathbb{R}^n$, where ν is a radial vector (and so $\tau_1, \cdots, \tau_{n-1}$ are orthogonal to ν). Then on $B_t \sim B_{t-\delta/2}$,

$$D_{\tau_i} w_m(r\omega) = \frac{t-r}{\delta/2} \cdot \frac{t}{r} D_{\tau_i} v(t\omega) + \frac{r-(t-\delta/2)}{\delta/2} \cdot \frac{t}{r} D_{\tau_i} v_m(t\omega), \tag{5.46}$$

and so

$$|D_{\tau_i} w_m(r\omega)| \leqq c(1+|Dv_m(t\omega)|), \tag{5.47}$$

since $t/r \leqq 2$, $(t-r)/(\delta/2) \leqq 1$, $(r-(t-\delta/2)/(\delta/2) \leqq 1$, $|Dv| \leqq c$.

Similarly,

$$D_\nu w_m(r\omega) = \frac{2}{\delta}(v_m(t\omega) - v(t\omega)), \tag{5.48}$$

and so

$$|D_\nu w_m(r\omega)| \leqq c\delta^{-1}|v_m(t\omega) - v(t\omega)|. \tag{5.49}$$

It follows that

$$\begin{aligned} &\langle \Lambda_i D w_m(r\omega), \tau_1 \wedge \cdots \wedge \tau_i \rangle \\ &\quad = D_{\tau_1} w_m(r\omega) \wedge \cdots \wedge D_{\tau_i} w_m(r\omega) \\ &\quad = \bigwedge_{j=1}^{i} \left(\frac{t-r}{\delta/2} \cdot \frac{t}{r} D_{\tau_j} v(t\omega) + \frac{r-(t-\delta/2)}{\delta/2} \cdot \frac{t}{r} \cdot D_{\tau_j} v_m(t\omega) \right). \end{aligned}$$

On expanding we obtain terms of the form

$$f(D_{\tau_1} v(t\omega) \text{ or } D_{\tau_1} v_m(t\omega)) \wedge \cdots \wedge (D_{\tau_i} v(t\omega) \text{ or } D_{\tau_i} v_m(t\omega)),$$

where $|f| \leqq c$. Such terms are norm bounded by

$$c\left(1 + \sum_{j=1}^{i} |\Lambda_j D v_m(t\omega)|\right),$$

since $|D_{\tau_j} v(t\omega)| \leqq c$.

Similarly,

$$\begin{aligned} &\langle \Lambda_i D w_m(r\omega), \nu \wedge \tau_1 \wedge \cdots \wedge \tau_{i-1} \rangle \\ &\quad = D_\nu w_m(r\omega) \wedge D_{\tau_1} w_m(r\omega) \wedge \cdots \wedge D_{\tau_{i-1}} w_m(rw) \\ &\quad = \frac{2}{\delta}(v_m(t\omega) - v(t\omega)) \wedge \\ &\qquad \cdot \wedge \bigwedge_{j=1}^{i-1} \left(\frac{t-r}{\delta/2} \cdot \frac{t}{r} D_{\tau_j} v(t\omega) + \frac{r-(t-\delta/2)}{\delta/2} \cdot \frac{t}{r} D_{\tau_j} v_m(t\omega) \right). \end{aligned}$$

As before, such terms are norm bounded by

$$c\delta^{-1}|v_m(t\omega) - v(t\omega)|\left(1 + \sum_{j=1}^{i-1} |\Lambda_j D v_m(t\omega)|\right).$$

It follows from the above that on $B_t \sim B_{t-\delta/2}$

$$|\Lambda_i D w_m| \leqq c(1+\delta^{-1}|v-v_m|)\left(1 + \sum_{j=1}^{i-1} |\Lambda_j D v_m|\right) + c|\Lambda_i D v_m|, \tag{5.50}$$

where the left side is evaluated at $r\omega$ for $t-\delta/2 \leqq r \leqq t$, the right side is evaluated at $t\omega$, and $1 \leqq i \leqq k$.

Since $v_m = w_m$ on $B_1 \sim B_t$,

$$I^t_m(v_m) - I^t_m(w_m) \leqq 0.$$

Hence

$$\begin{aligned} I^t_m(v_m) - I^t_m(v) &= (I^t_m(v_m) - I^t_m(w_m)) + (I^t_m(w_m) - I^t_m(v)) \\ &\leqq I^t_m(w_m) - I^t_m(v) \\ &= \int_{B_t \sim B_{t-\delta}} \sum_{i=1}^{k} (F^i_m(Dw_m) - F^i_m(Dv)). \end{aligned} \tag{5.51}$$

From (5.12) we estimate for any $w \in \Lambda_k W^{1,s}(B_1)$,

$$\begin{aligned} |F^1_m(Dw)| &= \left| \left(\int_0^1 (1-\tau) D^2 F^1(A_m + \tau \lambda_m Dw)\, d\tau \right) Dw\, Dw \right| \\ &\leqq c(1 + \lambda_m^{s-2} |Dw|^{s-2}) |Dw|^2. \end{aligned} \tag{5.52}$$

For $2 \leqq i \leqq k$ we similarly have from (5.13) that

$$\begin{aligned} |F^i_m(Dw)| &= \lambda_m^{-2} \left| \left(\int_0^1 (1-\tau) D^2 F^i [\Lambda_i A_m + \tau(\Lambda_i(A_m + \lambda_m Dw) - \Lambda_i A_m)]\, d\tau \right) \right. \\ &\qquad \left. \cdot (\Lambda_i(A_m + \lambda_m Dw) - \Lambda_i A_m)(\Lambda_i(A_m + \lambda_m Dw) - \Lambda_i A_m) \right| \\ &\leqq c\lambda_m^{-2} (|\Lambda_i A_m|^{s-2} + |\Lambda_i(A_m + \lambda_m Dw) - \Lambda_i A_m|^{s-2}) \\ &\qquad \cdot |\Lambda_i(A_m + \lambda_m Dw) - \Lambda_i A_m|^2 \\ &\leqq c\lambda_m^{-2} \left(|\Lambda_i A_m|^{s-2} + \sum_{j=1}^{i} \lambda_m^{j(s-2)} |\Lambda_j Dw|^{s-2} \right) \left(\sum_{j=1}^{i} \lambda_m^{2j} |\Lambda_j Dw|^2 \right) \\ &\leqq c \sum_{j=1}^{i} |\Lambda_j A_m|^{s-2} \lambda_m^{2j-2} |\Lambda_j Dw|^2 + c \sum_{j=1}^{i} \lambda_m^{js-2} |\Lambda_j Dw|^s. \end{aligned} \tag{5.53}$$

From (5.47), (5.48), and (5.52), we calculate (with the left side evaluated at $r\omega$ and the right side at $t\omega$, where $t - \delta/2 \leqq r \leqq t$)

$$\begin{aligned} |F^1_m(Dw_m)| &\leqq c(1 + \delta^{-1}|v - v_m| + |Dv_m|)^2 \\ &\qquad + c\lambda_m^{s-2}(1 + \delta^{-1}|v - v_m| + |Dv_m|)^s \\ &\leqq c(1 + \delta^{-2}|v - v_m|^2 + |Dv_m|^2) \\ &\qquad + c\lambda_m^{s-2}(\delta^{-s}|v - v_m|^s + |Dv_m|^s). \end{aligned} \tag{5.54}$$

Similarly, for $2 \leqq i \leqq k$, we have from (5.50) and (5.53), with the same evaluation convention, that

$$\begin{aligned} |F^i_m(Dw_m)| &\leqq c \sum_{j=1}^{i} |\Lambda_j A_m|^{s-2} \lambda_m^{2j-2} \left[(1 + \delta^{-2}|v - v_m|^2) \right. \\ &\qquad \left. \cdot \left(1 + \sum_{l=0}^{j-1} |\Lambda_l Dv_m|^2 \right) + |\Lambda_j Dv_m|^2 \right] \\ &\quad + c \sum_{j=1}^{i} \lambda_m^{js-2} \left[(1 + \delta^{-s}|v - v_m|^s) \right. \\ &\qquad \left. \cdot \left(1 + \sum_{l=0}^{j-1} |\Lambda_l Dv_m|^s \right) + |\Lambda_j Dv_m|^s \right]. \end{aligned} \tag{5.55}$$

The first term on the right side is bounded by

$$c\sum_{j=1}^{i}|\Lambda_j A_m|^{s-2}\lambda_m^{2j-2}(1+\delta^{-2}|v-v_m|^2)$$

$$+c\sum_{l=0}^{i-1}|\Lambda_l A_m|^{s-2}\lambda_m^{2l}(1+\delta^{-2}|v-v_m|^2)|\Lambda_l Dv_m|^2$$

$$+c\sum_{j=1}^{i}|\Lambda_j A_m|^{s-2}\lambda_m^{2j-2}|\Lambda_j Dv_m|^2$$

(where we have used $|\Lambda_j A_m|\leqq c|\Lambda_l A_m|$ and $\lambda_m^{2j-2}\leqq c\lambda_m^{2l}$ for $1\leqq l\leqq j-1$)

$$\leqq c(1+\delta^{-2}|v-v_m|^2)$$

$$+c\lambda_m^2\delta^{-2}|v-v_m|^2\sum_{j=1}^{i-1}|\Lambda_j A_m|^{s-2}\lambda_m^{2j-2}|\Lambda_j Dv_m|^2$$

$$+c\sum_{j=1}^{i}|\Lambda_j A_m|^{s-2}\lambda_m^{2j-2}|\Lambda_j Dv_m|^2.$$

Similarly, the second term on the right side of (5.55) is bounded by

$$c\sum_{j=1}^{i}\lambda_m^{js-2}(1+\delta^{-5}|v-v_m|^s)+c\sum_{l=0}^{i-1}\lambda_m^{ls-2+s}(1+\delta^{-5}|v-v_m|^s)|\Lambda_l Dv_m|^s$$

$$+c\sum_{j=1}^{i}\lambda_m^{js-2}|\Lambda_j Dv_m|^s$$

$$\leqq c\lambda_m^{s-2}(1+\delta^{-s}|v-v_m|^s)+c\lambda_m^s\delta^{-s}|v-v_m|^s\sum_{j=1}^{i-1}\lambda_m^{js-2}|\Lambda_j Dv_m|^s$$

$$+c\sum_{j=1}^{i}\lambda_m^{js-2}|\Lambda_j Dv_m|^s.$$

It follows that for $2\leqq i\leqq k$

$$|F_m^i(Dw_m)|\leqq c\left(1+\delta^{-2}|v-v_m|^2+\sum_{j=1}^{i}|\Lambda_j A_m|^{s-2}\lambda_m^{2j-2}|\Lambda_j Dv_m|^2\right)$$

$$+c\lambda_m^2\delta^{-2}|v-v_m|^2\sum_{j=1}^{i-1}|\Lambda_j A_m|^{s-2}\lambda_m^{2j-2}|\Lambda_j Dv_m|^2$$

$$+c\left(\lambda_m^{s-2}\delta^{-s}|v-v_m|^s+\sum_{j=1}^{i}\lambda_m^{js-2}|\Lambda_j Dv_m|^s\right)$$

$$+c\lambda_m^s\delta^{-s}|v-v_m|^s\sum_{j=1}^{i-1}\lambda_m^{js-2}|\Lambda_j Dv_m|^s. \tag{5.56}$$

Remember that the left side is evaluated at $r\omega$ and the right side at $t\omega$, where $t-\delta/2\leqq r\leqq t$.

Applying Fubini's theorem to (5.54) and (5.56) it follows that

$$\begin{aligned}
&\int_{B_t\sim B_{t-\delta/2}} \sum_{i=1}^{k} F_m^i(Dw_m) \\
&\qquad \leqq c\delta \int_{\partial B_t} 1+\delta^{-2}|v-v_m|^2+|Dv_m|^2+\sum_{j=2}^{k}|\Lambda_j A_m|^{s-2}\lambda_m^{2j-2}|\Lambda_j Dv_m|^2 \\
&\qquad +c\delta^{-1}\lambda_m^2 \sup_{\partial B_t}|v-v_m|^2 \int_{\partial B_t}\sum_{j=1}^{k-1}|\Lambda_j A_m|^{s-2}\lambda_m^{2j-2}|\Lambda_j Dv_m|^2 \qquad (5.57)\\
&\qquad +c\delta \int_{\partial B_t}\lambda_m^{s-2}\delta^{-s}|v-v_m|^s+\lambda_m^{s-2}|Dv_m|^s+\sum_{j=2}^{k}\lambda_m^{js-2}|\Lambda_j Dv_m|^s \\
&\qquad +c\delta^{1-s}\lambda_m^s \sup_{\partial B_t}|v-v_m|^s \int_{\partial B_t}\sum_{j=1}^{k-1}\lambda_m^{js-2}|\Lambda_j Dv_m|^s.
\end{aligned}$$

It follows from (5.11) and Lemma 5.10 that for almost everywhere $t\in(\frac{1}{2},1)$ there exists $M_t<\infty$ and a subsequence depending on t for which

$$\begin{aligned}
&\int_{\partial B_t}|Dv_m|^2+\lambda_m^{s-2}|Dv_m|^s+\sum_{j=2}^{k}\lambda_m^{2j-2}|\Lambda_j A_m|^{s-2}|\Lambda_j Dv_m|^2 \\
&\qquad +\sum_{j=2}^{k}\lambda_m^{js-2}|\Lambda_j Dv_m|^s \leqq M_t. \qquad (5.58)
\end{aligned}$$

We may also assume from (5.11) that

$$\int_{\partial B_t}|v-v_m|^2+\lambda_m^{s-2}|v-v_m|^s\to 0. \qquad (5.59)$$

Finally, notice that $\lambda_m^{1-2/s}(v_m-v)\rightharpoonup 0$ in $W^{1,s}(B_1)$ (from (5.11)(ii), (iii)). It follows from (5.58), by using a countable dense set of test functions, that $\lambda_m^{1-2/s}(v_m-v)\rightharpoonup 0$ in $W^{1,s}(\partial B_t)$ for almost everywhere $t\in(\frac{1}{2},1)$. It then follows by the Sobolev compactness theorem, together with the fact that $s>n-1=\text{dimension}(\partial B_t)$, that

$$\lim_{m\to\infty}\lambda_m^{1-2/s}\sup_{\partial B_t}|v-v_m|=0 \qquad (5.60)$$

for almost everywhere $t\in(\frac{1}{2},1)$.

Remark 5.9. This is the main point in the proof where we require $s>n-1$.

It follows from (5.57)-(5.60) that

$$\limsup_{m\to\infty}\int_{B_t\sim B_{t-\delta/2}}\sum_{i=1}^{k}F_m^i(Dw_m)\leqq c\delta(1+M_t) \qquad (5.61)$$

for almost everywhere $t\in(\frac{1}{2},1)$ and some subsequence depending on t.

On the other hand, we easily see from (5.52), (5.53), and (5.45) that

$$\int_{B_{t-\delta/2}\sim B_{t-\delta}}\sum_{i=1}^{k}F_m^i(Dw_m)\leqq c\delta. \qquad (5.62)$$

Finally, from (5.52), (5.53), and (5.38) we see that

$$\left|\int_{B_t\sim B_{t-\delta}}\sum_{i=1}^{k}F_m^i(Dv)\right|\leqq c\delta. \qquad (5.63)$$

From (5.51) and the fact that $\delta\in(0,\frac{1}{4})$ is otherwise arbitrary, it now follows, for almost everywhere $t\in(\frac{1}{2},1)$ and for some subsequence depending on t, that (5.39) is true.

LEMMA 5.10. *Suppose* $\{u_m\}$ *is a sequence of* $L^1(B_1)$ *functions such that* $\int_{B_1}|u_m| \leqq M$, *where* M *is independent of* m. *Then for almost everywhere* $t \in (0,1)$ *there exists a constant* $M_t < \infty$ *and a subsequence depending in* t *such that*

$$I_m(t) := \int_{\partial B_t} |u_m| \leqq M_t.$$

Proof. If the claim were not true, there would exist $S \subset (0,1)$ with $\mathscr{L}^1(S) = \alpha > 0$ such that $t \in S$ implies $\lim_{m\to\infty} I_m(t) = \infty$.

Let

$$S_i = \left\{ t \in S : I_m(t) \geqq \frac{2M}{\alpha} \quad \text{if} \quad m \geqq i \right\}.$$

Since $S_i \uparrow S$ as $i \to \infty$, it follows $\mathscr{L}^1(S_j) > \alpha/2$ for some j. But then

$$\int_{B_1} |u_j| = \int_0^1 \int_{\partial B_t} |u_j| > \frac{\alpha}{2} \cdot \frac{2M}{\alpha} = M,$$

which contradicts the hypotheses.

This proves the lemma. □

D. An estimate of $I^t_m(v_m) - I^t_m(v)$ from below. By estimating $I^t_m(v_m) - I^t_m(v)$ from below, and using (5.39), we will establish that, for almost everywhere $t \in (\frac{1}{2}, 1)$ and for some subsequence depending on t (which can be taken to be the same subsequence as for (5.39)), the following hold:

$$\lim_{m\to\infty} \fint_{B_t} |Dv_m - Dv|^2 = 0, \tag{5.64}$$

$$\lim_{m\to\infty} \fint_{B_t} \lambda_m^{s-2} |Dv_m|^s = 0 \quad \text{if } s > 2, \tag{5.65}$$

$$\lim_{m\to\infty} \fint_{B_t} |\Lambda_i A_m|^{s-2} \lambda_m^{2i-2} |\Lambda_i Dv_m|^2 = 0 \quad \text{if } i = 2, \cdots, k \quad \text{and } s > 2, \tag{5.66}$$

$$\lim_{m\to\infty} \fint_{B_t} \lambda_m^{is-2} |\Lambda_i Dv_m|^s = 0 \quad \text{if } i = 2, \cdots, k. \tag{5.67}$$

Thus, weak convergence in (5.11) is converted into strong convergence.

To show these limits recall

$$I^t_m(v_m) - I^t_m(v) = \fint_{B_t} \sum_{i=1}^{k} F^i_m(Dv_m) - F^i_m(Dv). \tag{5.68}$$

Write

$$\begin{aligned} \fint_{B_t} F^1_m(Dv_m) - F^1_m(Dv) &= \fint_{B_t} F^1_m(Dv_m - Dv) \\ &\quad + \fint_{B_t} F^1_m(Dv_m) - F^1_m(Dv) - F^1_m(Dv_m - Dv) \\ &= A^1_m + B^1_m. \end{aligned} \tag{5.69}$$

From (5.12)

$$A_m^1 = \fint_{B_t} \left(\int_0^1 (1-\tau) D^2 F^1[A_m + \tau\lambda_m(Dv_m - Dv)]\, d\tau \right)(Dv_m - Dv)(Dv_m - Dv)$$

$$\geqq \gamma \fint \left(\int_0^1 (1-\tau)(1+|A_m + \tau\lambda_m(Dv_m - Dv)|^{s-2})\, d\tau \right)|Dv_m - Dv|^2$$

(using (H3)(b)).

But from [E, Lemma 8.1] there exists $\sigma > 0$ such that, for any two $n \times N$ matrices A and B,

$$\sigma(|A|^{s-2} + |B|^{s-2}) \leqq \int_0^1 (1-\tau)|A + \tau B|^{s-2}\, d\tau. \tag{5.70}$$

Thus, for some $\sigma > 0$,

$$\begin{aligned} A_m^1 &\geqq \sigma \fint_{B_t} (1 + |A_m|^{s-2} + \lambda_m^{s-2}|Dv_m - Dv|^{s-2})|Dv_m - Dv|^2 \\ &\geqq \sigma \fint_{B_t} |Dv_m - Dv|^2 + \lambda_m^{s-2}|Dv_m - Dv|^s. \end{aligned} \tag{5.71}$$

On the other hand, we can estimate B_m^1 in a similar way to that in [EG, eqn. (2.16)]. Namely,

$$B_m^1 = \fint_{B_t} \left(\int_0^1 DF_m^1(Dv + \tau(Dv_m - Dv))\, d\tau \right)(Dv_m - Dv)$$

$$- \fint_{B_t} \left(\int_0^1 DF_m^1(\tau(Dv_m - Dv))\, d\tau \right)(Dv_m - Dv)$$

(using the fact that $F_m^1(0) = 0$)

$$= \fint_{B_t} \left(\int_0^1 \int_0^1 D^2 F_m^1(\nu Dv + \tau(Dv_m - Dv))\, d\tau\, d\nu \right)(Dv_m - Dv)Dv$$

$$= \fint_{B_t} \left(\int_0^1 \int_0^1 D^2 F^1[A_m + \lambda_m(\nu Dv + \tau(Dv_m - Dv))]\, d\tau\, d\nu \right)$$

$$\cdot (Dv_m - Dv)Dv$$

(using (5.12)).

We now apply Lemma 5.5 with

$$f_m = \int_0^1 \int_0^1 D^2 F^1[A_m + \lambda_m(\nu Dv + \tau(Dv_m - Dv))]\, d\tau\, d\nu,$$

$$g_m = Dv_m - Dv.$$

Then $f_m \to D^2 F^1(A)$ almost everywhere and $g_m \rightharpoonup 0$ in L^2.

Also,

$$\begin{aligned} |f_m g_m| &\leqq c(1 + \lambda_m^{s-2}|Dv_m|^{s-2})(1 + |Dv_m|) \\ &\leqq c(1 + |Dv_m| + \lambda_m^{s-2}|Dv_m|^{s-1}) \\ &= c(1 + |Dv_m| + \lambda_m^{1-2/s}(\lambda_m^{s-2}|Dv_m|^s)^{(s-1)/s}). \end{aligned}$$

It follows that $f_m g_m \in L^{s/(s-1)}$ uniformly in m.

From Lemma 5.5 it now follows that

$$B_m^1 \to 0 \quad \text{as } m \to \infty. \tag{5.72}$$

We next consider $F_m^i(Dv_m) - F_m^i(Dv)$ for $2 \leqq i \leqq k$. These cases are treated a little differently from $i = 1$. We have

$$\begin{aligned}
&\fint_{B_t} F_m^i(Dv_m) - F_m^i(Dv) \\
&\quad = \lambda_m^{-2} \fint_{B_t} \left(\int_0^1 (1-\tau) D^2 F^i[\Lambda_i A_m + \tau(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m)]\, d\tau \right) \\
&\qquad \cdot (\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m)(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m) \\
&\qquad - \lambda_m^{-2} \fint_{B_t} \left(\int_0^1 (1-\tau) D^2 F^i[\Lambda_i A_m + \tau(\Lambda_i A_m + \lambda_m Dv) - \Lambda_i A_m)]\, d\tau \right) \\
&\qquad \cdot (\Lambda_i(A_m + \lambda_m Dv) - \Lambda_i A_m)(\Lambda_i(A_m + \lambda_m Dv) - \Lambda_i A_m)
\end{aligned}$$

(from (5.13))

$$\begin{aligned}
&= \lambda_m^{-2} \fint_{B_t} \left(\int_0^1 (1-\tau) D^2 F^i[\Lambda_i A_m + \tau(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m)]\, d\tau \right) \\
&\qquad \cdot [(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m)(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m) \\
&\qquad\quad - (\Lambda_i(A_m + \lambda_m Dv) - \Lambda_i A_m)(\Lambda_i(A_m + \lambda_m Dv) - \Lambda_i A_m)] \\
&\qquad + \lambda_m^{-2} \fint_{B_t} \left(\int_0^1 (1-\tau) D^2 F^i[\Lambda_i A_m + \tau(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m)] \right. \\
&\qquad\quad \left. - (1-\tau) D^2 F^i[\Lambda_i A_m + \tau(\Lambda_i(A_m + \lambda_m Dv) - \Lambda_i A_m)]\, d\tau \right) \\
&\qquad \cdot (\Lambda_i(A_m + \lambda_m Dv) - \Lambda_i A_m)(\Lambda_i(A_m + \lambda_m Dv) - \Lambda_i A_m) \\
&\quad = C_m^i + D_m^i.
\end{aligned} \tag{5.73}$$

Moreover,

$$\begin{aligned}
C_m^i &= \lambda_m^{-2} \fint_{B_t} \left(\int_0^1 (1-\tau) D^2 F^i[\Lambda_i A_m + \tau(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m)]\, d\tau \right) \\
&\qquad \cdot ([\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i(A_m + \lambda_m Dv)] \\
&\qquad\quad \cdot [\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i(A_m + \lambda_m Dv)] \\
&\qquad\quad + 2[\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i(A_m + \lambda_m Dv)] \\
&\qquad\quad \cdot [\Lambda_i(A_m + \lambda_m Dv) - \Lambda_i A_m]) \\
&\quad = E_m^i + F_m^i,
\end{aligned} \tag{5.74}$$

where we have used the identity

$$D_{q_\alpha q_\beta} F^i(\xi_\alpha \xi_\beta - \eta_\alpha \eta_\beta) = D_{q_\alpha q_\beta} F^i[(\xi_\alpha - \eta_\alpha)(\xi_\beta - \eta_\beta) + 2(\xi_\alpha - \eta_\alpha)\eta_\beta].$$

We have for some $\sigma > 0$,

$$E^i_m \geqq \gamma\lambda_m^{-2} \fint_{B_t} \left(\int_0^1 (1-\tau)|\Lambda_i A_m + \tau(\Lambda_i(A_m+\lambda_m Dv_m) - \Lambda_i A_m)|^{s-2}\, d\tau \right.$$
$$\left. \cdot |\Lambda_i(A_m+\lambda_m Dv_m) - \Lambda_i(A_m+\lambda_m Dv)|^2 \right)$$

(using (H4)(b))

$$\geqq \sigma\lambda_m^{-2} \fint_{B_t} (|\Lambda_i A_m|^{s-2} + |\Lambda_i(A_m+\lambda_m Dv_m) - \Lambda_i A_m|^{s-2})$$
$$\cdot |\Lambda_i(A_m+\lambda_m Dv_m) - \Lambda_i(A_m+\lambda_m Dv)|^2$$

(using (5.70))

$$\geqq \sigma \fint_{B_t} \left(|\Lambda_i A_m|^{s-2} + \lambda_m^{s-2} \left| \sum_{j=1}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv_m \right|^{s-2} \right)$$
$$\cdot \left| \sum_{j=1}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot (\Lambda_j Dv_m - \Lambda_j Dv) \right|^2. \tag{5.75}$$

Also,

$$F^i_m = 2 \fint_{B_t} \left(\int_0^1 (1-\tau) D^2 F^i [\Lambda_i A_m + \tau(\Lambda_i(A_m+\lambda_m Dv) - \Lambda_i A_m)]\, d\tau \right)$$
$$\cdot \left(\Lambda_{i-1} A_m \tilde{\odot} (Dv_m - Dv) + \sum_{j=2}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot (\Lambda_j Dv_m - \Lambda_j Dv) \right)$$
$$\cdot \left(\sum_{j=1}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv \right) \tag{5.76}$$
$$= G^i_m + H^i_m.$$

We now have from (5.68), (5.69), (5.71), (5.73), (5.74), and (5.75)-(5.76) that

$$I^t_m(v_m) - I^t_m(v) \geqq \sigma \fint_{B_t} (|Dv_m - Dv|^2 + \lambda_m^{s-2} |Dv_m - Dv|^s) + B^1_m$$
$$+ \sigma \sum_{i=2}^{k} \fint_{B_t} \left(|\Lambda_i A_m|^{s-2} + \lambda_m^{s-2} \left| \sum_{j=1}^{i} \lambda_m^j \Lambda_{i-j} A_m \odot \Lambda_j Dv_m \right|^{s-2} \right)$$
$$\cdot \left| \sum_{j=1}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot (\Lambda_j Dv_m - \Lambda_j Dv) \right|^2 \tag{5.77}$$
$$+ \sum_{i=2}^{k} (G^i_m + H^i_m + D^i_m).$$

We first apply Lemma 5.5 to G^i_m, with

$$f_m = 2 \int_0^1 (1-\tau) D^2 F^i [\Lambda_i(A_m + \tau(\Lambda_i(A_m+\lambda_m Dv_m) - \Lambda_i A_m)]\, d\tau),$$

$$g_m = \Lambda_{i-1} A_m \tilde{\odot} (Dv_m - Dv) \lambda_m^{j-1} \Lambda_{i-j} A_m, \quad \text{for each } j = 1, \cdots, i.$$

Then $f_m \to D^2F^i(\Lambda_i A)$ almost everywhere, and $g_m \rightharpoonup 0$ in L^2. Moreover,

$$\begin{aligned}
|f_m g_m| &\leqq c\left(|\Lambda_i A_m|^{s-2} + \sum_{j=1}^{i} \lambda_m^{j(s-2)} |\Lambda_{i-j} A_m|^{s-2} |\Lambda_j Dv_m|^{s-2}\right)(1+|Dv_m|) \\
&\leqq c\left(1+|Dv_m| + \sum_{j=1}^{i} \lambda_m^{j(s-2)} |\Lambda_j Dv_m|^{s-2} + \sum_{j=1}^{i} \lambda_m^{j(s-2)} |\Lambda_j Dv_m|^{s-2} |Dv_m|\right) \\
&\leqq c\Bigg(1+|Dv_m| + \lambda_m^{2-4/s} \sum_{j=1}^{i} (\lambda_m^{j-2/s} |\Lambda_j Dv_m|)^{s-2} \\
&\qquad + \lambda_m^{1-2/s} \sum_{j=1}^{i} (\lambda_m^{j-2/s} |\Lambda_j Dv_m|)^{s-2} (\lambda_m^{1-2/s} |Dv_m|)\Bigg) \\
&\leqq c\Bigg(1+|Dv_m| + \lambda_m^{2-4/s} \sum_{j=1}^{i} (\lambda_m^{js-2} |\Lambda_j Dv_m|^s)^{(s-2)/s} \\
&\qquad + \lambda_m^{1-2/s} \sum_{j=1}^{i} (\lambda_m^{j-2/s} |\Lambda_j Dv_m|)^{s-1} + \lambda_m^{1-2/s} (\lambda_m^{1-2/s} |Dv_m|)^{s-1}\Bigg).
\end{aligned}$$

It follows that $f_m g_m \in L^{s/(s-1)}$, uniformly in m. From Lemma 5.5 it follows that

$$G_m^i \to 0 \quad \text{as } m \to \infty. \tag{5.78}$$

We next claim

$$H_m^i \to 0 \quad \text{as } m \to \infty. \tag{5.79}$$

To see this, apply Lemma 5.5 with

$$\begin{aligned}
f_m &= 2\left(\int_0^1 (1-\tau) D^2F^i[\Lambda_i A_m + \tau(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m)]\, d\tau\right) \\
&\quad \cdot \left(\sum_{j=2}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot (\Lambda_j Dv_m - \Lambda_j Dv)\right) \lambda_m^{j-1} \Lambda_{i-j} A_m, \\
g_m &= 1,
\end{aligned}$$

where we fix $j \in \{1, \cdots, i\}$.

Then $f_m \to 0$ almost everywhere (using (5.11)). Moreover,

$$\begin{aligned}
|f_m| &\leqq c\left(|\Lambda_i A_m|^{s-2} + \sum_{j=1}^{i} \lambda_m^{j(s-2)} |\Lambda_j Dv_m|^{s-2}\right)\left(1 + \sum_{j=2}^{i} \lambda_m^{j-1} |\Lambda_j Dv_m|\right) \\
&\leqq c\Bigg[|\Lambda_i A_m|^{s-2}\left(1 + \sum_{j=2}^{i} \lambda_m^{j-1} |\Lambda_j Dv_m|\right) + \sum_{j=1}^{i} \lambda_m^{j(s-2)} |\Lambda_j Dv_m|^{s-2} \\
&\qquad + \sum_{j=1}^{i} \lambda_m^{-1} \lambda_m^{j(s-1)} |\Lambda_j Dv_m|^{s-1}\Bigg] \\
&\leqq c\Bigg[1 + \sum_{j=2}^{i} |\Lambda_j A_m|^{s/2-1} \lambda_m^{j-1} |\Lambda_j Dv_m| + \lambda_m^{2-4/s} \sum_{j=1}^{i} (\lambda_m^{js-2} |\Lambda_j Dv_m|^s)^{(s-2)/s} \\
&\qquad + \lambda_m^{1-2/s} \sum_{j=1}^{i} (\lambda_m^{js-2} |\Lambda_j Dv_m|^s)^{(s-1)/s}\Bigg].
\end{aligned}$$

It follows from (5.11) that $f_m \in L^{s/(s-1)}$, uniformly in m. From Lemma 5.5 we see that (5.79) holds.

Finally, we claim that

$$D^i_m \to 0 \quad \text{as } m \to \infty. \tag{5.80}$$

To show this, let

$$\begin{aligned} f_m = \Bigg[\int_0^1 & (1-\tau) D^2 F^i(\Lambda_i A_m + \tau(\Lambda_i(A_m + \lambda_m Dv_m) - \Lambda_i A_m)) \\ & -(1-\tau) D^2 F^i(\Lambda_i A_m + \tau(\Lambda_i(A_m + \lambda_m Dv) - \Lambda_i A_m))\, d\tau \Bigg] \\ & \cdot \left(\sum_{j=1}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv \right) \left(\sum_{j=1}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv \right). \end{aligned}$$

Then $f_m \to 0$ almost everywhere. Moreover,

$$\begin{aligned} |f_m| &\leqq c \left(|\Lambda_i A_m|^{s-2} + \sum_{j=1}^{i} \lambda_m^{j(s-2)} |\Lambda_j Dv_m|^{s-2} + \sum_{j=1}^{i} \lambda_m^{j(s-2)} |\Lambda_j Dv|^{s-2} \right) \\ &\leqq c \left(1 + \sum_{j=1}^{i} \lambda_m^{j(s-2)} |\Lambda_j Dv_m|^{s-2} \right) \\ &\leqq c \left(1 + \lambda_m^{2-4/s} \left(\sum_{j=1}^{i} \lambda_m^{js-2} |\Lambda_j Dv_m|^s \right)^{(s-2)/s} \right). \end{aligned}$$

Thus $f_m \in L^{s/(s-2)} (\in L^\infty$ if $s=2)$ uniformly in m.

If we apply Lemma 5.5 with $g_m \equiv 1$ we can now establish (5.80).

Combining (5.39) with (5.72) and (5.77)-(5.80), we have that for almost everywhere $t \in (\frac{1}{2}, 1)$ there exists a subsequence depending on t for which

$$\begin{aligned} 0 &\geqq \lim_{m\to\infty} (I^t_m(v_m) - I^t_m(v)) \\ &\geqq \lim_{m\to\infty} \Bigg(\fint_{B_t} |Dv_m - Dv|^2 + \lambda_m^{s-2} |Dv_m - Dv|^s \\ &\qquad + \sum_{i=2}^{k} \fint_{B_t} |\Lambda_i A_m|^{s-2} \left| \sum_{j=1}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot (\Lambda_j Dv_m - \Lambda_j Dv) \right|^2 \\ &\qquad + \sum_{i=2}^{k} \fint_{B_t} \lambda_m^{s-2} \left| \sum_{j=1}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv_m \right|^{s-2} \\ &\qquad \cdot \left| \sum_{j=1}^{i} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot (\Lambda_j Dv_m - \Lambda_j Dv) \right|^2 \Bigg). \end{aligned} \tag{5.81}$$

This now establishes (5.64) and (5.65). Moreover, from the triangle inequality,

$$\begin{aligned} \lim_{m\to\infty} \fint_{B_t} |\Lambda_i A_m|^{s-2} & \left| \Lambda_{i-1} A_m \tilde{\odot} (Dv_m - Dv) \right. \\ & \left. + \sum_{j=2}^{i-1} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv_m + \lambda_m^{i-1} \Lambda_i Dv_m \right|^2 = 0, \end{aligned} \tag{5.82}$$

$$\begin{aligned} \lim_{m\to\infty} \fint_{B_t} & \lambda_m^{s-2} \left| \sum_{j=1}^{i-1} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv_m + \lambda_m^{i-1} \Lambda_i Dv_m \right|^{s-2} \\ & \cdot \left| \sum_{j=1}^{i-1} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv_m + \lambda_m^{i-1} \Lambda_i Dv_m + \varphi^i_m \right|^2 = 0, \end{aligned} \tag{5.83}$$

where φ^i_m is a smooth function, uniformly bounded in m, and $i = 2, \cdots, k$.

From (5.82) with $i=2$, and (5.64), it follows (5.66) is true with $i=2$. Suppose (5.66) is true for $i = 2, \cdots, l (l < k)$. Then it follows from (5.82) with $i = l+1$, the fact

that $|\Lambda_{l+1}A_m| \leqq c|\Lambda_i A_m|$ if $i \leqq l$, and from (5.64), that (5.66) is true with $i = l+1$. It now follows that (5.66) holds for $i = 2, \cdots, k$.

We now claim that (5.67) is true for $i = 2, \cdots, k$.

First note that for $j = 1, \cdots, k$ we have

$$\fint_{B_t} \lambda_m^{s-2} |\lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv_m|^{s-2} |\varphi_m^i|^2 \leqq c \fint_{B_t} \lambda_m^{2-4/s} (\lambda_m^{j-2/s} |\Lambda_j Dv_m|)^{s-2} \to 0 \quad \text{as } m \to \infty,$$

since $s > 2$.

From this and (5.83) it follows that

$$\lim_{m\to\infty} \fint_{B_t} \lambda_m^{s-2} \left| \sum_{j=1}^{i-1} \lambda_m^{j-1} \Lambda_{i-j} A_m \odot \Lambda_j Dv_m + \lambda_m^{i-1} \Lambda_i Dv_m \right|^s = 0 \tag{5.84}$$

for $i = 2, \cdots, k$.

From (5.65) and (5.84) with $i = 2$, it follows that (5.67) is true with $i = 2$. Suppose (5.67) holds for $i = 2, \cdots, l (l < k)$. Then it follows from (5.84) with $i = l+1$ that (5.67) is true with $i = l+1$. This establishes (5.67) for $i = 2, \cdots, k$ in the case where $s > 2$.

Consider now the case where $s = 2$. Then directly from (5.81) we have in the case where $i = 2$ that

$$0 \geqq \lim_{m\to\infty} \left(\fint_{B_t} |Dv_m - Dv|^2 + |A_m \odot (Dv_m - Dv) + \lambda_m (\Lambda_2 Dv_m - \Lambda_2 Dv)|^2 \right).$$

Using (5.64) it follows that

$$0 = \lim_{m\to\infty} \fint_{B_t} \lambda_m^2 |\Lambda_2 Dv_m|^2.$$

Since $s = 2$ implies $n = 2$, and hence $\Lambda_i Dv_m$ is trivial if $i > 2$, it follows that (5.67) is now established in the case where $s = 2$.

This completes the proof of (5.64)-(5.67).

E. Completion of proof. Choose $t \in (\frac{1}{2}, 1)$ such that (5.64)-(5.67) are true for some subsequence.

In order to obtain a contradiction to (5.2) we first use (5.1) and (5.59) to compute

$$\begin{aligned} \lambda_m^{-2} U(x_m, \tau r_m) = & \fint_{B_\tau} |Dv_m - (Dv_m)_\tau|^2 + \fint_{B_\tau} \lambda_m^{s-2} |Dv_m - (Dv_m)_\tau|^s \\ & + \fint_{B_\tau} \sum_{i=2}^{k} |\Lambda_i (A_m + \lambda_m (Dv_m)_\tau)|^{s-2} \lambda_m^{2i-2} |\Lambda_i (Dv_m - (Dv_m)_\tau)|^2 \\ & + \fint_{B_\tau} \sum_{i=2}^{k} \lambda_m^{si-2} |\Lambda_i (Dv_m - (Dv_m)_\tau)|^s. \end{aligned} \tag{5.85}$$

Since $\tau \leqq \frac{1}{2} < t$, it follows from (5.64) that $Dv_m \to Dv$ in $L^2(B_\tau)$, and hence $(Dv_m) \to (Dv)_\tau$. Thus the integral on the right side of (5.85) converges to $f_{B_\tau} |Dv - (Dv)_\tau|^2$.

If $s > 2$ then $\lambda_m^{1-2/s} Dv_m \to 0$ in $L^s(B_\tau)$ from (5.65), and hence $\lambda_m^{1-2/s} Dv_m \to 0$. Thus the second integral on the right side converges to zero.

Next note that since $(Dv_m)_\tau \to (Dv)_\tau$, it follows that $\Lambda_i (Dv_m)_\tau$ is bounded uniformly in m for $i = 2, \cdots, k$ (with the bound possibly depending on τ).

Hence, for $i=2,\cdots,k$,

$$\fint_{B_\tau}|\Lambda_i(A_m+\lambda_m(Dv_m)_\tau|^{s-2}\lambda_m^{2i-2}|\Lambda_i(Dv_m-(Dv_m)_\tau)|^s$$

$$\leqq c\fint_{B_\tau}(|\Lambda_iA_m|^{s-2}+\lambda_m^{s-2})\lambda_m^{2i-2}\left(1+\sum_{j=1}^{i}|\Lambda_jDv_m|^2\right)$$

$$\leqq c\fint_{B_\tau}|\Lambda_iA_m|^{s-2}\lambda_m^{2i-2}+\sum_{j=1}^{i}|\Lambda_jA_m|^{s-2}\lambda_m^{2j-2}|\Lambda_jDv_m|^2$$

$$+\lambda_m^{s+2i-4}+\sum_{j=1}^{i}\lambda_m^{s+2j-4}|\Lambda_jDv_m|^2$$

(using $|\Lambda_iA_m|\leqq c|\Lambda_jA_m|$ if $1\leqq j\leqq i$)

$$\leqq c\fint_{B_\tau}|\Lambda_iA_m|^{s-2}\lambda_m^{2i-2}+\sum_{j=1}^{i}|\Lambda_jA_m|^{s-2}\lambda_m^{2j-2}|\Lambda_jDv_m|^2$$

$$+\lambda_m^{s+2i-4}+\lambda_m^{s(1-2/s)^2}\left(\sum_{j=1}^{i}\lambda_m^{j-2/s}|\Lambda_jDv_m|\right)^2$$

$$\to 0 \quad \text{as } m\to\infty$$

(using (5.66), (5.67), and the fact we may here assume that $s>2$).

Finally, for $i=2,\cdots,k$,

$$\fint_{B_\tau}\lambda_m^{is-2}|\Lambda_i(Dv_m-(Dv_m)_\tau)|^s\leqq c\fint_{B_\tau}\lambda_m^{is-2}\left(1+\sum_{j=1}^{i}|\Lambda_jDv_m|^s\right)$$

$$\to 0 \quad \text{as } m\to\infty,$$

from (5.67).

It follows that

$$\lambda_m^{-2}U(x_m,\tau r_m)\to\fint_{B_\tau}|Dv-(Dv)_\tau|^2$$

as $m\to\infty$. But

$$\fint_{B_\tau}|Dv-(Dv)_\tau|^2\leqq c\tau^2\sup_{B_\tau}|D^2v|^2$$

$$\leqq c_2\tau^2 \tag{5.86}$$

from (5.38), where $c_2=c_2(M)$. Hence

$$\lambda_m^{-2}U(x_m,\tau r_m)\leqq 2c_2\tau^2$$

for all sufficiently large m, contradicting (5.2) if we set $c_1=2c_2(M)$.

This contradiction completes the proof of Theorem 5.1. □

6. Proof of the main theorem. We first show how Theorem 5.1 is iterated in the standard manner. We continue to assume $u\in\Lambda_kW^{1,s}(\Omega)$ is a minimiser of $I[\cdot]$ and $I[\cdot]$ satisfies the hypotheses (H1)-(H4).

LEMMA 6.1. *Suppose* $0<\alpha<1$ *and* $M>0$. *Then there exists* $0<\tau<\frac{1}{2}$ *and* $\varepsilon>0$ *such that, for any* $B_r(x)\subset\Omega$, *if*

$$|(Du)_{x,r}|<M,$$
$$U(x,r)<\varepsilon,$$

then

$$U(x,\tau^l r)\leqq(\tau^l)^{2\alpha}U(x,r)$$

for each $l=1,2,\cdots$.

Proof. Using the notation of Theorem 5.1, first choose τ such that

$$0<\tau<\tfrac{1}{2},\quad c_1(2M)\tau^2\leqq\tau^{2\alpha},\quad \tau^\alpha<\tfrac{1}{2}. \tag{6.1}$$

Next, choose ε so that

$$0<\varepsilon\leqq\varepsilon_0(\tau,2M),\qquad \tau^{-n}\varepsilon^{1/2}\leqq\frac{M}{2}. \tag{6.2}$$

Suppose

$$B_r(x)\subset\Omega,\quad |(Du)_{x,r}|<M,\quad U(x,r)<\varepsilon. \tag{6.3}$$

We will prove by induction on $l\geqq0$ that

$$\begin{aligned}&|(Du)_{x,\tau^l r}|\leqq M\left(1+\frac{1}{2}+\cdots+\frac{1}{2^l}\right),\\ &U(x,\tau^l r)\leqq(\tau^l)^{2\alpha}U(x,r).\end{aligned} \tag{6.4}$$

First note that

$$\begin{aligned}|(Du)_{x,\tau^{l+1}r}-(Du)_{x,\tau^l r}|&\leqq\fint_{B(x,\tau^{l+1}r)}|Du-(Du)_{x,\tau^l r}|\\ &\leqq\tau^{-n}\fint_{B(x,\tau^l r)}|Du-(Du)_{x,\tau^l r}|\\ &\leqq\tau^{-n}[U(x,\tau^l r)]^{1/2}.\end{aligned} \tag{6.5}$$

Note that (6.4) is trivially true of $l=0$. Assume that it is true for some l. Then

$$\begin{aligned}|(Du)_{x,\tau^{l+1}r}|&\leqq|(Du)_{x,\tau^l r}|+|(Du)_{x,\tau^{l+1}r}-(Du)_{x,\tau^l r}|\\ &\leqq M\left(1+\frac{1}{2}+\cdots+\frac{1}{2^l}\right)+\tau^{-n}(U(x,\tau^l r))^{1/2}\\ &\leqq M\left(1+\frac{1}{2}+\cdots+\frac{1}{2^l}\right)+\tau^{-n}(\tau^l)^\alpha(U(x,r))^{1/2}\end{aligned}$$

(using (6.4))

$$\leqq M\left(1+\frac{1}{2^l}+\cdots+\frac{1}{2^l}+(\tau^l)^\alpha\frac{M}{2}\right)$$

(using (6.2), (6.3))

$$\leqq M\left(1+\frac{1}{2}+\cdots+\frac{1}{2^l}+\frac{1}{2^{l+1}}\right)$$

(using (6.1)).

Also, from (6.4), (6.2), (6.3) we can apply Theorem 5.1 with M and r there replaced by $2M$ and $\tau^l r$, to deduce

$$U(x, \tau^{l+1}r) \leqq c_1(2M)\tau^2 U(x, \tau^l r)$$
$$\leqq \tau^{2\alpha}(\tau^l)^{2\alpha} U(x, r)$$

(using (6.1) and (6.4))

$$= (\tau^{l+1})^{2\alpha} U(x, r).$$

The proof is now completed by using induction of l. □

LEMMA 6.2. *Suppose* $u \in \Lambda_k W^{1,s}(\Omega)$. *Then*

$$\lim_{r\to 0} \fint_{B(x,r)} |\Lambda_i(Du - (Du)_{x,r})|^s = 0$$

for almost everywhere $x \in \Omega$ *and for* $i = 1, \cdots, k$.

Proof. Let $\{A_\alpha\}_{\alpha\in S}$ be a countable dense subset of Hom $(\mathbb{R}^n; \mathbb{R}^N)$.

By considering the set of Lebesgue points of Du and of the functions $|\Lambda_i(Du - A_\alpha)|^s$ for $i = 1, \cdots, k$ and $\alpha \in S$, there exists $E \subset \Omega$ with $\mathscr{L}^n(\Omega \sim E) = 0$ such that

$$(Du)_{x,r} \to Du(x),$$

$$\fint_{B_{x,r}} |\Lambda_i(Du - A_\alpha)|^s \to |\Lambda_i(Du(x) - A_\alpha)|^s,$$

as $r \to 0$, for all $x \in E$, $\alpha \in S$, and $i = 1, \cdots, k$.

We have for all such x, α, and i,

$$\limsup_{r\to 0} \fint_{B(x,r)} |\Lambda_i(Du - (Du)_{x,r}|^s$$

$$= \limsup_{r\to 0} \fint_{B(x,r)} |\Lambda_i(Du - A_\alpha + A_\alpha - (Du)_{x,r}|^s$$

$$\leqq c \limsup_{r\to 0} \sum_{j=0}^{i} |\Lambda_j(A_\alpha - (Du)_{x,r})|^s \fint_{B(x,r)} |\Lambda_{i-j}(Du - A_\alpha)|^s$$

(using (2.17))

$$= c \sum_{j=0}^{i} |\Lambda_j(A_\alpha - Du(x))|^s |\Lambda_{i-j}(Du(x) - A_\alpha)|^s.$$

Since A_α is dense in Hom $(\mathbb{R}^n; \mathbb{R}^N)$, we can make the right side arbitrarily small. This establishes the lemma. □

6.1. Completion of proof of Main Theorem 4.1. Let

$$\Omega_0 = \Big\{ x \in \Omega: \lim_{r\to 0} (Du)_{x,r} = Du(x), \quad \text{and}$$

$$\lim_{r\to 0} \fint_{B(x,r)} |\Lambda_i(Du - (Du)_{x,r})|^s = 0 \quad \text{for } i = 1, \cdots, k \Big\}.$$

Then $\mathscr{L}^n(\Omega \sim \Omega_0) = 0$ by Lemma 6.2. Moreover, by standard arguments (e.g., [E] or [G]) it follows from Lemma 6.1 that Ω_0 is open and $u \in C^{1,\alpha}(\Omega_0)$. □

7. Concluding remarks. It is clear that the methods of the paper can be applied to other problems.

7.1. We could include quasi-convex, nonpolyconvex terms in (4.2) by allowing F^1 to be quasi convex. More precisely, the hypotheses (H1)-(H4) would remain unchanged, except that (H3)(b) would be replaced by the strict quasi-convexity condition

$$\int F^1(p+D\varphi) \geqq \int F^1(p)+\gamma(|D\varphi|^2+|D\varphi|^s))$$

for all $\varphi \in C_c^\infty(\mathbb{R}^n;\mathbb{R}^N)$ and all $p \in \mathbb{R}^{n\times N}$.

The F^1 term is then treated essentially as in [EG] to obtain (5.71). Note that (5.30) needs to be replaced by the appropriate Legendre-Hadamard condition.

7.2. Model problems of the form

$$I[u]=\int_\Omega \left(1+|Du|^2+\sum_{i=2}^k |\Lambda_i Du|^2\right)^{s/2},$$

or more generally of the form

$$I[u]=\int_\Omega |Du|^s+\left(1+|Du|^2+\sum_{i=2}^k |\Lambda_i Du|^2\right)^{s'/2},$$

where $s \geqq 2$ if $n=2$, $s>n-1$ if $n>2$, and $s' \geqq 2$, are easily treated by our methods. Such model problems were discussed (from an existence point of view) in [GMS, § 6.A].

In order to prove partial regularity, we take the quantity $U(x, r)$ to be

$$U(x,r)=\fint_{B(x,r)} |Du-(Du)_{x,r}|^2+|Du-(Du)_{x,r}|^{\max\{s,s'\}} + \sum_{i=2}^k [|\Lambda_i(Du-(Du)_{x,r})|^2+|\Lambda_i(Du-(Du)_{x,r}|^{s'}].$$

The arguments involved in this case, as opposed to the arguments involved in treating the model problem (1.3), are somewhat easier, as no degeneracy in ellipticity in the $\Lambda_i Du$ terms is allowed.

Partial regularity theory for such problems, but in the case $s>n$, has also been recently treated in [GMS2].

7.3. Many model problems of the form

$$\int |Du|^2+|Du|^{s_1}+\sum_{i=2}^k |\Lambda_i Du|^{s_i},$$

where $s_1 \geqq 2$ if $n=2$, $s_1>n-1$ if $n-2$, and $s_1 \geqq s_2 \geqq \cdots \geqq s_k \geqq 2$, are covered by the present arguments. More generally, we could include more than one term involving $\Lambda_i Du$. We have not checked, however, the appropriate most general conditions on the exponents.

Similarly, we could formulate more general structure conditions. It is not necessary that the integrand in (4.2) split as a sum of terms involving the $\Lambda_i Du$ separately. In this more general case, we define the normalised functional

$$I_m[w]=\int_\Omega F_m(Dw),$$

where F_m is obtained by linearisation about $(A_m, \Lambda_2 A_m, \cdots, \Lambda_k A_m)$.

Acknowledgments. The support of the "Comitato Nazionale per le Scienze Mathematiche" and the friendly environment of the Mathematics Department at the University of Salerno are gratefully acknowledged.

REFERENCES

[AF] E. ACERBI AND N. FUSCO, *Semicontinuity problems in the calculus of variations*, Arch. Rational Mech. Anal., 86 (1984), pp. 125-145.

[B] J. M. BALL, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337-403.

[E] L. C. EVANS, *Quasiconvexity and partial regularity in the calculus of variations*, Arch. Rational Mech. Anal., 95 (1986), pp. 227-252.

[EG] L. C. EVANS AND R. F. GARIEPY, *Blow-up, compactness and partial regularity in the calculus of variations*, Indiana U. Math. J., 36 (1987), pp. 361-371.

[FH] N. FUSCO AND J. E. HITCHINSON, *$C^{1,\alpha}$ partial regularity of functions minimising quasi-convex integrals*, Manuscripta Math., 54 (1985), pp. 121-143.

[FH2] ———, *Partial regularity for minimisers of certain functionals having non-quadratic growth*, Ann. Mat. Pura Appl. (IV), 155 (1989), pp. 1-24.

[G] M. GIAQUINTA. *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Anal. Math. Stud., 105, Princeton University Press, Princeton, NJ, 1983.

[GM] M. GIAQUINTA AND G. MODICA, *Partial regularity of minimizers of quasi-convex integrals*, Ann. Inst. H. Poincaré, 3 (1986), pp. 185-208.

[GMS] M. GIAQUINTA, G. MODICA, AND J. SOUČEK, *Cartesian currents, weak diffeomorphisms and nonlinear elasticity*, Arch. Rational Mech. Anal., 106 (1989), pp. 97-159.

[GMS2] ———, *Partial regularity of Cartesian currents which minimize certain variational integrals*, in Partial Differential Equations and the Calculus of Variations (Essays in Honor of Ennio De Giorgi), Birkhäuser, Boston, MA, 1989.

[M] C. B. MORREY, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, 1966.

[Mu] S. MÜLLER, *Weak continuity of determinants and nonlinear elasticity*, C.R. Acad Sci. Paris, 307 (1988), pp. 501-506.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1552-1574, November 1991

004

PERIOD DOUBLING WITH HIGHER-ORDER DEGENERACIES*

BRUCE B. PECKHAM† AND IOANNIS G. KEVREKIDIS‡

Abstract. A family of local diffeomorphisms of $\mathbf{R}^n$ can undergo a period doubling (flip) bifurcation as an eigenvalue of a fixed point passes through -1. This bifurcation is either supercritical or subcritical, depending on the sign of a coefficient determined by higher-order terms. If this coefficient is zero, the resulting bifurcation is "degenerate." The period doubling bifurcation with a single higher-order degeneracy is treated, as well as the more general degenerate period doubling bifurcation where a fixed point has -1 eigenvalue and any number of higher-order degeneracies. The main procedure is a Lyapunov-Schmidt reduction: period-2 orbits are shown to be in one-to-one correspondence with roots of the reduced "bifurcation function," which has $\mathbf{Z}_2$ symmetry. Illustrative examples of the occurrence of the singly degenerate period doubling in the context of periodically forced planar oscillators are also presented.

Key words. period doubling, bifurcation, bifurcation function, Lyapunov-Schmidt, $\mathbf{Z}_2$ symmetry

AMS(MOS) subject classifications. 39, 15

1. Introduction. This paper describes the local bifurcations that take place when we perturb a diffeomorphism $\mathbf{G}_0$ of $\mathbf{R}^n$ which has a fixed point with a single eigenvalue equal to -1. Since $\mathbf{G}_0$ has a nonhyperbolic fixed point, it is necessary to consider higher-order (nonlinear) terms in order to describe the phase portraits near the fixed point of the map $\mathbf{G}_0$, both by itself and also under perturbation in a family $\mathbf{G}_\mu$, $\mu \in \mathbf{R}^k$.

When G_0 is a map of $\mathbf{R}$, any even-order term in its Taylor series expansion can be eliminated by a change of variables. This is a direct result of the normal forms theorem. After eliminating the constant and second-order terms, the linear coefficient will be -1 and the sign of the resulting coefficient of the third-order term will determine whether G_0 will undergo a supercritical or subcritical period doubling (flip) bifurcation [Ar], [GH]. If the third-order coefficient should happen to be zero (a higher-order degeneracy), then the sign of the fifth-order term becomes important. Perturbations of the resulting map $(G_0(x) = -x + cx^5 + o(x^5),\ c \neq 0)$ produce a greater number of topologically distinct phase portraits than do perturbations of the nondegenerate $(G_0(x) = -x + cx^3 + o(x^3),\ c \neq 0)$ map. Two parameters are needed to fully capture all possible phase portraits near the (singly) degenerate map. By the same token, a degenerate bifurcation will generically occur only in families with at least two parrameters.

This discussion naturally extends to multiply degenerate period doubling maps: $G_0(x) = -x + cx^{2k+1} + o(x^{2k+1}),\ c \neq 0$ ($k-1$ times degenerate). These codimension-k bifurcations will generically occur only in families with at least k parameters.

In § 2, we consider the model $k-1$ times degenerate period doublings $f_0(x) = -x + \delta x^{2k+1}$ where $\delta = \pm 1$, and the corresponding model k-parameter unfoldings $f_\varepsilon(x) = -(\varepsilon_1 + 1)x - \varepsilon_2 x^3 - \cdots - \varepsilon_k x^{2k-1} + \delta x^{2k+1}$. We present the mathematical theory in § 3. We show that the period-2 orbits of the individual maps we study are in one-to-one correspondence with the zeros of a "reduced" bifurcation function. This bifurcation function is obtained by using a standard Lyapunov-Schmidt reduction. Because the

* Received by the editors January 2, 1990; accepted for publication (in revised form) January 28, 1991. This work was partially supported by National Science Foundation grants DMS 85-06634, DMS 86-00372-01, and DMS 87-03429, CBT 87-07090, ECS-8717787, and a David and Lucile Packard Foundation Fellowship.

† Department of Mathematics, Boston University, Boston, Massachusetts 02135. Present address, Department of Mathematics and Statistics, University of Minnesota, Duluth, Minnesota 55812.

‡ Department of Chemical Engineering, Princeton University, Princeton, New Jersey, 08544.

topological equivalence of the maps we study is determined by the period-2 orbits and their stability, knowledge of the corresponding bifurcation functions is sufficient to provide us with the topological classification of the original maps. When we consider a *family* of maps, the possible behaviors of the bifurcation functions are given by standard singularity theory. We need only interpret the singularity theory results in the bifurcation context of the original family of maps. In particular, we show that each family in the class of period doubling bifurcations that we treat is "equivalent" to one of the model families we describe in § 2.

The singly degenerate period doubling has a special significance in two-parameter families of maps such as those generated by periodically forced oscillators, which possess period-q "resonance horns" whose boundaries typically consist of saddlenode bifurcations for the qth iterate of the map. We and other researchers [KAS], [MSA], [P1], [P2], [P3], [SDCM], [VR] have repeatedly observed such a degenerate period doubling bifurcation on the boundaries of period-2 resonance horns. In § 4 we describe two models of periodically operated chemical reactors (a chemostat with simple predator-prey kinetics, and a continuous stirred tank reactor (CSTR) with a single irreversible exothermic reaction) where this bifurcation occurs.

The bifurcation diagrams we obtained for our degenerate period doublings turned out to be virtually the same as those for a Hopf bifurcation with higher-order degeneracies for a *flow* [GS], [Ta]. Consequently, work on the Hopf bifurcation suggested approaches to the period doubling problem. Our analysis in its final form parallels that of Golubitsky and Schaeffer [GS, Chap. VIII]. In particular, the use of the Lyapunov–Schmidt reduction to obtain a bifurcation function, as well as the unreduced function with which to start, was suggested by their exposition. Using the reduction on a "finite sequence space," however, appears to be a new idea in this paper. (We have since found out that Vanderbauwhede [Va] and Brown and Roberts [BR] have independently started using the Lyapunov–Schmidt reduction on finite sequence spaces in current research as well.) See also the bibliography in [GS] for the original references using the Lyapunov–Schmidt reduction and singularity theory to study the Hopf bifurcation for flows.

The Hopf problem for flows and our problem are analogous because both can be reduced to finding roots of the same $\mathbf{Z}_2$-symmetric bifurcation function. The period doubling problem, interestingly, turns out to be significantly easier to handle than the Hopf bifurcation. Many of the issues that [GS] had to treat simply did not appear in the period doubling analysis. Consequently, we are able to obtain slightly stronger stability information from the bifurcation than was obtained for the Hopf bifurcation in [GS]. We discuss the comparison with the Hopf bifurcation further in § 5.

To place our work in context, we provide Table 1, showing model unfoldings for bifurcations with higher-order degeneracies. The unfoldings in the table are not always exactly as in the corresponding reference, and the references are not intended to be complete. In all cases, $\boldsymbol{\varepsilon} \in \mathbf{R}^k$ is the unfolding parameter of the codimension-k bifurcation; $\delta = \pm 1$.

The most widely known higher-order degeneracy in Table 1 is the saddle-node (for either the flow or map) with a single higher-order degeneracy, commonly called the *cusp* bifurcation. The map and flow cases are exactly analogous. We will encounter saddlenode bifurcations with higher-order degeneracies in this paper for period-2 orbits, because they appear in the unfoldings of period doubling points with more than one higher-order degeneracy. Higher-order degeneracies in the Hopf bifurcation for maps, however, are much more complicated to treat than degeneracies in the Hopf bifurcation for flows. The map case includes not only all the subtleties of the flow

TABLE 1

Flows: Name	Vector field	Unfolding	References
Saddlenode	$x' = \delta x^{k+1}$	$x' = \varepsilon_1 + \varepsilon_2 x + \cdots + \varepsilon_k x^{k-1} + \delta x^{k+1}$	[Ar], [GH]
Hopf	$r' = \delta r^{2k+1}$ $\theta' = \omega + r^2$	$r' = \varepsilon_1 r + \varepsilon_2 r^3 + \cdots + \varepsilon_{2k} r^{2k-1} + \delta r^{2k+1}$ $\theta' = \omega + r^2$	[Ar], [GH], [GS], [Ta]
Maps: Name	**Map**	**Unfolding**	**References**
Saddlenode	$x \to x + \delta x^{k+1}$	$x \to \varepsilon_1 + (\varepsilon_2 + 1)x + \cdots + \varepsilon_{k-1} x^{k-1} + \delta x^{k+1}$	[Ar], [GH]
Hopf	$r \to \delta r^{2k+1}$ $\theta \to \theta + \omega + r^2$	$r \to \varepsilon_r r + \varepsilon_2 r^3 + \cdots + \varepsilon_{2k} r^{2k-1} + \delta r^{2k+1} + \text{h.o.t.}$ $\theta \to \theta + \omega + r^2 + \text{h.o.t.}$	[Ch]
Period Dblg	$x \to -x + \delta x^{2k+1}$	$x \to -(\varepsilon_1 + 1)x - \cdots - \varepsilon_k x^{2k-1} + \delta x^{2k+1}$	this paper

case, but also some monumental additional problems caused by resonant interaction of periodic orbits, and the existence of invariant sets other than equilibria and closed orbits. Chenciner [Ch] has performed much work on this problem. Note that the higher-order terms must appear, even in the model unfoldings.

We point out that [HW] provides a short description of the period doubling with a single higher-order degeneracy ($k = 2$ in Table 1). That model, but not the theorems in this paper, is relatively well known to bifurcation researchers.

2. The model period doubling families. This section is devoted to describing the bifurcations that take place in the specific families we use as our models. The new results, including the justification for choosing these particular families as models, are given in § 3. The interested reader may skip directly to that section, if desired. We do, however, make some effort in this section to prepare the groundwork for the techniques of § 3. In particular, we use the zeros of several "bifurcation functions" to help us describe the topological classification of our model families. These bifurcation functions will turn out to be special cases of the more general bifurcation functions obtained from the more general maps treated in § 3. (See Corollary 3.13.)

Recall that for maps of $\mathbf{R}$ having a fixed point with a -1 eigenvalue, the normal forms theorem [Ar], [GH] allows us to eliminate any even-order term by a change of variable. Thus the absence of even-order terms from our models should seem reasonable. Keep in mind that, because we are describing local bifurcations, we are only interested in the germs of our functions in phase × parameter space. The base point of all our model germs is the origin of $\mathbf{R} \times \mathbf{R}^k$.

DEFINITION 2.1. The local (near $(x, \boldsymbol{\varepsilon}) = (0, \mathbf{0})$) family

$$f_{\boldsymbol{\varepsilon};k,\delta}(x) := -(\varepsilon_1 + 1)x - \varepsilon_2 x^3 - \cdots - \varepsilon_k x^{2k-1} + \delta x^{2k+1}, \qquad \delta = \pm 1 \tag{2.2}$$

is called the *model local period doubling bifurcation family with $k-1$ higher-order degeneracies.* The map $f_{\mathbf{0};k,\delta}(x) = -x + \delta x^{2k+1}$ (for x near zero) is called the *model period doubling bifurcation map with $k-1$ higher-order degeneracies.*

Note that the parameter $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_k)$ is in $\mathbf{R}^k$, $k \geqq 1$. We will often drop the subscripts k and δ since their values are assumed to be fixed for a given family.

2.1. Individual maps: Stability of periodic orbits. We first describe the behavior of the map $f_{\boldsymbol{\varepsilon}} = f_{\boldsymbol{\varepsilon};k,\delta}$ for fixed values of the parameters. Zero is the unique fixed point near $x = 0$ of $f_{\boldsymbol{\varepsilon}}$ for any $\boldsymbol{\varepsilon}$ near $\mathbf{0}$. The fixed point is attracting for $\varepsilon_1 < 0$ and repelling for $\varepsilon_1 > 0$. Since $f_{\boldsymbol{\varepsilon}}^2$ is an orientation preserving diffeomorphism of $\mathbf{R}$, fixed points are

the only form of recurrence it can have. Once these fixed points of f_ε^2 have been located, the topological equivalence class of f_ε^2, and therefore that of the orientation reversing f_ε, is determined by the directions in which iterates of f_ε^2 progress in the intervals of $\mathbf{R}\backslash\{\text{fixed points of } f_\varepsilon^2\}$. (This can be proven by fundamental interval arguments.) For nonzero x, it is apparent that if $f_\varepsilon^2(x)-x=0$, then x is on a period-2 orbit for f_ε; if $f_\varepsilon^2(x)-x>0$, the orbit of x increases under iteration of f_ε^2; if $f_\varepsilon^2(x)-x<0$, the orbit of x decreases under iteration of f_ε^2. Since the behavior of f_ε is completely determined by the roots and sign of the function $f_\varepsilon^2(x)-x$, we call it a "bifurcation function" associated with the family $f_\varepsilon=f_{\varepsilon;k,\delta}$ in (2.2).

Even in the nondegenerate case ($k=1$ in $f_{\varepsilon;k,\delta}$), the second iterate f_ε^2 is somewhat cumbersome to handle. The algebra is greatly reduced by noticing that f_ε is an odd, or $\mathbf{Z}_2$-symmetric, function of x: $f_\varepsilon(-x)=-f_\varepsilon(x)$. The consequence is that x is a period-2 orbit if and only if $f_\varepsilon(x)=-x$. Thus $f_\varepsilon^2(x)-x=0$ is equivalent to $-f_\varepsilon(x)-x=0$. We have chosen to use $-f_\varepsilon(x)-x$ instead of $f_\varepsilon(x)+x$ because when they do not equal zero, $\operatorname{sgn}(f_\varepsilon^2(x)-x)=\operatorname{sgn}(-f_\varepsilon(x)-x)$. Thus the function $-x-f_\varepsilon(x)$ is also a bifurcation function for (2.2).

Furthermore, $-x-f_\varepsilon(x)=xP_\varepsilon(x^2)$, where

$$P_{\varepsilon;k,\delta}(u):=P_\varepsilon(u):=\varepsilon_1+\varepsilon_2 u+\cdots+\varepsilon_k u^{k-1}-\delta u^k. \tag{2.3}$$

Since $x=0$ is always a fixed point, the roots of $P_\varepsilon(x^2)$ with $x\neq 0$ are precisely the period-2 points. That is, each positive root r^2 of $P_\varepsilon(u)$ corresponds to the period-2 orbit $r\leftrightarrow -r$. For $x\neq 0$ the sign of $-f_\varepsilon(x)-x$ and therefore the sign of $f_\varepsilon^2(x)-x$ is determined by the sign of $P_\varepsilon(x^2)$. So the stability of the fixed point and any period-2 orbit is also determined by the sign of P_ε. Thus, $P_\varepsilon(u)$ becomes our third and simplest bifurcation function.

It may help to keep in mind Fig. 1a, where we graph the three bifurcation functions $f_\varepsilon^2(x)-x$, $-f_\varepsilon(x)-x$, and $P_\varepsilon(x^2)$ for a specific example: $(\varepsilon;k,\delta)=((.000016,.0024,.09),3,+1)$. Figure 1b shows the phase portrait for f_ε^2 that Fig. 1a determines.

2.2. Bifurcations. We are now ready to analyze the bifurcation sets in phase × parameter space ($\mathbf{R}\times\mathbf{R}^k$) that exist in the model families $f_{\varepsilon;k,\delta}$ for fixed values of k and δ. These consist of the nonhyperbolic fixed and period-2 points, possibly with higher-order degeneracies.

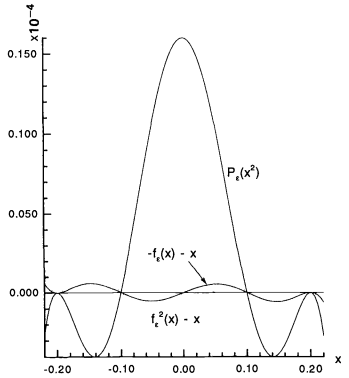


FIG. 1a. *Three bifurcation functions.*

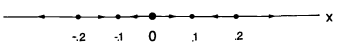


FIG. 1b. *The "flow" of* f_ε^2.

We will treat the fixed-point bifurcations first. Since f_ε is an orientation reversing diffeomorphism of $\mathbf{R}$, the only potential bifurcations for the unique fixed point zero are period doublings. From (2.2), the set in $\mathbf{R}\times\mathbf{R}^k$ of fixed points, which we call D^0, is $\{x=0\}$; the set of period doubling bifurcations is $D^1 := \{x=\varepsilon_1=0\}$; more generally, the set of period doubling bifurcations with at least $i-1$ higher-order degeneracies is apparently (look ahead to Definition 3.1—the model families in (2.2) are already in normal form on the center manifold) the codimension $i+1$ (dimension $k-i$) hyperplane given by

$$D^i := D^i_k = \{(x,\varepsilon)\in\mathbf{R}\times\mathbf{R}^k : x=\varepsilon_1=\cdots=\varepsilon_i=0\}, \qquad i=0,\cdots,k. \tag{2.4}$$

The superscripts have been chosen to indicate the codimension of the corresponding set when projected to the k-dimensional parameter space. The set of simple (nondegenerate) period doubling bifurcation parameters is thus, as usual, a codimension-1 set in the parameter space.

The nonhyperbolic period-2 points are treated by considering f^2_ε. Since f^2_ε is an orientation preserving diffeomorphism of $\mathbf{R}$, the only potential bifurcations for the period-2 orbits are saddlenodes, possibly with higher-order degeneracies. By definition [Ar], [GH], a map $g:\mathbf{R}\to\mathbf{R}$ has a saddlenode with $i-1$ higher-order degeneracies at y_0 if $g(y)-y$ has a zero of multiplicity $i+1$ at $y=y_0$. So the period-2 points in our models have saddlenode bifurcations with $i-1$ higher-order degeneracies at x_0 if, for a fixed value of ε, $f^2_\varepsilon(x)-x$ has a zero of multiplicity $i+1$ at $x=x_0\neq 0$. But $f^2_\varepsilon(x)-x$ having a zero of multiplicity $i+1$ at $x=x_0\neq 0$ is equivalent to $P_\varepsilon(x^2)$ having a zero of multiplicity $i+1$ at $x=x_0\neq 0$. If we define $S^0_{k,\delta}$ as the set of period-2 points and $S^j_{k,\delta}$ as the set of period-2 saddlenode points with at least $j-1$ higher-order degeneracies for $1\leqq j\leqq k-1$, then these sets are

$$S^j := S^j_{k,\delta} = \{(x,\varepsilon)\in\mathbf{R}\times\mathbf{R}^k : P^{(i)}_\varepsilon(x^2)=0 \text{ for } 0\leqq i\leqq j,\ x\neq 0\}. \tag{2.5}$$

2.3. The low codimension period doublings. We can now use (2.4), (2.5), and the sign of P_ε to determine the bifurcation diagrams and phase portraits for the codimension-k bifurcations with $k=1,2,3$.

***k* = 1.** When $k=1$ then $\varepsilon=\varepsilon_1$ and (2.2) becomes the *simple* (nondegenerate) *period doubling bifurcation*: $f_{\varepsilon;1,\delta}(x)=-(\varepsilon_1+1)x+\delta x^3$. $P_\varepsilon(u)=\varepsilon_1-\delta u$ and $P'_\varepsilon(u)=-\delta\neq 0$. From (2.5) we see that period-2 points exist whenever $\delta\varepsilon>0$ and are located on the parabola $x=\pm\sqrt{\delta\varepsilon}$. The period-2 orbits are stable for $\delta=+1$ and unstable for $\delta=-1$. Since $P'_\varepsilon(u)\neq 0$ all period-2 points are hyperbolic. A bifurcation diagram with three representative phase portraits for $\delta=+1$ is shown in Fig. 2. This is the *supercritical* case. The arrows on these phase portraits indicate the direction of travel of second iterates of f_ε. The same figure can be used for $\delta=-1$, the *subcritical* case, by reversing the direction of the ε-axis and the direction of the arrows on the phase portraits. Changing the arrow directions means that the stability of the fixed point and any period-2 orbits for $\delta=-1$ will be the reverse of the stability for $\delta=+1$.

***k* = 2.** In this case, which really motivated the whole paper, (2.2) represents the *singly degenerate period doubling bifurcation* $f_{\varepsilon;2,\delta}(x)=-(\varepsilon_1+1)x-\varepsilon_2x^3+\delta x^5$. Since the coefficient ε_2 of the x^3 term, which determines the criticality of the simple period doubling bifurcation, is allowed to change from positive to negative, we will have both supercritical and subcritical period doublings. All the fixed-point bifurcations have already been identified in (2.4). For the period-2 bifurcations, we use the bifurcation function $P_{\varepsilon;2,\delta}(u)=P_\varepsilon(u)=\varepsilon_1+\varepsilon_2u-\delta u^2$, so $P'_\varepsilon(u)=\varepsilon_2-2\delta u$ and $P''_\varepsilon(u)=-2\delta\neq 0$.

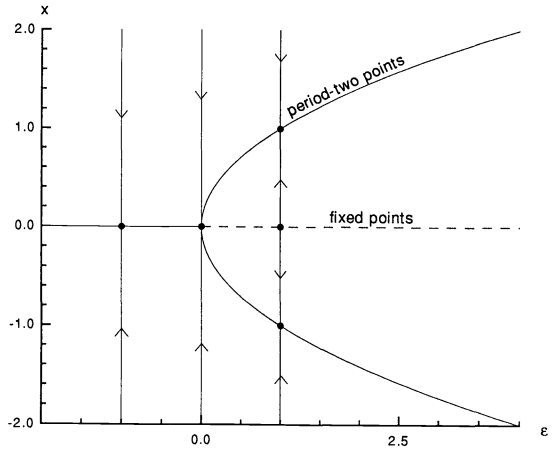


FIG. 2. *Supercritical simple period doubling.*

By (2.5), the period-2 points in $\mathbf{R}\times\mathbf{R}^k$ are $S^0=\{\varepsilon_1=-\varepsilon_2x^2+\delta x^4,\ \varepsilon_2\neq 2\delta x^2\}$ and they project to $\pi_\varepsilon(S^0)=\{\delta\varepsilon_1>0\}\cup\{\delta\varepsilon_2>0$ and $\varepsilon_2^2\geqq -4\delta\varepsilon_1\}$ in the ε-parameter plane $\mathbf{R}^2$. The nonhyperbolic period-2 points are all (nondegenerate) saddlenode bifurcations. They are given by $S^1=\{\varepsilon_1=-\delta x^4,\ \varepsilon_2=2\delta x^2,\ x\neq 0\}$ and project to $\pi_\varepsilon(S^1)=\{\varepsilon_1=(-\delta/4)\varepsilon_2^2,\ \delta\varepsilon_2>0\}$. The formulas for the projections to the ε parameter plane are obtained by eliminating x from the expressions for S^0 and S^1.

Figure 3 shows sketches of the above sets for $\delta=+1$ in phase $\times$ parameter space. The projections to parameter space are drawn on the fixed-point plane $\{x=0\}$. The surface $S^0_{2,+1}$ of period-2 points, the plane D_2^0 of fixed points, the period doubling line D_2^1, the saddlenode curve $S^1_{2,+1}$, and its projection $\pi_\varepsilon(S^1_{2,+1})$ to the ε parameter plane, drawn in the $\{x=0\}$ plane, are all indicated in the figure. Note that all the bifurcation points occur on the "folds" of the period-2 surface $S^0_{2,+1}$.

Various two-dimensional bifurcation diagrams (pieces of Fig. 3) are shown in Fig. 4: 4a gives the projection of the bifurcation sets S^1 (saddlenodes) and D^1 (period doublings) to the parameter space; the other three are representative one-parameter cuts of Fig. 3: 4b and 4c each have a fixed value for ε_2, while a small circular path

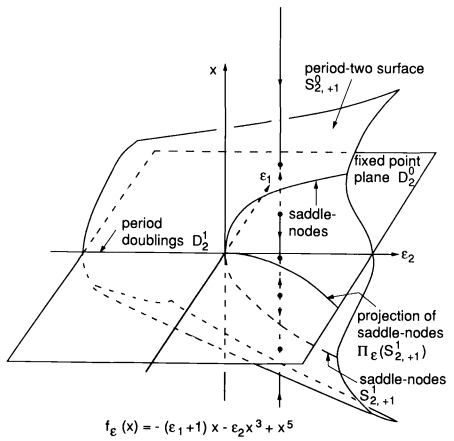


FIG. 3. *Singly degenerate period doubling.*

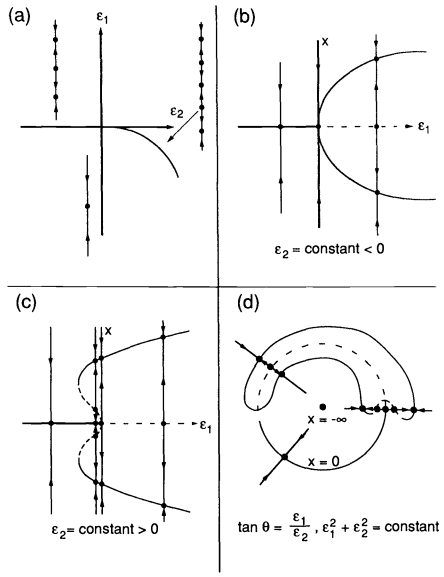


FIG. 4. *Aspects of singly degenerate period doubling.*

around the origin in the ε-plane yields 4d. Arrows all indicate the "flow" of the second iterate of $f_{\varepsilon;2,+1}$. (Compare Figs. 4b and 4c with Fig. 3.1 in [GS, p. 260]; compare Fig. 4d with Fig. 136 in [Ar, p. 283].)

As in the simple period doubling case, Fig. 3 and all Fig. 4 diagrams could be converted from the $\delta = +1$ case to the $\delta = -1$ case by reversing the directions of the ε_1 axis, the ε_2 axis, and the "flow" lines. The stability of the fixed point and all period-2 orbits is opposite for the two cases.

$k \geqq 3$. The program for computing the bifurcation submanifolds can obviously be continued for the model period doublings of any codimension. Because the computations are more lengthy but not much more enlightening, we merely list the results, with special attention to the $(k, \delta) = (3, +1)$ case.

The fixed-point bifurcation sets satisfy $D_k^0 \supset D_k^1 \supset \cdots \supset D_k^{k-1} \supset D_k^k$ where $D_k^{j-1} \backslash D_k^j$ is the codimension-j manifold in $\mathbf{R} \times \mathbf{R}^k$ of period doubling points with exactly $j-2$ higher-order degeneracies. Similarly, the period-2 bifurcation sets satisfy $S_{k,\delta}^0 \supset S_{k,\delta}^1 \supset \cdots \supset S_{k,\delta}^{k-2} \supset S_{k,\delta}^{k-1}$ where $S_{k,\delta}^{j-1} \backslash S_{k,\delta}^j$ is the codimension-j manifold in $\mathbf{R} \times \mathbf{R}^k$ of period-2 saddlenodes with exactly $j-2$ higher-order degeneracies. The set $S_{k,\delta}^{k-1}$ and its projection to parameter space have the explicit parametric representations

$$S_{k,\delta}^{k-1} = \left\{ (x, [\varepsilon_1, \cdots, \varepsilon_k]) = \left(x, -\delta \left[(-1)^k \binom{k}{k} x^{2k}, \cdots, -\binom{k}{3} x^6, \binom{k}{2} x^4, -\binom{k}{1} x^2 \right] \right), x \neq 0 \right\}, \tag{2.6}$$

$$\pi_\varepsilon(S_{k,\delta}^{k-1}) = \left\{ (\varepsilon_1, \cdots, \varepsilon_k) = \left((-1)^{k+1} \frac{\delta \binom{k}{k}}{k^k} \varepsilon_k^{2k-2}, \cdots, \frac{\delta \binom{k}{3}}{k^3} \varepsilon_k^4, \frac{-\delta \binom{k}{2}}{k^2} \varepsilon_k^2, \varepsilon_k \right), \delta \varepsilon_k > 0 \right\}. \tag{2.7}$$

When $k=3$ we obtain

$$S^0_{3,\delta}=\{(x,\varepsilon_1,\varepsilon_2,\varepsilon_3)=(x,-\varepsilon_2x^2-\varepsilon_3x^4+\delta x^6,\varepsilon_2,\varepsilon_3),\ x\neq 0\},$$
$$S^1_{3,\delta}=\{(x,\varepsilon_1,\varepsilon_2,\varepsilon_3)=(x,\varepsilon_3x^4-2\delta x^6,-2\varepsilon_3x^2+3\delta x^4,\varepsilon_3),\ x\neq 0\},$$
$$S^2_{3,\delta}=\{(x,\varepsilon_1,\varepsilon_2,\varepsilon_3)=(x,\delta x^6,-3\delta x^4,3\delta x^2),\ x\neq 0\}$$

(cf. (2.6)). $\pi_\varepsilon(S^0_{3,+1})$, the set of all parameter values with period-2 orbits, is described below.

$$\pi_\varepsilon(S^1_{3,+1})=\left\{(\varepsilon_1,\varepsilon_2,\varepsilon_3)=(\varepsilon_3[W_-(\varepsilon_2,\varepsilon_3)]^2+2[W_-(\varepsilon_2,\varepsilon_3)]^3,\varepsilon_2,\varepsilon_3),\ \varepsilon_3<0,\ 0<\varepsilon_2\leqq\frac{\varepsilon_3^2}{3}\right\}$$
$$\cup\left\{(\varepsilon_1,\varepsilon_2,\varepsilon_3)=(\varepsilon_3[W_+(\varepsilon_2,\varepsilon_3)]^2+2[W_+(\varepsilon_2,\varepsilon_3)]^3,\varepsilon_2,\varepsilon_3),\ \varepsilon_3\geqq 0,\ \varepsilon_2<0 \text{ or } \varepsilon_3<0,\ \varepsilon_2\leqq\frac{\varepsilon_3^2}{3}\right\},$$

where

$$W_\pm(\varepsilon_2,\varepsilon_3):=\frac{-\varepsilon_3\pm\sqrt{\varepsilon_3^3-3\varepsilon_2}}{3};$$
$$\pi_\varepsilon(S^2_{3,+1})=\left\{(\varepsilon_1,\varepsilon_2,\varepsilon_3)=\left(\frac{\varepsilon_3^4}{27},\frac{-\varepsilon_3^2}{3},\varepsilon_3\right),\ \varepsilon_3>0\right\}$$

(cf. (2.7)).

Because the full phase × parameter space is now four-dimensional, the best pictures we can draw are either "slices" or projections of the four-dimensional space. Figure 5 shows the slice corresponding to $\varepsilon_3\equiv$ constant <0. Note the appearance of the cusp point on the curve of saddlenodes, so named for its location on the projection of the saddlenode curve to the parameter plane. Such a point appears only for $k\geqq 3$. The slice corresponding to $\varepsilon_3\equiv$ constant >0 we do not show, because it is qualitatively the same as Fig. 3.

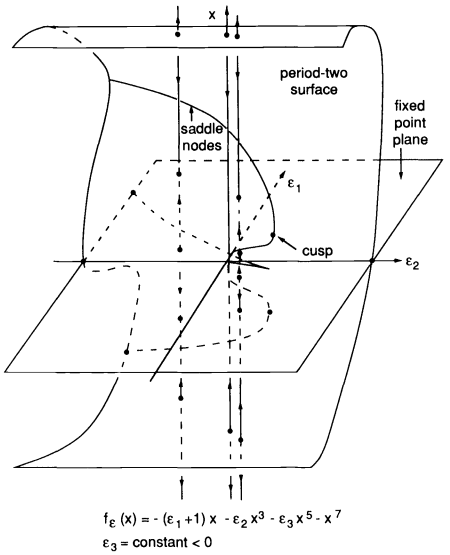


FIG. 5. *Doubly degenerate period doubling.*

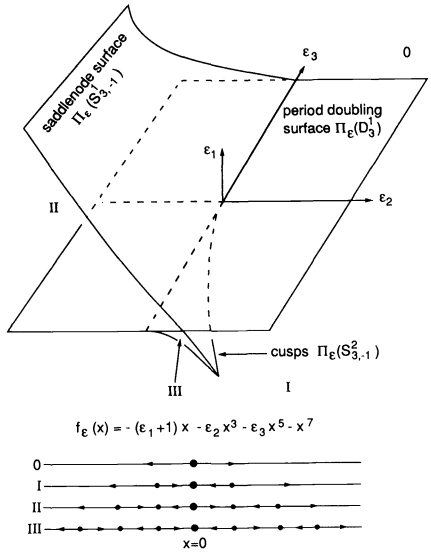


FIG. 6. *Doubly degenerate period doubling: parameter space.*

Figure 6 shows the saddlenode surfaces $\pi_\varepsilon(S^1_{3,+1})$, the cusp curve $\pi_\varepsilon(S^2_{3,+1})$ (where the two saddlenode surfaces, one defined with W_+ and the other with W_-, meet), and the period doubling plane $\pi_\varepsilon(D^1_3)$ in the three-dimensional parameter space. The set $\pi_\varepsilon(S^0_{3,+1})$ is bounded "above" in Fig. 6 by the higher of the saddlenode surfaces (inclusive) and the period doubling plane (not inclusive). Compare our Fig. 6 with Fig. 6 in [Ta]. Note that the plane $\{x=0, \varepsilon_3 = \text{const} < 0\}$ appears in both Fig. 5, as the fixed-point plane, and in Fig. 6, as the leading edge of the graph.

3. General period doubling families. In § 2 we analyzed the local topological behavior of the special families of diffeomorphisms of $\mathbf{R}: f_{\varepsilon;k,\delta}(x) = -(\varepsilon_1+1)x - \varepsilon_2 x^3 - \cdots - \varepsilon_k x^{2k-1} + \delta x^{2k+1}$. We now treat the more general case of a local family of diffeomorphisms of $\mathbf{R}^n$.

DEFINITION 3.1. Fix $k \geqq 1$. Let $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu}) = \mathbf{G}_{\boldsymbol{\mu}}(\mathbf{x})$ be a representative of the germ of a C^{2k+1} function satisfying

(1) $\mathbf{G}: U \to \mathbf{R}^n$, U is a neighborhood of $(\mathbf{x}_0, \boldsymbol{\mu}_0)$ in $\mathbf{R}^n \times \mathbf{R}^m$.

(2) $\mathbf{G}(\mathbf{x}_0, \boldsymbol{\mu}_0) = \mathbf{x}_0$.

(3) $D_{\mathbf{x}}\mathbf{G}(\mathbf{x}_0, \boldsymbol{\mu}_0)$ has a single eigenvalue of -1 and no other eigenvalues on the unit circle.

(4) On its one-dimensional center manifold, the map $\mathbf{G}_{\boldsymbol{\mu}_0}$ can be transformed by a C^{2k+1} change of coordinates to a C^{2k+1} map of the form $y \to -y + cy^{2k+1} + o(y^{2k+1})$, $c \neq 0$.

Then $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$ is a *local period doubling bifurcation family with* $k-1$ *higher-order degeneracies*, and $\mathbf{G}_{\boldsymbol{\mu}_0}$ is *a local period doubling bifurcation map with* $k-1$ *higher-order degeneracies.*

The main goal of this section is to establish Theorem 3.15, where we show that on its center manifold, every k-parameter period doubling bifurcation family with $k-1$ higher-order degeneracies is, at least generically, the "same" as one of the model families $f_{\varepsilon;k,\delta}$, where $\delta = \text{sign}(c)$. The main technical tools for Theorem 3.15 are the existence of a "$\mathbf{Z}_2$-symmetric bifurcation function" related to the original period doubling family (Theorem 3.3) and the universal unfolding theorem from $\mathbf{Z}_2$-singularity theory (Lemma 3.21). We are then able to compare $\mathbf{G}_{\boldsymbol{\mu}}$ to the appropriate model family via their respective bifurcation functions.

The Lyapunov-Schmidt reduction. Let $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$ be a period doubling family with any number of higher-order degeneracies. For simplicity, we will assume $(\mathbf{x}_0, \boldsymbol{\mu}_0) = (\mathbf{0}, \mathbf{0})$. As with our special functions $f_{\epsilon;k,\delta}$ in (2.2), the implicit function theorem guarantees that $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$ has a unique fixed point near $\mathbf{x} = \mathbf{0}$ for each $\boldsymbol{\mu}$ near zero. Having only one phase variable (along with the m parameters) on the center manifold implies that the only other local recurrence can be in the form of period-2 points [CMY].

The period-2 points (including the fixed point) of $\mathbf{G}$ are characterized by the roots of the function $\boldsymbol{\Phi}: \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n \times \mathbf{R}^n$ defined by

$$\boldsymbol{\Phi}(\mathbf{x}, \mathbf{y}, \boldsymbol{\mu}) := \boldsymbol{\Phi}_{\boldsymbol{\mu}}(\mathbf{x}, \mathbf{y}) := (\mathbf{y} - \mathbf{G}(\mathbf{x}, \boldsymbol{\mu}), \mathbf{x} - \mathbf{G}(\mathbf{y}, \boldsymbol{\mu})). \tag{3.2}$$

The reason this function turns out to be more useful than $\mathbf{G}_{\boldsymbol{\mu}}^2(\mathbf{x}) - \mathbf{x}$ is twofold: $\boldsymbol{\Phi}$ deals only with first iterates of $\mathbf{G}_{\boldsymbol{\mu}}$, and it has an obvious symmetry that will be quite useful. Specifically, $\boldsymbol{\Phi}_{\boldsymbol{\mu}} \mathfrak{R} = \mathfrak{R} \boldsymbol{\Phi}_{\boldsymbol{\mu}}$, where $\mathfrak{R}$ is the reflection that interchanges the variables $\mathbf{x}$ and $\mathbf{y}$ in both the domain and range of $\boldsymbol{\Phi}$. That is, $\boldsymbol{\Phi}_{\boldsymbol{\mu}} \mathfrak{R}(\mathbf{x}, \mathbf{y}) = \boldsymbol{\Phi}_{\boldsymbol{\mu}}(\mathbf{y}, \mathbf{x}) = (\mathbf{x} - \mathbf{G}(\mathbf{y}, \boldsymbol{\mu}), \mathbf{y} - \mathbf{G}(\mathbf{x}, \boldsymbol{\mu})) = \mathfrak{R}(\mathbf{y} - \mathbf{G}(\mathbf{x}, \boldsymbol{\mu}), \mathbf{x} - \mathbf{G}(\mathbf{y}, \boldsymbol{\mu})) = \mathfrak{R} \boldsymbol{\Phi}_{\boldsymbol{\mu}}(\mathbf{x}, \mathbf{y})$.

We now perform the Lyapunov-Schmidt reduction [GS, § I.3] on $\boldsymbol{\Phi}$ to get the following theorem. Although the theorem is stated for C^ρ functions, we will be interested mainly in the case $\rho = \infty$.

THEOREM 3.3. *Let $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$ be a C^ρ, $2k+1 \leqq \rho \leqq \infty$, local period doubling bifurcation family with $k-1$ higher-order degeneracies as in Definition* 3.1, *with $(\mathbf{x}_0, \boldsymbol{\mu}_0) = (\mathbf{0}, \mathbf{0})$. Define $\boldsymbol{\Phi}(\mathbf{x}, \mathbf{y}, \boldsymbol{\mu})$ by* (3.2). *Then there exists a C^ρ* bifurcation function *$b: \mathbf{R} \times \mathbf{R}^m \to \mathbf{R}$ of the form $b(s, \boldsymbol{\mu}) = sB(u, \boldsymbol{\mu})$, $u := s^2$, such that solutions of $\boldsymbol{\Phi}(\mathbf{x}, \mathbf{y}, \boldsymbol{\mu}) = \mathbf{0}$ for $(\mathbf{x}, \mathbf{y}, \boldsymbol{\mu})$ near $(\mathbf{0}, \mathbf{0}, \mathbf{0})$ are in one to one correspondence with solutions of $b(s, \boldsymbol{\mu}) = 0$ for $(s, \boldsymbol{\mu})$ near $(0, \mathbf{0})$.*

Proof. The Lyapunov-Schmidt reduction to prove Theorem 3.3 is standard [GS, § 1.3], but we include most of the computations since we will be interested in the specific bifurcation function we get via the reduction, as well as some of the intermediate functions defined in the proof.

Case 1: $x \in \mathbf{R}$. In standard coordinates, the linearization of $\boldsymbol{\Phi}_{\mathbf{0}}$ at $(0, 0)$ is $L := D_{x,y} \boldsymbol{\Phi}_0(0, 0) = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. Thus the kernel of L, $\ker L = \langle (1, -1) \rangle$, and range $L = \langle (1, 1) \rangle$. Note that $\mathbf{R}^2 = \ker L \oplus \text{range } L$ so that $E(x, y) := ((x+y)/\sqrt{2}, (x+y)/\sqrt{2})$ is the projection onto range L, and $(I - E)(x, y) := ((x-y)/\sqrt{2}, -(x-y)/\sqrt{2})$ is the projection onto $\ker L$. The equation $\boldsymbol{\Phi}(\mathbf{x}, \mathbf{y}, \boldsymbol{\mu}) = \mathbf{0}$, which we wish to solve, is equivalent to the two equations (with the $\sqrt{2}$ factor introduced for convenience):

$$\sqrt{2}\, E \boldsymbol{\Phi}(x, y, \boldsymbol{\mu}) = (0, 0), \tag{3.4a}$$

$$\sqrt{2}(I - E) \boldsymbol{\Phi}(x, y, \boldsymbol{\mu}) = (0, 0). \tag{3.4b}$$

These two equations are more conveniently expressed in coordinates with respect to the splitting $\mathbf{R}^2 = \ker L \oplus \text{range } L$. Formally, this can be defined by the change of coordinates from (x, y) with respect to the standard basis on $\mathbf{R}^2$ to (s, r) with respect to the new basis which we choose as $\{(1, -1), (1, 1)\}$. The coordinates are related by $x(1, 0) + y(0, 1) = s(1, -1) + r(1, 1)$, or $x = s + r$ and $y = r - s$. Since the s component of the new coordinate version of (3.4a) is automatically satisfied by definition of E, as is the r component of the new coordinate version of (3.4b), the two vector equations in (3.4) are equivalent to the two scalar equations

$$Q(s, r, \boldsymbol{\mu}) := \tfrac{1}{2}\{2r - G(s + r, \boldsymbol{\mu}) - G(-s + r, \boldsymbol{\mu})\} = 0, \tag{3.5a}$$

$$\tfrac{1}{2}\{-2s - G(s + r, \boldsymbol{\mu}) + G(-s + r, \boldsymbol{\mu})\} = 0. \tag{3.5b}$$

Equation (3.5a) is the r component of (3.4a); equation (3.5b) is the s component of (3.4b).

Since $Q(0,0,\mathbf{0})=0$ and $(\partial Q/\partial r)(0,0,\mathbf{0})=2\neq 0$, then the implicit function theorem implies that there exists a unique C^ρ function $R(s,\boldsymbol{\mu})$ satisfying $R(0,\mathbf{0})=0$ and $Q(s,R(s,\boldsymbol{\mu}),\boldsymbol{\mu})=0$ for $(s,\boldsymbol{\mu})$ near $(0,\mathbf{0})$. Plugging this new function $R(s,\boldsymbol{\mu})$ into the left-hand side of (3.5b), we get our reduced bifurcation function $b(s,\boldsymbol{\mu})$:

$$\begin{aligned} b(s,\boldsymbol{\mu}) &:= \tfrac{1}{2}\{-2s-G(s+R(s,\boldsymbol{\mu}),\boldsymbol{\mu})+G(-s+R(s,\boldsymbol{\mu}),\boldsymbol{\mu})\} \\ &= -G(s+R(s,\boldsymbol{\mu}),\boldsymbol{\mu})+(-s+R(s,\boldsymbol{\mu})). \end{aligned} \tag{3.6}$$

The latter form is obtained by substituting $Q(s,R(s,\boldsymbol{\mu}),\boldsymbol{\mu})=0$ from (3.5a) into the first line of (3.6).

It can be verified directly that $R(-s,\boldsymbol{\mu})=R(s,\boldsymbol{\mu})$, and therefore that $b(-s,\boldsymbol{\mu})=-b(s,\boldsymbol{\mu})$, but this is really a consequence of the equivariance of the original function $\boldsymbol{\Phi}$ with respect to the reflection $\mathfrak{R}$. This is because $b(s,\boldsymbol{\mu})$ is really the coordinate representation of a map from $\ker\boldsymbol{\Phi}\times\mathbf{R}^k$ to $\ker\boldsymbol{\Phi}$, and $\mathfrak{R}$ acts on $\ker\boldsymbol{\Phi}$ by $\mathfrak{R}(s(1,-1))=\mathfrak{R}(s,-s)=(-s,s)=-s(1,-1)$.

That $b(s,\boldsymbol{\mu})$ has the form $sB(s^2,\boldsymbol{\mu})$ is immediate from the odd symmetry of $b(s,\boldsymbol{\mu})$. The one-to-one correspondence between solutions of $b(s,\boldsymbol{\mu})=0$ and $\boldsymbol{\Phi}(x,y,\boldsymbol{\mu})=0$ is

$$(s,\boldsymbol{\mu})\leftrightarrow(s+R(s,\boldsymbol{\mu}),-s+R(s,\boldsymbol{\mu}),\boldsymbol{\mu}). \tag{3.7}$$

Note that if $s\neq 0$, solutions s and $-s$ correspond to the same period-2 orbit, but these are distinct solutions for $\boldsymbol{\Phi}:\boldsymbol{\Phi}(x,y,\boldsymbol{\mu})=0$ and $\boldsymbol{\Phi}(y,x,\boldsymbol{\mu})=0$. This completes the proof for $x\in\mathbf{R}$.

Case 2: $\mathbf{x}\in\mathbf{R}^n$, $n>1$, and the coordinates $\mathbf{x}=(x_1,\cdots,x_n)$ have been chosen with respect to the basis $\{\mathbf{e}_i\}_{i=1}^n$ so that matrix of $D_{\mathbf{x}}\mathbf{G}(\mathbf{0},\mathbf{0})$ has the block form $\hat{B}=\left(\begin{smallmatrix}-1 & 0\\ 0 & B\end{smallmatrix}\right)$, where B is an $(n-1)\times(n-1)$ matrix. The $2n\times 2n$ matrix of the linearization of $L=D_{\mathbf{x},\mathbf{y}}\boldsymbol{\Phi}_{\mathbf{0}}(\mathbf{0},\mathbf{0})$ with respect to the induced basis $\{\mathbf{f}_1,\cdots,\mathbf{f}_{2n}\}=\{(\mathbf{e}_1,\mathbf{0}),\cdots,(\mathbf{e}_n,\mathbf{0}),(\mathbf{0},\mathbf{e}_1),\cdots,(\mathbf{0},\mathbf{e}_n)\}$ becomes $L=\left(\begin{smallmatrix}-\hat{B} & I\\ I & -B\end{smallmatrix}\right)$. The first and $(n+1)$st rows of L are identical, but using the fact that no other eigenvalues are on the unit circle, it can be shown that the remaining rows are independent. (This would be easier to see if $\{\mathbf{e}_i\}$ were a basis putting $\hat{B}$ into Jordan canonical form.) So we still have the dimension of $\ker L=1$. In fact, $\ker L=\langle\mathbf{f}_1-\mathbf{f}_{n+1}\rangle$ and range $L=\langle\mathbf{f}_1+\mathbf{f}_{n+1},\mathbf{f}_2,\cdots,\mathbf{f}_n,\mathbf{f}_{n+2},\cdots,\mathbf{f}_{2n}\rangle$. We also still have $\mathbf{R}^{2n}=\ker L\oplus\text{range } L$. The coordinates with respect to this splitting are s on $\ker L$, and $(r,x_2,\cdots,x_n,y_2,\cdots,y_n)$ on range L, where $x_1=s+r$ and $y_1=r-s$, and $\mathbf{x}=(x_1,\cdots,x_n)$ and $\mathbf{y}=(y_1,\cdots,y_n)$ are in coordinates with respect to $\{e_i\}_{i=1}^n$. Solving $\sqrt{2}\mathbf{E}\boldsymbol{\Phi}(\mathbf{x},\mathbf{y},\boldsymbol{\mu})=0$ in the new coordinates is equivalent to $\mathbf{Q}=\mathbf{0}$, where

$$\begin{aligned} &\mathbf{Q}(s,r,x_2,\cdots,x_n,y_2,\cdots,y_n,\boldsymbol{\mu}) \\ &\quad := (\tfrac{1}{2}(2r-G_1(s+r,x_2,\cdots,x_n,\boldsymbol{\mu})-G_1(-s+r,y_2,\cdots,y_n,\boldsymbol{\mu})), \\ &\qquad y_2-G_2(s+r,x_2,\cdots,x_n,\boldsymbol{\mu}),\cdots,y_n-G_n(s+r,x_2,\cdots,x_n,\boldsymbol{\mu}), \\ &\qquad x_2-G_2(-s+r,y_2,\cdots,y_n,\boldsymbol{\mu}),\cdots,x_n-G_n(-s+r,y_2,\cdots,y_n,\boldsymbol{\mu}))=\mathbf{0}. \end{aligned} \tag{3.8}$$

This equation can be solved uniquely by the implicit function theorem for C^ρ functions $r,x_2,\cdots,x_n,y_2,\cdots,y_n$, all in terms of s and $\boldsymbol{\mu}$ in a neighborhood of $(s,\boldsymbol{\mu})=(0,\mathbf{0})$. We shall call these solutions $R(s,\boldsymbol{\mu})$, $X_j(s,\boldsymbol{\mu})$, and $Y_j(s,\boldsymbol{\mu})$. That is,

$$\mathbf{Q}(s,\mathbf{W}(s,\boldsymbol{\mu}),\boldsymbol{\mu})=\mathbf{0}, \tag{3.9}$$

where $\mathbf{W}(s,\boldsymbol{\mu}):=(R(s,\boldsymbol{\mu}),X_2(s,\boldsymbol{\mu}),\cdots,X_n(s,\boldsymbol{\mu}),Y_2(s,\boldsymbol{\mu}),\cdots,Y_n(s,\boldsymbol{\mu}))$. Differentiation of (3.9) with respect to s and using the block form of $D_{\mathbf{x}}\mathbf{G}(\mathbf{0},\mathbf{0})$ yields

$$\frac{\partial\mathbf{W}}{\partial s}=\mathbf{0}. \tag{3.10}$$

As for any Lyapunov-Schmidt reduction involving a symmetry, the symmetry $\mathfrak{R}$ is inherited as $\mathbf{W}\mathfrak{R}(s,\boldsymbol{\mu}) = \mathfrak{R}\mathbf{W}(s,\boldsymbol{\mu})$, interpreted as

$$(R(-s,\boldsymbol{\mu}), X_2(-s,\boldsymbol{\mu}), \cdots, X_n(-s,\boldsymbol{\mu}), Y_2(-s,\boldsymbol{\mu}), \cdots, Y_n(-s,\boldsymbol{\mu}))$$
$$= (R(s,\boldsymbol{\mu}), Y_2(s,\boldsymbol{\mu}), \cdots, Y_n(s,\boldsymbol{\mu}), X_2(s,\boldsymbol{\mu}), \cdots, X_n(s,\boldsymbol{\mu})).$$

Thus $Y_j(s,\boldsymbol{\mu}) = X_j(-s,\boldsymbol{\mu}), j = 2, \cdots, n$, and $R(s,\boldsymbol{\mu}) = R(-s,\boldsymbol{\mu})$. The bifurcation function, analogous to (3.6), is

$$\begin{aligned} b(s,\boldsymbol{\mu}) = \tfrac{1}{2}\{&-2s - G_1(s + R(s,\boldsymbol{\mu}), X_2(s,\boldsymbol{\mu}), \cdots, X_n(s,\boldsymbol{\mu}), \boldsymbol{\mu}) \\ &+ G_1(-s + R(s,\boldsymbol{\mu}), X_2(-s,\boldsymbol{\mu}), \cdots, X_n(-s,\boldsymbol{\mu}), \boldsymbol{\mu})\}, \end{aligned} \tag{3.11}$$

where $\mathbf{G} = (G_1, \cdots, G_n)$. It is clear from the first line, since $R(s,\boldsymbol{\mu}) = R(-s,\boldsymbol{\mu})$, that we still have our $\mathbf{Z}_2$-symmetric bifurcation function: $b(-s,\boldsymbol{\mu}) = -b(s,\boldsymbol{\mu})$. So $b(s,\boldsymbol{\mu})$ is still of the form $sB(s^2,\boldsymbol{\mu})$. The one-to-one correspondence between roots of $b(s,\boldsymbol{\mu})$ and $\boldsymbol{\Phi}(\mathbf{x},\mathbf{y},\boldsymbol{\mu})$, analogous to (3.7), is given by $(s,\boldsymbol{\mu}) \leftrightarrow (\mathbf{X}(s,\boldsymbol{\mu}), \mathbf{Y}(s,\boldsymbol{\mu}), \boldsymbol{\mu})$, where

$$\begin{aligned} &\mathbf{X}(s,\boldsymbol{\mu}) := (s + R(s,\boldsymbol{\mu}), X_2(s,\boldsymbol{\mu}), \cdots, X_n(s,\boldsymbol{\mu})), \\ &\mathbf{Y}(s,\boldsymbol{\mu}) := \mathbf{X}(-s,\boldsymbol{\mu}) = (-s + R(s,\boldsymbol{\mu}), X_2(-s,\boldsymbol{\mu}), \cdots, X_n(-s,\boldsymbol{\mu})). \end{aligned} \tag{3.12}$$

Thus the theorem is true for $\mathbf{x} \in \mathbf{R}^n$, with the assumed coordinate system.

Case 3: $\mathbf{x} \in \mathbf{R}^n$, $n > 1$. Change this general case into the special coordinate form of Case 2 by a linear change of variable. Then follow the procedure outlined in that case. □

We now prove two corollaries that give some insight into the mechanics of the Lyapunov-Schmidt reduction of Theorem 3.3.

COROLLARY 3.13. *For our model families* $f_{\boldsymbol{\varepsilon};k,\delta}(x) = -(\varepsilon_1 + 1)x - \varepsilon_2 x^3 - \cdots - \varepsilon_k x^{2k-1} + \delta x^{2k+1}$, *the bifurcation function* $b_f(s,\boldsymbol{\varepsilon}) = sB_f(s^2,\boldsymbol{\varepsilon}) = -f_\varepsilon(s) - s$. *Also,* $B_f(u,\boldsymbol{\varepsilon}) = P_{\boldsymbol{\varepsilon};k,\delta}(u)$, *where* $P_{\boldsymbol{\varepsilon};k,\delta}(u)$ *is as defined in* (2.3).

Proof. Because our model families $f_{\boldsymbol{\varepsilon};k,\delta}$ are odd, it is apparent from (3.5a) by letting $G(x,\boldsymbol{\varepsilon}) = f_{\boldsymbol{\varepsilon};k,\delta}(x)$ for any fixed values of k and δ that $Q(s,0,\boldsymbol{\varepsilon}) = 0$, so $R(s,\boldsymbol{\varepsilon}) = 0$ must be the unique solution to $Q(s,R(s,\boldsymbol{\varepsilon}),\boldsymbol{\varepsilon}) = 0$. Thus, from (3.6), $b(s,\boldsymbol{\varepsilon}) = sB_f(s^2,\boldsymbol{\varepsilon})$ becomes $-s - f_\varepsilon(s)$. But $= -s - f_\varepsilon(s) = \varepsilon_1 s + \varepsilon_2 s^3 + \cdots + \varepsilon_k s^{2k-1} - \delta s^{2k+1} = sP_{\boldsymbol{\varepsilon};k,\delta}(s^2)$. So $B_f(u,\boldsymbol{\varepsilon}) = P_{\boldsymbol{\varepsilon};k,\delta}(u)$. □

So the seemingly ad hoc method we used in § 2 to analyze our model families turns out to be merely a special case of the more general Lyapunov-Schmidt reduction.

COROLLARY 3.14. *Let* $\mathbf{G}(\mathbf{x},\boldsymbol{\mu})$ *be a local period doubling family with* $k-1$ *higher-order degeneracies at* $(\mathbf{x},\boldsymbol{\mu}) = (\mathbf{0},\mathbf{0}) \in \mathbf{R}^n \times \mathbf{R}^m$. *If* $\{(\mathbf{x},\boldsymbol{\mu}): x_2 = \cdots = x_n = 0\}$ *is the center manifold (instead of just the center eigenspace) of* $\mathbf{G}(\mathbf{x},\boldsymbol{\mu})$, *then*

(A) *the functions* $X_j(s,\boldsymbol{\mu})$ *and* $Y_j(s,\boldsymbol{\mu})$, *defined in* (3.12), *are zero for* $j = 2, 3, \cdots, n$,

(B) *the bifurcation function of* $[\mathbf{G}]$ = *the bifurcation function of* [$\mathbf{G}$ *restricted to its* $1+m$ *dimensional center manifold*].

Proof. (A). We can show there exists a solution to (3.8) with $x_j = y_j = 0$, $j = 2, \cdots, n$. By uniqueness of solutions, the functions $X_j(s,\boldsymbol{\mu})$ and $Y_j(s,\boldsymbol{\mu})$ *must* be zero for $j \geqq 2$.

(B) This follows from (A) by directly computing the two bifurcation functions using (3.8) and (3.11). □

Corollaries 3.13 and 3.14 suggest that using the Lyapunov-Schmidt reduction to obtain the bifurcation function $b_{\mathbf{G}}(s,\boldsymbol{\mu})$ should be compared to the more topological alternative of obtaining a bifurcation function $-\tilde{f}_{\boldsymbol{\mu}}(s) - s$ from $\mathbf{G}(\mathbf{x},\boldsymbol{\mu})$ by the following steps:

(1) Restrict $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$ to its $1+m$ dimensional center mainfold: $f(x_1, \boldsymbol{\mu}) := f_{\boldsymbol{\mu}}(x_1) := G_1((x_1, \mathbf{H}(x_1, \boldsymbol{\mu})), \boldsymbol{\mu})$, where the center manifold is the graph of $\mathbf{H}: \mathbf{R} \times \mathbf{R}^m \to \mathbf{R}^{n-1}$.

(2) Put the resulting function into its normal form $\tilde{f}(s, \boldsymbol{\mu}) := \tilde{f}_{\boldsymbol{\mu}}(s) := h_{\boldsymbol{\mu}} \circ f_{\boldsymbol{\mu}} \circ h_{\boldsymbol{\mu}}^{-1}(s)$, where $h(x_1, \boldsymbol{\mu}) := h_{\boldsymbol{\mu}}(x_1)$ are the coordinate changes to put $f_{\boldsymbol{\mu}}$ into its normal form $\tilde{f}_{\boldsymbol{\mu}}$.

(3) Use the resulting odd symmetry to replace the bifurcation function $\tilde{f}_{\boldsymbol{\mu}}^2(s) - s$ with the simpler function $-\tilde{f}_{\boldsymbol{\mu}}(s) - s$.

Besides being a single step, the Lyapunov-Schmidt reduction has another major advantage over the center manifold/normal forms technique. Although the normal forms theorem guarantees a polynomial change of coordinates to put $f_{\boldsymbol{\mu}}(x_1)$ into its normal form up to any finite order, the existence of a coordinate change to eliminate *all* even-order terms in x_1 is not guaranteed. Thus step (2) above may not even be possible. On the other hand, if we put the function $f_{\boldsymbol{\mu}}(x_1)$ into its normal form only up to some finite order, step (3) would not be possible because the resulting function would be odd only up to that finite order. Note also that the original function $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$ being C^∞ does not imply that its center manifold realization is C^∞. The Lyapunov-Schmidt bifurcation function $b_{\mathbf{G}}(s, \boldsymbol{\mu})$, however, is C^∞.

Universality of the model families. We now use Theorem 3.3 and some standard results from singularity theory to show that the model unfoldings we considered in Chapter 2 are "universal unfoldings." More specifically, we prove that, when restricted to a center manifold, any map in a local period doubling family is topologically equivalent to one of the model family maps. If certain nondegeneracy conditions are satisfied, the whole family of center manifold maps will be "equivalent" to one of the model families. Our notion of equivalence is embodied in the statement of the theorem.

We use the following notation. Let $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$ be any C^∞ period doubling family with $k-1$ higher-order degeneracies at $(\mathbf{0}, \mathbf{0})$. Let $b_{\mathbf{G}}(s, \boldsymbol{\mu}) = sB_{\mathbf{G}}(s^2, \boldsymbol{\mu})$ be a bifurcation function obtained from $\mathbf{G}$ as in Theorem 3.3. Assume x_1 is a coordinate along the eigenspace corresponding to the -1 eigenvalue for the fixed point $\mathbf{x} = \mathbf{0}$ for $\boldsymbol{\mu} = \mathbf{0}$. Let $g_{\boldsymbol{\mu}}(x_1) := g(x_1, \boldsymbol{\mu})$ be the realization of $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$ on its $1+m$ dimensional center manifold. By Definition 3.1, the center manifold map in normal form up to order $2k+1$ for $\boldsymbol{\mu} = \mathbf{0}$ is $y \to -y + cy^{2k+1} + o(y^{2k+1})$. Let $f_{\boldsymbol{\varepsilon}}(z) := f_{\boldsymbol{\varepsilon};k,\delta}(z)$ be the model family $-(\varepsilon_1 + 1)z - \varepsilon_2 z^3 - \cdots - \varepsilon_k z^{2k-1} + \delta z^{2k+1}$, where $\delta = -\operatorname{sgn}(c)$. Recall the definitions of the bifurcation sets D_k^i and $S_{k,\delta}^i$ in (2.4) and (2.5) for the model families $f_{\boldsymbol{\varepsilon};k,\delta}$. We now analogously define the bifurcation sets for $\mathbf{G}$.

$D_g^0 = \{(x_1, \boldsymbol{\mu}) \in \mathbf{R} \times \mathbf{R}^m : x_1$ is a fixed point for $g_{\boldsymbol{\mu}}\}$,

$D_g^i = \{(x_1, \boldsymbol{\mu}) \in \mathbf{R} \times \mathbf{R}^m : x_1$ is a fixed point for $g_{\boldsymbol{\mu}}$ with eigenvalue -1 and at least $i-1$ higher-order degeneracies$\}$ for $i \geqq 1$,

$S_g^0 = \{(x_1, \boldsymbol{\mu}) \in \mathbf{R} \times \mathbf{R}^m : x_1$ is a period-2 point for $g_{\boldsymbol{\mu}}\}$,

$S_g^i = \{(x_1, \boldsymbol{\mu}) \in \mathbf{R} \times \mathbf{R}^m : x_1$ is a period-2 point for $g_{\boldsymbol{\mu}}$ with eigenvalue 1 and at least $i-1$ higher-order degeneracies$\}$ for $i \geqq 1$.

THEOREM 3.15. *Let $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$ be a C^∞ period doubling family with $k-1$ higher-order degeneracies. Define its center manifold representation $g_{\boldsymbol{\mu}}(x_1)$ and the model family $f_{\boldsymbol{\varepsilon}}(z)$ as in the above paragraph. Assume the "eigenvalue crossing condition:" $\nabla_{\boldsymbol{\mu}} \lambda(\boldsymbol{\mu}) \neq \mathbf{0}$, where $\lambda(\boldsymbol{\mu})$ is the eigenvalue of the unique fixed point of $g_{\boldsymbol{\mu}}$. Then*

(a) *There exists a neighborhood N of $(0, \mathbf{0})$ in $\mathbf{R} \times \mathbf{R}^m$ and a C^∞ function $\boldsymbol{\Psi}: N \to \{\mathbf{R} \times \mathbf{R}^k\} : (x_1, \boldsymbol{\mu}) \to (z, \boldsymbol{\varepsilon})$ of the form $\boldsymbol{\Psi}(x_1, \boldsymbol{\mu}) = (Z_{\boldsymbol{\mu}}(x_1), \psi(\boldsymbol{\mu}))$ with the following properties:*

(1) $\boldsymbol{\Psi}: (0, \mathbf{0}) \to (0, \mathbf{0})$.

(2) *For each fixed parameter value $\boldsymbol{\mu}$, $g_{\boldsymbol{\mu}}(x_1)$ restricted to the neighborhood N and $f_{\psi(\boldsymbol{\mu})}(z)$ restricted to $\boldsymbol{\Psi}(N)$ are topologically conjugate to each other.*

(3) Ψ *maps fixed points, period-2 points, and bifurcation manifolds of g to fixed points, period-2 points, and corresponding bifurcation manifolds of f, respectively.* (*That is,* $\Psi: D_g^i \to D_k^i$ *for* $i=0,\cdots,k$, *and* $S_g^i \to S_{k,\delta}^i$ *for* $i=0,\cdots,k-1$.)

(b) *Let k and* δ *be fixed. Any family that can replace* $f_{\varepsilon:k,\delta}$ *in Theorem* 3.15(a) *must have at least k parameters.* (*This justifies calling the period doubling bifurcation with* $k-1$ *higher-order degeneracies a codimension-k bifurcation.*)

(c) *If* $\mu \in \mathbf{R}^k$ *and*

$$\left\{ \nabla_\mu \left(\frac{\partial^i B_G(u,\mu)}{\partial u^i} \right) \Big|_{(0,0)}, \qquad i=0,\cdots,k-1 \right\}$$

is independent, then ψ *and* Ψ *are* C^∞ *diffeomorphisms.*

Before beginning the proof of this theorem, we make the following comments:

(1) Recall that in the proof of Theorem 3.3, $b_G(s,\mu)$ and therefore $B_G(s,\mu)$ were defined using the implicit function theorem. Although this means the bifurcation functions and their derivatives are not usually computable, their values at $(s,\mu)=(0,\mathbf{0})$ are computable. (See, for example, Lemma 3.16.) Consequently, the nondegeneracy conditions in part (c) of the theorem *are* computable.

(2) The nondegeneracy conditions in part (c) will generically be true. Thus, for a generic k-parameter family of maps, Ψ will be a diffeomorphism. Since the C^∞ diffeomorphism Ψ preserves the bifurcation sets, and the bifurcation sets for the models are analytic, this is what guarantees that the bifurcation manifolds will all be C^∞ and that the pictures obtained from applications (see § 4) all "look like" the bifurcation pictures obtained from the model families in § 2. In particular, the orders of tangency of corresponding bifurcation manifolds will be the same as in the model families. In the codimension-2 case, with only one higher-order degeneracy, the projection to the parameter space of the bifurcation manifolds will always (generically) show a curve of saddlenodes for the second iterate of the map being tangent to a period doubling curve where it terminates. (Look ahead to Figs. 7–9 in comparison to the model family bifurcation diagrams in Figs. 3 and 4.)

(3) Note that the center eigenspace coordinate x_1 can be replaced by any phase space coordinate not perpendicular to x_1 by a one-dimensional linear change of coordinates independent of the parameter. Consequently, any generic phase variable coordinate can be used in place of a center eigenspace coordinate x_1 in drawing the bifurcation sets. This is exactly what was done to obtain Figs. 7–9.

(4) This is a technical comment comparing our notion of "equivalence" implied by the existence in the theorem of the function Ψ to the oft-used notion of "topological conjugacy." Recall that $g(x_1,\mu)$ and $f(z,\varepsilon)$ are (locally) topologically conjugate families if there exists a local homeomorphism $\Phi(x_1,\mu)=(h_\mu(x_1),\phi(\mu))$ such that $g_\mu = h_\mu^{-1} \circ f_{\psi(\mu)} \circ h_\mu$. If the individual topological conjugacies $h_\mu(x_1)$ do not necessarily vary continuously with respect to the parameter μ, then the families are said to be "mildly topologically conjugate" [NPT]. Because Theorem 3.15 guarantees that $g_\mu(x_1)$ and $f_{\psi(\mu)}(z)$ will be topologically conjugate to each other for each fixed value of μ, our equivalence implies the two families $g(x_1,\mu)$ and $f(z,\varepsilon)$ are at least mildly topologically conjugate (by letting $\phi=\psi$) as long as the parameter space map $\psi(\mu)$ is a homeomorphism.

We point out that although the conjugacies $h_\mu(x_1)$ and the functions $Z_\mu(x_1)$ of the theorem are not the same, they are related. Specifically, they will agree on all the bifurcation sets D_g^i and S_g^i. This includes the fixed and period-2 sets. Thus, when restricted to the bifurcation sets, $h_\mu(x_1)$ will not only vary continuously with respect to the parameter μ, but will also be C^∞.

Consequently, when the parameter space map $\psi(\mu)$ is a diffeomorphism, the existence of the function Ψ of Theorem 3.15 is a stronger property than mild topological conjugacy but not comparable to topological conjugacy. Topological conjugacies have the stronger property that the individual conjugacies $h_\mu(x_1)$ should vary continuously with the parameter; our equivalence has the stronger property that the function Ψ is a (C^∞) diffeomorphism, and consequently that the individual conjugacies $h_\mu(x_1)$ *restricted to the bifurcation surfaces* are also diffeomorphisms.

The rest of this section is devoted to the proof of Theorem 3.15. We begin with the following lemmas.

LEMMA 3.16. *If* $x \in \mathbf{R}$ *and* $c \neq 0$, *then* $G(x, \mathbf{0}) = -x + cx^{2k+1} + o(x^{2k+1})$ *implies* $b_G(s, \mathbf{0}) = -cs^{2k+1} + o(x^{2k+1})$.

Proof. We differentiate the definition of $b_G(s, \mu)$ in (3.6), using the derivatives of $R(s, \mu)$ at $(0, \mathbf{0})$, which we obtain from (3.5a) by repeated implicit differentiation. Since R is even in s, we immediately know that $(\partial^j R/\partial s^j)(0, \mathbf{0}) = 0$ for odd j. We also know from the proof of Theorem 3.3 that $R(0, \mathbf{0}) = 0$. It is relatively straightforward to show that the implicit differentiation yields

$$\frac{\partial^2 \mathrm{R}}{\partial \mathrm{s}^2}(0, \mathbf{0}) = \frac{1}{2}\frac{\partial^2 G}{\partial x^2}(0, \mathbf{0}),$$

$$\frac{\partial^4 R}{\partial s^4}(0, \mathbf{0}) = \frac{1}{2}\frac{\partial^4 G}{\partial x^4}(0, \mathbf{0}) + \frac{3}{4}\frac{\partial^2 G}{\partial x^2}(0, \mathbf{0})\left\{2\frac{\partial^3 G}{\partial x^3}(0, \mathbf{0}) + \frac{1}{2}\left[\frac{\partial^2 G}{\partial x^2}(0, \mathbf{0})\right]^2\right\}.$$

In general,

$$\frac{\partial^k R}{\partial s^k}(0, \mathbf{0}) = \frac{1}{2}\frac{\partial^k G}{\partial x^k}(0, \mathbf{0}) + \cdots,$$

where the omitted terms all have factors of $(\partial^j G/\partial x^j)(0, \mathbf{0})$ with $2 \leqq j \leqq k-1$.

Using these derivatives, and the fact that $b_G(s, \mu)$ is odd in r (so that all even derivatives of b_G with respect to r vanish), we obtain

$$\frac{\partial b_G}{\partial r}(0, \mathbf{0}) = 0,$$

$$\frac{\partial^3 b_G}{\partial r^3}(0, \mathbf{0}) = -\frac{\partial^3 G}{\partial x^3}(0, \mathbf{0}) - \frac{3}{2}\left\{\frac{\partial^2 G}{\partial x^2}(0, \mathbf{0})\right\}^2, \tag{3.17}$$

$$\begin{aligned}\frac{\partial^5 b_G}{\partial r^5}(0, \mathbf{0}) = &-\frac{\partial^5 G}{\partial x^5}(0, \mathbf{0}) - 5\frac{\partial^4 G}{\partial x^4}(0, \mathbf{0})\frac{\partial^2 G}{\partial x^2}(0, \mathbf{0}) - \frac{15}{4}\frac{\partial^3 G}{\partial x^3}(0, \mathbf{0})\left(\frac{\partial^2 G}{\partial x^2}(0, \mathbf{0})\right)^2 \\ &-5\left[\frac{1}{2}\frac{\partial^4 G}{\partial x^4}(0, \mathbf{0}) + \frac{3}{4}\frac{\partial^2 G}{\partial x^2}(0, \mathbf{0})\left\{2\frac{\partial^3 G}{\partial x^3}(0, \mathbf{0}) + \frac{1}{2}\left(\frac{\partial^2 G}{\partial x^2}(0, \mathbf{0})\right)^2\right\}\right].\end{aligned} \tag{3.18}$$

The expressions for the seventh-order derivative are not pretty. In general, however, we have the relation

$$\frac{\partial^k b_G}{\partial r^k}(0, \mathbf{0}) = -\frac{\partial^k G}{\partial x^k}(0, \mathbf{0}) + \cdots,$$

where the omitted terms all have factors of $(\partial^j G/\partial x^j)(0, \mathbf{0})$ with $2 \leqq j \leqq k-1$.

The lemma follows immediately. □

Note. The sign of (3.17) determines the criticality of the nondegenerate period doubling bifurcation. If it is negative, the bifurcation is supercritical; if it is positive, the bifurcation is subcritical; if it equals zero, there is at least one higher-order

degeneracy. If both (3.17) and (3.18) are zero, there are at least two higher-order degeneracies.

LEMMA 3.19. *Let the C^∞ period doubling family $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$, its center manifold realization $g_{\boldsymbol{\mu}}(x_1)$, and the bifurcation function $b_{\mathbf{G}}(s, \boldsymbol{\mu})$ be as in the paragraph preceding Theorem 3.15. Then there exists a neighborhood N of $(0, \mathbf{0})$ in $\mathbf{R}\times\mathbf{R}^m$ such that for $(s, \boldsymbol{\mu}) \in N$, $g^2_{\boldsymbol{\mu}}(x_1) - x_1$ has the same sign as $b_{\mathbf{G}}(s, \boldsymbol{\mu})$, where $(s, \boldsymbol{\mu})$ and $(x_1, \boldsymbol{\mu})$ are related by the C^∞ diffeomorphism $(x_1, \boldsymbol{\mu}) = (s + R(s, \boldsymbol{\mu}), \boldsymbol{\mu})$ (as in (3.12)).*

Furthermore, for each fixed $\boldsymbol{\mu}$, the multiplicity of the corresponding zeros of $g^2_{\boldsymbol{\mu}}(x_1) - x_1$ and $b_{\mathbf{G}}(s, \boldsymbol{\mu})$ is the same.

Proof. Theorem 3.3 guarantees that roots of $G^2_{\boldsymbol{\mu}}(\mathbf{x}) - \mathbf{x}$ are in one-to-one correspondence with roots of $b_{\mathbf{G}}(s, \boldsymbol{\mu})$. Since roots of $G^2_{\boldsymbol{\mu}}(\mathbf{x}) - \mathbf{x}$ must be on the center manifold of $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$, the roots of $g^2_{\boldsymbol{\mu}}(x_1) - x_1$ must also be in one-to-one correspondence with roots of $\mathbf{G}^2_{\boldsymbol{\mu}}(\mathbf{x}) - \mathbf{x}$, and therefore with roots of $b_{\mathbf{G}}(s, \boldsymbol{\mu})$. The correspondence is indicated by (3.12) in the proof of Theorem 3.3:

$$(3.20)\quad s \leftrightarrow \mathbf{x} = \mathbf{X}(s, \boldsymbol{\mu}) = (s + R(s, \boldsymbol{\mu}), X_2(s, \boldsymbol{\mu}), \cdots, X_n(s, \boldsymbol{\mu})) \leftrightarrow x_1 = s + R(s, \boldsymbol{\mu}).$$

For each fixed $\boldsymbol{\mu}$, the multiplicities of corresponding roots of $g^2_{\boldsymbol{\mu}}(x_1) - x_1$ and $b_{\mathbf{G}}(s, \boldsymbol{\mu})$ must be the same, because if they are not, then a perturbation of $\mathbf{G}$ could be made so that their roots would not correspond. (It can be shown that an arbitrarily C^∞ small perturbation of $\mathbf{G}(\mathbf{x}, \boldsymbol{\mu})$ can be chosen to perturb $g^2_{\boldsymbol{\mu}}(x_1) - x_1$ or $b_{\mathbf{G}}(s, \boldsymbol{\mu})$ from a zero of multiplicity ρ to a function with ρ distinct real roots.)

We have left only to show that the signs of the two functions are equal. Since for fixed $\boldsymbol{\mu}$ we already have the zeros and their multiplicities corresponding for $g^2_{\boldsymbol{\mu}}(x_1) - x_1$ and $b_{\mathbf{G}}(s, \boldsymbol{\mu})$, and since these two functions are perturbations of $g^2_{\mathbf{0}}(x_1) - x_1$ and $b_{\mathbf{G}}(s, \mathbf{0})$, respectively, the signs will be the same for $g^2_{\boldsymbol{\mu}}(x_1) - x_1$ and $b_{\mathbf{G}}(s, \boldsymbol{\mu})$ if and only if the signs of the leading coefficients of $g^2_{\mathbf{0}}(x_1) - x_1$ and $b_{\mathbf{G}}(s, \mathbf{0})$ are the same.

According to Definition 3.1, if $x \in \mathbf{R}$ then in normal form up to order $2k+1$, $G(x, \mathbf{0}) = -x + cx^{2k+1} + o(x^{2k+1})$, $c \neq 0$. This makes $G^2_{\mathbf{0}}(x) - x = g^2_{\mathbf{0}}(x) - x = -2cx^{2k+1} + o(x^{2k+1})$. Lemma 3.16 implies $b_G(s, \mathbf{0}) = -cs^{2k+1} + o(s^{2k+1})$. If $\boldsymbol{\mu} = \mathbf{0}$, then $s = 0$ corresponds to $x = x_1 = 0 + R(0, \mathbf{0}) = 0$, so the signs of the leading coefficients of $g^2_{\mathbf{0}}(x) - x$ and $b_G(s, \mathbf{0})$ correspond. If $x \in \mathbf{R}$ but $G(x, \mathbf{0}) = g(x, \mathbf{0})$ is not in normal form up to order $2k+1$, a near identity polynomial change of coordinates $x = h(y)$ can put $g_{\boldsymbol{\mu}}(x)$ into this normal form. That is, $\tilde{g}_{\mathbf{0}}(y) := h^{-1}(g_{\mathbf{0}}(h(y)))$ is in normal form up to order $2k+1$. By perturbation arguments as in the second paragraph of this proof, the multiplicity of the zeros of $\tilde{g}^2_{\mathbf{0}}(y) - y$, $g^2_{\mathbf{0}}(x) - x$, $b_g(s, \mathbf{0})$, and $b_{\tilde{g}}(\tilde{s}, \mathbf{0})$ must all be the same. The same logic works along a whole path of coordinate changes from h_t, $t \in [0, 1]$, from the $h_0 :=$ identity to $h_1 := h$. Therefore, by continuity, the sign of the leading coefficient of $\tilde{g}^2_{\mathbf{0}}(y) - y$ and $g^2_{\mathbf{0}}(x) - x$ must be the same, as must be the sign of the leading coefficient of $b_g(s, \mathbf{0})$ and $b_{\tilde{g}}(\tilde{s}, \mathbf{0})$. Since the sign of the leading coefficients of $\tilde{g}^2_{\mathbf{0}}(y) - y$ and $b_{\tilde{g}}(\tilde{s}, \mathbf{0})$ are equal by the previous paragraph, this forces the sign of the leading coefficients of $g^2_{\mathbf{0}}(x) - x$ and $b_g(s, \mathbf{0})$ to be the same.

If $x \in \mathbf{R}^n$ with $n > 1$, then the realization of $\mathbf{G}$ on its center manifold can also be obtained by a near identity change of coordinates. So by a continuity argument similar to that in the paragraph above, the leading coefficient of $g^2_{\mathbf{0}}(x_1) - x_1$ will have the same sign as the leading coefficient of $b_{\mathbf{G}}(s, \mathbf{0})$. □

One consequence of Lemma 3.19 is that the period doubling map with $k-1$ higher-order degeneracies can be alternatively characterized by

$$\left.\frac{\partial^i B_{\mathbf{G}}(u, \boldsymbol{\mu})}{\partial u^i}\right|_{(0,\mathbf{0})} = 0 \quad \text{for } i = 0, \cdots, k-1,$$

but

$$\left.\frac{\partial^k B_{\mathbf{G}}(u,\boldsymbol{\mu})}{\partial u^k}\right|_{(0,\mathbf{0})} \neq 0.$$

Another consequence is that the sign of $b_{\mathbf{G}}$ or $B_{\mathbf{G}}$ can be used to determine stability of the fixed and period-2 orbits of $\mathbf{G}(\mathbf{x},\boldsymbol{\mu})$ and $g(x_1,\boldsymbol{\mu})$. It is usually more practical, however, to do this by eigenvalue computations, especially because, as mentioned after the statement of Lemma 3.19, the bifurcation functions are defined via the implicit function theorem.

Technical note. Lemma 3.19 and Theorem 3.15 are both stated under the assumption that the coordinate x_1 is already a coordinate on the center eigenspace for $\boldsymbol{\mu}=\mathbf{0}$. When $\mathbf{G}(\mathbf{x},\boldsymbol{\mu})$ does not originally come in this form, there is some leeway in choosing x_1. Its choice, however, involves a change of coordinates from the given form of $\mathbf{G}(\mathbf{x},\boldsymbol{\mu})$. If the change of coordinates is orientation preserving, a path to the identity argument as in the last two paragraphs of the proof of Lemma 3.19 can be used to show that the leading coefficient of $g_{\mathbf{0}}^2(x_1)-x_1$ will have the same sign as the leading coefficient of $b_{\mathbf{G}}(s,\mathbf{0})$. The case of an orientation reversing change of coordinates is converted to the orientation preserving case by noting that the change of variables $x_1\to -x_1$ leaves $b_g(s,\boldsymbol{\mu})$ the same and leaves the leading coefficient of $g_{\mathbf{0}}^2(x_1)-x_1$ the same.

This note shows that even though the bifurcation function constructed in the proof of Theorem 3.3 is not necessarily unique (there is a choice of coordinates made in reducing Case 3 to Case 2), the zeros, including multiplicities, and signs at corresponding nonzero points of any two bifurcation functions arising from the same original function must all be equal.

We now recall the universal unfolding theorem for $\mathbf{Z}_2$-symmetric bifurcation functions.

LEMMA 3.21. *Define the k-parameter family of $\mathbf{Z}_2$-symmetric bifurcation functions* $U(S,\boldsymbol{\varepsilon}):=\varepsilon_0 S+\varepsilon_1 S^3+\cdots+\varepsilon_k S^{2k-1}+\delta S^{2k+1}$, $\delta=\pm 1$. *Let $V(s,\boldsymbol{\mu})$ be any family of $\mathbf{Z}_2$-symmetric bifurcation functions satisfying* $V(s,\mathbf{0})=cs^{2k+1}+\cdots$, *with* $\operatorname{sgn}(c)=\delta$, *and*

$$\nabla_{\boldsymbol{\mu}}\left(\frac{\partial^i V(s,\boldsymbol{\mu})}{\partial s^i}\right)\Bigg|_{(0,\mathbf{0})} \neq \mathbf{0}.$$

Then in a neighborhood of $(0,\mathbf{0})$, *there exist C^∞ functions M, Σ, and $\boldsymbol{\phi}$ such that*

$$V(s,\boldsymbol{\mu})=M(s,\boldsymbol{\mu})U(\Sigma(s,\boldsymbol{\mu}),\boldsymbol{\phi}(\boldsymbol{\mu})) \tag{3.22}$$

with

$$M(s,\boldsymbol{\mu})>0,\ (\partial\Sigma/\partial s)(s,\mathbf{0})>0,\ \Sigma(s,\mathbf{0})=0,$$

$$\boldsymbol{\phi}(\mathbf{0})=\mathbf{0},\ M(-s,\boldsymbol{\mu})=M(s,\boldsymbol{\mu}),\ \Sigma(-s,\boldsymbol{\mu})=-\Sigma(s,\boldsymbol{\mu}).$$

Furthermore, there is no family having the properties of $U(S,\boldsymbol{\varepsilon})$ with fewer than k parameters.

Proof. Combine Proposition 2.14 [GS, p. 256] and Proposition 3.4 [GS, p. 259]. □

Proof of Theorem 3.15. (a) Recall from the paragraph preceding the statement of Theorem 3.15 that $g(x_1,\boldsymbol{\mu})$ is the center manifold realization of $\mathbf{G}(\mathbf{x},\boldsymbol{\mu})$ and $f(z,\boldsymbol{\varepsilon})$ is the appropriate model family. We will define the function $\boldsymbol{\Psi}$ so that the sign of $g_{\boldsymbol{\mu}}^2(x_1)-x_1$ will be the same as the sign of $f_{\boldsymbol{\varepsilon}}^2(z)-z$ for $(z,\boldsymbol{\varepsilon})=\boldsymbol{\Psi}(x_1,\boldsymbol{\mu})$. As previously

noted in § 2.1, this will guarantee that g_μ and f_ε will be topologically conjugate to each other for fixed values of the parameters (and appropriately restricted neighborhoods).

Let $b_G(s, \mu)$ and $b_f(S, \varepsilon)$ be the bifurcation functions determined from $\mathbf{G}(\mathbf{x}, \mu)$ and $f(z, \varepsilon)$, respectively, as in the proof of Theorem 3.3. Let $R^G(s, \mu)$ and $R^f(S, \varepsilon)$ be the respective functions defined following (3.9), with the superscripts added to distinguish the R's arising from the different functions $\mathbf{G}$ and f.

By Lemma 3.19, $g_\mu^2(x_1) - x_1$ has the same sign as $b_G(s, \mu)$, where (s, μ) and (x_1, μ) are related by the diffeomorphism $(x_1, \mu) = (s + R^G(s, \mu), \mu)$. Also by Lemma 3.19, $f_\varepsilon^2(z) - z$ has the same sign as $b_f(S, \varepsilon)$, where (S, ε) and (z, ε) are related by the diffeomorphism $(z, \varepsilon) = (S + R^f(S, \varepsilon), \varepsilon) = (S, \varepsilon)$. This last equality follows from the proof of Corollary 3.13, where we showed that $R^f(S, \varepsilon) = 0$.

Also, by Corollary 3.13, $b_f(S, \varepsilon) = \varepsilon_0 S + \varepsilon_1 S^3 + \cdots + \varepsilon_k S^{2k-1} + \delta S^{2k+1}$, which equals $U(S, \varepsilon)$ as defined in Lemma 3.21. Lemma 3.21 can therefore be used to show that there exist functions Σ and ϕ such that $b_G(s, \mu)$ and $b_f(S, \varepsilon)$ have the same sign for $(S, \varepsilon) = (\Sigma(s, \mu), \phi(\mu))$. Note that this C^∞ map will be a diffeomorphism if $\phi(\mu)$ is a diffeomorphism.

Combining the results of the two paragraphs above, we see that the signs of $g_\mu^2(x_1) - x_1$, $b_G(s, \mu)$, $b_f(S, \varepsilon)$ and $f_\varepsilon^2(z) - z$ are all the same for $x_1 = s + R^G(s, \mu)$, $(S, \varepsilon) = (\Sigma(s, \mu), \phi(\mu))$, and $S = z$. These relationships define the map $\Psi(x_1, \mu)$ by the composition

$$(x_1, \mu) \to (s, \mu) \to (S, \varepsilon) \to (z, \varepsilon). \tag{3.23}$$

Each map in the composition is C^∞ in a neighborhood of $(0, \mathbf{0})$ and each fixes $(0, \mathbf{0})$. Therefore the same is true of Ψ. This establishes (a)(1) and (a)(2) of Theorem 3.15. Part (a)(3) is true because each map in (3.23) preserves not only the zeros but also their multiplicities. (This is true for the first and third maps by Lemma 3.19, and for the middle map by (3.22).)

(b) If there existed a family that could replace f_ε in Theorem 3.15(a), then its bifurcation function would be a "universal unfolding" in the space of $\mathbf{Z}_2$ bifurcation functions with fewer than k parameters. This would contradict the last sentence of the universal unfolding theorem for $\mathbf{Z}_2$-symmetric bifurcation functions, Lemma 3.21.

(c) The condition that

$$\left\{ \nabla_\mu \left(\frac{\partial^i B_G(u, \mu)}{\partial u^i} \right) \Big|_{(0,\mathbf{0})}, \qquad i = 0, \cdots, k-1 \right\}$$

be independent is equivalent to the Jacobian determinant $|\partial \varepsilon_i / \partial \mu_j|_{\mu=\mathbf{0}} \neq 0$ *and therefore is equivalent to the map* $\varepsilon = \psi(\mu)$ *being a local diffeomorphism. In this case* Ψ *is also a local diffeomorphism.* □

4. Applications. Theorem 3.15 states that any period doubling diffeomorphism with $k-1$ higher-order degeneracies is equivalent, both in terms of its topological behavior under iteration (restricted to its center manifold) and in terms of its bifurcation sets, to one of our model families of § 2. In order to support these theoretical results, we used a version of the continuation routine AUTO [DK] that we adapted for use with maps to investigate two examples where we knew a period doubling with a higher-order degeneracy to exist. Both are two-parameter families of maps generated by flows of periodically forced planar oscillators. The stroboscopic map and its derivatives were calculated using ODESSA [LK]. Because our applications involved only two parameters, we would not expect to see a period doubling with more than the single higher-order degeneracy. The bifurcation diagrams we produced from these

applications should be compared to Figs. 3 and 4 for our model period doubling map with a single higher-order degeneracy.

4.1. Resonance horns in forced oscillators. Consider a system of two autonomous coupled nonlinear ODE's

$$d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}, p), \qquad \mathbf{f}: \mathbf{R}^2 \times \mathbf{R} \to \mathbf{R}^2,$$

where $p \in \mathbf{R}$ is a parameter. Assume that for $p = p_0$ the system above has an asymptotically attracting closed orbit with frequency ω_0. Consider the two-parameter family of forced oscillators

$$d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}, p_0 + \alpha g(\omega t)),$$

where α and ω are the parameters (α is the amplitude of the forcing and g has period $T = 1/\omega$). A more convenient second parameter is the ratio ω/ω_0 of the forcing to the natural frequency. Taking the time T return map of this flow (sometimes referred to as the stroboscopic map) gives us a two-parameter family of invertible, orientation preserving maps of the plane. The asymptotic attractivity of the limit cycle of the unforced oscillator guarantees the existence of a normally hyperbolic attracting invariant circle for small forcing amplitude α. According to standard circle map theory [Ar], [Ha], we expect resonance horns (also called entrainment regions of Arnol'd tongues) entering the first quadrant of the ω/ω_0-α parameter plane for every rational value of ω/ω_0. The boundaries of the "q/p resonance horn" emanating from $\omega/\omega_0 = q/p$ are saddlenode bifurcation points for the qth iterate of the map. Inside this q/p resonance horn, the corresponding map has at least one (typically two: a stable and unstable pair) period-q orbit. In particular, we are interested in the situation where $q = 2$, when the boundaries of the $2/p$ horns are saddlenode bifurcations for the second iterate of the map. In continuing these saddlenode curves towards higher values of α, we have repeatedly found them to terminate at a degenerate period doubling where they collide with a period doubling curve. (This was a much easier and less expensive ways of locating the degenerate period doubling points than the method suggested by Definition 3.1 or comment 1 following the statement of Theorem 3.15. To compute the normal form of a map on its center manifold and/or $(\partial^i B_G(u, \boldsymbol{\mu})/\partial u^i)|_{(0,\mathbf{0})}$, we would need higher derivatives of the stroboscopic map generated by numerically integrating the forced oscillator flows.)

Figures 7 and 8 show various features of the period doubling with a single higher-order degeneracy in the context of a 2/3 resonance horn for our first system of periodically forced ODEs:

$$\frac{dx_1}{d\tau} = -(p_0 + \alpha \cos(\omega\tau))x_1 + \frac{ax_2}{b + x_2} x_1,$$

$$\frac{dx_2}{d\tau} = -(p_0 + \alpha \cos(\omega\tau))x_2 + \frac{z_f - x_1 - x_2}{1 + z_f - x_1 - x_2} x_2 - \frac{ax_2}{b + x_2} x_1.$$

These ODEs model a predator-prey system (protozoa preying on bacteria in a chemostat). Here x_1 is the dimensionless concentration of protozoa, x_2 is the dimensionless concentration of bacteria, and z_f is the dimensionless feed concentration of a substrate on which the bacteria grow with Monod-type kinetics [PK]. The parameter we vary periodically is the flow rate of the chemostat. The autonomous system for $a = 0.4$, $b = 2.8125$, $z_f = 12.4$, and $p_0 = 0.2$ has a single attracting limit cycle of period $T = 18.999$ units of dimensionless time τ.

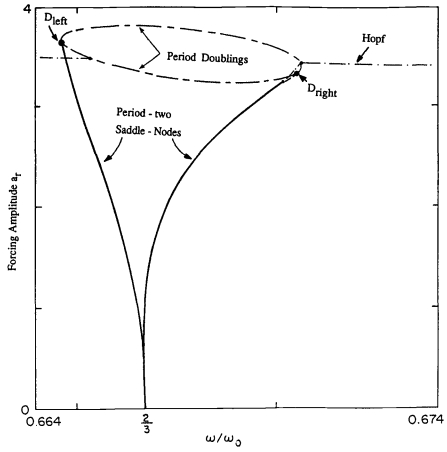


FIG. 7. *Predator-prey parameter plane.*

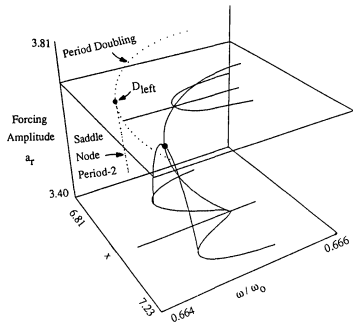


FIG. 8. *Predator-prey: singly degenerate period doubling.*

Figure 7 shows the boundaries of the 2/3 resonance horn for this model ($a_r = \alpha/0.00265$). As we follow both sides of the horn boundary towards higher values of α we encounter degenerate period doubling points D_{left} and D_{right}. Figure 8 is a three-dimensional representation of the full four-dimensional phase × parameter space of the solution surface and the codimension-1 bifurcation curves in the neighborhood of D_{left}. Compare this diagram to Fig. 3.

Another example where we also observed this phenomenon is the Continuous Stirred Tank Reactor (CSTR) in which a simple exothermic reaction $A \rightarrow B$ takes place. This classical chemical reaction engineering system can be modeled by the following set of dimensionless ODEs:

$$\frac{dx_1}{d\tau} = -x_1 + Da(1-x_1)\exp(x_2),$$

$$\frac{dx_2}{d\tau} = -x_2 + B\,\mathrm{Da}\,(1-x_2)\exp(x_2) + \beta(T_c - x_2),$$

where x_1 is a dimensionless concentration of reactant A, x_2 is a dimensionless temperature, and Da (the Damkoehler number), B (the dimensionless heat of reaction),

$T_c = T_{c,0} + \alpha \cos(\omega\tau)$ (the coolant temperature), and β (the dimensionless heat transfer coefficient between the reactor and the coolant fluid) are parameters. For $B = 22$, $\mathrm{Da} = 0.085$, $\beta = 3$, and $T_{c,0} = 0$ the autonomous system ($\alpha = 0$) has an attracting limit cycle of period $T_0 = 1.094996$ surrounding an unstable steady state. In a previous publication [KAS] degenerate period doublings were observed on both $2/p$ horns studied (the 2/1 and the 2/3 horns). Figure 9 is a three-dimensional representation of the full four-dimensional phase × parameter space of the solution surface and the codimension-1 bifurcation curves in the neighborhood of the equivalent of the D_{right} point of Fig. 7 for the 2/1 resonance horn of the periodically forced CSTR ($a_r = 0.063036$). Compare Fig. 9 also to Fig. 3.

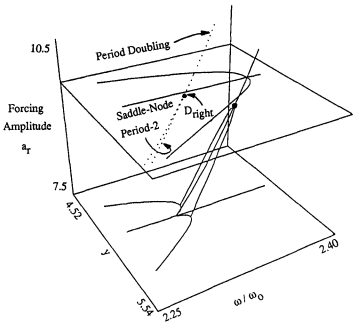


FIG. 9. *Forced* CSTR: *singly degenerate period doubling.*

Recent studies by McKarnin, Schmidt, and Aris [MSA] (a periodically forced surface reaction model), Schreiber et al. [SDCM] (a periodically forced Brusselator), as well as by Vance and Ross [VR] (a periodically forced CSTR) have also repeatedly revealed degenerate period doublings on the boundaries of $2/p$ resonance horns. This bifurcation appears therefore to be ubiquitous in models of periodically forced oscillators arising in various disciplines.

4.2. High-amplitude closing of the resonance horns. In our example (Fig. 7), as well as in the numerous studies of periodically forced oscillators we referred to above, the phenomenon of high-amplitude "closing" of the $2/p$, and generally of the q/p resonance horns was observed. It has been shown that this "closing" phenomenon implies the existence of certain codimension-2 bifurcations for the maps [AMKA], [P1], [P2], [P3]. In most horns, the boundary consists of codimension-1 saddlenode bifurcation curves for the qth iterate of the map along with certain codimension-2 points on these curves. For a $2/p$-horn, however, this boundary typically changes from a saddlenode curve for the 2nd iterate of the map to a period doubling curve in order for the horn to close. The point at which they change is the codimension-2 degenerate period doubling point.

See the references above for details and [Ga] for a related analytical study.

5. Discussion.

5.1. The Hopf bifurcation with higher-order degeneracies. As we mentioned in the introduction, certain higher-order degeneracies in the Hopf bifurcation for flows generate bifurcation diagrams almost identical to those for the period doubling bifurcation with higher-order degeneracies. This is not surprising if we look at the model

flows of Table 1:

$$r' = \varepsilon_1 r + \varepsilon_2 r^3 + \cdots + \varepsilon_{2k-1} r^{2k-1} + \delta r^{2k+1},$$

$$\theta' = \omega + r^2.$$

Circular limit cycles exist whenever r satisfies $r(\varepsilon_1 + \varepsilon_2 r^2 + \cdots + \varepsilon_{2k-1} r^{2k-2} + \delta r^{2k}) = 0$. That is, the roots of this function determine the topological phase portraits of the corresponding flows. But this function is precisely $rP_{\varepsilon;k,\delta}(r^2)$, the bifurcation function we defined in (2.3) and used for our model period doublings in § 2. In both cases, the root at $r = 0$ corresponds to a "center" fixed point; other roots correspond to limit cycles for the Hopf flow and period-2 orbits for the period doubling map. Roots of higher multiplicity determine higher codimension bifurcation manifolds in both cases.

To prove that the general Hopf bifurcations are all like the above models, Golubitsky and Schaeffer ([GS] and references therein) define a function, analogous to Φ in § 3, whose roots determine the limit cycles for a given flow. Among several factors complicating the Hopf analysis are the facts that Φ is defined on an infinite-dimensional function space and that its kernel is two-dimensional. After performing a Lyapunov-Schmidt reduction on this function, however, they obtain the same "reduced" bifurcation function as we obtained in Theorem 3.3. That is, both problems can be reduced to finding roots of the *same* bifurcation function.

We illustrate a more geometric connection between the Hopf bifurcation for flows and the period doubling bifurcation for some fixed parameter value in Fig. 10. The flow in $\mathbf{R}^2$ induces a map in $\mathbf{R}^1$ by taking a return map of the flow along a line (not a ray) through the origin. (Let the origin be a fixed point of the map.) Limit cycles of the flow correspond to period-2 orbits of the induced map.

5.2. Other "finite sequence spaces." We characterized period-2 points of $\mathbf{G}(\mathbf{x})$ in this paper as roots of the function $\Phi(\mathbf{x}, \mathbf{y}) = (\mathbf{y} - \mathbf{G}(\mathbf{x}), \mathbf{x} - \mathbf{G}(\mathbf{y}))$ and then used the Lyapunov-Schmidt procedure to reduce $\Phi = \mathbf{0}$ to a simpler system. Brown and Roberts [BR] and Vanderbauwhede [Va] have recently used Lyapunov-Schmidt reduction for functions on similar "finite sequence spaces" whose roots characterize periodic points of periods other than 2. In general, a period-k orbit $\{\mathbf{x}^1, \cdots, \mathbf{x}^k\}$ of $\mathbf{G}: \mathbf{R}^n \to \mathbf{R}^n$ is characterized as a root of the function $\Phi: (\mathbf{R}^n)^k \to (\mathbf{R}^n)^k$ defined by $\Phi(\mathbf{x}^1, \cdots, \mathbf{x}^k) = (\mathbf{x}^2 - \mathbf{G}(\mathbf{x}^1), \mathbf{x}^3 - \mathbf{G}(\mathbf{x}^2), \cdots, \mathbf{x}^1 - G(\mathbf{x}^k))$. The Lyapunov-Schmidt reduction starts from this function.

Acknowledgments. The authors acknowledge useful discussions and suggestions by Professors D. Aronson, M. Golubitsky, M. Krupa, R. McGehee, and A. Vanderbauwhede, and the referee. The hospitality of the Center of Nonlinear Studies at the Los Alamos National Laboratory is also gratefully acknowledged.

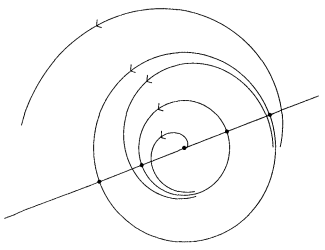


FIG. 10. *Period doubling and Hopf bifurcations.*

REFERENCES

[Ar] V. I. ARNOL'D, *Geometrical methods in the theory of ordinary differential equations*, Grundlehren, 250, Springer-Verlag, New York, 1983.

[AMKA] D. G. ARONSON, R. P. MCGEHEE, I. G. KEVREKIDIS, AND R. ARIS, *Entrainment regions for periodically forced oscillators*, Phys. Rev. A, 33 (1986), pp. 2190–2192.

[BR] A. G. BROWN AND R. M. ROBERTS, *Subharmonic bifurcations of equivariant maps*, in preparation.

[Ch] A. CHENCINER, *Bifurcations de points fixes elliptiques* II, Invent. Math., 80 (1985), pp. 81–106.

[DK] E. J. DOEDEL AND J. P. KERNEVEZ, AUTO: *Software for continuation and bifurcation problems in ordinary differential equations (including the* AUTO 86 *User Manual)*, Report, Applied Mathematics, California Institute of Technology, Pasadena, CA, 1986.

[CMY] S. N. CHOW, J. MALLET-PARET, AND J. YORKE, *A Periodic Orbit Invariant Which Is a Bifurcation Invariant*, Lecture Notes in Math. 1007, Springer-Verlag, New York, 1983, pp. 100–131.

[Ga] J. M. GAMBAUDO, *Perturbation of a Hopf bifurcation by an external time-periodic forcing*, J. Differential Equations, 57 (1985), pp. 172–199.

[GS] M. GOLUBITSKY AND D. SCHAEFFER, *Singularities and Groups in Bifurcation Theory*, Vol. I, Appl. Math. Sci., 51, Springer-Verlag, New York, 1985.

[GH] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Appl. Math. Sci., 42, Springer-Verlag, New York, 1983.

[Ha] G. R. HALL, *Resonance zones in two-parameter families of circle homeomorphisms*, SIAM J. Math. Anal., 15 (1984), pp. 1075–1081.

[HW] P. HOLMES AND D. WHITLEY, *Bifurcations of one- and two-dimensional maps*, Philos. Trans. Roy. Soc. London Ser. A, 311 (1984), pp. 43–102.

[LK] J. R. LEIS AND M. A. KRAMER, ODESSA—*An ordinary differential equation solver with explicit simultaneous sensitivity analysis*, ACM Trans. Math. Software, 14 (1985), pp. 61–67.

[KAS] I. G. KEVREKIDIS, R. ARIS, AND L. D. SCHMIDT, *The stirred tank forced*, Chem. Engrg. Sci., 41 (1986), pp 1549–1560.

[MSA] M. A. MCKARNIN, L. D. SCHMIDT, AND R. ARIS, *Forced oscillations of a self-oscillating bimolecular surface reaction model*, Proc. Roy. Soc. London Ser. A, 417 (1988), pp. 363–388.

[NPT] S. NEWHOUSE, J. PALIS, AND F. TAKENS, *Bifurcations and stability of families of diffeomorphisms*, Publ. Math. IHES, 1982.

[PK] S. PAVLOU AND I. G. KEVREKIDIS, *Microbial predation in a periodically operated chemostat*, Math. Biosci., in press.

[P1] B. B. PECKHAM, *Global properties of resonance surfaces and resonance horns*, Ph.D. thesis, University of Minnesota, Minneapolis, MN, 1988.

[P2] ———, *The necessity of the Hopf bifurcation in periodically forced oscillators*, Nonlinearity, 3 (1990), pp. 261–280.

[P3] ———, *Typical bifurcation diagrams for periodically forced oscillators*, in preparation.

[SDCM] I. SCHREIBER, M. DOLNIK, P. CHEE, AND M. MAREK, *Resonance behavior in two-parameter families of periodically forced oscillators*, Phys. Lett. A, 128 (1988), pp. 66–70.

[Ta] F. TAKENS, *Unfoldings of certain singularities of vectorfields: Generalized Hopf bifurcations*, J. Differential Equations, 14 (1973), pp. 476–493.

[VR] W. VANCE AND J. ROSS, *A detailed study of a forced chemical oscillator: Arnol'd tongues and bifurcation sets*, J. Chem. Phys., 91 (1989), pp. 7654–7670.

[Va] A. VANDERBAUWHEDE, *Branching of periodic solutions in time reversible systems, Geometry and Analysis in Nonlinear Dynamics*, Res. Notes in Math., Longman Pitman, Boston, MA, to appear.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1575-1583, November 1991

005

VARIATIONAL PRINCIPLES WITHOUT DEFINITENESS CONDITIONS*

PAUL BINDING†‡ AND QIANG YE†§

Abstract. Variational characterizations of eigentuples are discussed for a self-adjoint problem $Ax = \lambda Bx$ in a Hilbert space H. If the pair (A, B) is simultaneously real diagonable in a certain sense, and if B is 1-1, then all eigenvalues obey certain minimum and maximum principles for a generalised Rayleigh quotient. Sup-inf and inf-sup principles are established for eigenvalues which obey certain interlacing inequalities. Applications are made to both finite- and infinite-dimensional cases.

Key words. minimax principles, indefinite eigenvalue problems

AMS(MOS) subject classification. 49G05

1. Introduction. The study of pairs of self-adjoint operators has a long history. For example, Weierstrass [25] produced a canonical form for a pair of quadratic forms in finite dimensions, and this subject has been revived in recent years, often in the setting of self-adjoint operators on (perhaps indefinite) inner product spaces (cf. [17], [11]). The latter topic was given a firm foundation in infinite dimensions by Pontryagin, Krein, and subsequent workers (cf. [7], [15]). Differential eigenvalue equations involving indefinite operators have been the subject of much recent interest, as evidenced by [18].

This considerable activity has been largely devoted to extending classical results from definite to nondefinite pairs. The pair (A, B) of operators in a Hilbert space H is *self-adjoint* (respectively, *definite*) if A and B are self-adjoint (respectively, admit a definite linear combination). Here we shall discuss variational principles for eigenvalues of a class of pairs with no definiteness assumptions. In definite cases (say with B definite), such principles usually involve the generalised Rayleigh quotient

$$r(x) = a(x)/b(x)$$

where

$$a(x) = (x, Ax), \qquad b(x) = (x, Bx), \tag{1.1}$$

and $r(x)$ is defined for all nonzero $x \in D(A) \cap D(B)$. On the other hand, if B is indefinite, then $b(x)$ necessarily vanishes for certain nonzero x, and then the very definition of $r(x)$ needs modification. It is perhaps for this reason that few variational principles exist for nondefinite pairs. There is recent finite-dimensional work [2] but it seems unrelated to ours. An approach to definite pairs via indefinite inner product spaces was given by Phillips [22] and generalised by Textorius [23] to possibly singular pencils. This approach involves triple extrema, and we hope to return to this elsewhere, but here we shall concentrate on single and double extrema.

Eigenvalue equations

$$Ax = \lambda Bx \tag{1.2}$$

involving nondefinite pairs (A, B) arise in various applications. One of the earliest comes from separation of variables in boundary value problems for waves that are

* Received by the editors October 2; 1989; accepted for publication (in revised form) January 28, 1991.

† Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4.

‡ The research of this author was supported by a National Sciences and Engineering Research Council of Canada Operating Grant.

§ The research of this author was supported by a University of Calgary Research Fellowship.

electromagnetic, acoustic, etc. This leads to two (or more) parameter problems like $Ax=(\lambda B+\mu C)x$ (cf. [24]). Depending on μ/λ and the boundary conditions for A, this problem is generally nondefinite. Another application concerns quantum theory of crystal lattices, involving operators A of the form $-\Delta+q$, where q is a periodic potential [16]. Recently Deift, Hempel, and others have made several investigations of this topic, B representing the potential of an impurity (cf. [8]). Another seminal nondefinite application of (1.2) lies in the stability analysis of population genetics models, where A is a Neumann Laplacian and B is a linearized natural selection term (cf. [10]). Much subsequent work has been carried out on abstract versions, e.g., [14]. We also note that quadratic eigenvalue problems can frequently be put into a nondefinite form (1.2) satisfying our hypotheses. We cite Greenlee's work [12] which applies to certain problems in fluid dynamics, and we shall analyse a problem from mechanics in §4.

For such problems, it turns out that all but finitely many of the real eigenvalues are of either positive or negative "type," depending on the sign of one of the quadratic forms in (1.1) (usually $b(x)$) at eigenvectors, and those of positive (respectively, negative) "type" accumulate only at $+\infty$ (respectively, $-\infty$). For example, the real eigenvalues λ_j of the regular Sturm–Liouville problem

$$(-py')'+qy=\lambda wy \qquad (p>0) \tag{1.3}$$

(with indefinite weight w) may be indexed so that $\lambda_j\to\pm\infty$ as $j\to\pm\infty$ and λ_j corresponds, for $|j|\geq j_1$, to an eigenfunction y with $j-1$ internal zeros. For $j\geq j_1$ (respectively, $j\leq -j_1$), λ_j is of positive (respectively, negative) "type." For a history of such results, see Mingarelli [20], who calls j_1 (which we take to be minimal) the Haupt index after his 1915 paper [13]. For recent work on abstract generalisations of (1.3), including various matrix and partial different equation problems, we refer to [5]. Similar considerations apply to other aspects of (1.3): for large $|j|$, the λ_j (and the corresponding eigenfunctions) behave much as for the (left) definite case. As an example, the eigenfunctions for $|j|\geq j_1$ span a subspace K of finite codimension, and K is a Hilbert space in the inner product generated by a of (1.1) (cf. [9] and [6] for an abstract version).

The importance of the regularity index j_1 for variational principles was pointed out in [19] for matrix problems, and in [1] for the problem

$$(-\Delta+q)y=\lambda wy, \tag{1.4}$$

where Δ is the n-dimensional Laplacian with Dirichlet boundary conditions. Specifically, if a certain diagonability condition holds, then for $j\geq j_1$ the eigenvalues λ_j are of positive "type" and can be characterized by

$$\lambda_j=\sup\{\inf\{r(x):x\in C^+\cap S\}:\operatorname{codim} S=k\}, \tag{1.5}$$

with a similar formula for $j\leq -j_1$ where the eigenvalues are of negative "type." Here

$$C^{\pm}=\{x\in D(A)\cap D(B):\pm b(x)>0\}, \tag{1.6}$$

S is a subspace of H, and k depends on j. On the other hand, it has been conjectured (in private communication) that no such characterization is possible for eigenvalues λ_j with $|j|<j_1$. We shall refer to the latter as "overlap" eigenvalues, since those of positive and negative "type" are no longer separated, and indeed $a(x)=b(x)=0$ is possible, so some eigenvalues may not have a defined "type."

We shall see that (1.5) is valid under quite general circumstances, and does indeed characterize some (although not all) "overlap" eigenvalues. Various related "dual" formulae will also be given. Section 2 contains preliminary results, including a general

minimum principle, valid for all eigenvalues. This gives a complete recursive characterization of eigenvalues for certain pairs. Section 3 is devoted to double extremum principles like (1.5), with simple (mostly finite-dimensional) illustrations. Section 4 contains applications to various matrix, differential, and integral equation problems, including (1.3) and (1.4).

2. Preliminaries. Let us start with our basic assumption, which is the existence of a linear homeomorphism T which "simultaneously diagonalises" A and B. More precisely, we assume the following:

(A1) There exist

(i) an integer index set J, not containing zero,

(ii) a bounded linear operator T on H with bounded inverse,

(iii) a complete orthonormal basis $e_j (j \in J)$ of H,

(iv) real α_j and real β_j taking the sign of j, such that $T^*ATe_j = \alpha_j e_j$, $T^*BTe_j = \beta_j e_j$, $j \in J$.

Remarks. Assumptions (ii) and (iii) mean that the elements Te_j form a Riesz basis of H. We stress that explicit knowledge of T is unnecessary; only its existence is required, and sufficient conditions on A and B will be given in §4. Also (iv) is a "real diagonability" condition; it is straightforward to relax this to all but finitely many $j \in J$, using constructions given in [21] for the span of the remaining Te_j. For notational simplicity, however, we shall adopt (iv) as stated.

We also need some definitions before stating the minimum principle. A complex number λ is an *eigenvalue* of (A, B) if the *eigenspace* $E_\lambda := N(A - \lambda B)$ is nontrivial, and x is a $b\pm$ *eigenvector* if $x \in E_\lambda \cap C^\pm$; see (1.6). We write $\lambda_j = \alpha_j/\beta_j$, and for any real λ we define index sets

$$I_\lambda^\pm = \{j: \pm(\alpha_j - \lambda\beta_j) \geq 0\}, \qquad I_\lambda^0 = I_\lambda^+ \cap I_\lambda^- = \{j: \lambda = \lambda_j\}$$

and subspaces

$$S_\lambda^\pm = \text{Span}\,\{Te_j: j \in I_\lambda^\pm\}, \qquad S_\lambda^0 = \text{Span}\,\{Te_j: j \in I_\lambda^0\},$$

where Span means closed linear span. This notation allows considerable freedom in the ordering of the λ_j, e.g., in the case of several accumulation points. In our examples, $\lambda_j \leq \lambda_{j+1}$ will hold separately for positive and negative j (although $\lambda_1 < \lambda_{-1}$ in the case of "overlap" eigenvalues). As an illustration, suppose that b is definite on each separate eigenspace. If λ is an eigenvalue of (A, B), then S_λ^+ turns out to be the Span of those E_μ on which b is positive definite and $\mu \geq \lambda$, together with those on which b is negative definite and $\mu \leq \lambda$. Also $E_\lambda = S_\lambda^+ \cap S_\lambda^- = S_\lambda^0$.

THEOREM 2.1.

(i) $I_\lambda^0 \neq \emptyset \Leftrightarrow \lambda$ *is an eigenvalue of* (A, B).

(ii) *If* λ *admits a* $b\pm$ *eigenvector* x, *then*

$$\lambda = \min\{r(x): x \in S_\lambda^\pm \cap C^\pm\} \tag{2.1}$$

and x *is a corresponding minimiser.*

(iii) E_λ *is spanned by the set of minimisers* (*using both possible signs*) *in* (2.1).

Proof.

(i) If $I_\lambda^0 \neq \emptyset$, then $\alpha_j = \lambda\beta_j$ for some j. Thus

$$(A - \lambda B)Te_j = T^{-*}(\alpha_j - \lambda\beta_j)e_j = 0.$$

Conversely, if $0 \neq x \in E_\lambda$ and

$$T^{-1}x = y = \sum_{j \in J} y_j e_j, \tag{2.2}$$

then

$$0 = T^*(A-\lambda B)Ty = \sum_{j\in J} (\alpha_j - \lambda\beta_j) y_j e_j. \tag{2.3}$$

If $I_\lambda^0 = \emptyset$, then the inner product of (2.3) with e_k gives $y_k = 0$. Thus $y = 0$, so we obtain the contradiction $x = 0$.

(ii) If $x \in S_\lambda^+ \cap C^+$, then (2.2) gives

$$r(x) = \sum_{j\in I_\lambda^+} \alpha_j |y_j|^2 / \sum_{j\in I_\lambda^+} \beta_j |y_j|^2 \geq \lambda. \tag{2.4}$$

If, moreover, $x \in E_\lambda \cap C^+$, then (2.3) implies $y_j = 0$ unless $j \in I_\lambda^0$, so

$$x \in S_\lambda^0 \cap C^+ \subseteq S_\lambda^+ \cap C^+.$$

It follows that $\geq$ may be replaced by $=$ in (2.4), and the proof is complete for the + sign case. For the − sign case, we replace (A, B) by $(-A, -B)$ and apply the + sign case.

(iii) I_λ^0 partitions into two subsets, one positive and one negative. These subsets generate $b\pm$ eigenvectors Te_j, which by (ii) are minimisers for the two sign cases. Moreover these minimisers span S_λ^0, which as we saw above equals E_λ. □

COROLLARY 2.2. *Equation* (2.1) *may be replaced by*

$$\lambda = \max\{r(x): x \in S_\lambda^\pm \cap C^\pm\}.$$

Proof. Replace A by $-A$ and apply Theorem 2.1(ii). □

Let us briefly examine the consequences for definite pencils in finite dimensions. If B is (say positive) definite (the "right" definite case), then the S_λ^- are trivial, and the S_λ^+ nest. In fact all eigenvalues for (A, B) may be calculated recursively via (2.1), beginning with the least eigenvalue λ, for which $S_\lambda^+ = H$. This is the standard "generalised Rayleigh quotient" recursive characterization, since $C^+ = H\backslash\{0\}$. If instead $A - \lambda_0 B$ is (say positive) definite then b takes the sign of $\lambda - \lambda_0$ on E_λ. Thus again (2.1) characterizes the eigenvalues for (A, B) recursively, beginning with the least $> \lambda_0$ (for which $S_\lambda^+ = H$) and the greatest $< \lambda_0$ (for which $S_\lambda^- = H$). Similar considerations apply in infinite dimensions provided the extreme eigenvalues above exist, but the recursions stop at accumulation points of eigenvalues. See §4 for applications.

We shall need two specific results for perhaps nondiagonable pencils in two dimensions, and we shall first examine the role Theorem 2.1 plays in the indefinite, but real diagonable, case. Recall the notation $\lambda_j = \alpha_j/\beta_j$.

LEMMA 2.3 *Let* $\dim H = 2$ *and* $J = \{-1, 1\}$.

(i) *If* $\lambda_{-1} \leq \lambda_1$ *then* $\min\{r(x): x \in C^+\} = \lambda_1$.

If $\lambda_{-1} > \lambda_1$ *then*

(ii) $\inf\{r(x): x \in C^+\} = -\infty$ *and*

(iii) $\max\{r(x): x \in C^+\} = \lambda_1$.

Proof. Lemma 2.3 (i) follows from Theorem 2.1 since $S_{\lambda_1}^+ = H$. For (ii) we follow (2.2) to give

$$r(x) = \beta_{-1}^{-1}\alpha_{-1} + \beta_1(\lambda_1 - \lambda_{-1})|y_1|^2 / (\beta_{-1}|y_{-1}|^2 + \beta_1|y_1|^2). \tag{2.5}$$

Now let $y_{-1} = 1$, $y_1 \in \mathbb{R}$ and $\beta_1 y_1^2 \downarrow -\beta_{-1}$ to give the result.

Lemma 2.3(iii) follows from Corollary 2.2 since $S_{\lambda_1}^- = H$. □

The other characterizations degenerate for this problem, e.g., $\lambda_1 = \min\{r(x): x \in E_\mu\}$ in (ii).

We can now derive the two special results that we need later. Real diagonability of the corresponding pencils is not assumed.

COROLLARY 2.4. *Let* α_j, β_j *be as in Lemma* 2.3, *and* $\alpha_0 \in \mathbb{R}$. *If* $\lambda_{-1} > \lambda_1$ *and*

$$\alpha(y) := \alpha_{-1} y_{-1}^2 + \alpha_1 y_1^2 + \alpha_0 y_{-1} y_1, \tag{2.6}$$

then

$$\inf\{\alpha(y)/(\beta_{-1} y_{-1}^2 + \beta_1 y_1^2) : \beta_{-1} y_{-1}^2 + \beta_1 y_1^2 > 0, (y_{-1}, y_1) \in \mathbb{R}^2\} = -\infty.$$

Proof. This follows from Lemma 2.3 (ii) provided we choose the sign of y_1 in (2.5) so that $\alpha_0 y_1 \le 0$. □

For the second result, we retain the index set $J = \{-1, 1\}$ for convenience, even though we assume $\beta_{-1} = 0$.

LEMMA 2.5. *Let* $\alpha_{-1} < 0 < \beta_1$; $\alpha_1, \alpha_0, \beta_0 \in \mathbb{R}$ *and* $\alpha(y)$ *be as in* (2.6). *Then*

$$\inf\{\alpha(y)/(\beta_1 y_1^2 + \beta_0 y_{-1} y_1) : \beta_1 y_1^2 + \beta_0 y_{-1} y_1 > 0, (y_{-1}, y_1) \in \mathbb{R}^2\} = -\infty.$$

Proof. We set $y_1 = \beta_0$ if $\beta_0 \neq 0$ and $y_1 = 1$ if $\beta_0 = 0$. It follows that

$$\beta_1 y_1^2 + \beta_0 y_{-1} y_1 > 0$$

for all $y_{-1} > 0$, and as $y_{-1} \to \infty$, $\alpha(y)/y_{-1}^2 \to \alpha_{-1}$. This suffices for the conclusion. □

3. Double extremum principles. We need one more finite-dimensional result in preparation for the variational principle.

LEMMA 3.1. *Let* dim $H = n < \infty$, $b(u_0) < 0$ *and let* J *have* p *positive elements. Then there exists a* p*-dimensional subspace* U, *orthogonal to* Bu_0, *with* b *positive definite on* U.

Proof. Let u_{-j}, $0 \le j \le n - p + 1$, be a basis for a maximal b-negative subspace S. The elements Bu_{-j} are linearly independent, so their span BS has dimension $n - p$. By the law of inertia, b is positive definite on $U = (BS)^{\perp}$. □

In order to state the variational principle efficiently, we shall split the index sets $I_\lambda^\pm$ into two. Specifically, we define

$$J_{\ge\lambda}^{\pm} = \{\pm j > 0 : \lambda_j \ge \lambda\}, \qquad T_{\ge\lambda}^{\pm} = \text{Span}\{Te_j : j \in J_{\ge\lambda}^{\pm}\}$$

with similar definitions for $T_{\le\lambda}^{\pm}$, $T_{<\lambda}^{\pm}$, and $T_{>\lambda}^{\pm}$. Note $S_\lambda^\pm = T_{\ge\lambda}^{\pm} + T_{\le\lambda}^{\mp}$ (cf. the illustration before Theorem 2.1).

THEOREM 3.2. *Let* λ *admit a* $b\pm$ *eigenvector, and suppose* $\mu > \lambda$, *where*

$$\dim T_{<\lambda}^{\pm} = k < \infty, \qquad \dim T_{\ge\mu}^{\mp} = l < \infty,$$

$T_{>\lambda}^{\mp} \cap T_{<\mu}^{\mp} = \{0\}$ *and* $\dim(T_{\ge\lambda}^{\pm} \cap T_{<\mu}^{\pm}) > l$. *Then* $\lambda = \sigma_{k+l}^{\pm}$, *where*

$$\sigma_m^{\pm} = \sup\{\inf\{r(x) : x \in S \cap C^{\pm}\}; \text{codim } S = m\}. \tag{3.1}$$

The dual result involves replacing A by $-A$, and is as follows.

COROLLARY 3.3. *This is the same as for Theorem* 3.2, *with the following changes: all inequalities* (*except* "$m \ge l$") *are reversed, and* "sup" *is interchanged with* "inf."

Remark. Since T is 1-1, it need not be known explicitly to determine dim $T_{<\lambda}^{\pm}$, etc. Roughly, the dimension conditions state (in the case of a $b+$ eigenvector at λ) that there must be enough $\lambda_j^+ \in [\lambda, \mu[$ to "cancel" any $\lambda_j^- \ge \mu$.

Before proving Theorem 3.2, we shall discuss some simple illustrations. Suppose first that dim $H = 4$, and $J = \{\pm 1, \pm 2\}$. We shall concentrate on the positive sign case of (3.1).

Example 3.4. If $\lambda_{-2} < \lambda_{-1} < \lambda_1 < \lambda_2$, then the pencil is definite (there is no "overlap"), $\sigma_0^+ = \lambda_1$ (with $k = l = 0$), in agreement with (2.1), and $\sigma_1^+ = \lambda_2$ (with $k = 1$, $l = 0$), in agreement with [19].

Example 3.5. If $\lambda_{-2} < \lambda_1 < \lambda_2 < \lambda_{-1}$ then the pencil is indefinite, and $\sigma_1^+ = \lambda_1$, with $k = 0$, $l = 1$. We know of no other results that characterize "overlap" eigenvalues like λ_1.

More generally, we have the following result.

THEOREM 3.6. *If* $\dim H = 4$, *then all quantities* σ_j^+ *are either infinite or else characterize eigenvalues via Theorem* 3.2 *or Corollary* 3.3.

We shall not prove this here, since there are many cases. We may note, however, that $\sigma_2^+ = \sigma_3^+ = \infty$ in Examples 3.4 and 3.5, while $\sigma_0^+ = -\infty$ in Example 3.5. Also (3.1) gives $\sigma_1^- = \lambda_{-1}$ with $k = 1$, $l = 0$, in Example 3.5, and the remaining eigenvalues λ_j can be characterized via Corollary 3.3 in the above exaples.

The simplest infinite-dimensional situations with $l = 0$ and $l = 1$ are obtained by expanding the sequences λ_j in Examples 3.4 and 3.5, using negative j to the left and positive j to the right.

Another infinite-dimensional case may be of interest. We assume that B has only finitely many (counting multiciplicity) negative eigenvalues, i.e., there are only finitely many negative β_j. If $J = J_1 \cup J_2 \cup N$ where J_1, J_2 consist of (possibly infinitely many) positive integers, N consists of finitely many negative integers, and

$$\lambda_i < \lambda_j \quad \text{for any } i \in J_1 \quad \text{and} \quad j \in J_2 \cup N,$$

then the eigenvalues $\{\lambda_i : i \in J_1\}$ are characterized by Theorem 3.2. Note that if $N = \varnothing$, i.e., B is positive definite, then the classical minimax theorem characterizes λ_i for $i \in J_1$. So this result says that the minimax theorem remains valid even in the presence of negative index eigenvalues provided that they do not destroy the extreme structure of $\{\lambda_i : i \in J_1\}$. There is, however, an index shift in the characterization.

Proof of Theorem 3.2. The negative sign case in (3.1) may be reduced to the positive sign case by the transformation $(A, B) \to (-A, -B)$, so we shall discuss only the positive sign case. Then Theorem 2.1 gives

$$\lambda = \min \{r(x) : x \in C^+ \cap S_\lambda^+\},$$

and since $T(T^{-1}S_\lambda^+)^\perp = T_{<\lambda}^+ + T_{>\lambda}^-$ has dimension $k + l$, it will suffice to prove that for any S with codimension $k + l$,

$$\text{there exists } x \in S \cap C^+ \text{ such that } r(x) \le \lambda. \tag{3.2}$$

Now let

$$\Sigma \subseteq T_{\le\lambda}^+ + T_{\ge\mu}^- \tag{3.3}$$

be a subspace of (finite) dimension $> k + l$, so there is a nonzero $u_0 = v + w \in \Sigma \cap S$, where $v \in T_{\le\lambda}^+$ and $w \in T_{\ge\mu}^-$. Since $r(v) \le \lambda$, $r(w) \ge \mu$, and $b(v) > 0 > b(w)$, it follows that

$$a(v) \le \lambda b(v) \quad a(w) \le \mu b(w) \quad \text{and} \quad (v, Aw) = (v, Bw) = 0. \tag{3.4}$$

We must consider three possible cases.

Case 1. $b(u_0) > 0$. Then (3.3) gives

$$r(u_0) \le [\lambda b(v) + \mu b(w)]/[b(v) + b(w)]$$

and the right side $\le \lambda$ by Lemma 2.3 (iii), since $\lambda < \mu$. Thus $x = u_0$ suffices for (3.2).

Case 2. $b(u_0) = 0$. By assumption, $\dim T_{<\mu}^+ > k + l$, so there is nonzero $x \in S \cap T_{<\mu}^+$. Choose $x = y_{-1}u_0 + y_1 z$, where $(y_{-1}, y_1) \in \mathbb{R}^2$. Now $b(z) > 0$ and by (3.4)

$$a(u_0) = a(v) + a(w) < \mu b(v) + \mu b(w) = \mu b(u_0), \tag{3.5}$$

so $a(u_0)<0$. Thus

$$r(x)=[a(u_0)y_{-1}^2+a(z)y_1^2+2Re(u_0, Az)y_{-1}y_1]/[b(z)y_1^2+2Re(u_0, Bz)y_{-1}y_1]$$

can be made $\leq\lambda$ by virtue of Lemma 2.5, and (3.2) is satisfied.

Case 3. $b(u_0)<0$. Lemma 3.1, with H replaced by Σ (3.3) gives a b-positive subspace $U\subseteq\Sigma$ of dimension $>k$, orthogonal t Bu_0. Now choose a subspace $V\subset T^+_{\geq\lambda}\cap T^+_{<\mu}$ of finite dimension so that $Q=U+V$ has dimension $>k+l$. Evidently Q is b-positive and orthogonal to Bu_0, and so there is $q\in Q\cap S$ such that

$$b(q)>0 \quad \text{and} \quad (q, Bu_0)=0. \tag{3.6}$$

Moreover dim $Q<\infty$, so $q\in T^+_{\leq\nu}+T^-_{\geq\mu}$ for some $\nu<\mu$, and arguing as for Case 1 we conclude $r(q)\leq\nu$, i.e.,

$$a(q)\leq\nu b(q). \tag{3.7}$$

Now choose $x=y_{-1}u_0+y_1q$, where $(y_{-1}, y_1)\in\mathbb{R}^2$. We may then use (3.5), (3.6), (3.7), and $\nu<\mu$ to show that

$$r(x)=[a(u_0)y_{-1}^2+a(q)y_1^2+2Re(u_0, Aq)y_{-1}y_1]/[b(u_0)y_{-1}^2+b(q)y_1^2]$$

can be made $\leq\lambda$ by virtue of Corollary 2.4, and again (3.2) is satisfied. □

4. Applications. Our starting point is a special case of a canonical form in [4].

THEOREM 4.1. *Suppose $A=I+C$ and B is* 1-1 *with B, C compact symmetric on H. Then there exists a linear homeomorphism T of H, and subspaces F, G such that* dim $F<\infty$, G_A (i.e., *G with the inner product generated by A*) *is a Hilbert space and, relative to the* (*A-orthogonal*) *direct sum $H=F\oplus G$, $T^*AT=A_F\oplus I_G$, $T^*BT=B_F\oplus Q$, where Q is* 1-1, *compact and symmetric on G.*

Explicit constructions are given in [4], A_F and B_F being the Gram operators (in the usual inner product) for A and B on F. Thus assumption (A1) in § 2 reduces to real diagonability of the finite-dimensional operator $B_F^{-1}A_F$, but as we mentioned in the remark after (A1) this diagonability is unnecessary if we use the constructions of [21]. Essentially, therefore, our results hold for (1.2) with A, B as in Theorem 4.1. For more general situations, e.g., with B merely bounded and symmetric, see [4].

In finite dimensions, the results of [19, Thms. 2.1, 2.2] correspond to the case $l=0$ of Theorem 3.2 and Corollary 3.3, and the classical situation of definite pencils is a special case of this. The bounded operator setting given above applies directly to integral equations with symmetric kernels. We remark that there are frequently several problem formulations, which affect (β_j) signs, λ_j accumulation points, etc. For example, the eigenvalues λ of a compact symmetric operator Γ can be obtained via the pair (Γ, I), which is not covered by Theorem 4.1, although (A1) is satisfied. The reciprocals λ^{-1} can be obtained via the pair (I, Γ), which does fit Theorem 4.1 and (A1). All (nonzero) eigenvalues are characterized via Theorem 3.2 and Corollary 3.3: for (Γ, I), $C^{\pm}=H$, while for (I, Γ), $C^{\pm}$ are defined by Γ.

It may also be noted that Theorem 4.1 can also be used to treat (1.2) in the case $A=P+C$, where P is positive definite with a compact inverse and B, C are bounded. Then

$$P^{-1/2}AP^{-1/2}=I+C_1 \quad \text{and} \quad P^{-1/2}BP^{-1/2}=B_1 \tag{4.1}$$

with C_1 and B_1 satisfying the hypothesis of Theorem 4.1, and $P^{-1/2}$ can be combined with T. We stress that T (and $P^{1/2}$), which are difficult to calculate in general, which, need not be determined explicitly: for example codim S equals j if and only if codim (TS) equals j. This applies to certain types of unbounded operator equations, including

various self-adjoint uniformly elliptic boundary value problems like (1.3) and (1.4). Specifically, B and C are the multiplication operators associated with w and q, respectively, and $Py=(-py')'$ for (1.3) and $Py=-\Delta y$ for (1.4). Then the above conditions follow if all coefficients are L_∞ on a compact domain for the equation, and $1/p$ is L_1. We then obtain, for example, [1, Cor. 2.9] in the case $l=0$ of Theorem 3.2.

Finally, let us examine a gyroscopic stabilization problem from mechanics [3], which leads to a quadratic eigenvalue problem

$$(\lambda^2 I+\lambda iG+C)y=0 \tag{4.2}$$

where $G=-G^*$ and $C=C^*>0$. Using a standard linearization, we may rewrite this in the (nondefinite) form (1.2) where

$$A=\begin{bmatrix} iG & I \\ I & 0 \end{bmatrix} \quad \text{and} \quad N=\begin{bmatrix} C & 0 \\ 0 & -I \end{bmatrix}.$$

Under the (gyroscopic stabilizability) condition $G^2+(kI+k^{-1}C)^2<0$ for some $k>0$, it is shown in [3] that each eigenvalue of (4.2), and hence of (1.2), is real and of definite type. Thus the real diagonability condition is satisfied, i.e., assumption (A1) holds as stated.

It is also shown in [3] that if iG is $n\times n$ and has p positive eigenvalues, then the eigenvalues for (4.2), i.e., for (1.2), satisfy

$$\lambda_{-n}\le\cdots\le\lambda_{p-n-1}<\lambda_1\le\cdots\le\lambda_p<0<\lambda_{p-n}\le\cdots\le\lambda_{-1}<\lambda_{p+1}\le\cdots\le\lambda_n.$$

Thus our results characterize λ_j for $j<p-n$, $j>p$, and $1\le j\le 2p-n$ if $2p>n$ (respectively, $-1\ge j\ge n-2p$ if $n>2p$).

REFERENCES

[1] W. ALLEGRETTO AND A. B. MINGARELLI, *Boundary problems of the second order with an indefinite weight function*, J. Reine Angew. Math., 398 (1989), pp. 1-24.

[2] G. AUCHMUTY, *Variational principles for eigenvalues of nonsymmetric matrices*, SIAM J. Appl. Math., 10 (1989), pp. 105-117.

[3] L. BARKWELL, P. LANCASTER, AND A. S. MARKUS, *Gyroscopically stabilized system: A class of quadratic eigenvalue problems with real spectrum*, preprint.

[4] P. A. BINDING, *A canonical form for self-adjoint pencils in Hilbert space*, Integral Equations Operator Theory, 12 (1989), pp. 324-342.

[5] P. A. BINDING AND P. J. BROWNE, *Applications of two parameter spectral theory to symmetric generalised eigenvalue problems*, Appl. Anal., 29 (1988), pp. 108-142.

[6] P. A. BINDING AND K. SEDDIGHI, *On root vectors of self-adjoint pencils*, J. Funct. Anal., 70 (1987), pp. 117-125.

[7] J. BOGNÁR, *Indefinite Inner Product Spaces*, Springer-Verlag, Berlin, New York, 1974.

[8] P. A. DEIFT AND R. HEMPEL, *On the existence of eigenvalues of the Schrödinger operator H-λW in a gap of σ(H)*, Comm. Math. Phys., 103 (1986), pp. 461-490.

[9] K. DAHO AND H. LANGER, *Sturm-Liouville operators with an indefinite weight function*, Proc. Roy. Soc. Edinburgh, 78A (1977), pp. 161-191.

[10] W. H. FLEMING, *A selection-migration model in population genetics*, J. Math. Biol., 2 (1975), pp. 219-233.

[11] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrices and Indefinite Scalar Products*, Birkhäuser, Basel, 1983.

[12] W. M. GREENLEE, *Double unconditional bases associated with a quadratic characteristic parameter problem*, J. Funct. Anal., 15 (1974), pp. 306-339.

[13] O. HAUPT, *Über eine Methode zum Bewweise von Oszillationstheoreme*, Mat. Ann., 76 (1915), pp. 67-104.

[14] P. HESS AND T. KATO, *On some linear and nonlinear eigenvalue problem with an indefinite weight function*, Comm. Partial Differential Equations, 5 (1980), pp. 999-1030.

[15] I. S. IOHVIDOV, M. G. KREIN, AND H. LANGER, *Indefinite Inner Product Spaces*, Akademie-Verlag, Berlin, 1982.

[16] R. KRONIG AND W. G. PENNEY, *Quantum mechanics of electrons in crystal lattices*, Proc. Roy. Soc. London, 130 (1931), pp. 499–513.

[17] B. KÅGSTRÖM AND A. RUHE, EDS. *Matrix Pencils*, Lecture Notes in Math. 973, Springer-Verlag, Berlin, New York, 1983.

[18] H. G. KAPER AND A. ZETTL, EDS., *Proc.* 1984 *Workshop: Spectral Theory of Sturm–Liouville Differential Operators*, Argonne National Laboratory, Argonne, IL, 1984.

[19] P. LANCASTER AND Q. YE, *Variational properties and Rayleigh quotient algorithms for symmetric matrix pencils*, Operator Theory: Adv. Appl., 40 (1989), pp. 247–278.

[20] A. B. MINGARELLI, *A survey of the regular weighted Sturm–Liouville problem: The non-definite case*, Proc. Workshop on Applied Differential Equations, Tsinghua Univ., Beijing, 1985.

[21] B. NAJMAN AND Q. YE, *A minimax characterization of eigenvalues of Hermitian pencils*, Linear Algebra Appl., 144 (1991), pp. 217–230.

[22] R. S. PHILLIPS, *A minimax characterization for the eigenvalues of a positive symmetric operator in a space with indefinite metric*, J. Fac. Sci. Univ. Tokyo Sect. 1A (Math.), 17 (1970), pp. 51–59.

[23] B. TEXTORIUS, *Minimaxprinzipe zur Bestimmung der Eigenwerte J-nichtnegativer Operatoren*, Math. Scand., 35 (1974), pp. 105–114.

[24] H. VOLKMER, *Multiparameter Eigenvalue Problems and Expansion Theorems*, Lecture Notes in Math. 1356, Springer-Verlag, Berlin, New York, 1988.

[25] K. WEIERSTRASS, *Zur Theorie der quadratischen und bilinearen Formen*, Monats. Akad. Wiss. Berlin (1868), pp. 310–338.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1584–1630, November 1991

006

PERSISTENCE OF INVARIANT TORI IN SYSTEMS OF COUPLED OSCILLATORS. II. DEGENERATE PROBLEMS*

MASAJI WATANABE†‡ AND HANS G. OTHMER†

Abstract. This paper continues the study of persistence of invariant tori in a class of differential equations that describe indirectly- or capacitatively-coupled systems of nonlinear oscillators. In a previous paper the authors proved that when the unperturbed torus is hyperbolic it persists under weak coupling, and the flow on the perturbed tori was analyzed. In this paper the persistence of invariant tori is proved in certain degenerate cases in which the invariant manifold in the unperturbed problem is not normally hyperbolic. This is done by proving the existence of fixed points in appropriate Banach spaces.

Key words. coupled oscillators, invariant tori

AMS(MOS) subject classifications. 34C35, 34A34, 34C15, 34C45

1. Introduction.

A number of applications in physics and biology lead to a system of the form

$$\begin{aligned} \frac{dx_i}{dt} &= f_i(x_i) + \delta P(x_0 - x_i), \qquad i = 1, \cdots, N, \\ \frac{dx_0}{dt} &= \epsilon\delta P\left(\frac{1}{N}\sum_{i=1}^{N} x_i - x_0\right) \end{aligned} \tag{1.1}$$

(cf. [16], hereafter referred to as Part I, for references to the applications). In (1.1) P is an $n \times n$ constant matrix of permeability coefficients or conductances, and the parameters ϵ^{-1} and δ measure the relative capacity of the subunits and the coupling strength, respectively. In the absence of coupling the evolution in the ith subunit is governed by the n-dimensional system

$$\frac{dx_i}{dt} = f_i(x_i), \tag{1.2}$$

and it is assumed that there is a nonconstant periodic solution

$$x_i = \eta_i(t) \tag{1.3}$$

with least period $T_i > 0$ of (1.2) for $i = 1, \cdots, N$. The variable x_0 represents the state of the coupling medium, through which the subunits are coupled [13], [12].

In this paper we study the following generalization of (1.1):

$$\begin{aligned} \frac{dx_i}{dt} &= f_i(x_i) + \delta g_i(x_0, x_1, \cdots, x_N, \delta), \qquad i = 1, \cdots, N, \\ \frac{dx_0}{dt} &= \epsilon\delta\left(Bx_0 + \sum_{i=1}^{N} C_i x_i\right). \end{aligned} \tag{1.4}$$

* Received by the editors February 15, 1990; accepted for publication (in revised form) January 21, 1991. This research was supported in part by National Institute of Health grant GM29123.

† Department of Mathematics, University of Utah, Salt Lake City, Utah 84112.

‡ Present address, Institute of Computational Fluid Dynamics, 1-22-3 Haramachi, Meguro-ku, Tokyo 152, Japan.

Here $x_i \in \mathbb{R}^{n_i}$, $i = 1, \cdots, N$, $x_0 \in \mathbb{R}^{N_2}$, B is an $N_2 \times N_2$ matrix, and C_i is an $N_2 \times n_i$ matrix, and thus (1.4) is a system in $\mathbb{R}^{N_2+M}$, where $M = \sum_{i=1}^{N} n_i$. We assume that each n_i-dimensional system (1.2) has a nonconstant periodic solution (1.3) with least period $T_i > 0$, and we make the following assumptions about (1.4).

ASSUMPTION 1.1. (a) $\epsilon = \epsilon_0 \delta^{-p}$ *for* $\epsilon_0 > 0, \delta \geq 0$, *and* $p \in (0,1)$.

(b) *For* $i = 1, \cdots, N$,

$$f_i : \mathbb{R}^{n_i} \longrightarrow \mathbb{R}^{n_i}, \qquad g_i : \mathbb{R}^{N_2+M+1} \longrightarrow \mathbb{R}^{n_i}$$

are $(k+1)$*-times continuously differentiable with* $k \geq 2$.

(c) *For each* $i = 1, \cdots, N$, 1 *is a simple multiplier of*

$$\frac{dx_i}{dt} = Df_i(\eta_i(t))x_i,$$

and the remaining $n_i - 1$ *multipliers have modulus less than* 1.

(d) *The eigenvalues of* B *have negative real parts.*

In Part I we studied the persistence of invariant tori for the singular ($p > 1$) and regular ($p = 1$) cases, where standard results could be used to prove the existence of invariant tori for small coupling that lie near certain invariant tori that exist in the uncoupled system. In this paper we study the degenerate case $p \in (0,1)$, and as we shall see shortly, this case is far more difficult due to the fact that the coupling vanishes faster than the capacitance.

The M-dimensional system, which consists of (1.2) for $i = 1, \cdots, N$, has the N-dimensional invariant torus T_0^N in $\mathbb{R}^M$ defined by

$$x_i = \eta_i(\theta_i), \qquad i = 1, \cdots, N. \tag{1.5}$$

Therefore, when $0 < p < 1$, $T_0^N \times \mathbb{R}^{N_2}$ is an invariant manifold of the unperturbed system ($\delta = 0$) of (1.4) that consists of the N_2-parameter family of N-dimensional invariant tori $\mathcal{T}_{0,c}^N$,

$$\begin{aligned} x_0 &= c, \qquad c \in \mathbb{R}^{N_2}, \\ x_i &= \eta_i(\theta_i), \qquad i = 1, \cdots, N. \end{aligned} \tag{1.6}$$

The simplest example arises for $N = N_2 = 1, n_1 = 2$, in which case the invariant manifold at $\delta = 0$ is the cylinder $\eta \times \mathbb{R} \subset \mathbb{R}^3$.

The problem we address in this paper is that of determining which, if any, of the invariant tori in (1.6) persist for $\delta > 0$. The standard results [4], [9] on persistence of invariant manifolds for flows are not applicable to the degenerate problem that arises from (1.4) at $\delta = 0$, because the compactness of the unperturbed invariant manifold is essential for the proofs. In fact, the invariant manifold (1.6) that exists for the unperturbed problem does not persist, and the question is which of the invariant tori persists. Thus the problem is in essence a bifurcation problem. We address this problem in §3, where we show that the invariant torus given by

$$\begin{aligned} x_0 &= -B^{-1} \sum_{i=1}^{N} C_i \frac{1}{T_i} \int_0^{T_i} \eta_i(s)\, ds, \\ x_i &= \eta_i(\theta_i), \qquad i = 1, \cdots, N \end{aligned} \tag{1.7}$$

persists for all small $\delta > 0$. That is, a one-parameter family of invariant tori bifurcates from (1.6) under Assumption 1.1.

In the following section we show that, depending on the value of p, (1.4) can be transformed into one of the following systems of generalized phase-amplitude equations.

(a) $1 - 1/k \leq p < 1$,

$$\begin{aligned} \frac{d\sigma}{dt} &= \omega + \mathcal{S}(\sigma, w, z, \mu, \rho), \\ \frac{dw}{dt} &= A(\Omega^{-1}\sigma)w + \mathcal{W}(\sigma, w, z, \mu, \rho), \\ \frac{dz}{dt} &= \rho\mu\left[Qz + \mathcal{Z}(\sigma, w, z, \mu, \rho)\right]. \end{aligned} \tag{1.8}$$

(b) $0 < p < 1$,

$$\begin{aligned} \frac{d\sigma}{dt} &= \omega + \mathcal{S}(\sigma, w, z, \mu), \\ \frac{dw}{dt} &= A(\Omega^{-1}\sigma)w + \mathcal{W}(\sigma, w, z, \mu), \\ \frac{dz}{dt} &= \mu\left[Qz + \mathcal{Z}(\sigma, w, z, \mu)\right]. \end{aligned} \tag{1.9}$$

Here σ and w are scaled variables in the phase-amplitude coordinates and z is the deviation of x_0 from the average over T_0^N; thus $\sigma \in \mathbb{R}^N$, $w \in \mathbb{R}^{N_1}$ with $N_1 = M - N$, $z \in \mathbb{R}^{N_2}$. The parameters μ and ρ arise from a scaling of δ (cf. (2.9) and (2.10)), and $\mu \longrightarrow 0$ as $\delta \longrightarrow 0$. In both cases the invariant tori correspond to invariant manifolds of (1.8) and (1.9), and in both cases they are smooth in the natural parameterization for δ fixed. Furthermore, in the first case the tori are also smooth in a fractional power of δ.

The existence and the smoothness of invariant tori are established in §3 where we construct invariant manifolds of (1.8) and (1.9) that correspond to invariant tori of (1.4). This is done as follows. Let I be an open interval. For (1.8) we look for a pair of smooth mappings $q_\rho : \mathbb{R}^N \times I \longrightarrow \mathbb{R}^{N_1}$, $r_\rho : \mathbb{R}^N \times I \longrightarrow \mathbb{R}^{N_2}$, for some ρ such that for each $\mu \in I$, q_ρ and r_ρ are 2π-periodic in each component of $\sigma \in \mathbb{R}^N$, and the manifold defined by $w = q_\rho(\sigma, \mu)$, $z = r_\rho(\sigma, \mu)$ is an invariant manifold of (1.8). Similarly, for (1.9) we look for smooth mappings $q_\mu : \mathbb{R}^N \longrightarrow \mathbb{R}^{N_1}$, $r_\mu : \mathbb{R}^N \longrightarrow \mathbb{R}^{N_2}$, for all small $\mu > 0$ such that q_μ and r_μ are 2π-periodic in each component of $\sigma \in \mathbb{R}^N$, and the manifold defined by $w = q_\mu(\sigma)$, $z = r_\mu(\sigma)$ is an invariant manifold of (1.9). Such pairs are called integral manifolds in [5] and [6], and as we show in §3, these give rise to invariant tori of (1.4). Furthermore, the smoothness of the invariant tori is determined by the smoothness of the integral manifolds. In particular, integral manifolds that are smooth in the natural parameterization give rise to smooth invariant manifolds, and those that are also smooth in μ give rise to smooth invariant manifolds that vary smoothly with δ.

In §4 we study the asymptotic behavior of solutions of (1.8) in a neighborhood of the invariant manifold $w = q_\rho(\sigma, \mu)$, $z = r_\rho(\sigma, \mu)$. It is shown that for sufficiently

small $\mu > 0$ and $\rho > 0$, there is a neighborhood of the manifold in which solutions of (1.8) are attracted to the manifold at an exponential rate in time. We state and prove a similar result for (1.9). These results completely describe the behavior of solutions of (1.4) in a neighborhood of the perturbed torus. That is, for each small $\delta > 0$, there is a neighborhood of the perturbed torus in which solutions of (1.4) are attracted to it at an exponential rate in time. These neighborhoods shrink to (1.7) as $\delta \longrightarrow 0$ and the rates at which solutions are attracted to the perturbed tori tend to zero as $\delta \longrightarrow 0$.

The smoothness of integral manifolds in related systems was first studied carefully by Kelly in [11], where systems of the following type are analyzed.

$$\begin{aligned} \frac{d\theta}{dt} &= a + \epsilon\Theta(\theta, x, z, \epsilon), \\ \frac{dx}{dt} &= Ax + X(\theta, x, z, \epsilon), \\ \frac{dz}{dt} &= Cz + Z(\theta, x, z, \epsilon). \end{aligned} \tag{1.10}$$

The basic assumption made there is the following.

ASSUMPTION 1.2. (a) *A and C are constant square matrices in real canonical form, A has eigenvalues with negative real parts, C has eigenvalues with positive real parts, a is a constant vector.*

(b) *Θ, X, and Z are defined and k-times continuously differentiable in*

$$N_\delta = \{(\theta, x, z, \epsilon) : \theta \text{ arbitrary}, \|x\| + \|z\| + |\epsilon| < \delta\},$$

and ω_i-periodic in θ_i.

(c) *θ, X, and Z, and the Jacobian of X and Z vanish when $(x, z, \epsilon) = 0$.*

Under this assumption it is shown that (1.10) has an integral manifold $x = u(\theta, \epsilon)$, $z = w(\theta, \epsilon)$ which is k-times continuously differentiable. Note that the matrix A in (1.10) is a constant matrix, whereas we have studied in [15] and in this paper the existence and smoothness of integral manifolds of similar systems in which the counterpart is a function of one of the dependent variables. Furthermore, (1.10) does not degenerate at $\epsilon = 0$.

Systems related to (1.8), (1.9), or (2.7) are studied in [3], [5]–[7], [11], and [14]. For example, in [3] Diliberto considers the following systems:

$$\begin{aligned} \frac{d\theta}{dt} &= l + \Theta_1(\theta, y, \epsilon) + \Theta_2(\theta, \epsilon), \\ \frac{dy}{dt} &= B(\theta, \epsilon)y + a(\theta, \epsilon) + Y(\theta, y, \epsilon)y, \end{aligned} \tag{1.11}$$

$$\begin{aligned} \frac{d\theta}{dt} &= l + \epsilon\left[\Theta_1(\theta, y, \epsilon) + \Theta_2(\theta, \epsilon)\right], \\ \frac{dy}{dt} &= \epsilon\left[B(\theta, \epsilon)y + a(\theta, \epsilon) + Y(\theta, y, \epsilon)y\right], \end{aligned} \tag{1.12}$$

under the following assumption.

ASSUMPTION 1.3. (a) $\theta \in \mathbb{R}^N$, $y \in \mathbb{R}^M$, *and* l *is a constant vector.* $\Theta_1(\theta, y, \epsilon)$, $\Theta_2(\theta, \epsilon)$, $B(\theta, \epsilon)$, $a(\theta, \epsilon)$, *and* $Y(\theta, y, \epsilon)$ *are* ω_i*-periodic in* θ_i, *continuously differentiable in* (θ, y), *and continuous in* (θ, y, ϵ).

(b) $Y(\theta, 0, \epsilon) \equiv 0$ *and* $a(\theta, \epsilon) \longrightarrow 0$ *as* $\epsilon \longrightarrow 0$.

(c) *There is* $b > 0$ *such that* $(B(\theta, 0)y, y) \geq b\|y\|_2^2$ *for all* $y \in \mathbb{R}^M$, *and* $\lambda(\theta) < b$ *for all* θ, *where* $\lambda(\theta)$ *is an eigenvalue of*

$$\tfrac{1}{2}[\Omega(\theta) + \Omega^T(\theta)]$$

with

$$\Omega(\theta) = \frac{\partial \Theta_2}{\partial \theta}(\theta, 0).$$

It is shown that under this assumption (1.11) or (1.12) has an integral manifold $y = S(\theta, \epsilon)$, where S is continuous and $S(\theta, \epsilon) \longrightarrow 0$ as $\epsilon \longrightarrow 0$. The system at (1.12) degenerates at $\epsilon = 0$, as do the systems studied here, but the former lacks the normal component that appears in our equations. Furthermore, hypotheses like those in Assumption 1.3 are difficult to check in the original form (1.4) of the equations we study.

Systems similar to (1.8) or (1.9) have also been studied by Hale. The existence of Lipschitz continuous integral manifolds is established in [5] for the system

$$\begin{aligned} \frac{d\theta}{dt} &= d(\epsilon) + \Theta(t, \theta, y, x, \epsilon), \\ \frac{dy}{dt} &= Ay + Y(t, \theta, y, x, \epsilon), \\ \frac{dz}{dt} &= \epsilon C z + \epsilon Z(t, \theta, y, z, \epsilon). \end{aligned} \tag{1.13}$$

Similar results are obtained in [6] and [7]. However, the matrix A in (1.13) is a constant matrix, whereas the counterparts which appear in (1.8) and (1.9) are functions of one of the dependent variables. Moreover, some of the technical assumptions made for (1.13) are not met by the systems related to our degenerate problem, and therefore we cannot apply the existence result in [5] directly. Finally, the smoothness of the manifolds with respect to parameters is not studied in detail there, but smoothness with respect to parameters is important when investigating the flow on the manifolds.

2. Reduction of the equations. In this section we introduce transformations that convert (1.4) to suitable forms for the degenerate case. We first review the transformations introduced in [16]. For each $i = 1, \cdots, N$, there is an $n_i \times (n_i - 1)$-matrix $\Phi_i(\theta_i)$ each entry of which is a $(k+1)$-times continuously differentiable function of θ_i, with the following properties:

$$\begin{aligned} \Phi_i(\theta_i + T_i) &= \Phi_i(\theta_i), \\ \Phi_i(\theta_i)^T \Phi_i(\theta_i) &= I_{(n_i-1)\times(n_i-1)}, \\ f_i(\eta_i(\theta_i))^T \Phi_i(\theta_i) &= O_{1\times(n_i-1)} \end{aligned}$$

for all $\theta_i \in \mathbb{R}$. Then, for each x_i in a neighborhood of the orbit of (1.3), there are $\theta_i \in \mathbb{R}$ and $y_i \in \mathbb{R}^{n_i-1}$ such that

$$\eta_i(\theta_i) + \Phi_i(\theta_i)y_i - x_i = 0. \tag{2.1}$$

Equation (2.1) defines a transformation between a neighborhood of the torus T_0^N and $\mathbb{R} \times \mathcal{G}$ where $\mathcal{G}$ is a neighborhood of the origin in $\mathbb{R}^{N_1}$. Thus we can convert (1.4) to the system for θ, y, and x_0, where

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_N \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad y_i \in \mathbb{R}^{n_i-1}, \quad i = 1, \cdots, N. \tag{2.2}$$

We can write x_0 as the average plus a deviation by defining $a : \mathbb{R}^N \times \mathbb{R} \longrightarrow \mathbb{R}^{N_2}$,

$$a(\theta, \mu) = \begin{cases} \mu \sum_{i=1}^{N} \left[I - e^{\mu B T_i}\right]^{-1} \int_0^{T_i} e^{-\mu B(s-T_i)} C_i \eta_i(\theta_i + s)\, ds, & \mu \neq 0, \\ -B^{-1} \sum_{i=1}^{N} C_i \dfrac{1}{T_i} \int_0^{T_i} \eta_i(s)\, ds, & \mu = 0, \end{cases} \tag{2.3}$$

and setting

$$x_0 = a(\theta, \epsilon\delta) + z. \tag{2.4}$$

Finally, define an N-dimensional vector ω and an $N \times N$-diagonal matrix Ω by

$$\omega = \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_N \end{pmatrix} \text{ and } \Omega = \begin{bmatrix} \omega_1 & & \\ & \ddots & \\ & & \omega_N \end{bmatrix}, \text{ where } \omega_i = \frac{2\pi}{T_i},\ i = 1, \cdots, N, \tag{2.5}$$

and let

$$\sigma = \Omega\theta. \tag{2.6}$$

Using (2.1), (2.4), and (2.6), we obtain from (1.4) the following system of ordinary differential equations for σ, y, and z.

$$\begin{aligned} \frac{d\sigma}{dt} &= \omega + S(\sigma, y, z, \delta, \epsilon\delta), \\ \frac{dy}{dt} &= A(\Omega^{-1}\sigma)y + Y(\sigma, y, z, \delta, \epsilon\delta), \\ \frac{dz}{dt} &= \epsilon\delta \left[Bz + Z(\sigma, y, z, \delta, \epsilon\delta)\right]. \end{aligned} \tag{2.7}$$

Here $A(\theta)$ is the $N_1 \times N_1$ matrix defined by

$$A(\theta) = \bigoplus_{i=1}^{N} A_i(\theta_i),$$

where $A_i(\theta_i)$, $i = 1, \cdots, N$, is the $(n_i - 1) \times (n_i - 1)$ matrix defined by

$$A_i(\theta_i) = \Phi_i(\theta_i)^T [Df_i(\eta_i(\theta_i))\Phi_i(\theta_i) - \Phi_i'(\theta_i)].$$

S, Y, and Z are defined by

$$\begin{aligned}
S(\sigma, y, z, \delta, \mu) &= \Omega\Theta(\Omega^{-1}\sigma, y, z + a(\Omega^{-1}\sigma, \mu), \delta), \\
Y(\sigma, y, z, \delta, \mu) &= \mathcal{Y}(\Omega^{-1}\sigma, y, z + a(\Omega^{-1}\sigma, \mu), \delta), \\
Z(\sigma, y, z, \delta, \mu) &= C(\Omega^{-1}\sigma)y - b(\Omega^{-1}\sigma, \mu)\Theta(\Omega^{-1}\sigma, y, z + a(\Omega^{-1}\sigma, \mu), \delta),
\end{aligned}$$

where

$$\begin{aligned}
\Theta_i(\theta, y, x_0, \delta) = {} & \frac{f_i(\eta_i(\theta_i))^T f_i(\eta_i(\theta_i) + \Phi_i(\theta_i)y_i)}{f_i(\eta_i(\theta_i))^T [f_i(\eta_i(\theta_i)) + \Phi_i'(\theta_i)y_i]} - 1 \\
& + \delta \frac{f_i(\eta_i(\theta_i))^T g_i(x_0, \eta_1(\theta_1) + \Phi_1(\theta_1)y_1, \cdots, \eta_N(\theta_N) + \Phi_N(\theta_N)y_N, \delta)}{f_i(\eta_i(\theta_i))^T [f_i(\eta_i(\theta_i)) + \Phi_i'(\theta_i)y_i]}, \\
\mathcal{Y}_i(\theta, y, x_0, \delta) = {} & \Phi_i(\theta_i)^T [f_i(\eta_i(\theta_i) + \Phi_i(\theta_i)y_i) \\
& \qquad - Df_i(\eta_i(\theta_i))\Phi_i(\theta_i)y_i - \Theta_i(\theta, y, x_0, \delta)\Phi_i'(\theta_i)y_i] \\
& + \delta \Phi_i(\theta_i)^T g_i(x_0, \eta_1(\theta_1) + \Phi_1(\theta_1)y_1, \cdots, \eta_N(\theta_N) + \Phi_N(\theta_N)y_N, \delta),
\end{aligned}$$

and $\Theta(\theta, y, x_0, \delta) = (\Theta_1(\theta, y, x_0, \delta), \cdots, \Theta_N(\theta, y, x_0, \delta))^T$ and $\mathcal{Y}(\theta, y, x_0, \delta) = (\mathcal{Y}_1(\theta, y, x_0, \delta), \cdots, \mathcal{Y}_N(\theta, y, x_0, \delta))^T$. Finally, $C(\theta)$ is the $N_2 \times N_1$-matrix defined by

$$C(\theta) = [C_1\Phi_1(\theta_1) \cdots C_N\Phi_N(\theta_N)],$$

and $b(\theta, \mu)$ is the $N_2 \times N$-matrix defined by

$$\mu b(\theta, \mu) = \frac{\partial a}{\partial \theta}(\theta, \mu).$$

Under Assumption 1.1(b), each entry of $A(\theta)$ is a k-times continuously differentiable function of θ and a T_i-periodic function of θ_i for each $i = 1, \cdots, N$. It follows that each entry of $A(\Omega^{-1}\sigma)$ is a k-times continuously differentiable function of σ and 2π-periodic in each component of σ. Under Assumption 1.1(c), if $\Psi_i(\theta)$ is a fundamental matrix solution of

$$\frac{dy_i}{d\theta} = A_i(\theta)y_i,$$

then there are positive constants α_i and H_i such that

$$\|\Psi_i(\theta)\Psi_i^{-1}(\theta_0)\| \leq H_i e^{-\alpha_i(\theta - \theta_0)} \quad \text{for } \theta \geq \theta_0.$$

Here and hereafter we define the norm $\|x\|$ of $x \in \mathbb{R}^n$ by

$$\|x\| = \sum_{i=1}^{n} |x_i|,$$

where x_i is the ith component of x. The functions $S : \mathbb{R}^N \times \mathcal{G} \times \mathbb{R}^{N_2} \times \mathbb{R}^2 \longrightarrow \mathbb{R}^N$, $Y : \mathbb{R}^N \times \mathcal{G} \times \mathbb{R}^{N_2} \times \mathbb{R}^2 \longrightarrow \mathbb{R}^{N_1}$, and $Z : \mathbb{R}^N \times \mathcal{G} \times \mathbb{R}^{N_2} \times \mathbb{R}^2 \longrightarrow \mathbb{R}^{N_2}$ are k-times continuously differentiable. Moreover, they are 2π-periodic in each component of σ, and

$$\begin{aligned} S(\sigma, y, z, \delta, \mu) &\sim \mathcal{O}(\|y\| + |\delta|), \\ Y(\sigma, y, z, \delta, \mu) &\sim \mathcal{O}(\|y\|^2 + |\delta|), \\ Z(\sigma, y, z, \delta, \mu) &\sim \mathcal{O}(\|y\| + |\delta|). \end{aligned} \tag{2.8}$$

Remark 1. The Floquet reduction could be used to transform (2.7) further to a system in which the second equation becomes a perturbation of a linear system with constant coefficients. However, the periodicity in some of the components of σ might have to be doubled in the vector field obtained by the reduction, if the reduced system is to have real variables. In particular, the period in σ_i is doubled when the linear system

$$\frac{dy_i}{d\theta_i} = A_i(\theta_i) y_i$$

has a real negative multiplier. This reduction leads to no simplification in the analysis and is not used here.

Next we introduce scalings for y and δ that depend on p.

$$1 - \frac{1}{k} \leq p < 1 \quad : \quad y = \rho^{\kappa} \mu^{\nu} w, \quad \delta = \rho^{\nu} \mu^{\nu}, \quad 1 < \kappa < \nu, \tag{2.9}$$

$$0 < p < 1 \quad : \quad y = \mu^{\zeta} w, \quad \delta = \mu^{\nu}, \quad 1 < \zeta < \nu, \tag{2.10}$$

where

$$\nu = \frac{1}{1-p}. \tag{2.11}$$

Using (2.9) and (2.10), we obtain from (2.7) the following systems of equations (2.12) and (2.13), respectively.

$$\begin{aligned} \frac{d\sigma}{dt} &= \omega + S\left(\sigma, \rho^{\kappa} \mu^{\nu} w, z, \rho^{\nu} \mu^{\nu}, \epsilon_0 \rho \mu\right), \\ \frac{dw}{dt} &= A(\Omega^{-1}\sigma) w + \rho^{-\kappa} \mu^{-\nu} Y\left(\sigma, \rho^{\kappa} \mu^{\nu} w, z, \rho^{\nu} \mu^{\nu}, \epsilon_0 \rho \mu\right), \\ \frac{dz}{dt} &= \rho \mu \left[\epsilon_0 B z + \epsilon_0 Z\left(\sigma, \rho^{\kappa} \mu^{\nu} w, z, \rho^{\nu} \mu^{\nu}, \epsilon_0 \rho \mu\right)\right]. \end{aligned} \tag{2.12}$$

$$
\begin{aligned}
\frac{d\sigma}{dt} &= \omega + S\left(\sigma, \mu^{\zeta} w, z, \mu^{\nu}, \epsilon_0 \mu\right), \\
\frac{dw}{dt} &= A(\Omega^{-1}\sigma)w + \mu^{-\zeta} Y\left(\sigma, \mu^{\zeta} w, z, \mu^{\nu}, \epsilon_0 \mu\right), \\
\frac{dz}{dt} &= \mu\left[\epsilon_0 B z + \epsilon_0 Z\left(\sigma, \mu^{\zeta} w, z, \mu^{\nu}, \epsilon_0 \mu\right)\right].
\end{aligned} \tag{2.13}
$$

We have chosen the scalings (2.9) and (2.10) in order for the vector fields in (2.12) and (2.13) to satisfy the following conditions. In view of (2.8), the right-hand side of the first equation in (2.12) consists of ω and the term whose derivatives with respect to σ, w, z, and μ of order less than or equal to k are of the order $o(\rho)$ as $\rho \longrightarrow 0$. The right-hand side of the second equation and the terms inside the brackets of the third equation consist of the linear terms and the terms whose derivatives tend to zero as $\rho \longrightarrow 0$. The derivatives of the nonlinear terms in (2.13) with respect to σ, w, and z satisfy similar order estimates with respect to μ.

Equations (2.12) and (2.13) are special cases of the equations given at (1.8) and (1.9), respectively, and as was indicated above, they satisfy the conditions stated in Assumption 2.1 below. Hereafter we analyze (1.8) and (1.9) under Assumption 2.1, but we note that (2.12) satisfies Assumption 2.1 (a)–(e), while (2.13) satisfies Assumption 2.1 (a)–(c) and (f).

ASSUMPTION 2.1. (a) *$A(\theta)$ is an $N_1 \times N_1$-matrix that has the form*

$$A(\theta) = \bigoplus_{i=1}^{N} A_i(\theta_i),$$

where for each i, $A_i(\theta_i)$ is an $(n_i - 1) \times (n_i - 1)$-matrix each entry of which is a k-times continuously differentiable T_i-periodic function of θ_i and there are $\alpha_i, H_i > 0$ such that if $\Psi_i(\theta)$ is a fundamental matrix solution of

$$\frac{dy_i}{d\theta} = A_i(\theta) y_i,$$

there are positive constants α_i and H_i such that

$$\|\Psi_i(\theta)\Psi_i^{-1}(\theta_0)\| \leq H_i e^{-\alpha_i(\theta-\theta_0)} \quad \text{for } \theta \geq \theta_0.$$

(b) *The spectrum of Q lies in the left-half complex plane. Thus there are positive constants β_2 and K_2 such that for $t \geq s$,*

$$\left\| e^{Q(t-s)} \right\| \leq K_2 e^{-\beta_2(t-s)}.$$

(c) *The N-dimensional vector ω and the $N \times N$-matrix Ω are defined by* (2.5).

(d) *There are neighborhoods F_1 and F_2 of the origin in $\mathbb{R}^{N_1}$ and $\mathbb{R}^{N_2}$, respectively, and open intervals I and I_1 such that if $E = \mathbb{R}^N \times F_1 \times F_2 \times I$,*

$$
\begin{aligned}
\mathcal{S} &: E \times I_1 \longrightarrow \mathbb{R}^N, \\
\mathcal{W} &: E \times I_1 \longrightarrow \mathbb{R}^{N_1}, \\
\mathcal{Z} &: E \times I_1 \longrightarrow \mathbb{R}^{N_2}.
\end{aligned}
$$

For each $\rho \in I_1$, $\mathcal{S}$, $\mathcal{W}$, *and* $\mathcal{Z}$ *are* 2π*-periodic in each component of* σ *and* k*-times continuously differentiable in* E. *There are* $\lambda(\rho)$ *and* $\lambda_i(\rho)$, $i = 1, 2$, *such that if* $0 \leq m_1 + m_2 + m_3 + m_4 \leq k$

$$\left\| \frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{S}}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}}(\sigma, w, z, \mu, \rho) \right\| \leq \lambda(\rho)\mu^{k-m_4},$$

$$\left\| \frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{W}}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}}(\sigma, w, z, \mu, \rho) \right\| \leq \lambda_1(\rho),$$

$$\left\| \frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{Z}}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}}(\sigma, w, z, \mu, \rho) \right\| \leq \lambda_2(\rho)\mu^{k-m_4}$$

for all $(\sigma, w, z, \mu) \in E$, *and*

$$\frac{\lambda(\rho)}{\rho} \longrightarrow 0,$$
$$\lambda_i(\rho) \longrightarrow 0, \qquad i = 1, 2$$

as $\rho \longrightarrow 0$.

(e)

$$\mathcal{W}(\sigma, 0, 0, \mu, \rho) \longrightarrow 0, \quad \mathcal{Z}(\sigma, 0, 0, \mu, \rho) \longrightarrow 0,$$

as $\mu \longrightarrow 0$ *for all* $\sigma \in \mathbb{R}^N$ *and* $\rho \in I_1$.

(f) *There are neighborhoods* F_1 *and* F_2 *of the origin in* $\mathbb{R}^{N_1}$ *and* $\mathbb{R}^{N_2}$, *respectively, and an open interval* $I = (0, \mu_0)$ *such that if* $E = \mathbb{R}^N \times F_1 \times F_2$,

$$\mathcal{S} : E \times I \longrightarrow \mathbb{R}^N, \quad \mathcal{W} : E \times I \longrightarrow \mathbb{R}^{N_1}, \quad \mathcal{Z} : E \times I \longrightarrow \mathbb{R}^{N_2}.$$

For each $\mu \in I$, $\mathcal{S}$, $\mathcal{W}$, *and* $\mathcal{Z}$ *are* 2π*-periodic each component of* σ *and* k*-times continuously differentiable in* E, *and there are* $\lambda(\mu)$ *and* $\lambda_i(\mu)$, $i = 1, 2$, *such that if* $0 \leq m_1 + m_2 + m_3 \leq k$,

$$\left\| \frac{\partial^{m_1+m_2+m_3}\mathcal{S}}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}}(\sigma, w, z, \mu) \right\| \leq \lambda(\mu),$$

$$\left\| \frac{\partial^{m_1+m_2+m_3}\mathcal{W}}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}}(\sigma, w, z, \mu) \right\| \leq \lambda_1(\mu),$$

$$\left\| \frac{\partial^{m_1+m_2+m_3}\mathcal{Z}}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}}(\sigma, w, z, \mu) \right\| \leq \lambda_2(\mu)$$

for all $(\sigma, w, x) \in E$, *and*

$$\frac{\lambda(\mu)}{\mu}, \lambda_1(\mu), \frac{\lambda_2(\mu)}{\mu} \longrightarrow 0$$

as $\mu \longrightarrow 0$.

Remark 2. In what follows we assume without loss of generality that $I \subseteq (-1, 1)$. We also assume that $0 \in I$ whenever the statement $\mu \longrightarrow 0$ appears.

We summarize the results concerning the transformations for the degenerate problem in the following proposition.

PROPOSITION 2.2. *Suppose that the conditions stated in Assumption* 1.1(a), (b), *and* (d) *are satisfied with* $0 < p < 1$. *Using the transformations defined by* (2.1), (2.4), *and* (2.6), *we obtain from* (1.4) *the system of ordinary differential equations* (2.7) *for* σ, y, *and* z. *Further, when* $1 - 1/k \leq p < 1$, *we obtain* (2.12) *from* (2.7) *using* (2.9). *Equation* (2.12) *has the form* (1.8) *and satisfies the conditions stated in Assumption* 2.1(a)–(e). *When* $0 < p < 1$, *we obtain* (2.13) *from* (2.7) *using* (2.10). *Equation* (2.13) *has the form* (1.9) *and satisfies the conditions stated in Assumption* 2.1(a)–(c) *and* (f).

We recover the solutions of (1.4) from the solutions of (2.12) and (2.13) by the transformations (2.1), (2.4), (2.6), (2.9), and (2.10). Therefore integral manifolds of (2.12) and (2.13) generate invariant tori via these transformations. We summarize the relation between the integral manifolds of (2.13) and invariant tori of (1.4) in the following proposition.

PROPOSITION 2.3. (a) *If the pair* $w = q_\rho(\sigma, \mu)$, $z = r_\rho(\sigma, \mu)$ *is an integral manifold of* (2.12), *then*

$$
\begin{aligned}
x_i &= \eta_i(\omega_i^{-1}\sigma_i) + \rho^\kappa \mu^\nu \Phi_i(\omega_i^{-1}\sigma_i) q_{\rho,i}(\sigma, \mu), \qquad i = 1, \cdots, N, \\
x_0 &= r_\rho(\sigma, \mu) + a(\Omega^{-1}\sigma, \epsilon_0 \rho \mu), \\
\mu &= \frac{\delta^{1-p}}{\rho}
\end{aligned}
\tag{2.14}
$$

is an invariant torus of (1.4) *for* $1 - 1/k \leq p < 1$.

(b) *If the pair* $w = q_\mu(\sigma)$, $z = r_\mu(\sigma)$ *is an integral manifold of* (2.13), *then*

$$
\begin{aligned}
x_i &= \eta_i(\omega_i^{-1}\sigma_i) + \mu^\zeta \Phi_i(\omega_i^{-1}\sigma_i) q_{\mu,i}(\sigma), \qquad i = 1, \cdots, N, \\
x_0 &= r_\mu(\sigma) + a(\Omega^{-1}\sigma, \epsilon_0 \mu), \\
\mu &= \delta^{1-p}
\end{aligned}
\tag{2.15}
$$

is an invariant torus of (1.4) *for* $0 < p < 1$.

$q_{\rho,i}(\sigma, \mu)$ *and* $q_{\mu,i}(\sigma)$ *are the components of* $q_\rho(\sigma, \mu)$ *and* $q_\mu(\sigma)$, *respectively, which correspond to components* y_i *of* y *in* (2.2).

Remark 3. In view of (2.3), the one-parameter families of invariant tori given in Proposition 2.3 bifurcate from (1.7) when the integral manifolds tend to zero as $\mu \longrightarrow 0$.

The existence of integral manifolds for (1.8) and (1.9) under Assumption 2.1 will be established in §3.

3. Existence of invariant tori in the perturbed system. In this section we prove the existence and smoothness of integral manifolds for (1.8) under Assumption 2.1(a)–(e). We also prove their existence and smoothness for (1.9) under Assumption 2.1(a)–(c) and (f). The technique which we use to construct the integral manifolds is an extension of those used in [5], [3], and [11]. The construction is done as follows. We define a map on a closed subset of a suitable Banach space of functions, and show that the map has a fixed point. This fixed point is an integral manifold of (1.8) or (1.9), as the case may be.

Let $\mathcal{PC}^k(\mathbb{R}^m \times I, \mathbb{R}^n)$ be the set of all k-times continuously differentiable mappings $s : \mathbb{R}^m \times I \longrightarrow \mathbb{R}^n$ endowed with the norm

$$\|s\|_k = \max_{0 \le i+j \le k} \sup \left\{ \left\| \frac{\partial^{i+j} s}{\partial \sigma^i \partial \mu^j}(\sigma, \mu) \right\| : \sigma \in \mathbb{R}^m, \mu \in I \right\} < \infty$$

and such that

$$s(\sigma + 2\pi e^i, \mu) = s(\sigma, \mu) \quad \text{for all } \sigma \in \mathbb{R}^m \text{ and } i = 1, \cdots, m,$$

where e^i is the ith unit coordinate vector in $\mathbb{R}^m$. Then $\mathcal{PC}^k(\mathbb{R}^m \times I, \mathbb{R}^n)$ is a Banach space (cf. [2], [11]). Let $\mathcal{B}^k_\Delta(I, m, n)$ be the closed ball in $\mathcal{PC}^k(\mathbb{R}^m \times I, \mathbb{R}^n)$ with radius Δ and center at the origin. For (1.8), let $\phi(t, \sigma^0, \mu, q, r, \rho)$ be the solution of

$$\frac{d\sigma}{dt} = \omega + \mathcal{S}(\sigma, q(\sigma, \mu), r(\sigma, \mu), \mu, \rho), \qquad \phi(0, \sigma^0, \mu, q, r, \rho) = \sigma^0, \tag{3.1}$$

where $(q, r) \in \mathcal{B}^k_\Delta(I, N, N_1) \times \mathcal{B}^k_\Delta(I, m, n)$. We will choose I, Δ, and ρ in such a way that this definition makes sense. Then let $\Lambda(t, t_0, \sigma^0, \mu, q, r, \rho)$ be the fundamental matrix solution of

$$\frac{dw}{dt} = A(\Omega^{-1}\phi(t, \sigma^0, \mu, q, r, \rho))w, \qquad \Lambda(t_0, t_0, \sigma^0, \mu, q, r, \rho) = I_{N_1 \times N_1}. \tag{3.2}$$

The properties of $\phi(t, \sigma^0, \mu, q, r, \rho)$ and $\Lambda(t, t_0, \sigma^0, \mu, q, r, \rho)$ are studied in the Appendix. By Lemmas 5.1–5.11, which we prove in the Appendix, $\phi(t, \sigma^0, \mu, q, r, \rho)$ and $\Lambda(t, t_0, \sigma^0, \mu, q, r, \rho)$ are k-times continuously differentiable with respect to t, t_0, σ^0, and μ, and have the following properties.

PROPERTY 3.1. There are positive constants ρ_1, $K_{l,m}$, $\widehat{K}_{l,m}$, $C_{l,m}$, $\widehat{C}_{l,m}$, H, and β_1 such that for $0 < \rho \le \rho_1$ and $s \le 0$ the following holds.

(a)

$$\left\| \frac{\partial \phi}{\partial \sigma^0}(s, \sigma^0, \mu, q, r, \rho) \right\| \le e^{-\lambda(\rho)\mu^k(1+2\Delta)s}.$$

(b) For $2 \le l + 2m$ and $l + m \le k$,

$$\left\| \frac{\partial^{l+m}\phi}{\partial \sigma^0_{i_1} \cdots \partial \sigma^0_{i_l} \partial \mu^m}(s, \sigma^0, \mu, q, r, \rho) \right\|$$

$$\le K_{l,m}\mu^{-m} \left\{ \sum_{i=1}^{l+2m-1} \left[\lambda(\rho)\mu^k(-s)\right]^i \right\} e^{-(l+m)\lambda(\rho)\mu^k(1+2\Delta)s}.$$

(c) For $0 \le l + m \le k - 1$,

$$\left\| \frac{\partial^{l+m}\phi}{\partial \sigma^0_{i_1} \cdots \partial \sigma^0_{i_l} \partial \mu^m}(s, \sigma^0, \mu, q_1, r_1, \rho) - \frac{\partial^{l+m}\phi}{\partial \sigma^0_{i_1} \cdots \partial \sigma^0_{i_l} \partial \mu^m}(s, \sigma^0, \mu, q_2, r_2, \rho) \right\|$$

$$\le \widehat{K}_{l,m} \left(\|q_1 - q_2\|_{l+m} + \|r_1 - r_2\|_{l+m} \right)$$

$$\times \mu^{-m} \left\{ \sum_{i=1}^{l+2m+1} \left[\lambda(\rho)\mu^k(-s)\right]^i \right\} e^{-(l+m+1)\lambda(\rho)\mu^k(1+2\Delta)s}.$$

(d)

$$\|\Lambda(0,s,\sigma^0,\mu,q,r,\rho)\| \le K_1 e^{[\beta_1 - H\lambda(\rho)\mu^k]s}.$$

(e) For $1 \le l+m \le k$,

$$\left\|\frac{\partial^{l+m}\Lambda}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(0,s,\sigma^0,\mu,q,r,\rho)\right\| \le C_{l,m}\left[\sum_{i=1}^{l+2m}(-s)^i\right]e^{\{\beta_1-[H+(l+m)(1+2\Delta)]\lambda(\rho)\mu^k\}s}.$$

(f) For $0 \le l+m \le k-1$,

$$\left\|\frac{\partial^{l+m}\Lambda}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(0,s,\sigma^0,\mu,q_1,r_1,\rho) - \frac{\partial^{l+m}\Lambda}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(0,s,\sigma^0,\mu,q_2,r_2,\rho)\right\|$$
$$\le \widehat{C}_{l,m}\left(\|q_1-q_2\|_{l+m} + \|r_1-r_2\|_{l+m}\right)$$
$$\times\left[\sum_{i=2}^{l+2m+2}(-s)^i\right]e^{\{\beta_1-[H+(l+m+1)(1+2\Delta)]\lambda(\rho)\mu^k\}s}.$$

Estimates in Property 3.1(a)–(c) will be crucial in the contraction mapping argument used later.

To establish the existence of integral manifolds we define a mapping $S_\rho = (S_{\rho,1}, S_{\rho,2})$ on $\mathcal{B}^k_\Delta(I,N,N_1)\times\mathcal{B}^k_\Delta(I,N,N_2)$ by

$$\begin{aligned} S_{\rho,1}(q,r)(\sigma,\mu) &= \int_{-\infty}^0 G_1(s,\sigma,\mu,q,r,\rho)\,ds, \\ S_{\rho,2}(q,r)(\sigma,\mu) &= \int_{-\infty}^0 G_2(s,\sigma,\mu,q,r,\rho)\,ds, \end{aligned} \tag{3.3}$$

where

$$\begin{aligned} G_1(s,\sigma,\mu,q,r,\rho) &= \Lambda(0,s,\sigma,\mu,q,r,\rho)\mathcal{W}(\phi(s),q(\phi(s),\mu),r(\phi(s),\mu),\mu,\rho), \\ G_2(s,\sigma,\mu,q,r,\rho) &= \rho\mu e^{-\rho\mu Q s}\mathcal{Z}(\phi(s),q(\phi(s),\mu),r(\phi(s),\mu),\mu,\rho) \\ \phi(s) &= \phi(s,\sigma,\mu,q,r,\rho). \end{aligned} \tag{3.4}$$

We will show that for sufficiently small $|\rho|$, S_ρ has a fixed point in $\mathcal{PC}^{k-1}(\mathbb{R}^N\times I,\mathbb{R}^{N_1})\times\mathcal{PC}^{k-1}(\mathbb{R}^N\times I,\mathbb{R}^{N_2})$, and that this fixed point is an integral manifold of (1.8).

In the remainder of this section we assume that $0<\rho\le\rho_1$ is fixed and ρ_1 is sufficiently small. By Assumption 2.1(d), we may also assume without loss of generality that

$$\lambda(\rho) \le \rho \quad \text{for } 0<\rho\le\rho_1. \tag{3.5}$$

We summarize the results, which concern some estimates associated with $G_1(s,\sigma,\mu,q,r,\rho)$ and $G_2(s,\sigma,\mu,q,r,\rho)$, in the following lemmas. The proofs are given in the Appendix.

LEMMA 3.2. *For $0 \le l+m \le k$ there is a positive constant K such that for $s \le 0$,*

$$\left\| \frac{\partial^{l+m} G_1}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^m}(s,\sigma,\mu,q,r,\rho) \right\| \le K\lambda_1(\rho)\left[\sum_{i=0}^{l+2m}(-s)^i\right] e^{\{\beta_1-[H+(l+m)(1+2\Delta)]\lambda(\rho)\mu^k\}s}, \tag{3.6}$$

$$\left\| \frac{\partial^{l+m} G_2}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^m}(s,\sigma,\mu,q,r,\rho) \right\| \le K\lambda_2(\rho)\mu^{k-m}\left[\sum_{i=0}^{l+2m}(\rho\mu)^{i+1}(-s)^i\right] e^{[\rho\mu\beta_2-(l+m)\lambda(\rho)\mu^k(1+2\Delta)]s}. \tag{3.7}$$

LEMMA 3.3. *For $0 \le l+m \le k-1$ there is a positive constant C such that for $s \le 0$,*

$$\left\| \frac{\partial^{l+m} G_1}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^m}(s,\sigma,\mu,q_1,r_1,\rho) - \frac{\partial^{l+m} G_1}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^m}(s,\sigma,\mu,q_2,r_2,\rho) \right\|$$

$$\le C\lambda_1(\rho)\left(\|q_1-q_2\|_{l+m}+\|r_1-r_2\|_{l+m}\right) \times\left[\sum_{i=0}^{l+2m+2}(-s)^i\right] e^{\{\beta_1-[H+(l+m+1)(1+2\Delta)]\lambda(\rho)\mu^k\}s}, \tag{3.8}$$

$$\left\| \frac{\partial^{l+m} G_2}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^m}(s,\sigma,\mu,q_1,r_1,\rho) - \frac{\partial^{l+m} G_2}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^m}(s,\sigma,\mu,q_2,r_2,\rho) \right\|$$

$$\le C\lambda_2(\rho)\mu^{k-m}\left(\|q_1-q_2\|_{l+m}+\|r_1-r_2\|_{l+m}\right) \times\left[\sum_{i=0}^{l+2m+2}(\rho\mu)^{i+1}(-s)^i\right] e^{[\rho\mu\beta_2-(l+m+1)\lambda(\rho)\mu^k(1+2\Delta)]s}. \tag{3.9}$$

Now we are ready to state and prove one of the main results.

THEOREM 3.4. *Under Assumptions* 2.1(a)–(d), *there is $\rho_0 > 0$ such that for $0 < \rho \le \rho_0$ S_ρ, defined by* (3.3) *and* (3.4) *maps $\mathcal{B}^k_\Delta(I,N,N_1)\times\mathcal{B}^k_\Delta(I,N,N_2)$ into itself and has a unique fixed point (q_ρ,r_ρ) in the closure of $\mathcal{B}^k_\Delta(I,N,N_1)\times\mathcal{B}^k_\Delta(I,N,N_2)$ in $\mathcal{PC}^{k-1}(\mathbb{R}^N\times I,\mathbb{R}^{N_1})\times\mathcal{PC}^{k-1}(\mathbb{R}^N\times I,\mathbb{R}^{N_2})$. (q_ρ,r_ρ) is an integral manifold of* (1.8) *such that, if $0 \le l+m \le k-1$,*

$$\frac{\partial^{l+m} r_\rho}{\partial\sigma^l\partial\mu^m}(\sigma,\mu) \longrightarrow 0 \tag{3.10}$$

as $\mu \longrightarrow 0$. *Moreover,* q_ρ *and* r_ρ *satisfy the system of first-order quasilinear partial differential equations*

$$\begin{aligned} &\frac{\partial q_\rho}{\partial \sigma}\,[\omega + \mathcal{S}(\sigma^\star, q_\rho, r_\rho, \mu, \rho)] = A(\Omega^{-1}\sigma^\star)q_\rho + \mathcal{W}(\sigma^\star, q_\rho, r_\rho, \mu, \rho), \\ &\frac{\partial r_\rho}{\partial \sigma}\,[\omega + \mathcal{S}(\sigma^\star, q_\rho, r_\rho, \mu, \rho)] = \rho\mu\,[Qr_\rho + \mathcal{Z}(\sigma^\star, q_\rho, r_\rho, \mu, \rho)] \end{aligned} \tag{3.11}$$

$(at\ \sigma = \sigma^\star)$. *In addition, under Assumption* 2.1(e),

$$q_\rho(\sigma, \mu) \longrightarrow 0 \quad as\ \mu \longrightarrow 0 \quad for\ all\ \sigma \in \mathbb{R}^N. \tag{3.12}$$

Proof. Choose ρ_0 such that $0 < \rho_0 \leq \rho_1$ and for $0 < \rho \leq \rho_0$,

$$\beta_1 \geq 2[H + k(1 + 2\Delta)]\lambda(\rho), \tag{3.13}$$

$$\beta_2 \geq 2k(1 + 2\Delta)\frac{\lambda(\rho)}{\rho}, \tag{3.14}$$

$$K\lambda_j(\rho)\left[\sum_{i=0}^{2k} \frac{2^{i+1}i!}{\beta_j^{i+1}}\right] \leq \Delta, \qquad j = 1, 2, \tag{3.15}$$

$$C\lambda_j(\rho)\left[\sum_{i=0}^{2k} \frac{2^{i+1}i!}{\beta_j^{i+1}}\right] \leq \frac{1}{4}, \qquad j = 1, 2. \tag{3.16}$$

By Lemma 5.1, (3.13), and (3.14), $S_\rho(q, r) \in \mathcal{PC}^k(\mathbb{R}^N \times I, \mathbb{R}^{N_1}) \times \mathcal{PC}^k(\mathbb{R}^N \times I, \mathbb{R}^{N_2})$ (cf. [11]). By Lemma 3.2 and (3.15), $S_\rho(q, r) \in \mathcal{B}_\Delta^k(I, N, N_1) \times \mathcal{B}_\Delta^k(I, N, N_2)$. By Lemma 3.3 and (3.16), S_ρ is a contraction in $\mathcal{PC}^{k-1}(\mathbb{R}^N \times I, \mathbb{R}^{N_1}) \times \mathcal{PC}^{k-1}(\mathbb{R}^N \times I, \mathbb{R}^{N_2})$. The existence of a pair (q_ρ, r_ρ) now follows from these facts.

It is shown in [3] that such a pair (q_ρ, r_ρ) defines an integral manifold. However the proof is easy and we include it for completeness. It is easily shown that

$$\phi(s, \phi(t, \sigma^0, \mu, q, r, \rho), \mu, q, r, \rho) = \phi(s + t, \sigma^0, \mu, q, r, \rho),$$

$$\Lambda(0, s, \phi(t, \sigma^0, \mu, q, r, \rho), \mu, q, r, \rho) = \Lambda(t, s + t, \sigma^0, \mu, q, r, \rho).$$

It follows that

$$\begin{aligned} q(\phi(t), \mu) &= \int_{-\infty}^{0} \Lambda(t, s + t)\mathcal{W}(\phi(s + t), q(\phi(s + t), \mu), r(\phi(s + t), \mu), \mu, \rho)\, ds \\ &= \int_{-\infty}^{t} \Lambda(t, s)\mathcal{W}(\phi(s), q(\phi(s), \mu), r(\phi(s), \mu), \mu, \rho)\, ds, \end{aligned}$$

$$\begin{aligned} r(\phi(t), \mu) &= \int_{-\infty}^{0} e^{-Qs}\mathcal{Z}(\phi(s + t), q(\phi(s + t), \mu), r(\phi(s + t), \mu), \mu, \rho)\, ds \\ &= \int_{-\infty}^{t} e^{Q(t-s)}\mathcal{Z}(\phi(s), q(\phi(s), \mu), r(\phi(s), \mu), \mu, \rho)\, ds. \end{aligned}$$

This shows that

$$\sigma = \phi(t, \sigma^0, \mu, q, r, \rho), \quad w = q(\phi(t, \sigma^0, \mu, q, r, \rho), \mu), \quad z = r(\phi(t, \sigma^0, \mu, q, r, \rho), \mu)$$

is a solution of (1.8) (cf. [1]). Thus (q_ρ, r_ρ) is an integral manifold. Moreover, by Lemma 3.2,

$$\left\| \frac{\partial^{l+m} r_\rho}{\partial \sigma^l \partial \mu^m}(\sigma, \mu) \right\| \leq \Delta \mu^{k-m} \tag{3.17}$$

for $0 \leq l + m \leq k - 1$. Now (3.10) follows from (3.17). Equation (3.11) follows from the fact that the manifold defined by $w = q_\rho(\sigma, \mu)$, $z = r_\rho(\sigma, \mu)$ is an invariant manifold of (1.8) (cf. Part I). When $\mathcal{W}$ and $\mathcal{Z}$ satisfy Assumption 2.1(e), we consider the restriction of S_ρ to the closed subspace of $\mathcal{PC}^k(\mathbb{R}^N \times I, \mathbb{R}^{N_1}) \times \mathcal{PC}^k(\mathbb{R}^N \times I, \mathbb{R}^{N_2})$ consisting of all (q, r) such that

$$q(\sigma, \mu) \longrightarrow 0, \qquad r(\sigma, \mu) \longrightarrow 0,$$

as $\mu \longrightarrow 0$ and obtain (3.12). □

COROLLARY 3.5. *Under Assumption* 1.1, *for* $1 - 1/k \leq p < 1$, *there is a* $\delta_1 > 0$ *such that for fixed* $\delta \in (0, \delta_1)$, (1.4) *has a* C^{k-1} *invariant torus given by Proposition* 2.3(a) *and Theorem* 3.4. *The invariant tori are* $(k-1)$*-times continuously differentiable in* δ^{1-p}, *and they bifurcate from the invariant torus of the unperturbed system defined by* (1.7).

The construction of an integral manifold of (1.9) is similar. Let $\mathcal{PC}^k(\mathbb{R}^m, \mathbb{R}^n)$ be the Banach space of all k-times continuously differentiable mappings $s : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ such that

$$\|s\|_k = \max_{0 \leq i \leq k} \sup \left\{ \left\| \frac{\partial^i s}{\partial \sigma^i}(\sigma) \right\| : \sigma \in \mathbb{R}^m \right\} < \infty,$$

and for each $i = 1, \cdots, m$,

$$s(\sigma + 2\pi e^i) = s(\sigma) \quad \text{for all } \sigma \in \mathbb{R}^m.$$

Let $\mathcal{B}^k_\Delta(m, n)$ be the closed ball in $\mathcal{PC}^k(\mathbb{R}^m, \mathbb{R}^n)$ with radius Δ and center at the origin. Let $\phi(t, \sigma^0, q, r, \mu)$ be the solution of

$$\frac{d\sigma}{dt} = \omega + \mathcal{S}(\sigma, q(\sigma), r(\sigma), \mu), \qquad \phi(0, \sigma^0, q, r, \mu) = \sigma^0, \tag{3.18}$$

where $(q, r) \in \mathcal{B}^k_\Delta(N, N_1) \times \mathcal{B}^k_\Delta(N, N_2)$. Then let $\Lambda(t, t_0, \sigma^0, q, r, \mu)$ be the fundamental matrix solution of

$$\frac{dw}{dt} = A(\Omega^{-1}\phi(t, \sigma^0, q, r, \mu))w, \qquad \Lambda(t_0, t_0, \sigma^0, q, r, \mu) = I_{N_1 \times N_1}. \tag{3.19}$$

We define a mapping $S_\mu = (S_{\mu,1}, S_{\mu,2})$ on $\mathcal{B}^k_\Delta(N, N_1) \times \mathcal{B}^k_\Delta(N, N_2)$ by

$$
\begin{aligned}
S_{\mu,1}(q,r)(\sigma) &= \int_{-\infty}^{0} G_1(s,\sigma,q,r,\mu)\,ds, \\
S_{\mu,2}(q,r)(\sigma) &= \int_{-\infty}^{0} G_2(s,\sigma,q,r,\mu)\,ds,
\end{aligned} \tag{3.20}
$$

where

$$
\begin{aligned}
G_1(s,\sigma,q,r,\mu) &= \Lambda(0,s,\sigma,q,r,\mu)\mathcal{W}(\phi(s),q(\phi(s)),r(\phi(s)),\mu), \\
G_2(s,\sigma,q,r,\mu) &= \mu e^{-\mu Q s}\mathcal{Z}(\phi(s),q(\phi(s)),r(\phi(s)),\mu), \\
\phi(s) &= \phi(s,\sigma,q,r,\mu).
\end{aligned} \tag{3.21}
$$

We will show that for all sufficiently small $\mu > 0$, S_μ has a fixed point in $\mathcal{PC}^{k-1}(\mathbb{R}^N, \mathbb{R}^{N_1}) \times \mathcal{PC}^{k-1}(\mathbb{R}^N, \mathbb{R}^{N_2})$ and that this fixed point is an integral manifold of (1.9).

THEOREM 3.6. *Under Assumptions* 2.1(a)–(c) *and* (f), *there is* $\mu_0 > 0$ *such that for* $0 < \mu \le \mu_0$ S_μ, *defined by* (3.20) *and* (3.21), *maps* $\mathcal{B}_\Delta^k(N,N_1) \times \mathcal{B}_\Delta^k(N,N_2)$ *into itself and has a unique fixed point* (q_μ, r_μ) *in the closure of* $\mathcal{B}_\Delta^k(N,N_1) \times \mathcal{B}_\Delta^k(N,N_2)$ *in* $\mathcal{PC}^{k-1}(\mathbb{R}^N, \mathbb{R}^{N_1}) \times \mathcal{PC}^{k-1}(\mathbb{R}^N, \mathbb{R}^{N_2})$. (q_μ, r_μ) *is an integral manifold of* (1.9) *such that if* $0 \le l \le k-1$

$$
\frac{\partial^l q_\mu}{\partial \sigma^l}(\sigma) \longrightarrow 0, \qquad \frac{\partial^l r_\mu}{\partial \sigma^l}(\sigma) \longrightarrow 0,
$$

as $\mu \longrightarrow 0$. *Moreover,* q_μ *and* r_μ *satisfy the system of first-order quasilinear partial differential equations*

$$
\frac{\partial q_\mu}{\partial \sigma}\left[\omega + \mathcal{S}(\sigma^\star, q_\mu, r_\mu, \mu)\right] = A(\Omega^{-1}\sigma^\star)q_\mu + \mathcal{W}(\sigma^\star, q_\mu, r_\mu, \mu),
$$

$$
\frac{\partial r_\mu}{\partial \sigma}\left[\omega + \mathcal{S}(\sigma^\star, q_\mu, r_\mu, \mu)\right] = \mu\left[Q r_\mu + \mathcal{Z}(\sigma^\star, q_\mu, r_\mu, \mu)\right]
$$

$(at\ \sigma = \sigma^\star)$.

Proof. One derives estimates similar to those given in Lemmas 3.2 and 3.3 using Property 5.13 given in Appendix and then repeats the argument in the proof of Theorem 3.4. □

COROLLARY 3.7. *Under Assumption* 1.1, *for* $0 < p < 1$, *there is a* $\delta_1 > 0$ *such that for* $0 < \delta < \delta_1$, (1.4) *has a* C^{k-1} *invariant torus given by Proposition* 2.3(b) *and Theorem* 3.6. *These invariant tori bifurcate from the invariant torus of the unperturbed system defined by* (1.7).

Remark 4. If $\epsilon = \epsilon_0\delta^{-p}$ and $1 - 1/k \le p < 1$, and $w = q_\rho(\sigma,\mu)$, $x = r_\rho(\sigma,\mu)$ is an integral manifold of (2.12) whose existence is proven in Theorem 3.4, then by (2.9),

$$
y = \rho^\kappa \mu^\nu q_\rho(\sigma,\mu), \quad z = r_\rho(\sigma,\mu), \quad \delta = \rho^\nu \mu^\nu \tag{3.22}
$$

is an integral manifold of (2.7). Also, if $\epsilon = \epsilon_0\delta^{-p}$ and $0 < p < 1$, and $w = q_\mu(\sigma)$, $x = r_\mu(\sigma)$ is an integral manifold of (2.13) whose existence is proven in Theorem 3.6, then by (2.10),

$$y = \mu^{\zeta} q_{\mu}(\sigma), \quad z = r_{\mu}(\sigma), \quad \delta = \mu^{\nu} \tag{3.23}$$

is an integral manifold of (2.7). The integral manifolds for the systems of degenerate type (1.12) and (1.13) are at most Lipschitz continuous in θ and continuous with respect to parameters, whereas (3.22) is $(k-1)$-times continuously differentiable in θ and δ, and (3.23) is $(k-1)$-times continuously differentiable in θ.

4. Asymptotic behavior of solutions near integral manifolds. In this section we study the behavior of solutions of (1.8) in a neighborhood of the integral manifold $w = q_{\rho}(\sigma, \mu)$, $z = r_{\rho}(\sigma, \mu)$. We assume that $I \subset \mathbb{R}^{+}$ and show that the integral manifold is asymptotically stable for small $\rho > 0$ in the sense that there is a $\rho_d > 0$ such that for each $\rho \in (0, \rho_d)$ and $\mu \in I$, there is a neighborhood of the integral manifold such that any solution of (1.8) through a point in the neighborhood approaches the manifold at a exponential rate in time (cf. Proposition 4.4). This leads to the following conclusion. For all small $\delta > 0$, there is a neighborhood of the unperturbed torus (1.7) such that the distance between any solution of (1.4) through a point in the neighborhood and its projection onto the torus defined by (2.14) decays at a exponential rate in time (cf. (4.33)). This result is stated in Theorem 4.5. We present similar results concerning the behavior of solutions of (1.9) near the integral manifold $w = q_{\mu}(\sigma)$, $z = r_{\mu}(\sigma)$ and the behavior of solutions of (1.4) near the corresponding invariant tori in Proposition 4.7 and Theorem 4.5.

Let

$$w = u + q_{\rho}(\sigma, \mu), \qquad z = v + r_{\rho}(\sigma, \mu).$$

Then (1.8) becomes

$$\begin{aligned}
\frac{d\sigma}{dt} &= \omega + \mathcal{S}_0(\sigma, u, v, \mu, \rho), \\
\frac{du}{dt} &= A(\Omega^{-1}\sigma)u + U_1(\sigma, u, v, \mu, \rho)u + U_2(\sigma, u, v, \mu, \rho)v, \\
\frac{dv}{dt} &= \rho\mu Q v + V_1(\sigma, u, v, \mu, \rho)u + V_2(\sigma, u, v, \mu, \rho)v,
\end{aligned} \tag{4.1}$$

where

$$\begin{aligned}
\mathcal{S}_0(\sigma, u, v, \mu, \rho) &= \mathcal{S}\left(\sigma, u + q_{\rho}(\sigma, \mu), v + r_{\rho}(\sigma, \mu), \mu, \rho\right) \\
&= \mu\Sigma(\sigma, \mu, \rho) + S_1(\sigma, u, v, \mu, \rho)u + S_2(\sigma, u, v, \mu, \rho)v,
\end{aligned}$$

$$\begin{aligned}
\Sigma(\sigma, \mu, \rho) = \int_0^1 \Bigg[&\frac{\partial \mathcal{S}}{\partial w}(\sigma, q_{\rho}(\sigma, s\mu), r_{\rho}(\sigma, s\mu), s\mu, \rho) \frac{\partial q_{\rho}}{\partial \mu}(\sigma, s\mu) \\
&+ \frac{\partial \mathcal{S}}{\partial z}(\sigma, q_{\rho}(\sigma, s\mu), r_{\rho}(\sigma, s\mu), s\mu, \rho) \frac{\partial r_{\rho}}{\partial \mu}(\sigma, s\mu) \\
&+ \frac{\partial \mathcal{S}}{\partial \mu}(\sigma, q_{\rho}(\sigma, s\mu), r_{\rho}(\sigma, s\mu), s\mu, \rho) \Bigg] \, ds,
\end{aligned}$$

$$S_1(\sigma,u,v,\mu,\rho)=\int_0^1 \frac{\partial \mathcal{S}}{\partial w}(\sigma,su+q_\rho(\sigma,\mu),sv+r_\rho(\sigma,\mu),\mu,\rho)\,ds,$$

$$S_2(\sigma,u,v,\mu,\rho)=\int_0^1 \frac{\partial \mathcal{S}}{\partial z}(\sigma,su+q_\rho(\sigma,\mu),sv+r_\rho(\sigma,\mu),\mu,\rho)\,ds,$$

and

$$U_1(\sigma,u,v,\mu,\rho)=\int_0^1 \frac{\partial \mathcal{W}}{\partial w}(\sigma,su+q_\rho(\sigma,\mu),sv+r_\rho(\sigma,\mu),\mu,\rho)\,ds$$
$$-\frac{\partial q_\rho}{\partial \sigma}(\sigma,\mu)S_1(\sigma,u,v,\mu,\rho),$$

$$U_2(\sigma,u,v,\mu,\rho)=\int_0^1 \frac{\partial \mathcal{W}}{\partial z}(\sigma,su+q_\rho(\sigma,\mu),sv+r_\rho(\sigma,\mu),\mu,\rho)\,ds$$
$$-\frac{\partial q_\rho}{\partial \sigma}(\sigma,\mu)S_2(\sigma,u,v,\mu,\rho),$$

$$V_1(\sigma,u,v,\mu,\rho)=\rho\mu\int_0^1 \frac{\partial \mathcal{Z}}{\partial w}(\sigma,su+q_\rho(\sigma,\mu),sv+r_\rho(\sigma,\mu),\mu,\rho)\,ds$$
$$-\frac{\partial r_\rho}{\partial \sigma}(\sigma,\mu)S_1(\sigma,u,v,\mu,\rho),$$

$$V_2(\sigma,u,v,\mu,\rho)=\rho\mu\int_0^1 \frac{\partial \mathcal{Z}}{\partial z}(\sigma,su+q_\rho(\sigma,\mu),sv+r_\rho(\sigma,\mu),\mu,\rho)\,ds$$
$$-\frac{\partial r_\rho}{\partial \sigma}(\sigma,\mu)S_2(\sigma,u,v,\mu,\rho).$$

Denote by $S_n(x,\varepsilon)$ the open ball of radius ε in $\mathbb{R}^n$ centered at x. Choose a neighborhood of the origin in $\mathbb{R}^{N_1+N_2}$, which we call $\mathcal{D}$, such that

$$\overline{S_{N_1}(0,\Delta)}\times\overline{S_{N_2}(0,\Delta)}+\overline{\mathcal{D}}\subset F_1\times F_2. \tag{4.2}$$

Then for $\rho\in(0,\rho_0)$, $\mathcal{S}_0$, U_1, U_2, V_1, and V_2 are defined and continuously differentiable in $\mathbb{R}^N\times\mathcal{D}\times I$. Choose a positive number ε_0 such that

$$\overline{S_{N_1+N_2}(0,\varepsilon_0)}\subset\mathcal{D}. \tag{4.3}$$

For $\mu\in I$, $\rho\in(0,\rho_0)$, and $(u,v)\in\overline{S_{N_1+N_2}(0,\varepsilon_0)}$,

$$\|\mathcal{S}_0(\sigma,u,v,\mu,\rho)\|\le\lambda(\rho)\mu^k, \tag{4.4}$$

$$\begin{aligned}&\|U_i(\sigma,u,v,\mu,\rho)\|\le\lambda_1(\rho)+\lambda(\rho)\mu^k\Delta, \qquad i=1,2,\\ &\|V_i(\sigma,u,v,\mu,\rho)\|\le\lambda_2(\rho)\rho\mu+\lambda(\rho)\mu^k\Delta, \qquad i=1,2.\end{aligned} \tag{4.5}$$

Let

$$\begin{aligned} \sigma &= \sigma(t) = \sigma(t, \sigma^\star, u^\star, v^\star, \mu, \rho), \\ u &= u(t) = u(t, \sigma^\star, u^\star, v^\star, \mu, \rho), \\ v &= v(t) = v(t, \sigma^\star, u^\star, v^\star, \mu, \rho) \end{aligned} \tag{4.6}$$

be the solution of (4.1) with initial value

$$\sigma(0, \sigma^\star, u^\star, v^\star, \mu, \rho) = \sigma^\star, \qquad u(0, \sigma^\star, u^\star, v^\star, \mu, \rho) = u^\star,$$
$$v(0, \sigma^\star, u^\star, v^\star, \mu, \rho) = v^\star.$$

If $(u^\star, v^\star) \in S_{N_1+N_2}(0, \varepsilon_0)$, there is $t_1 > 0$ such that (4.6) exists and $(u(t), v(t)) \in S_{N_1+N_2}(0, \varepsilon_0)$ for $t \in [0, t_1)$. Let $u = \Xi(t, t_0) = \Xi(t, t_0, \sigma^\star, u^\star, v^\star, \mu, \rho)$ be the fundamental matrix solution of

$$\frac{du}{dt} = A(\Omega^{-1}\sigma(t, \sigma^\star, u^\star, v^\star, \mu, \rho))u, \qquad \Xi(t_0, t_0, \sigma^\star, u^\star, v^\star, \mu, \rho) = I_{N_1 \times N_1}.$$

We assert that $\Xi(t, t_0)$ decays exponentially as t increases in Lemma 4.1. The proof of this lemma is similar to the one for Lemma 5.8 in the Appendix and is left to the reader.

LEMMA 4.1.

$$\|\Xi(t, t_0, \sigma^\star, u^\star, v^\star, \mu, \rho)\| \le K_1 e^{-\left[\beta_1 - H\lambda(\rho)\mu^k\right](t-t_0)} \quad \textit{for } t_0 \le t.$$

We can write $u(t)$ and $v(t)$ in the form

$$\begin{aligned} u(t) = \Xi(t, 0)u^\star + \int_0^t \Xi(t, s) \, &[U_1(\sigma(s), u(s), v(s), \mu, \rho)u(s) \\ &+ U_2(\sigma(s), u(s), v(s), \mu, \rho)v(s)] \, ds, \\ v(t) = e^{\rho\mu Q t}v^\star + \int_0^t e^{\rho\mu Q(t-s)} \, &[V_1(\sigma(s), u(s), v(s), \mu, \rho)u(s) \\ &+ V_2(\sigma(s), u(s), v(s), \mu, \rho)v(s)] \, ds, \end{aligned}$$

and now (4.4), (4.5), and Lemma 4.1 lead to the estimate

$$\begin{aligned} \|u(t)\| &\le K_1\|u^\star\|e^{-\left[\beta_1 - H\lambda(\rho)\mu^k\right]t} \\ &\quad + \int_0^t K_1 \left[\lambda_1(\rho) + \lambda(\rho)\mu^k\Delta\right] e^{-\left[\beta_1 - H\lambda(\rho)\mu^k\right](t-s)} \left(\|u(s)\| + \|v(s)\|\right) ds, \\ \|v(t)\| &\le K_2\|v^\star\|e^{-\rho\mu\beta_2 t} + \int_0^t K_2 \left[\lambda_2(\rho)\rho\mu + \lambda(\rho)\mu^k\Delta\right] \\ &\quad \times e^{-\rho\mu\beta_2(t-s)} \left(\|u(s)\| + \|v(s)\|\right) ds. \end{aligned} \tag{4.7}$$

We prove the following two lemmas in order to show that the functions $u(t)$ and $v(t)$, which satisfy (4.7), decay exponentially as t increases when μ and ρ are sufficiently small.

LEMMA 4.2. *Suppose that h is a continuous nonnegative function on the closed interval $[t_0, t_1]$ such that*

$$h(t) \le e^{A_1(t-t_0)} \left\{ B_1 + C_1 \int_{t_0}^{t} e^{(A_2-A_1)(s-t_0)} \left[B_2 + C_2 \int_{t_0}^{s} e^{-A_2(\tau-t_0)} h(\tau)\, d\tau \right] ds \right\}$$

for all $t \in [t_0, t_1]$, where A_i, B_i, and C_i are real numbers and $C_i \ge 0$ for $i = 1$ and 2. Then

$$h(t) \le e^{\mathcal{R}_0(t-t_0)} \left[B_1 + (B_2C_1 - B_1 r) \frac{1 - e^{-\mathcal{R}(t-t_0)}}{\mathcal{R}} \right] \quad \text{for all } t \in [t_0, t_1], \tag{4.8}$$

where

$$\mathcal{R} = \sqrt{(A_2 - A_1)^2 + 4C_1C_2} = |A_2 - A_1| + 2\mathcal{I},$$

$$\mathcal{R}_0 = \frac{A_1 + A_2 + R}{2} = \max\{A_1, A_2\} + \mathcal{I},$$

$$r = \frac{A_2 - A_1 + R}{2} = \frac{A_2 - A_1 + |A_2 - A_1|}{2} + \mathcal{I},$$

$$\mathcal{I} = \int_0^1 \frac{C_1C_2\, ds}{\sqrt{(A_2 - A_1)^2 + 4C_1C_2 s}}.$$

Proof. Let

$$G(t) = B_1 + C_1 \int_{t_0}^{t} e^{(A_2-A_1)(s-t_0)} \left[B_2 + C_2 \int_{t_0}^{s} e^{-A_2(\tau-t_0)} h(\tau)\, d\tau \right] ds. \tag{4.9}$$

Then

$$h(t) \le e^{A_1(t-t_0)} G(t). \tag{4.10}$$

On the other hand, $G(t)$ is twice continuously differentiable on (t_0, t_1). Moreover,

$$G'(t) = C_1 e^{(A_2-A_1)(t-t_0)} \left[B_2 + C_2 \int_{t_0}^{t} e^{-A_2(\tau-t_0)} h(\tau)\, d\tau \right]. \tag{4.11}$$

Differentiating (4.11) again and using (4.10), we find that

$$\begin{aligned} G''(t) &= (A_2 - A_1)G'(t) + C_1C_2 e^{-A_1(t-t_0)} h(t) \\ &\le (A_2 - A_1)G'(t) + C_1C_2 G(t). \end{aligned} \tag{4.12}$$

Let

$$G(t) = e^{r(t-t_0)}H(t). \tag{4.13}$$

Then using (4.12), we obtain

$$H''(t) + \mathcal{R}H'(t) \leq 0.$$

It follows that

$$H'(t) \leq H'(t_0)e^{-\mathcal{R}(t-t_0)}$$

and

$$H(t) \leq H(t_0) + H'(t_0)\frac{1 - e^{-\mathcal{R}(t-t_0)}}{\mathcal{R}}. \tag{4.14}$$

From (4.9) and (4.13), we find that

$$H(t_0) = G(t_0) = B_1. \tag{4.15}$$

From (4.11), (4.13), and (4.15), we obtain

$$H'(t_0) = G'(t_0) - rH(t_0) = B_2C_1 - B_1r. \tag{4.16}$$

On the other hand, according to (4.10) and (4.13),

$$h(t) \leq e^{A_1(t-t_0)}G(t) \leq e^{(A_1+r)(t-t_0)}H(t). \tag{4.17}$$

Now (4.8) follows from (4.14), (4.15), (4.16), and (4.17). □

LEMMA 4.3. *Suppose that h_1 and h_2 are continuous nonnegative functions on the closed interval $[t_0, t_1]$ such that for $t \in [t_0, t_1]$,*

$$h_1(t) \leq e^{-a_1(t-t_0)}\left\{b_1 + c_1\int_{t_0}^{t} e^{a_1(s-t_0)}\left[h_1(s) + h_2(s)\right]\,ds\right\},$$

$$h_2(t) \leq e^{-a_2(t-t_0)}\left\{b_2 + c_2\int_{t_0}^{t} e^{a_2(s-t_0)}\left[h_1(s) + h_2(s)\right]\,ds\right\},$$

where a_i, b_i, and c_i are real numbers and $c_i \geq 0$. Then for $t \in [t_0, t_1]$,

$$h_1(t) \leq e^{-R_0(t-t_0)}\left[b_1 + (b_2c_1 - b_1r_1)\frac{1 - e^{-R(t-t_0)}}{R}\right], \tag{4.18}$$

$$h_2(t) \leq e^{-R_0(t-t_0)}\left[b_2 + (b_1c_2 - b_2r_2)\frac{1 - e^{-R(t-t_0)}}{R}\right], \tag{4.19}$$

where

$$R = \sqrt{[a_1 - c_1 - (a_2 - c_2)]^2 + 4c_1c_2} = |a_1 - c_1 - (a_2 - c_2)| + 2\mathcal{I}_0,$$

$$r_1 = \frac{a_1 - c_1 - (a_2 - c_2) + R}{2} = \frac{a_1 - c_1 - (a_2 - c_2) + |a_1 - c_1 - (a_2 - c_2)|}{2} + \mathcal{I}_0,$$

$$r_2 = \frac{a_2 - c_2 - (a_1 - c_1) + R}{2} = \frac{a_2 - c_2 - (a_1 - c_1) + |a_1 - c_1 - (a_2 - c_2)|}{2} + \mathcal{I}_0,$$

$$R_0 = a_1 - c_1 - r_1 = a_2 - c_2 - r_2 = \min\{a_1 - c_1, a_2 - c_2\} - \mathcal{I}_0,$$

$$\mathcal{I}_0 = \int_0^1 \frac{c_1c_2\,ds}{\sqrt{[a_1 - c_1 - (a_2 - c_2)]^2 + 4c_1c_2s}}.$$

Proof. It can easily be shown that h_1 and h_2 satisfy

$$h_1(t) \leq e^{-(a_1-c_1)(t-t_0)}\left[b_1 + c_1\int_{t_0}^{t} e^{(a_1-c_1)(s-t_0)}h_2(s)\,ds\right], \tag{4.20}$$

$$h_2(t) \leq e^{-(a_2-c_2)(t-t_0)}\left[b_2 + c_2\int_{t_0}^{t} e^{(a_2-c_2)(s-t_0)}h_1(s)\,ds\right]. \tag{4.21}$$

Substituting (4.21) in (4.20), we obtain

$$h_1(t) \leq e^{-(a_1-c_1)(t-t_0)}\left\{b_1 + c_1\int_{t_0}^{t} e^{[a_1-c_1-(a_2-c_2)](s-t_0)} \times \left[b_2 + c_2\int_{t_0}^{s} e^{(a_2-c_2)(\tau-t_0)}h_1(\tau)\,d\tau\right] ds\right\}.$$

Set

$$A_i = -(a_i - c_i), \qquad B_i = b_i, \qquad C_i = c_i, \qquad i = 1, 2.$$

Now (4.18) follows from Lemma 4.2. A similar argument leads to (4.19). □

We set

$$\begin{aligned} a_1 &= \beta_1 - \lambda(\rho)\mu^k(2\mu\Delta + 1 + \varepsilon_0)H, \qquad b_1 = K_1\|u^\star\|, \\ c_1 &= K_1\left[\lambda_1(\rho) + \lambda(\rho)\mu^k\Delta\right], \\ a_2 &= \rho\mu\beta_2, \qquad b_2 = K_2\|v^\star\|, \\ c_2 &= K_2\left[\lambda_2(\rho)\rho\mu + \lambda(\rho)\mu^k\Delta\right]. \end{aligned} \tag{4.22}$$

Then for all sufficiently small $\rho > 0$, $a_1 - c_1 - (a_2 - c_2)$ is positive and bounded away from zero for all $\mu \in I$ uniformly. R_0 now becomes

$$R_0 = a_2 - c_2 - c_1 c_2 \int_0^1 \frac{ds}{\sqrt{[a_1 - c_1 - (a_2 - c_2)]^2 + 4c_1 c_2 s}}. \tag{4.23}$$

Since

$$a_2 - c_2 = \rho\mu \left\{ \beta_2 - K_2 \left[\lambda_2(\rho) + \frac{\lambda(\rho)}{\rho} \mu^{k-1} \Delta \right] \right\},$$

$$c_1 c_2 = \rho\mu K_1 K_2 \left[\lambda_1(\rho) + \lambda(\rho)\mu^k \Delta \right] \left[\lambda_2(\rho) + \frac{\lambda(\rho)}{\rho} \mu^{k-1} \Delta \right],$$

and the integral in (4.23) is bounded, there is a $\alpha_d > 0$ such that for all small $\rho > 0$, R_0 defined by (4.22) and (4.23) satisfies

$$R_0 > \rho\mu\alpha_d \tag{4.24}$$

for all $\mu \in I$. On the other hand,

$$\begin{aligned} & b_1 + b_2 + (b_2 c_1 - b_1 r_1 + b_1 c_2 - b_2 r_2) \frac{1 - e^{-Rt}}{R} \\ &= b_1 \left[1 + (c_2 - r_1) \frac{1 - e^{-Rt}}{R} \right] + b_2 \left[1 + (c_1 - r_2) \frac{1 - e^{-Rt}}{R} \right] \\ &= \|u^\star\| K_1 \left[1 + (c_2 - r_1) \frac{1 - e^{-Rt}}{R} \right] + \|v^\star\| K_2 \left[1 + (c_1 - r_2) \frac{1 - e^{-Rt}}{R} \right]. \end{aligned}$$

It is easily seen that there is $L_1 > 0$ such that for all small $\rho > 0$,

$$\max \left\{ K_1 \left[1 + (c_2 - r_1) \frac{1 - e^{-Rt}}{R} \right], K_2 \left[1 + (c_1 - r_2) \frac{1 - e^{-Rt}}{R} \right] \right\} \le L_1 \tag{4.25}$$

for all $\mu \in I$ and $t \ge 0$. It follows from Lemma 4.3 that

$$\begin{aligned} \|u(t)\| + \|v(t)\| &\le e^{-R_0 t} \left[b_1 + b_2 + (b_2 c_1 - b_1 r_1 + b_1 c_2 - b_2 r_2) \frac{1 - e^{-Rt}}{R} \right] \\ &\le L_1 \left(\|u^\star\| + \|v^\star\| \right) e^{-\rho\mu\alpha_d t}. \end{aligned} \tag{4.26}$$

For all sufficiently small $\rho > 0$, if

$$(u^\star, v^\star) \in S_{N_1+N_2} \left(0, \frac{\varepsilon_0}{L_1} \right),$$

$$(u(t), v(t)) \in S_{N_1+N_2}(0, \varepsilon_0) \quad \text{for all } t \in [0, t_1).$$

For such $u^\star$ and $v^\star$, $\sigma(t)$, $u(t)$, and $v(t)$ exist and (4.26) is valid for all $t \in [0, \infty)$ (cf. [10]). In particular $u(t)$ and $v(t)$ decay exponentially. We summarize this result in terms of solutions of (1.8) in Proposition 4.4.

PROPOSITION 4.4. *Suppose that*

$$\begin{aligned} \sigma &= \sigma(t) = \sigma(t, \sigma^\star, w^\star, z^\star, \mu, \rho), \\ w &= w(t) = w(t, \sigma^\star, w^\star, z^\star, \mu, \rho), \\ z &= z(t) = z(t, \sigma^\star, w^\star, z^\star, \mu, \rho) \end{aligned} \tag{4.27}$$

is the solution of (1.8) *with the initial value*

$$\begin{aligned} &\sigma(0, \sigma^\star, w^\star, z^\star, \mu, \rho) = \sigma^\star, \quad w(0, \sigma^\star, w^\star, z^\star, \mu, \rho) = w^\star, \\ &z(0, \sigma^\star, w^\star, z^\star, \mu, \rho) = z^\star. \end{aligned} \tag{4.28}$$

Then there are positive constants ρ_d, α_d, ε_d, *and* L_1 *such that for all* $0 < \rho < \rho_d$, *if*

$$\|w^\star - q_\rho(\sigma^\star, \mu)\| + \|z^\star - r_\rho(\sigma^\star, \mu)\| < \varepsilon_d,$$

then (4.27) *exists for all* $t \geq 0$ *and*

$$\begin{aligned} &\|w(t) - q_\rho(\sigma(t), \mu)\| + \|z(t) - r_\rho(\sigma(t), \mu)\| \\ &\qquad \leq L_1 \left(\|w^\star - q_\rho(\sigma^\star, \mu)\| + \|z^\star - r_\rho(\sigma^\star, \mu)\|\right) e^{-\rho\mu\alpha_d t}. \end{aligned} \tag{4.29}$$

Proposition 4.4 now enables us to determine the asymptotic behavior of solutions of (1.4) near the invariant torus defined by (2.14). Let $\mathcal{F}_\delta$ denote the set of all points $(x_0, x_1, \cdots, x_N)$ in $\mathbb{R}^{M+N_2}$ given by

$$\begin{aligned} &x_i^\star = \eta_i(\omega_i^{-1}\sigma_i^\star) + \rho^\kappa \mu^\nu \Phi_i(\omega_i^{-1}\sigma_i^\star) w_i^\star, \qquad i = 1, \cdots, N, \\ &x_0^\star = z^\star + a(\Omega^{-1}\sigma^\star, \epsilon_0 \rho\mu) \quad \text{for } 0 \leq \sigma_i^\star < 2\pi \text{ and } \|y^\star\| + \|z^\star\| < \frac{\varepsilon_d}{2}, \end{aligned} \tag{4.30}$$

where the relationship between δ, ρ, μ, κ, and ν is given by (2.9) and (2.11). Then $\mathcal{F}_\delta$ is a neighborhood of the unperturbed torus defined by (1.6). In view of (3.10) and (3.12), we may assume that for $\rho \in (0, \rho_d)$ and $\mu \in I$,

$$\|q_\rho(\sigma, \mu)\| + \|r_\rho(\sigma, \mu)\| < \frac{\varepsilon_d}{2} \quad \text{for all } \sigma \in \mathbb{R}^N.$$

We will show that when $\delta > 0$ is sufficiently small, the solution of (1.4) through any point in $\mathcal{F}_\delta$ is attracted to the invariant torus defined by (2.14). Let

$$x_i = \phi_i(t) = \phi_i\left(t, x_0^\star, x_1^\star, \cdots, x_N^\star, \delta\right), \qquad i = 0, \cdots, N \tag{4.31}$$

be the solution of (1.4) with the initial value

$$\phi_i\left(0, x_0^\star, x_1^\star, \cdots, x_N^\star, \delta\right) = x_i^\star, \qquad i = 0, \cdots, N. \tag{4.32}$$

Let

$$L_2 = \max_{1\le i\le N} \sup_{0\le\theta_i\le T_i} \{\|\Phi_i(\theta_i)\|\}.$$

If $x_i^\star$, $i = 0, \cdots, N$, are given by (4.30), then

$$\phi_i(t) = \eta_i(\omega_i^{-1}\sigma_i(t)) + \rho^\kappa\mu^\nu\Phi_i(\omega_i^{-1}\sigma_i(t))w_i(t), \qquad i = 1, \cdots, N,$$
$$\phi_0(t) = z(t) + a(\Omega^{-1}\sigma(t), \epsilon_0\rho\mu),$$

where $\sigma = \sigma(t)$, $w = w(t)$, $z = z(t)$, defined in (4.27) and (4.28), is the solution of (2.12) with $\mu = \delta^{1-p}/\rho$. It follows that

$$\begin{aligned}
&\sum_{i=1}^N \left\|\phi_i(t) - \left[\eta_i(\omega_i^{-1}\sigma_i(t)) + \rho^\kappa\mu^\nu\Phi_i(\omega_i^{-1}\sigma_i(t))q_{\rho,i}(\sigma(t),\mu)\right]\right\| \\
&\quad + \left\|\phi_0(t) - \left[r_\rho(\sigma(t),\mu) + a(\Omega^{-1}\sigma(t),\epsilon_0\rho\mu)\right]\right\| \\
&\qquad \le \rho^\kappa\mu^\nu L_2 \sum_{i=1}^N \|w_i(t) - q_{\rho,i}(\sigma(t),\mu)\| + \|z(t) - r_\rho(\sigma(t),\mu)\| \\
&\qquad \le (\rho^\kappa\mu^\nu L_2 + 1)\left(\|w(t) - q_\rho(\sigma(t),\mu)\| + \|z(t) - r_\rho(\sigma(t),\mu)\|\right) \\
&\qquad \le (\rho^\kappa\mu^\nu L_2 + 1)\, L_1 \left(\|w^\star - q_\rho(\sigma^\star,\mu)\| + \|z^\star - r_\rho(\sigma^\star,\mu)\|\right) e^{-\rho\mu\alpha_d t}.
\end{aligned}$$

We summarize this result in Theorem 4.5.

THEOREM 4.5. *Under Assumption* 1.1 *with* $1 - 1/k \le p < 1$, *there are positive numbers* δ_d, ε_d, α_d, *and* L_d *with the following properties. We define a region* $\mathcal{F}_\delta$ *by* (4.30). *Then* $\mathcal{F}_\delta$ *is a neighborhood of the unperturbed torus defined by* (1.7). *For* $0 < \delta \le \delta_d$ *the solution of* (1.4) *through a point in* $\mathcal{F}_\delta$ *is attracted to the invariant torus defined by* (2.14) *in the following sense.*

$$\begin{aligned}
&\sum_{i=1}^N \left\|\phi_i(t) - \left[\eta_i(\omega_i^{-1}\sigma_i(t)) + \rho^\kappa\mu^\nu\Phi_i(\omega_i^{-1}\sigma_i(t))q_{\rho,i}(\sigma(t),\mu)\right]\right\| \\
&\quad + \left\|\phi_0(t) - \left[r_\rho(\sigma(t),\mu) + a(\Omega^{-1}\sigma(t),\epsilon_0\rho\mu)\right]\right\| \\
&\qquad \le L_d \left(\|w^\star - q_\rho(\sigma^\star,\mu)\| + \|z^\star - r_\rho(\sigma^\star,\mu)\|\right) e^{-\rho\mu\alpha_d t}.
\end{aligned} \tag{4.33}$$

In (4.33), $\phi_i(t)$, $i = 0, \cdots, N$, *are the solutions of* (1.4) *defined in* (4.31) *and* (4.32). $\sigma(t)$, $w(t)$, *and* $z(t)$ *are the solutions of* (2.12) *defined in* (4.27) *and* (4.28). *Here the relationship between* $x_i^\star$, $i = 0, \cdots, N$, *and* $(\sigma^\star, y^\star, z^\star)$ *is given in* (4.30). (q_ρ, r_ρ) *is an integral manifold of* (2.12) *that defines an invariant torus of* (1.4) *via* (2.14).

Next we study the behavior of solutions of (1.9) near the integral manifold $w = q_\mu(\sigma)$, $z = r_\mu(\sigma)$. Let

$$w = u + q_\mu(\sigma),$$
$$z = v + r_\mu(\sigma).$$

Then (1.9) becomes

$$\begin{aligned}
\frac{d\sigma}{dt} &= \omega + \mathcal{S}_0(\sigma, u, v, \mu), \\
\frac{du}{dt} &= A(\Omega^{-1}\sigma)u + U_1(\sigma, u, v, \mu)u + U_2(\sigma, u, v, \mu)v, \\
\frac{dv}{dt} &= \mu Q v + V_1(\sigma, u, v, \mu)u + V_2(\sigma, u, v, \mu)v,
\end{aligned} \tag{4.34}$$

where

$$\begin{aligned}
\mathcal{S}_0(\sigma, u, v, \mu) &= \mathcal{S}(\sigma, u + q_\mu(\sigma), v + r_\mu(\sigma), \mu) \\
&= \Sigma(\sigma, \mu) + S_1(\sigma, u, v, \mu)u + S_2(\sigma, u, v, \mu)v, \\
\Sigma(\sigma, \mu) &= \mathcal{S}(\sigma, q_\mu(\sigma), r_\mu(\sigma), \mu), \\
S_1(\sigma, u, v, \mu) &= \int_0^1 \frac{\partial S}{\partial w}(\sigma, su + q_\mu(\sigma), sv + r_\mu(\sigma), \mu)\, ds, \\
S_2(\sigma, u, v, \mu) &= \int_0^1 \frac{\partial S}{\partial z}(\sigma, su + q_\mu(\sigma), sv + r_\mu(\sigma), \mu)\, ds,
\end{aligned}$$

and

$$\begin{aligned}
U_1(\sigma, u, v, \mu) &= \int_0^1 \frac{\partial \mathcal{W}}{\partial w}(\sigma, su + q_\mu(\sigma), sv + r_\mu(\sigma), \mu)\, ds - \frac{\partial q_\mu}{\partial \sigma}(\sigma) S_1(\sigma, u, v, \mu), \\
U_2(\sigma, u, v, \mu) &= \int_0^1 \frac{\partial \mathcal{W}}{\partial z}(\sigma, su + q_\mu(\sigma), sv + r_\mu(\sigma), \mu)\, ds - \frac{\partial q_\mu}{\partial \sigma}(\sigma) S_2(\sigma, u, v, \mu), \\
V_1(\sigma, u, v, \mu) &= \mu \int_0^1 \frac{\partial \mathcal{Z}}{\partial w}(\sigma, su + q_\mu(\sigma), sv + r_\mu(\sigma), \mu)\, ds - \frac{\partial r_\mu}{\partial \sigma}(\sigma) S_1(\sigma, u, v, \mu), \\
V_2(\sigma, u, v, \mu) &= \mu \int_0^1 \frac{\partial \mathcal{Z}}{\partial z}(\sigma, su + q_\mu(\sigma), sv + r_\mu(\sigma), \mu)\, ds - \frac{\partial r_\mu}{\partial \sigma}(\sigma) S_2(\sigma, u, v, \mu).
\end{aligned}$$

Choose a neighborhood of the origin $\mathcal{D}$ in $\mathbb{R}^{N_1+N_2}$ that satisfies (4.2). Then for $\mu \in I$, $\mathcal{S}_0$, U_1, U_2, V_1, and V_2 are defined and continuously differentiable in $\mathbb{R}^N \times \mathcal{D}$. Choose a positive number ε_0 that satisfies (4.3). For $\mu \in I$ and $(u, v) \in \overline{S_{N_1+N_2}(0, \varepsilon_0)}$,

$$\|\mathcal{S}_0(\sigma, u, v, \mu)\| \leq \lambda(\mu), \tag{4.35}$$

$$\begin{aligned}
\|U_i(\sigma, u, v, \mu)\| &\leq \lambda_1(\mu) + \lambda(\mu)\Delta, \qquad i = 1, 2, \\
\|V_i(\sigma, u, v, \mu)\| &\leq \lambda_2(\mu)\mu + \lambda(\mu)\Delta, \qquad i = 1, 2.
\end{aligned} \tag{4.36}$$

Let

$$\begin{aligned} \sigma &= \sigma(t) = \sigma(t, \sigma^\star, u^\star, v^\star, \mu), \\ u &= u(t) = u(t, \sigma^\star, u^\star, v^\star, \mu), \\ v &= v(t) = v(t, \sigma^\star, u^\star, v^\star, \mu) \end{aligned} \tag{4.37}$$

be the solution of (4.34) with initial value

$$\sigma(0, \sigma^\star, u^\star, v^\star, \mu) = \sigma^\star, \quad u(0, \sigma^\star, u^\star, v^\star, \mu) = u^\star, \quad v(0, \sigma^\star, u^\star, v^\star, \mu) = v^\star.$$

If $(u^\star, v^\star) \in S_{N_1+N_2}(0, \varepsilon_0)$, there is $t_1 > 0$ such that (4.37) exists and $(u(t), v(t)) \in S_{N_1+N_2}(0, \varepsilon_0)$ for $t \in [0, t_1)$. Let $u = \Xi(t, t_0) = \Xi(t, t_0, \sigma^\star, u^\star, v^\star, \mu)$ be the fundamental matrix solution of

$$\frac{du}{dt} = A(\Omega^{-1}\sigma(t, \sigma^\star, u^\star, v^\star, \mu))u,$$

$$\Xi(t_0, t_0, \sigma^\star, u^\star, v^\star, \mu) = I_{N_1 \times N_1}.$$

We may assume without loss of generality that for $\mu \in I$,

$$\lambda(\mu) < \omega_i, \qquad i = 1, \cdots, N.$$

It is shown in Lemma 4.6 that $\Xi(t, t_0)$ decays exponentially as t increases. Again its proof is similar to the one in Lemma 5.8 and is left to the reader.

LEMMA 4.6.

$$\|\Xi(t, t_0, \sigma^\star, u^\star, v^\star, \mu)\| \le K_1 e^{-[\beta_1 - \lambda(\mu)H](t-t_0)} \quad \textit{for } t_0 \le t.$$

Write $u(t)$ and $v(t)$ in the form

$$\begin{aligned} u(t) = \Xi(t, 0)u^\star + \int_0^t \Xi(t, s) \, [&U_1(\sigma(s), u(s), v(s), \mu)u(s) \\ &+ U_2(\sigma(s), u(s), v(s), \mu)v(s)] \, ds, \end{aligned}$$

$$\begin{aligned} v(t) = e^{\mu Q t}v^\star + \int_0^t e^{\mu Q(t-s)} \, [&V_1(\sigma(s), u(s), v(s), \mu)u(s) \\ &+ V_2(\sigma(s), u(s), v(s), \mu)v(s)] \, ds. \end{aligned}$$

Equations (4.36), (4.37), and Lemma 4.6 lead to

$$\begin{aligned} \|u(t)\| &\le K_1\|u^\star\|e^{-[\beta_1 - \lambda(\mu)H]t} \\ &+ \int_0^t K_1 \, [\lambda_1(\mu) + \lambda(\mu)\Delta] \, e^{-[\beta_1 - \lambda(\mu)H](t-s)} \, (\|u(s)\| + \|v(s)\|) \, ds, \\ \|v(t)\| &\le K_2\|v^\star\|e^{-\mu\beta_2 t} + \int_0^t K_2 \, [\lambda_2(\mu)\mu + \lambda(\mu)\Delta] \, e^{-\mu\beta_2(t-s)} \, (\|u(s)\| + \|v(s)\|) \, ds. \end{aligned}$$

Now let

$$
\begin{aligned}
a_1 &= \beta_1 - \lambda(\mu)H, \qquad b_1 = K_1\|u^\star\|, \\
c_1 &= K_1\left[\lambda_1(\mu) + \lambda(\mu)\Delta\right], \\
a_2 &= \mu\beta_2, \qquad b_2 = K_2\|v^\star\|, \\
c_2 &= K_2\left[\lambda_2(\mu)\mu + \lambda(\mu)\Delta\right].
\end{aligned} \tag{4.38}
$$

Then using an argument similar to the one leading to (4.24) and (4.25), we conclude that there is a $\tilde{\alpha}_d > 0$ such that for all small $\mu > 0$, R_0 defined by (4.23) and (4.38) satisfies

$$R_0 > \mu\tilde{\alpha}_d$$

for all $\mu \in I$, and that there is $\widetilde{L}_1 > 0$ such that for all small $\mu > 0$,

$$\max\left\{K_1\left[1 + (c_2 - r_1)\frac{1 - e^{-Rt}}{R}\right], K_2\left[1 + (c_1 - r_2)\frac{1 - e^{-Rt}}{R}\right]\right\} \leq \widetilde{L}_1$$

for all $t \geq 0$. It follows from Lemma 4.3 that

$$
\begin{aligned}
\|u(t)\| + \|v(t)\| &\leq e^{-R_0 t}\left[b_1 + b_2 + (b_2c_1 - b_1r_1 + b_1c_2 - b_2r_2)\frac{1 - e^{-Rt}}{R}\right], \\
&\leq \widetilde{L}_1\left(\|u^\star\| + \|v^\star\|\right)e^{-\mu\tilde{\alpha}_d t}.
\end{aligned} \tag{4.39}
$$

For all sufficiently small $\rho > 0$, if

$$
\begin{aligned}
(u^\star, v^\star) &\in S_{N_1+N_2}\left(0, \frac{\varepsilon_0}{\widetilde{L}_1}\right), \\
(u(t), v(t)) &\in S_{N_1+N_2}(0, \varepsilon_0) \quad \text{for all } t \in [0, t_1).
\end{aligned}
$$

For such $u^\star$ and $v^\star$, $\sigma(t)$, $u(t)$, and $v(t)$ exist and (4.39) is valid for all $t \in [0, \infty)$ (cf. [10]). In particular $u(t)$ and $v(t)$ decay exponentially. We summarize this result in Proposition 4.7.

PROPOSITION 4.7. *Suppose that*

$$
\begin{aligned}
\sigma &= \sigma(t) = \sigma(t, \sigma^\star, w^\star, z^\star, \mu), \\
w &= w(t) = w(t, \sigma^\star, w^\star, z^\star, \mu), \\
z &= z(t) = z(t, \sigma^\star, w^\star, z^\star, \mu)
\end{aligned} \tag{4.40}
$$

is the solution of (1.9) *with the initial value*

$$
\begin{aligned}
\sigma(0, \sigma^\star, w^\star, z^\star, \mu) &= \sigma^\star, \\
w(0, \sigma^\star, w^\star, z^\star, \mu) &= w^\star, \\
z(0, \sigma^\star, w^\star, z^\star, \mu) &= z^\star.
\end{aligned} \tag{4.41}
$$

Then there are positive constants μ_d, $\tilde{\alpha}_d$, $\tilde{\varepsilon}_d$, *and* $\widetilde{L}_1$ *such that for all* $\mu \in \mu_d$, *if*

$$\|w^\star - q_\mu(\sigma^\star)\| + \|z^\star - r_\mu(\sigma^\star)\| < \tilde{\varepsilon}_d,$$

then (4.40) *exists for all* $t \geq 0$ *and*

$$\begin{aligned}\|w(t) - q_\mu(\sigma(t))\| + \|z(t) - r_\mu(\sigma(t))\| \\ \leq \widetilde{L}_1 \left(\|w^\star - q_\mu(\sigma^\star)\| + \|z^\star - r_\mu(\sigma^\star)\|\right) e^{-\mu\tilde{\alpha}_d t}.\end{aligned}$$

Next we study the behavior of solutions of (1.4) near the invariant torus defined by (2.15). Let $\widetilde{\mathcal{F}}_\delta$ denote the set of all points $(x_0, x_1, \cdots, x_N)$ in $\mathbb{R}^{M+N_2}$ given by

$$\begin{aligned} x_i^\star &= \eta_i(\omega_i^{-1}\sigma_i^\star) + \mu^\zeta \Phi_i(\omega_i^{-1}\sigma_i^\star) w_i^\star, \qquad i = 1, \cdots, N, \\ x_0^\star &= z^\star + a(\Omega^{-1}\sigma^\star, \epsilon_0\mu) \quad \text{for } 0 \leq \sigma_i^\star < 2\pi \text{ and } \|y^\star\| + \|z^\star\| < \frac{\tilde{\varepsilon}_d}{2}, \end{aligned} \tag{4.42}$$

where the relationship between δ, μ, ζ, and ν is given by (2.10) and (2.11). Then $\widetilde{\mathcal{F}}_\delta$ is a neighborhood of the unperturbed torus defined by (1.6). In view of (3.10) and (3.12), we may assume that for a sufficiently small $\delta > 0$,

$$\|q_\mu(\sigma)\| + \|r_\mu(\sigma)\| < \frac{\tilde{\varepsilon}_d}{2} \quad \text{for all } \sigma \in \mathbb{R}^N.$$

We will show that the solution of (1.4) through any point in $\widetilde{\mathcal{F}}_\delta$ is attracted to the invariant torus. Consider the solution of (1.4) defined in (4.31) and (4.32). If $x_i^\star$, $i = 0, \cdots, N$, are given by (4.42), then

$$\begin{aligned} \phi_i(t) &= \eta_i(\omega_i^{-1}\sigma_i(t)) + \mu^\zeta \Phi_i(\omega_i^{-1}\sigma_i(t)) w_i(t), \qquad i = 1, \cdots, N, \\ \phi_0(t) &= z(t) + a(\Omega^{-1}\sigma(t), \epsilon_0\mu), \end{aligned}$$

and it follows that

$$\begin{aligned} &\sum_{i=1}^N \left\|\phi_i(t) - \left[\eta_i(\omega_i^{-1}\sigma_i(t)) + \mu^\zeta \Phi_i(\omega_i^{-1}\sigma_i(t)) q_{\mu,i}(\sigma(t))\right]\right\| \\ &\qquad + \left\|\phi_0(t) - \left[r_\mu(\sigma(t)) + a(\Omega^{-1}\sigma(t), \epsilon_0\mu)\right]\right\| \\ &\leq \mu^\zeta L_2 \sum_{i=1}^N \|w_i(t) - q_{\mu,i}(\sigma(t))\| + \|z(t) - r_\mu(\sigma(t))\| \\ &\leq \left(\mu^\zeta L_2 + 1\right) \left(\|w(t) - q_\mu(\sigma(t))\| + \|z(t) - r_\mu(\sigma(t))\|\right) \\ &\leq \left(\mu^\zeta L_2 + 1\right) \widetilde{L}_1 \left(\|w^\star - q_\mu(\sigma^\star)\| + \|z^\star - r_\mu(\sigma^\star)\|\right) e^{-\mu\tilde{\alpha}_d t}. \end{aligned}$$

We summarize this result in Theorem 4.8.

THEOREM 4.8. *Under Assumption* 1.1 *with* $0 < p < 1$, *there are positive numbers* $\tilde{\delta}_d$, $\tilde{\varepsilon}_d$, $\tilde{\alpha}_d$, *and* $\widetilde{L}_d$ *with the following properties. We define a region* $\widetilde{\mathcal{F}}_\delta$ *by* (4.42). *Then* $\widetilde{\mathcal{F}}_\delta$ *is a neighborhood of the unperturbed torus defined by* (1.7). *For* $0 < \delta \leq \tilde{\delta}_d$ *the solution of* (1.4) *through a point in* $\widetilde{\mathcal{F}}_\delta$ *is attracted to the invariant torus defined by* (2.15) *in the following sense.*

$$
\begin{aligned}
&\sum_{i=1}^{N} \left\| \phi_i(t) - \left[\eta_i(\omega_i^{-1}\sigma_i(t)) + \mu^{\zeta}\Phi_i(\omega_i^{-1}\sigma_i(t)) q_{\mu,i}(\sigma(t)) \right] \right\| \\
&\qquad + \left\| \phi_0(t) - \left[r_\mu(\sigma(t)) + a(\Omega^{-1}\sigma(t), \epsilon_0\mu) \right] \right\| \\
&\qquad\qquad \leq \widetilde{L}_d \left(\|w^\star - q_\mu(\sigma^\star)\| + \|z^\star - r_\mu(\sigma^\star)\| \right) e^{-\mu\tilde{\alpha}_d t}.
\end{aligned} \tag{4.43}
$$

In (4.43), $\phi_i(t)$, $i = 0, \cdots, N$, *are the solutions of* (1.4) *defined in* (4.31) *and* (4.32). $\sigma(t)$, $w(t)$, *and* $z(t)$ *are the solutions of* (1.9) *defined in* (4.40) *and* (4.41). *Here the relationship between* $x_i^\star$, $i = 0, \cdots, N$, *and* $(\sigma^\star, y^\star, z^\star)$ *is given in* (4.42). (q_μ, r_μ) *is an integral manifold of* (2.13) *that defines an invariant torus of* (1.4) *via* (2.15).

5. Appendix.

5.1. Properties of flows along periodic surfaces. In this Appendix we will prove a number of technical lemmas that are used to prove the existence of integral manifolds in §3. Specifically, in Lemmas 5.1–5.11, we will develop several properties of $\phi(t, \sigma^0, \mu, q, r, \rho)$ and $\Lambda(t, t_0, \sigma^0, \mu, q, r, \rho)$, which are defined as solutions of (3.1) and (3.2), respectively, under Assumption 2.1(a)–(e). These properties are used in the existence proofs for (1.8). Similar results for $\phi(t, \sigma^0, q, r, \mu)$ and $\Lambda(t, t_0, \sigma^0, q, r, \mu)$, defined as solutions of (3.18) and (3.19), respectively, under Assumption 2.1(a–c) and (f) are stated in Property 5.13 and they will be applied to (1.9).

Choose Δ such that

$$
\overline{S_{N_i}(0, \Delta)} \subset F_i, \qquad i = 1, 2.
$$

Recall that $\phi(t, \sigma^0, \mu, q, r, \rho)$ is a solution of (3.1) where $\mathcal{S}$ satisfies Assumption 2.1(e) and $(q, r) \in B_\Delta^k(\mathbb{R}^N \times I, \mathbb{R}^{N_1}) \times B_\Delta^k(\mathbb{R}^N \times I, \mathbb{R}^{N_2})$. For the following Lemmas 5.1–5.11 assume that $(q, r), (q_1, r_1), (q_2, r_2) \in B_\Delta^k(\mathbb{R}^N \times I, \mathbb{R}^{N_1}) \times B_\Delta^k(\mathbb{R}^N \times I, \mathbb{R}^{N_2})$. The proof of Lemma 5.1 is left to the reader.

LEMMA 5.1. *For each* $\rho \in I_1$, $\phi(t, \sigma^0, \mu, q, r, \rho)$ *exists for all* $t \in \mathbb{R}$, $\sigma^0 \in \mathbb{R}^N$, *and* $\mu \in I$, *and it is* k*-times continuously differentiable with respect to* t, σ^0, *and* μ. *Moreover, for* $i = 1, \cdots, N$,

$$
\phi(t, \sigma^0 + 2\pi e^i, \mu, q, r, \rho) = \phi(t, \sigma^0, \mu, q, r, \rho) + 2\pi e^i.
$$

Define

$$
\Upsilon(t, s, \sigma^0, \mu, q, r, \rho) = \frac{\partial \phi}{\partial \sigma^0}(t, \sigma^0, \mu, q, r, \rho) \frac{\partial \phi^{-1}}{\partial \sigma^0}(s, \sigma^0, \mu, q, r, \rho).
$$

LEMMA 5.2. *For all* $t, s \in \mathbb{R}$,

$$
\|\Upsilon(t, s, \sigma^0, \mu, q, r, \rho)\| \leq e^{\lambda(\rho)\mu^k(1+2\Delta)|t-s|}. \tag{5.1}
$$

In particular, for $t \leq 0$,

$$
\left\| \frac{\partial \phi}{\partial \sigma^0}(t, \sigma^0, \mu, q, r, \rho) \right\| \leq e^{-\lambda(\rho)\mu^k(1+2\Delta)t}. \tag{5.2}
$$

Proof. $x = \Upsilon(t, s, \sigma^0, \mu, q, r, \rho)$ satisfies the following linear system.

$$\frac{dx}{dt} = R(t)x, \qquad \Upsilon(s, s, \sigma^0, \mu, q, r, \rho) = I,$$

where

$$\begin{aligned} R(t) = {} & \frac{\partial \mathcal{S}}{\partial \sigma}(\phi(t), q(\phi(t), \mu), r(\phi(t), \mu), \mu, \rho) \\ & + \frac{\partial \mathcal{S}}{\partial w}(\phi(t), q(\phi(t), \mu), r(\phi(t), \mu), \mu, \rho) \frac{\partial q}{\partial \sigma}(\phi(t), \mu) \\ & + \frac{\partial \mathcal{S}}{\partial z}(\phi(t), q(\phi(t), \mu), r(\phi(t), \mu), \mu, \rho) \frac{\partial r}{\partial \sigma}(\phi(t), \mu), \end{aligned} \tag{5.3}$$

$$\phi(t) = \phi(t, \sigma^0, \mu, q, r, \rho).$$

Therefore

$$\Upsilon(t, s, \sigma^0, \mu, q, r, \rho) = I + \int_s^t R(u)\Upsilon(u, s, \sigma^0, \mu, q, r, \rho)\, du$$

and it follows from Assumption 2.1(d) that

$$\left\|\Upsilon(t, s, \sigma^0, \mu, q, r, \rho)\right\| \le 1 + \left| \int_s^t \lambda(\rho)\mu^k(1 + 2\Delta) \left\|\Upsilon(u, s, \sigma^0, \mu, q, r, \rho)\right\| du \right|.$$

Now (5.1) follows from Gronwall's inequality (cf. [1]) and (5.2) follows from (5.1) by setting $s = 0$. □

LEMMA 5.3. *For $2 \le l + 2m$ and $l + m \le k$ there is a positive constant $K_{l,m}$ such that for $t \le 0$,*

$$\begin{aligned} & \left\| \frac{\partial^{l+m}\phi}{\partial \sigma_{i_1}^0 \cdots \partial \sigma_{i_l}^0 \partial \mu^m}(t, \sigma^0, \mu, q, r, \rho) \right\| \\ & \le K_{l,m}\mu^{-m} \left\{ \sum_{i=1}^{l+2m-1} \left[\lambda(\rho)\mu^k(-t)\right]^i \right\} e^{-(l+m)\lambda(\rho)\mu^k(1+2\Delta)t}. \end{aligned} \tag{5.4}$$

Proof. It is easily shown inductively that for $2 \le l + 2m$ and $l + m \le k$

$$x = \frac{\partial^{l+m}\phi}{\partial \sigma_{i_1}^0 \cdots \partial \sigma_{i_l}^0 \partial \mu^m}(t, \sigma^0, \mu, q, r, \rho)$$

satisfies the following nonhomogeneous linear system:

$$\begin{aligned} & \frac{dx}{dt} = R(t)x + f_{(i_1, \cdots, i_l, m)}(t), \\ & \frac{\partial^{l+m}\phi}{\partial \sigma_{i_1}^0 \cdots \partial \sigma_{i_l}^0 \partial \mu^m}(0, \sigma^0, \mu, q, r, \rho) = 0. \end{aligned} \tag{5.5}$$

System (5.5) has the following properties.

PROPERTY 5.4. Matrix $R(t)$ is given by (5.3). $f_{(i_1,\cdots,i_l,m)}(t)$ is a (finite) sum of terms of the form

$$\frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{S}}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}}a_1\cdots a_{m_1}B_1\cdots B_{m_2}C_1\cdots C_{m_3},$$

where

$$\frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{S}}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}} = \frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{S}}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}}(\phi(t),q(\phi(t),\mu),r(\phi(t),\mu),\mu,\rho),$$

$$a_v = \frac{\partial^{l_{1,v}+m_{1,v}}\phi}{\partial\sigma^0_{j_{v,1}}\cdots\partial\sigma^0_{j_{v,l_{1,v}}}\partial\mu^{m_{1,v}}}(t), \qquad v=1,\cdots,m_1,$$

$$B_u = \frac{\partial^{\nu_{2,u}+\chi_{2,u}}q}{\partial\sigma^{\nu_{2,u}}\mu^{\chi_{2,u}}}(\phi(t),\mu)b_{u,1}\cdots b_{u,\nu_{2,u}}, \qquad u=1,\cdots,m_2,$$

$$b_{u,v} = \frac{\partial^{l_{2,u,v}+m_{2,u,v}}\phi}{\partial\sigma^0_{j_{2,u,v,1}}\cdots\partial\sigma^0_{j_{2,u,v,l_{2,u,v}}}\partial\mu^{m_{2,u,v}}}(t), \qquad v=1,\cdots,\nu_{2,u},$$

$$C_u = \frac{\partial^{\nu_{3,u}+\chi_{3,u}}q}{\partial\sigma^{\nu_{3,u}}\mu^{\chi_{3,u}}}(\phi(t),\mu)c_{u,1}\cdots c_{u,\nu_{3,u}}, \qquad u=1,\cdots,m_3,$$

$$c_{u,v} = \frac{\partial^{l_{3,u,v}+m_{3,u,v}}\phi}{\partial\sigma^0_{j_{3,u,v,1}}\cdots\partial\sigma^0_{j_{3,u,v,l_{3,u,v}}}\partial\mu^{m_{3,u,v}}}(t), \qquad v=1,\cdots,\nu_{3,u},$$

$$0\le m_i\le l+m,\ i=1,\cdots,4, \qquad 1\le m_1+m_2+m_3+m_4\le l+m,$$

$$1\le l_{1,v}+m_{1,v},\ l_{2,u,v}+m_{2,u,v},\ l_{3,u,v}+m_{3,u,v}\le l+m-1,$$

$$\sum_{v=1}^{m_1}l_{1,v}+\sum_{u=1}^{m_2}\sum_{v=1}^{\nu_{2,u}}l_{2,u,v}+\sum_{u=1}^{m_3}\sum_{v=1}^{\nu_{3,u}}l_{3,u,v}=l,$$

$$m_4+\sum_{v=1}^{m_1}m_{1,v}+\sum_{u=1}^{m_2}\left(\sum_{v=1}^{\nu_{2,u}}m_{2,u,v}+\chi_{2,u}\right)+\sum_{u=1}^{m_3}\left(\sum_{v=1}^{\nu_{3,u}}m_{3,u,v}+\chi_{3,u}\right)=m.$$

Note that

$$\text{(5.6)}\qquad \text{if } m_4+\sum_{u=1}^{m_2}\chi_{2,u}+\sum_{u=1}^{m_3}\chi_{3,u}=0 \quad \text{then } m_1+\sum_{u=1}^{m_2}\nu_{2,u}+\sum_{u=1}^{m_3}\nu_{3,u}\ge 2.$$

By the variation of constants formula (cf. [7], [8]),

$$\text{(5.7)}\qquad \frac{\partial^{l+m}\phi}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(t,\sigma^0,\mu,q,r,\rho)=\int_0^t\Upsilon(t,s,\sigma^0,\mu,q,r,\rho)f_{(i_1,\cdots,i_l,m)}(s)\,ds.$$

Now we prove (5.4) inductively. By Lemma 5.2, Property 5.4, and (5.6), there is $K_{2,0}>0$ such that for $t\le 0$,

$$\left\|f_{(i_1,i_2,0)}(t)\right\| \le K_{2,0}\lambda(\rho)\mu^k e^{-2\lambda(\rho)\mu^k(1+2\Delta)t}.$$

By Lemma 5.2 and (5.7)

$$\begin{aligned}\left\|\frac{\partial^2\phi}{\partial\sigma^0_{i_1}\partial\sigma^0_{i_2}}(t,\sigma^0,\mu,q,r,\rho)\right\| &\le \int_t^0 \|\Upsilon(t,s,\sigma^0,\mu,q,r,\rho)\|\|f_{(i_1,i_l,0)}(s)\|\,ds\\ &\le K_{2,0}\lambda(\rho)\mu^k(-t)e^{-2\lambda(\rho)\mu^k(1+2\Delta)t}.\end{aligned}$$

Again by Property 5.4, there is $K_{0,1} > 0$ such that

$$\left\|f_{(1)}(t)\right\| \le K_{0,1}\lambda(\rho)\mu^{k-1},$$

and by Lemma 5.2 and (5.7),

$$\begin{aligned}\left\|\frac{\partial\phi}{\partial\mu}(t,\sigma^0,\mu,q,r,\rho)\right\| &\le \int_t^0 \|\Upsilon(t,s,\sigma^0,\mu,q,r,\rho)\|\|f_{(1)}(s)\|\,ds\\ &\le \frac{K_{0,1}\lambda(\rho)\mu^{k-1}}{\lambda(\rho)\mu^k(1+2\Delta)}\left(e^{-\lambda(\rho)\mu^k(1+2\Delta)t}-1\right)\\ &\le K_{0,1}\lambda(\rho)\mu^{k-1}(-t)e^{-\lambda(\rho)\mu^k(1+2\Delta)t}.\end{aligned}$$

Thus (5.4) holds for $l + 2m = 2$.

Assume that (5.4) holds for all $\hat{l}$ and $\hat{m}$ such that $2 \le \hat{l}+2\hat{m}$ and $\hat{l}+\hat{m} \le l+m-1$. By Lemma 5.2, Property 5.4, and (5.6) there is $K_{l,m} > 0$ such that for $t \le 0$,

$$\left\|f_{(i_1,\cdots,i_l,m)}(t)\right\| \le K_{l,m}\lambda(\rho)\mu^{k-m}\left\{\sum_{i=0}^{l+2m-2}\left[\lambda(\rho)\mu^k(-t)\right]^i\right\}e^{-(l+m)\lambda(\rho)\mu^k(1+2\Delta)t}.$$

Now (5.4) follows from (5.7). □

LEMMA 5.5. *For $0 \le l + m \le k - 1$ there is a positive constant $\widehat{K}_{l,m}$ such that for $t \le 0$,*

$$\begin{aligned}&\left\|\frac{\partial^{l+m}\phi}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(t,\sigma^0,\mu,q_1,r_1,\rho) - \frac{\partial^{l+m}\phi}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(t,\sigma^0,\mu,q_2,r_2,\rho)\right\|\\ &\le \widehat{K}_{l,m}\left(\|q_1-q_2\|_{l+m} + \|r_1-r_2\|_{l+m}\right)\\ &\quad\times\mu^{-m}\left\{\sum_{i=1}^{l+2m+1}\left[\lambda(\rho)\mu^k(-t)\right]^i\right\}e^{-(l+m+1)\lambda(\rho)\mu^k(1+2\Delta)t}.\end{aligned} \tag{5.8}$$

Proof. The proof is by induction. Let

$$\phi_1(t) = \phi(t,\sigma^0,\mu,q_1,r_1,\rho), \qquad \phi_2(t) = \phi(t,\sigma^0,\mu,q_2,r_2,\rho).$$

Since

$$\phi_1(t) - \phi_2(t) = \int_0^t [\mathcal{S}_1 - \mathcal{S}_2]\, ds,$$

where

$$\mathcal{S}_1 = \mathcal{S}(\phi_1(s), q_1(\phi_1(s), \mu), r_1(\phi_1(s), \mu), \mu, \rho),$$

$$\mathcal{S}_2 = \mathcal{S}(\phi_2(s), q_2(\phi_2(s), \mu), r_2(\phi_2(s), \mu), \mu, \rho),$$

and

$$\|\mathcal{S}_1 - \mathcal{S}_2\| \leq \lambda(\rho)\mu^k \left[\|q_1 - q_2\|_0 + \|r_1 - r_2\|_0 + (1 + 2\Delta) \|\phi_1(s) - \phi_2(s)\|\right],$$

it follows that

$$\|\phi_1(t) - \phi_2(t)\| \leq \lambda(\rho)\mu^k \left(\|q_1 - q_2\|_0 + \|r_1 - r_2\|_0\right)(-t)$$

$$+ \int_t^0 \lambda(\rho)\mu^k(1 + 2\Delta) \|\phi_1(s) - \phi_2(s)\|\, ds.$$

By Gronwall's inequality

$$\|\phi_1(t) - \phi_2(t)\|$$

$$\leq \frac{\lambda(\rho)\mu^k \left(\|q_1 - q_2\|_0 + \|r_1 - r_2\|_0\right)}{\lambda(\rho)\mu^k(1 + 2\Delta)} \left(e^{-\lambda(\rho)\mu^k(1+2\Delta)t} - 1\right)$$

$$\leq \left(\|q_1 - q_2\|_0 + \|r_1 - r_2\|_0\right) \lambda(\rho)\mu^k(-t)e^{-\lambda(\rho)\mu^k(1+2\Delta)t}.$$

Therefore (5.8) holds for $l + m = 0$.

It is easily shown that for $1 \leq l + m \leq k - 1$,

$$x = \frac{\partial^{l+m}\phi_1}{\partial\sigma_{i_1}^0 \cdots \partial\sigma_{i_l}^0 \partial\mu^m}(t) - \frac{\partial^{l+m}\phi_2}{\partial\sigma_{i_1}^0 \cdots \partial\sigma_{i_l}^0 \partial\mu^m}(t)$$

satisfies the following nonhomogeneous linear system:

$$\begin{aligned} &\frac{dx}{dt} = R_1(t)x + \widehat{f}_{(i_1,\cdots,i_l,m)}(t), \\ &\frac{\partial^{l+m}\phi_1}{\partial\sigma_{i_1}^0 \cdots \partial\sigma_{i_l}^0 \partial\mu^m}(0) - \frac{\partial^{l+m}\phi_2}{\partial\sigma_{i_1}^0 \cdots \partial\sigma_{i_l}^0 \partial\mu^m}(0) = 0. \end{aligned} \tag{5.9}$$

System (5.9) has the following properties.

PROPERTY 5.6.

$$R_1(t) = \frac{\partial\mathcal{S}}{\partial\sigma}(\phi_1(t), q_1(\phi_1(t), \mu), r_1(\phi_1(t), \mu), \mu, \rho)$$

$$+ \frac{\partial\mathcal{S}}{\partial w}(\phi_1(t), q_1(\phi_1(t), \mu), r_1(\phi_1(t), \mu), \mu, \rho) \frac{\partial q_1}{\partial\sigma}(\phi_1(t), \mu)$$

$$+ \frac{\partial\mathcal{S}}{\partial z}(\phi_1(t), q_1(\phi_1(t), \mu), r_1(\phi_1(t), \mu), \mu, \rho) \frac{\partial r_1}{\partial\sigma}(\phi_1(t), \mu)$$

and $\widehat{f}_{(i_1,\cdots,i_l,m)}(t)$ is a (finite) sum of terms of the forms

$$(5.10)\quad \frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{S}_1}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}}a_1^1\cdots a_{m_1}^1B_1^1\cdots B_{m_2}^1C_1^2\cdots C_{u-1}^2\left(C_u^1-C_u^2\right)C_{u+1}^1\cdots C_{m_3}^1,$$

$$(5.11)\quad \frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{S}_1}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}}a_1^1\cdots a_{m_1}^1B_1^2\cdots B_{u-1}^2\left(B_u^1-B_u^2\right)B_{u+1}^1\cdots B_{m_2}^1C_1^2\cdots C_{m_3}^2,$$

$$(5.12)\quad \frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{S}_1}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}}a_1^2\cdots a_{v-1}^2\left(a_v^1-a_v^2\right)a_{v+1}^1\cdots a_{m_1}^1B_1^2\cdots B_{m_2}^2C_1^2\cdots C_{m_3}^2,$$

$$(5.13)\quad \left(\frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{S}_1}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}}-\frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{S}_2}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}}\right)a_1^2\cdots a_{m_1}^2B_1^2\cdots B_{m_2}^2C_1^2\cdots C_{m_3}^2,$$

where for $i=1,2$,

$$a_v^i=\frac{\partial^{l_{1,v}+m_{1,v}}\phi_i}{\partial\sigma_{j_{v,1}}^0\cdots\partial\sigma_{j_{v,l_{1,v}}}^0\partial\mu^{m_{1,v}}}(t),\qquad v=1,\cdots,m_1,$$

$$B_u^i=\frac{\partial^{\nu_{2,u}+\chi_{2,u}}q_i}{\partial\sigma^{\nu_{2,u}}\mu^{\chi_{2,u}}}(\phi_i(t),\mu)b_{u,1}^i\cdots b_{u,\nu_{2,u}}^i,\qquad u=1,\cdots,m_2,$$

$$b_{u,v}^i=\frac{\partial^{l_{2,u,v}+m_{2,u,v}}\phi_i}{\partial\sigma_{j_{2,u,v,1}}^0\cdots\partial\sigma_{j_{2,u,v,l_{2,u,v}}}^0\partial\mu^{m_{2,u,v}}}(t),\qquad v=1,\cdots,\nu_{2,u},$$

$$C_u^i=\frac{\partial^{\nu_{3,u}+\chi_{3,u}}q}{\partial\sigma^{\nu_{3,u}}\mu^{\chi_{3,u}}}(\phi_i(t),\mu)c_{u,1}^i\cdots c_{u,\nu_{3,u}}^i,\qquad u=1,\cdots,m_3,$$

$$c_{u,v}^i=\frac{\partial^{l_{3,u,v}+m_{3,u,v}}\phi_i}{\partial\sigma_{j_{3,u,v,1}}^0\cdots\partial\sigma_{j_{3,u,v,l_{3,u,v}}}^0\partial\mu^{m_{3,u,v}}}(t),\qquad v=1,\cdots,\nu_{3,u},$$

$$0\le m_i\le l+m,\quad i=1,\cdots,4,\quad 1\le m_1+m_2+m_3+m_4\le l+m,$$

$$\sum_{v=1}^{m_1}l_{1,v}+\sum_{u=1}^{m_2}\sum_{v=1}^{\nu_{2,u}}l_{2,u,v}+\sum_{u=1}^{m_3}\sum_{v=1}^{\nu_{3,u}}l_{3,u,v}=l,$$

$$m_4+\sum_{v=1}^{m_1}m_{1,v}+\sum_{u=1}^{m_2}\left(\sum_{v=1}^{\nu_{2,u}}m_{2,u,v}+\chi_{2,u}\right)+\sum_{u=1}^{m_3}\left(\sum_{v=1}^{\nu_{3,u}}m_{3,u,v}+\chi_{3,u}\right)=m.$$

In (5.10), (5.11), and (5.12)

$$1\le l_{1,v}+m_{1,v}\le l+m-1,\qquad v=1,\cdots,m_1,$$

$$1\le l_{2,u,v}+m_{2,u,v}\le l+m-1,\quad u=1,\cdots,m_2,\quad v=1,\cdots,\nu_{2,u},$$

$$1\le l_{3,u,v}+m_{3,u,v}\le l+m-1,\quad u=1,\cdots,m_3,\quad v=1,\cdots,\nu_{3,u}.$$

In (5.13)

$$1 \le l_{1,v} + m_{1,v} \le l + m, \ v = 1, \cdots, m_1,$$
$$1 \le l_{2,u,v} + m_{2,u,v} \le l + m, \ u = 1, \cdots, m_2, \ v = 1, \cdots, \nu_{2,u},$$
$$1 \le l_{3,u,v} + m_{3,u,v} \le l + m, \ u = 1, \cdots, m_3, \ v = 1, \cdots, \nu_{3,u}.$$

On the other hand, by the variation of constants formula,

$$\begin{aligned} &\frac{\partial^{l+m}\phi}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(t,\sigma^0,\mu,q_1,r_1,\rho) - \frac{\partial^{l+m}\phi}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(t,\sigma^0,\mu,q_2,r_2,\rho) \\ &= \int_0^t \Upsilon(t,s,\sigma^0,\mu,q_1,r_1,\rho)\widehat{f}_{(i_1,\cdots,i_l,m)}(s)\,ds. \end{aligned} \tag{5.14}$$

Now assume that (5.8) holds for all $\widehat{l}$ and $\widehat{m}$ such that $\widehat{l} + \widehat{m} \le l + m - 1$. By Lemmas 5.2 and 5.3, and Property 5.6 there is $\widehat{K}_{l,m} > 0$ such that for $t \le 0$,

$$\begin{aligned} \left\|\widehat{f}_{(i_1,\cdots,i_l,m)}(t)\right\| \ &\le \ \widehat{K}_{l,m}\left(\|q_1 - q_2\|_{l+m} + \|r_1 - r_2\|_{l+m}\right)\lambda(\rho)\mu^{k-m} \\ &\quad \times \left\{\sum_{i=0}^{l+2m}\left[\lambda(\rho)\mu^k(-t)\right]^i\right\} e^{-(l+m+1)\lambda(\rho)\mu^k(1+2\Delta)t}. \end{aligned} \tag{5.15}$$

Now (5.8) follows from (5.14) and (5.15). □

Next we will state and prove some properties of $\Lambda(t, t_0, \sigma^0, \mu, q, r, \rho)$, which is defined as the fundamental matrix solution of (3.2), where $A(\theta)$ satisfies Assumption 2.1(a). The proof of Lemma 5.7 is left to the reader.

LEMMA 5.7. *For each $\rho \in I_1$, $\Lambda(t, t_0, \sigma^0, \mu, q, r, \rho)$ exists for all $t, t_0 \in \mathbb{R}$, $\sigma^0 \in \mathbb{R}^N$, and $\mu \in I$, and it is k-times continuously differentiable with respect to t, t_0, σ^0, and μ, and T_i-periodic in σ^0_i for $i = 1, \cdots, N$.*

Let

$$K_1 = \max_{1\le i\le N}\{H_i\}, \quad \beta_1 = \min_{1\le i\le N}\{\alpha_i\}, \quad H = \max_{1\le i\le N}\left\{\frac{\alpha_i + H_i\|A_i\|}{\omega_i}\right\}.$$

By Assumption 2.1(d), there is $\rho_1 > 0$ such that for $|\rho| \le \rho_1$,

$$\lambda(\rho)\mu^k \le \frac{\beta_1}{H} \quad \text{for all } \mu \in I, \tag{5.16}$$

and

$$\lambda(\rho) < \omega_i, \qquad i = 1, \cdots, N. \tag{5.17}$$

In the remainder of this section we assume that $|\rho| \le \rho_1$.

LEMMA 5.8. *For $t_0 \le t$,*

$$\|\Lambda(t,t_0,\sigma^0,\mu,q,r,\rho)\| \le K_1 e^{-[\beta_1 - H\lambda(\rho)\mu^k](t-t_0)}. \tag{5.18}$$

Proof. Let $\Lambda_i(t,t_0,\sigma^0,\mu,q,r,\rho)$ be the fundamental matrix solution of

$$\begin{aligned} &\frac{dw_i}{dt} = A_i(\omega_i^{-1}\phi_i(t,\sigma^0,\mu,q,r,\rho))w_i, \\ &\Lambda_i(t_0,t_0,\sigma^0,\mu,q,r,\rho) = I. \end{aligned} \tag{5.19}$$

By (5.17)

$$\frac{d\phi_i}{dt}(t,\sigma^0,\mu,q,r,\rho) > 0.$$

Therefore we can take $\theta_i = \omega_i^{-1}\phi_i(t,\sigma^0,\mu,r,s,\rho)$ as an independent variable in (5.19) and obtain

$$\begin{aligned} \frac{dw_i}{d\theta_i} &= \frac{dt}{d\theta_i}\frac{dw_i}{dt} \\ &= \frac{\omega_i}{\omega_i + \mathcal{S}_i(\phi(t),q(\phi(t),\mu),r(\phi(t),\mu),\mu,\rho)} A_i(\theta_i)w_i \\ &= A_i(\theta_i)w_i - \frac{\mathcal{S}_i(\phi(t),q(\phi(t),\mu),r(\phi(t),\mu),\mu,\rho)}{\omega_i + \mathcal{S}_i(\phi(t),q(\phi(t),\mu),r(\phi(t),\mu),\mu,\rho)} A_i(\theta_i)w_i. \end{aligned}$$

By the variation of constants formula, Λ_i, as a function of θ_i, is given by

$$\begin{aligned} \Lambda_i(\theta_i) &= \Psi_i(\theta_i)\Psi_i^{-1}(\theta_i^0)\Lambda_i(\theta_i^0) \\ &\quad - \int_{\theta_i^0}^{\theta_i} \Psi_i(\theta_i)\Psi_i^{-1}(s)\frac{\mathcal{S}_i(\phi(t),q(\phi(t),\mu),r(\phi(t),\mu),\mu,\rho)}{\omega_i + \mathcal{S}_i(\phi(t),q(\phi(t),\mu),r(\phi(t),\mu),\mu,\rho)} A_i(s)\Lambda_i(s)\,ds. \end{aligned}$$

Suppose $\theta_i^0 = \omega_i^{-1}\phi_i(t_0,\sigma^0,q,r,\mu,\rho)$. Then $\Lambda_i(\theta_i^0) = I$. Moreover for $t_0 \le t$, $\theta_i^0 \le \theta_i$. Therefore

$$\begin{aligned} \|\Lambda_i(\theta_i)\| &\le H_i e^{-\alpha_i(\theta_i - \theta_i^0)} \\ &\quad + \frac{H_i\,\|A_i\|\,\lambda(\rho)\mu^k}{\omega_i - \lambda(\rho)\mu^k} \int_{\theta_i^0}^{\theta_i} e^{-\alpha_i(\theta_i - s)}\,\|\Lambda_i(s)\|\,ds. \end{aligned}$$

By Gronwall's inequality we obtain

$$\begin{aligned} \|\Lambda_i(\theta_i)\| &\le H_i e^{-\left[\alpha_i - H_i\|A_i\|\lambda(\rho)\mu^k/(\omega_i - \lambda(\rho)\mu^k)\right](\theta_i - \theta_i^0)} \\ &= H_i e^{-\left[\alpha_i\omega_i - (\alpha_i + H_i\|A_i\|)\lambda(\rho)\mu^k\right](\theta_i - \theta_i^0)/(\omega_i - \lambda(\rho)\mu^k)}. \end{aligned}$$

Since

$$\begin{aligned}
&\phi_i(t,\sigma^0,\mu,q,r,\rho) - \phi_i(t_0,\sigma^0,\mu,q,r,\rho)\\
&\quad = \omega_i(t-t_0) + \int_{t_0}^{t} \mathcal{S}_i(\phi(s),q(\phi(s),\mu),r(\phi(s),\mu),\mu,\rho)\,ds\\
&\quad \geq (\omega_i - \lambda(\rho)\mu^k)(t-t_0),
\end{aligned}$$

it follows from (5.16) that for $t_0 \leq t$,

$$\begin{aligned}
\left\|\Lambda_i(t,t_0,\sigma^0,\mu,q,r,\rho)\right\| &\leq H_i e^{-\left[\alpha_i - (\alpha_i + H_i\|A_i\|)/\omega_i\lambda(\rho)\mu^k\right](t-t_0)}\\
&\leq K_1 e^{-[\beta_1 - H\lambda(\rho)\mu^k](t-t_0)}.
\end{aligned}$$

Now (5.18) follows from the fact that

$$\Lambda(t,t_0,\sigma^0,\mu,q,r,\rho) = \bigoplus_{i=1}^{N} \Lambda_i(t,t_0,\sigma^0,\mu,q,r,\rho). \qquad \square$$

LEMMA 5.9. *For $1 \leq l+m \leq k$ there is a positive constant $C_{l,m}$ such that for $t_0 \leq t \leq 0$,*

$$\begin{aligned}
&\left\|\frac{\partial^{l+m}\Lambda}{\partial\sigma_{i_1}^0 \cdots \partial\sigma_{i_l}^0 \partial\mu^m}(t,t_0,\sigma^0,\mu,q,r,\rho)\right\|\\
&\quad \leq C_{l,m}\left[\sum_{i=1}^{l+2m}(-t_0)^i\right] e^{-[\beta_1 - H\lambda(\rho)\mu^k](t-t_0) - (l+m)\lambda(\rho)\mu^k(1+2\Delta)t_0}.
\end{aligned} \tag{5.20}$$

Proof. It is easily shown inductively that for $1 \leq l+m \leq k$

$$x = \frac{\partial^{l+m}\Lambda}{\partial\sigma_{i_1}^0 \cdots \partial\sigma_{i_l}^0 \partial\mu^m}(t,t_0,\sigma^0,\mu,q,r,\rho)$$

satisfies the following nonhomogeneous linear system:

$$\begin{aligned}
&\frac{dx}{dt} = A(\Omega^{-1}\phi(t,\sigma^0,\mu,q,r,\rho))x + g_{(i_1,\cdots,i_l,m)}(t),\\
&\frac{\partial^{l+m}\Lambda}{\partial\sigma_{i_1}^0 \cdots \partial\sigma_{i_l}^0 \partial\mu^m}(t_0,t_0,\sigma^0,\mu,q,r,\rho) = 0.
\end{aligned} \tag{5.21}$$

System (5.21) has the following properties.

PROPERTY 5.10. $g_{(i_1,\cdots,i_l,m)}(t)$ is a (finite) sum of terms of the form

$$\frac{\partial^n A}{\partial\theta^n}(\Omega^{-1}\phi(t))a_1 \cdots a_n b_{n+1},$$

where

$$a_v = \Omega^{-1} \frac{\partial^{l_v+m_v}\phi}{\partial\sigma^0_{j_{v,1}} \cdots \partial\sigma^0_{j_{v,l_{1,v}}} \partial\mu^{m_v}}(t), \qquad v = 1, \cdots, n,$$

$$b_{n+1} = \frac{\partial^{l_{n+1}+m_{n+1}}\Lambda}{\partial\sigma^0_{j_{n+1,1}} \cdots \partial\sigma^0_{j_{n+1,l_{n+1}}} \partial\mu^{m_{n+1}}}(t, t_0, \sigma^0, \mu, q, r, \rho),$$

$$1 \le n \le l + m,$$

$$1 \le l_v + m_v \le l + m, \qquad v = 1, \cdots, n,$$

$$0 \le l_{n+1} + m_{n+1} \le l + m - 1,$$

$$\sum_{v=1}^{n+1} l_v = l, \qquad \sum_{v=1}^{n+1} m_v = m.$$

By the variation of constants formula,

$$\frac{\partial^{l+m}\Lambda}{\partial\sigma^0_{i_1} \cdots \partial\sigma^0_{i_l} \partial\mu^m}(t, t_0, \sigma^0, \mu, q, r, \rho) = \int_0^t \Lambda(t, s, \sigma^0, \mu, q, r, \rho) g_{(i_1,\cdots,i_l,m)}(s)\, ds. \tag{5.22}$$

Now we prove (5.20) inductively using Lemmas 5.2, 5.3, and 5.8, Property 5.10, and (5.16) and (5.22). □

LEMMA 5.11. *For $0 \le l + m \le k - 1$ there is a positive constant $\widehat{C}_{l,m}$ such that for $t_0 \le t \le 0$,*

$$\left\| \frac{\partial^{l+m}\Lambda}{\partial\sigma^0_{i_1} \cdots \partial\sigma^0_{i_l} \partial\mu^m}(t, t_0, \sigma^0, \mu, q_1, r_1, \rho) - \frac{\partial^{l+m}\Lambda}{\partial\sigma^0_{i_1} \cdots \partial\sigma^0_{i_l} \partial\mu^m}(t, t_0, \sigma^0, \mu, q_2, r_2, \rho) \right\|$$

$$\le \widehat{C}_{l,m} \left(\|q_1 - q_2\|_{l+m} + \|r_1 - r_2\|_{l+m} \right) \tag{5.23}$$

$$\times \left[\sum_{i=2}^{l+2m+2} (-t_0)^i \right] e^{-[\beta_1 - H\lambda(\rho)\mu^k](t-t_0) - (l+m+1)\lambda(\rho)\mu^k(1+2\Delta)t_0}.$$

Proof. Let

$$\Lambda_1(t, t_0) = \Lambda(t, t_0, \sigma^0, \mu, q_1, r_1, \rho), \qquad \Lambda_2(t, t_0) = \Lambda(t, t_0, \sigma^0, \mu, q_2, r_2, \rho).$$

We easily show that, for $0 \le l + m \le k - 1$,

$$x = \frac{\partial^{l+m}\Lambda_1}{\partial\sigma^0_{i_1} \cdots \partial\sigma^0_{i_l} \partial\mu^m}(t, t_0) - \frac{\partial^{l+m}\Lambda_2}{\partial\sigma^0_{i_1} \cdots \partial\sigma^0_{i_l} \partial\mu^m}(t, t_0)$$

satisfies the following nonhomogeneous linear system:

$$\begin{aligned}&\frac{dx}{dt} = A(\Omega^{-1}\phi_1(t))x + \widehat{g}_{(i_1,\cdots,i_l,m)}(t),\\ &\frac{\partial^{l+m}\Lambda_1}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(t_0,t_0) - \frac{\partial^{l+m}\Lambda_2}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(t_0,t_0) = 0.\end{aligned} \tag{5.24}$$

System (5.24) has the following properties.

PROPERTY 5.12. $\widehat{g}_{(i_1,\cdots,i_l,m)}(t)$ is a (finite) sum of terms of the forms

$$\frac{\partial^n A}{\partial\sigma^n}(\Omega^{-1}\phi_1(t))a^1_1\cdots a^1_n\left(b^1_{n+1} - b^2_{n+1}\right), \tag{5.25}$$

$$\frac{\partial^n A}{\partial\sigma^n}(\Omega^{-1}\phi_1(t))a^2_1\cdots a^2_{v-1}\left(a^1_v - a^2_v\right)a^1_{v+1}\cdots a^1_n b^2_{n+1}, \tag{5.26}$$

$$\left[\frac{\partial^n A}{\partial\sigma^n}(\Omega^{-1}\phi_1(t)) - \frac{\partial^n A}{\partial\sigma^n}(\Omega^{-1}\phi_2(t))\right]a^2_1\cdots a^2_n b^2_{n+1}, \tag{5.27}$$

where for $i = 1, 2$,

$$\begin{aligned}&a^i_v = \Omega^{-1}\frac{\partial^{l_v+m_v}\phi_i}{\partial\sigma^0_{j_{v,1}}\cdots\partial\sigma^0_{j_{v,l_v}}\partial\mu^{m_v}}(t), \qquad v = 1,\cdots,n,\\ &b^i_{n+1} = \frac{\partial^{l_{n+1}+m_{n+1}}\Lambda_i}{\partial\sigma^0_{j_{n+1,1}}\cdots\partial\sigma^0_{j_{n+1,l_{n+1}}}\partial\mu^{m_{n+1}}}(t,t_0),\\ &\sum_{v=1}^{n+1} l_v = l, \qquad \sum_{v=1}^{n+1} m_v = m.\end{aligned}$$

In (5.25) and (5.26)

$$\begin{aligned}&1 \le n \le l+m,\\ &1 \le l_v + m_v \le l+m, \qquad v = 1,\cdots,n,\\ &0 \le l_{n+1} + m_{n+1} \le l+m-1.\end{aligned}$$

In (5.27)

$$\begin{aligned}&0 \le n \le l+m,\\ &1 \le l_v + m_v \le l+m, \qquad v = 1,\cdots,n,\\ &n = 0 \quad \text{if and only if } l_{n+1} = l \text{ and } m_{n+1} = m.\end{aligned}$$

By the variation of constants formula,

$$\begin{aligned}&\frac{\partial^{l+m}\Lambda_1}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(t,t_0) - \frac{\partial^{l+m}\Lambda_2}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}\partial\mu^m}(t,t_0)\\ &\quad = \int_0^t \Lambda(t,s,\sigma^0,\mu,q_1,r_1,\rho)\widehat{g}_{(i_1,\cdots,i_l,m)}(s)\,ds.\end{aligned} \tag{5.28}$$

Now we prove (5.23) inductively using Lemmas 5.2, 5.3, 5.5, 5.8, and 5.9, Property 5.12, and (5.16) and (5.28). □

Recall that $\phi(t,\sigma^0,q,r,\mu)$ is a solution of (3.18) where $\mathcal{S}$ satisfies Assumption 2.1(f) and $(q,r)\in B^k_\Delta(\mathbb{R}^N,\mathbb{R}^{N_1})\times B^k_\Delta(\mathbb{R}^N,\mathbb{R}^{N_2})$, and $\Lambda(t,t_0,\sigma^0,q,r,\mu)$ is the fundamental matrix solution of (3.19) where $A(\theta)$ satisfies Assumption 2.1(a). Suppose that $(q,r),(q_1,r_1),(q_2,r_2)\in B^k_\Delta(\mathbb{R}^N,\mathbb{R}^{N_1})\times B^k_\Delta(\mathbb{R}^N,\mathbb{R}^{N_2})$. We only state the following properties of $\phi(t,\sigma^0,q,r,\mu)$ and $\Lambda(t,t_0,\sigma^0,q,r,\mu)$ because the proofs are similar to those of Lemmas 5.1–5.11 and even simpler.

PROPERTY 5.13. (a) For each $\mu\in I$, $\phi(t,\sigma^0,q,r,\mu)$ exists for all $t\in\mathbb{R}$, $\sigma^0\in\mathbb{R}^N$, and it is k-times continuously differentiable with respect to t and σ^0. Moreover, for $i=1,\cdots,N$

$$\phi(t,\sigma^0+2\pi e^i,q,r,\mu)=\phi(t,\sigma^0,q,r,\mu)+2\pi e^i.$$

(b) If

$$\Upsilon(t,s,\sigma^0,q,r,\mu)=\frac{\partial\phi}{\partial\sigma^0}(t,\sigma^0,q,r,\mu)\frac{\partial\phi^{-1}}{\partial\sigma^0}(s,\sigma^0,q,r,\mu),$$

then for all $t,s\in\mathbb{R}$,

$$\|\Upsilon(t,s,\sigma^0,q,r,\mu)\|\le e^{\lambda(\mu)(1+2\Delta)|t-s|}.$$

In particular, for $t\le 0$,

$$\left\|\frac{\partial\phi}{\partial\sigma^0}(t,\sigma^0,q,r,\mu)\right\|\le e^{-\lambda(\mu)(1+2\Delta)t}.$$

(c) For $2\le l\le k$ there is a positive constant K_l such that for $t\le 0$,

$$\left\|\frac{\partial^l\phi}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}}(t,\sigma^0,q,r,\mu)\right\|\le K_l\left\{\sum_{i=1}^{l-1}[\lambda(\mu)(-t)]^i\right\}e^{-l\lambda(\mu)(1+2\Delta)t}.$$

(d) For $0\le l\le k-1$ there is a positive constant $\widehat{K}_l$ such that for $t\le 0$,

$$\left\|\frac{\partial^l\phi}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}}(t,\sigma^0,q_1,r_1,\mu)-\frac{\partial^l\phi}{\partial\sigma^0_{i_1}\cdots\partial\sigma^0_{i_l}}(t,\sigma^0,q_2,r_2,\mu)\right\|$$

$$\le\widehat{K}_l\left(\|q_1-q_2\|_l+\|r_1-r_2\|_l\right)\left\{\sum_{i=1}^{l+1}[\lambda(\mu)(-t)]^i\right\}e^{-(l+1)\lambda(\mu)(1+2\Delta)t}.$$

(e) For each $\mu\in I$, $\Lambda(t,t_0,\sigma^0,q,r,\mu)$ exists for all t and $t_0\in\mathbb{R}$, and $\sigma^0\in\mathbb{R}^N$. It is k-times continuously differentiable with respect to t, t_0, and σ^0, and T_i-periodic in σ^0_i for $i=1,\cdots,N$.

(f) If

$$\lambda(\mu)\le\frac{\beta_1}{H},$$

then for $t_0\le t$

$$\left\|\Lambda(t,t_0,\sigma^0,q,r,\mu)\right\| \le He^{-[\beta_1 - H\lambda(\mu)](t-t_0)}.$$

(g) For $1 \le l \le k$ there is a positive constant C_l such that for $t_0 \le t \le 0$,

$$\left\|\frac{\partial^l \Lambda}{\partial\sigma_{i_1}^0 \cdots \partial\sigma_{i_l}^0}(t,t_0,\sigma^0,q,r,\mu)\right\| \le C_l \left[\sum_{i=1}^{l}(-t_0)^i\right] e^{-[\beta_1 - H\lambda(\mu)](t-t_0) - l\lambda(\mu)(1+2\Delta)t_0}$$

(h) For $0 \le l \le k-1$ there is a positive constant $\widehat{C}_l$ such that for $t_0 \le t \le 0$,

$$\left\|\frac{\partial^l \Lambda}{\partial\sigma_{i_1}^0 \cdots \partial\sigma_{i_l}^0}(t,t_0,\sigma^0,q_1,r_1,\mu) - \frac{\partial^l \Lambda}{\partial\sigma_{i_1}^0 \cdots \partial\sigma_{i_l}^0}(t,t_0,\sigma^0,q_2,r_2,\mu)\right\| \le \widehat{C}_l\left(\|q_1-q_2\|_l + \|r_1-r_2\|_l\right)\left[\sum_{i=2}^{l+2}(-t_0)^i\right] e^{-[\beta_1 - H\lambda(\mu)](t-t_0) - (l+1)\lambda(\mu)(1+2\Delta)t_0}.$$

5.2. The proofs of Lemmas 3.1 and 3.2. In this section we prove the lemmas in §3 that lead to Theorem 3.4. Recall that $G_1(s,\sigma,\mu,q,r,\rho)$ and $G_2(s,\sigma,\mu,q,r,\rho)$ are defined by (3.4) where $\phi(t,\sigma^0,\mu,q,r,\rho)$ and $\Lambda(t,t_0,\sigma^0,\mu,q,r,\rho)$ are the solutions of (3.1) and (3.2), respectively.

Proof of Lemma 3.1.

$$\frac{\partial^{l+m} G_1}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^m}(s,\sigma,\mu,q,r,\rho)$$

is a (finite) sum of terms of the form

$$\frac{\partial^{l_1+j_1}\Lambda}{\partial\sigma_{i_{1,1}}\cdots\partial\sigma_{i_{1,l_1}}\partial\mu^{j_1}}\,\frac{\partial^{l_2+j_2}\mathcal{W}}{\partial\sigma_{i_{2,1}}\cdots\partial\sigma_{i_{2,l_2}}\partial\mu^{j_2}},$$

$$l_1 + l_2 = l, \qquad j_1 + j_2 = m, \tag{5.29}$$

$$0 \le l_1, \quad l_2 \le l, \quad 0 \le j_1, \quad j_2 \le m,$$

where Λ and $\mathcal{W}$ are as in (3.4). The second factor of (5.29) is in turn a (finite) sum of terms of the form

$$\frac{\partial^{m_1+m_2+m_3+m_4}\mathcal{W}}{\partial\sigma^{m_1}\partial w^{m_2}\partial z^{m_3}\partial\mu^{m_4}} a_1\cdots a_{m_1} B_1\cdots B_{m_2} C_1\cdots C_{m_3}, \tag{5.30}$$

where

$$a_v = \frac{\partial^{l_{1,v}+m_{1,v}}\phi}{\partial\sigma_{j_{v,1}}\cdots\partial\sigma_{j_{v,l_{1,v}}}\partial\mu^{m_{1,v}}}(s), \qquad v = 1,\cdots,m_1,$$

$$B_u = \frac{\partial^{\nu_{2,u}+\chi_{2,u}}q}{\partial\sigma^{\nu_{2,u}}\mu^{\chi_{2,u}}}(\phi(s),\mu)b_{u,1}\cdots b_{u,\nu_{2,u}}, \qquad u = 1,\cdots,m_2,$$

$$b_{u,v} = \frac{\partial^{l_{2,u,v}+m_{2,u,v}}\phi}{\partial\sigma_{j_{2,u,v,1}}\cdots\partial\sigma_{j_{2,u,v,l_{2,u,v}}}\partial\mu^{m_{2,u,v}}}(s), \qquad v = 1,\cdots,\nu_{2,u},$$

$$C_u = \frac{\partial^{\nu_{3,u}+\chi_{3,u}}q}{\partial\sigma^{\nu_{3,u}}\mu^{\chi_{3,u}}}(\phi(s),\mu)c_{u,1}\cdots c_{u,\nu_{3,u}}, \qquad u = 1,\cdots,m_3,$$

$$c_{u,v} = \frac{\partial^{l_{3,u,v}+m_{3,u,v}}\phi}{\partial\sigma_{j_{3,u,v,1}}\cdots\partial\sigma_{j_{3,u,v,l_{3,u,v}}}\partial\mu^{m_{3,u,v}}}(s), \qquad v = 1,\cdots,\nu_{3,u},$$

$$0 \le m_i \le l_2 + j_2, \quad i = 1,\cdots,4, \quad 1 \le m_1+m_2+m_3+m_4 \le l_2 + j_2,$$

$$\sum_{v=1}^{m_1} l_{1,v} + \sum_{u=1}^{m_2}\sum_{v=1}^{\nu_{2,u}} l_{2,u,v} + \sum_{u=1}^{m_3}\sum_{v=1}^{\nu_{3,u}} l_{3,u,v} = l_2,$$

$$m_4 + \sum_{v=1}^{m_1} m_{1,v} + \sum_{u=1}^{m_2}\left(\sum_{v=1}^{\nu_{2,u}} m_{2,u,v} + \chi_{2,u}\right) + \sum_{u=1}^{m_3}\left(\sum_{v=1}^{\nu_{3,u}} m_{3,u,v} + \chi_{3,u}\right) = j_2.$$

Now (3.6) follows from (5.29), (5.30), and Property 3.1(a)–(e).

Note that

$$\frac{\partial^{l+m}G_2}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^m}(s,\sigma,\mu,q,r,\rho)$$

is a (finite) sum of terms of the form

$$\frac{\partial^{m_1}}{\partial\mu^{m_1}}\left(\rho\mu e^{-\rho\mu Qs}\right)\frac{\partial^{l+m_2}\mathcal{Z}}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^{m_2}},$$

$$m_1 + m_2 = m, \quad 0 \le m_1, \quad m_2 \le m,$$

where $\mathcal{Z}$ is as given in (3.4). Now we show inductively that, for $0 \le m$,

$$\frac{\partial^m}{\partial\mu^m}\left(\rho\mu e^{-\rho\mu Qs}\right) = (-1)^{m-1}\left[m\rho^m(Qs)^{m-1} - \rho^{m+1}\mu(Qs)^m\right]e^{-\rho\mu Qs},$$

and it follows that there is $\widehat{H}_m > 0$ such that for $s \le 0$,

$$\text{(5.31)} \quad \left\|\frac{\partial^m}{\partial\mu^m}\left(\rho\mu e^{-\rho\mu Qs}\right)\right\| \le \begin{cases} K_2\rho\mu e^{\rho\mu\beta_2 s}, & m = 0, \\ \widehat{H}_m\left[\rho^m(-s)^{m-1} + \rho^{m+1}\mu(-s)^m\right]e^{\rho\mu\beta_2 s}, & m > 0. \end{cases}$$

On the other hand, using Property 3.1(a) and (b) and an analysis similar to the one in the first part of the proof, we obtain the estimate

(5.32)

$$\left\| \frac{\partial^{l+m_2} \mathcal{Z}}{\partial \sigma_{i_1} \cdots \partial \sigma_{i_l} \partial \mu^{m_2}} \right\| \leq \widehat{K} \lambda_2(\rho) \mu^{k-m_2} \left\{ \sum_{i=0}^{l+2m_2} \left[\lambda(\rho) \mu^k (-s) \right]^i \right\} e^{-(l+m_2)\lambda(\rho)\mu^k (1+2\Delta)s}.$$

Now (3.7) follows from (3.5), (5.31), and (5.32). □

Proof of Lemma 3.3.

$$\frac{\partial^{l+m} G_1}{\partial \sigma_{i_1} \cdots \partial \sigma_{i_l} \partial \mu^m}(s, \sigma, \mu, q_1, r_1, \rho) - \frac{\partial^{l+m} G_1}{\partial \sigma_{i_1} \cdots \partial \sigma_{i_l} \partial \mu^m}(s, \sigma, \mu, q_2, r_2, \rho)$$

is a (finite) sum of terms of the form

$$\frac{\partial^{l_1+j_1} \Lambda_1}{\partial \sigma_{i_{1,1}} \cdots \partial \sigma_{i_{1,l_1}} \partial \mu^{j_1}} \frac{\partial^{l_2+j_2} \mathcal{W}_1}{\partial \sigma_{i_{2,1}} \cdots \partial \sigma_{i_{2,l_2}} \partial \mu^{j_2}}$$

$$- \frac{\partial^{l_1+j_1} \Lambda_2}{\partial \sigma_{i_{1,1}} \cdots \partial \sigma_{i_{1,l_1}} \partial \mu^{j_1}} \frac{\partial^{l_2+j_2} \mathcal{W}_2}{\partial \sigma_{i_{2,1}} \cdots \partial \sigma_{i_{2,l_2}} \partial \mu^{j_2}}$$

$$= \frac{\partial^{l_1+j_1} \Lambda_1}{\partial \sigma_{i_{1,1}} \cdots \partial \sigma_{i_{1,l_1}} \partial \mu^{j_1}} \left(\frac{\partial^{l_2+j_2} \mathcal{W}_1}{\partial \sigma_{i_{2,1}} \cdots \partial \sigma_{i_{2,l_2}} \partial \mu^{j_2}} - \frac{\partial^{l_2+j_2} \mathcal{W}_2}{\partial \sigma_{i_{2,1}} \cdots \partial \sigma_{i_{2,l_2}} \partial \mu^{j_2}} \right)$$

$$+ \left(\frac{\partial^{l_1+j_1} \Lambda_1}{\partial \sigma_{i_{1,1}} \cdots \partial \sigma_{i_{1,l_1}} \partial \mu^{j_1}} - \frac{\partial^{l_1+j_1} \Lambda_2}{\partial \sigma_{i_{1,1}} \cdots \partial \sigma_{i_{1,l_1}} \partial \mu^{j_1}} \right) \frac{\partial^{l_2+j_2} \mathcal{W}_2}{\partial \sigma_{i_{2,1}} \cdots \partial \sigma_{i_{2,l_2}} \partial \mu^{j_2}},$$

$$l_1 + l_2 = l, \qquad j_1 + j_2 = m,$$

$$0 \leq l_1, \quad l_2 \leq l, \quad 0 \leq j_1, j_2 \leq m,$$

where for $i = 1, 2$,

$$\Lambda_i = \Lambda(0, s, \sigma, \mu, q_i, r_i, \rho), \qquad \mathcal{W}_i = \mathcal{W}(\tau_i(s), q_i(\tau_i(s), \mu), r_i(\tau_i(s), \mu), \mu, \rho).$$

On the other hand,

$$\frac{\partial^{l_2+j_2} \mathcal{W}_1}{\partial \sigma_{i_{2,1}} \cdots \partial \sigma_{i_{2,l_2}} \partial \mu^{j_2}} - \frac{\partial^{l_2+j_2} \mathcal{W}_2}{\partial \sigma_{i_{2,1}} \cdots \partial \sigma_{i_{2,l_2}} \partial \mu^{j_2}}$$

is in turn a (finite) sum of terms of the forms

$$\frac{\partial^{m_1+m_2+m_3+m_4} \mathcal{W}_1}{\partial \sigma^{m_1} \partial w^{m_2} \partial z^{m_3} \partial \mu^{m_4}} a_1^1 \cdots a_{m_1}^1 B_1^1 \cdots B_{m_2}^1 C_1^2 \cdots C_{u-1}^2 \left(C_u^1 - C_u^2 \right) C_{u+1}^1 \cdots C_{m_3}^1,$$

$$\frac{\partial^{m_1+m_2+m_3+m_4} \mathcal{W}_1}{\partial \sigma^{m_1} \partial w^{m_2} \partial z^{m_3} \partial \mu^{m_4}} a_1^1 \cdots a_{m_1}^1 B_1^2 \cdots B_{u-1}^2 \left(B_u^1 - B_u^2 \right) B_{u+1}^1 \cdots B_{m_2}^1 C_1^2 \cdots C_{m_3}^2,$$

$$\frac{\partial^{m_1+m_2+m_3+m_4} \mathcal{W}_1}{\partial \sigma^{m_1} \partial w^{m_2} \partial z^{m_3} \partial \mu^{m_4}} a_1^2 \cdots a_{v-1}^2 \left(a_v^1 - a_v^2 \right) a_{v+1}^1 \cdots a_{m_1}^1 B_1^2 \cdots B_{m_2}^2 C_1^2 \cdots C_{m_3}^2,$$

$$\left(\frac{\partial^{m_1+m_2+m_3+m_4} \mathcal{W}_1}{\partial \sigma^{m_1} \partial w^{m_2} \partial z^{m_3} \partial \mu^{m_4}} - \frac{\partial^{m_1+m_2+m_3+m_4} \mathcal{W}_2}{\partial \sigma^{m_1} \partial w^{m_2} \partial z^{m_3} \partial \mu^{m_4}} \right) a_1^2 \cdots a_{m_1}^2 B_1^2 \cdots B_{m_2}^2 C_1^2 \cdots C_{m_3}^2,$$

where

$$a_v^i = \frac{\partial^{l_{1,v}+m_{1,v}}\tau_i}{\partial\sigma_{j_{v,1}}\cdots\partial\sigma_{j_{v,l_{1,v}}}\partial\mu^{m_{1,v}}}(s), \qquad v=1,\cdots,m_1,$$

$$B_u^i = \frac{\partial^{\nu_{2,u}+\chi_{2,u}}q_i}{\partial\sigma^{\nu_{2,u}}\mu^{\chi_{2,u}}}(\tau_i(s),\mu)b_{u,1}^i\cdots b_{u,\nu_{2,u}}^i, \qquad u=1,\cdots,m_2,$$

$$b_{u,v}^i = \frac{\partial^{l_{2,u,v}+m_{2,u,v}}\tau_i}{\partial\sigma_{j_{2,u,v,1}}\cdots\partial\sigma_{j_{2,u,v,l_{2,u,v}}}\partial\mu^{m_{2,u,v}}}(s), \qquad v=1,\cdots,\nu_{2,u},$$

$$C_u^i = \frac{\partial^{\nu_{3,u}+\chi_{3,u}}q}{\partial\sigma^{\nu_{3,u}}\mu^{\chi_{3,u}}}(\tau_i(s),\mu)c_{u,1}^i\cdots c_{u,\nu_{3,u}}^i, \qquad u=1,\cdots,m_3,$$

$$c_{u,v}^i = \frac{\partial^{l_{3,u,v}+m_{3,u,v}}\tau_i}{\partial\sigma_{j_{3,u,v,1}}\cdots\partial\sigma_{j_{3,u,v,l_{3,u,v}}}\partial\mu^{m_{3,u,v}}}(s), \qquad v=1,\cdots,\nu_{3,u},$$

$$0\le m_i\le l_2+j_2, \quad i=1,\cdots,4, \quad 1\le m_1+m_2+m_3+m_4\le l_2+j_2,$$

$$\sum_{v=1}^{m_1} l_{1,v}+\sum_{u=1}^{m_2}\sum_{v=1}^{\nu_{2,u}} l_{2,u,v}+\sum_{u=1}^{m_3}\sum_{v=1}^{\nu_{3,u}} l_{3,u,v}=l_2,$$

$$m_4+\sum_{v=1}^{m_1} m_{1,v}+\sum_{u=1}^{m_2}\left(\sum_{v=1}^{\nu_{2,u}} m_{2,u,v}+\chi_{2,u}\right)+\sum_{u=1}^{m_3}\left(\sum_{v=1}^{\nu_{3,u}} m_{3,u,v}+\chi_{3,u}\right)=j_2.$$

Now (3.8) follows from this fact and Property 3.1.

Observe that

$$\frac{\partial^{l+m}G_2}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^m}(s,\sigma,\mu,q_1,r_1,\rho)-\frac{\partial^{l+m}G_2}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^m}(s,\sigma,\mu,q_2,r_2,\rho)$$

is a (finite) sum of terms of the form

$$\frac{\partial^{m_1}}{\partial\mu^{m_1}}\left(\rho\mu e^{-\rho\mu Qs}\right)\left(\frac{\partial^{l+m_2}\mathcal{Z}_1}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^{m_2}}-\frac{\partial^{l+m_2}\mathcal{Z}_2}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^{m_2}}\right),$$

$$m_1+m_2=m, \quad 0\le m_1,m_2\le m.$$

An analysis similar to the one in the first part of the proof leads to the estimate

$$\begin{aligned}\left\|\frac{\partial^{l+m_2}\mathcal{Z}_1}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^{m_2}}-\frac{\partial^{l+m_2}\mathcal{Z}_2}{\partial\sigma_{i_1}\cdots\partial\sigma_{i_l}\partial\mu^{m_2}}\right\| &\le \widehat{K}\lambda_2(\rho)\mu^{k-m_2}\left(\|q_1-q_2\|_{l+m_2}+\|r_1-r_2\|_{l+m_2}\right)\\ &\times\left\{\sum_{i=0}^{l+2m_2+2}\left[\lambda(\rho)\mu^k(-s)\right]^i\right\}e^{-(l+m_2+1)\lambda(\rho)\mu^k(1+2\Delta)s}.\end{aligned} \tag{5.33}$$

Now (3.9) follows from (3.5), (5.31), and (5.33). □

REFERENCES

[1] W. A. COPPEL, *Stability and Asymptotic Behavior of Differential Equations*, D. C. Heath, Boston, 1965.

[2] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, London, 1969.

[3] S. P. DILIBERTO, *New results on periodic surfaces and the averaging principle*, in Proc. U.S.-Japan Seminar on Differential and Functional Equations, W. A. Harris and Y. Sibuya, eds., Benjamin, New York, 1967, pp. 49–87.

[4] N. FENICHEL, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971) pp. 193–226.

[5] J. K. HALE, *Integral manifolds of perturbed differential systems*, Ann. Math., 73 (1961) pp. 496–531.

[6] _____, *Oscillations in Nonlinear Systems*, McGraw–Hill, New York, 1963.

[7] _____, *Ordinary Differential Equations*, John Wiley, New York, 1969.

[8] P. HARTMAN, *Ordinary Differential Equations*, Birkhäuser, Boston, 1982.

[9] M. W. HIRSCH, C.C. PUGH, AND M. SHUB, *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag, New York, 1977.

[10] M. W. HIRSCH AND S. SMALE, *Differential Equations, Dynamical Systems and Linear Algebra*, Academic Press, New York, 1974.

[11] A. KELLY, *The stable, center-stable, center, center-unstable, and unstable manifolds*, in Transversal Mappings and Flows, Benjamin, New York, 1967.

[12] H. G. OTHMER, *Synchronization, phase-locking and other phenomena in coupled cells*, in Temporal Order, L. Rensing and N. I. Jaeger, eds., Springer-Verlag, Heidelberg, 1985, pp. 130–143.

[13] H. G. OTHMER AND J. ALDRIDGE, *The effects of cell density and metabolite flux on cellular dynamics*, J. Math. Biol., 5 (1978) pp. 169–200.

[14] R. J. SACKER, *A new approach to the perturbation theory of invariant surfaces*, Comm. Pure Appl. Math., XVIII (1965), pp. 717–732.

[15] M. WATANABE, *Bifurcation of invariant tori and periodic solutions in systems of coupled oscillators*, Ph.D. thesis, University of Utah, Salt Lake City, UT, 1987.

[16] M. WATANABE AND H. G. OTHMER, *Persistence of invariant tori in systems of coupled oscillators*, I: *Regular and singular problems*, Differential and Integral Equations, 4 (1991), pp. 331–368.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1631-1650, November 1991

007

THE EXISTENCE OF INFINITELY MANY TRAVELING FRONT AND BACK WAVES IN THE FITZHUGH–NAGUMO EQUATIONS*

BO DENG†

Abstract. Consideration is given to the FitzHugh-Nagumo equations of bistable type. The existence of traveling front and back waves with any finite number of pulses is proved. The speed of such a multiple pulse wave is characterized by its number of pulses: the more pulses it has, the slower it travels. Traveling impulse and traveling train solutions are also found. These traveling waves arise from the bifurcation of a doubly twisted front-back wave loop. The method is based on the theory of heteroclinic loop bifurcation, the geometric theory of singular perturbation and the Melnikov method.

Key words. traveling wave, twisted heteroclinic loop, singular perturbation, Melnikov integral

AMS(MOS) subject classifications. primary 35K57; secondary 34B99, 34C28, 34C45, 34D15

1. Introduction.

Consider the FitzHugh-Nagumo equations

$$v_t = v_{xx} + f(v) - w, \qquad w_t = \varepsilon(v - \gamma w), \tag{1.1}$$

where $f(v) = v(v-a)(1-v)$ is a cubic polynomial and $0 < a < \frac{1}{2}$, $\varepsilon > 0$, and $\gamma > 0$ are parameters. A solution $(v, w)(x, t)$ that is bounded over $x \in \mathbb{R}$ and $t \in \mathbb{R}$ is called a traveling wave if it is a function of one variable and there is a constant c so that $(v, w)(x, t) = (v, w)(x + ct)$.

The simplest traveling waves might be constant, or steady-state solutions. Depending on the value of γ, there are one, two, or three steady states (cf. Fig. 1.1), which are the intersections of the nullclines $w = f(v)$ and $v = \gamma w$. In this paper, however, we are interested in the case when γ is greater than the critical value $\gamma_1 := v_{\max}/f_{\max}$ and there are three steady states. We restrict our attention further to the leftmost and rightmost stable states, which we denote by $\mathbf{a}_1$ and $\mathbf{a}_2$ in Fig. 1.1, respectively, and

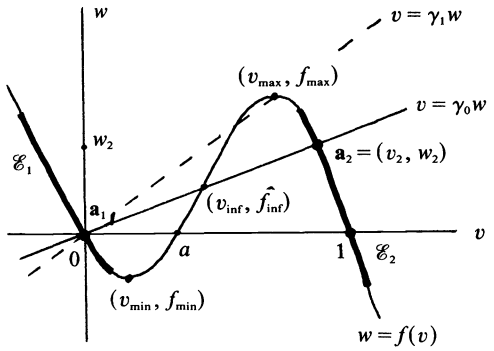


FIG. 1.1. *γ_0 is chosen so that $\mathbf{a}_1$ and $\mathbf{a}_2$ are symmetric with respect to the inflection point $(v_{\inf}, f_{\inf})$ of the cubic curve $w = f(v)$. The thickly drawn segments $\mathscr{E}_1$ and $\mathscr{E}_2$ on $w = f(v)$ do not contain the extreme points but they are long enough so that $\mathscr{E}_1$ contains $\mathbf{a}_1 = 0$ and the intersect point $\{w = w_2\} \cap \{w = f(v)\}$ and $\mathscr{E}_2$ contains $\mathbf{a}_2$ and $(1, 0)$ as interior points, respectively.*

* Received by the editors September 28, 1989; accepted for publication (in revised form) October 30, 1990.

† Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0323.

study nonconstant traveling waves connecting $\mathbf{a}_1$ and $\mathbf{a}_2$. The various types of connecting waves considered here are as follows. A traveling wave $(v, w)(x+ct)$ is said to be a *traveling front* if

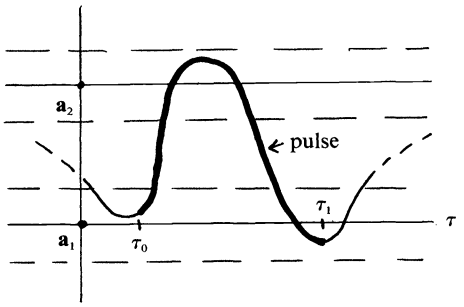
$$\lim_{t\to-\infty}(v, w)(x+ct)=\mathbf{a}_1 \quad \text{and} \quad \lim_{t\to+\infty}(v, w)(x+ct)=\mathbf{a}_2,$$

exist for all x. Likewise, by a *traveling back* wave we mean the same limits exist except that the first limit is $\mathbf{a}_2$ while the second limit is $\mathbf{a}_1$. A traveling wave solution is said to be an *impulse of* $\mathbf{a}_1$ if

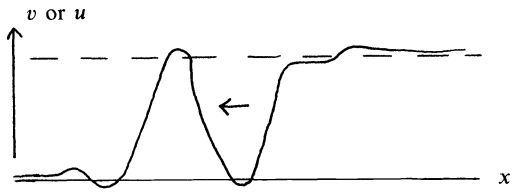
$$\lim_{\tau\to\pm\infty}(v, w)(\tau)=\mathbf{a}_1.$$

An *impulse of* $\mathbf{a}_2$ is analogously defined. Last, a traveling wave solution is said to be a *traveling train* if $(v, w)(\tau)$ is periodic in τ.

We further characterize the waves of the same type according to their numbers of pulses contained. To be precise, choose and fix a neighborhood of each $\mathbf{a}_1$ and $\mathbf{a}_2$ for the equations. A traveling wave has a *pulse* from $\mathbf{a}_1$ if there is a closed interval $\tau_0 \leqq \tau \leqq \tau_1$ such that $(v, w)(\tau)$ arises from the chosen vicinity of $\mathbf{a}_1$, enters into the vicinity of $\mathbf{a}_2$, and only afterwards falls back to the vicinity of $\mathbf{a}_1$ as τ increases from τ_0 to τ_1 (cf. Fig. 1.2). A pulse from $\mathbf{a}_2$ is defined similarly. A front (back, respectively) wave is called *k-front* (k-back, respectively) wave if it has k pulses from $\mathbf{a}_1$. A front (back, respectively) without a pulse is referred to as a simple front (back, respectively). An impulse of $\mathbf{a}_i$ is called a *k-impulse* if it has k pulses from $\mathbf{a}_i$. A traveling train is



(a)



(b)

FIG. 1.2. (a) *One pulse (thickly drawn curve) forms when the traveling wave jumps from a given neighborhood of* $\mathbf{a}_1$ *into a given neighborhood of* $\mathbf{a}_2$ *and only afterwards drops into the neighborhood of* $\mathbf{a}_1$ *again over an interval* $\tau\in[\tau_0, \tau_1]$.

(b) *When the traveling speed* $c>0$, *a 1-front here moves to the left with time.*

called a *k-train* if within the minimum period of the periodic traveling wave there are k pulses. An impulse or a traveling train is simple if there is only one pulse. We emphasize that the comparison among fronts or other types of waves in terms of the number of pulses makes sense only if the neighborhoods of $\mathbf{a}_1$ and $\mathbf{a}_2$ are fixed, but they are allowed to vary with parameters.

The profiles of a given traveling wave are simply the graphs of v and w over the real line (cf. Fig. 1.2). Thus, a traveling wave moves to the left with time if the traveling velocity c is positive. Likewise, it travels to the right if $c<0$. Also, it is trivial to verify that if $(v, w)(x+ct)$ is a traveling wave then $(\bar{v}, \bar{w})(x+(-c)t) := (v, w)(-x+ct)$ is another traveling wave solution, traveling in the opposite direction. For this reason, we only consider the existence of leftward traveling waves (i.e., with $c>0$) from now on.

By smoothness in this paper we mean differentiability of as many times as needed. Our main result is the following theorem.

THEOREM 1.1. *Let $0<a<\frac{1}{2}$ be fixed in the FitzHugh–Nagumo equations* (1.1). *There exists a small ε_0 and two smooth functions $\gamma(\varepsilon)$ and $\delta(\varepsilon)$, $0 \leqq \varepsilon \leqq \varepsilon_0$, such that the following is satisfied for all $0<\varepsilon<\varepsilon_0$.*

(a) *On the relevant (γ, c)-parameter space there are two smooth curves $c = c_{i,0}(\gamma)$ defined on the interval $|\gamma-\gamma(\varepsilon)|<\delta(\varepsilon)$ and $i=1, 2$ ($c_{i,0}$ here and all other curves below are smoothly parametrized by ε also, but ε is usually suppressed for simplicity) such that* (1.1) *has a simple front wave of speed $c_{1,0}(\gamma)$ and a simple back wave of speed $c_{2,0}(\gamma)$.*

(b) *There is a sequence $\{c_{1,k}(\gamma)\}_{k=1}^{\infty}$ of smooth curves of the left half interval $0<\gamma(\varepsilon)-\gamma<\delta(\varepsilon)$ such that* (1.1) *has a k-front wave of speed $c_{1,k}(\gamma)$ for every $k=1, 2, \cdots$ and $0<\gamma(\varepsilon)-\gamma<\delta(\varepsilon)$. Similarly, there is a sequence $\{c_{2,k}(\gamma)\}_{k=1}^{\infty}$ of smooth curves of the right half interval $0<\gamma-\gamma(\varepsilon)<\delta(\varepsilon)$ such that* (1.1) *has a k-back wave of speed $c_{2,k}(\gamma)$ for every $k=1, 2, \cdots$ and $0<\gamma-\gamma(\varepsilon)<\delta(\varepsilon)$.*

(c) *There is a smooth curve $c_{1,\infty}(\gamma)$ of the left half interval $0<\gamma(\varepsilon)-\gamma<\delta(\varepsilon)$ such that* (1.1) *has a simple impulse wave of $\mathbf{a}_1$ with speed $c_{1,\infty}(\gamma)$ for every $0<\gamma(\varepsilon)-\gamma<\delta(\varepsilon)$. Similarly, there is a smooth curve $c_{2,\infty}(\gamma)$ of the right half interval $0<\gamma-\gamma(\varepsilon)<\delta(\varepsilon)$ such that* (1.1) *has a simple impulse wave of $\mathbf{a}_2$ with speed $c_{2,\infty}(\gamma)$ for every $0<\gamma-\gamma(\varepsilon)<\delta(\varepsilon)$.*

(d) *The simple front and back wave curves $c_{i,0}(\gamma)$ intersect transversely at $\gamma(\varepsilon)$. The intersection point, $(\gamma(\varepsilon), c(\varepsilon))$, is smooth in ε. At $\varepsilon=0$, $(\gamma(0), c(0)) = (9/(2-a)(1-2a), (1-2a)/\sqrt{2}) := (\gamma_0, c_0)$. The slopes of $c_{i,0}(\gamma)$ satisfy $c'_{1,0}(\gamma_0)=0$ and $c'_{2,0}(\gamma_0)<0$, respectively, at $\varepsilon=0$. Moreover, for fixed $\varepsilon>0$, γ and $i=1, 2$, the sequence $\{c_{i,k}(\gamma)\}$ is monotone decreasing in $k=0, 1, 2, \cdots$ and converges to the corresponding impulse curve $c_{i,\infty}(\gamma)$ as $k\to\infty$. Furthermore, every $c_{1,k}$ curve is asymptotically tangent to the $c_{1,0}$ curve from the left of $\gamma(\varepsilon)$ as $\gamma\to\gamma(\varepsilon)^-$ and, similarly, every $c_{2,k}$ curve is asymptotically tangent to the $c_{2,0}$ curve from the right of $\gamma(\varepsilon)$ as $\gamma\to\gamma(\varepsilon)^+$.*

(e) *There is a neighborhood of $(\gamma(\varepsilon), c(\varepsilon))$ in the (γ, c)-parameter space for each $0<\varepsilon<\varepsilon_0$ such that* (1.1) *has a simple traveling train solution for some speed c if and only if (γ, c) is in this neighborhood and lies below the curve $c_{1,\infty} \cup c_{2,\infty} \cup (\gamma(\varepsilon), c(\varepsilon))$. See Fig.* 1.3.

It would be of mathematical interest to assume that the recovering variable w also diffuses slightly along the spatial line. This leads to the consideration of the following system of reaction diffusion equations:

$$v_t = v_{xx} + f(v) - w, \qquad w_t = \kappa w_{xx} + \varepsilon(v - \gamma w). \tag{1.2}$$

This is the same as the FitzHugh–Nagumo equations (1.1) except for a small diffusion term κw_{xx} with $|\kappa| \ll 1$ in the w-equation. It is easy to see that the steady states of (1.2),

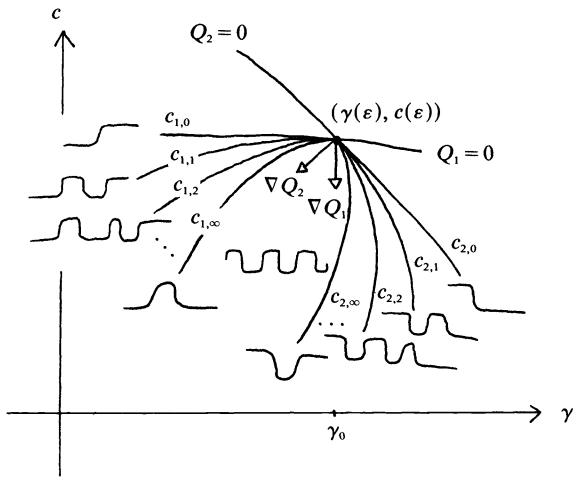


FIG. 1.3. *The bifurcation diagram of Theorem* 1.1. Q_1 *and* Q_2 *here are the Melnikov functions associated with the simple front wave and the simple back wave, respectively. The simple front (back, respectively) wave curve* $c_{1,0}$ ($c_{2,0}$, *respectively) is the level set of* $Q_1=0$ ($Q_2=0$, *respectively*).

in particular, the steady states $\mathbf{a}_1$ and $\mathbf{a}_2$, are the same as those of (1.1) and all the definitions for traveling waves of different types considered above can be directly extended for (1.2). As we will see later, the following theorem is a direct consequence of Theorem 1.1 in the context of singular perturbation.

THEOREM 1.2. *Let* $0<a<\frac{1}{2}$ *be fixed in* (1.2). *There are small constants* ε_0 *and* κ_0 *and smooth functions* $\gamma(\varepsilon,\kappa)$, $c(\varepsilon,\kappa)$, *and* $\delta(\varepsilon,\kappa)$ *of* $0\leqq\varepsilon\leqq\varepsilon_0$ *and* $|\kappa|\leqq\kappa_0$ *such that for* $0<\varepsilon<\varepsilon_0$ *and* $|\kappa|\leqq\kappa_0$ *all the conclusions of Theorem* 1.1 *are satisfied when* $\gamma(\varepsilon,\kappa)$, $c(\varepsilon,\kappa)$, *and* $\delta(\varepsilon,\kappa)$ *are substituted for* $\gamma(\varepsilon)$, $c(\varepsilon)$, *and* $\delta(\varepsilon)$, *respectively. Moreover,* $\gamma(0,\kappa)=\gamma_0$ *and* $c(0,\kappa)=c_0$, *the same constants as in Theorem* 1.1. *In this case, of course, all the curves* $c_{i,k}$ *depend smoothly on* κ *as well.*

The FitzHugh–Nagumo equations have been studied extensively for the last two decades. This system of reaction-diffusion equations is a qualitative model for several applications including nerve conduction (Hodgkin and Huxley (1952), FitzHugh (1961), and Nagumo, Arimoto, and Yoshizawa (1963)), neuronal interactions at the population level (Wilson and Cowan (1972)), chemical and biochemical reaction (Ortoleva and Ross (1975)), as well as electronic transmission lines (Nagumo, Arimoto, and Yoshizawa (1963)). The references quoted here only reflect the author's limited understanding on this subject. For the case when $\gamma=0$, Hastings (1974), (1976) and Casten, Cohen, and Lagerstrom (1975) have studied the existence of impulse and traveling train solutions. Rinzel and Keller (1973) have studied the same problem except that the function f is replaced by a piecewise continuous function $f=H(v-a)-v$ with H to be the Heaviside step function. For the case when ε, a, γ are all small, Hastings (1982) and Evans, Fenichel, and Feroe (1982) have studied the existence of impulse solutions of double pulses and traveling trains of multiple pulses. Their results are closely related to the saddle-focus homoclinic bifurcation theorem by Sil'nikov (1970). Feroe (1982) and, more recently, Wang (1988I) have also studied this type of phenomenon with the piecewise linear function f. For large γ, Carpenter (1977) has studied the existence of traveling front and back waves as well as impulse and traveling train solutions through a constructive singular perturbation approach. Rinzel and Terman (1982) have also considered the same problem for the piecewise

linear case. Indeed, as pointed out by one of the referees of this paper, the existence of traveling waves for (1.1) has been the subject of many other researchers including Aronson and Weinberger (1975), Conley (1975), Greenberg (1973), Keener (1980), Langer (1980), McKean (1970), Pauwelussen (1980), Rauch and Smoller (1978), and probably many more. With only a few exceptions, all the traveling wave solutions investigated so far can be characterized as simple wave solutions in our terminology. The multiple pulse front and back waves obtained here have not previously been proved to exist, nor investigated numerically.

The proof of the theorem is based on three important theories in dynamical systems, namely, the bifurcation theory of a doubly twisted heteroclinic loop by Deng (1991), the geometric singular perturbation theory by Fenichel (1979), and the Melnikov method. Although the idea of using the Melnikov integral together with singular perturbation theory can be found in Kokubu, Nishiura, and Oka (1988) and Lin (1989), our singular perturbation approach is quite different. Whereas other researchers emphasize the "singular" aspects of the problems which inevitably lead to techniques like asymptotic expansion, matching principle, etc., we only need to address the "regular" aspects of the singular perturbation problems. This point of view is taken from Fenichel (1979) which asserts that a singular perturbation problem is essentially a regular perturbation problem in terms of invariant manifold theory, in particular, the center manifold theory. Extending his idea via invariant manifold theory to connecting orbits, we naturally see some connections between the singular perturbation theory and the Melnikov method.

This paper is organized as follows. In § 2, we will state the heteroclinic loop bifurcation theory from Deng (1991) from which the proof of our main result Theorem 1.1 will be derived. The remaining sections are devoted to verifying all the conditions of that theorem. Specifically, we will introduce the Melnikov function (which was called the separation function by Kokubu, Nishiura, and Oka (1988)) in § 3. In §§ 4–8, all the nondegenerate conditions of Theorem 2.1 will be verified. At first sight, these conditions may appear next to impossible to check. However, since they are all generic with respect to the existence of a heteroclinic loop, it is not too surprising to see that the existence of the twisted loop, which only requires our extended Melnikov method, indeed contains enough information for its genericity. In § 9, Theorem 1.2 will be proved based on the singular perturbation method and the proof of Theorem 1.1.

In Theorem 1.1(e) it appears that the neighborhood around $(\gamma(\varepsilon), c(\varepsilon))$ where the traveling train solutions may occur, as well as the lengths of those curves for traveling front, back and impulse waves, depend on the parameter ε. In fact, they can be made uniform for all small $\varepsilon > 0$. Unfortunately, we cannot show this fact in this paper. It requires some nontrivial modification and generalization of our Theorem 2.1 to singularly perturbed systems. The same comment also applies to Theorem 1.2 with respect to $\varepsilon > 0$ and $|\kappa| \ll 1$.

This paper was originally inspired by the work of Rinzel and Terman (1982). David Terman helped me understand their work correctly. This made it possible for me to find the twist structure of the front-back wave loop at the bifurcation point (γ_0, c_0) through their numerical bifurcation diagram for the simple front, simple back, and simple impulse waves (cf. Fig. 1.3). Finally, let us mention the other equally important motivation that lies beyond the scope of this paper. We would like to eventually prove that all the multiple pulse front and back waves found here are stable with respect to the PDEs (1.1) and (1.2). The fact that the steady states $\mathbf{a}_1$, $\mathbf{a}_2$, the simple front, simple back, simple impulse, and simple train solutions near the bifurcation point (γ_0, c_0) are all stable is well known. (See, e.g., Rinzel and Terman (1982)

and the references therein. See also Wang (1988II) for the stability problem of multiple pulse impulses for the piecewise linear case.)

2. The bifurcation of a twisted heteroclinic loop. Recall that a traveling wave (v, w) is a function of one variable τ. Thus, if we let $u := v'$, where the prime denotes the derivative in τ, then (v, u, w) satisfies the following system of first-order ordinary differential equations:

$$v' = u, \quad u' = cu - f(v) + w, \quad w' = \frac{\varepsilon}{c}(v - \gamma w). \tag{2.1}$$

For $\varepsilon > 0$, it is trivial to check that this system has three equilibrium points when $\gamma > \gamma_1$ (cf. Fig. 1.1). They are those points (v, u, w) satisfying that $u = 0$ and (v, w) equals to the steady states of (1.1) discussed in the Introduction. For this reason and for simplicity, we will denote throughout the equilibria with $u = 0$, $(v, w) = \mathbf{a}_i$ just by $\mathbf{a}_i$ alone. Also, as the counterpart of traveling front of (1.1), a solution of (2.1) is call a *heteroclinic* orbit from $\mathbf{a}_1$ to $\mathbf{a}_2$ if

$$\lim_{\tau \to -\infty} (v, u, w)(\tau) = \mathbf{a}_1 \quad \text{and} \quad \lim_{\tau \to +\infty} (v, u, w)(\tau) = \mathbf{a}_2.$$

A heteroclinic orbit from $\mathbf{a}_2$ to $\mathbf{a}_1$ is defined analogously. On the other hand, a solution of (2.1) is called a *homoclinic* orbit to $\mathbf{a}_1$ if

$$\lim_{|\tau| \to \infty} (v, u, w)(\tau) = \mathbf{a}_1.$$

A similar definition applies to a homoclinic orbit to $\mathbf{a}_2$. Note that a heteroclinic orbit of (2.1) from $\mathbf{a}_1$ to $\mathbf{a}_2$ (respectively, from $\mathbf{a}_2$ to $\mathbf{a}_1$) gives rise to a traveling front (respectively, back) wave of (1.1) while a homoclinic orbit of (2.1) gives rise to an impulse wave of (1.1). In a similar way as in the previous section, we can define *k-heteroclinic*, *k-homoclinic*, and *k-periodic* orbits with respect to some neighborhoods of $\mathbf{a}_1$ and $\mathbf{a}_2$ for (2.1). We leave this to the reader. Therefore, our strategy to prove Theorem 1.1 is to prove the same theorem except that (1.1) is replaced by (2.1) and the traveling fronts, etc., are replaced by heteroclinic orbits, etc., respectively. To do this, we will apply a theorem on the bifurcations of a twisted heteroclinic loop from Deng (1991) (cf. Theorem B of Deng (1991)). Although that theorem as well as the result on singular perturbation by Fenichel and the method of Melnikov integral are available for any finite-dimensional system, we will treat them only in $\mathbb{R}^3$ here for simplicity. Also, we need to warn the reader in advance that the theorem we are about to state is the time-reversed version of Theorem B of Deng (1991). Note that upon time reversal, the stable manifold becomes the unstable manifold and vice versa. Also, a heteroclinic orbit from $\mathbf{a}_1$ to $\mathbf{a}_2$ becomes a heteroclinic orbit from $\mathbf{a}_2$ to $\mathbf{a}_1$, and so on. But a homoclinic orbit of $\mathbf{a}_1$ remains the same.

To state the theorem we begin with its hypotheses (2.2a)-(2.2e). Since the bifurcation problem to be discussed is of codimension two, we will only include the so-called relevant parameter α ($\alpha = (\gamma, c)$ in our case) in a vector field $X^\alpha := X^\alpha(x)$, where $x \in \mathbb{R}^3$ and $\alpha \in \mathbb{R}^2$. We certainly allow other parameters (say ε in our case) to be included in the vector field but we will usually suppress them unless otherwise indicated.

(2.2a) *The relative expansion of* $\mathbf{a}_i$. Let X^α denote a family of vector fields in $\mathbb{R}^3$ parametrized by a relevant parameter α in $\mathbb{R}^2$. Suppose $\mathbf{a}_i = \mathbf{a}_i(\alpha)$, $i = 1, 2$ are hyperbolic equilibrium points of X^α for all α and are relatively expansive in the sense that the eigenvalues λ_j of the linearization $DX^\alpha(\mathbf{a}_i)$ satisfy

$$\lambda_1 < \lambda_2 < 0 < \lambda_3 \quad \text{and} \quad \lambda_3 + \lambda_2 > 0.$$

Of course, λ_j here also depends on $\mathbf{a}_i$ for $i=1, 2$. The eigenvalues λ_2 and λ_3 are called the principal stable and the principal unstable eigenvalues, respectively. Their eigenvectors are thus referred to as the principal stable and the principal unstable eigenvectors, respectively (see (1.2) of Deng (1991)).

(2.2b) *The nondegeneracy of a heteroclinic loop.* There is a parameter value α_0 so that the equation $x' = X^{\alpha}(x)$ has a nondegenerate heteroclinic loop. Specifically, there exists a simple heteroclinic orbit $\mathbf{z}_1^*$ from $\mathbf{a}_1$ to $\mathbf{a}_2$ and a simple heteroclinic orbit $\mathbf{z}_2^*$ from $\mathbf{a}_2$ to $\mathbf{a}_1$ at the same parameter α_0. Moreover, by nondegeneracy we mean that the following two conditions are satisfied. First, $\mathbf{z}_i^*(\tau)$ is asymptotically tangent to the principal stable eigenvector of $\mathbf{a}_j$, $j \neq i$ as $\tau \to +\infty$ and the principal unstable eigenvector of $\mathbf{a}_i$ as $\tau \to -\infty$, respectively. Second, the following strong inclination conditions hold:

(2.2b′) $$\lim_{\tau \to -\infty} T_{\mathbf{z}_i^*(\tau)} W_j^s = T_{\mathbf{a}_i} W_i^u + T_{\mathbf{a}_i} W_i^{ss},$$

where $i, j = 1, 2$ and $i \neq j$, $T_{\mathbf{p}} W$ means the tangent space of a given manifold W at a base point $\mathbf{p} \in W$. Also, W_i^s, W_i^u, W_i^{ss} is the standard notation for the stable, unstable, and strong stable manifolds of $\mathbf{a}_i$, respectively. They are two-dimensional, one-dimensional, and one-dimensional, respectively, in this case (see (1.5) and (1.7) of Deng (1991)).

(2.2c) *The double twist of a nondegenerate heteroclinic loop.* Let $\mathbf{z}_1^*$ and $\mathbf{z}_2^*$ form a nondegenerate heteroclinic loop. Let

$$\mathbf{e}_i^- = \lim_{\tau \to -\infty} \frac{\mathbf{z}_i^*(\tau) - \mathbf{a}_i}{\|\mathbf{z}_i^*(\tau) - \mathbf{a}_i\|}, \qquad \mathbf{e}_j^+ = \lim_{\tau \to +\infty} \frac{\mathbf{z}_i^*(\tau) - \mathbf{a}_j}{\|\mathbf{z}_i^*(\tau) - \mathbf{a}_j\|}$$

be the unit principal unstable and stable eigenvectors along which the heteroclinic orbit $\mathbf{z}_i^*$ comes from $\mathbf{a}_i$ and goes towards $\mathbf{a}_j$, respectively, then $\mathbf{e}_i^+$ and $\mathbf{e}_j^-$ point to opposite sides of $T_{\mathbf{z}_i^*(\tau)} W_j^s$ at $\tau \to -\infty$ and $\tau \to +\infty$, respectively. Here, $i, j = 1, 2$, $i \neq j$ (cf. Fig. 2.1 and see (1.9) and Definition 1.1 of Deng (1991)).

Remark. The definition of twisted heteroclinic orbit has much in common with that of twisted homoclinic orbit. The geometric notion of twisted homoclinic orbit was given by Deng (1989) and Chow, Deng, and Fiedler (1990), based on the author's strong λ-lemma. It was inspired by a work of Yanagida (1987).

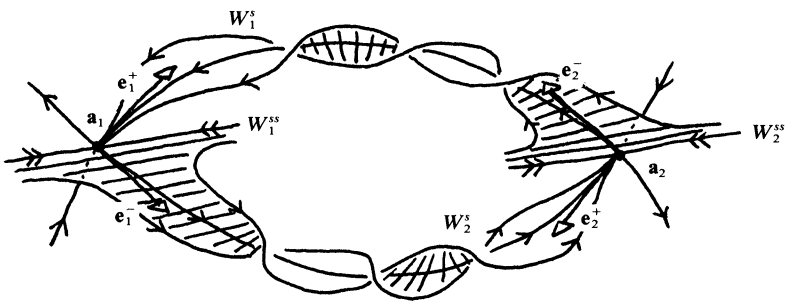


FIG. 2.1. *A nondegenerate and doubly twisted heteroclinic loop in* $\mathbb{R}^3$. $\mathbf{e}_i^-$ *and* $\mathbf{e}_j^+$ *point to the opposite sides of* W_i^s.

(2.2d) *The continuation of* $\mathbf{z}_i^*$. There exist two curves 0-het$_1$ and 0-het$_2$ in the parameter space $\alpha \in \mathbb{R}^2$ which intersect at α_0 transversely so that when $\alpha \in$ 0-het$_i$ there is a simple heteroclinic orbit $\mathbf{z}_{i,\alpha}$ from $\mathbf{a}_i$ to $\mathbf{a}_j$ and $\mathbf{z}_{i,\alpha}$ is the continuation of $\mathbf{z}_i^*$ in the sense that $\mathbf{z}_{i,\alpha_0} = \mathbf{z}_i^*$ and $\mathbf{z}_{i,\alpha}(\tau)$ are continuous in τ and α (see (1.1a) of Deng (1991)).

(2.2e) *The transverse crossing of the stable and unstable manifolds along* $\mathbf{z}_i^*$. Let $Q_i(\alpha)$ be the Melnikov function defined in the next section; then the gradient vectors $\nabla Q_1(\alpha_0)$ and $\nabla Q_2(\alpha_0)$ are linearly independent (see (1.10b) of Deng (1991)). We remark that since the discussions of Melnikov functions in § 3 is independent of what follows, we may certainly find out the precise definition before continuing.

THEOREM 2.1. *Suppose conditions* (2.2a)-(2.2e) *are satisfied. Then the following holds.*

(a) *There is a sequence* $\{k\text{-het}_1\}_{k=1}^{\infty}$ *of smooth curves in* $\mathbb{R}^2$ *such that for every* $k = 1, 2, \cdots$ *the equation* $x' = X^{\alpha}(x)$ *has a k-heteroclinic orbit from* $\mathbf{a}_1$ *to* $\mathbf{a}_2$ *if and only if* $\alpha \in k\text{-het}_1$. *Similarly, there is a sequence* $\{k\text{-het}_2\}_{k=1}^{\infty}$ *of smooth curves in* $\mathbb{R}^2$ *such that for every* $k = 1, 2, \cdots$ *there is a k-heteroclinic orbit from* $\mathbf{a}_2$ *to* $\mathbf{a}_1$ *if and only if* $\alpha \in k\text{-het}_2$.

(b) *There is a smooth curve* hom$_i$ *for each* $i = 1, 2$ *such that there is a simple homoclinic orbit of* $\mathbf{a}_i$ *if and only if* $\alpha \in$ hom$_i$.

(c) *The curve* 0-het$_i$ *is simply the level set of* $Q_i = 0$, *and it is divided by the other* 0-het$_j$ *curve into two parts. Let* 0-het$_i^+$ *be the half of the* 0-het$_i$ *curve that points to the gradient direction of* ∇Q_j (*cf. Fig.* 2.2). *Then all the curves* $\{k\text{-het}_1\}_{k=1}^{\infty}$ *together with*

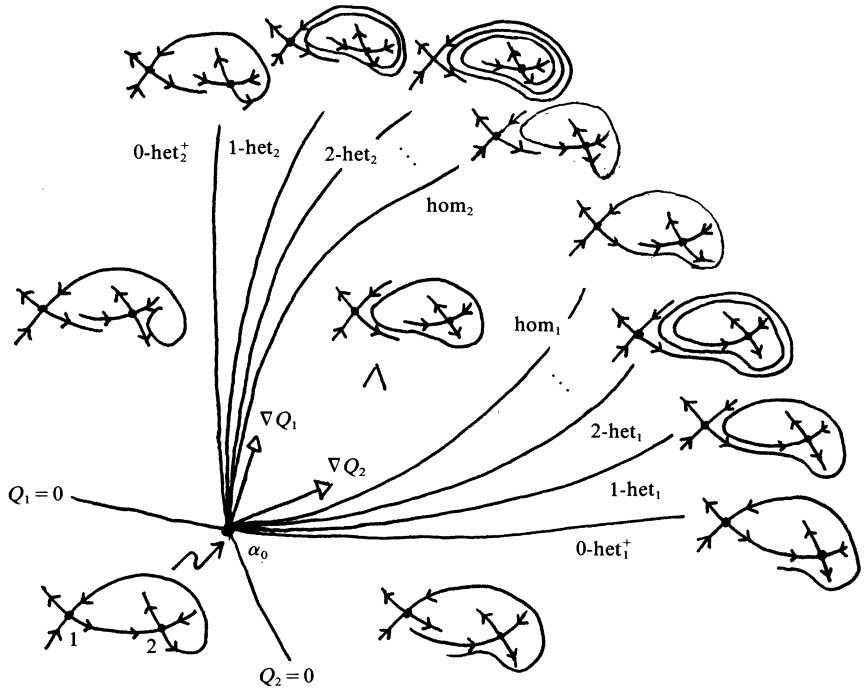


FIG. 2.2. *The bifurcation diagram of Theorem* 2.1. *The doubly twisted heteroclinic loop is drawn in* $\mathbb{R}^2$. ∇Q_i *and* 0-het$_j^+$ *point to the same side of the* 0-het$_i$ *curve which is the level set* $Q_i = 0$ *and* Q_i *is the Melnikov function for the primary connection from* $\mathbf{a}_i$ *to* $\mathbf{a}_j$ *when* $\alpha = \alpha_0$.

hom_1 *are in the sector bounded by* 0-het_1^+ *and* 0-het_2^+ *and converge to* α_0 *asymptotically tangent to* 0-het_1^+. *Moreover,* $\{k\text{-ket}_1\}_{k=1}^{\infty}$ *lies between the* 0-het_1^+ *and* hom_1 *curves and converges to the* hom_1 *curve monotonely in* $k=1, 2, \cdots$ (*cf. Fig.* 2.2). *Analogous result holds for the sequence* $\{h\text{-het}_2\}_{k=1}^{\infty}$ *and the* hom_2 *curve.*

(d) *Let* Λ *be the sector bounded by the homoclinic curves* hom_1 *and* hom_2. *Then there is a simple periodic solution of the vector field* X^α *if and only if* $\alpha \in \Lambda$.

Proof. The theorem is the same as Theorem B of Deng (1991) provided that the vector field X^α is replaced by its time-reversed vector field $-X^\alpha$. Thus the proof is complete. What follows is meant to help the reader pin down the parallel comparisons between the Melnikov functions from condition (2.2e) and from that theorem, respectively. First, the Melnikov function Q_j here is essentially a positive constant multiple of $Q_{i\alpha}^{(1)}$ there. Second, the condition (1.10b) of Theorem B is replaced by $\partial Q_{j\alpha}^{(1)}/\partial\alpha_j > 0$ which was really what we used in the proof of Theorem B of Deng (1991). Note that the 0-het_i curve here is the α_i-axis there, which is the level set $Q_{j\alpha}^{(1)}=0$, and that the positive direction of α_i in Theorem B, which is the same as 0-het_i^+ here, was chosen in correspondence with $\partial Q_{i\alpha}^{(1)}/\partial\alpha_i > 0$. This positive derivative in turn is equivalent to that $\nabla Q_{i\alpha}^{(1)}$ and 0-het_i^+ point to the same side of the level set $Q_{i\alpha}^{(1)}=0$. Also, the linear independence of $\nabla Q_{1\alpha}^{(1)}$ and $\nabla Q_{2\alpha}^{(1)}$ is equivalent to the transverse intersection of 0-het_1 and 0-het_2 at α_0. □

3. The Melnikov method. Let $X^\alpha(x)$ be a sufficiently smooth vector field in $\mathbb{R}^3$ with parameter α. Let $\mathbf{a}_1$ and $\mathbf{a}_2$ be two equilibrium points having a two-dimensional stable manifold $W^s(\mathbf{a}_i, \alpha)$ and a one-dimensional unstable manifold $W^u(\mathbf{a}_i, \alpha)$, respectively. In what follows we will write W^s, or $W^s(\mathbf{a}_i)$ interchangeably for the stable manifold and so on for simplicity, provided that there is no confusion involved. Suppose at some α_0 there is a heteroclinic orbit $\mathbf{z}^*$ from $\mathbf{a}_1$ to $\mathbf{a}_2$ (an identical consideration can also be given to $\mathbf{a}_2$ to $\mathbf{a}_1$ connections). Then it must be $\mathbf{z}^* \subset W^u(\mathbf{a}_1) \cap W^s(\mathbf{a}_2)$ at α_0. We would like to know how the heteroclinic connection $\mathbf{z}^*$ changes with the parameter. In many applications, the Melnikov method presented below is very useful for attacking this problem.

Naturally, we would like to examine how the "signed distance" between the stable and unstable manifolds changes with the parameter. To implement this intuitive idea, we choose and fix a point $\mathbf{z}_0^* \in \mathbf{z}^*$ from the orbit and a two-dimensional plane Σ which is perpendicular to $\mathbf{z}^{*\prime}$ and through $\mathbf{z}_0^*$. The intersection $\Sigma \cap W^s$ is necessarily a curve, whereas $\Sigma \cap W^u$ is just a point for every α near α_0. Choose and fix a vector $\mathbf{e}$ on Σ that is perpendicular to the curve $\Sigma \cap W^s$ at $\mathbf{z}_0^*$ and α_0. Let l be a straight line that goes through the point $\mathbf{p}^u(\alpha) := \Sigma \cap W^u(\mathbf{a}_1, \alpha)$ and that is of the direction of $\mathbf{e}$. Then l must intersect the stable manifold $\Sigma \cap W^s(\mathbf{a}_2, \alpha)$ at a unique point $\mathbf{p}^s(\alpha)$ for α sufficiently close to α_0. $\mathbf{p}^u(\alpha)$ and $\mathbf{p}^s(\alpha)$ can be chosen differentiable and satisfying $\mathbf{p}^s(\alpha_0) = \mathbf{p}^u(\alpha_0) = \mathbf{z}_0^*$. Now, there must be a smooth function $Q(\alpha)$ such that

$$\mathbf{p}^s(\alpha) - \mathbf{p}^u(\alpha) = Q(\alpha)\mathbf{e} \quad \text{or} \quad Q(\alpha) = (\mathbf{p}^s(\alpha) - \mathbf{p}^u(\alpha)) \cdot \mathbf{e}/\|\mathbf{e}\|^2. \tag{3.1}$$

See Fig. 3.1. The function $Q(\alpha)$ serves what we called the "signed distance" between $W^s(\mathbf{a}_2)$ and $W^u(\mathbf{a}_1)$ above. We will call it the Melnikov function (or the separation function by Kokubu, Nishiura, and Oka (1988)). We are interested in the solutions of $Q(\alpha)=0$ since that precisely gives rise to those parameters at which there is a heteroclinic orbit from $\mathbf{a}_1$ to $\mathbf{a}_2$.

Several modifications can be made in the construction of $Q(\alpha)$ above. The requirements that Σ be perpendicular to $\mathbf{z}^*$ and that $\mathbf{e}$ be perpendicular to $\Sigma \cap W^s$ are not necessary. For instance, take the case where Σ is the same as above but $\mathbf{e}$ is replaced

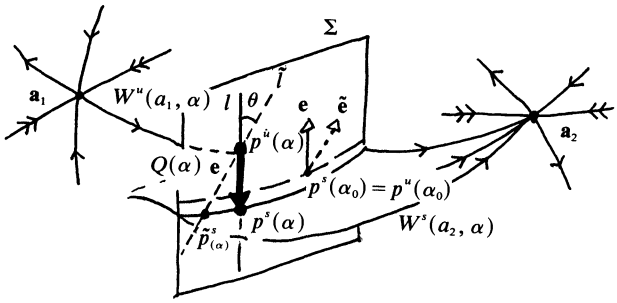


FIG. 3.1. *The long dashed curve represents the intersection curve of the stable manifold of* $\mathbf{a}_2$ *at the bifurcation point* $\alpha = \alpha_0$ *and the plane* Σ. *The Melnikov function* $Q(\alpha)$ *is defined as* $Q(\alpha) = (\mathbf{p}^s(\alpha) - \mathbf{p}^u(\alpha))\mathbf{e}/\|\mathbf{e}\|^2$.

by another vector $\tilde{\mathbf{e}}$ that is just transverse to the stable manifold curve $\Sigma \cap W^s(\mathbf{a}_2, \alpha_0)$ at $\mathbf{z}_0^*$. Similarly, let $\tilde{l}$ be the line through $\mathbf{p}^u(\alpha)$ that is parallel to $\tilde{\mathbf{e}}$, let $\tilde{\mathbf{p}}^s(\alpha)$ be the corresponding intersection point $\tilde{l} \cap \Sigma \cap W^s(\mathbf{a}_2, \alpha)$, and let $\tilde{Q}(\alpha) = (\tilde{\mathbf{p}}^s(\alpha) - \mathbf{p}^u(\alpha)) \cdot \tilde{\mathbf{e}}/\|\tilde{\mathbf{e}}\|^2$ be the Melnikov function. Then it is easy to see from Fig. 3.1 that there is a nonzero constant β independent of α so that

$$\tilde{Q}(\alpha) = \beta Q(\alpha) + o(|\tilde{\mathbf{p}}^s(\alpha) - \mathbf{p}^s(\alpha)|).$$

In fact, $\beta = \|\mathbf{e}\| \cos\theta/\|\tilde{\mathbf{e}}\|$ and β is positive (respectively, negative) if $\mathbf{e}$ and $\tilde{\mathbf{e}}$ point to the same (respectively, opposite) side of $W^s(\mathbf{a}_2) \cap \Sigma$, where θ is the angle between $\mathbf{e}$ and $\tilde{\mathbf{e}}$ (cf. Fig. 3.1). It is clear that if we want to solve the equation $Q(\alpha) = 0$ by the implicit function theorem it is important to know the behavior of the partial derivatives of Q at the bifurcation point $\alpha = \alpha_0$. But, this alternative definition $\tilde{Q}$ simply says that both functions are essentially the same in the sense that

$$\frac{\partial \tilde{Q}(\alpha_0)}{\partial \alpha} = \beta \frac{\partial Q(\alpha_0)}{\partial \alpha}. \tag{3.2}$$

The same conclusion also holds true if we relax the choice of Σ to be a plane transverse, instead of perpendicular, to the heteroclinic orbit $\mathbf{z}^*$. We also remark that the definition of the Melnikov function above is not necessarily just restricted to vector fields in $\mathbb{R}^3$. It can be easily extended to stable and unstable manifolds of any finite dimensions under the condition that they are in general position along $\mathbf{z}_0^*$. Moreover, in the case of nonhyperbolic equilibrium points, we can analogously define the Melnikov function between $W^u(\mathbf{a}_1, \alpha)$ and $W^{cs}(\mathbf{a}_2, \alpha)$. Another extension we will need later is the "signed distance" between points of a center unstable manifold $W^{cu}(\mathbf{a}_1, \alpha)$ and a center stable manifold $W^{cs}(\mathbf{a}_2, \alpha)$ whose dimensions satisfy, for our consideration only, dim W^{cs} = dim $W^{cu} = 2$ and dim $W^c = 1$. Let Σ and $\mathbf{e}$ be the same as above. Then $\Sigma \cap W^{cu}(\mathbf{a}_1, \alpha)$ is a curve too. Suppose this curve is parametrized by $w \in [-1, 1]$ so that $w = 0$ always corresponds to the intersection point $\Sigma \cap W^u(\mathbf{a}_1, \alpha)$. Let $\mathbf{p}^{cu}(\alpha, w)$ be a given point from $\Sigma \cap W^{cu}$ and $\mathbf{p}^{cs}(\alpha, w)$ be the corresponding point of $l \cap \Sigma \cap W^{cs}$, where l is the line through $\mathbf{p}^{cu}(\alpha, w)$ with direction $\mathbf{e}$. Define

$$q(\alpha, w) = (\mathbf{p}^{cs}(\alpha, w) - \mathbf{p}^{cu}(\alpha, w)) \cdot \mathbf{e}/\|\mathbf{e}\|^2, \tag{3.3}$$

which represents a differential along the $\mathbf{e}$ direction from W^{cu} to W^{cs}. Of course, we have $q(\alpha, 0) = Q(\alpha)$, where $Q(\alpha)$ is the Melnikov function between $W^u(\mathbf{a}_1)$ and $W^{cs}(\mathbf{a}_2)$. The purpose to introduce this function $q(\alpha, s)$ is to relate the condition

$$\frac{\partial q(\alpha_0, 0)}{\partial w} \neq 0$$

to the transverse intersection of $W^{cu}(\mathbf{a}_1, \alpha_0)$ and $W^{cs}(\mathbf{a}_2, \alpha_0)$ along the connection $\mathbf{z}^*$ at α_0. The other useful property is

$$\frac{\partial Q(\alpha_0)}{\partial \alpha} = \frac{\partial q(\alpha_0, 0)}{\partial \alpha}, \tag{3.4}$$

which will be used later in computing the derivative of $q(\cdot, 0)$.

Now, let us return to Theorem 2.1, in particular, the choice of the Melnikov functions $Q_1(\alpha)$ and $Q_2(\alpha)$ from the hypothesis (2.2e). Note that we explicitly talked about the directions $\nabla Q_i(\alpha)$ which determine the bifurcation directions of those multiple pulse heteroclinic orbits in our main theorem. But, on the other hand, we have the freedom of choosing either $\mathbf{e}$ or $-\mathbf{e}$, which is also transverse to $W^s(\mathbf{a}_2, \alpha_0) \cap \Sigma$, in the definition of the Melnikov function $Q(\alpha)$. Thus, from now on we will specify the direction $\mathbf{e}$. To this end, recall the strong inclination limit (2.2b′). From that condition, we can easily conclude that the principal stable unit eigenvector $\mathbf{e}_1^+$ is transverse to the stable manifold $W^s(\mathbf{a}_2, \alpha_0)$ near $\mathbf{a}_1$. Thus, $\mathbf{e}_1^+$ defines an orientation for $W^s(\mathbf{a}_2, \alpha_0)$. Now, in the definition of $Q_1(\alpha)$, choose a vector $\mathbf{e}$ which points to the same side of $W^s(\mathbf{a}_2)$ at $\mathbf{z}_0^*$ as $\mathbf{e}_1^+$ does up to the flow homotopy. Indeed, when $\mathbf{z}_0^*$ is sufficiently close to $\mathbf{a}_1$, we can simply let $\mathbf{e} = \mathbf{e}_1^+$. And, as mentioned earlier, $Q_1(\alpha)$ can be chosen to be (or essentially to be) a positive constant multiple of $Q_{2\alpha}^{(1)}$ in the proof of Theorem B of Deng (1991), or (5.41), (5.43) of Chow, Deng, and Terman (1990).

So much for the theoretical aspect of the Melnikov function $Q(\alpha)$. When an application comes, what really matters is the so-called Melnikov integral which provides us with a computable formula for the derivative $\partial Q(\alpha_0)/\partial \alpha$. We introduce this integral below.

Without loss of generality, let Σ and $\mathbf{e}$ be perpendicular to $\mathbf{z}^*$ and $\Sigma \cap W^s$, respectively. Let the orbit $\mathbf{z}^*(\tau)$ be parametrized so that $\mathbf{z}^*(0) = \mathbf{z}_0^*$ and $\mathbf{z}^*(\tau)$ satisfies the equation $x' = X^{\alpha_0}(x)$. Consider the variational equation $y' = DX^{\alpha_0}(\mathbf{z}^*(\tau))y$ along $\mathbf{z}^*$ and its adjoint equation $y' = -(DX^{\alpha_0}(\mathbf{z}^*(\tau))^T y$. Then there is a unique bounded solution $\varphi(\tau)$, $\tau \in \mathbb{R}$, of the adjoint equation with the initial condition $\varphi(0) = \mathbf{e}$ (see, e.g., Palmer (1984)). It is well known that

$$\frac{\partial Q}{\partial \alpha}(\alpha_0) = -\int_{-\infty}^{\infty} \varphi(\tau) \cdot \frac{\partial X^{\alpha_0}(\mathbf{z}^*(\tau))}{\partial \alpha}\, d\tau,$$

where $Q(\alpha) = (\mathbf{p}^s(\alpha) - \mathbf{p}^u(\alpha)) \cdot \mathbf{e}/\|\mathbf{e}\|^2$ (see Holmes (1980), Palmer (1984), or Guckenheimer and Holmes (1983)). In particular, when the vector field is two-dimensional and $\mathbf{e}$ is chosen to be of the orthogonal vector $(-z_2^{*\prime}(0), z_1^{*\prime}(0))$ of $\mathbf{z}^{*\prime}(0)$ where $\mathbf{z}^{*\prime}(0) := (z_1^{*\prime}(0), z_2^{*\prime}(0))$ is the component form, it can be directly checked that

$$\varphi(\tau) = \exp\left[-\int_0^{\tau} \operatorname{tr} DX^{\alpha_0}(\mathbf{z}^*(s))\, ds\right](-z_2^{*\prime}(\tau), z_1^{*\prime}(\tau)),$$

where $\operatorname{tr} A$ means the trace of a given square matrix A. See, e.g., Melnikov (1964), Holmes (1980), and Palmer (1984). Note that $\mathbf{e}$ here is uniquely determined (cf. Fig. 3.2). In summary, we have

$$\frac{\partial Q(\alpha_0)}{\partial \alpha} = -\int_{-\infty}^{\infty} \exp\left[-\int_0^{\tau} \operatorname{tr} DX^{\alpha_0}(\mathbf{z}^*(s))\, ds\right](-z_2^{*\prime}(\tau), z_1^{*\prime}(\tau)) \cdot \frac{\partial X^{\alpha_0}(\mathbf{z}^*(\tau))}{\partial \alpha}\, d\tau, \tag{3.5}$$

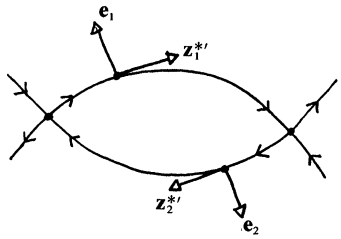


FIG. 3.2. *The unique choice of the vector* **e**.

which is referred to as the Melnikov integral. Last, let us remark that in light of (3.2) we will sometimes slightly abuse the notation by writing $\partial\tilde{Q}(\alpha_0)/\partial\alpha$ as the same Melnikov integral as above provided that the directions **e** and $\tilde{\mathbf{e}}$ point to the same side of the stable manifold. A justification for this is based on the statement of Theorem 2.1 that only the signs of the derivatives of a Melnikov function really count.

4. Proof of conditions (2.2a, d). Beginning with this section, we will show that the hypotheses (2.2a–e) of Theorem 2.1 are satisfied for the reduced FitzHugh–Nagumo equation (2.1). It is straightforward to see that the condition (2.2d) for the continuation of the simple heteroclinic orbits is superfluous. Indeed, it is implied by the existence of the simple heteroclinic orbits $\mathbf{z}_i^*$ at α_0 from condition (2.2b) and the linear independence of $\nabla Q_1(\alpha_0)$ and $\nabla Q_2(\alpha_0)$. The reason to include it in the statement of Theorem 2.1 is simply for a convenient parallel comparison between that theorem and Theorem B of Deng (1991). Thus, the condition (2.2d) may now be removed from our checklist.

To show condition (2.2a), recall the equilibria $\mathbf{a}_i = (v_i, 0, w_i)$ from § 1 and consider the linearization of (2.1) at $\mathbf{a}_i$

$$V' = U, \quad U' = cU - f'(v_i)V + W, \quad W' = \frac{\varepsilon}{c}(V - \gamma W).$$

The corresponding characteristic equation is

$$\Delta(\lambda, \varepsilon) = \lambda(c-\lambda)\left(\frac{\varepsilon\gamma}{c}+\lambda\right) - f'(v_i)\left(\frac{\varepsilon\gamma}{c}+\lambda\right) + \frac{\varepsilon}{c} = 0.$$

When $\varepsilon = 0$, it is straightforward to check that since $f'(v_i) < 0$ for $i = 1, 2$ (cf. Fig. 1.1), $\Delta(\lambda, 0) = 0$ has roots

$$\lambda_1 = \frac{c - \sqrt{c^2 - 4f'(v_i)}}{2} < \lambda_2 = 0 < \lambda_3 = \frac{c + \sqrt{c^2 - 4f'(v_i)}}{2}. \tag{4.1}$$

Since $\Delta(0, \varepsilon) = -f'(v_i)\varepsilon\gamma/c + \varepsilon/c > 0$ for $\varepsilon > 0$ and $\partial\Delta(0, 0)/\partial\lambda = -f'(v_i) > 0$, the second root λ_2 must move to the left of the origin while λ_1 and λ_3 stay uniformly away from the origin for small $\varepsilon > 0$. Hence $\lambda_1(\varepsilon) < \lambda_2(\varepsilon) < 0 < \lambda_3(\varepsilon)$ and $\lambda_3(\varepsilon) + \lambda_2(\varepsilon) > 0$ for small $\varepsilon > 0$ by continuity. This proves condition (2.2a) that the equilibrium points are relatively expansive by definition.

5. Proof of condition (2.2b). The methods used in this section include the geometric theory of singular perturbations by Fenichel (1979) and the Melnikov method discussed above. It is a rather long section but it contains all the information we will need for verifying the remaining conditions (2.2c–e).

Let us begin with the singular perturbation. The singular parameter for (2.1) is ε. When $\varepsilon = 0$, it becomes

$$v' = u, \quad u' = cu - f(v) + w, \quad w' = 0. \tag{5.1a}$$

Note that the variable w can be regarded as a parameter and the entire cubic curve $w = f(v)$ on the plane $u = 0$ consists of equilibrium points of (5.1a). Let $\mathscr{E}_i$ be a bounded, connected and closed segment on the cubic curve $w = f(v)$ which contains $\mathbf{a}_i$ and the intersection point $\{w = w_j\} \cap \{w = f(v)\}$ but does not contain any of the extreme points of the cubic curve (cf. Fig. 1.1). It is easy to check that the linearization at $\mathscr{E}_i$ is

$$V' = U, \quad U' = cU - f'(v)V + W, \quad W' = 0,$$

where v are those points that $\{w = f(v)\} \subset \mathscr{E}_i$, and thus $f'(v) < 0$ since $\mathscr{E}_i$ does not contain any extreme point. Similar to (4.1), the roots for the characteristic equation are the same as λ_j in (4.1) except that $f'(v_i)$ there is now replaced by $f'(v)$. It is also straightforward to check that the corresponding eigenvectors are

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ \lambda_1 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 0 \\ -f'(v) \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ \lambda_3 \\ 0 \end{pmatrix}. \tag{5.1b}$$

See Fig. 5.1. Since λ_1 and λ_3 are strictly nonzero, the eigendirections $\mathbf{v}_1$ and $\mathbf{v}_3$ are normal to $\mathscr{E}_i$. Therefore, $\mathscr{E}_i$ is normally hyperbolic according to Fenichel (1979). Thus, by Theorem 9.1 of Fenichel (1979), there exists a global center-stable manifold $W^{cs}_\varepsilon(\mathscr{E}_i)$ and a global center-unstable manifold $W^{cu}_\varepsilon(\mathscr{E}_i)$. These manifolds are smooth in ε for $|\varepsilon| \ll 1$. Moreover, when $\varepsilon > 0$, W^{cs}_ε is precisely the stable manifold $W^s(\mathbf{a}_i, \varepsilon)$ of $\mathbf{a}_i$ while the unstable manifold $W^u(\mathbf{a}_i, \varepsilon)$ of $\mathbf{a}_i$ is the unstable fiber, called $\mathscr{F}^u_\varepsilon$ by Fenichel (1979), through $\mathbf{a}_i$. $W^u(\mathbf{a}_i, \varepsilon)$ is also smooth in ε for $|\varepsilon| \ll 1$. Furthermore, all the invariant manifolds above depend smoothly on all other parameters. Thus, our strategy now becomes to show the existence of a connection, i.e., $W^u(\mathbf{a}_i, \varepsilon) \cap W^{cs}_\varepsilon(\mathscr{E}_j) \neq \emptyset$, when $\varepsilon = 0$ and to show the continuation of that connection for $\varepsilon > 0$ by the Melnikov method.

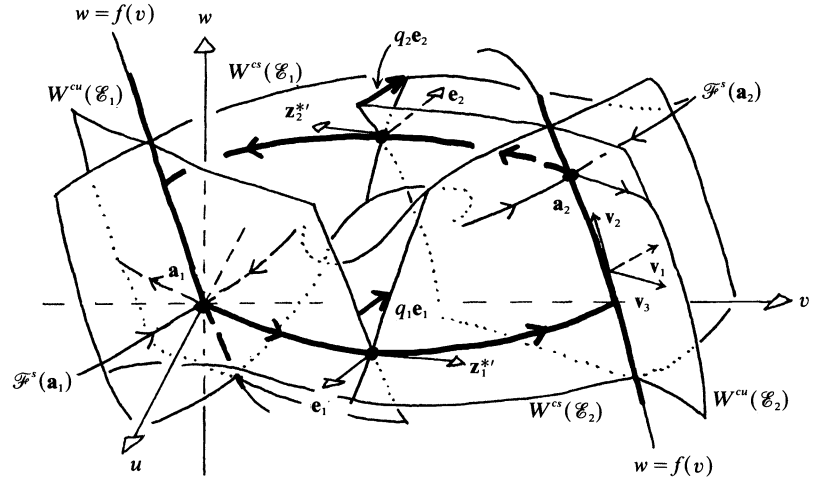


FIG. 5.1. *The strange loop at $\varepsilon = 0$. The two vertical heavy curves are $\mathscr{E}_1$ and $\mathscr{E}_2$, respectively. At the critical parameter $(\gamma, c) = (\gamma_0, c_0)$, $\mathbf{z}_1^*$ lies on the plane $\{w = 0\}$ while $\mathbf{z}_2^*$ lies on the plane $\{w = w_2\}$, connecting $\mathscr{E}_1$ and $\mathscr{E}_2$. At the appropriately perturbed point $(\gamma, c) = (\gamma(\varepsilon), c(\varepsilon))$ with $\varepsilon > 0$, $\mathbf{z}_1^*$ becomes a connection from $\mathbf{a}_1$ to $\mathbf{a}_2$, and it connects opposite sides of $W^{cs}(\mathscr{E}_1)$, which is the stable manifold of $\mathbf{a}_1$ after the perturbation $\varepsilon > 0$. Similarly, $\mathbf{z}_2^*$ connects opposite sides of $W^{cs}(\mathscr{E}_2)$ when $(\gamma, c) = (\gamma(\varepsilon), c(\varepsilon))$ and $\varepsilon > 0$.*

The existence of the desired connections when $\varepsilon = 0$ is given in the Appendix. The useful properties are summarized as follows. There is a connection $\mathbf{z}_1^* = (v_1^*, u_1^*, 0)$ from $\mathbf{a}_1 = 0$ to the intersection point $\{w = 0\} \cap \mathscr{E}_2 = (1, 0, 0)$ for (5.1a) when $c = c_0 = (1-2a)/\sqrt{2}$ for all γ and $v_1^* > 0$, $u_1^* > 0$. For the same c_0, when $\gamma = \gamma_0 = 9/(2-1)(1-2a)$ there is a connection $\mathbf{z}_2^* = (v_2^*, u_2^*, w_2)$ from $\mathbf{a}_2 = (v_2, 0, w_2)$ back to the intersection point $\{w = w_2\} \cap \mathscr{E}_1$ for (5.1a) with $u_2^* < 0$, where $v_2 = \gamma_0 w_2$, $w_2 = f(v_2)$ and (γ_0, c_0) is the same as in Theorem 1.1(d).

Next, let us show that these two connections can be continued for parameters near $(\varepsilon, \gamma, c) = (0, \gamma_0, c_0)$ via the Melnikov method. We consider this question for the $\mathbf{z}_1^*$ connection first. Let $\mathbf{z}_0^*$ be an arbitrarily chosen point from $\mathbf{z}_1^*$, and let Σ be the corresponding plane perpendicular to $\mathbf{z}_1^*$ at $\mathbf{z}_0^*$. Without loss of generality, let $\mathbf{z}_1^*(0) = \mathbf{z}_0^*$ and $\mathbf{z}_1^*(\tau) = (v_1^*, u_1^*, 0)(\tau)$ up to time translation along the solution. Since w in (5.1a) can be regarded as a parameter, the center-stable manifold $W_0^{cs}(\mathscr{E}_2) \cap \Sigma$ of $\mathscr{E}_2$ on Σ can be parametrized by w. Let $\mathbf{e}$ be the vector $(-u_1^{*\prime}(0), v_1^{*\prime}(0), 0)$ (which is labeled as $\mathbf{e}_1$ in Fig. 5.1) on $\{w = 0\}$ as discussed in the last section for vector field in $\mathbb{R}^2$; then $\mathbf{e}$ must be transverse to $W_0^{cs}(\mathscr{E}_2)$ near $\mathbf{z}_0^*$ since it points forwards (cf. Fig. 5.1). Let $Q_1(\varepsilon, \gamma, c)$ be the corresponding Melnikov function for $W^u(\mathbf{a}_1, \varepsilon)$ and $W_\varepsilon^{cs}(\mathscr{E}_2)$ (we will see later in § 7 that the direction $\mathbf{e}$ is indeed consistent with the requirement of § 3 that it points to the same side of $W^s(\mathbf{a}_2)$ as $\mathbf{e}_1^+$ does). On the other hand, as mentioned earlier, since w can be viewed as a parameter for the two-dimensional system

$$v' = u, \quad u' = cu - f(v) + w, \tag{5.2}$$

we can define a Melnikov function $Q(c, w)$ for the connection (v_1^*, u_1^*) at the parameters $w = 0$ and $c = c_0$ with respect to the corresponding straight line $\Sigma \cap \{w = 0\}$ and the same direction $\mathbf{e}$ on $\{w = 0\}$. It is easy to see that $(\partial/\partial c) Q_1(0, \gamma_0, c_0) = (\partial/\partial c) Q(c_0, 0)$. Thus, by (3.2) and the Melnikov integral (3.5) we have, up to a positive constant multiple,

$$\begin{aligned} \frac{\partial Q_1}{\partial c}(0, \gamma_0, c_0) &= -\int_{-\infty}^{\infty} \exp(-c\tau)(-u_1^{*\prime}(\tau), v_1^{*\prime}(\tau)) \cdot (0, u_1^*(\tau))\, d\tau \\ &= -\int_{-\infty}^{\infty} \exp(-c\tau) u_1^*(\tau)^2\, d\tau < 0. \end{aligned} \tag{5.3a}$$

Therefore, by the implicit function theorem, the solutions of $Q_1 = 0$ can be expressed as a function of $c = c_{1,0}(\gamma, \varepsilon)$, or $c = c_{1,0}(\gamma)$ for short, near $(0, \gamma_0, c_0)$. Moreover, an identical argument yields

$$\frac{\partial Q_1}{\partial \gamma}(0, \gamma_0, c_0) = 0. \tag{5.3b}$$

Therefore,

$$\frac{\partial c_{1,0}(\gamma_0, 0)}{\partial \gamma} = 0 \quad \text{or} \quad c_{1,0}'(\gamma_0) = 0 \tag{5.3c}$$

by the implicit function theorem. Similarly, define the other Melnikov function $Q_2(\varepsilon, \gamma, c)$ for $W^u(\mathbf{a}_2, \varepsilon)$ and $W_\varepsilon^{cs}(\mathscr{E}_1)$ with the same kind of choices of Σ and $\mathbf{e}$ except that $(v_1^*, u_1^*, 0)$ above is replaced by (v_2^*, u_2^*, w_2), where w_2 is a constant satisfying $v_2 = \gamma_0 w_2$, $w_2 = f(v_2)$. By the same argument,

$$\frac{\partial Q_2}{\partial c}(0, \gamma_0, c_0) = -\int_{-\infty}^{\infty} \exp(-c\tau) u_2^*(\tau)^2\, d\tau < 0. \tag{5.4a}$$

The only difference is that $\partial Q_2/\partial\gamma\ (0, \gamma_0, c_0) = \partial Q/\partial w\ (c_0, w_2) \cdot \partial w_2/\partial\gamma < 0$. Indeed, since w_2 is strictly decreasing in γ (cf. Fig. 1.1) and, by the Melnikov integral (3.5),

$$\frac{\partial Q}{\partial w}(c_0, w_2) = -\int_{-\infty}^{\infty} \exp(-c\tau)(-u_2^{*\prime}(\tau), v_2^{*\prime}(\tau)) \cdot (0, 1)\, d\tau$$

$$= -\int_{-\infty}^{\infty} \exp(-c\tau) u_2^*(\tau)\, d\tau > 0$$

for $u_2^* < 0$, we have

$$\frac{\partial Q_2}{\partial\gamma}(0, \gamma_0, c_0) < 0. \tag{5.4b}$$

Thus, by the implicit function theorem the solution function $c = c_{2,0}(\gamma, \varepsilon)$, or $c_{2,0}(\gamma)$, of $Q_2 = 0$ satisfies

$$\frac{\partial c_{2,0}(\gamma_0, 0)}{\partial\gamma} < 0 \quad \text{or} \quad c'_{2,0}(\gamma_0) < 0. \tag{5.4c}$$

This and (5.3c) show that $c_{1,0}$ and $c_{2,0}$ intersect transversely near (γ_0, c_0) for small ε. More precisely, the $c_{2,0}$ curve crosses the $c_{1,0}$ curve transversely from above into below as γ increases through γ_0. Let the intersection be $(\gamma(\varepsilon), c(\varepsilon))$. Then at $(\gamma(\varepsilon), c(\varepsilon))$ with $\varepsilon > 0$ there exists a heteroclinic loop connecting $\mathbf{a}_1$ and $\mathbf{a}_2$.

Next, to show the heteroclinic orbit $\mathbf{z}_1^*$ of (5.1a) converges to $\mathbf{a}_2$ along the principal stable eigenvector of $\mathbf{a}_2$ we only need to show that $\mathbf{z}_1^*$ is not contained in the strong stable manifold $W^{ss}(\mathbf{a}_2, \varepsilon)$ of $\mathbf{a}_2$ which is the stable fiber, $\mathscr{F}_\varepsilon^s$, through $\mathbf{a}_2$ on the center-stable manifold of $\mathscr{E}_2$. Since in limit $\varepsilon = 0$ $\mathscr{F}_\varepsilon^s$ lies on $\{w = w_2\}$ while $\mathbf{z}_1^*$ lies on $\{w = 0\}$ and $w_2 > 0$ because $\gamma_0 = 9/(2-a)(1-2a) \neq 0$ for $0 < a < \frac{1}{2}$ and $w_2 = f(v_2)$, $v_2 = \gamma_0 w_2 \neq 0$, $\mathscr{F}_\varepsilon^s$ still stays away from $\mathbf{z}_1^*$ uniformly for small $\varepsilon > 0$ by continuity. Similar arguments apply to the $\mathbf{z}_2^*$ connection.

Last, to show the strong inclination property (2.2b′) along $\mathbf{z}_1^*$ we show first that the center-stable and the center-unstable manifolds of (5.1a) intersect transversely along $\mathbf{z}_1^*$ at the limit $\varepsilon = 0$ and then we will relate this transverse intersection to the strong inclination property for $\varepsilon > 0$. Recall that $\mathscr{E}_1$ and $\mathscr{E}_2$ lie on $\{u = 0\}$ and can be parametrized by w so that when $w = 0$, $c = c_0$ there is the connection $\mathbf{z}_1^*$ from $\mathbf{a}_1$ to $\mathscr{E}_2 \cap \{w = 0\}$. Recall the definition of the differential function $q(\varepsilon, \gamma, c, w)$ along the forward pointing direction $\mathbf{e}$ from § 3, where $\mathbf{e}$ is the same as in the definition of the Melnikov function Q_1 above in this section (cf. Fig. 5.1). As mentioned earlier in § 3, we need to show

$$\frac{\partial q(0, \gamma, c_0, 0)}{\partial w} \neq 0$$

in order to prove the transverse intersection of $W^{cs}(\mathscr{E}_2)$ and $W^{cu}(\mathscr{E}_1)$ along $\mathbf{z}_1^*$. However, by treating (5.1a) as a two-dimensional system (5.2) again, it is easy to see that the differential q between W^{cs} and W^{cu} for the three-dimensional system (5.1a) at $\varepsilon = 0$ is precisely the Melnikov function $Q(c, w)$ of (5.2). Thus, $\partial q(0, \gamma, c_0, 0)/\partial w = \partial Q(c_0, 0)/\partial w$. Hence, by (3.2) and the Melnikov integral (3.5) we have

$$\frac{\partial q(0, \gamma, c_0, 0)}{\partial w} = -\int_{-\infty}^{\infty} \exp(-c\tau)(-u_1^{*\prime}(\tau), v_1^{*\prime}(\tau)) \cdot (0, 1)\, d\tau$$

$$= -\int_{-\infty}^{\infty} \exp(-c\tau) u_1^*(\tau)\, d\tau < 0 \tag{5.5a}$$

since $u_1^*(\tau) > 0$. Similarly, for the other connection $\mathbf{z}_2^*$ from $\mathbf{a}_2$ back to $\mathscr{E}_1$, we can show

$$\frac{\partial q(0, \gamma, c_0, w_2)}{\partial w} = -\int_{-\infty}^{\infty} \exp(-c\tau) u_2^*(\tau)\, d\tau > 0 \tag{5.5b}$$

since $u_2^*(\tau) < 0$. This shows that by definition $W^{cu}(\mathscr{E}_i)$ and $W^{cs}(\mathscr{E}_j)$ intersect transversely and the corresponding strong inclination property is satisfied by the strong λ-lemma of Deng (1989), (1990). By "strong inclination property" for nonhyperbolic system we mean that the same limit as (2.2b′) exists except that W_j^s, W_i^u, and W_i^{ss} in the formula are understood as $W^{cs}(\mathbf{a}_j)$, $\mathscr{F}_0^u(\mathbf{a}_i)$, and $\mathscr{F}_0^s(\mathbf{a}_i)$, respectively. Because the strong inclination property is generic, that is, it persists for those small perturbations of the vector field along which there exists the continuation of the connection $\mathbf{z}_i^*$, the strong inclination property (2.2b′) also holds true for sufficiently small $\varepsilon > 0$. This completes the proof for condition (2.2b).

Remark. The nonzeroness of the Melnikov integrals in (5.3a), (5.4a), (5.5a), and (5.5b) have also been derived by Lin (1989).

6. Proof of condition (2.2c). Recall that the connection $\mathbf{z}_1^*$ is twisted if the principal eigenvectors $\mathbf{e}_2^-$ and $\mathbf{e}_1^+$ along which the other connection $\mathbf{z}_2^*$ comes from $\mathbf{a}_2$ and goes to $\mathbf{a}_1$, respectively, point to opposite sides of the stable manifold $W^s(\mathbf{a}_2)$ of $\mathbf{a}_2$ along $\mathbf{z}_1^*$. Using the same strategy as above, let us examine the limiting case $\varepsilon = 0$ first. Note that in limit $\varepsilon = 0$, $\mathbf{e}_1^+$ is of the direction of the center eigenvector $\mathbf{v}_2$ of $\mathbf{a}_1$ and $\mathbf{e}_2^-$ is of the direction of the unstable tangent fiber $-\mathbf{v}_3$ of $\mathbf{a}_2$ (cf. Fig. 5.1). Thus, it suffices to show $\mathbf{v}_2$ of $\mathbf{a}_1$ and $-\mathbf{v}_3$ of $\mathbf{a}_2$ point to opposite sides of the center stable manifold $W^{cs}(\mathscr{E}_2)$ of $\mathscr{E}_2$. To show this, recall the differential function q from the previous section and the property (5.5a). $\partial q/\partial w < 0$ implies that $W^{cu}(\mathscr{E}_1)$ and $W^{cs}(\mathscr{E}_2)$ must split in such a way that when $w > 0$, $W^{cs}(\mathscr{E}_2)$ lies behind $W^{cu}(\mathscr{E}_1)$ near $\mathbf{z}_1^*$ (cf. Fig. 5.1). Thus, $\mathbf{v}_2$ of $\mathbf{a}_1$ and $\mathbf{v}_3$ of the intersection point $\mathscr{E}_2 \cap \{w = 0\}$ point to the same side of $W^{cs}(\mathscr{E}_2)$. Since $\mathbf{v}_3$ of $\mathscr{E}_2 \cap \{w = 0\}$ and $-\mathbf{v}_3$ of $\mathbf{a}_2$ point to opposite sides of $W^{cs}(\mathscr{E}_2)$, the desired result is proved. Similar arguments show that $\mathbf{z}_2^*$ is twisted. This proves condition (2.2c).

7. Proof of condition (2.2e). Recall from (5.3a, b) that when $(\varepsilon, \gamma, c) = (0, \gamma_0, c_0)$, $\partial Q_1/\partial c < 0$, $\partial Q_1/\partial \gamma = 0$. Also recall from (5.4a, b) that $\partial Q_2/\partial c < 0$, $\partial Q_2/\partial \gamma < 0$. It obviously follows that ∇Q_1 and ∇Q_2 are linearly independent for small $\varepsilon > 0$, where the gradient operator ∇ is taken with respect to the relevant parameters γ and c. Last, from the proof of the twist of $\mathbf{z}_i^*$ above, it is easy to see that the vectors $\mathbf{e}_i$ and $\mathbf{e}_i^+$ do point to the same side of $W^s(\mathbf{a}_j)$ (cf. Fig. 5.1). Hence, the Melnikov function Q_i, $i = 1$, 2, satisfies the specific requirement with respect to the direction $\mathbf{e}_i$ in § 3. This proves condition (2.2e).

Remark. Back in 1980, Langer proved that the stable and unstable manifolds cross transversely with the velocity parameter c for small $\varepsilon > 0$. This is also implied by the nonzeroness of $\partial Q_i/\partial c$, which also indicates the direction in which the transversal crossing takes place.

8. Proof of Theorem 1.1. As discussed in § 2, to prove Theorem 1.1 we only need to show that Theorem 2.1 is applicable to the reduced FitzHugh–Nagumo equation (2.1). We have shown in §§ 4–7 above that the hypotheses of Theorem 2.1 are all satisfied. Thus, it is only left to determine the bifurcation directions for the curves of the multiple pulse front and back waves by Theorem 2.1(c). Again, recall from (5.3a, b) and (5.4a, b), or from the previous section, that $\partial Q_1/\partial c < 0$, $\partial Q_1/\partial \gamma = 0$, and $\partial Q_2/\partial c < 0$,

$\partial Q_2/\partial\gamma < 0$ at $(\varepsilon, \gamma, c) = (0, \gamma_0, c_0)$. That is, ∇Q_1 and ∇Q_2 point downward at $(\gamma, c) = (\gamma(\varepsilon), c(\varepsilon))$ on the (γ, c)-plane for small $\varepsilon \geqq 0$. Therefore, all the interesting bifurcations take place in the southwest sector bounded by $Q_1 = 0$ and $Q_2 = 0$ in the (γ, c)-plane. This completes the proof. □

9. Proof of Theorem 1.2. To prove this theorem it suffices to show that the reduced traveling wave equations when restricted to the center manifold with respect to the singular parameter κ are "almost" the same as the FitzHugh-Nagumo equations (1.1). To begin with, recast (1.2) into the traveling wave equations with $\tau = x + ct$:

$$v' = u, \quad u' = cu - f(v) + w, \quad w' = x, \quad \kappa x' = cx - \varepsilon(v - \gamma w). \tag{9.1}$$

Treating κ as a singular parameter and writing this equation in terms of the fast time $t := \tau/\kappa$ variable, we have

$$\dot{v} = \kappa u, \quad \dot{u} = \kappa(cu - f(v) + w), \quad \dot{w} = \kappa x, \quad \dot{x} = cx - \varepsilon(v - \gamma w), \tag{9.2}$$

where the dot means the derivative in t. Note that when $\kappa = 0$ the equilibrium points of (9.2) consist of the entire three-dimensional manifold:

$$\mathscr{E}_0 := \left\{ x = \frac{\varepsilon}{c}(v - \gamma w) \right\}.$$

The linearization of (9.2) at $\mathscr{E}_0$ when $\kappa = 0$ is $\dot{V} = \dot{U} = \dot{W} = 0$ and $\dot{X} = cX$. It has only one nonzero eigenvalue $c > 0$ and the corresponding eigenvector $(0, 0, 0, 1)$ is normal to the manifold $\mathscr{E}_0$. For simplicity, let $\mathscr{E}_0$ also denote a sufficiently large, connected, and compact set in what follows. Therefore, in the context of Fenichel (1979) the invariant manifold $\mathscr{E}_0$ is normally hyperbolic. Also by Theorem 9.1 of Fenichel (1979) again there is a center manifold

$$\mathscr{E}_\kappa := \{x = h(v, u, w, \kappa)\}$$

in a neighborhood of $\mathscr{E}_0$ for all $|\kappa| \ll 1$, and it is a smooth continuation of $\mathscr{E}_0$, namely,

$$h(v, u, w, 0) = \frac{\varepsilon}{c}(v - \gamma w).$$

Here, the other parameters are suppressed from the expression of h for simplicity. It is very important to note from (9.2) that when $\varepsilon = 0$, $\{x = 0\}$ is always an invariant manifold regardless of the parameter κ. This implies that the function h can be chosen so that

$$h(v, u, w, \kappa) = \frac{\varepsilon}{c}(v - \gamma w) + O(\varepsilon\kappa). \tag{9.3}$$

According to the singular perturbation theory of Fenichel (1979) this manifold $\mathscr{E}_\kappa$ is invariant for both the slow and fast equations (9.1) and (9.2) for all $|\kappa| \ll 1$. Now, recasting (9.1) on the center manifold $\mathscr{E}_\kappa$ yields

$$v' = u, \quad u' = cu - f(v) + w, \quad w' = \frac{\varepsilon}{c}(v - \gamma w) + O(\varepsilon\kappa), \tag{9.4}$$

because of $w' = w = h(v, u, w, \kappa)$ and the estimate (9.3). Note that this is exactly the same FitzHugh-Nagumo equation (1.1) except for a perturbation term $O(\varepsilon\kappa)$ to the w equation. Now, it is easy to see that all the analysis for (2.1) in the previous sections applies to (9.4) as well. More specifically, ε is the singular parameter, γ and c are the relevant parameters. The additional parameter κ represents a trivial direction of

perturbation. That is to say, it will not change any qualitative structure of the system with respect to the heteroclinic loop bifurcation of Theorem 2.1, neither the singular perturbation structure in terms of ε nor the Melnikov method. This completes the proof. □

Appendix. The result presented below is taken from McKean (1970) and Casten, Cohen, and Lagerstrom (1975). Consider (5.1a). Since $w' = 0$, we can treat w as a parameter. For $f_{\min} < w_0 < f_{\max}$, there are three roots for $-f(v) + w_0 = 0$. Denote them by $\beta_1 < \beta_2 < \beta_3$ which implicitly depend on w_0 (cf. Fig. A). Then (5.1a) is equivalent to

$$u\frac{du}{dv} = cu + (v-\beta_1)(v-\beta_2)(v-\beta_3),$$

since $d\tau = dv/u$. Now, it is straightforward to check that $u = \lambda(v-\beta_1)(\beta_3 - v)$ with $\lambda = \pm 1/\sqrt{2}$, $c = \lambda(\beta_1 + \beta_3 - 2\beta_2)$ is a polynomial solution going through β_1 and β_3. In particular, when $w_0 = 0$, $\beta_1 = 0$, $\beta_2 = a$, $\beta_3 = 1$, and $\lambda = 1/\sqrt{2}$ we obtain a connection with a positive speed $c_{1,0} = (1-2a)/\sqrt{2}$. Denote the corresponding solution by (v_1^*, u_1^*); then $1 > v_1^* > 0$, $u_1^* > 0$. This implies that the connection is from $\mathbf{a}_1$ to $\mathscr{C}_2 \cap \{w = 0\}$ since $v_1^{*\prime} = u_1^* > 0$. On the other hand, choosing $\lambda = -1/\sqrt{2}$, we obtain another connecting orbit (v_2^*, u_2^*) for $c_{2,0} = \lambda(\beta_1 + \beta_3 - 2\beta_2)$. But, in this case $u_2^* < 0$ since $\beta_1 < v_2^* < \beta_3$; thus $v_2^{*\prime} < 0$. This implies that the connection is from $\mathscr{C}_2 \cap \{w = w_0\}$ back to $\mathscr{C}_1 \cap \{w = w_0\}$. Thus, it is only left to determine whether the corresponding speed c is positive and equal to $c_{1,0}$ at some γ in order to obtain a "loop." Because the cubic curve $w = f(v)$ is symmetric with respect to the inflection point $(v_{\inf}, f_{\inf}) := ((1+a)/3, (1+a)(1-2a)(2-1)/27)$ of $w = f(v)$, we can choose (v_2, w_2) from the curve $w = f(v)$ to be the point that is symmetric to the origin (0, 0) with respect to the inflection point. Because of this symmetry $c_{2,0} = -(\beta_1 + \beta_3 - 2\beta_2)/\sqrt{2} = (1-2a)/\sqrt{2} = c_{1,0}$ at $w_0 = w_2$. A direct computation yields $\gamma_0 := v_2/w_2 = 9/(2-a)(1-2a)$, which is the same as in Theorem 1.1(d). Certainly, for this parameter γ_0, the connection (v_2^*, u_2^*) is from $\mathbf{a}_2$ back to $\mathscr{C}_1 \cap \{w = w_2\}$.

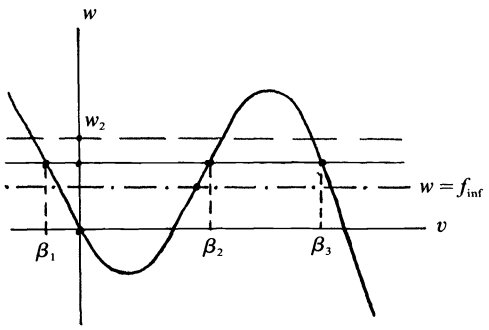


FIG. A. $\beta_1 < \beta_2 < \beta_3$ *are the roots of* $w_0 - f(v) = 0$.

REFERENCES

D. G. ARONSON AND H. F. WEINBERGER, *Nonlinear diffusion in population genetics, combustion and nerve propagation*, in Proc. Tulane Program in Partial Differential Equations and Related Topics, Lecture Notes in Math. 446, Springer-Verlag, Berlin, 1975, pp. 5-49.

G. A. CARPENTER, *A geometric approach to singular perturbation problems with applications to nerve impulse equations*, J. Differential Equations, 23 (1977), pp. 335-367.

R. H. CASTEN, H. COHEN, AND P. LAGERSTROM, *Perturbation analysis of an approximation to Hodgkin-Huxley theory*, Quart. Appl. Math., 32 (1975), pp. 365-402.

S.-N. CHOW, B. DENG, AND B. FIEDLER, *Homoclinic bifurcation at resonant eigenvalues*, J. Dynamical Systems and Differential Equations, 2 (1990), pp. 177-244.

S.-N. CHOW, B. DENG, AND D. TERMAN, *The bifurcation of a homoclinic and periodic orbits from two heteroclinic orbits*, SIAM J. Math. Anal., 21 (1990), pp. 179-204.

C. CONLEY, *On traveling wave solutions of nonlinear diffusion equations*, Tech. Report 1492, Mathematics Research Center, University of Wisconsin, Madison, WI, 1975.

B. DENG, *The bifurcations of countable connections from a twisted heteroclinic loop*, SIAM J. Math. Anal., 22 (1991), pp. 653-679.

———, *Homoclinic bifurcations with nonhyperbolic equilibria*, SIAM J. Math. Anal., 21 (1990), pp. 693-720.

———, *The Sil'nikov problem, exponential expansion, strong λ-lemma, C^1-linearization and homoclinic bifurcation*, J. Differential Equations, 79 (1989), pp. 189-231.

J. W. EVANS, N. FENICHEL, AND J. A. FEROE, *Double impulse solutions in nerve axon equations*, SIAM J. Appl. Math., 42 (1982), pp. 219-234.

N. FENICHEL, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53-98.

J. A. FEROE, *Existence and stability of multiple impulse solutions of a nerve equation*, SIAM J. Appl. Math., 42 (1982), pp. 219-234.

R. FITZHUGH, *Impulses and physiological states in theoretical models of nerve membrane*, Biophys. J., 1 (1961), pp. 445-466.

J. M. GREENBERG, *A note on Nagumo equation*, Quart. J. Math. (Oxford), 24 (1973), pp. 307-314.

J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

S. P. HASTINGS, *The existence of periodic solutions to Nagumo's equation*, Quart. J. Math., 25 (1974), pp. 368-378.

———, *On the existence of homoclinic and periodic orbits for the FitzHugh-Nagumo equations*, Quart. J. Math., 27 (1976), pp. 123-134.

———, *Single and multiple pulse waves for the FitzHugh-Nagumo equations*, SIAM J. Appl. Math., 42 (1982), pp. 247-260.

A. L. HODGKIN AND A. F. HUXLEY, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol. (London), 117, 500 (1952).

P. HOLMES, *Averaging and chaotic motions in forced oscillations*, SIAM J. Appl. Math., 38 (1980), pp. 65-80.

H. KOKUBU, Y. NISHIURA, AND H. OKA, *Heteroclinic and homoclinic bifurcation in bistable reaction diffusion systems*, preprint KSU/ICS 88-08, 1988.

R. LANGER, *Existence of homoclinic travelling wave solutions to the FitzHugh-Nagumo equations*, Ph.D. thesis, Northeastern University, 1980.

X.-B. LIN, *Heteroclinic bifurcation and singular perturbed boundary value problems*, preprint, 1989.

H. P. MCKEAN, JR., *Nagumo's equation*, Adv. in Math., 4 (1970), pp. 209-223.

V. K. MELNIKOV, *On the stability of the center for time periodic perturbations*, Trans. Moscow Math. Soc., 12 (1964), pp. 1-57.

J. S. NAGUMO, S. ARIMOTO, AND S. YOSHIZAWA, *An active pulse transmission line simulating nerve axon*, Proc. IRE, 50 (1963), pp. 2061-2070.

P. ORTOLEVA AND J. ROSS, *Theory of propagation of discontinuities in kinetic systems with multiple time scales: fronts, front multiplicity, and pulses*, J. Chem. Phys., 63 (1975), pp. 3398-3408.

K. J. PALMER, *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55 (1984), pp. 225-265.

J. P. PAUWELUSSEN, *Heteroclinic waves of the FitzHugh-Nagumo equations*, preprint, 1980.

J. RAUCH AND J. SMOLLER, *Qualitative theory of FitzHugh-Nagumo equations*, Adv. in Math., 27 (1978), pp. 12-44.

J. RINZEL AND J. B. KELLER, *Traveling wave solutions of a nerve conduction equation*, Biophys. J., 12 (1973), pp. 1313-1337.

J. RINZEL AND D. TERMAN, *Propagation phenomena in a bistable reaction-diffusion system*, SIAM J. Appl. Math., 42 (1982), pp. 1111-1137.

L. P. SIL'NIKOV, *A contribution to the problem of the structure of an extended neighborhood of a rough equilibrium state of saddle-focus type*, Math. USSR Sb., 10 (1970), pp. 91-102.

W.-P. WANG, *Multiple impulse solutions to McKean's caricature of the nerve equation. I—Existence*, Comm. Pure. Appl. Math., 41 (1988), pp. 71-103.

W.-P. WANG, *Multiple impulse solutions to McKean's caricature of the nerve equation.* II—*Stability*, Comm. Pure. Appl. Math., 41 (1988), pp. 997–1025.

H. R. WILSON AND J. D. COWAN, *Excitatory and inhibitory interactions in localized populations of model neurons*, Biophys. J., 12 (1972), pp. 1–24.

E. YANAGIDA, *Branching of double pulse solutions from single pulse solutions in nerve axon equations*, J. Differential Equations, 66 (1987), pp. 243–262.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1651-1678, November 1991

008

STABILITY ANALYSIS OF STATIONARY SOLUTIONS OF BISTABLE REACTION-VARIABLE DIFFUSION SYSTEMS*

HIDEO IKEDA† AND MASAYASU MIMURA‡

Abstract. From a wave-blocking viewpoint, a bistable reaction-variable diffusion system is considered:

$$u_t = (D(z)u_z)_z + \frac{1}{\sigma} f(u, v),$$

$$z \in \mathbf{R}, \quad t > 0.$$

$$v_t = (D(z)v_z)_z + \sigma g(u, v),$$

Depending on σ and $D(z)$, there exist stationary solutions that connect two stable states at $z = \pm\infty$. By using the linearized stability criterion, the role of these solutions in wave-blocking phenomena is discussed.

Key words. stability, variable diffusion, singular perturbation, wave-blocking

AMS(MOS) subject classifications. 35B25, 35B40, 35K57

1. Introduction.

In this paper, we consider the following reaction-diffusion system:

$$u_t = (D(z)u_z)_z + \frac{1}{\sigma} f(u, v),$$

$$z \in \mathbf{R}, \quad t > 0, \tag{1.1}$$

$$v_t = (D(z)v_z)_z + \sigma g(u, v),$$

where $D(z)$ is the *variable* diffusion rates of u and v. For the nonlinearities f and g, we assume that there exist three constant steady states as in Fig. 1: Two of them $(u_\pm, v_\pm)$ are stable, while the other (u_0, v_0) is unstable. $1/\sigma^2$ is the ratio of the reaction rates of u and v. If σ is small, u reacts faster than does v. Conversely, if σ is large, the situation is vice versa. In chemical terms, (1.1) describes the situation where two ion components u and v diffuse in one-dimensional heterogeneous medium and react with each other in bistable dynamics.

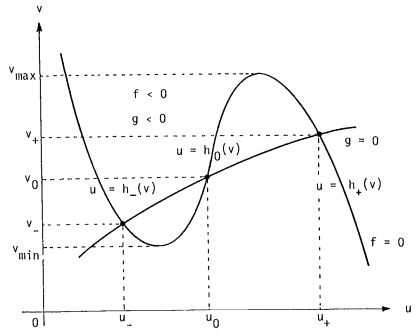


FIG. 1 *Functional forms of* $f = 0$ *and* $g = 0$.

* Received by the editors January 8, 1990; accepted for publication December 12, 1990.
† Department of Mathematics, Toyama University, Toyama 930, Japan.
‡ Department of Mathematics, Hiroshima University, Hiroshima 730, Japan.

In studying the influence of variable diffusion on qualitative properties of solutions, there are many works on scalar equations. The main concerns are with the existence, stability, and bifurcation properties of *nonconstant* stationary solutions. Along these lines, [M], [Y], [CH], [FH], and [FP] treat these problems in a finite interval or on the whole line. From an application viewpoint, in [P] and [K] wave-blocking phenomena due to the effect of a change in geometry of a nerve axon on the propagation of potential waves are studied.

Here we are interested in these problems for a system of equations (1.1), when the variable diffusion rate $D(z)=d(z/\varepsilon)$, where $d(\xi)$ is a strictly monotone $C^1(\mathbf{R})$ function satisfying

$$d(\xi)=\begin{cases} d & \text{as } \xi\to-\infty, \\ 1 & \text{as } \xi\to+\infty \end{cases} \tag{1.2}$$

and ε is a small positive constant. That is, this situation indicates that the diffusion rate $D(z)$ has an abrupt change at $z=0$. Let

$$x=\tau(z)\equiv\int_0^z ds/d\left(\frac{s}{\varepsilon}\right).$$

Clearly, τ maps $\mathbf{R}$ into $\mathbf{R}$ and its inverse τ^{-1} is well defined. Introducing x as the independent variable into (1.1), we have

$$\begin{aligned} u_t &= \delta_\varepsilon(x)u_{xx}+\frac{1}{\sigma}f(u,v), \\ & \qquad\qquad x\in\mathbf{R}, \quad t>0, \\ v_t &= \delta_\varepsilon(x)v_{xx}+\sigma g(u,v), \end{aligned} \tag{1.3}$$

where $\delta_\varepsilon(x)=1/d(\tau^{-1}(x)/\varepsilon)$. Set $\varepsilon=0$. Then (1.3) reduces to the following limiting system:

$$\begin{aligned} u_t &= \delta_0(x)u_{xx}+\frac{1}{\sigma}f(u,v), \\ & \qquad\qquad x\in\mathbf{R}, \quad t>0, \\ v_t &= \delta_0(x)v_{xx}+\sigma g(u,v), \end{aligned} \tag{1.4}$$

where

$$\delta_0(x)=\begin{cases} 1/d, & x<0, \\ 1, & x>0. \end{cases}$$

When $\varepsilon>0$ is sufficiently small, perturbation techniques suggest that (1.4) becomes a nice approximation to (1.3) in some sense. So we study the above limiting system (1.4).

When $d=1$ or $\delta_0(x)=1$ (homogeneous medium), (1.4) reduces to a usual reaction-diffusion system

$$\begin{aligned} u_t &= u_{xx}+\frac{1}{\sigma}f(u,v), \\ & \qquad\qquad x\in\mathbf{R}, \quad t>0. \\ v_t &= v_{xx}+\sigma g(u,v), \end{aligned} \tag{1.5}$$

For (1.5), we could expect that there are traveling wave solutions connecting one stable state (u_-, v_-) at $x=-\infty$ and the other (u_+, v_+) at $x=+\infty$. In fact, when f and g are specified as

$$\begin{aligned} f(u, v) &= u(1-u)(u-a)-v, \\ g(u, v) &= u-\gamma v, \end{aligned} \tag{1.6}$$

there are three constant steady states $(0, 0)$, $(\bar{u}_+, \bar{u}_+/\gamma)$, and $(\bar{u}_-, \bar{u}_-/\gamma)$ for some values of a and γ, where $\bar{u}_-$ and $\bar{u}_+$ $(\bar{u}_- < \bar{u}_+)$ are two solutions of $(1-u)(u-a)=1/\gamma$. Numerical simulations show that two different types of traveling wave solutions connecting $(0, 0)$ at $x=-\infty$ and $(\bar{u}_+, \bar{u}_+/\gamma)$ at $x=+\infty$ coexist: one is a *front* wave that propagates to the left direction, and the other is a *back* wave that goes in the opposite direction (see Fig. 2). We state briefly the results obtained in [IM2]. Suppose that f and g take the form (1.6) for simplicity and γ_0, γ_1, and γ_2 are the critical values as shown in Fig. 4. The qualitative property of traveling waves depends on the value of σ. When σ is sufficiently large, there is a unique stable traveling wave solution for fixed $\gamma > \gamma_0$ as in Fig. 5(i). On the other hand, when σ is sufficiently small, there are three solutions for fixed $\gamma \in (\gamma_0, \gamma_2)$ as in Fig. 5(ii); two of them are stable, whereas the other is unstable.

What happens in the case when $d \neq 1$ (heterogeneous medium)? As in Fig. 3, numerical simulations show three typical behaviours when σ is sufficiently small, where a, σ, and γ are the same values as in Fig. 2: (i) When d is close to 1, we could expect that the front and back waves pass through the point of discontinuity $x=0$ and propagate to the left and right directions, respectively (see Fig. 3(i)). (ii) When d is

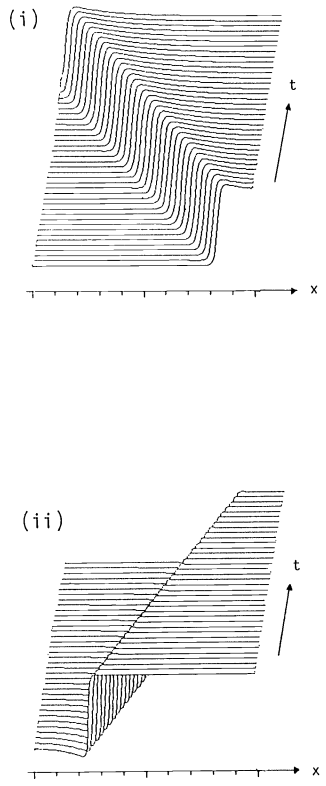


FIG. 2. (i) *Traveling front wave.* (ii) *Traveling back wave* ($a=0.25$, $\sigma=0.1$ *and* $\gamma=10.5$).

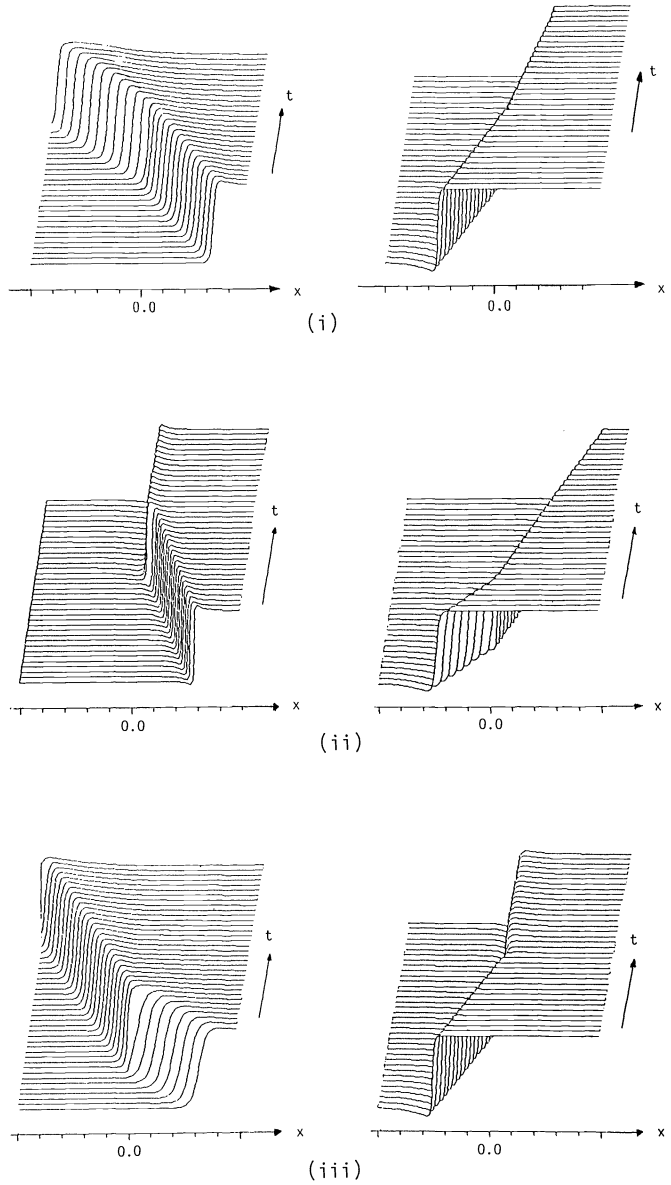


FIG. 3. *Numerical computations for wave-blocking phenomena*: (i) *When d is close to* 1, *the front and back waves pass through*; (ii) *When d is large, the front wave is blocked, whereas the back wave passes through*; (iii) *When d is small, the front wave passes through, whereas the back wave is blocked.*

large, the back wave passes through the point $x = 0$ and propagates to the right direction, while the front wave is blocked at $x = 0$ (see Fig. 3 (ii)). (iii) When d is small, the situation is opposite to that in (ii) (see Fig. 3(iii)). The last two cases (ii) and (iii) indicate that the traveling waves are blocked and there appear *new* stationary solutions. These evidences clearly show that such wave-blocking phenomena are caused by the heterogeneity of $\delta_0(x)$.

For the problem of wave-blocking phenomena in a heterogeneous medium, the dependency on d of stationary solutions is considered in [IM1] when σ is sufficiently

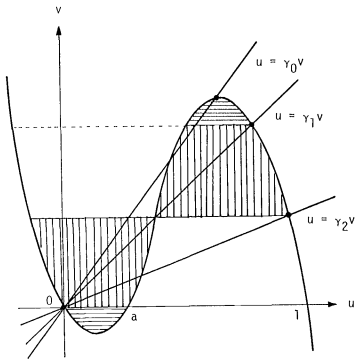


FIG. 4. *Critical values of* γ.

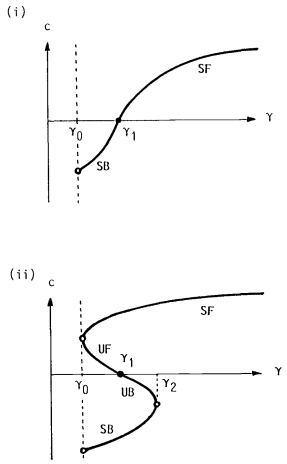


FIG. 5. *The global diagrams of traveling waves with respect to* γ. *The symbols SF, SB, UF, and UB stand for a stable front branch, a stable back branch, an unstable front branch, and an unstable back branch, respectively*: (i) σ *is sufficiently large*; (ii) σ *is sufficiently small.*

large or small. Suppose that σ is sufficiently large. If γ is fixed to satisfy $\gamma \in (\gamma_1, \infty)$, there exist two stationary solutions for large d (see Fig. 6(i)) and if γ is fixed to satisfy $\gamma \in (\gamma_0, \gamma_1)$, there exist two solutions for small d (see Fig. 6(ii)). On the other hand, if σ is sufficiently small, then the situation is different from the above. If γ is fixed to satisfy $\gamma \in (\gamma_2, \infty)$, then there exist two stationary solutions for large d (see Fig. 7(i)). However, if γ is fixed to satisfy $\gamma \in (\gamma_0, \gamma_2)$, there exist two solutions for large as well as small d (see Figs. 7(ii) and 7(iii)). This evidence shows that the qualitative property of stationary solutions depends on the value of σ as well as d.

Integrating these and using complementary numerical results, we arrive at the following conjecture. For simplicity, we consider the case when σ is sufficiently small and γ satisfies $\gamma \in (\gamma_1, \gamma_2)$. Under this situation, there are one front wave and one back wave, which are stable, in a homogeneous medium. For large fixed d, the front wave is blocked at $x = 0$ and one of the stationary solutions $(\underline{U}^*, \underline{V}^*)$ acts as its *barrier*, while the other $(\bar{U}^*, \bar{V}^*)$ acts as a *separator* between the barrier and the traveling like front wave (in a heterogeneous medium) as in Fig. 3(ii), whereas the back wave passes through. On the other hand, for small fixed d, the back wave is blocked and one of the stationary solutions $(\bar{U}_*, \bar{V}_*)$ acts as its barrier, while the other $(\underline{U}_*, \underline{V}_*)$ acts as

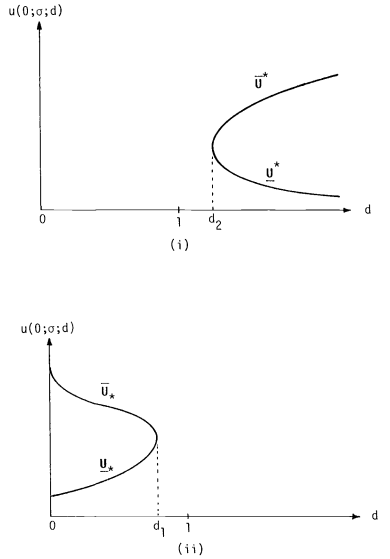


FIG. 6. *The schematic global diagrams of stationary solutions of* (1.1) *with* (1.4) *when* σ *is sufficiently large*: (i) $\gamma > \gamma_1$; (ii) $\gamma_0 < \gamma < \gamma_1$.

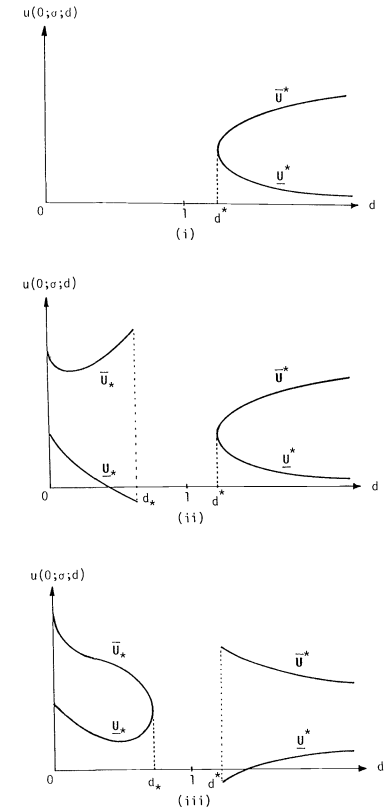


FIG. 7. *The schematic global diagrams of stationary solutions of* (1.1) *with* (1.4) *when* σ *is sufficiently small*: (i) $\gamma > \gamma_2$; (ii) $\gamma_1 < \gamma < \gamma_2$; (iii) $\gamma_0 < \gamma < \gamma_1$.

a separator between the barrier and the traveling like back wave as in Fig. 3(iii), whereas the front wave passes through. For intermediate fixed d, there are no stationary solutions and two waves pass through the discontinuous point $x=0$ as in Fig. 3(i).

Motivated by this evidence, we study the stability of the stationary solutions constructed in [IM1]. When σ is sufficiently small and γ satisfies $\gamma\in(\gamma_1, \gamma_2)$, for instance, the result is stated as follows. For large d, the $(\underline{U}^*, \underline{V}^*)$-branch is stable and the $(\bar{U}^*, \bar{V}^*)$-branch is unstable, while for small d, the $(\bar{U}_*, \bar{V}_*)$-branch is stable and the $(\underline{U}_*, \underline{V}_*)$-branch is unstable (see Fig. 7(ii)). The details will be discussed in § 3.

Since $\delta_0(x)$ is discontinuous at $x=0$, let us define weak solutions $(u, v)(t, x)$ of (1.4) by (u, v), which satisfies the following equations:

$$\begin{aligned} &u_t=\delta_0(x)u_{xx}+\frac{1}{\sigma}f(u, v),\\ &\qquad\qquad x\in\mathbf{R}\backslash\{0\},\quad t>0 \qquad (1.7)_a\\ &v_t=\delta_0(x)v_{xx}+\sigma g(u, v), \end{aligned}$$

with boundary conditions

$$(u, v)(t, \pm\infty)=(u_\pm, v_\pm) \qquad (1.7)_b$$

and compatibility conditions

$$\begin{aligned} &(u, v)(t, -0)=(u, v)(t, +0),\\ &\qquad\qquad t>0. \qquad (1.7)_c\\ &(u_x, v_x)(t, -0)=(u_x, v_x)(t, +0), \end{aligned}$$

We now study the spectral analysis for the linearized equations of $(1.7)_a$ around the specified stationary solution obtained in [IM1]. The spectrum of the linearized operator consists of essential spectrum and isolated eigenvalues. Since the essential spectrum is strictly bounded away from the imaginary axis to the left (see the Appendix), it is sufficient for the stability argument to study the distribution of isolated eigenvalues only. Our assertion is that there is only one eigenvalue that essentially determines the stability and its sign corresponds, in a one-to-one manner, to that of the Jacobian of the matching condition which is defined in constructing solutions by use of singular perturbation methods (see (2.14)). Here we discuss only the case where σ is sufficiently small. The case where σ is sufficiently large is briefly stated in the final section.

We first state the following assumptions on the nonlinearities f and g (see Fig. 1):

(A1) $f=0$ is S-shaped and consists of three branches $u=h_-(v)$, $h_0(v)$, and $h_+(v)$ $(h_-(v)\leqq h_0(v)\leqq h_+(v))$, and $g=0$ intersects each branch once. That is, there is only one intersection point on each branch $u=h_-(v)$ (respectively, $u=h_0(v)$ or $u=h_+(v)$), which is denoted by (u_-, v_-) (respectively, (u_0, v_0) or $(u_+, v_+))(v_-<v_0<v_+)$. The signs of f and g are both negative in the region above the curves $f=0$ and $g=0$.

(A2) $\mathcal{I}(v)=\int_{h_-(v)}^{h_+(v)}f(u, v)\,du$ has a unique isolated zero at $v^*\in(v_{\min}, v_{\max})$.

(A3) $f_u(h_\pm(v), v)<0$ for $v\in[v_-, v_+]$, $g(h_-(v), v)<0<g(h_+(v), v)$ for $v\in(v_-, v_+)$, and $g_u(h_\pm(v), v)h'_\pm(v)+g_v(h_\pm(v), v)<0$ at $v=v_\pm$.

(A4) $f_v(u, v)<0$ for $(u, v)\in\{(u, v)\,|\,h_-(v)\leqq u\leqq h_+(v),\ v_-\leqq v\leqq v_+\}$, $g_u(u, v)>0$ and $g_v(u, v)<0$ at $(u, v)=(u_\pm, v_\pm)$.

(A5) $g_v(u, v)<0$ for $(u, v)\in\{(u, v)\,|\,g(u, v)=0,\ v_-\leqq v\leqq v_+\}$.

Throughout this paper, we shall use the following function spaces. For $I=\mathbf{R}_-$, $\mathbf{R}_+$, or $\mathbf{R}$, positive numbers σ and μ, and an integer n, let

$$X^n_{\mu,\sigma}(I)=\left\{u\in C^n(I)\,\middle|\,\|u\|_{X^n_{\mu,\sigma}(I)}\equiv\sum_{i=0}^{n}\sup_{x\in I}\left|e^{\mu|x|}\left(\sigma\frac{d}{dx}\right)^i u(x)\right|<+\infty\right\},$$

$$\mathring{X}^n_{\mu,\sigma}(I)=\{u\in X^n_{\mu,\sigma}(I)\,|\,u(0)=0\},$$

$$BC^n(I)=\{\text{the set of functions all of whose derivatives of order}\leqq n\text{ are bounded and uniformly continuous on }I\}.$$

2. Construction of stationary solutions. We assume that σ is sufficiently small. First, we shall summarize the existence result for stationary solutions of (1.7) which is stated in [IM1]. Putting $y=\sqrt{\sigma}\,x$, we rewrite the stationary problem of (1.7) as

$$\begin{aligned}&(u,v)\in\mathcal{D}\times\mathcal{D},\\&\sigma^2\delta_0(y)u_{yy}+f(u,v)=0,\\&\qquad\qquad\qquad\qquad y\in\mathbf{R}/\{0\},\\&\delta_0(y)v_{yy}+g(u,v)=0,\\&u(\pm\infty)=u_\pm,\qquad v(\pm\infty)=v_\pm,\end{aligned}\tag{2.1}$$

where $\mathcal{D}\equiv\{u\in BC^1(\mathbf{R})\cap BC^2((-\infty,0])\cap BC^2([0,\infty))\,|\,\delta_0(y)u_{yy}\in BC(\mathbf{R})\}$. Let us seek a solution of (2.1) $(u,v)(y;\sigma;d)\in\mathcal{D}\times\mathcal{D}$ under (A1)-(A4). Since $\delta_0(y)(d\neq 1)$ has a discontinuity at $y=0$, we separate the problem (2.1) into the following two parts:

$$\begin{aligned}&\sigma^2\delta_0(y)u^\pm_{yy}+f(u^\pm,v^\pm)=0,\\&\qquad\qquad\qquad\qquad y\in\mathbf{R}_\pm,\\&\delta_0(y)v^\pm_{yy}+g(u^\pm,v^\pm)=0,\\&u^\pm(\pm\infty)=u_\pm,\qquad u^\pm(0)=\alpha,\\&v^\pm(\pm\infty)=v_\pm,\qquad v^\pm(0)=\beta,\end{aligned}\tag{2.2}_\pm$$

where α and β are constants that will be specified later. Using classical singular perturbation techniques, we construct solutions of $(2.2)_\pm$ in each subinterval $\mathbf{R}_\pm$, and then determine α and β so as to match these solutions in the C^1-sense at $y=0$. In § 2.1, we construct *outer solutions* that are approximate solutions to $(2.2)_\pm$ outside a layer region, and in §2.2, we construct *inner solutions* that are approximations to $(2.2)_\pm$ in a boundary layer region. In § 2.3, using these approximate solutions, we construct *singular limit solutions* as $\sigma\downarrow 0$, which become nice approximations to (2.1) uniformly on $\mathbf{R}$. Finally, in § 2.4, by using the singular limit solutions we prove the existence of solutions of (2.1) for a sufficiently small (but not zero) σ and draw the global picture of solution structures with respect to d. We emphasize that it is inherited from that of the singular limit solutions as $\sigma\downarrow 0$.

2.1. Outer solutions. We consider the reduced problems, by putting $\sigma=0$ in $(2.2)_\pm$. These are described by

$$\begin{aligned}&f(u^\pm,v^\pm)=0,\\&\qquad\qquad\qquad y\in\mathbf{R}_\pm,\\&\delta_0(y)v^\pm_{yy}+g(u^\pm,v^\pm)=0,\\&v^\pm(\pm\infty)=v_\pm,\qquad v^\pm(0)=\beta.\end{aligned}\tag{2.3}_\pm$$

Since $f(u^{\pm}, v^{\pm}) = 0$ are reduced to $u^{\pm} = h_{\pm}(v^{\pm})$ by the conditions $v^{\pm}(\pm\infty) = v_{\pm}$, $(2.3)_{\pm}$ can be rewritten as

$$(2.4)_{\pm} \qquad \begin{aligned} &\delta_0(y)v^{\pm}_{yy} + g(h_{\pm}(v^{\pm}), v^{\pm}) = 0, \qquad y \in \mathbf{R}_{\pm}, \\ &v^{\pm}(\pm\infty) = v_{\pm}, \qquad v^{\pm}(0) = \beta. \end{aligned}$$

Here β is arbitrarily fixed to satisfy $\beta \in (v_-, v_+)$.

LEMMA 2.1. *Equations* $(2.4)_{\pm}$ *have unique strictly monotone increasing solutions* $V_0^+(y; \beta)(y \in \mathbf{R}_+)$ *and* $V_0^-(y; \beta; d)(y \in \mathbf{R}_-)$ *satisfying*

$$|V_0^+(y; \beta) - v_+| \in X^2_{\mu_+,1}(\mathbf{R}_+), |V_0^-(y; \beta; d) - v_-| \in X^2_{\mu_-(d),1}(\mathbf{R}_-),$$

where $\mu_+ = \sqrt{-D_+(v_+)}$ *and* $\mu_-(d) = \sqrt{-dD_-(v_-)}$ *with* $D_{\pm}(v) = g_u(h_{\pm}(v), v)h'_{\pm}(v) + g_v(h_{\pm}(v), v)$. *Moreover,* $V_0^+(y; \beta)$ *(respectively,* $V_0^-(y; \beta; d)$*) is continuous with respect to* β *in the* $X^2_{\mu_+,1}(\mathbf{R}_+)$ *(respectively,* $X^2_{\mu_-(d),1}(\mathbf{R}_-)$*)-topology and satisfies*

$$\frac{d}{dy} V_0^+(0; \beta) = \sqrt{2 \int_{\beta}^{v_+} g(h_+(v), v)\, dv}$$

$$\left(\textit{resp.}, \frac{d}{dy} V_0^-(0; \beta; d) = \sqrt{2d \int_{\beta}^{v_-} g(h_-(v), v)\, dv} \right).$$

Using $V_0^+(y; \beta)$ and $V_0^-(y; \beta; d)$, we define $U_0^+(y; \beta)$ and $U_0^-(y; \beta; d)$ by

$$U_0^+(y; \beta) \equiv h_+(V_0^+(y; \beta)), \qquad y \in \mathbf{R}_+$$

and

$$U_0^-(y; \beta; d) \equiv h_-(V_0^-(y; \beta; d)), \qquad y \in \mathbf{R}_-,$$

respectively. Although $U_0^+(y; \beta)$ and $U_0^-(y; \beta; d)$ do not satisfy the boundary conditions at $y = 0$, we expect that when σ is small, $(U_0^{\pm}, V_0^{\pm})$ become nice approximations to solutions of $(2.2)_{\pm}$ outside a neighborhood of $y = 0$, which are called the *outer solutions* of $(2.2)_{\pm}$.

2.2. Inner solutions. We introduce $W_0^{\pm}(\xi)$ with the stretched variable $\xi = y/\sigma$ such that $(U_0^{\pm} + W_0^{\pm}, V_0^{\pm})$ become approximations to $(2.2)_{\pm}$ in a neighborhood of $y = 0$. Substituting $(U_0^{\pm} + W_0^{\pm}, V_0^{\pm})$ into $(2.2)_{\pm}$ and then setting $\sigma = 0$, we have the following problems with respect to $W_0^{\pm}$:

$$(2.5)_{\pm} \qquad \begin{aligned} &\delta_0(\xi)(W_0^{\pm})_{\xi\xi} + f(h_{\pm}(\beta) + W_0^{\pm}, \beta) = 0, \qquad \xi \in \mathbf{R}_{\pm}, \\ &W_0^{\pm}(\pm\infty) = 0, \qquad W_0^{\pm}(0) = \alpha - h_{\pm}(\beta). \end{aligned}$$

When $v^* < v_+$, $\omega_0(\beta)$ is uniquely determined for any $\beta \in [v^*, v_+]$ by $\int_{\omega_0(\beta)}^{h_+(\beta)} f(u, \beta)\, du = 0$. When $v_- < v^*$, $\omega_1(\beta)$ is uniquely determined for any $\beta \in [v_-, v^*]$ by $\int_{h_-(\beta)}^{\omega_1(\beta)} f(u, \beta)\, du = 0$. Using $\omega_0(\beta)$ and $\omega_1(\beta)$, we define Σ_1, Σ_2, and Σ_3 in (α, β)-space in $\mathbf{R}^2$ as follows. For $v_- < v^* < v_+$,

$$\Sigma_1 \equiv \{(\alpha, \beta) \in \mathbf{R}^2 \mid v^* \leqq \beta \leqq v_+, \omega_0(\beta) < \alpha < h_+(\beta) \text{ or } v_- \leqq \beta \leqq v^*, h_-(\beta) < \alpha < \omega_1(\beta)\};$$

for $v^* \leqq v_-$,

$$\Sigma_2 \equiv \{(\alpha, \beta) \in \mathbf{R}^2 \mid v_- \leqq \beta \leqq v_+, \omega_0(\beta) < \alpha < h_+(\beta)\};$$

for $v_+ \leqq v^*$,

$$\Sigma_3 \equiv \{(\alpha, \beta) \in \mathbf{R}^2 \mid v_- \leqq \beta \leqq v_+, h_-(\beta) < \alpha < \omega_1(\beta)\}.$$

LEMMA 2.2. *For any* $(\alpha, \beta) \in \Sigma_i$ $(i = 1, 2, 3)$, the $(2.5)_{\pm}$ *have unique strictly monotone increasing solutions* $W_0^+(\xi; \alpha, \beta)(\xi \in \mathbf{R}_+)$ *and* $W_0^-(\xi; \alpha, \beta; d)(\xi \in \mathbf{R}_-)$ *satisfying*

$$|W_0^+(\xi; \alpha, \beta)| \in X^2_{\tau_+(\beta),1}(\mathbf{R}_+), \qquad |W_0^-(\xi; \alpha, \beta; d)| \in X^2_{\tau_-(\beta; d),1}(\mathbf{R}_-),$$

where $\tau_+(\beta) = \sqrt{-f_u(h_+(\beta), \beta)}$ *and* $\tau_-(\beta; d) = \sqrt{-df_u(h_-(\beta), \beta)}$. *Moreover,* $W_0^+(\xi; \alpha, \beta)$ *(respectively,* $W_0^{\nabla}(\xi; \alpha, \beta; d)$*) is continuous with respect to* (α, β) *in the* $X^2_{\tau_+(\beta),1}(\mathbf{R}_+)$ *(respectively,* $X^2_{\tau_-(\beta; d),1}(\mathbf{R}_-)$*)-topology and satisfies*

$$\frac{d}{d\xi} W_0^+(0; \alpha, \beta) = \sqrt{2 \int_{\alpha}^{h_+(\beta)} f(u, \beta)\, du}$$

$$\left(\textit{resp.}, \frac{d}{d\xi} W_0^-(0; \alpha, \beta; d) = \sqrt{2d \int_{\alpha}^{h_-(\beta)} f(u, \beta)\, du} \right).$$

2.3. Singular limit solutions. In the previous sections, we constructed the lowest-order approximations $(U_0^+(y;\beta) + W_0^+(y/\sigma; \alpha, \beta), V_0^+(y; \beta))$ and $(U_0^-(y; \beta; d) + W_0^-(y/\sigma; \alpha, \beta; d), V_0^-(y; \beta; d))$ to the problems $(2.2)_{\pm}$, respectively. In order to construct an approximate solution to (2.1), which belongs to $\mathscr{D} \times \mathscr{D}$, we choose α and β such that $(U_0^+(y; \beta) + W_0^+(y/\sigma; \alpha, \beta), V_0^+(y; \beta))$ and $(U_0^-(y; \beta; d) + W_0^-(y/\sigma; \alpha, \beta; d), V_0^-(y; \beta; d))$ are matched to give C^1-continuity at $y = 0$. That is, we determine α and β as functions of d such that

$$\begin{aligned} &\Phi_0(\alpha, \beta; d) \equiv \frac{d}{d\xi} W_0^+(0; \alpha, \beta) - \frac{d}{d\xi} W_0^-(0; \alpha, \beta; d) = 0, \\ &\Psi_0(\beta; d) \equiv \frac{d}{dy} V_0^+(0; \beta) - \frac{d}{dy} V_0^-(0; \beta; d) = 0. \end{aligned} \tag{2.6}$$

We call the relations $\Phi_0(\alpha, \beta; d) = 0$ and $\Psi_0(\beta; d) = 0$ the *inner* and *outer matching conditions*, respectively. By Lemma 2.1, we find that

$$\Psi_0(\beta; d) = \sqrt{2 \int_{\beta}^{v_+} g(h_+(v), v)\, dv} - \sqrt{2d \int_{\beta}^{v_-} g(h_-(v), v)\, dv}.$$

Then we directly obtain the following result.

LEMMA 2.3. $\Psi_0(\beta; d) = 0$ *has a unique root* $\beta = \beta_0(d)$ *for all* $d > 0$, *which is strictly monotone decreasing and satisfies*

$$\lim_{d \downarrow 0} \beta_0(d) = v_+ \quad \textit{and} \quad \lim_{d \uparrow \infty} \beta_0(d) = v_-.$$

On the other hand, by Lemma 2.2, $\Phi_0(\alpha, \beta; d) = 0$ is rewritten as

$$d \int_{\alpha}^{h_-(\beta)} f(u, \beta)\, du - \int_{\alpha}^{h_+(\beta)} f(u, \beta)\, du = 0,$$

that is,

$$d = d_0(\alpha, \beta) = \frac{\displaystyle\int_{\alpha}^{h_+(\beta)} f(u, \beta)\, du}{\displaystyle\int_{\alpha}^{h_-(\beta)} f(u, \beta)\, du}. \tag{2.7}$$

LEMMA 2.4. *$d_0(\alpha, \beta)$ is a function defined in each Σ_i $(i=1,2,3)$ and satisfies $(\partial/\partial\beta)d_0(\alpha,\beta)<0$. Moreover:*

(i) *For any fixed β satisfying $v^*<\beta<v_+$, $d_0(\alpha,\beta)$ is less than 1 and monotone increasing in $(\omega_0(\beta), h_0(\beta))$ and monotone decreasing in $(h_0(\beta), h_+(\beta))$ and satisfies $\lim_{\alpha\downarrow\omega_0(\beta)} d_0(\alpha,\beta)=0=\lim_{\alpha\uparrow h_+(\beta)} d_0(\alpha,\beta)$ and $\lim_{\alpha\uparrow h_+(\beta)}(\partial/\partial\alpha)d_0(\alpha,\beta)=0=(\partial/\partial\alpha)d_0(h_0(\beta),\beta)$;*

(ii) *For any fixed β satisfying $v_-<\beta<v^*$, $d_0(\alpha,\beta)$ is greater than 1 and monotone decreasing in $(h_-(\beta), h_0(\beta))$ and monotone increasing in $(h_0(\beta), \omega_1(\beta))$ and satisfies $\lim_{\alpha\downarrow h_-(\beta)} d_0(\alpha,\beta)=+\infty=\lim_{\alpha\uparrow\omega_1(\beta)} d_0(\alpha,\beta)$, and $(\partial/\partial\alpha)d_0(h_0(\beta),\beta)=0$;*

(iii) *For $\beta=v^*$, $d_0(\alpha,\beta)=1$ in $(h_-(\beta), h_+(\beta))$.*

Consequently, by Lemma 2.3, it is sufficient to determine $\alpha=\alpha(d)$ such that

$$\Phi_0(\alpha, \beta_0(d); d)=0 \tag{2.8}$$

holds, which is obtained by Lemma 2.4. Combining the above results, we obtain Lemma 2.5.

LEMMA 2.5. (i) *When $v_-<v^*<v_+$, the following three cases occur:*

(a) *If $\beta_0(1)>v^*$, there are two points d_1 and d_2 $(d_1<1<d_2)$ such that (2.8) has two roots $\alpha=\bar\alpha_1(d)$ and $\alpha=\underline\alpha_1(d)$ with $\bar\alpha_1(d)>\underline\alpha_1(d)$ and $\bar\alpha_1(d_1)=\underline\alpha_1(d_1)$ defined for $d\in(0,d_1]$ and has two roots $\alpha=\bar\alpha_2(d)$ and $\alpha=\underline\alpha_2(d)$ with $\bar\alpha_2(d)>\underline\alpha_2(d)$ defined for $d\in[d_2,\infty)$;*

(b) *If $\beta_0(1)<v^*$, there are two points d_1 and d_2 $(d_1<1<d_2)$ such that (2.8) has two roots $\alpha=\bar\alpha_1(d)$ and $\alpha=\underline\alpha_1(d)$ with $\bar\alpha_1(d)>\underline\alpha_1(d)$ defined for $d\in(0,d_1]$ and has two roots $\alpha=\bar\alpha_2(d)$ and $\alpha=\underline\alpha_2(d)$ with $\bar\alpha_2(d)>\underline\alpha_2(d)$ and $\bar\alpha_2(d_2)=\underline\alpha_2(d_2)$ defined for $d\in[d_2,\infty)$;*

(c) *If $\beta_0(1)=v^*$, (2.8) has two roots $\alpha=\bar\alpha_1(d)$ and $\alpha=\underline\alpha_1(d)$ with $\bar\alpha_1(d)>\underline\alpha_1(d)$ defined for $d\in(0,1]$ and has two roots $\alpha=\bar\alpha_2(d)$ and $\alpha=\underline\alpha_2(d)$ with $\bar\alpha_2(d)>\underline\alpha_2(d)$ defined for $d\in[1,\infty)$. Especially when $d=1$, $\bar\alpha_1(1)=\bar\alpha_2(1)$, $\underline\alpha_1(1)=\underline\alpha_2(1)$, and (2.8) also holds for any $\alpha\in(\bar\alpha_1(1), \underline\alpha_1(1))$.*

(ii) *When $v^*\leqq v_-$, there is $d_1(<1)$ such that (2.8) has two roots $\alpha=\bar\alpha_1(d)$ and $\alpha=\underline\alpha_1(d)$ with $\bar\alpha_1(d)>\underline\alpha_1(d)$ and $\bar\alpha_1(d_1)=\underline\alpha_1(d_1)$ defined for $d\in(0,d_1]$.*

(iii) *When $v_+\leqq v^*$, there is d_2 (>1) such that (2.8) has two roots $\alpha=\bar\alpha_2(d)$ and $\alpha=\underline\alpha_2(d)$ with $\bar\alpha_2(d)>\underline\alpha_2(d)$ and $\bar\alpha_2(d_2)=\underline\alpha_2(d_2)$ defined for $d\in[d_2,\infty)$.*

For any $(\alpha^*(d), \beta^*(d))$ satisfying the inner and outer matching conditions (2.6), we define

$$U_0(y;\sigma;d)\equiv\begin{cases} U_0^+(y;\beta^*(d))+W_0^+\left(\dfrac{y}{\sigma};\alpha^*(d),\beta^*(d)\right), & y\in\mathbf{R}_+,\\ U_0^-(y;\beta^*(d);d)+W_0^-\left(\dfrac{y}{\sigma};\alpha^*(d),\beta^*(d);d\right), & y\in\mathbf{R}_-\end{cases}$$

and

$$V_0(y;\sigma;d)\equiv\begin{cases} V_0^+(y;\beta^*(d)), & y\in\mathbf{R}_+,\\ V_0^-(y;\beta^*(d);d), & y\in\mathbf{R}_-,\end{cases}$$

which becomes the lowest-order approximate solution uniformly on $\mathbf{R}$. We call $(U_0, V_0)(y;\sigma;d)$ a *singular limit solution of* (2.1). By Lemma 2.5, we obtain Theorem 2.1.

THEOREM 2.1. *Suppose that* (A1)-(A4) *hold and d_1 and d_2 are the same constants as in Lemma* 2.5.

(i) *When* $v_- < v^* < v_+$, *the following three cases occur*:

(a) *If* $\beta_0(1) > v^*$, (2.1) *has two singular limit solutions*

$$(\bar{U}_{0,*}, \bar{V}_{0,*})(y;\sigma;d), \quad (\underline{U}_{0,*}, \underline{V}_{0,*})(y;\sigma;d) \quad \textit{for } d \in (0, d_1],$$
$$(\bar{U}_0^*, \bar{V}_0^*)(y;\sigma;d), \quad (\underline{U}_0^*, \underline{V}_0^*)(y;\sigma;d) \quad \textit{for } d \in [d_2, \infty);$$

(b) *If* $\beta_0(1) < v^*$, (2.1) *has two singular limit solutions*

$$(\bar{U}_{0,*}, \bar{V}_{0,*})(y;\sigma;d), \quad (\underline{U}_{0,*}, \underline{V}_{0,*})(y;\sigma;d) \quad \textit{for } d \in (0, d_1],$$
$$(\bar{U}_0^*, \bar{V}_0^*)(y;\sigma;d), \quad (\underline{U}_0^*, \underline{V}_0^*)(y;\sigma;d) \quad \textit{for } d \in [d_2, \infty);$$

(c) *If* $\beta_0(1) = v^*$, (2.1) *has two singular limit solutions*

$$(\bar{U}_{0,*}, \bar{V}_{0,*})(y;\sigma;d), \quad (\underline{U}_{0,*}, \underline{V}_{0,*})(y;\sigma;d) \quad \textit{for } d \in (0, 1],$$
$$(\bar{U}_0^*, \bar{V}_0^*)(y;\sigma;d), \quad (\underline{U}_0^*, \underline{V}_0^*)(y;\sigma;d) \quad \textit{for } d \in [1, \infty);$$

(ii) *When* $v^* \leqq v_-$, (2.1) *has two singular limit solutions* $(\bar{U}_{0,*}, \bar{V}_{0,*})(y;\sigma;d)$ *and* $(\underline{U}_{0,*}, \underline{V}_{0,*})(y;\sigma;d)$ *for* $d \in (0, d_1]$.

(iii) *When* $v_+ \leqq v^*$, (2.1) *has two singular limit solutions* $(\bar{U}_0^*, \bar{V}_0^*)(y;\sigma;d)$ *and* $(\underline{U}_0^*, \underline{V}_0^*)(y;\sigma;d)$ *for* $d \in [d_2, \infty)$.

2.4. Existence theorem. We construct solutions of (2.1) for small σ, which tend to the above singular limit solutions as $\sigma \downarrow 0$ in a suitable topology. We assume that $(\alpha^*(d), \beta^*(d))$ satisfies the inner and outer matching conditions

$$\Phi_0(\alpha^*(d), \beta^*(d); d) = 0 = \Psi_0(\beta^*(d); d) \tag{2.9}$$

and moreover

$$\frac{\partial}{\partial \alpha} \Phi_0(\alpha^*(d), \beta^*(d); d) \neq 0. \tag{2.10}$$

In order to prove the existence of solutions of (2.1) for small but not zero σ, we need to assume (2.10). For any fixed $(\alpha, \beta) \in \Lambda_\nu \equiv \{(\alpha, \beta) \,|\, |\alpha - \alpha^*(d)| + |\beta - \beta^*(d)| \leqq \nu\}$, we first look for solutions of the problems $(2.2)_\pm$, which take the forms

$$\begin{aligned} u^+(y;\sigma;\alpha,\beta) &= U_0^+(y;\beta) + W_0^+\left(\frac{y}{\sigma};\alpha,\beta\right) + r^+(y;\sigma;\alpha,\beta) \\ &\quad + h_+(V_0^+(y;\beta))s^+(y;\sigma;\alpha,\beta), \\ v^+(y;\sigma;\alpha,\beta) &= V_0^+(y;\beta) + \sigma^2 Y^+\left(\frac{y}{\sigma};\alpha,\beta\right) + s^+(y;\sigma;\alpha,\beta), \end{aligned} \tag{2.11}_+$$

and

$$\begin{aligned} u^-(y;\sigma;\alpha,\beta;d) &= U_0^-(y;\beta;d) + W_0^-\left(\frac{y}{\sigma};\alpha,\beta;d\right) + r^-(y;\sigma;\alpha,\beta;d) \\ &\quad + h_-(V_0^-(y;\beta;d))s^-(y;\sigma;\alpha,\beta;d), \\ v^-(y;\sigma;\alpha,\beta;d) &= V_0^-(y;\beta;d) + \sigma^2 Y^-\left(\frac{y}{\sigma};\alpha,\beta;d\right) + s^-(y;\sigma;\alpha,\beta;d), \end{aligned} \tag{2.11}_-$$

where

$$Y^+\left(\frac{y}{\sigma};\alpha,\beta\right) = Y_0^+\left(\frac{y}{\sigma};\alpha,\beta\right) - e^{-\mu y} Y_0^+(0;\alpha,\beta)$$

with

$$Y_0^+(\xi;\alpha,\beta)=-\int_{\xi}^{+\infty}\int_{\eta}^{+\infty}\{g(h_+(\beta)+W_0^+(\zeta;\alpha,\beta),\beta)-g(h_+(\beta),\beta)\}\,d\zeta\,d\eta$$

and

$$Y^-\left(\frac{y}{\sigma};\alpha,\beta;d\right)=Y_0^-\left(\frac{y}{\sigma};\alpha,\beta;d\right)-e^{\mu y}Y_0^-(0;\alpha,\beta;d)$$

with

$$Y_0^-(\xi;\alpha,\beta;d)=-d\int_{-\infty}^{\xi}\int_{-\infty}^{\eta}\{g(h_-(\beta)+W_0^-(\zeta;\alpha,\beta;d),\beta)-g(h_-(\beta),\beta)\}\,d\zeta\,d\eta$$

and μ is an arbitrarily fixed number satisfying $\mu>\max(\mu_+,\mu_-(d))$. Let us define a function space $\mathring{X}_\sigma(\mathbf{R}_\pm)$ by

$$\mathring{X}_\sigma(\mathbf{R}_\pm)\equiv\mathring{X}^2_{\rho,\sigma}(\mathbf{R}_\pm)\times\mathring{X}^2_{\rho,1}(\mathbf{R}_\pm)$$

for any fixed constant ρ $(0<\rho<\min(\mu_+,\mu_-(d)))$. Then we obtain the following result for the remainders $t^\pm=(r^\pm,s^\pm)$.

LEMMA 2.6. *There are $\sigma_1>0$ and $\nu_1>0$ such that for any $\sigma\in(0,\sigma_1)$ and $(\alpha,\beta)\in\Lambda_{\nu_1}$, there exist $t^\pm(\sigma;\alpha,\beta)\in\mathring{X}_\sigma(\mathbf{R}_\pm)$ for which $(2.11)_\pm$ satisfy $(2.2)_\pm$. Moreover, $t^\pm(\sigma;\alpha,\beta)$, $(\partial t^\pm/\partial\alpha)(\sigma;\alpha,\beta)$ and $(\partial t^\pm/\partial\beta)(\sigma;\alpha,\beta)$ are uniformly continuous with respect to $(\sigma,\alpha,\beta)\in(0,\sigma_1)\times\Lambda_{\nu_1}$ in the $\mathring{X}_\sigma(\mathbf{R}_\pm)$-topology and satisfy*

$$\|t^\pm(\sigma;\alpha,\beta)\|_{\mathring{X}_\sigma(\mathbf{R}_\pm)}=o(1),$$

$$\left\|\frac{\partial t^\pm}{\partial\alpha}(\sigma;\alpha,\beta)\right\|_{\mathring{X}_\sigma(\mathbf{R}_\pm)}=o(1),$$

$$\left\|\frac{\partial t^\pm}{\partial\beta}(\sigma;\alpha,\beta)\right\|_{\mathring{X}_\sigma(\mathbf{R}_\pm)}=o(1)$$

as $\sigma\downarrow 0$ uniformly in $(\alpha,\beta)\in\Lambda_{\nu_1}$.

Finally, we construct a solution of (2.1) on the whole interval $\mathbf{R}$ by matching $(u^+,v^+)(y;\sigma;\alpha,\beta)$ and $(u^-,v^-)(y;\sigma;\alpha,\beta;d)$ at $y=0$ in the C^1-sense. For this purpose, we define

$$\begin{aligned}\Phi(\sigma;\alpha,\beta;d)&\equiv\sigma\frac{d}{dy}u^+(0;\sigma;\alpha,\beta)-\sigma\frac{d}{dy}u^-(0;\sigma;\alpha,\beta;d),\\ \Psi(\sigma;\alpha,\beta;d)&\equiv\frac{d}{dy}v^+(0;\sigma;\alpha,\beta)-\frac{d}{dy}v^-(0;\sigma;\alpha,\beta;d),\end{aligned}\tag{2.12}$$

and determine α and β as functions of σ and d such that the relations

$$\Phi(\sigma;\alpha,\beta;d)=0=\Psi(\sigma;\alpha,\beta;d)\tag{2.13}$$

hold. We call (2.13) the *matching condition*. We extend Φ and Ψ continuously so as to be defined for $\sigma=0$. Putting $\sigma=0$ in (2.12), we obtain by Lemma 2.6 that

$$\Phi(0;\alpha,\beta;d)=\Phi_0(\alpha,\beta;d),$$

$$\Psi(0;\alpha,\beta;d)=\Psi_0(\beta;d).$$

Here we define $\mathbf{I}(d)$ by using the *Jacobian of the matching condition* (2.13) at $\sigma=0$ as follows:

$$\mathbf{I}(d)=\operatorname{sign}\{\mathbf{J}(d)\} \tag{2.14}$$

with

$$\mathbf{J}(d)\equiv \det\begin{bmatrix} \dfrac{\partial}{\partial\alpha}\Phi_0(\alpha^*(d),\beta^*(d);d) & \dfrac{\partial}{\partial\beta}\Phi_0(\alpha^*(d),\beta^*(d);d) \\ \dfrac{\partial}{\partial\alpha}\Psi_0(\beta^*(d);d) & \dfrac{\partial}{\partial\beta}\Psi_0(\beta^*(d);d) \end{bmatrix},$$

where

$$\operatorname{sign}\{x\}=\begin{cases} +1 & \text{for } x>0, \\ 0 & \text{for } x=0, \\ -1 & \text{for } x<0. \end{cases}$$

Since $\alpha^*(d)$ and $\beta^*(d)$ satisfy (2.9) and (2.10), we find by Lemmas 2.1 and 2.2 that

$$\frac{\partial}{\partial\beta}\Psi_0(\beta^*(d);d)=\frac{dg(h_-(\beta^*(d)),\beta^*(d))-g(h_+(\beta^*(d)),\beta^*(d))}{\sqrt{2\int_{\beta^*(d)}^{v_+} g(h_+(s),s)\,ds}}<0$$

and

$$\frac{\partial}{\partial\alpha}\Phi_0(\alpha^*(d),\beta^*(d);d)=\frac{(d-1)f(\alpha^*(d),\beta^*(d))}{\sqrt{2\int_{\alpha^*(d)}^{h_+(\beta^*(d))} f(u,\beta^*(d))\,du}}. \tag{2.15}$$

We note that $\mathbf{I}(d)$ is essentially determined by the inner matching condition. Since $\mathbf{I}(d)\neq 0$ follows from assumption (2.10), the implicit function theorem can be applied to (2.13). That is, there is $\sigma_2>0$ such that there uniquely exist continuous functions $\alpha(\sigma;d)$ and $\beta(\sigma;d)$ satisfying (2.13) for $\sigma\in[0,\sigma_2)$ and $\lim_{\sigma\downarrow 0}\alpha(\sigma;d)=\alpha^*(d)$ and $\lim_{\sigma\downarrow 0}\beta(\sigma;d)=\beta^*(d)$. Thus, we obtain the desired theorem.

THEOREM 2.2. *Suppose that* (A1)-(A4) *hold and d is fixed arbitrarily such that* $(\alpha^*(d),\beta^*(d))$ *satisfies* (2.9) *and* (2.10), *that is,* $\mathbf{I}(d)\neq 0$. *Then for any* $\sigma\in(0,\sigma_2)$ *there exists a solution* $(u,v)(y;\sigma;d)$ *of* (2.1) *corresponding to the singular limit solution* (U_0,V_0) *such that*

$$\|u(y;\sigma;d)-U_0(y;\sigma;d)\|_{X^1_{\rho,\sigma}(\mathbf{R})}+\|v(y;\sigma;d)-V_0(y;\sigma;d)\|_{X^1_{\rho,1}(\mathbf{R})}\to 0$$

holds as $\sigma\downarrow 0$.

COROLLARY 2.1. *Suppose that* (A1)-(A4) *hold and* d_1 *and* d_2 *are the same constants as in Lemma* 2.5.

(i) *When* $v_-<v^*<v_+$, *the following three cases occur*:

(a) If $\beta_0(1)>v^*$, (2.1) *has two solutions*

$(\bar U_*,\bar V_*)(y;\sigma;d)(\mathbf{I}(d)=1)$, $(\underline U_*,\underline V_*)(y;\sigma;d)(\mathbf{I}(d)=-1)$ *for* $d\in(0,d_1)$,

$(\bar U^*,\bar V^*)(y;\sigma;d)(\mathbf{I}(d)=-1)$, $(\underline U^*,\underline V^*)(y;\sigma;d)(\mathbf{I}(d)=1)$ *for* $d\in[d_2,\infty)$;

(b) *If* $\beta_0(1)<v^*$, (2.1) *has two solutions*

$(\bar U_*,\bar V_*)(y;\sigma;d)(\mathbf{I}(d)=1)$, $(\underline U_*,\underline V_*)(y;\sigma;d)(\mathbf{I}(d)=-1)$ *for* $d\in(0,d_1]$,

$(\bar U^*,\bar V^*)(y;\sigma;d)(\mathbf{I}(d)=-1)$, $(\underline U^*,\underline V^*)(y;\sigma;d)(\mathbf{I}(d)=1)$ *for* $d\in(d_2,\infty)$;

(c) *If* $\beta_0(1)=v^*$, (2.1) *has two solutions*

$(\bar U_*,\bar V_*)(y;\sigma;d)(\mathbf{I}(d)=1)$, $(\underline U_*,\underline V_*)(y;\sigma;d)(\mathbf{I}(d)=-1)$ *for* $d\in(0,1)$,

$(\bar U^*,\bar V^*)(y;\sigma;d)(\mathbf{I}(d)=-1)$, $(\underline U^*,\underline V^*)(y;\sigma;d)(\mathbf{I}(d)=1)$ *for* $d\in(1,\infty)$;

(ii) *When* $v^* \leqq v_-$, (2.1) *has two solutions* $(\bar{U}_*, \bar{V}_*)(y; \sigma; d)(\mathbf{I}(d) = 1)$ *and* $(\underline{U}_*, \underline{V}_*)(y; \sigma; d)(\mathbf{I}(d) = -1)$ *for* $d \in (0, d_1)$.

(iii) *When* $v_+ \leqq v^*$, (2.1) *has two solutions* $(\bar{U}^*, \bar{V}^*)(y; \sigma; d)(\mathbf{I}(d) = -1)$ *and* $(\underline{U}^*, \underline{V}^*)(y; \sigma; d)(\mathbf{I}(d) = 1)$ *for* $d \in (d_2, \infty)$ (*see Fig.* 8).

Remark 2.1. If f and g are given by (1.6), the cases where $\gamma > \gamma_2$, $\gamma \in (\gamma_1, \gamma_2)$, and $\gamma \in (\gamma_0, \gamma_1)$ correspond to (iii) $(v_+ \leqq v^*)$, (i)(b) $(v_- < v^* < v_+, \beta_0(1) < v^*)$, and (i)(a) $(v_- < v^* < V_+, \beta_0(1) > v^*)$, respectively.

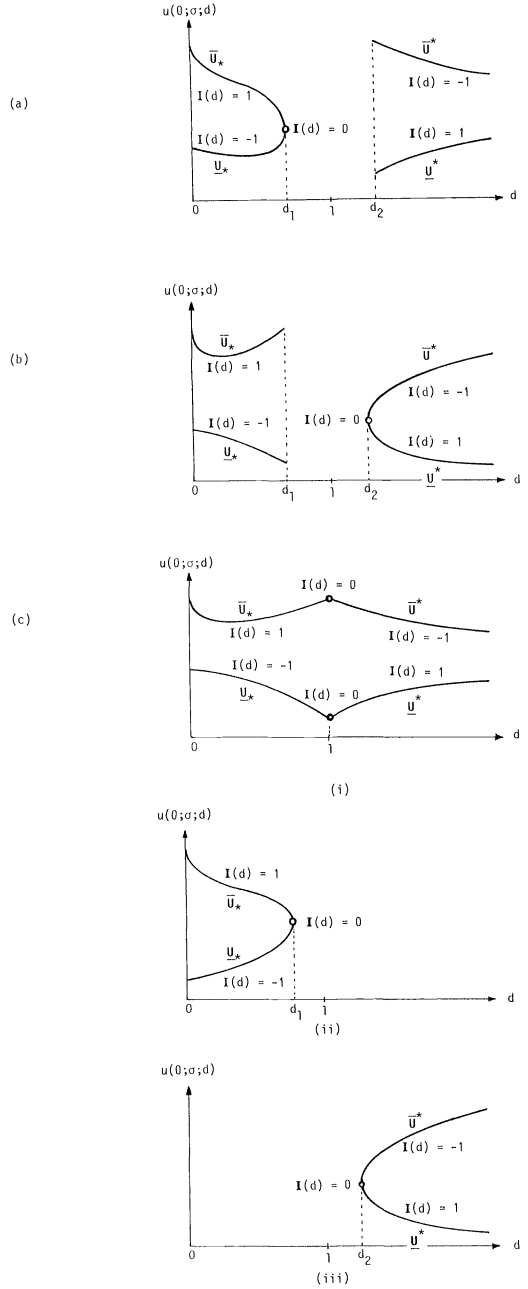


FIG. 8. *Global pictures of stationary solutions of* (1.1) *and these* $\mathbf{I}(d)$: (i) $v_- < v^* < v_+$: (a) $\beta_0(1) > v^*$, (b) $\beta_0(1) < v^*$, (c) $\beta_0(1) = v^*$; (ii) $v^* \leqq v_-$; (iii) $v_+ \leqq v^*$.

3. Stability of stationary solutions. Our aim in this section is to show that under (A1)-(A4), $\mathbf{I}(d)$ corresponds in a one-to-one manner to the stability of stationary solutions of (1.7) constructed in § 2, that is, the stationary solution is *stable* (respectively, *unstable*) if and only if $\mathbf{I}(d)$ is $+1$ (respectively, -1). We remark that this result is closely related to the index argument on stability properties of nerve equations [Ev1], [Ev2], [AGJ], [GJ], [I].

3.1. Preliminaries for stability analysis. Let $(U, V)(x; \sigma; d)$ be a stationary solution of (1.7). We will determine its stability by the linearized stability criterion. The linearized eigenvalue problem of (1.7) at $(U, V)(x; \sigma; d)$ is given by

$$\mathscr{L}^{\sigma,d}\begin{bmatrix}\varphi\\ \psi\end{bmatrix} = \lambda\begin{bmatrix}\varphi\\ \psi\end{bmatrix}, \qquad x \in \mathbf{R}\backslash\{0\}, \tag{3.1}$$

where

$$\mathscr{L}^{\sigma,d} \equiv \begin{bmatrix} \delta_0(x)\dfrac{d^2}{dx^2} + \dfrac{1}{\sigma} f_u^{\sigma,d} & \dfrac{1}{\sigma} f_v^{\sigma,d} \\ \sigma g_u^{\sigma,d} & \delta_0(x)\dfrac{d^2}{dx^2} + \sigma g_v^{\sigma,d} \end{bmatrix}.$$

Here $f_u^{\sigma,d}$, $f_v^{\sigma,d}$, $g_u^{\sigma,d}$, and $g_v^{\sigma,d}$ denote the partial derivatives of f and g evaluated at $(U, V)(x; \sigma; d)$. The underlying space for (3.1) is taken as $BC(\mathbf{R}) \times BC(\mathbf{R})$ with

$$\mathscr{D}(\mathscr{L}^{\sigma,d}) = \mathscr{D} \times \mathscr{D}.$$

Since $\mathscr{L}^{\sigma,d}$ is a sectorial operator, its spectral analysis assures the nonlinear stability or instability of the stationary solution (for instance, see [H]). Therefore it suffices to consider the following two problems:

(i) Distribution of the *essential spectrum*;

(ii) Distribution of *isolated eigenvalues.*

Noting that $(u_\pm, v_\pm)$ are both stable constant solutions, we first show the following proposition.

PROPOSITION 3.1. *Assume* (A1)-(A4). *Then there exists a positive constant* l_1 *independently of* d *and* σ *such that*

$$\mathrm{Re}\,\{\textit{essential spectrum of } (3.1)\} \leqq -\sigma l_1$$

holds for any $d > 0$ *and small* $\sigma > 0$.

The proof is given in the Appendix.

Since the essential spectrum of (3.1) is strictly bounded away from the imaginary axis, the stability property of the stationary solution is studied by the distribution of the isolated eigenvalues. The complex number λ is called an eigenvalue of (3.1) if and only if (3.1) has a nontrivial solution $(\varphi, \psi)(x; \sigma; d; \lambda)$ belonging to $BC(\mathbf{R}) \times BC(\mathbf{R})$. Therefore the eigenfunction $(\varphi, \psi)(x; \sigma; d; \lambda)$ must belong to $\mathscr{D}(\mathscr{L}^{\sigma,d})$. By the definition of $\mathscr{D}$, the eigenvalue problem (3.1) can be rewritten equivalently as

$$\begin{aligned} &\delta_0(x)\varphi_{xx}^{\pm} + \frac{1}{\sigma} f_u^{\sigma,d}\varphi^{\pm} + \frac{1}{\sigma} f_v^{\sigma,d}\psi^{\pm} = \lambda\varphi^{\pm}, \qquad x \in \mathbf{R}_\pm \\ &\delta_0(x)\psi_{xx}^{\pm} + \sigma g_u^{\sigma,d}\varphi^{\pm} + \sigma g_v^{\sigma,d}\psi^{\pm} = \lambda\psi^{\pm}, \end{aligned} \tag{3.2}_\pm$$

with compatibility conditions

$$\begin{aligned} &\varphi^+(0; \sigma; d; \lambda) = \varphi^-(0; \sigma; d; \lambda), \qquad \psi^+(0; \sigma; d; \lambda) = \psi^-(0; \sigma; d; \lambda), \\ &\varphi_x^+(0; \sigma; d; \lambda) = \varphi_x^-(0; \sigma; d; \lambda), \qquad \psi_x^+(0; \sigma; d; \lambda) = \psi_x^-(0; \sigma; d; \lambda), \end{aligned} \tag{3.3}$$

where $\varphi^+(x;\sigma;d;\lambda)$, $\psi^+(x;\sigma;d;\lambda)\in BC^2([0,\infty))$, and $\varphi^-(x;\sigma;d;\lambda)$, $\psi^-(x;\sigma;d;\lambda)\in BC^2((-\infty,0])$. Then, the eigenfunctions φ and ψ of (3.1) are, respectively, represented as

$$\varphi(x;\sigma;d;\lambda)=\begin{cases}\varphi^+(x;\sigma;d;\lambda), & x\geqq 0,\\ \varphi^-(x;\sigma;d;\lambda), & x\leqq 0,\end{cases}$$

and

$$\psi(x;\sigma;d;\lambda)=\begin{cases}\psi^+(x;\sigma;d;\lambda), & x\geqq 0,\\ \psi^-(x;\sigma;d;\lambda), & x\leqq 0.\end{cases}$$

Let us rewrite the problems $(3.2)_\pm$, (3.3) as the four-dimensional systems

$$(3.4)_\pm\qquad \begin{aligned}&\frac{dv_1^\pm}{dx}=v_2^\pm,\\ &\frac{dv_2^\pm}{dx}=\left(\lambda-\frac{1}{\sigma}f_u^{\sigma,d}\right)v_1^\pm/\delta_0(x)-\frac{1}{\sigma}f_v^{\sigma,d}v_3^\pm/\delta_0(x),\\ &\frac{dv_3^\pm}{dx}=v_4^\pm,\\ &\frac{dv_4^\pm}{dx}=-\sigma g_u^{\sigma,d}v_1^\pm/\delta_0(x)+(\lambda-\sigma g_v^{\sigma,d})v_3^\pm/\delta_0(x),\end{aligned}\qquad x\in\mathbf{R}_\pm,$$

$$(3.5)\qquad \begin{aligned}&v_1^+(0;\sigma;d;\lambda)=v_1^-(0;\sigma;d;\lambda),\qquad v_3^+(0;\sigma;d;\lambda)=v_3^-(0;\sigma;d;\lambda),\\ &v_2^+(0;\sigma;d;\lambda)=v_2^-(0;\sigma;d;\lambda),\qquad v_4^+(0;\sigma;d;\lambda)=v_4^-(0;\sigma;d;\lambda),\end{aligned}$$

where $v_1^\pm=\varphi^\pm$, $v_2^\pm=\varphi_x^\pm$, $v_3^\pm=\psi^\pm$, and $v_4^\pm=\psi_x^\pm$. Using $\mathbf{V}^\pm=(v_1^\pm,v_2^\pm,v_3^\pm,v_4^\pm)^t$, we simply write $(3.4)_\pm$ as

$$(3.6)_\pm\qquad \frac{d}{dx}\mathbf{V}^\pm=A(x;\sigma;d;\lambda)\mathbf{V}^\pm,\qquad x\in\mathbf{R}_\pm.$$

Since the stationary solution $(U,V)(x;\sigma;d)$ approaches the equilibrium states $(u_\pm,v_\pm)$ with exponential order as $x\to\pm\infty$, the asymptotic behaviour of eigenfunctions for large $|x|$ can be studied by using the limiting systems of $(3.6)_\pm$ with constant coefficients

$$(3.7)_\pm\qquad \frac{d}{dx}\mathbf{V}^\pm=A(\pm\infty;\sigma;d;\lambda)\mathbf{V}^\pm,\qquad x\in\mathbf{R}_\pm,$$

where

$$A(\pm\infty;\sigma;d;\lambda)=\begin{bmatrix}0 & 1 & 0 & 0\\ (\lambda-(1/\sigma)f_u(u_\pm,v_\pm))/\delta_0(\pm\infty) & 0 & -(1/\sigma)f_v(u_\pm,v_\pm)/\delta_0(\pm\infty) & 0\\ 0 & 0 & 0 & 1\\ -\sigma g_u(u_\pm,v_\pm)/\delta_0(\pm\infty) & 0 & (\lambda-\sigma g_v(u_\pm,v_\pm))/\delta_0(\pm\infty) & 0\end{bmatrix}$$

with $\delta_0(+\infty)=1$ and $\delta_0(-\infty)=1/d$. Let $\mu_i^\pm(\sigma;d;\lambda)$ $(i=1,2,3,4)$ satisfying $\mathrm{Re}\,\{\mu_1^\pm\}\leqq\mathrm{Re}\,\{\mu_2^\pm\}\leqq\mathrm{Re}\,\{\mu_3^\pm\}\leqq\mathrm{Re}\,\{\mu_4^\pm\}$ denote the eigenvalues of $A(\pm\infty;\sigma;d;\lambda)$.

LEMMA 3.1. *There exists a positive constant l_2 independently of σ and d such that*

$$\begin{aligned}\mathrm{Re}\,\{\mu_1^{\pm}(\sigma; d; \lambda)\} \leqq \mathrm{Re}\,\{\mu_2^{\pm}(\sigma; d; \lambda)\} &< 0 \\ &< \mathrm{Re}\,\{\mu_3^{\pm}(\sigma; d; \lambda)\} \leqq \mathrm{Re}\,\{\mu_4^{\pm}(\sigma; d; \lambda)\}\end{aligned}$$

hold for all $\lambda \in \mathbf{C}_{l_2} \equiv \{\lambda \in \mathbf{C} \,|\, \mathrm{Re}\,\{\lambda\} \geqq -l_2\}$.

This lemma can be proved in a way similarly to Lemma 4.1 in [I], so we omit the proof.

By virtue of Lemma 3.1 and $(3.7)_{\pm}$, $(3.6)_{+}$ has just two linearly independent solutions $\mathbf{V}_i^{+}(x; \sigma; d; \lambda)$ $(i = 1, 2)$ for any $\lambda \in \mathbf{C}_{l_2}$, which satisfy

$$\mathbf{V}_i^{+}(x; \sigma; d; \lambda) \to \mathbf{0} \quad \text{as } x \to \infty \qquad (i = 1, 2)$$

and $(3.6)_{-}$ has just two linearly independent solutions $\mathbf{V}_i^{-}(x; \sigma; d; \lambda)$ $(i = 1, 2)$ for any $\lambda \in \mathbf{C}_{l_2}$, which satisfy

$$\mathbf{V}_i^{-}(x; \sigma; d; \lambda) \to \mathbf{0} \quad \text{as } x \to -\infty \qquad (i = 1, 2).$$

By a short calculation, we can find that $\mu_i^{\pm}(\sigma; d; \lambda)$ $(i = 1, 2, 3, 4)$ and $\mathbf{V}_i^{\pm}(x; \sigma; d; \lambda)$ $(i = 1, 2)$ depend analytically on $\lambda \in \mathbf{C}_{l_2}$.

A nontrivial solution of $(3.4)_{\pm}$

$$\mathbf{V}(x; \sigma; d; \lambda) \equiv \begin{cases} \mathbf{V}^{+}(x; \sigma; d; \lambda), & x \geqq 0, \\ \mathbf{V}^{-}(x; \sigma; d; \lambda), & x \leqq 0, \end{cases}$$

which corresponds to an eigenvalue $\lambda \in \mathbf{C}_{l_2}$, must satisfy

$$\mathbf{V}^{\pm}(x; \sigma; d; \lambda) \to \mathbf{0} \quad \text{as } x \to \pm\infty.$$

Hence $\mathbf{V}^{\pm}(x; \sigma; d; \lambda)$ can be written as

$$\mathbf{V}^{\pm}(x; \sigma; d; \lambda) = \sum_{i=1}^{2} \alpha_i^{\pm} \mathbf{V}_i^{\pm}(x; \sigma; d; \lambda), \qquad x \in \mathbf{R}_{\pm}$$

for some constants $\alpha_i^{\pm}$ $(i = 1, 2)$. Conditions (3.5) lead to

$$\sum_{i=1}^{2} \alpha_i^{+} \mathbf{V}_i^{+}(0; \sigma; d; \lambda) = \sum_{i=1}^{2} \alpha_i^{-} \mathbf{V}_i^{-}(0; \sigma; d; \lambda).$$

That is, $\lambda = \lambda_0$ is an eigenvalue of (3.1) if and only if the vectors $\mathbf{V}_i^{\pm}(0; \sigma; d; \lambda)$ $(i = 1, 2)$ are not linearly independent when $\lambda = \lambda_0$. Let

$$(3.8) \quad g(\sigma; d; \lambda) \equiv \det[\mathbf{V}_1^{+}(0; \sigma; d; \lambda), \mathbf{V}_2^{+}(0; \sigma; d; \lambda), \mathbf{V}_1^{-}(0; \sigma; d; \lambda), \mathbf{V}_2^{-}(0; \sigma; d; \lambda)].$$

We find that $g(\sigma; d; \lambda)$ is an analytic function of $\lambda \in \mathbf{C}_{l_2}$ and we have the following lemma.

LEMMA 3.2. *For any $\lambda \in \mathbf{C}_{l_2}$, λ is an eigenvalue of* (3.1) *if and only if $g(\sigma; d; \lambda) = 0$ holds.*

3.2. Relation between I(d) and stability of stationary solutions. In order to examine the distribution of eigenvalues, that is, to find solutions λ of $g(\sigma; d; \lambda) = 0$, we construct linearly independent solutions $\mathbf{V}_i^{\pm}(x; \sigma; d; \lambda)$ $(i = 1, 2)$ satisfying

$$(3.9)_{\pm} \qquad \frac{d}{dx}\mathbf{V}_i^{\pm} = A(x; \sigma; d; \lambda)\mathbf{V}_i^{\pm}, \qquad x \in \mathbf{R}_{\pm},$$

$$(3.10)_{\pm} \qquad \mathbf{V}_i^{\pm}(x; \sigma; d; \lambda) \to \mathbf{0} \quad \text{as } x \to \pm\infty.$$

First, we solve the following problems for $(\varphi^{\pm}, \psi^{\pm})$ in $\mathbf{R}_{\pm}$:

$$
(3.11)_{\pm}\qquad
\begin{aligned}
&\delta_0(x)\varphi^{\pm}_{xx}+\frac{1}{\sigma}f^{\sigma,d}_u\varphi^{\pm}+\frac{1}{\sigma}f^{\sigma,d}_v\psi^{\pm}=\lambda\varphi^{\pm}, && x\in\mathbf{R}_{\pm},\\
&\delta_0(x)\psi^{\pm}_{xx}+\sigma g^{\sigma,d}_u\varphi^{\pm}+\sigma g^{\sigma,d}_v\psi^{\pm}=\lambda\psi^{\pm},\\
&\varphi^{\pm}(\pm\infty)=0, \qquad \varphi^{\pm}(0)=a,\\
&\psi^{\pm}(\pm\infty)=0, \qquad \psi^{\pm}(0)=b,
\end{aligned}
$$

where a and b are arbitrarily fixed constants, and denote their solutions by $(\varphi^{\pm}, \psi^{\pm})$-$(x; \sigma; d; \lambda; a, b)$. Letting

$$
(3.12)\qquad
\begin{aligned}
\mathbf{V}^{\pm}_1(x;\sigma;d;\lambda)&=\begin{bmatrix}\varphi^{\pm}(x;\sigma;d;\lambda;1,0)\\ \varphi^{\pm}_x(x;\sigma;d;\lambda;1,0)\\ \psi^{\pm}(x;\sigma;d;\lambda;1,0)\\ \psi^{\pm}_x(x;\sigma;d;\lambda;1,0)\end{bmatrix} && (x\in\mathbf{R}_{\pm}),\\
\mathbf{V}^{\pm}_2(x;\sigma;d;\lambda)&=\begin{bmatrix}\varphi^{\pm}(x;\sigma;d;\lambda;0,1)\\ \varphi^{\pm}_x(x;\sigma;d;\lambda;0,1)\\ \psi^{\pm}(x;\sigma;d;\lambda;0,1)\\ \psi^{\pm}_x(x;\sigma;d;\lambda;0,1)\end{bmatrix} && (x\in\mathbf{R}_{\pm}),
\end{aligned}
$$

we find that $\mathbf{V}^{+}_i(x;\sigma;d;\lambda)$ $(i=1,2)$ are linearly independent solutions of $(3.9)_+$, $(3.10)_+$, and that $\mathbf{V}^{-}_i(x;\sigma;d;\lambda)$ $(i=1,2)$ are linearly independent solutions of $(3.9)_-$, $(3.10)_-$. Making use of $\mathbf{V}^{\pm}_i(x;\sigma;d;\lambda)$ $(i=1,2)$, we are able to seek λ satisfying $g(\sigma;d;\lambda)=0$ and arrive at the following stability theorem.

THEOREM 3.1. *Assume* (A1)-(A4) *and let* $(U_0, V_0)(x;\sigma;d)$ *and* $(U, V)(x;\sigma;d)$ *be an arbitrary singular limit solution and the corresponding stationary solution constructed in* § 2 *and define* $\mathbf{I}(d)$ *by* (2.14). *Then there exists a positive constant* l_3 *such that for any* $\lambda\in\mathbf{C}_{\sigma l_3}$ *the following hold: If* $\mathbf{I}(d)=1$, *the equation* $g(\sigma;d;\lambda)=0$ *has no roots, that is,* $(U, V)(x;\sigma;d)$ *is stable. Conversely, if* $\mathbf{I}(d)=-1$, $g(\sigma;d;\lambda)=0$ *has a unique simple positive root* $\lambda(\sigma;d)$, *that is,* $(U, V)(x;\sigma;d)$ *is unstable* (*see Fig.* 8).

The proof will be given in the next section.

4. Proof of Theorem 3.1. In this section, we shall represent the explicit form of the function $g(\sigma;d;\lambda)$ and then find solutions λ of $g(\sigma;d;\lambda)=0$.

First, setting $y=\sqrt{\sigma}\,x$, $(3.11)_{\pm}$ can be rewritten as

$$
(4.1)_{\pm}\qquad
\begin{aligned}
&\sigma^2\delta_0(y)\Phi^{\pm}_{yy}+\hat{f}^{\sigma,d}_u\Phi^{\pm}+\hat{f}^{\sigma,d}_v\Psi^{\pm}=\sigma^2\mu\Phi^{\pm}, && y\in\mathbf{R}_{\pm},\\
&\delta_0(y)\Psi^{\pm}_{yy}+\hat{g}^{\sigma,d}_u\Phi^{\pm}+\hat{g}^{\sigma,d}_v\Psi^{\pm}=\mu\Psi^{\pm},\\
&\Phi^{\pm}(\pm\infty)=0, \qquad \Phi^{\pm}(0)=1,\\
&\Psi^{\pm}(\pm\infty)=0, \qquad \Psi^{\pm}(0)=b,
\end{aligned}
$$

where $\mu=\lambda/\sigma$, $\hat{f}^{\sigma,d}_u=f_u(U(y/\sqrt{\sigma};\sigma;d), V(y/\sqrt{\sigma};\sigma;d))$, and $\hat{f}^{\sigma,d}_v$, $\hat{g}^{\sigma,d}_u$, and $\hat{g}^{\sigma,d}_v$ are similarly defined. Here we may set $a=1$. For solutions $(\Phi^{\pm}, \Psi^{\pm})(y;\sigma;d;\mu;b)$ and $(\varphi^{\pm}, \psi^{\pm})(x;\sigma;d;\lambda;a,b)$ of $(4.1)_{\pm}$ and $(3.11)_{\pm}$, respectively, we have the following obvious relations:

$$
(4.2)_{\pm}\qquad
\begin{aligned}
(\varphi^{\pm}, \psi^{\pm})(x;\sigma;d;\lambda;1,0)&=(\Phi^{\pm}, \Psi^{\pm})(\sqrt{\sigma}\,x;\sigma;d;\lambda/\sigma;0),\\
(\varphi^{\pm}, \psi^{\pm})(x;\sigma;d;\lambda;0,1)&=(\Phi^{\pm}, \Psi^{\pm})(\sqrt{\sigma}\,x;\sigma;d;\lambda/\sigma;1)\\
&\quad-(\Phi^{\pm}, \Psi^{\pm})(\sqrt{\sigma}x;\sigma;d;\lambda/\sigma;0).
\end{aligned}
$$

By using singular perturbation techniques, we solve $(4.1)_{\pm}$ for any $\mu \in \mathbf{C}_{l_3}$, where l_3 is a positive constant specified later. To do so, the discussion will be divided into four cases according to the dependency of μ on σ:

I. $\mu = \mu(\sigma) = O(1)$ (that is, $\lambda(\sigma) = O(\sigma)$) as $\sigma \downarrow 0$.

For the other three cases, we have $|\mu(\sigma)| \uparrow \infty$ as $\sigma \downarrow 0$, and we know by Lemma 1.1.1 of [Ec] that there exists a real positive and continuous function $\omega(\sigma)$ satisfying $\omega(\sigma) \uparrow \infty$ as $\sigma \downarrow 0$ such that $\mu(\sigma)$ is represented as

$$\mu(\sigma) = \omega(\sigma)\hat{\mu}(\sigma), \tag{4.3}$$

where $\hat{\mu}(\sigma)$ satisfies $|\hat{\mu}(0)| \neq 0$ (that is, under the assumption $\mu \in \mathbf{C}_{l_3}$, it holds that either $\operatorname{Re}\{\hat{\mu}(0)\} > 0$ or if $\operatorname{Re}\{\hat{\mu}(0)\} = 0$, $\operatorname{Im}\{\hat{\mu}(0)\} \neq 0$). Then, with $z = \sqrt{\omega(\sigma)}\, y$, $(4.1)_{\pm}$ are rewritten as

$$\begin{aligned}
&\sigma^2\omega(\sigma)\delta_0(z)\Phi^{\pm}_{zz} + \tilde{f}^{\sigma,d}_u\Phi^{\pm} + \tilde{f}^{\sigma,d}_v\Psi^{\pm} = \sigma^2\omega(\sigma)\hat{\mu}(\sigma)\Phi^{\pm}, \\
&\delta_0(z)\Psi^{\pm}_{zz} + \frac{\tilde{g}^{\sigma,d}_u}{\omega(\sigma)}\Phi^{\pm} + \frac{\tilde{g}^{\sigma,d}_v}{\omega(\sigma)}\Psi^{\pm} = \hat{\mu}(\sigma)\Psi^{\pm}, \qquad z \in \mathbf{R}_{\pm}, \\
&\Phi^{\pm}(\pm\infty) = 0, \qquad \Phi^{\pm}(0) = 1, \\
&\Psi^{\pm}(\pm\infty) = 0, \qquad \Psi^{\pm}(0) = b,
\end{aligned} \tag{4.4}_{\pm}$$

where $\tilde{f}^{\sigma,d}_u = f_u(U(z/\sqrt{\sigma\omega(\sigma)}; \sigma; d), V(z/\sqrt{\sigma\omega(\sigma)}; \sigma; d))$ and $\tilde{f}^{\sigma,d}_v$, $\tilde{g}^{\sigma,d}_u$, and $\tilde{g}^{\sigma,d}_v$ are similarly defined. We discuss $(4.4)_{\pm}$ by dividing the coefficient of $\Phi^{\pm}_{zz}$, say $\sigma^2\omega(\sigma)$, into the following three cases:

II. $\sigma^2\omega(\sigma) \downarrow 0$ and $\omega(\sigma) \uparrow \infty$ (that is, $\sigma\lambda(\sigma) \downarrow 0$ and $\lambda(\sigma)/\sigma \uparrow \infty$) as $\sigma \downarrow 0$;

III. $\sigma^2\omega(\sigma) \to \kappa$ for some positive constant κ (that is, $\sigma\lambda(\sigma)$ is bounded but does not converge to zero) as $\sigma \downarrow 0$;

IV. $\sigma^2\omega(\sigma) \uparrow \infty$ (that is, $\sigma\lambda(\sigma) \uparrow \infty$) as $\sigma \downarrow 0$.

Case I. $\mu(\sigma) = O(1)$ as $\sigma \downarrow 0$. In order to construct approximate solutions to $(4.1)_{\pm}$, we formally put $\sigma = 0$ in $(4.1)_{\pm}$. The resulting equations are

$$\begin{aligned}
&\hat{f}^{0,\pm}_u\Phi^{\pm}_0 + \hat{f}^{0,\pm}_v\Psi^{\pm}_0 = 0, \\
&\delta_0(y)(\Psi^{\pm}_0)_{yy} + \hat{g}^{0,\pm}_u\Phi^{\pm}_0 + \hat{g}^{0,\pm}_v\Psi^{\pm}_0 = \mu(0)\Psi^{\pm}_0, \qquad y \in \mathbf{R}_{\pm}, \\
&\Psi^{\pm}_0(\pm\infty) = 0, \qquad \Psi^{\pm}_0(0) = b,
\end{aligned} \tag{4.5}_{\pm}$$

where $\hat{f}^{0,\pm}_u = f_u(U^{\pm}_0(y; \beta^*(d)), V^{\pm}_0(y; \beta^*(d)))$ with the outer solutions $(U^{\pm}_0, V^{\pm}_0)(y; \beta^*(d))$ obtained in § 2.1. $\hat{f}^{0,\pm}_v$, $\hat{g}^{0,\pm}_u$, and $\hat{g}^{0,\pm}_v$ are similarly defined. By (A3), the first equations in $(4.5)_{\pm}$ are written as

$$\Phi^{\pm}_0 = -\hat{f}^{0,\pm}_v\Psi^{\pm}_0/\hat{f}^{0,\pm}_u.$$

Then $(4.5)_{\pm}$ are reduced to the following equations:

$$\begin{aligned}
&\delta_0(y)(\Psi^{\pm}_0)_{yy} + \{H^{\pm}(y; d) - \mu(0)\}\Psi^{\pm}_0 = 0, \qquad y \in \mathbf{R}_{\pm}, \\
&\Psi^{\pm}_0(\pm\infty) = 0, \qquad \Psi^{\pm}_0(0) = b,
\end{aligned} \tag{4.6}_{\pm}$$

where $H^{\pm}(y; d) = (\hat{f}^{0,\pm}_u\hat{g}^{0,\pm}_v - \hat{f}^{0,\pm}_v\hat{g}^{0,\pm}_u)/\hat{f}^{0,\pm}_u$. The following lemma is very useful for our purposes.

LEMMA 4.1. *Consider the problems for* $\Pi^{\pm}$:

$$\begin{aligned}
&\delta_0(y)\Pi^{\pm}_{yy} + \{H^{\pm}(y; d) - \mu\}\Pi^{\pm} = 0, \qquad y \in \mathbf{R}_{\pm}, \\
&\Pi^{\pm}(\pm\infty) = 0, \qquad \Pi^{\pm}(0) = 1.
\end{aligned} \tag{4.7}_{\pm}$$

Then there exists a positive constant l_3 such that for any $\mu \in \mathbf{C}_{l_3} = \{\mu \in \mathbf{C} \mid \mathrm{Re}\,\{\mu\} \geqq -l_3\}$, the $(4.7)_\pm$ have unique solutions $\Pi^\pm(y; d; \mu)$.

The proof is given in the Appendix.

We thus find that, for any $\mu(0) \in \mathbf{C}_{l_3}$, the $(4.5)_\pm$ have unique solutions

$$\Psi_0^\pm(y; d; \mu(0); b) = b\Pi^\pm(y; d; \mu(0))$$

and

$$\Phi_0^\pm(y; d; \mu(0); b) = -\hat{f}_v^{0,\pm}\Psi_0^\pm(y; d; \mu(0); b)/\hat{f}_u^{0,\pm}.$$

By using these functions $(\Phi_0^\pm, \Psi_0^\pm)$, we now construct solutions $(\Phi^\pm, \Psi^\pm)$ of $(4.1)_\pm$. However, $\Phi_0^\pm(y; d; \mu(0); b)$ do not satisfy the boundary conditions at $y = 0$, so that we must modify these by adding other approximate solutions in a neighborhood of $y = 0$. That is, we introduce correction terms $\Gamma_0^\pm(y/\sigma)$ such that $(\Phi_0^\pm + \Gamma_0^\pm, \Psi_0^\pm)$ become approximate solutions to $(4.1)_\pm$ uniformly on $\mathbf{R}_\pm$. Substituting $(\Phi_0^\pm + \Gamma_0^\pm, \Psi_0^\pm)$ into $(4.1)_\pm$ and setting $\sigma = 0$, we have the following problems:

$$(4.8)_\pm \quad \begin{aligned} &\delta_0(\xi)(\Gamma_0^\pm)_{\xi\xi} + f_u(h_\pm(\beta^*(d)) + W_0^\pm, \beta^*(d))\Gamma_0^\pm = bP^\pm(\xi; d), \qquad \xi \in \mathbf{R}, \\ &\Gamma_0^\pm(\pm\infty) = 0, \qquad \Gamma_0^\pm(0) = 1 - bh'_\pm(\beta^*(d)), \end{aligned}$$

where $\xi = y/\sigma$, $W_0^\pm$ are the inner solutions obtained in § 2.2. With the representation

$$\begin{aligned} P^\pm(\xi; d) = h'_\pm(\beta^*(d))\{&f_u(h_\pm(\beta^*(d)), \beta^*(d)) - f_u(h_\pm(\beta^*(d)) + W_0^\pm, \beta^*(d))\} \\ &+ \{f_v(h_\pm(\beta^*(d)), \beta^*(d)) - f_v(h_\pm(\beta^*(d)) + W_0^\pm, \beta^*(d))\}, \end{aligned}$$

solutions of $(4.8)_\pm$ are represented as

$$\begin{aligned} \Gamma_0^\pm(\xi; d; b) = &\{1 - bh'_\pm(\beta^*(d))\}\frac{(W_0^\pm)_\xi(\xi)}{(W_0^\pm)_\xi(0)} \\ &- b(W_0^\pm)_\xi(\xi)\int_0^\xi ((W_0^\pm)_\eta(\eta))^{-2}\int_\eta^{\pm\infty}(W_0^\pm)_\zeta(\zeta)P^\pm(\zeta; d)\,d\zeta\,d\eta. \end{aligned}$$

Using the functions $(\hat{\Phi}_0^\pm, \hat{\Psi}_0^\pm)$ defined by

$$\hat{\Phi}_0^\pm(y; \sigma; d; \mu(0); b) = \Phi_0^\pm(y; d; \mu(0); b) + \Gamma_0^\pm\left(\frac{y}{\sigma}; d; b\right),$$

$$\hat{\Psi}_0^\pm(y; \sigma; d; \mu(0); b) = \Psi_0^\pm(y; d; \mu(0); b),$$

we can obtain solutions $(\Phi^\pm, \Psi^\pm)(y; \sigma; d; \mu(\sigma); b)$ of $(4.1)_\pm$ for any $\mu(\sigma)$ satisfying $\mu(0) \in \mathbf{C}_{l_3}$, which satisfy

$$\|\Phi^\pm - \hat{\Phi}_0^\pm\|_{\hat{X}^2_{\rho,\sigma}(\mathbf{R}_\pm)} + \|\Psi^\pm - \hat{\Psi}_0^\pm\|_{\hat{X}^2_{\rho,1}(\mathbf{R}_\pm)} \to 0 \quad \text{as } \sigma \downarrow 0.$$

The proof can be done by using standard singular perturbation techniques (see § 2 or [IM1]). By (3.12) and $(4.2)_\pm$, (3.8) is represented as

$$\begin{aligned} g(\sigma; d; \lambda(\sigma)) &= g(\sigma; d; \sigma\mu(\sigma)) \\ &= -\{(\Gamma_0^+)_\xi(0; d; 0) - (\Gamma_0^-)_\xi(0; d; 0)\} \\ &\quad \times \{(\Psi_0^+)_y(0; d; \mu(0); 1) - (\Psi_0^-)_y(0; d; \mu(0); 1)\} + o(1) \end{aligned}$$

as $\sigma \downarrow 0$. From the condition (2.10), it follows that

$$\begin{aligned} (\Gamma_0^+)_\xi(0; d; 0) - (\Gamma_0^-)_\xi(0; d; 0) &= \{(W_0^+)_{\xi\xi}(0) - (W_0^-)_{\xi\xi}(0)\}/(W_0^-)_\xi(0) \\ &= (d-1)f(\alpha^*(d), \beta^*(d))/(W_0^-)_\xi(0) \\ &\neq 0. \end{aligned} \tag{4.9}$$

Here, with

$$\mu = \mu^R + i\mu^I \quad \text{and} \quad \Psi_0^{\pm} = P^{\pm} + iQ^{\pm}$$

we define $G(d; \mu)$ by

$$G(d; \mu) = (\Psi_0^+)_y(0; d; \mu; 1) - (\Psi_0^-)_y(0; d; \mu; 1)$$

and show $G(d; \mu) \neq 0$ for any $\mu \in \mathbf{C}_{l_3}$. We rewrite $(4.6)_{\pm}$ with $b = 1$ as the following problems for $(P^{\pm}, Q^{\pm})(y; d; \mu; 1)$:

$$(4.10)_{\pm} \qquad \begin{aligned} &\delta_0(y)P^{\pm}_{yy} + \{H^{\pm}(y; d) - \mu^R\}P^{\pm} + \mu^I Q^{\pm} = 0, \qquad y \in \mathbf{R}_{\pm}, \\ &P^{\pm}(\pm\infty) = 0, \qquad P^{\pm}(0) = 1 \end{aligned}$$

and

$$(4.11)_{\pm} \qquad \begin{aligned} &\delta_0(y)Q^{\pm}_{yy} + \{H^{\pm}(y; d) - \mu^R\}Q^{\pm} - \mu^I P^{\pm} = 0, \qquad y \in \mathbf{R}_{\pm}, \\ &Q^{\pm}(\pm\infty) = 0, \qquad Q^{\pm}(0) = 0. \end{aligned}$$

Multiplying $(4.10)_{\pm}$ by $Q^{\pm}$ and integrating these over the intervals $\mathbf{R}_{\pm}$, we have

$$(4.12)_{\pm} \qquad \begin{aligned} &-\int_0^{\pm\infty} \delta_0(y)(P^{\pm})_y(Q^{\pm})_y \, dy + \int_0^{\pm\infty} \{H^{\pm}(y; d) - \mu^R\}P^{\pm}Q^{\pm} \, dy \\ &\quad + \mu^I \int_0^{\pm\infty} (Q^{\pm})^2 \, dy = 0. \end{aligned}$$

Similarly, multiplying $(4.11)_{\pm}$ by $P^{\pm}$ and integrating these over the intervals $\mathbf{R}_{\pm}$, we obtain

$$(4.13)_{\pm} \qquad \begin{aligned} &-\delta_0(\pm 0)(Q^{\pm})_y(0; d; \mu; 1) - \int_0^{\pm\infty} \delta_0(y)(P^{\pm})_y(Q^{\pm})_y \, dy \\ &\quad + \int_0^{\pm\infty} \{H^{\pm}(y; d) - \mu^R\}P^{\pm}Q^{\pm} \, dy - \mu^I \int_0^{\pm\infty} (P^{\pm})^2 \, dy = 0. \end{aligned}$$

Combining $(4.12)_{\pm}$ with $(4.13)_{\pm}$, we know that

$$\mu^I \int_0^{\pm\infty} \{(P^{\pm})^2 + (Q^{\pm})^2\} \, dy = -\delta_0(\pm 0)(Q^{\pm})_y(0; d; \mu; 1).$$

We thus find

$$\begin{aligned} \operatorname{Im}\{G(d; \mu)\} &= (Q^+)_y(0; d; \mu; 1) - (Q^-)_y(0; d; \mu; 1) \\ &= -\mu^I \left[\frac{1}{\delta_0(-0)} \int_{-\infty}^{0} \{(P^-)^2 + (Q^-)^2\} \, dy \right. \\ &\qquad \left. + \frac{1}{\delta_0(+0)} \int_0^{+\infty} \{(P^+)^2 + (Q^+)^2\} \, dy \right], \end{aligned}$$

which implies that $G(d; \mu) \neq 0$ if $\mu^I \neq 0$.

Next we consider the case $\mu^I = 0$; that is, μ is a real number. Letting $R^{\pm}(y; d; \mu) = (\partial/\partial\mu)\Psi_0^{\pm}(y; d; \mu; 1)$, we obtain the following equations:

$$(4.14)_{\pm} \qquad \begin{aligned} &\delta_0(y)R^{\pm}_{yy} + \{H^{\pm}(y; d) - \mu\}R^{\pm} = \Psi_0^{\pm}, \qquad y \in \mathbf{R}_{\pm}, \\ &R^{\pm}(\pm\infty) = 0, \qquad R^{\pm}(0) = 0. \end{aligned}$$

Multiplying $(4.14)_{\pm}$ by $\Psi_0^{\pm}$ and integrating these over $\mathbf{R}_{\pm}$, we have

$$-\delta_0(\pm 0)(R^{\pm})_y(0; d; \mu) = \int_0^{\pm\infty} (\Psi_0^{\pm})^2 \, dy.$$

Thus, we know that

$$\begin{aligned}\frac{d}{d\mu} G(d; \mu) &= \frac{\partial}{\partial \mu}(\Psi_0^+)_y(0; d; \mu; 1) - \frac{\partial}{\partial \mu}(\Psi_0^-)_y(0; d; \mu; 1) \\ &= -\left[\frac{1}{\delta_0(-0)} \int_{-\infty}^{0} (\Psi_0^-)^2 \, dy + \frac{1}{\delta_0(+0)} \int_0^{+\infty} (\Psi_0^+)^2 \, dy\right] \\ &< 0\end{aligned}$$

holds for any $\mu > -l_3$. On the other hand, we obtain

$$\begin{aligned}G(d; 0) &= (\Psi_0^+)_y(0; d; 0; 1) - (\Psi_0^-)_y(0; d; 0; 1) \\ &= \{(V_0^+)_{yy}(0; \beta^*(d)) - (V_0^-)_{yy}(0; \beta^*(d))\}/(V_0^-)_y(0; \beta^*(d)) \\ &= \{-g(h_+(\beta^*(d)), \beta^*(d))/\delta_0(+0) \\ &\quad + g(h_-(\beta^*(d)), \beta^*(d))/\delta_0(-0)\}/(V_0^-)_y(0; \beta^*(d)) \\ &< 0\end{aligned}$$

(see § 2.1). Combining these inequalities, we find that

$$G(d; \mu) \neq 0 \quad \text{for any } \mu > -l_3$$

(if necessary, we choose the positive constant l_3 to be smaller). Therefore we find that

$$G(d; \mu) \neq 0 \quad \text{for any } \mu(\sigma) \text{ satisfying } \mu(0) \in \mathbf{C}_{l_3} \tag{4.15}$$

so that from (4.9)

$$g(\sigma; d; \lambda(\sigma)) \neq 0$$

holds for any $\lambda(\sigma) = \sigma\mu(\sigma)$ satisfying $\mu(0) \in \mathbf{C}_{l_3}$, when σ is sufficiently small.

Case II. $\sigma^2\omega(\sigma) \downarrow 0$ and $\omega(\sigma) \uparrow \infty$ as $\sigma \downarrow 0$. Since $(4.1)_{\pm}$ and $(4.4)_{\pm}$ are quite similar forms, applying the same method as that in Case I to $(4.4)_{\pm}$, we can conclude that

$$g(\sigma; d; \lambda(\sigma)) \neq 0$$

for any $\lambda(\sigma)$ satisfying $\sigma\lambda(\sigma) \downarrow 0$ and $\lambda(\sigma)/\sigma \uparrow \infty$ as $\sigma \downarrow 0$.

Case III. $\sigma^2\omega(\sigma) \to \kappa$ as $\sigma \downarrow 0$. For this case, $(4.4)_{\pm}$ fall into regularly perturbed problems, because the coefficients of $\Phi_{zz}^{\pm}$ do not degenerate as $\sigma \downarrow 0$. First we construct approximate solutions to $(4.4)_{\pm}$. Using the transformation $\xi = z/\sqrt{\kappa}$ and putting $\sigma = 0$ in $(4.4)_{\pm}$, we obtain the following systems:

$$\begin{aligned}&\delta_0(\xi)(\Phi_0^{\pm})_{\xi\xi} + f_u(h_{\pm}(\beta^*(d)) + W_0^{\pm}, \beta^*(d))\Phi_0^{\pm} \\ &\qquad + f_v(h_{\pm}(\beta^*(d)) + W_0^{\pm}, \beta^*(d))\Psi_0^{\pm} = \kappa\hat{\mu}(0)\Phi_0^{\pm}, \qquad \xi \in \mathbf{R}_{\pm}, \\ &\delta_0(\xi)(\Psi_0^{\pm})_{\xi\xi} = \kappa\hat{\mu}(0)\Psi_0^{\pm}, \\ &\Phi_0^{\pm}(\pm\infty) = 0, \qquad \Phi_0^{\pm}(0) = 1, \\ &\Psi_0^{\pm}(\pm\infty) = 0, \qquad \Psi_0^{\pm}(0) = b.\end{aligned} \tag{4.16}_{\pm}$$

Noting that $\hat{\mu}(0)$ satisfies

$$\operatorname{Re}\{\hat{\mu}(0)\} \geqq 0 \quad \text{and} \quad |\hat{\mu}(0)| \neq 0, \tag{4.17}$$

we find that

$$\Psi_0^{\pm}(\xi; d; \hat{\mu}(0); b) = b \exp\{\mp \sqrt{\kappa\hat{\mu}(0)/\delta_0(\pm 0)}\, \xi\}.$$

Therefore, $(4.16)_{\pm}$ are reduced to

$$(4.18)_{\pm} \qquad \begin{aligned} &\delta_0(\xi)(\Phi_0^{\pm})_{\xi\xi} + \{f_u(h_{\pm}(\beta^*(d)) + W_0^{\pm}, \beta^*(d)) - \kappa\hat{\mu}(0)\}\Phi_0^{\pm} \\ &\quad = -f_v(h_{\pm}(\beta^*(d)) + W_0^{\pm}, \beta^*(d))\Psi_0^{\pm}, \qquad \xi \in \mathbf{R}_{\pm}, \\ &\Phi_0^{\pm}(\pm\infty) = 0, \qquad \Phi_0^{\pm}(0) = 1. \end{aligned}$$

Applying results similar to Lemma 4.1 to the systems

$$\begin{aligned} &\delta_0(\xi)\Pi_{\xi\xi}^{\pm} + \{f_u(h_{\pm}(\beta^*(d)) + W_0^{\pm}, \beta^*(d)) - \kappa\mu\}\Pi^{\pm} = 0, \qquad \xi \in \mathbf{R}_{\pm}, \\ &\Pi^{\pm}(\pm\infty) = 0, \qquad \Pi^{\pm}(0) = 1, \end{aligned}$$

we find that $(4.18)_{\pm}$ have unique solutions $\Phi_0^{\pm}(\xi; d; \hat{\mu}(0); b)$ for any $\hat{\mu}(0) \in \mathbf{C}$ satisfying (4.17). According to a standard perturbation process, for any $\hat{\mu}(0)$ satisfying (4.17), we can find exact solutions $(\Phi^{\pm}, \Psi^{\pm})(z; \sigma; d; \hat{\mu}(\sigma); b)$ of $(4.4)_{\pm}$ which satisfy

$$\|\Phi^{\pm}(z) - \Phi_0^{\pm}(z/\sqrt{\kappa})\|_{\hat{X}^2_{\rho,1}(\mathbf{R}_{\pm})} + \|\Psi^{\pm}(z) - \Psi_0^{\pm}(z/\sqrt{\kappa})\|_{\hat{X}^2_{\rho,1}(\mathbf{R}_{\pm})} \to 0 \quad \text{as } \sigma \downarrow 0.$$

In a way similar to that in Case I, we obtain

$$\begin{aligned} g(\sigma; d; \lambda(\sigma)) &= g(\sigma; d; \sigma\omega(\sigma)\hat{\mu}(\sigma)) \\ &= -\frac{\sigma\omega(\sigma)}{\kappa}[\{(\Phi_0^+)_{\xi}(0; d; \hat{\mu}(0); 0) - (\Phi_0^-)_{\xi}(0; d; \hat{\mu}(0); 0)\} \\ &\qquad \times \{(\Psi_0^+)_{\xi}(0; d; \hat{\mu}(0); 1) - (\Psi_0^-)_{\xi}(0; d; \hat{\mu}(0); 1)\} + o(1)] \\ &= \frac{\sigma\omega(\sigma)}{\kappa}\left[\{(\Phi_0^+)_{\xi}(0; d; \hat{\mu}(0); 0) - (\Phi_0^-)_{\xi}(0; d; \hat{\mu}(0); 0)\} \right. \\ &\qquad \left. \times \left\{\sqrt{\frac{\kappa\hat{\mu}(0)}{\delta_0(+0)}} + \sqrt{\frac{\kappa\hat{\mu}(0)}{\delta_0(-0)}}\right\} + o(1)\right] \end{aligned}$$

as $\sigma \downarrow 0$. Put

$$\begin{aligned} \hat{g}(\sigma; d; \mu) &= \{(\Phi_0^+)_{\xi}(0; d; \mu; 0) - (\Phi_0^-)_{\xi}(0; d; \mu; 0)\} \\ &\qquad \times \{\sqrt{\kappa\mu/\delta_0(+0)} + \sqrt{\kappa\mu/\delta_0(-0)}\} + o(1). \end{aligned}$$

Since $\hat{g}(\sigma; d; \mu)$ can be extended continuously so as to be defined for $\sigma = 0$, we obtain

$$\begin{aligned} \hat{g}(0; d; \mu) &= \{(\Phi_0^+)_{\xi}(0; d; \mu; 0) - (\Phi_0^-)_{\xi}(0; d; \mu; 0)\} \\ &\qquad \times \{\sqrt{\kappa\mu/\delta_0(+0)} + \sqrt{\kappa\mu/\delta_0(-0)}\}. \end{aligned}$$

In a way similar to that in Case I, we know that

$$\hat{g}(\sigma; d; \mu) \neq 0 \quad \text{for } \operatorname{Im}\{\mu\} \neq 0.$$

That is, $\hat{g}(\sigma; d; \mu) = 0$ has no roots μ satisfying $\operatorname{Im}\{\mu\} \neq 0$. Then it suffices to consider the case when μ is a real number; that is, for any $\mu > 0$, we find roots μ of $\hat{g}(0; d; \mu) = 0$. For any $\mu > 0$, $\hat{g}(0; d; \mu) = 0$ is equivalent to the equation

$$G(d; \mu) \equiv (\Phi_0^+)_{\xi}(0; d; \mu; 0) - (\Phi_0^-)_{\xi}(0; d; \mu; 0) = 0.$$

Therefore, we consider $G(d;\mu)=0$ instead of $\hat{g}(0;d;\mu)=0$. By the same method as that in Case I, we can show that

$$\frac{d}{d\mu}G(d;\mu)<0 \quad \text{for any } \mu>0,$$

and

$$G(d;\mu)=-\sqrt{\kappa\mu/\delta_0(+0)}-\sqrt{\kappa\mu/\delta_0(-0)}+O(1)\to-\infty \quad \text{as } \mu\uparrow\infty.$$

Moreover it holds that

$$\begin{aligned}G(d;0)&=(\Phi_0^+)_\xi(0;d;0;0)-(\Phi_0^-)_\xi(0;d;0;0)\\&=\{(W_0^+)_{\xi\xi}(0;\alpha^*(d),\beta^*(d))\\&\qquad-(W_0^-)_{\xi\xi}(0;\alpha^*(d),\beta^*(d))\}/(W_0^-)_\xi(0;\alpha^*(d),\beta^*(d))\\&=(d-1)f(\alpha^*(d),\beta^*(d))/(W_0^-)_\xi(0;\alpha^*(d),\beta^*(d)).\end{aligned}$$

Using the relations (2.14) and (2.15), we conclude that

$$\operatorname{sign}\{G(d;0)\}=-\mathbf{I}(d).$$

Then we directly have the following.

LEMMA 4.2. *If* $\mathbf{I}(d)=1$, $\hat{g}(0;d;\mu)\neq 0$ *for any* $\mu>0$. *Conversely, if* $\mathbf{I}(d)=-1$, *there exists a unique simple positive number* $\mu^*(d)$ *satisfying* $\hat{g}(0;d;\mu^*(d))=0$.

Applying the implicit function theorem to $\hat{g}(\sigma;d;\mu)=0$, it follows from Lemma 4.2 that

(i) When $\mathbf{I}(d)=1$, $\hat{g}(\sigma;d;\hat{\mu}(\sigma))\neq 0$ for any $\hat{\mu}(\sigma)\in\mathbf{C}$ satisfying (4.17);

(ii) When $\mathbf{I}(d)=-1$, $\hat{g}(\sigma;d;\hat{\mu}(\sigma))=0$ has a unique simple positive root $\hat{\mu}(d;\sigma)$ satisfying (4.17) such that $\lim_{\sigma\to 0}\hat{\mu}(d;\sigma)=\mu^*(d)$.

Therefore, if $\mathbf{I}(d)=1$, $g(\sigma;d;\lambda(\sigma))\neq 0$ holds for any $\lambda(\sigma)=\sigma\omega(\sigma)\hat{\mu}(\sigma)$ satisfying (4.17), while, if $\mathbf{I}(d)=-1$, $g(\sigma;d;\lambda(\sigma))=0$ has a unique simple positive root $\lambda=\lambda(\sigma)=\sigma\omega(\sigma)\hat{\mu}(d;\sigma)$.

Case IV. $\sigma^2\omega(\sigma)\uparrow\infty$ as $\sigma\downarrow 0$. In the same spirit of the proof of Lemma 4.3 in Case I, we can prove that

$$g(\sigma;d;\lambda(\sigma))\neq 0$$

for any $\lambda(\sigma)$ satisfying $\sigma\lambda(\sigma)\uparrow\infty$ as $\sigma\downarrow 0$. So we leave the proof to the reader.

Combining the above four cases, we are able to complete the proof of Theorem 3.1.

5. Concluding remarks. In the previous sections, we have not mentioned the case where σ is sufficiently large. Here under the assumptions (A1) and (A5), we briefly state the results on the stability as well as the existence of stationary solutions for this case. Let $v=h(u)$ be the function uniquely defined by the relation $g(u,h(u))=0$ (see (A5)) and let $S=\int_{u_-}^{u_+}f(u,h(u))\,du$. In [IM1], under the assumptions (A1) and (A5), we have shown the following existence result:

(i) When $S<0$, there exists d_1 (<1) such that (1.7) has two stationary solutions $(\bar{U}_*,\bar{V}_*)$ and $(\underline{U}_*,\underline{V}_*)(\bar{U}_*>\underline{U}_*)$ for any $d\in(0,d_1)$;

(ii) When $S>0$, there exists d_2 (>1) such that (1.7) has two stationary solutions $(\bar{U}^*,\bar{V}^*)$ and $(\underline{U}^*,\underline{V}^*)(\bar{U}^*>\underline{U}^*)$ for any $d\in(d_2,\infty)$ (see Fig. 9).

Using the same method as that in §§3 and 4, we can show that when $S<0$, the upper branch corresponding to $(\bar{U}_*,\bar{V}_*)$ is stable, while the lower one corresponding to $(\underline{U}_*,\underline{V}_*)$ is unstable. Conversely, when $S>0$, the lower one corresponding to $(\underline{U}^*,\underline{V}^*)$ is stable, whereas the upper one corresponding to $(\bar{U}^*,\bar{V}^*)$ is unstable.

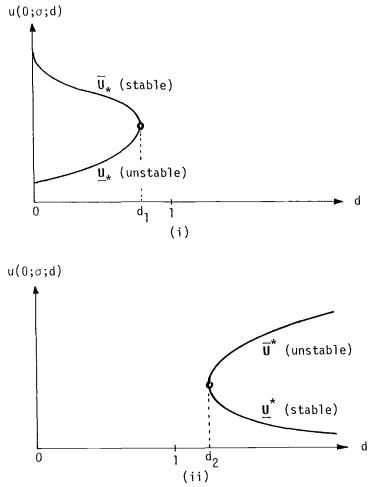


FIG. 9. *Global pictures of stationary solutions of* (1.1) *and their stability*; (i) $S<0$; (ii) $S>0$.

In this paper, motivated by the wave-blocking phenomena, we have discussed the stability property of stationary solutions for the special case when the reaction rates of u and v are totally different; in other words, σ is very large or small. Unfortunately, we are not able to discuss here the case when the reaction rates are almost equal. This is a future problem.

Appendix.

Proof of Proposition 3.1. The location of the essential spectrum for the operator $\mathscr{L}^{\varepsilon,d}$ is contained in the union of the following two sets:

$$S^{\pm}=\{\lambda\in\mathbf{C}\,|\,\det(-\nu^2D^{\pm}+N^{\pm}-\lambda I)=0,\ \nu\in\mathbf{R}\},$$

where

$$D^{\pm}=\begin{bmatrix}\delta_0(\pm\infty) & 0\\ 0 & \delta_0(\pm\infty)\end{bmatrix},\qquad N^{\pm}=\begin{bmatrix}f_u^{\pm}/\sigma & f_v^{\pm}/\sigma\\ \sigma g_u^{\pm} & \sigma g_v^{\pm}\end{bmatrix},$$

$f_u^{\pm}=f_u(u_{\pm},v_{\pm})$ and $f_v^{\pm}$, $g_u^{\pm}$, and $g_v^{\pm}$ are similarly defined. From the relations

$$\det(-\nu^2D^{\pm}+N^{\pm}-\lambda I)=0,$$

it follows that

$$\lambda^2+\left(2\nu^2\delta_0(\pm\infty)-\frac{1}{\sigma}f_u^{\pm}-\sigma g_v^{\pm}\right)\lambda+\left(\nu^2\delta_0(\pm\infty)-\frac{1}{\sigma}f_u^{\pm}\right)(\nu^2\delta_0(\pm\infty)-\sigma g_v^{\pm})-f_v^{\pm}g_u^{\pm}=0.$$

The roots of these equations are given by

$$\lambda=\frac{1}{2}\left[-2\nu^2\delta_0(\pm\infty)+\frac{1}{\sigma}f_u^{\pm}+\sigma g_v^{\pm}+\sqrt{\left(\frac{1}{\sigma}f_u^{\pm}-\sigma g_v^{\pm}\right)^2+4f_v^{\pm}g_u^{\pm}}\right]$$

or

$$\lambda=\frac{1}{2}\left[-2\nu^2\delta_0(\pm\infty)+\frac{1}{\sigma}f_u^{\pm}+\sigma g_v^{\pm}-\sqrt{\left(\frac{1}{\sigma}f_u^{\pm}-\sigma g_v^{\pm}\right)^2+4f_v^{\pm}g_u^{\pm}}\right].$$

Using the assumptions (A3) and (A4), we find that for small $\sigma>0$,

$$\begin{aligned}\operatorname{Re}\lambda&\leqq\frac{1}{2}\left[-2\nu^2\delta_0(\pm\infty)+\frac{1}{\sigma}f_u^{\pm}+\sigma g_v^{\pm}+\left|\frac{1}{\sigma}f_u^{\pm}-\sigma g_v^{\pm}\right|\right]\\&=-\nu^2\delta_0(\pm\infty)+\sigma g_v^{\pm}.\end{aligned}$$

That is, for any $\nu \in \mathbf{R}$ and small $\sigma > 0$

$$\operatorname{Re} \lambda \leqq \sigma g_{\nu}^{\pm}$$

hold. Then we can conclude Proposition 3.1 with $l_1 = -\max\{g_{\nu}^{+}, g_{\nu}^{-}\}(>0)$. □

Proof of Lemma 4.1. Proving Lemma 4.1 is equivalent to showing the following.

LEMMA A1. *Consider the homogeneous problems*

$$(A1)_{\pm} \qquad \begin{aligned} &\delta_0(y)(\Pi_0^{\pm})_{yy} + \{H^{\pm}(y; d) - \mu\}\Pi_0^{\pm} = 0, \qquad y \in \mathbf{R}_{\pm}, \\ &\Pi_0^{\pm}(\pm\infty) = 0, \qquad \Pi_0^{\pm}(0) = 0. \end{aligned}$$

Then there exists a positive constant l such that for any $\mu \in \mathbf{C}_l$, *the* $(A1)_{\pm}$ *have only trivial solutions.*

Proof. Let $L^{\pm}$ be linear operators defined by

$$L^{\pm}u^{\pm} \equiv \delta_0(y)u_{yy}^{\pm} + H^{\pm}(y; d)u^{\pm}.$$

According to the Sturm-Liouville theory, all the eigenvalues of $(A1)_{\pm}$ are real and simple. Furthermore, it is known that the eigenfunctions $u_0^{\pm}$ corresponding to the largest eigenvalues $\mu_0^{\pm}$ have no zeros in $\mathbf{R}_{\pm}$, respectively. Without loss of generality, we assume $u_0^{\pm} \geqq 0$ in $\mathbf{R}_{\pm}$. That is,

$$(A2)_{\pm} \qquad \begin{aligned} &\delta_0(y)(u_0^{\pm})_{yy} + \{H^{\pm}(y; d) - \mu_0^{\pm}\}u_0^{\pm} = 0, \qquad y \in \mathbf{R}_{\pm}, \\ &u_0^{\pm}(\pm\infty) = 0, \qquad u_0^{\pm}(0) = 0 \end{aligned}$$

hold. On the other hand, the derivatives of the outer solutions $V_0^{\pm}$, say $P^{\pm} \equiv (V_0^{\pm})_y > 0$, satisfy the following equations:

$$(A3)_{\pm} \qquad \begin{aligned} &\delta_0(y)P_{yy}^{\pm} + H^{\pm}(y; d)P^{\pm} = 0, \qquad y \in \mathbf{R}_{\pm}, \\ &P^{\pm}(\pm\infty) = 0, \qquad P^{\pm}(0) = (V_0^{\pm})_y(0; \beta^*(d)). \end{aligned}$$

Using the above relations $(A2)_{\pm}$ and $(A3)_{\pm}$, we obtain

$$\begin{aligned} \mu_0^{\pm} \int_0^{\pm\infty} u_0^{\pm} P^{\pm}\, dy &= \int_0^{\pm\infty} \{\delta_0(y)(u_0^{\pm})_{yy} + H^{\pm}(y; d)u_0^{\pm}\}P^{\pm}\, dy \\ &= -\delta_0(\pm 0)(u_0^{\pm})_y(0)P^{\pm}(0) + \int_0^{\pm\infty} \{\delta_0(y)P_{yy}^{\pm} + H^{\pm}(y; d)P^{\pm}\}u_0^{\pm}\, dy \\ &= -\delta_0(\pm 0)(u_0^{\pm})_y(0)P^{\pm}(0). \end{aligned}$$

Then

$$\mu_0^{\pm} = -\delta_0(\pm 0)(u_0^{\pm})_y(0)P^{\pm}(0) \Big/ \int_0^{\pm\infty} u_0^{\pm} P^{\pm}\, dy$$

hold. Noting that $(u_0^{+})_y(0) > 0$, $(u_0^{-})_y(0) < 0$, $P^{\pm}(0) > 0$ and $u_0^{\pm}(y)P^{\pm}(y) \geqq 0$ in $\mathbf{R}_{\pm}$, we find that

$$\mu_0^{\pm} < 0.$$

Thus if we put $l = \max\{\mu_0^{-}, \mu_0^{+}\}$, Lemma A1 is proved. □

REFERENCES

[AGJ] J. ALEXANDER, R. GARDNER, AND C. JONES, *A topological invariant arising in the stability analysis of travelling waves*, preprint.

[CH] M. CHIPOT AND J. K. HALE, *Stable equilibria with variable diffusion*, in Proc. Nonlinear Partial Differential Equations Conference, Durham, NC, June 20-26, 1982, AMS Contemporary Mathematics Series, American Mathematical Society, Providence, RI, 1983.

[Ec] W. ECKHAUS, *Asymptotic Analysis of Singular Perturbation*, North-Holland, Amsterdam, 1979.

[Ev1] J. W. EVANS, *Nerve axon equations, III: stability of the nerve impulse*, Indiana Univ. Math. J., 22 (1972), pp. 577-593.

[Ev2] ———, *Nerve axon equations, IV: the stable and the unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169-1190.

[FH] G. FUSCO AND J. K. HALE, *Stable equilibria in a scalar parabolic equation with variable diffusion*, SIAM J. Math. Anal., 16 (1985), pp. 1152-1164.

[FP] P. C. FIFE AND L. A. PELETIER, *Clines induced by variable selection and migration*, Proc. Roy. Soc. London. Ser. B, 214 (1981), pp. 99-123.

[GJ] R. GARDNER AND C. JONES, *Stability of traveling wave solutions of diffusive predator-prey systems*, preprint.

[H] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 80, Springer-Verlag, Berlin, New York, 1981.

[I] H. IKEDA, *Singular perturbation approach to stability properties of traveling wave solutions of reaction-diffusion systems*, Hiroshima Math. J., 19 (1989), pp. 587-630.

[IM1] H. IKEDA AND M. MIMURA, *Wave-blocking phenomena in bistable reaction-diffusion systems*, SIAM J. Appl. Math., 49 (1989), pp. 515-538.

[IM2] ———, *Multiple traveling wave solutions in some bistable reaction-diffusion systems*, in preparation.

[K] J. P. KEENER, *Causes of propagation failure in excitable media*, in Temporal Disorder in Human Oscillatory Systems, L. Rensing, U. Heiden, and M. Mackey, eds., Springer-Verlag, Berlin, New York, 1987, pp. 134-140.

[M] H. MATANO, *Convergence of solutions of one-dimensional semilinear parabolic equations*, J. Math. Kyoto Univ., 18 (1978), pp. 224-243.

[P] J. P. PAUWELUSSEN, *Nerve impulse propagation in a branching nerve system: a simple model*, Phys. D, 4 (1981), pp. 67-88.

[Y] E. YANAGIDA, *Stability of stationary distributions in a space-dependent population growth process*, J. Math. Biol., 15 (1982), pp. 37-50.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1679-1701, November 1991

009

INTEGRAL EQUATION METHODS IN A QUASI-PERIODIC DIFFRACTION PROBLEM FOR THE TIME-HARMONIC MAXWELL'S EQUATIONS*

J. C. NEDELEC† AND F. STARLING†

Abstract. The problem of the time-harmonic Maxwell's equations is considered in the exterior in R^3 of a domain which has a doubly periodic structure in a plane of R^3. It is proved that except for a discrete set of value of the frequency, this problem has a unique solution.

Key words. Maxwell equations, doubly periodic domain, integral equation

AMS(MOS) subject classifications. 45B05, 45P05, 78A95, 78A25

Introduction. There is an increasing interest in the study of radiation patterns created by the new periodic structures of phased arrays antennas. These antennas consist in the superposition of a great number of identical electromagnetic horns. The calculus of the radiation pattern of such a horn, when it is considered separately from the others, can be handled by integral equations techniques (see [1]). However, the complete calculus by this method, taking account of all mutual influences, leads to a great number of unknowns and consequently to very heavy computations. Here, we present a mathematical model of such a problem that shall lead us, later on, to its numerical resolution. It is based on a reduction to a "quasi-periodic" formulation (see [4]), which enables us to focus the analysis of one "elementary scatterer." On that starting point, we modify the methods in [1] to get existence and uniqueness results, which shall justify a rough numerical method. New mathematical difficulties, which differ from those found in [1], appear in our analysis and are to be be related to the kind of model discussed.

The paper is organized as follows. In § 1, we introduce the mathematical model and framework to be considered. Section 2 is devoted to the analysis of uniqueness properties for this model. Finally, § 3 gives the integral equation formulation of the problem, which provides existence results and will lead to numerical computations.

1. The mathematical model. This section will lead us to a precise mathematical framework for our electromagnetic diffraction problem. We begin with a general description of the geometrical setting of the problem.

1.1. Geometrical structure and notation. If Y and Y' are elements of $\mathbb{C}^\nu$, $Y = (y_1, \cdots, y_\nu)$, $Y' = (y'_1, \cdots, y'_\nu)$, we will use the following general notation:

$\bar{Y} := (\bar{y}_1, \cdots, \bar{y}_\nu)$ is the complex conjugate of Y,
$Y.\ Y' := y_1 y'_1 + \cdots + y_\nu y'_\nu$ is the scalar product of Y, Y',
$Y.\ \overline{Y'} := y_1 \bar{y}'_1 + \cdots + y_\nu \bar{y}'_\nu$ is the Hermitian scalar product of Y, Y',
$|Y| := (Y.\ \bar{Y})^{1/2} = (y_1 \bar{y}_1 + \cdots + y_\nu \bar{y}_\nu)^{1/2}$ is the modulus of Y.

Let $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ be an orthonormal basis of $\mathbb{R}^3$, related to a system of coordinates (x_1, x_2, x_3).

Let ω be a bounded open regular set of $\mathbb{R}^3$, with connected complement in $\mathbb{R}^3$; ω will be our basic scatterer, repeated periodically in a plane of $\mathbb{R}^3$.

Indeed, we will consider a "doubly periodic" scatterer, consisting in the union of identical scatterers, translated from ω in the following way.

* Received by the editors May 2, 1988; accepted for publication (in revised form) July 4, 1990.
† Centre de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau Cedex, France.

We assume that (d_1, d_2) are two positive reals such that

$$\forall I = (i_1, i_2) \in \mathbb{Z}^2 \quad \text{if } \omega_I := \omega + i_1 d_1 \mathbf{e}_1 + i_2 d_2 \mathbf{e}_2;$$

then

$$\omega_1 \cap \omega_j = \varnothing \quad \text{if } I \neq J, \quad J \in \mathbb{Z}^2. \tag{1.1}$$

Our doubly periodic scatterer is defined by the disjoint union

$$S := \bigcup_{J \in \mathbb{Z}^2} \omega_J$$

(d_1, d_2) hence represents the double periodicity of the structure. (See Fig. 1.) This geometrical structure leads us to the following splitting in the system of coordinates, which will widely simplify the notation: we will write

$$(x_1, x_2, x_3) = (X, Z)$$

where X stands for the $\mathbb{R}^2$ element (x_1, x_2) and Z stands for the $\mathbb{R}$ element x_3.

Property (1.1) implies that we can choose an origin O in $\mathbb{R}^3$ such that ω is completely embedded in the "cylindrical" set of $\mathbb{R}^3$:

$$\Omega :=]-d_1/2, d_1/2[\times]-d_2/2, d_2/2[\times \mathbb{R}.$$

We will call this set the "elementary cell."

We will see in § 1.4 that the diffraction problem, which is set on $\mathbb{R}^3 \backslash \bar{S}$, can be reduced to a boundary value problem which is set only in the complement of ω with respect to the "elementary cell" Ω. To this end, we define such an exterior domain by

$$\Omega^{\text{ext}} := \Omega \backslash \bar{\omega},$$

and we will also need to truncate this domain Ω^{ext} to obtain some compactness properties in § 2; thus, we set, for every real ρ,

$$\Omega_\rho := \Omega \cap \{(X, Z) \in \mathbb{R}^d / |Z| < \rho\},$$

$$\Omega_\rho^{\text{ext}} := \Omega^{\text{ext}} \cap \{(X, Z) \in \mathbb{R}^d / |Z| < \rho\}.$$

As $\bar{\omega}$ is bounded, we can assume that it is embedded in a set

$$\{(X, Z) \in \mathbb{R}^3 / |Z| < \rho_0\}$$

for a positive real ρ_0. It follows that Ω_ρ contains $\bar{\omega}$ if $\rho \geqq \rho_0$.

We also define sections Σ^ρ of the "elementary cell" Ω by setting

$$\Sigma^\rho :=]-d_1/2, d_1/2[\times]-d_2/2, d_2/2[\times \{Z \in \mathbb{R}^3 / |Z| = \rho\}$$

for every real $\rho \geqq \rho_0$.

Finally, we set $Q :=]-d_1/2, d_1/2[\times]-d_2/2, d_2/2[$.

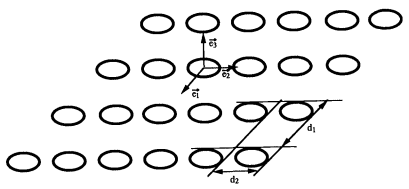


FIG. 1

1.2. Maxwell's equations. Generally speaking, the determination of the time-harmonic diffracted field by a perfectly conducting body gives rise to the following kind of boundary value problem, where the time dependence, in $\exp(-i\omega_0 t)$ (with $i=\sqrt{-1}$), is dropped by linearity:

(Max. 0) Find $(\mathbf{e}, \mathbf{h})$ such that

$$\begin{aligned} &\operatorname{curl}\mathbf{e} - i\omega_0\mu\mathbf{h} = \mathbf{0} && \text{in the exterior domain,}\\ &\operatorname{curl}\mathbf{h} + i\omega_0\varepsilon\mathbf{e} = \mathbf{0} && \text{in the exterior domain,}\\ &\mathbf{n}\times\mathbf{e} = -\mathbf{n}\times\mathbf{e}^{\mathrm{inc}} && \text{on the conducting body.}\end{aligned}$$

$(\mathbf{e}, \mathbf{h})$ satisfies an outgoing wave condition at infinity.

Here $\mathbf{e}^{\mathrm{inc}}$ is the electric part of the incident electromagnetic field; k, the wave number of the phenomenon, is related to the pulsation of the incident wave ω_0 by the formula $k=\omega_0\sqrt{\varepsilon\mu}$ with ε and μ the electromagnetic characteristics of the exterior domain; $\mathbf{n}$ is the outwardly directed normal on the body.

Following [1], we can eliminate $\mathbf{h}$ from these equations to obtain the equivalent problem:

(Max. 1) Find $\mathbf{e}$ such that

$$\begin{aligned} &\Delta\mathbf{e} + k^2\mathbf{e} = \mathbf{0} && \text{in the exterior domain,}\\ &\operatorname{div}\mathbf{e} = 0 && \text{in the exterior domain,}\\ &\mathbf{n}\times\mathbf{e} = -\mathbf{n}\times\mathbf{e}^{\mathrm{inc}} && \text{on the conducting body.}\end{aligned}$$

$\mathbf{e}$ satisfies an outgoing wave condition at infinity.

In this form, the diffraction problem has been widely analysed and its mathematical properties are well known (see [1], [2]; part of our study will be based on this knowledge). However, our doubly periodic scattering problem has not, strictly speaking, a physical meaning because we assume that the scatterer is of infinite extent. Hence, to give a precise meaning to problems (Max. 0) and (Max 1), we must add some specific conditions on the behaviour of the diffracted fields, and also specify an appropriate outgoing wave condition, generalizing the usual ones.

First of all, we restrict our study to the case where the incident field has a plane wave structure. Other kinds of excitations of such a system are possible. Thus, we write, in a fixed system of coordinates (X, Z),

$$\mathbf{e}^{\mathrm{inc}}(X, Z) = \mathbf{e}_0^{\mathrm{inc}} \exp(i\mathbf{K}\cdot X)\exp(i\mathbf{K}_z\cdot Z) \tag{1.2}$$

where the wave vector of $\mathbf{e}^{\mathrm{inc}}$ is $\mathbf{k}=\mathbf{K}+\mathbf{K}_z$, $|\mathbf{k}|=k$, and $\mathbf{K}\in\operatorname{Vect}(\mathbf{e}_1, \mathbf{e}_2)$, $\mathbf{K}_z\in\operatorname{Vect}(\mathbf{e}_3)$.

Consequently, the tangential component of our diffracted field, which is matched with the incident field on conductors, will satisfy the following "quasi-periodicity" condition:

For all $J=(j_1, j_2)\in\mathbb{Z}^2$, if $X_J := (j_1 d_1, j_2 d_2)$,
$(\mathbf{n}\times\mathbf{e})(X+X_J, Z) = (\mathbf{n}\times\mathbf{e})(X, Z)\exp(i\mathbf{K}\cdot X_J)$.

This suggests searching for a diffracted field $\mathbf{e}$ satisfying such a "quasi-periodicity" condition not only on the conductors, and for the tangential component of the field, but everywhere in $\mathbb{R}^3\backslash\bar{S}$ and for the whole field $\mathbf{e}$. In fact, we can guess that this requirement is necessary, for, assuming that the solution of our problem is in some sense unique, an X_J translation of the scatterer gives rise to the two following kinds of solutions:

$$\mathbf{e}^J(X, Z) = \mathbf{e}(X+X_J, Z) \quad \text{(translation of the scatter),}$$

$$\mathbf{e}^*(X, Z) = \mathbf{e}(X, Z)\exp(i\mathbf{K}\cdot X_J) \quad \text{(modification of the incident wave),}$$

so that writing $\mathbf{e}^J(X, Z) = \mathbf{e}^*(X, Z)$, we find again our "quasi-periodicity" condition:

$$\text{For all } J \text{ in } \mathbb{Z}^2, \quad \mathbf{e}(X + X_J, Z) = \mathbf{e}(X, Z) \exp(i\mathbf{K} \cdot X_J). \tag{1.3}$$

We can now state our "quasi-periodic" boundary value problem, in a more specific form:

(1.4) Find $\mathbf{e}$ defined on $\mathbb{R}^3 \backslash \bar{S}$ such that

$$\begin{aligned} &\Delta \mathbf{e} + k^2 \mathbf{e} = \mathbf{0} && \text{in } \mathbb{R}^3 \backslash \bar{S}, \\ &\operatorname{div} \mathbf{e} = 0 && \text{in } \mathbb{R}^3 \backslash \bar{S}, \\ &\mathbf{n} \times \mathbf{e} = -\mathbf{n} \times \mathbf{e}^{\text{inc}} && \text{on each conductor.} \end{aligned}$$

$\mathbf{e}$ satisfies (1.3) and a suitable outgoing wave condition at infinity.

Here $\mathbf{e}^{\text{inc}}$ has the form (1.2).

At this point, a precise meaning for the outgoing wave condition can be derived. It is easily seen that a function $\mathbf{e}$ satisfying conditions (1.4) can be expanded in a "quasi-periodic" Fourier series for sufficiently large $|Z|$:

$$\mathbf{e}(X, Z) = \sum_{J \in \mathbb{Z}^2} \exp(i(\mathbf{K} + \mathbf{K}_J) \cdot X) \mathbf{e}_J(Z), \qquad |Z| > \rho_0 \tag{1.5}$$

where

$$\mathbf{K}_J := (2\pi j_1 / d_1, 2\pi j_2 / d_2), \qquad J = (j_1, j_2)$$

and

$$\mathbf{e}_J(Z) := \frac{1}{d_1 d_2} \int_Q \mathbf{e}(X, Z) \exp(-i(\mathbf{K} + \mathbf{K}_J) \cdot X) \, d\sigma_X.$$

This series (1.5) is absolutely convergent and termwise infinitely differentiable, for $\mathbf{e}$, solving problem (1.4), is necessarily very regular.

Furthermore, each component $\mathbf{e}_J(\cdot)$ must satisfy

$$\frac{\partial^2}{\partial Z^2} \mathbf{e}_J(Z) + (k^2 - |\mathbf{K} + \mathbf{K}_J|^2) \mathbf{e}_J(Z) = \mathbf{0} \quad \text{for } |Z| > \rho_0. \tag{1.6}$$

The general solution of (1.6) can, of course, be given explicitly; however, having in mind that (1.6) is a Helmholtz equation set in a classical framework, we will select the different kinds of solutions by the usual Sommerfeld's radiation conditions:

$$\begin{aligned} &\mathbf{e}_J(Z) = O(1) \quad \text{for large } |Z| \text{ and all } J \text{ in } \mathbb{Z}^2, \\ &\frac{\partial}{\partial |Z|} \mathbf{e}_J(Z) - i\mu_J(k) \mathbf{e}_J(Z) = o(1) \quad \text{for large } |Z| \text{ and all } J \text{ in } \mathbb{Z}^2 \end{aligned} \tag{1.7}$$

where

$$\begin{aligned} &\mu_J(k) := (k^2 - |\mathbf{K} + \mathbf{K}_J|^2)^{1/2} && \text{if } k \geqq |\mathbf{K} + \mathbf{K}_J|, \\ &\mu_J(k) := i(|\mathbf{K} + \mathbf{K}_J|^2 - k^2)^{1/2} && \text{if } k < |\mathbf{K} + \mathbf{K}_J|. \end{aligned} \tag{1.8}$$

1.3. Functional spaces. In this section, we describe the functional setting needed to state our boundary value problem in a suitable mathematical form.

We shall use standard notation for the usual functional spaces: $L^2(G)$ is the space of complex square integrable functions defined on G, where G stands either for an open regular set of $\mathbb{R}^3$, or for the regular boundary of such a domain. The norm and scalar product on $L^2(G)$ will be denoted, as usual, by $|\cdot|_{0,G}$ and $(\cdot|\cdot)_{0,G}$.

$C_0^\infty(G)$ (respectively, $C_0^\infty(\bar{G})$) is the space of regular functions defined on G (respectively, $\bar{G}$), with compact support in G (respectively, $\bar{G}$).

$H^s(G)$ is the usual Sobolev space of order s, $s \in \mathbb{R}$.

$H_\Delta^1(G)$ is the set of functions u in $H^1(G)$, with Δu in $L^2(G)$.

$L^{2,\mathrm{loc}}(G)$ (respectively, $H^{s,\mathrm{loc}}(G)$, $H_\Delta^{1,\mathrm{loc}}(G)$) is the set of functions which are in $L^2(G_R)$ (respectively, $H^s(G_R)$, $H_\Delta^1(G_R)$), where $G_R := G \cap \{(X, Z) \in \mathbb{R}^3 / |(X, Z)| < R\}$ for every positive R.

We need to define Sobolev spaces of "quasi-periodic" functions. To this end, let $C_K^\infty(\mathbb{R}^3)$ be the space of those scalar functions $u \in C^\infty(\mathbb{R}^3)$ satisfying:

(i) u has compact support in Z, i.e., Supp $u \subset \{|Z| < \rho\}$ for some real positive ρ.

(ii) $u(X + X_J, Z) = u(X, Z) \exp(i\mathbf{K} \cdot X_J)$ for all $J = (j_1, j_2) \in \mathbb{Z}^3$, with $X_J = (j_1 d_1, j_2 d_2)$.

For an open set $\vartheta \subset \mathbb{R}^3$, $C_K^\infty(\vartheta)$ will be the space of restrictions to ϑ of functions of $C_K^\infty(\mathbb{R}^3)$, and $C_{0K}^\infty(\vartheta)$ the subset of $C_K^\infty(\vartheta)$ of those functions with compact support in ϑ.

This enables us to consider the following "quasi-periodic" Sobolev space, closure of $C_K^\infty(\vartheta)$ in $H^1(\vartheta)$, i.e.,

$$H_K^1(\vartheta) := \overline{C_K^\infty(\vartheta)}^{H^1(\vartheta)}.$$

It is not difficult to see that, equipped with the usual $H^1(\vartheta)$ norm and Hilbertian scalar product denoted by $|\cdot|_{1,\vartheta}$ and $(\cdot|\cdot)_{1,\vartheta}$, $H_K^1(\vartheta)$ becomes a Hilbert space.

To reduce our boundary value problem to one set only in the "elementary cell" Ω, we need a space of functions, defined on Ω^{ext}, which can be extended to $\mathbb{R}^3 \backslash \bar{S}$ as "quasi-periodic" functions, in a sufficiently smooth way, with respect to conditions dictated by the partial differential equations. To this end we define:

-The "quasi-periodic" extension of $u \in L^{2,\mathrm{loc}}(\Omega^{\mathrm{ext}})$, say $u^{(K)} \in L^{2,\mathrm{loc}}(\mathbb{R}^3 \backslash \bar{S})$, defined by

$u^{(K)}$ is given in $\Omega^{\mathrm{ext}} + X_J$, for each J in $\mathbb{Z}^2$, by
$u^{(K)}(X + X_J, Z) := u(X, Z) \exp(i\mathbf{K} \cdot X_J)$, for almost every (X, Z) in Ω^{ext}.

-The space $H_{\Delta,K}^1(\Omega^{\mathrm{ext}})$, of functions which are in $H_K^1(\Omega^{\mathrm{ext}}) \cap H_\Delta^1(\Omega^{\mathrm{ext}})$, and for which the "quasi-periodic" extension $u^{(K)}$ belongs to $H_\Delta^{1,\mathrm{loc}}(\mathbb{R}^3 \backslash S)$. This space, equipped with the natural norm $\|.\|_{\Delta,\Omega^{\mathrm{ext}}}$ given by

$$\|u\|_{\Delta,\Omega^{\mathrm{ext}}}^2 = |\Delta u|_{0,\Omega^{\mathrm{ext}}}^2 + \|u\|_{1,\Omega^{\mathrm{ext}}}^2,$$

is a Hilbert space. Two facts which will be useful in the sequel can be shown: first, $H_{\Delta,K}^1(\Omega^{\mathrm{ext}})$ can be given a weak characterization by means of a Green formula (see proof of Proposition 1.1); second, $C_K^\infty(\Omega^{\mathrm{ext}})$ is a dense subset of the Hilbert space $H_{\Delta,K}^1(\Omega^{\mathrm{ext}})$.

We will also use the Fréchet spaces of functions that are locally in $H_K^1(\Omega^{\mathrm{ext}})$ or in $H_{\Delta,K}^1(\Omega^{\mathrm{ext}})$, in the same sense as above. They will be, as above, designated by $H_K^{1,\mathrm{loc}}(\Omega^{\mathrm{ext}})$ or $H_{\Delta,K}^{1,\mathrm{loc}}(\Omega^{\mathrm{ext}})$.

For vector-valued functions, defined on an open set ϑ of $\mathbb{R}^3$, we will use the following notation. $F(\vartheta, \mathbb{C}^3)$ will always stand for the space of complex fields $\mathbf{e} := (e^1, e^2, e^3)$, with $e^i \in F(\vartheta)$, where F is any of the functional spaces described above. Note that with our choice of a basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, we can identify $F(\vartheta, \mathbb{C}^3)$ with $\{F(\vartheta)\}^3$.

For vector fields defined on $\partial\omega$, we will write $\mathbf{H}^s(\partial\omega) := \{H^2(\partial\omega)\}^3$, and following [1], we will split $\mathbf{H}^s(\partial\omega)$ in two subspaces of tangential fields and normal fields to $\partial\omega$: if $\mathbf{n}$ is the unit normal vector of $\partial\omega$, we have

$$\mathbf{e} := \Pi\mathbf{e} + (\mathbf{e} \cdot \mathbf{n})\mathbf{n}$$

where Π is the normal projection on the tangent plane to $\partial\omega$. Thus we can set

$$TH^s(\partial\omega) := \{\mathbf{e} \in \mathbf{H}^s(\partial\omega)/(\mathbf{e}\cdot\mathbf{n}) = 0\},$$
$$NH^s(\partial\omega) := \{\mathbf{e} \in \mathbf{H}^s(\partial\omega)/\mathbf{e} = \mathbf{e}\mathbf{n},\ e \in H^s(\partial\omega)\},$$

and so

$$\mathbf{H}^s(\partial\omega) = TH^s(\partial\omega) \oplus NH^s(\partial\omega).$$

For the norms, we will use the standard Sobolev spaces notation, noting that all the new "quasi-periodic" spaces that we have defined are Hilbert spaces when equipped with the norms issued from the usual Sobolev spaces.

Finally, we recall some facts about Green formulae and traces. For sufficiently smooth $\mathbf{e}$ and $\mathbf{e}^*$, defined on a regular open set ϑ of $\mathbb{R}^3$, the following formulae hold:

$$\begin{aligned}\int_{\vartheta} [\Delta\mathbf{e}\cdot\overline{\mathbf{e}^*} + \operatorname{curl}\mathbf{e}\cdot\overline{\operatorname{curl}\mathbf{e}^*} + \operatorname{div}\mathbf{e}\,\overline{\operatorname{div}\mathbf{e}^*}]\,dv \\ = \int_{\partial\vartheta} [(\operatorname{curl}\mathbf{e}\times\mathbf{n})\cdot\overline{\mathbf{e}^*} + \gamma_0 \operatorname{div}\mathbf{e}(\mathbf{n}\cdot\overline{\mathbf{e}^*})]\,d\sigma,\end{aligned} \tag{1.9}$$

$$\int_{\vartheta} \Delta\mathbf{e}\cdot\overline{\mathbf{e}^*}\,dv = -\sum_{\alpha=1}^{3}\int_{\vartheta} \nabla e^{\alpha}\cdot\overline{\nabla e^{*\alpha}}\,dv + \int_{\partial\vartheta} \gamma_1\mathbf{e}\cdot\overline{\gamma_0\mathbf{e}}\,d\sigma. \tag{1.10}$$

We now give a precise formulation of these properties in our "q.p." framework.

PROPOSITION 1.1. *Let* $\mathbf{e}$ *and* $\mathbf{e}^*$ *belong to* $C_K^{\infty}(\mathbb{R}^3, \mathbb{C}^3)$. *Then* (1.9) *and* (1.10) *can be written*

$$\begin{aligned}\int_{\Omega^{\mathrm{ext}}} [\Delta\mathbf{e}\cdot\overline{\mathbf{e}^*} + \operatorname{curl}\mathbf{e}\cdot\overline{\operatorname{curl}\mathbf{e}^*} + \operatorname{div}\mathbf{e}\cdot\overline{\operatorname{div}\mathbf{e}^*}]\,dv \\ = \int_{\partial\omega} [(\operatorname{curl}\mathbf{e}\times\mathbf{n})\cdot\overline{\mathbf{e}^*} + \gamma_0 \operatorname{div}\mathbf{e}(\mathbf{n}\cdot\overline{\mathbf{e}^*})]\,d\sigma,\end{aligned} \tag{1.11}$$

$$\int_{\Omega^{\mathrm{ext}}} \Delta\mathbf{e}\cdot\overline{\mathbf{e}^*}\,dv = -\sum_{\alpha=1}^{3}\int_{\Omega^{\mathrm{ext}}} \nabla e^{\alpha}\cdot\overline{\nabla e^{*\alpha}}\,dv + \int_{\partial\omega} \gamma_1\mathbf{e}\cdot\overline{\gamma_0\mathbf{e}^*}\,d\sigma. \tag{1.12}$$

Moreover, these formulae can be extended to $\mathbf{e}$ *in* $H^1_{\Delta,K}(\Omega^{\mathrm{ext}}, \mathbb{C}^3)$ *and* $\mathbf{e}^*$ *in* $H^1_K(\Omega^{\mathrm{ext}}, \mathbb{C}^3)$. (*The integrals over* $\partial\omega$ *are then to be considered in a duality sense.*)

Remark. As mentioned earlier, we will show in fact that (1.12) leads to the following characterization of $H^1_{\Delta,K}(\Omega^{\mathrm{ext}}, \mathbb{C}^3)$:

$$\begin{aligned}H^1_{\Delta,K}(\Omega^{\mathrm{ext}}, \mathbb{C}^3) = \{\mathbf{e} \in H^1_K(\Omega^{\mathrm{ext}}, \mathbb{C}^3) \cap H^1_{\Delta}(\Omega^{\mathrm{ext}}, \mathbb{C}^3)/ \\ (1.12) \text{ holds for every } \mathbf{e}^* \in H^1_K(\Omega^{\mathrm{ext}}, \mathbb{C}^3)\}.\end{aligned}$$

Proof. $\partial\Omega^{\mathrm{ext}}$ can be written in the following form:

$$\partial\Omega^{\mathrm{ext}} = \partial\omega \cup \partial\Omega \quad \text{where } \partial\Omega \text{ is a lateral boundary, i.e.,}$$

$$\partial\Omega = \{-d_1/2, d_1/2\}\times[-d_2/2, d_2/2]\times\mathbb{R} \cup \{-d_2/2, d_2/2\}\times[-d_1/2, d_1/2]\times\mathbb{R}.$$

It is easy to see that "q.p." conditions for regular functions imply that the sum over $\partial\Omega$ vanishes. To extend these results to $H^1_{\Delta,K}(\Omega^{\mathrm{ext}}, \mathbb{C}^3)$ and $H^1_K(\Omega^{\mathrm{ext}}, \mathbb{C}^3)$ functions, we have first to give a precise meaning for the traces on $\partial\omega$ and $\partial\Omega$:

Using a compactly supported smooth extension of the normal field $\mathbf{n}$, defined on $\partial\omega$, to a neighbourhood of $\partial\omega$, with the property that $|\mathbf{n}| \equiv 1$ near $\partial\omega$, for $\mathbf{e}$ in $H^1_K(\Omega^{\mathrm{ext}}, \mathbb{C}^3) \cap H^1_{\Delta}(\Omega^{\mathrm{ext}}, \mathbb{C}^3)$, we can define the traces

$$\gamma_0 \operatorname{curl}\mathbf{e}\times\mathbf{n} \quad \text{in } TH^{-1/2}(\partial\omega), \qquad \gamma_1(\boldsymbol{n}\cdot E) \quad \text{in } H^{-1/2}(\partial\omega).$$

Here $\gamma_1(\cdot)$ stands for the second trace or normal derivative of a function defined in a neighbourhood of $\partial\omega$. As usual, for $\mathbf{e}$ in $H^1_K(\Omega^{\text{ext}}, \mathbb{C}^3) \cap H^1_\Delta(\Omega^{\text{ext}}, \mathbb{C}^3)$, we can define the traces

$$\gamma_0 \operatorname{div} \mathbf{e} \quad \text{in } H^{-1/2}(\partial\omega), \qquad \gamma_1 \mathbf{e} \quad \text{in } H^{-1/2}(\partial\omega).$$

Note that only the local regularity near $\partial\omega$ is needed to define the traces above.

In an equivalent way, we can define the same kind of traces on $\partial\Omega$:

$$\gamma_0 \operatorname{curl} \mathbf{e} \times \mathbf{n} \quad \text{in } TH^{-1/2}(\partial\Omega), \qquad \gamma_1(\mathbf{n} \cdot E) \quad \text{in } H^{-1/2}(\partial\Omega),$$

$$\gamma_0 \operatorname{div} \mathbf{e} \quad \text{in } H^{-1/2}(\partial\Omega), \qquad \gamma_1 \mathbf{e} \quad \text{in } H^{-1/2}(\partial\Omega).$$

The presence of "corners" and the infinite extent of this boundary is handled by a suitable localization of trace definitions.

Formulae (1.9) and (1.10) are used in the setting of the above definitions, and are naturally extended for functions in $H^1_K(\Omega^{\text{ext}}, \mathbb{C}^3) \cap H^1_\Delta(\Omega^{\text{ext}}, \mathbb{C}^3)$. We have to prove that if $\mathbf{e}$ is, moreover, in $H^1_{\Delta,K}(\Omega^{\text{ext}}, \mathbb{C}^3)$ then there is no contribution from $\partial\Omega$. In fact, (1.11) (respectively, (1.12)) gives the following characterization of $H^1_{\Delta,K}(\Omega^{\text{ext}}, \mathbb{C}^3)$:

$$H^1_{\Delta,K}(\Omega^{\text{ext}}, \mathbb{C}^3) = \{\mathbf{e} \in H^1_K(\Omega^{\text{ext}}, \mathbb{C}^3) \cap H^1_\Delta(\Omega^{\text{ext}}, \mathbb{C}^3)/(1.11) \text{ (respectively, (1.12)) holds for every } \mathbf{e}^* \text{ in } H^1_K(\Omega^{\text{ext}}, \mathbb{C}^3)\}.$$

To see this, we write

$$\int_{\partial\Omega} \gamma_1 \mathbf{e} \cdot \overline{\mathbf{e}^*}\, d\sigma = \int_{\Gamma_1} [\gamma_1 \mathbf{e}^{(K)}]_1 \cdot \overline{\mathbf{e}^*}\, d\sigma + \int_{\Gamma_2} [\gamma_1 \mathbf{e}^{(K)}]_2 \cdot \overline{\mathbf{e}^*}\, d\sigma$$

where

$\Gamma_1 := \{-d_1/2\} \times]-d_2/2, d_2/2[\times \mathbb{R}$,

$\Gamma_2 :=]-d_1/2, d_1/2[\times \{-d_2/2\} \times \mathbb{R}$,

$[\cdot]_i$ is the jump across Γ_i of functions with traces on these boundaries:

$$[\mathbf{u}]_1 := \mathbf{u}((-d_1/2)^+, x_2, x_3) - \mathbf{u}((-d/2)^-, x_2, x_3),$$

$$[\mathbf{u}]_2 := \mathbf{u}(x_1, (-d_2/2)^+, x_3) - \mathbf{u}(x_1, (-d_1/2)^-, x_3).$$

The + or − signs represent the two sides of each boundary $\partial\Omega$, oriented by the axes $(0, \mathbf{e}_1)$ and $(0, \mathbf{e}_2)$ (see Fig. 2).

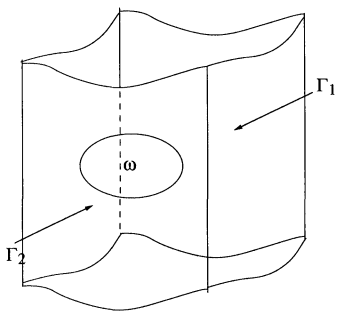


FIG. 2

It is not difficult to see that $\mathbf{e}^{(K)}$ is in $H^1_\Delta(\mathbb{R}^3\backslash\bar{S}, \mathbb{C}^3)$ if and only if the jumps $[\gamma_1\mathbf{e}^{(K)}]$ vanish, yielding the required result. □

These formulae are also valid if we take Ω^{ext}_ρ instead of Ω^{ext} but we then must consider contributions from Σ^ρ.

We now come to some facts involving traces and differential geometry tools. The following formulae can be used for smooth $\mathbf{e}$, $\mathbf{e}^*$:

$$\gamma_0 \operatorname{div} \mathbf{e} = \operatorname{div}_{\partial\omega}(\Pi\mathbf{e}) + 2H\mathbf{e}\cdot\mathbf{n} + \gamma_1(\mathbf{e}\cdot\mathbf{n}), \tag{1.13}$$

$$\mathbf{n}\cdot\mathbf{e}^*\gamma_0 \operatorname{div}\mathbf{e} = 2H\mathbf{e}\cdot\mathbf{e}^* + \gamma_1\mathbf{e}\cdot\overline{\gamma_0\mathbf{e}^*} \tag{1.14}$$

when $\mathbf{e}$ and $\mathbf{e}^*$ satisfy $\Pi\mathbf{e} = \Pi\mathbf{e}^* = \mathbf{0}$. Here, $\operatorname{div}_{\partial\omega}(\ )$ is the superficial divergence of a tangential field to $\partial\omega$, Π is the normal projection operator on the tangential plane of $\partial\omega$, and H is the mean value of the curvature at each point of the smooth surface $\partial\omega$. These formulae can be extended, by density, to fields $\mathbf{e}$ in $H^1_K(\Omega^{\text{ext}}, \mathbb{C}^3) \cap H^1_\Delta(\Omega^{\text{ext}}, \mathbb{C}^3)$ and $\mathbf{e}^*$ in $H^1_K(\Omega^{\text{ext}}, \mathbb{C}^3)$.

For a detailed analysis of all these properties, we refer the reader to [1].

1.4. The boundary value problem. We now have all the elements needed to set a mathematical boundary value problem, describing the physical situation, and which can be handled by standard functional analysis tools, related to Sobolev spaces formulations.

Our "whole space" problem will be:

(1.5) Find $\mathbf{e}$ in $H^{1,\text{loc}}_K(\mathbb{R}^3\backslash\bar{S}, \mathbb{C}^3) \cap H^{1,\text{loc}}_\Delta(\mathbb{R}^3\backslash\bar{S}, \mathbb{C}^3)$ such that

$$\begin{aligned} &\Delta\mathbf{e} + k^2\mathbf{e} = 0 && \text{in } \mathbb{R}^3\backslash\bar{S},\\ &\operatorname{div}\mathbf{e} = 0 && \text{in } \mathbb{R}^3\backslash\bar{S},\\ &\mathbf{n}\times\mathbf{e} = -\mathbf{n}\times\mathbf{e}^{\text{inc}} && \text{on } \partial S. \end{aligned}$$

$\mathbf{e}$ satisfies outgoing wave conditions (1.7).

We show that this problem is equivalent to the following one, which is set only in the "elementary cell":

(1.16) Find $\mathbf{e}$ in $H^{1,\text{loc}}_{\Delta,K}(\Omega^{\text{ext}}, \mathbb{C}^3)$ such that

$$\begin{aligned} &\Delta\mathbf{e} + k^2\mathbf{e} = 0 && \text{in } \Omega^{\text{ext}},\\ &\operatorname{div}\mathbf{e} = 0 && \text{in } \Omega^{\text{ext}},\\ &\mathbf{n}\times\mathbf{e} = -\mathbf{n}\times\mathbf{e}^{\text{inc}} && \text{on } \partial\omega. \end{aligned}$$

$\mathbf{e}$ satisfies outgoing wave conditions (1.7).

Indeed, $\mathbf{e}$ satisfying (1.15) can be extended by "quasi-periodicity" and its "quasi-periodic" extension $\mathbf{e}^{(K)}$ is easily found to solve (1.16). Conversely, the restriction to Ω^{ext} of any solution of (1.16) is clearly in $H^{1,\text{loc}}_{\Delta,K}(\Omega^{\text{ext}}, \mathbb{C}^3)$, and so solves (1.15).

Our study will now be focused on the analysis of problem (1.19) or on a slight extension of it:

(1.17) Find $\mathbf{e}$ in $H^{1,\text{loc}}_{\Delta,K}(\Omega^{\text{ext}}, \mathbb{C}^3)$ such that

$$\begin{aligned} &\Delta\mathbf{e} + k^2\mathbf{e} = 0 && \text{in } \Omega^{\text{ext}},\\ &\operatorname{div}\mathbf{e} = 0 && \text{in } \Omega^{\text{ext}},\\ &\mathbf{n}\times\mathbf{e} = \mathbf{c} \in TH^{1/2}(\partial\omega). \end{aligned}$$

$\mathbf{e}$ satisfies outgoing wave conditions (1.7).

2. Uniqueness properties. Our purpose is to establish the following uniqueness theorem.

THEOREM 2.1. *The problem*:

(P_k) *Find* $\mathbf{e}$ *in* $H^{1,\mathrm{loc}}_{\Delta,K}(\Omega^{\mathrm{ext}})$ *such that*

$$\Delta\mathbf{e}+k^2\mathbf{e}=\mathbf{0} \quad \text{in } \Omega^{\mathrm{ext}},$$
$$\operatorname{div}\mathbf{e}=0 \quad \text{in } \Omega^{\mathrm{ext}},$$
$$\mathbf{n}\times\mathbf{e}=\mathbf{0} \quad \text{on } \partial\omega.$$

$\mathbf{e}$ *satisfies the outgoing wave conditions* (1.13)

can have nontrivial solutions only if k *belongs to a set* K_{sing} *of exceptional values.* K_{sing} *is a countable set of positive reals that can be ordered in a nondecreasing sequence* $(k_n)_{n\geq 0}$, *tending to* $+\infty$.

Theorem 2.1 will be proved after some preliminary work: we will show that we can identify the restriction to $\Omega^{\mathrm{ext}}_\rho$ of a solution of (P_k) with an eigenvector of an operator acting on $L^2(\Omega^{\mathrm{ext}}_\rho)$, for which a complete spectral decomposition can be done.

To see how the truncated problem must be set, we begin by an analysis of some properties of a solution of (P_k).

2.1. Properties of solutions of (P_k). We begin with the following proposition.

PROPOSITION 2.2. *Let* $\mathbf{e}$ *be a solution of* (P_k). *Then* $\mathbf{e}$ *belongs to* $C^\infty(\overline{\Omega^{\mathrm{ext}}})$, *and the following expansion holds for* $|Z|>\rho_0$:

$$\mathbf{e}(X,Z)=\sum_{J\in\mathbb{Z}^2}\exp\left(i(\mathbf{K}+\mathbf{K}_J)\cdot X\right)\mathbf{e}_J(Z)$$

where the $\mathbf{e}_J$*'s are given by*

$$\mathbf{e}_J(Z)=\mathbf{e}_J^{0+}\exp\left(i\mu_J Z\right), \qquad Z>\rho_0,$$
$$\mathbf{e}_J(Z)=\mathbf{e}_J^{0-}\exp\left(-i\mu_J Z\right), \qquad Z<-\rho_0.$$

This series is absolutely convergent and termwise infinitely differentiable.

Proof. Regularity properties are local properties; thus, it can be easily deduced from the usual regularity properties (cf. [1], [2], for example), that

$$\mathbf{e}\in C^\infty(\Omega\backslash\omega,\mathbb{C}^3) \quad (\text{regularity up to } \partial\omega).$$

Having in mind then that the "q.p." extension of $\mathbf{e}$, $\mathbf{e}^{(k)}$ satisfies the Helmholtz equation in $\mathbb{R}^3\backslash S$, and using elliptic regularity properties for this equation again, but in $\mathbb{R}^3\backslash S$, we see that regularity goes up to the lateral part of the boundary, that is to say, $\partial\Omega$, and so we have

$$\mathbf{e}\in C^\infty(\overline{\Omega^{\mathrm{ext}}},\mathbb{C}^3).$$

The expansion can then be established as in § 1, having in mind that the regularity properties of $\mathbf{e}$ ensure that the series has the requested convergence. □

LEMMA 2.3. *Let* $\mathbf{e}$ *solve* (P_k) *and set*

$$J_k^+ := \{J\in\mathbb{Z}^2/|\mathbf{K}+\mathbf{K}_J|\geq k\},$$
$$J_k^- := \{J\in\mathbb{Z}^2/|\mathbf{K}+\mathbf{K}_J|<k\}.$$

Then $\mathbf{e}_J(Z)=0$ *for* $J\in J_k^-$ *and* $|Z|>\rho_0$; *hence, we have* $\mathbf{e}\in H^1_{\Delta,K}(\Omega^{\mathrm{ext}},\mathbb{C}^3)$, *except the case where there is a* J *in* J_k^+ *such that* $|\mathbf{K}+\mathbf{K}_J|=k$.

Proof. We apply Green formula in its usual form (see § 1.3) to the functions $\mathbf{e}$ and $\mathbf{e}$ in the set $\Omega^{\mathrm{ext}}_\rho$, for an arbitrarily chosen $\rho>\rho_0$:

$$0=\int_{\Omega^{\mathrm{ext}}_\rho}(\Delta\mathbf{e}\cdot\bar{\mathbf{e}}-\Delta\mathbf{e}\cdot\mathbf{e})\,dv=\int_{\partial\omega\cup\Sigma^\rho}(\gamma_1\mathbf{e}\cdot\overline{\gamma_0\mathbf{e}}-\overline{\gamma_1\mathbf{e}}\cdot\gamma_0\mathbf{e})\,d\sigma. \tag{2.1}$$

We treat separately the contributions from $\partial\omega$ and Σ^ρ to the boundary integral in (2.1).

On $\partial\omega$, the first trace of div $\mathbf{e}$, γ_0 div $\mathbf{e}$ vanishes. The use of (1.17) then yields

$$0 = 2H|\mathbf{e}|^2 + \gamma_1\mathbf{e} \cdot \overline{\gamma_0\mathbf{e}}. \tag{2.2}$$

Hence $\gamma_1\mathbf{e} \cdot \overline{\gamma_0\mathbf{e}}$ is real and the contribution of $\partial\omega$ to (2.1) cancels.

We now consider the contribution from Σ^ρ to the boundary integral in (2.1); thanks to the expansion of Proposition 2.2, we have

$$\int_{\Sigma^\rho} \gamma_1\mathbf{e} \cdot \overline{\gamma_0\mathbf{e}}\, d\sigma = d_1 d_2 \sum_{J\in\mathbb{Z}^2} i\mu_J(|\mathbf{e}_J^{0+}|^2 + |\mathbf{e}_J^{0-}|^2) \exp(i(\mu_J - \overline{\mu_J})\rho).$$

Now, we see that

$$J \in J_k^+ \Rightarrow i\mu_J \exp(i(\mu_J - \overline{\mu_J})\rho) \in \mathbb{R}.$$

This, with (2.1), implies that

$$2d_1 d_2 \sum_{J\in J_k^-} i\mu_J(|\mathbf{e}_J^{0+}|^2 + |\mathbf{e}_J^{0-}|^2) = 0.$$

Thus we find that

$$|\mathbf{e}_J^{0+}| = |\mathbf{e}_J^{0-}| = 0 \quad \text{for all } J \text{ in } J_k^-.$$

Hence $\mathbf{e}_J(Z) = 0$ for all J in J_k^- and $|Z| > \rho_0$.

If there is no index $J \in \mathbb{Z}^2$ such that $|\mathbf{K} + \mathbf{K}_J| = k$, the fact that $\mathbf{e}$ is in $H^1_{\Delta,K}(\Omega^{\text{ext}})$ follows easily from the exponential decrease of $\mathbf{e}_J(Z)$ for J in J_k^+. □

Lemma 2.3 shows, in this last case, that a solution of (P_k) appears as an eigenfunction of the unbounded operator of $\mathbf{L}^2(\Omega^{\text{ext}}, \mathbb{C}^3)$, A, defined by

$$D(A) := \{\mathbf{e} \in H^1_{\Delta,K}(\Omega^{\text{ext}})/\mathbf{n}\times\mathbf{e} = \mathbf{0} \text{ on } \partial\omega;\ \gamma_0 \text{ div } \mathbf{e} = 0 \text{ on } \partial\omega\},$$

$$A\mathbf{e} := -\Delta\mathbf{e} \quad \text{for } \mathbf{e} \quad \text{in } D(A).$$

We will not study directly the spectrum of A. However, the main properties of A are summarized in the following proposition which will be given without its proof.

PROPOSITION 2.4. *A is a self-adjoint operator on $\mathbf{L}^2(\Omega^{\text{ext}}; C^3)$, with dense domain. Its spectrum $\Sigma(A)$ is such that*

$$\Sigma(A) \subset [0; +\infty[,$$

$$\Sigma_{\text{ess}}(A) = [m; +\infty[$$

where

$$m = \min_{J\in Z^2} \{|K + K_J|\}.$$

Theorem 2.1 states that A has at most a sequence tending to $+\infty$ of eigenvalues in its spectrum.

2.2. The truncated problem. We now turn to the study of a problem set in a truncated domain Ω_ρ^{ext} for $\rho \geqq \rho_0$. To this end, we introduce Hilbert spaces of "q.p." functions defined on Ω_ρ^{ext}, which are given by

$$C_K^\infty(\Omega_\rho^{\text{ext}}, \mathbb{C}^3) = \{\mathbf{e}_{|\Omega_\rho^{\text{ext}}} : \mathbf{e} \in C_K^\infty(\Omega^{\text{ext}}, \mathbb{C}^3)\},$$

$$H_K^1(\Omega_\rho^{\text{ext}}, \mathbb{C}^3) = \{\mathbf{e}_{|\Omega_\rho^{\text{ext}}} : \mathbf{e} \in H_K^1(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)\},$$

$$H^1_{\Delta,K}(\Omega_\rho^{\text{ext}}, \mathbb{C}^3) = \{\mathbf{e}_{|\Omega_\rho^{\text{ext}}} : \mathbf{e} \in H^1_{\Delta,K}(\Omega^{\text{ext}}, \mathbb{C}^3)\}.$$

These spaces are endowed with their natural norms. We begin by giving some trace properties on Σ^ρ for functions in $H^1_K(\Omega^{\text{ext}}_\rho, \mathbb{C}^3)$ and $H^1_{\Delta,K}(\Omega^{\text{ext}}_\rho, \mathbb{C}^3)$, which will enable us to set the appropriate truncation relations on Σ^ρ. The usual trace properties are slightly complicated by the decomposition of $\partial\Omega^{\text{ext}}_\rho$ in three parts of different "nature": $\partial\Omega^{\text{ext}}_\rho = \partial\omega \cup \Sigma^\rho \cup \Gamma_1$, where Γ_1 is the lateral part of the boundary i.e., $\Gamma_1 := \partial\Omega^{\text{ext}}_\rho \cap \partial\Omega$. (See Fig. 3.)

We show that we can extend a "q.p." version of the first trace γ_0, well defined for functions in $C^\infty_K(\Omega^{\text{ext}}_\rho, \mathbb{C}^3)$, to $H^1_K(\Omega^{\text{ext}}_\rho \mathbb{C}^3)$, with values in a fractionary order Sobolev space of "q.p." functions over Σ^ρ. We then get similar results for second traces of functions in $H^1_{\Delta,K}(\Omega^{\text{ext}}_\rho, \mathbb{C}^3)$, which will be defined in a fractionary order Sobolev space of "q.p." distributions on Σ^ρ.

Let $\mathbf{e}$ be in $C^\infty_K(\Omega^{\text{ext}}_\rho, \mathbb{C}^3)$ and φ be a $C^\infty(\mathbb{R})$ function, with $\varphi \equiv 1$ near $Z = \pm\rho$ vanishing for $|Z| < \rho_0$. Consider the "q.p." expansion of $\mathbf{e}$; we have

$$|\mathbf{e}_J(\rho)|^2 = 2\,\text{Re}\left(\int_{\rho_0}^{+\rho} \frac{\partial}{\partial z}\mathbf{e}_J(z)\cdot\overline{\mathbf{e}_J(z)}\varphi(z)\,dz\right) + \int_{\rho_0}^{+\rho}\frac{\partial}{\partial z}\varphi(z)|\mathbf{e}_J(z)|^2\,dz.$$

Hence there exist two constants C and C' such that

$$|\mathbf{e}_J(\rho)|^2 \leqq 2C|\mathbf{e}_J|_{L^2(\rho_0,\rho)} \times \left|\frac{\partial}{\partial z}\mathbf{e}_J\right|_{L^2(\rho_0,\rho)} + C'|\mathbf{e}_J|^2_{L^2(\rho_0,\rho)}.$$

So, multiplying this inequality by $|\mathbf{K}+\mathbf{K}_J|$ and adding the corresponding terms, we have

$$\sum_{J\in\mathbb{Z}^2}|\mathbf{K}+\mathbf{K}_J||\mathbf{e}_J(\rho)|^2 \leqq C\left(\sum_{J\in\mathbb{Z}^2}|\mathbf{K}+\mathbf{K}_J|^2|\mathbf{e}_J|^2_{L^2(\rho_0,\rho)} + \sum_{J\in\mathbb{Z}^2}\left|\frac{\partial}{\partial z}\mathbf{e}_J\right|^2_{L^2(\rho_0,\rho)}\right) + C'\sum_{J\in\mathbb{Z}^2}|\mathbf{K}+\mathbf{K}_J||\mathbf{e}_J|^2_{L^2(\rho_0,\rho)}.$$

Thus we conclude that

$$\sum_{J\in\mathbb{Z}^2}|\mathbf{K}+\mathbf{K}_J||\mathbf{e}_J(\rho)|^2 \leqq C''|\mathbf{e}|^2_{1,\Omega^{\text{ext}}_\rho}$$

and the same estimate holds for $Z = -\rho$.

For $\mathbf{e}$ in $L^2(\Sigma^\rho, \mathbb{C}^3)$ we set

$$|\mathbf{e}|^2_{1/2,\Sigma^\rho} := \sum_{J\in\mathbb{Z}^2}(1+|\mathbf{K}+\mathbf{K}_J|^2)^{1/2}(|\mathbf{e}^+_J|^2+|\mathbf{e}^-_J|^2) \tag{2.3}$$

where

$$\mathbf{e}^+_J \text{ (respectively, } \mathbf{e}^-_J) := \int_{\Sigma^\rho\cap\{z=+\rho\,(\text{respectively, } z=-\rho)\}} \mathbf{e}(X)\exp\left(-i(\mathbf{K}+\mathbf{K}_J)\cdot X\right)d\sigma_X$$

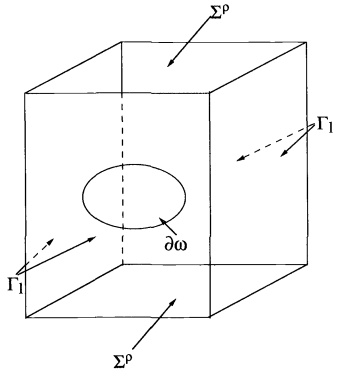


FIG. 3

and we define $H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3)$ as the space of functions $\mathbf{e}$ in $L^2(\Sigma^\rho, \mathbb{C}^3)$ for which $|\mathbf{e}|_{1/2,\Sigma^\rho} < +\infty$.

$H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3)$ can be given a Hilbert space structure, by means of formula (2.3). It is then clear that we can define a continuous trace γ_0 on Σ^ρ for functions of $H_K^1(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$, in the space $H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3)$, extending by density formula (2.3).

Now, for functions in $H_{\Delta,K}^1(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$ we carry out the same program and show that we can define a second trace in the space $H_K^{-1/2}(\Sigma^\rho, \mathbb{C}^3)$, defined similarly to $H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3)$ by

$$H_K^{-1/2}(\Sigma^\rho, \mathbb{C}^3) = \left\{ \mathbf{v} = \sum_{J \in Z^2} \mathbf{v}_J \exp\left(i(K + K_J) \cdot X\right) : \sum_{J \in Z^2} |\mathbf{v}_J|^2 (1 + |K + K_J|^2)^{-1/2} < +\infty \right\}$$

where the summation over Z^2 of the oscillating exponentials is to be understood in a distributional sense ("q.p." distributions on the "torus" Σ^ρ). $H_K^{-1/2}(\Sigma^\rho, \mathbb{C}^3)$ can be identified with $(H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3))'$, the dual space of $H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3)$, when $L^2(\Sigma^\rho, \mathbb{C}^3)$ is identified with its own dual. The (anti-)duality pairing between these spaces is obviously given by

$$\langle \mathbf{v}', \mathbf{v} \rangle_{-1/2,1/2} = \sum_{J \in Z^2} \mathbf{v}_J' \mathbf{v}_J.$$

Let $\mathbf{e}$ be in $C_K^\infty(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$ and take φ as above; we have

$$\left| \frac{\partial}{\partial z} \mathbf{e}_J(\rho) \right|^2 = 2 \operatorname{Re} \left(\int_{\rho_0}^{+\rho} \frac{\partial^2}{\partial z^2} \mathbf{e}_J(z) \cdot \frac{\partial}{\partial z} \overline{\mathbf{e}_J(z)} \varphi(z)\, dz \right) + \int_{\rho_0}^{+\rho} \frac{\partial}{\partial z} \varphi(z) \left| \frac{\partial}{\partial z} \mathbf{e}_J(z) \right|^2 dz.$$

Hence

$$\left| \frac{\partial}{\partial z} \mathbf{e}_J(\rho) \right|^2 \leqq 2C \left| \frac{\partial^2}{\partial z^2} \mathbf{e}_J \right|_{L^2(\rho_0,\rho)} \times \left| \frac{\partial}{\partial z} \mathbf{e}_J \right|_{L^2(\rho_0,\rho)} + C' \left| \frac{\partial}{\partial z} \mathbf{e}_J \right|^2_{L^2(\rho_0,\rho)}.$$

So, dividing this inequality by $p_J := (1 + |\mathbf{K} + \mathbf{K}_J|^2)^{1/2}$ and adding the corresponding terms, we have

$$\begin{aligned} \sum_{J \in \mathbb{Z}^2} \frac{1}{p_J} \left| \frac{\partial}{\partial z} \mathbf{e}_J(\rho) \right|^2 &\leqq C \left(\sum_{J \in \mathbb{Z}^2} \frac{1}{p_J^2} \left| \frac{\partial^2}{\partial z^2} \mathbf{e}_J \right|^2_{L^2(\rho_0,\rho)} + \sum_{J \in \mathbb{Z}^2} \left| \frac{\partial}{\partial z} \mathbf{e}_J \right|^2_{L^2(\rho_0,\rho)} \right) \\ &\quad + C' \sum_{J \in \mathbb{Z}^2} \frac{1}{p_J} |\mathbf{e}_J|^2_{L^2(\rho_0,\rho)}. \end{aligned} \tag{2.4}$$

As $\mathbf{e}$ is very regular, each one of the sums in the right member of (2.4) makes sense; we now try to get estimates in $H_{\Delta,K}^1(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$ as follows. We have

$$\begin{aligned} &\frac{1}{p_J^2} \left| \frac{\partial^2}{\partial z^2} \mathbf{e}_J \right|^2_{L^2(\rho_0,\rho)} \\ &\quad \leqq \frac{3}{2} \left(\frac{1}{p_J^2} \left| \frac{\partial^2}{\partial z^2} \mathbf{e}_J - |\mathbf{K} + \mathbf{K}_J|^2 \mathbf{e}_J \right|^2_{L^2(\rho_0,\rho)} + \frac{|\mathbf{K} + \mathbf{K}_J|^4}{p_J^2} |\mathbf{e}_J|^2_{L^2(\rho_0,\rho)} \right) \end{aligned}$$

and

$$\Delta \mathbf{e}(X, Z) = \sum_{J \in \mathbb{Z}^2} \left(\frac{\partial^2}{\partial z^2} \mathbf{e}_J - |\mathbf{K} + \mathbf{K}_J|^2 \mathbf{e}_J \right) \exp\left(i(\mathbf{K} + \mathbf{K}_J) \cdot X\right) \quad \text{for } \rho \geqq |Z| \leqq \rho_0,$$

is in $L^2(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$. Thus, we conclude that

$$\sum_{J \in \mathbb{Z}^2} (1 + |\mathbf{K} + \mathbf{K}_J|^2)^{-1/2} \left| \frac{\partial}{\partial z} \mathbf{e}_J(\rho) \right|^2 \leqq C''(|\mathbf{e}|^2_{1,\Omega_\rho^{\text{ext}}} + |\Delta \mathbf{e}|^2_{0,\Omega_\rho^{\text{ext}}})$$

and the same estimate holds for $Z = -\rho$.

This shows that the mapping $\mathbf{e} \to (\partial \mathbf{e}/\partial \mathbf{n}) = \gamma_1 \mathbf{e}$ can be continuously extended to $H^1_{\Delta,K}(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$ with values in $H_K^{-1/2}(\Sigma^\rho, \mathbb{C}^3)$. Here we have used the fact, pointed out in § 1.3, that $H^1_K(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$ is the closure of $C^\infty_K(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$ for the $\|.\|_{\Delta,\Omega_\rho^{\text{ext}}}$ norm.

We now come to the definition of an operator, acting on the space $H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3)$ with values in $(H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3))'$ given by the following proposition ($(\cdot)^+$ (respectively, $(\cdot)^-$), is the value on $Z = +\rho$ (respectively, $Z = -\rho$), of a function defined on Σ^ρ).

PROPOSITION 2.5. *Let $C^\infty_K(\Sigma^\rho, \mathbb{C}^3)$ be the set of regular* "q.p." *functions of Σ^ρ. Then T_k defined for $\mathbf{e}$ in $C^\infty_K(\Sigma^\rho, \mathbb{C}^3)$ by*

$$(T_k \mathbf{e})^+ = i \sum_{J \in J_k^+} \mu_J(k) \mathbf{e}_J^+ \exp\left(i(\mathbf{K}+\mathbf{K}_J)\cdot X\right),$$

$$(T_k \mathbf{e})^- = i \sum_{J \in J_k^+} \mu_J(k) \mathbf{e}_J^- \exp\left(i(\mathbf{K}+\mathbf{K}_J)\cdot X\right)$$

can be extended to a continuous linear operator from $H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3)$ to the set $(H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3))'$.

Moreover, T_k satisfies the following relations. For all

$$\mathbf{e}, \mathbf{e}^* \in \{H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3)\}^2:$$

(i) $$\langle T_k \mathbf{e}, \mathbf{e}^* \rangle = \langle T_k \mathbf{e}^*, \mathbf{e} \rangle,$$

(ii) $$\langle T_k \mathbf{e}, \mathbf{e} \rangle \leqq 0.$$

$\langle \cdot, \cdot \rangle$ stands for the duality brackets between $(H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3))'$ and $H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3)$.

Proof. Take smooth $\mathbf{e}$ and $\mathbf{e}^*$ and consider

$$\int_{\Sigma^\rho} T_k \mathbf{e} \cdot \overline{\mathbf{e}^*}\, d\sigma_X = -d_1 d_2 \sum_{J \in J_k^+} |\mu_J(k)| (\mathbf{e}_J^+ \cdot \overline{\mathbf{e}_J^{+*}} + \mathbf{e}_J^- \cdot \overline{\mathbf{e}_J^{-*}}).$$

We clearly have

$$\left| \sum_{J \in J_k^+} \left| \mu_J(k) |\mathbf{e}_J^+ \cdot \overline{\mathbf{e}_J^{+*}}|^2 \leqq \left(\sum_{J \in J_k^+} |\mu_J(k)| |\mathbf{e}_J^+|^2 \right) \left(\sum_{J \in J_k^+} |\mu_J(k)| |\mathbf{e}_J^{+*}|^2 \right) \right.$$

and the same estimate holds for the terms with a minus sign. But

$$|\mu_J(k)| \leqq |\mathbf{K} + \mathbf{K}_J|$$

and so,

$$|\langle T_k \mathbf{e}, \mathbf{e}^* \rangle| \leqq c |\mathbf{e}|_{1/2,\Sigma^\rho} |\mathbf{e}^*|_{1/2,\Sigma^\rho}.$$

This estimate ends the proof of the first part of Proposition 2.5. Properties (i) and (ii) are immediate for regular functions and can be extended by density for functions in $H_K^{1/2}(\Sigma^\rho, \mathbb{C}^3)$. □

The following proposition justifies the introduction of T.

PROPOSITION 2.6. *Let $\mathbf{e}$ solve (P_k); then the traces on Σ^ρ of the restriction to Ω_ρ^{ext} of $\mathbf{e}$ satisfy*

$$\gamma_1 \mathbf{e} = T(\gamma_0 \mathbf{e}). \tag{2.5}$$

Proof. We just have to consider the expansion of $\mathbf{e}$ given by Proposition 2.2 and take its normal derivative on Σ^ρ. Using then Proposition 2.2, we see that there is no component $\mathbf{e}_J(z)$, with $J \in J_k^-$. This enables us to write (2.5). □

We are now led to the study of an operator A_k, defined as a realization of $-\Delta$ on Ω_ρ^{ext}, satisfying (2.5) on Σ^ρ, the electric boundary value condition on $\partial\omega$, and "q.p." conditions on the lateral boundary.

$$D(A_k) := \{\mathbf{e} \in H^1_{\Delta,K}(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)/\mathbf{n}\times\mathbf{e} = \mathbf{0} \text{ on } \partial\omega;$$
$$\gamma_0 \operatorname{div} \mathbf{e} = 0 \text{ on } \partial\omega\colon \gamma_1\mathbf{e} = T(\gamma_0\mathbf{e}) \text{ on } \Sigma^\rho\},$$
$$A_k\mathbf{e} := -\Delta\mathbf{e} \quad \text{for } \mathbf{e} \text{ in } D(A_k).$$

Thanks to Proposition 1.1 and the above trace properties, we can state the following characterization of the space $H^1_{\Delta,K}(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$, which will make easier the analysis of A_k.

$$H^1_{\Delta,K}(\Omega_\rho^{\text{ext}}, \mathbb{C}^3) := \Big\{ \mathbf{e} \in H^1_K(\Omega_\rho^{\text{ext}}, \mathbb{C}^3) \cap H^1_\Delta(\Omega_\rho^{\text{ext}}, \mathbb{C}^3) \Big/$$
$$\forall\, \mathbf{e}^* \in H^1_K(\Omega_\rho^{\text{ext}}, \mathbb{C}^3) \int_{\Omega_\rho^{\text{ext}}} \Delta\mathbf{e} \cdot \overline{\mathbf{e}^*}\, dv$$
$$= -\sum_{\alpha=1}^{3} \int_{\Omega_\rho^{\text{ext}}} \nabla e^\alpha \cdot \overline{\nabla e^\alpha}\, dv + \int_{\partial\omega \cup \Sigma^\rho} \gamma_1\mathbf{e} \cdot \overline{\gamma_0\mathbf{e}^*}\, d\sigma \Big\}.$$

Here, the sum over $\partial\omega \cup \Sigma^\rho$ is to be understood in a duality sense.

We will prove the following theorem.

THEOREM 2.7. *A_k is a self-adjoint, densely defined operator, with compact resolvent.*

Proof. The set of smooth functions vanishing in a neighbourhood of $\partial\Omega_\rho^{\text{ext}}$ is dense in $L^2(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$ and is embedded in $D(A_k)$.

Take $\mathbf{e}$ and $\mathbf{e}^*$ in $D(A_k)$; using Green formula we get

$$\int_{\Omega_\rho^{\text{ext}}} (\Delta\mathbf{e} \cdot \overline{\mathbf{e}^*} - \overline{\Delta\mathbf{e}^*} \cdot \mathbf{e})\, dv = \int_{\partial\omega \cup \Sigma^\rho} (\gamma_1\mathbf{e} \cdot \overline{\gamma_0\mathbf{e}^*} - \overline{\gamma_1\mathbf{e}^*} \cdot \gamma_0\mathbf{e})\, d\sigma. \tag{2.6}$$

Using formulae (1.16), (1.17) on $\partial\omega$, we see that

$$\gamma_1\mathbf{e} \cdot \overline{\gamma_0\mathbf{e}^*} - \overline{\gamma_1\mathbf{e}^*} \cdot \gamma_0\mathbf{e} = -2H\gamma_0\mathbf{e} \cdot \overline{\gamma_0\mathbf{e}^*} + 2H\overline{\gamma_0\mathbf{e}^*} \cdot \gamma_0\mathbf{e} = 0 \quad \text{on } \partial\omega.$$

On Σ^ρ, boundary relations enable us to write

$$\int_{\Sigma^\rho} (\gamma_1\mathbf{e} \cdot \overline{\gamma_0\mathbf{e}^*} - \overline{\gamma_1\mathbf{e}^*} \cdot \gamma_0\mathbf{e})\, d\sigma = \langle T_k\mathbf{e}, \mathbf{e}^*\rangle - \overline{\langle T_k\mathbf{e}^*, \mathbf{e}\rangle} = 0.$$

Thus the left-hand side of (2.6) vanishes and A_k is symmetric.

We show that, for δ sufficiently large, $A_k + \delta I$ is maximal positive, so that self-adjointness will follow standard arguments. For this purpose, let us analyse the following problem:

(2.7) For $\mathbf{f}$ in $L^2(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$, find $\mathbf{e}$ in $D(A_k)$ such that

$$A_k\mathbf{e} + \gamma\mathbf{e} = \mathbf{f}.$$

Problem (2.7) has one and only one solution, for γ large enough, which is given by the solution of a variational problem set as follows.

We define

$$\mathbf{V}(\Omega_\rho^{\text{ext}}) := \{\mathbf{e} \in H^1_K(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)/\mathbf{n}\times\mathbf{e} = \mathbf{0} \text{ on } \partial\omega\}$$

and

$$a_k(\mathbf{e}, \mathbf{e}^*) := \sum_{\alpha=1}^{3} (\nabla e^\alpha \,|\, \nabla e^{*\alpha})_{0,\Omega_\rho^{\text{ext}}} - \langle T_k\mathbf{e}, \mathbf{e}^*\rangle - 2\int_{\partial\omega} H(\mathbf{e} \cdot \overline{\mathbf{e}^*})\, d\sigma.$$

Then, a variational formulation of (2.7) is:

(2.8) Find $\mathbf{e}$ in $\mathbf{V}(\Omega_\rho^{\text{ext}})$ such that, for all $\mathbf{e}^*$ in $\mathbf{V}(\Omega_\rho^{\text{ext}})$,

$$a_k(\mathbf{e}, \mathbf{e}^*) + \gamma(\mathbf{e}|\mathbf{e}^*)_{0,\Omega_\rho^{\text{ext}}} = (\mathbf{f}|\mathbf{e}^*)_{0,\Omega_\rho^{\text{ext}}}.$$

Lemma 2.8, which will be proved later, gives the properties of (2.8).

LEMMA 2.8. *There is a positive real γ^* such that, for all $\gamma \geqq \gamma^*$, we have that $a_k(\cdot,\cdot)+\gamma(\cdot|\cdot)_{0,\Omega_\rho^{\text{ext}}}$ is a continuous coercive sesquilinear form on the Hilbert space $(\mathbf{V}(\Omega_\rho^{\text{ext}}), \|\cdot\|_{1,\Omega_\rho^{\text{ext}}})$.*

Lemma 2.8 and the Lax-Milgram theorem enable us to state that problem (2.8) has one and only one solution, for a given $\mathbf{f}$ in $L^2(\Omega_\rho^{\text{ext}})$ and $\gamma \geqq \gamma^*$.

It remains to show that the solution of (2.6) belongs to $D(A_k)$, and leads to (2.7). This is a matter of classical tools in the analysis of variational problems, and we omit the proof.

$D(A_k)$ is compactly embedded in $L^2(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$; thus, A_k has compact resolvent. This ends the proof of Theorem 2.7. □

Proof of Lemma 2.8. In view of Proposition 2.5 and trace results of § 1, it is clear that $a_k(\cdot,\cdot)$ is well defined and continuous on $(\mathbf{V}_k(\Omega_\rho^{\text{ext}}), \|\cdot\|_{1,\Omega_\rho^{\text{ext}}})$. We now prove that $a_k(\cdot,\cdot)+\gamma(\cdot|\cdot)_{0,\Omega_\rho^{\text{ext}}}$ is coercive, with respect to the norm $\|\cdot\|_{1,\Omega_\rho^{\text{ext}}}$. Thanks to Proposition 2.5(ii), we have

$$a_k(\mathbf{e}, \mathbf{e}) + \gamma(\mathbf{e}|\mathbf{e})_{0,\Omega_\rho^{\text{ext}}} \geqq |\mathbf{e}|^2_{1,\Omega_\rho^{\text{ext}}} + \gamma|\mathbf{e}|^2_{0,\Omega_\rho^{\text{ext}}} \int_{\partial\omega} 2H|\mathbf{e}|^2 \, d\sigma.$$

But for every $\varepsilon > 0$, we have the estimate (see [7])

$$|\mathbf{e}|^2_{0,\partial\omega} \leqq \varepsilon|\mathbf{e}|^2_{1,\Omega_\rho^{\text{ext}}} + c(\varepsilon)|\mathbf{e}|^2_{0,\Omega_\rho^{\text{ext}}}.$$

Thus for γ large enough, $a_k(\mathbf{e}, \mathbf{e})+\gamma(\mathbf{e}|\mathbf{e})_{0,\Omega_\rho^{\text{ext}}} \geqq \alpha\|\mathbf{e}\|^2_{1,\Omega_\rho^{\text{ext}}}$ for a positive real α. □

A_k is the m-sectorial operator related to the sesquilinear form $a_k(\cdot,\cdot)$ in the sense of Kato's first representation theorem (see [6]).

Our next step is to get the spectral structure of A_k. We can apply to A_k standard results on self-adjoint operators, bounded from below and with compact resolvent. Thus we can state the following theorem (see [5]).

THEOREM 2.9. *There is a complete orthonormal basis of $L^2(\Omega_\rho^{\text{ext}}, \mathbb{C}^3)$, $\{\bar{\Phi}_n\}$, in $D(A_k)$ such that*

$$A_k\bar{\Phi}_n = \lambda_n\bar{\Phi}_n \quad \text{with } \lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n \uparrow +\infty$$

and $\lambda_n = \lambda_n(A_k)$ is given by the Min-Max *principle, which can be written in the following form*:

$$\lambda_n(A_k) = \operatorname*{Max}_{\mathbf{e}_1,\cdots,\mathbf{e}_n} \left\{ \operatorname*{Min}_{\substack{\mathbf{e}\in[\mathbf{e}_1,\cdots,\mathbf{e}_n]^\perp \\ |\mathbf{e}|_{0,\Omega_\rho^{\text{ext}}}=1;\ \mathbf{e}\in\mathbf{V}(\Omega_\rho^{\text{ext}})}} a_k(\mathbf{e}, \mathbf{e}) \right\} \tag{2.9}$$

Proof (see [5, Thm. XIII.64]). Here, we write the Min-Max principle in its "form" version [5, Thm. XIII.2], having in mind that the form domain of $a_k(\cdot,\cdot)$ is just $\mathbf{V}(\Omega_\rho^{\text{ext}})$. □

The Min-Max principle allows us to study the behaviour of $\lambda_n(A_k)$ as a function of k as follows.

PROPOSITION 2.10. *For each n, $\lambda_n(A_k)$ is a decreasing continuous function of k.*

Proof. Let $\mathbf{e}$ belong to $\mathbf{V}(\Omega_\rho^{\text{ext}})$; the dependence on k of $a_k(\mathbf{e}, \mathbf{e})$ is given by the sum

$$i \sum_{J\in J_k^+} \mu_J(k)(|\mathbf{e}_J^+|^2 + |\mathbf{e}_J^-|^2) \qquad (\in \mathbb{R}). \tag{2.10}$$

But clearly each term in (2.10) is a decreasing function of k; moreover, the sets $(J_k^+)_k$ decrease when k increases so that the sum (2.10) has the same property. This behaviour is conserved when we consider the Min-Max (2.9).

Take $\mathbf{e}$ in $\mathbf{V}(\Omega_\rho^{\text{ext}})$, with $|\mathbf{e}|_{0,\Omega_\rho^{\text{ext}}} = 1$. Then

$$a_k(\mathbf{e},\mathbf{e}) - a_{k'}(\mathbf{e},\mathbf{e}) = i \sum_{J \in J_k^+} (\mu_J(k) - \mu_J(k'))(|\mathbf{e}_J^+|^2 + |\mathbf{e}_J^-|^2).$$

We next consider the family $\hat{\mu}_J(k)$ of the extensions of $\mu_J(k)$ by 0, when $k \geqq |\mathbf{K} + \mathbf{K}_J|$. It is easily seen that the resulting family $\hat{\mu}_J(k)$ is equicontinuous. Hence for every $\varepsilon > 0$ there is an $\eta(\varepsilon)$ such that

$$|k - k'| \leqq \eta(\varepsilon) \Rightarrow \forall J \in \mathbb{Z}^2, \qquad |\mu_J(k) - \mu_J(k')| \leqq \varepsilon.$$

Hence, if $|k - k'| \leqq \eta(\varepsilon)$,

$$a_k(\mathbf{e},\mathbf{e}) - a_{k'}(\mathbf{e},\mathbf{e}) \leqq \varepsilon \sum_{J \in \mathbb{Z}^2} (|\mathbf{e}_J^+|^2 + |\mathbf{e}_J^-|^2).$$

Now choosing γ as in Lemma 2.8, we get

$$a_k(\mathbf{e},\mathbf{e}) \leqq a_{k'}(\mathbf{e},\mathbf{e}) + \varepsilon c(a_{k'}(\mathbf{e},\mathbf{e}) + \gamma)$$

where c is a constant independent of $\mathbf{e}$. For ε small enough, taking the Max-Min of this inequality, we obtain

$$\lambda_n(A_k) \leqq \lambda_n(A_{k'})(1 + c\varepsilon) + \varepsilon c\gamma.$$

We can also write

$$\lambda_n(A_{k'})(1 - c\varepsilon) \leqq \lambda_n(A_k) + \varepsilon c\gamma.$$

Hence

$$|\lambda_n(A_{k'}) - \lambda_n(A_k)| \leqq \varepsilon c |\lambda_n(A_{k'})| + \varepsilon\gamma c,$$

which yields the continuity of $\lambda_n(A_k)$ as a function of k.

In fact, it can be shown that $\lambda_n(A_k)$ is a piecewise analytical function of k, by using results of [6] on holomorphic families of operators for a suitable extension of A_k for complex values of k.

LEMMA 2.11. *The set K_{sing} of real numbers k for which there is an $n \in \mathbb{N}$ such that*

$$\lambda_n(A_k) = k^2 \tag{2.12}$$

is, at most, an increasing sequence $(k_n)_n$ tending to $+\infty$.

Proof. The situation is now understood after the analysis of Fig. 4. We just look after the crossings of the graphs of $k \to k^2$ and $k \to \lambda_n(A_k)$, for the different values of

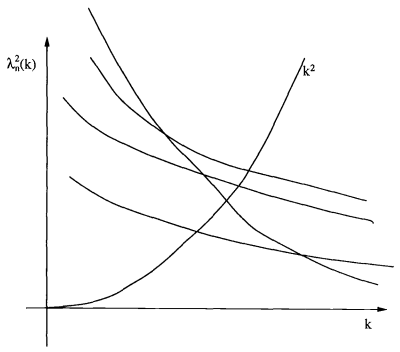


FIG. 4. *Different curves $k \to \lambda_n(k)$ and $k \to k^2$.*

n. The properties of $k \to \lambda_n(A_k)$ imply that, for each n, there exists a unique value of k satisfying $\lambda_n(A_k) = k^2$. We denote this value by k_n. Suppose now that the sequence $(k_n)_{n\in N}$ has an accumulation point at k_0^2. We can then extract a subsequence $(k_{n_p})_{n_p\in N}$ such that

$$\lim_{p\to\infty} \lambda_{n_p}(A_{k_{n_p}}) = \lim_{p\to\infty} k_{n_p}^2 = k_0^2.$$

Then

$$|\lambda_{n_p}(A_{k_0}) - k_0^2| \leqq |\lambda_{n_p}(A_{k_{n_p}}) - \lambda_{n_p}(A_{k_0})| + |\lambda_{n_p}(A_{k_{n_p}}) - k_0^2|.$$

Due to the uniformity in n of continuity estimates for $\lambda_n(A_k)$, we obtain that if $k_{n_p} - k_0$ is sufficiently small, i.e., $\partial\omega$ for p great enough, the value of $|\lambda_{n_p}(A_{k_0}) - k_0^2|$ can be made arbitrarily small. This contradicts the results of Theorem 2.9. □

The proof of Theorem 2.1 now becomes natural.

Proof of Theorem 2.1. Let $\mathbf{e}$ solve (P_k); we have already seen that $\mathbf{e}_{|\Omega_\rho^{ext}}$ is in $D(A_k)$. But the relation $\Delta\mathbf{e} + k^2\mathbf{e} = 0$ holds on Ω_ρ^{ext}, so that $\mathbf{e}$ is an eigenvector of A_k. Consequently, k is one of the values satisfying (2.12), and Lemma 2.11 gives the required result. □

2.3. Concluding remarks. We now want to underline some facts about our method. The construction of § 2.2 enables us to identify a solution of (1.16) with an eigenvector of A_k. The question now is, does an eigenvector of A_k give rise to a solution of (P_k)? If this were the case, we could state that there are really values of k for which problem (1.16) is not uniquely solvable. But eigenvectors of A_k cannot always be extended to the whole Ω^{ext}, satisfying the conditions dictated by the partial differential equations.

The first critical point is that we do not know if $\mathbf{e}_J(\overset{+}{-}\rho) = 0$ for J in J_k^-, so that we can extend each $\mathbf{e}_J(\cdot)$, J in J_k^- by zero, as is required by Lemma 2.2. Even if these conditions are satisfied, we then have to prove that the divergence of the extended eigenvector is zero, and this problem is related to that of the eigenvalues for the scalar case, with Dirichlet boundary conditions. Uniqueness properties for an arbitrarily smooth scatterer ω in this case have already been analysed, and have led to the same conclusions as Theorem 2.1: there is at most an increasing sequence of values k for which uniqueness is not satisfied, for the scalar Dirichlet problem; moreover, an eigenvector of this case gives, by considering its gradient field, an eigenvector for the Maxwell's equations problem (1.16). Besides, if ω satisfies a geometrical condition related to convexity properties we know that the scalar Dirichlet problem has no eigenvalue (no such criterion is known for the Neumann case). Granted these results, we can state that, if we are in the situation decribed above and if k^2 is not an eigenvalue of the scalar Dirichlet problem, then the extended eigenvector is actually a solution of (P_k).

3. Green kernel and integral equation formulation. It is known [1], [2], [8] that the determination of the diffracted field by a perfectly conducting body ω can be reduced to that of the surface currents $\mathbf{p}$ and charges λ on $\partial\omega$, which satisfy an integral equation, involving a suitable kernel, dictated by conditions at infinity. In this section we show that the same program can be carried out for the "quasi-periodic" diffraction problem, deriving by the way the expression of the requested Green kernel.

3.1. The Green kernel. We look for a function $\mathbf{G}^{qp}(X, Z)$ defined on Ω, which is "quasi periodic" and satisfies radiation conditions in the sense of § 1, and which is also an elementary solution of the Helmholtz equation in Ω. To this end we consider

the family

$$(\chi_J)_{J\in\mathbb{Z}^2};\ \chi_J := \frac{1}{i\mu_J(k)} \exp(i\mu_J(k)|Z|) \tag{3.1}$$

where we assume that $\mu_J(k) \neq 0$ ($|\mathbf{K}+\mathbf{K}_J| \neq k$) for all J in $\mathbb{Z}^2$.

χ_J is an elementary solution of the Helmholtz equation (1.6) with the radiation conditions (1.7).

We then consider formally the sum

$$\mathbf{G}^{qp}(X, Z) := \frac{1}{2d_1d_2} \sum_{J\in\mathbb{Z}^2} \chi_J(Z) \exp(-i(\mathbf{K}+\mathbf{K}_J)\cdot X). \tag{3.2}$$

PROPOSITION 3.1. *Assume that for every $J \in \mathbb{Z}^2 |\mathbf{K}+\mathbf{K}_J| \neq k$. Then* (3.2) *defines an $L^{2,\mathrm{loc}}(\mathbb{R}^3)$ function, which satisfies*:

(i) $\Delta\mathbf{G}^{qp} + k^2\mathbf{G}^{qp} = \sum_{J\in\mathbb{Z}^2} \delta(X_J, 0) \exp(i\mathbf{K}\cdot X_J)$ *in* $(C_0^\infty(\mathbb{R}^3))'$,

(ii) $\Delta\mathbf{G}^{qp} + k^2\mathbf{G}^{qp} = \delta_K$ *in* $(C_K^\infty(\Omega))'$,

(iii) $\mathbf{G}^{qp}$ *is a C^∞ function in $\bar{\Omega}\backslash\{(0,0,0)\}$, which satisfies "quasi-periodicity" conditions and radiation conditions.*

Here, δ_K is the Dirac distribution of the dual space of $C_K^\infty(\Omega)$, defined by $\langle\delta_K, \varphi\rangle := \varphi(0,0,0)$ for all $\varphi \in C_K^\infty(\Omega)$ where $\langle\cdot,\cdot\rangle$ stands for the duality bracket between $(C_K^\infty(\Omega))'$ and $C_K^\infty(\Omega)$.

Remark. We have to eliminate the case $|\mathbf{K}+\mathbf{K}_J| = k$, because a Green function satisfying radiation conditions cannot be constructed in this way.

Proof. Let J be in J_k^+, and consider

$$\xi_J(\rho) := \int_{-\rho}^{+\rho} |\chi_J|^2\, dZ \quad \text{for } \rho > 0.$$

Elementary calculus shows that

$$\xi_J(\rho) = 2/|\mu_J(k)|^3(1-\exp(-2\rho|\mu_J(k)|)).$$

Hence the family $(\xi_J(\rho))_{J\in J_k^+}$ is summable even if we take $\rho = +\infty$. For J in J_k^-, we also define

$$\xi_J(\rho) := \int_{-\rho}^{+\rho} |\chi_J|^2\, dZ \quad \text{for } \rho > 0$$

$$= 2\frac{\rho}{|\mu_J(k)|^2}.$$

But there is only a finite number of such terms, so that $(\xi_J(\rho))_{J\in\mathbb{Z}^2}$ is a summable family.

Now it is clear that for every finite part $F \subset \mathbb{Z}^2$

$$\left|\sum_{J\in F} \chi_J(Z) \exp(-i(\mathbf{K}+\mathbf{K}_J)\cdot X)\right|_{0,\Omega_\rho}^2 = d_1d_2 \sum_{J\in F} \xi_J(\rho) \leqq d_1d_2 \sum_{J\in\mathbb{Z}^2} \xi_J(\rho).$$

We conclude that $\mathbf{G}^{qp}$ is in $L^{2,\mathrm{loc}}(\Omega)$, and so (3.1) is almost everywhere summable in Ω. Now (3.2) defines a "q.p." function almost everywhere on $\mathbb{R}^3$, which is clearly in $L^{2,\mathrm{loc}}(\mathbb{R}^3)$.

We now come to point (i); indeed we have in a distributional sense the following expansion:

$$\Delta\mathbf{G}^{qp} + k^2\mathbf{G}^{qp} = 1/d_1d_2 \sum_{j\in\mathbb{Z}^2} \exp(-i(\mathbf{K}+\mathbf{K}_J)\cdot X) \otimes \delta_{Z=0} \quad \text{in } (C_0^\infty(\mathbb{R}^3))'.$$

Hence for φ in $C_0^\infty(\mathbb{R}^3)$, we get

$$\langle \Delta \mathbf{G}^{qp} + k^2 \mathbf{G}^{qp}, \varphi \rangle = 1/d_1 d_2 \sum_{J \in \mathbb{Z}^2} \int_{\mathbb{R}^2} \exp(-i(\mathbf{K}+\mathbf{K}_J)\cdot X)\varphi(X,0)\, dX$$

$$= 1/d_1 d_2 \sum_{J \in \mathbb{Z}^2} F_X(\varphi)(\mathbf{K}+\mathbf{K}_J, 0)$$

$$= \sum_{J \in \mathbb{Z}^2} \varphi(X_J, 0) \exp(-i\mathbf{K}\cdot X_J)$$

by using Poisson's formula for the Fourier transform in the X variable. Thus, we obtain (i).

In (ii), we consider $\mathbf{G}^{qp}$ as a distribution that is "q.p." in the X variable and acts as a usual distribution on the Z variable. "q.p." distributions are defined similarly to periodic distributions (see [10]), noting that, if $F(X)$ is a "q.p." function, $F(X)\exp(-iK\cdot X)$ is a periodic function. To prove (ii), we just note that δ_K has the following expansion in $(C_K^\infty(\Omega))'$:

$$\delta_K = 1/d_1 d_2 \sum_{J \in \mathbb{Z}^2} \exp(-i(\mathbf{K}+\mathbf{K}_J)\cdot X) \otimes \delta_{Z=0}$$

and that is exactly what we get, applying $(\Delta + k^2)$ to $\mathbf{G}^{qp}$. □

PROPOSITION 3.2. *Let $\mathbf{G}^0$ be a usual elementary solution of the Helmholtz equation*:

$$\mathbf{G}^0((X,Z)) := \frac{\exp(ik|(X,Z)|)}{4\pi|(X,Z)|}.$$

Then $\mathbf{G}^r := \mathbf{G}^0 - \mathbf{G}^{qp}$ is a C^∞ function in every open set which does not contain a point $(X_J, 0)$, $J \in \mathbb{Z}^2$, $j \neq (0,0)$. Hence $\mathbf{G}^{qp}$ is a C^∞ function in $\bar{\Omega}\backslash\{(0,0,0)\}$ and satisfies radiation conditions (1.7).

Proof. In an open set ϑ which does not contain a point $(X_J, 0)$, $\mathbf{G}^r$ satisfies

$$\Delta \mathbf{G}^r + k^2 \mathbf{G}^r = 0 \quad \text{in } (C^\infty(\vartheta))'.$$

But $\mathbf{G}^r$ is in $L^{2,\mathrm{loc}}(\mathbb{R}^3)$ so that, applying classical results on elliptic regularity for Laplace operator, we obtain that $\mathbf{G}^r$ is a $C^\infty(\vartheta)$ function. The rest of Proposition 3.2 is easy, because we know explicitly the behaviour of $\mathbf{G}^0$, and radiation conditions follow from the definition of $\mathbf{G}^{qp}$. □

3.2. Integral formulations and existence results. Let $\mathbf{e}$ be defined almost everywhere in Ω by the following: $\mathbf{e}$ is a solution of (1.16) in Ω^{ext}, and $\mathbf{e}$ is a solution of the interior problem related to (1.16) in ω.

Find $\mathbf{e}$ in $H^1_\Delta(\omega, \mathbb{C}^3)$ such that

$$\begin{aligned} &\Delta \mathbf{e} + k^2 \mathbf{e} = \mathbf{0} && \text{in } \omega, \\ &\operatorname{div} \mathbf{e} = 0 && \text{in } \omega, \\ &\mathbf{n} \times \mathbf{e} = \mathbf{c} && \text{on } \partial\omega. \end{aligned}$$

Regularity results yield $\mathbf{e} \in C^\infty(\bar{\omega}, \mathbb{C}^3) \cup C_K^\infty(\overline{\Omega^{\mathrm{ext}}}, \mathbb{C}^3)$.

We can then state the following proposition, giving a representation formula for $\mathbf{e}$.

PROPOSITION 3.3. $\mathbf{e}$ *admits the representation*

$$\mathbf{e}(X,Z) = -\operatorname{grad} v(\lambda)(X,Z) + \mathbf{A}(\mathbf{p})(X,Z) \quad \textit{for all } (X,Z) \quad \textit{in } \Omega^{\mathrm{ext}} \cup \omega, \tag{3.3}$$

$$v(\lambda)(X,Z) := \int_{\partial\omega} \mathbf{G}^{qp}(X'-X, Z'-Z)\lambda(X',Z')\, d\sigma', \tag{3.4}$$

$$\mathbf{A}(\mathbf{p})(X, Z) := \int_{\partial\omega} \mathbf{G}^{qp}(X'-X, Z'-Z)\mathbf{p}(X', Z')\, d\sigma', \quad \textit{and} \tag{3.5}$$

$$\lambda := -[\mathbf{e}\cdot n], \tag{3.6}$$

$$\mathbf{p} := [\operatorname{curl}\mathbf{e}\times n]. \tag{3.7}$$

Here $[\cdot]$ *stands for the jump across* $\partial\omega$ *of a function with exterior and interior traces at* $\partial\omega$.

Proof. We apply Green formula to the vector fields $\mathbf{G}^{qp}\mathbf{e}^*$ and $\mathbf{e}$, where $\mathbf{e}^*$ is an arbitrary constant vector of $\mathbb{R}^3$ in the sets ω and Ω_ρ^{ext}, respectively. This yields (after some tedious but classical computations) the following. If $|Z| < \rho$, then

$$\begin{aligned}\mathbf{e}(X, Z)\cdot\mathbf{e}^* = {} & (-\operatorname{grad} v(\lambda)(X, Z) + \mathbf{A}(\mathbf{p})(X, Z))\cdot\mathbf{e}^* \\ & + \int_{\Sigma^\rho} ((\nabla_X\cdot\mathbf{G}^{qp}(X'-X, Z'-Z)\times\mathbf{e}^*)\times\mathbf{e}_3)\cdot\mathbf{e}(X', Z')\, d\sigma' \\ & + \int_{\Sigma^\rho} (\nabla_X\cdot\mathbf{G}^{qp}(X'-X, Z'-Z)\cdot\mathbf{e}^*)(\mathbf{e}_3\cdot\mathbf{e}^*(X', Z'))\, d\sigma' \\ & - \int_{\Sigma^\rho} (\mathbf{G}^{qp}(X'-X, Z'-Z)\mathbf{e}^*)\cdot(\operatorname{curl}\mathbf{e}(X', Z')\times\mathbf{e}_3)\, d\sigma'.\end{aligned}$$

Thus, we only have to show that our radiation conditions imply that the contribution from Σ^ρ vanishes: again this is only a matter of computation and we omit the proof. □

We now give a more precise framework for the integral equation analysis, which is to be related to the study done in [2]. We begin by recalling some results that can be found in [2].

To this end, we set

$$\begin{aligned}\mathbf{H} &:= TH^{-1/2}(\partial\omega), \\ \mathbf{X} &:= \{\mathbf{p}\in H;\ \operatorname{div}_{\partial\omega}\mathbf{p}\in H^{-1/2}(\partial\omega)\}, \\ \mathbf{M} &:= \{\lambda\in H^{-1/2}(\partial\omega);\ \langle\lambda, 1\rangle = 0\}, \\ \mathbf{L} &:= \{\lambda\in M;\ \lambda\in H^{1/2}(\partial\omega)\}\end{aligned}$$

(these spaces are endowed with their natural norms), and we define the following operators, for smooth $\mathbf{p}$ and λ:

$$\mathbf{A}^0(\mathbf{p})(X, Z) := \int_{\partial\omega} \mathbf{G}^0(X'-X, Z'-Z)\mathbf{p}(X', Z')\, d\sigma,$$

$$v^0(\lambda)(X, Z) := \int_{\partial\omega} \mathbf{G}^0(X'-X, Z'-Z)\lambda(X', Z')\, d\sigma',$$

$$\mathbf{C}(\rho)(X, Z) := \int_{\partial\omega} \frac{1}{4\pi|(X'-X, Z'-Z)|}\rho(X', Z')\, d\sigma'.$$

From [9], we can state that $\mathbf{C}(\cdot)$ defines an isomorphism from $H^s(\partial\omega)$ onto $H^{s+1}(\partial\omega)$, for every positive real s.

From [2], we can state the following theorem.

THEOREM (Bendali). *The operator defined for smooth* $(\mathbf{p}, \lambda)$ *in* $\mathbf{X}\times\mathbf{L}$ *by*

$$\Lambda(\mathbf{p}, \lambda) := (\mathbf{A}^0(\mathbf{p}) - \operatorname{grad}_{\partial\omega}(v^0(\lambda)), \mathbf{C}(\operatorname{div}_{\partial\omega}\mathbf{p} + k^2\lambda))$$

can be extended to a continuous linear operator from $\mathbf{X}\times\mathbf{L}$ *onto* $TH^{1/2}(\partial\omega)\times\mathbf{M}'$. *Moreover,* Λ *is an isomorphism from* $\mathbf{X}\times\mathbf{L}$ *onto* $TH^{1/2}(\partial\omega)\times\mathbf{M}'$.

As a corollary of this result we state the following corollary.

COROLLARY. *The operator defined for smooth* $\mathbf{p}$ *in* $\mathbf{X}$ *by*

$$\tilde{\Lambda}(\mathbf{p}) := \mathbf{A}^0(\mathbf{p}) + 1/k^2 \operatorname{grad}_{\partial\omega}(v^0(\operatorname{div}_{\partial\omega}\mathbf{p}))$$

can be extended to an isomorphism from $\mathbf{X}_+ := \{\mathbf{p}\in\mathbf{X};\ \operatorname{div}_{\partial\omega}\mathbf{p}\in H^{1/2}(\partial\omega)\}$ *onto* $TH^{1/2}(\partial\omega)$. *The norm on* $\mathbf{X}^+$ is $\|\mathbf{p}\|_{\mathbf{X}_+} = \|\mathbf{p}\|_{-1/2,\partial\omega} + \|\operatorname{div}_{\partial\omega}\mathbf{p}\|_{1/2,\partial\omega}$.

Proof. We want to solve

$$\tilde{\Lambda}(\mathbf{p}) = \mathbf{c} \quad \text{for } \mathbf{c} \quad \text{in } TH^{1/2}(\partial\omega).$$

Let $(\mathbf{p}, \lambda)$ be the solution of

$$\Lambda(\mathbf{p}, \lambda) = (\mathbf{c}, 0);$$

then

$$\operatorname{div}_{\partial\omega}\mathbf{p} = -k^2\lambda \in H^{1/2}(\partial\omega)$$

and hence

$$\mathbf{p}\in\mathbf{X}_+ \quad \text{and} \quad \tilde{\Lambda}(\mathbf{p}) = \mathbf{c}$$

Now, $\tilde{\Lambda}$ is clearly continuous and injective; this follows from the properties of Λ. □

PROPOSITION 3.4. *The operator* Λ^{qp}, *defined for smooth* $\mathbf{p}$ *in* $\mathbf{X}_+$ *by*

$$\Lambda^{qp}(\mathbf{p}) = 1/k^2 \operatorname{grad}_{\partial\omega}(v^{qp}(\operatorname{div}_{\partial\omega}\mathbf{p})) + \mathbf{A}^{qp}(\mathbf{p}), \tag{3.8}$$

where

$$v^{qp}(\lambda)(X, Z) := \int_{\partial\omega} \mathbf{G}^{qp}(X'-X, Z'-Z)\lambda(X', Z')\, d\sigma', \tag{3.9}$$

$$\mathbf{A}^{qp}(\mathbf{p})(X, Z) := \int_{\partial\omega} \mathbf{G}^{qp}(X'-X, Z'-Z)\mathbf{p}(X', Z')\, d\sigma' \tag{3.10}$$

can be extended to a bounded linear operator from $\mathbf{X}_+$ *to* $TH^{1/2}(\partial\omega)$.

Let $\mathbf{c}$ *be in* $TH^{1/2}(\partial\omega)$, *and let* $\mathbf{p}$ *in* $\mathbf{X}_+$ *satisfy the integral equation*

$$\Lambda^{qp}(\mathbf{p}) = \mathbf{c}. \tag{3.11}$$

Then $\mathbf{e}$ *defined by*

$$\mathbf{e}(X, Z) = -\operatorname{grad}(v^{qp}(\operatorname{div}_{\partial\omega}\mathbf{p}))(X, Z) + \mathbf{A}^{qp}(\mathbf{p})(X, Z) \tag{3.12}$$

for all (X, Z) *in* $\Omega^{\text{ext}}\cup\omega$, *is a solution of* (1.16), *and of the related interior problem, and we have*

$$\mathbf{p} = [\operatorname{curl}\mathbf{e}\times\mathbf{n}]. \tag{3.13}$$

Proof. We use the splitting of $\mathbf{G}^{qp}$:

$$\mathbf{G}^{qp} = \mathbf{G}^0 + \mathbf{G}^r.$$

This decomposition leads us to define analogously to (3.9) and (3.10) for the kernel $\mathbf{G}^r$:

$$v^r(\lambda)(X, Z) := \int_{\partial\omega} \mathbf{G}^r(X'-X, Z'-Z)\lambda(X', Z')\, d\sigma',$$

$$\mathbf{A}^r(\mathbf{p})(X, Z) := \int_{\partial\omega} \mathbf{G}^r(X'-X, Z'-Z)\mathbf{p}(X', Z')\, d\sigma'.$$

From the Bendali theorem and its corollary we know that the mapping defined for smooth $\mathbf{p}$ in $\mathbf{X}$, by

$$\tilde{\Lambda}^0(\mathbf{p}) := 1/k^2 \operatorname{grad}_{\partial\omega} v^0(\operatorname{div}_{\partial\omega} \mathbf{p}) + \mathbf{A}^0(\mathbf{p})$$

can be extended to a bounded linear operator from $\mathbf{X}_+$ to $TH^{1/2}(\partial\omega)$.

Now it is easily seen that the regular kernel $\mathbf{G}^r$ defines a regularizing operator, which maps $\mathbf{X}_+$ on $TC^\infty(\partial\omega)$ (smooth tangential fields on $\partial\omega$), by

$$\Lambda^r(\mathbf{p}) := 1/k^2 \operatorname{grad}_{\partial\omega} v^r(\operatorname{div}_{\partial\omega} \mathbf{p}) + \mathbf{A}^r(\mathbf{p}).$$

Thus $\Lambda^{qp} = \Lambda^0 + \Lambda^r$ can be extended to a bounded linear operator from $\mathbf{X}_+$ to $TH^{1/2}(\partial\omega)$.

Thanks to the splitting $\mathbf{G}^{qp} = \mathbf{G}^0 + \mathbf{G}^r$, it is already seen that $\mathbf{e}(X, Z)$, defined by (3.12), satisfies

$$\mathbf{e} \text{ is in } H^{1,\mathrm{loc}}_{\Delta,K}(\Omega^{\mathrm{ext}}, \mathbb{C}^3),$$

$$\Delta\mathbf{e} + k^2\mathbf{e} = 0 \quad \text{in } \Omega^{\mathrm{ext}} \cup \omega.$$

Classical potential relations can be applied to the kernel $\mathbf{G}^0$, giving the continuity of the tangential component of (3.12) across $\partial\omega$, having in mind that the regular kernel $\mathbf{G}^r$ introduces no discontinuity across $\partial\omega$:

$$\begin{aligned}\Pi\mathbf{e} &= \Pi(1/k^2 \operatorname{grad} v^{qp}(\operatorname{div}_{\partial\omega} \mathbf{p}) + \mathbf{A}^{qp}(\mathbf{p})) \\ &= 1/k^2 \operatorname{grad}_{\partial\omega} v^{qp}(\operatorname{div}_{\partial\omega} \mathbf{p}) + \mathbf{A}^{qp}(\mathbf{p}) = \mathbf{c}.\end{aligned}$$

Moreover, radiation conditions are satisfied: indeed, for $z \geqq \rho_0$, for example, the following expansion holds:

$$\mathbf{e}(X, Z) = \sum_{j \in \mathbb{Z}^2} \mathbf{e}_J^{0+} \exp(i(\mathbf{K} + \mathbf{K}_J) \cdot X) \exp(iZ\mu_J(k))$$

where $\mathbf{e}_J^{0+}$ is the constant vector of $\mathbb{C}^3$ given by

$$\begin{aligned}\mathbf{e}_J^{0+} := \Bigg(-\frac{1}{k^2} \Bigg[\int_{\partial\omega} & \frac{\exp(-i(\mathbf{K} + \mathbf{K}_J) \cdot X')}{i\mu_J(k)} \\ & \cdot \exp(iZ'\mu_J(k)) \operatorname{div}_{\partial\omega} \mathbf{p}(X', z')\, d\sigma' \Bigg] (i(\mathbf{K} + \mathbf{K}_J) + i\mu_J(k)\mathbf{e}_3) \\ + \int_{\partial\omega} & \frac{\exp(-i(\mathbf{K} + \mathbf{K}_J) \cdot X')}{i\mu_J(k)} \exp(iZ'\mu_J(k)) \mathbf{p}(X', z')\, d\sigma \Bigg)\end{aligned}$$

and the series is absolutely convergent and termwise infinitely differentiable.

We just have now to check div $\mathbf{e}$:

$$\begin{aligned}\operatorname{div} \mathbf{e} &= 1/k^2 \Delta v^{qp}(\operatorname{div}_{\partial\omega} \mathbf{p}) + \operatorname{div}(\mathbf{A}^{qp}(\mathbf{p})) \\ &= -v^{qp}(\operatorname{div}_{\partial\omega} \mathbf{p}) + \operatorname{div}(\mathbf{A}^{qp}(\mathbf{p}))\end{aligned} \tag{3.14}$$

and $\operatorname{div}(\mathbf{A}^{qp}(\mathbf{p}))$ is given by

$$\begin{aligned}\operatorname{div}(\mathbf{A}^{qp}(\mathbf{p}))(X, Z) &:= \int_{\partial\omega} \nabla_X \mathbf{G}^{qp}(X' - X, Z' - Z)\mathbf{p}(X', Z')\, d\sigma' \\ &= -\int_{\partial\omega} \nabla_{\partial\omega, X} \cdot \mathbf{G}^{qp}(X' - X, Z' - Z)\mathbf{p}(X', Z')\, d\sigma'\end{aligned}$$

and using the Stokes formula:

$$\int_{\partial\omega} \nabla_{\partial\omega X} \cdot \mathbf{G}^{qp}(X'-X, Z'-Z)\mathbf{p}(X', Z')\, d\sigma'$$

$$= -\int_{\partial\omega} \mathbf{G}^{qp}(X'-X, Z'-Z)\, \mathrm{div}_{\partial\omega}\, \mathbf{p}(X', Z')\, d\sigma',$$

and so (3.14) vanishes.

To prove (3.13), we just recall that the regular kernel $\mathbf{G}^r$ induces no jump across $\partial\omega$, and then we apply results of [2]. □

Equation (3.11) defines an integral equation for the unknown $\mathbf{p}$; it gives rise to solutions of (1.16) and of its related interior problem in ω, so that we can use the uniqueness results of § 2 in our analysis.

THEOREM 3.5. *Let* $\mathbf{N}$ *be defined by*

$$\mathbf{N} := \{r \in \mathbb{R}_+;\ (r \in K_{\text{sing}})\}$$
$$\cup \{-r^2 \text{ is an eigenvalue of } \Delta \text{ in } \omega \text{ with Dirichlet homogenous condition}\}$$
$$\cup \{\exists J \in \mathbb{Z}^2 / r = |\mathbf{K} + \mathbf{K}_J|)\}.$$

Suppose that $k \notin \mathbf{N}$; *then the integral equation* (3.11) *is uniquely solvable, and so gives rise to a unique solution to problem* (1.16).

Proof. Our purpose is to show that (3.11) can be handled by a Fredholm alternative technique. As $TC^\infty(\partial\omega)$ is compactly embedded in $TH^{1/2}(\partial\omega)$, Λ^r is found to be a compact perturbation of Λ^0. Hence we can now state the conclusions of the Fredholm alternative if we show that first part of Fredholm alternative holds when k is not in $\mathbf{N}$.

Let us suppose that

$$\Lambda^{qp}(\mathbf{p}) = \mathbf{0}.$$

By uniqueness results of § 2, applied to the vector field $\mathbf{e}$ given in Ω_ρ^{ext} by (3.12), we have $\mathbf{e} = \mathbf{0}$ in Ω_ρ^{ext}. In ω, the hypothesis that $-k^2$ is not an eigenvalue of the Laplace operator with homogenous Dirichlet boundary condition ensures that $\mathbf{e} = \mathbf{0}$ in ω (see [2], [9]). Thus $\mathbf{p} = [\text{curl } \mathbf{e} \times \mathbf{n}] = \mathbf{0}$. Fredholm's alternative then implies that (3.11) has one and only one solution. □

REFERENCES

[1] A. BENDALI, *Numerical analysis of the exterior boundary value problem for the time harmonic Maxwell equation by a boundary finite element method*, Math. Comp., 43 (1984), pp. 29-46.

[2] ———, *Problème aux limites exterieur et interieur pour le systeme de Maxwell en regime harmonique*, Rapport interne 50, C.M.A.P., Ecole Polytechnique, Palaiseau, France, 1980.

[3] C. MULLER, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer-Verlag, Berlin, 1969.

[4] H. D. ALBER, *A quasi periodic boundary value problem for the Laplacian and the continuation of its resolvent*, Proc. Roy. Soc. Edinburgh Sect. A, (1979), pp. 251-272.

[5] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics*, Vol. 4, Academic Press, New York, 1979.

[6] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1976.

[7] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogenes et applications*, Vol. 1, Dunod, Paris, 1968.

[8] D. J. POGGIO AND E. K. MILLER, *Solutions of three-dimensional scattering problems*, in Computer Techniques for Electromagnetics, R. Mittra, ed., Pergamon Press, New York, 1973.

[9] J. GIROIRE, *Integral equation methods for exterior problems for the Helmholtz equation*, Rapport interne 40, C.M.A.P., Ecole Polytechnique, Palaiseau, France, 1978.

[10] L. SCHWARTZ, *Théorie des distributions*, Hermann, Paris, 1976.

[11] C. H. WILCOX, *Scattering Theory for Diffraction Gratings*, Springer-Verlag, Berlin, 1980.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1702–1722, November 1991

010

DIFFUSION OF FLUID IN A FISSURED MEDIUM WITH MICROSTRUCTURE*

R. E. SHOWALTER† AND N. J. WALKINGTON‡

Abstract. A system of quasilinear degenerate parabolic equations arising in the modeling of diffusion in a fissured medium is studied. There is one such equation in the local cell coordinates at each point of the medium, and these are coupled through a similar equation in the global coordinates. It is shown that the initial boundary value problems are well posed in the appropriate spaces.

Key words. porous medium, double porosity, degenerate parabolic system

AMS(MOS) subject classifications. 35K55, 35K65

1. Introduction.

We shall study the Cauchy–Dirichlet problem for degenerate parabolic systems of the form

$$\frac{\partial}{\partial t}a(u) - \vec{\nabla}\cdot\widetilde{A}(x,\vec{\nabla}u) + \int_{\Gamma_x}\widetilde{B}(x,s,\vec{\nabla}_y U)\cdot\vec{\nu}\,ds \ni f, \qquad x\in\Omega, \tag{1.1a}$$

$$\frac{\partial}{\partial t}b(U) - \vec{\nabla}_y\cdot\widetilde{B}(x,y,\vec{\nabla}_y U) \ni F, \qquad y\in\Omega_x, \tag{1.1b}$$

$$\widetilde{B}(x,y,\vec{\nabla}_y U)\cdot\vec{\nu} + \mu\big(U(x,y,t) - u(x,t)\big) \ni 0, \qquad y\in\Gamma_x. \tag{1.1c}$$

Here Ω is a domain in $\mathbb{R}^n$ and for each value of the macrovariable $x\in\Omega$ is specified a domain Ω_x with boundary Γ_x for the microvariable $y\in\Omega_x$. Each of a, b, μ is a maximal monotone graph. These graphs are not necessarily strictly increasing; they may be piecewise constant or multivalued. The elliptic operators in (1.1a) and (1.1b) are of p-Laplacian type, i.e., they are nonlinear in the gradient of degree $p-1>0$ and $q-1>0$, respectively, with $\frac{1}{q}+\frac{1}{n}\geq\frac{1}{p}$, so some specific degeneracy is also permitted here. Certain first-order spatial derivatives can be added to (1.1a) and (1.1b) with no difficulty, and corresponding problems with constraints, i.e., variational inequalities, can be treated similarly. A particular example important for applications is the linear constraint

$$U(x,y,t) = u(x,t), \qquad y\in\Gamma_x, \quad x\in\Omega \tag{1.1c$'$}$$

which then replaces (1.1c). The system (1.1) with $\mu(s)=\frac{1}{\varepsilon}|s|^{q-2}s$ is called a *regularized microstructure model*, and (1.1a), (1.1b), (1.1c$'$) is the corresponding *matched microstructure model* in which (formally) $\varepsilon\to 0$. An example of such a system as a model for the flow of a fluid (liquid or gas) through a fractured medium will be given below. In such a context, (1.1a) prescribes the flow on the global scale of the fissure system and (1.1b) gives the flow on the microscale of the individual cell at a specific

*Received by the editors April 23, 1990; accepted for publication (in revised form) January 30, 1991. This work was supported by grants from the National Science Foundation and the Office of Naval Research.

†Department of Mathematics, University of Texas, Austin, Texas 78712.

‡Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

point x in the fissure system. The transfer of fluid between the cells and surrounding medium is prescribed by (1.1c) or (1.1c′). A major objective is to accurately model this fluid exchange between the cells and fissures.

Systems of the form (1.1) were developed in [21], [22], [10] in physical chemistry as models for diffusion through a medium with a prescribed microstructure. Similar systems arose in soil science [5], [14] and in reservoir models for fractured media [11], [16]. An existence-uniqueness theory for linear problems which exploits the strong parabolic structure of the system was given in [24]. Alternatively it is possible to eliminate U and obtain a single functional differential equation for u in the simpler space $L^2(\Omega)$, but the structure of the equation then obstructs the optimal parabolic type results [18]. Also see [13] for a nonlinear system with reaction–diffusion local effects.

These systems also arise from methods of homogenization. There an exact model is assumed periodic and described by a parabolic equation with periodic coefficients corresponding to the properties of the two components, the cells and fissures. The limit of this highly singular problem as the period tends to zero is the system (1.1), which is thereby justified as an approximation for the exact model. Homogenization theory provides not only a justification of the linear case of (1.1) as a model but also a means of calculating the coefficients in (1.1) in terms of those of the exact model, and a deeper analysis may describe the convergence itself [25], [17], [2], [3]. Here we study the nonlinear system directly. The task of determining the coefficients in (1.1) directly from, e.g., boundary observations, is an intriguing open problem.

The plan of this paper is as follows. In § 2 we shall give the precise description and resolution of the stationary problem in a variational formulation by monotone operators from Banach spaces to their duals. In order to achieve this we describe first the relevant Sobolev spaces, the continuous direct sums of these spaces, and the distributed trace and constant functionals which occur in the system. The operators are monotone functions or multivalued subgradients and serve as models for nonlinear elliptic equations in divergence form. We develop an abstract Green's theorem to describe the resolution of the variational form as the sum of a partial differential equation and a complementary boundary operator. Then sufficient conditions of coercivity type are given to assert the existence of generalized solutions of the variational equations. In § 3 we describe the restriction of our system to appropriate products of L^r spaces. The Hilbert space case, $r = 2$, serves not only as a convenient starting point but also leads to the generalized accretive estimates we shall need for the singular case of (1.1) in which a or b is not only nonlinear but multivalued. The stationary operator for (1.1) is shown to be m-accretive in the L^1 space, so we obtain a *generalized solution* in the sense of the nonlinear semigroup theory for general Banach spaces. As an intermediate step we shall show the special case of $a = b =$ identity is resolved as a *strong solution* in every L^r space, $1 < r < \infty$, and also in appropriate dual Sobolev spaces.

In order to motivate the system (1.1), let us consider the flow of a fluid through a fissured medium. This is assumed to be a structure of porous and permeable blocks or cells which are separated from each other by a highly developed system of fissures. The majority of fluid transport will occur along flow paths through the fissure system, and the relative volume of the cell structure is much larger than that of the fissure system. There is assumed to be no direct flow between adjacent cells, since they are individually isolated by the fissures, but the dynamics of the flux exchanged between each cell and its surrounding fissures is a major aspect of the model. The distributed microstructure models that we develop here contain explicitly the local geometry of

the cell matrix at each point of the fissure system, and they thereby reflect more accurately the flux exchange on the microscale of the individual cells across their intricate interface.

Let the flow region Ω be a bounded domain in $\mathbb{R}^n$ with boundary $\Gamma = \partial\Omega$. Let $\rho(x,t)$ and $p(x,t)$ be the *density* and *pressure*, respectively, at $x \in \Omega$ and $t > 0$, each being obtained by averaging over an appropriately small neighborhood of x. At each such x let there be given a cell Ω_x, a bounded domain in $\mathbb{R}^n$ with smooth boundary $\Gamma_x = \partial\Omega_x$. The collection of these Ω_x, $x \in \Omega$, is the distribution of blocks or cells in the structure. Within each Ω_x there is fluid of density $\tilde{\rho}(x,y,t)$ and pressure $\tilde{p}(x,y,t)$, respectively, for $y \in \Omega_x$, $t > 0$. The conservation of fluid mass in the fissure system yields the global diffusion equation

$$\frac{\partial}{\partial t}\big(\rho + a_0(p)\big) - \sum_{j=1}^{n} \frac{\partial}{\partial x_j}\left(\rho k_j\left(\rho\frac{\partial p}{\partial x_j}\right)\frac{\partial p}{\partial x_j}\right) + q(x,t) = f(x,t), \qquad x \in \Omega, \tag{1.2a}$$

in which the total concentration $\rho + a_0(p)$ includes adsorption or capillary effects, the function k_j gives the permeability of the fissure system in the jth coordinate direction, $q(x,t)$ is the density of mass flow of fluid into the cell Ω_x at x, and f is the density of fluid sources. Similarly, we have within each cell

$$\frac{\partial}{\partial t}\big(\tilde{\rho} + b_0(\tilde{p})\big) = \sum_{j=1}^{n} \frac{\partial}{\partial y_j}\left(\tilde{\rho}\tilde{k}_j\left(\tilde{\rho}\frac{\partial \tilde{p}}{\partial y_j}\right)\frac{\partial \tilde{p}}{\partial y_j}\right), \qquad y \in \Omega_x, \tag{1.2b}$$

where b_0 denotes adsorption or capillary effects and the function $\tilde{k}_j$ gives the local cell permeability. Assume the flux across the cell boundary is driven by the pressure difference and is also proportional to the *average density* $\bar{\rho}$ on that pressure interval. Thus, we have the interface condition

$$\sum_{j=1}^{n} \tilde{\rho}\tilde{k}_j\left(\tilde{\rho}\frac{\partial \tilde{p}}{\partial y_j}\right)\frac{\partial \tilde{p}}{\partial y_j}\nu_j + \mu\big(\bar{\rho}(\tilde{p} - p)\big) \ni 0, \qquad y \in \Gamma_x, \tag{1.2c}$$

where $\vec{\nu}$ is the unit outward normal on Γ_x and μ is the relation between the flux across the interface and the density-weighted pressure difference as indicated. The total mass flow into the cell is given by

$$q(x,t) = \int_{\Gamma_x} \tilde{\rho} \sum_{j=1}^{n} \tilde{k}_j\left(\tilde{\rho}\frac{\partial \tilde{p}}{\partial y_j}\right)\frac{\partial \tilde{p}}{\partial y_j}\nu_j \, ds. \tag{1.2d}$$

In order to complete the dynamical system we need only to add a boundary condition on Γ to (1.2a) and to postulate the *state equation*

$$\rho = s(p) \tag{1.2e}$$

for the fluid in the fissure and cell systems. Here $s(\cdot)$ is a given monotone function (or graph) determined by the fluid.

In order to place (1.2) in a more convenient form, we introduce the monotone function

$$S(w) \equiv \int_0^w s(r)\,dr$$

and the corresponding flow potentials for the fluid in the fissures and cells

$$u = S(p), \qquad U = S(\tilde{p})\ .$$

In these variables with a change of notation the system (1.2) can be written in the form (1.1) together with boundary conditions on Γ for u or $\widetilde{A}(\nabla u)\cdot\nu$ and initial conditions at $t = 0$ on $a(u)$, $b(U)$. Note that the average density on the pressure interval $p, \tilde{p}$ is given by

$$\bar{\rho} = \frac{1}{p-\tilde{p}}\int_{\tilde{p}}^{p} s(r)\,dr = \frac{u-U}{p-\tilde{p}}.$$

As an alternative to (1.2c), we could require that $\tilde{p} = p$ on Γ_x and this leads to (1.1c′) in place of (1.1c). Finally, we note that the classical Forchheimer-type corrections to the Darcy law for fluids lead to the case $p = q = \frac{3}{2}$.

2. The variational formulation. We begin by stating and resolving the stationary forms of our systems. Let Ω be a bounded domain in $\mathbb{R}^n$ with smooth boundary, $\Gamma = \partial\Omega$. Let $1 \le p < \infty$ and denote by $L^p(\Omega)$ the space of pth power-integrable functions on Ω, by $L^\infty(\Omega)$ the essentially bounded measurable functions, and the duality pairing by

$$(u, f)_{L(\Omega)} = \int_\Omega u(x)f(x)\,dx, \qquad u \in L^p(\Omega), \quad f \in L^{p'}(\Omega),$$

for any pair of conjugate powers, $\frac{1}{p} + \frac{1}{p'} = 1$. Let $C_0^\infty(\Omega)$ denote the space of infinitely differentiable functions with compact support in Ω. $W^{m,p}(\Omega)$ is the Banach space of functions in $L^p(\Omega)$ for which each partial derivative up to order m belongs to $L^p(\Omega)$, and $W_0^{m,p}(\Omega)$ is the closure of $C_0^\infty(\Omega)$ in $W^{m,p}(\Omega)$. See [1] for information on these Sobolev spaces. In addition, we shall be given for each $x \in \Omega$ a bounded domain Ω_x which lies locally on one side of its smooth boundary Γ_x. Let $1 < q < \infty$ and denote by $\gamma_x : W^{1,q}(\Omega_x) \to L^q(\Gamma_x)$ the *trace* map which assigns boundary values. Let T_x be the range of γ_x; this is a Banach space with the norm induced by γ_x from $W^{1,q}(\Omega_x)$. Since Γ_x is smooth, there is a unit outward normal $\nu_x(s)$ at each $s \in \Gamma_x$. Finally, we define $W_x^{1,q}(\Omega_x)$ to be that closed subspace consisting of those $\varphi \in W^{1,q}(\Omega_x)$ with $\gamma_x\varphi \in \mathbb{R}$, i.e., each $\gamma_x(\varphi)$ is *constant* almost everywhere on Γ_x. We shall denote by $\vec{\nabla}_y$ the gradient on $W^{1,q}(\Omega_x)$ and by $\vec{\nabla}$ the gradient on $W^{1,p}(\Omega)$.

The essential construction to be used below is an example of a *continuous direct sum* of Banach spaces. The special case that is adequate for our purposes can be described as follows. Let S be a measure space and consider the product (measure) space $Q = \Omega \times S$, where Ω has Lebesgue measure. If $U \in L^q(Q)$ then from the Fubini theorem it follows that $U(x)(z) \equiv U(x, z)$, $x \in \Omega$, $z \in S$ defines $U(x) \in L^q(S)$ at almost everywhere $x \in \Omega$, and for each $\Phi \in L^{q'}(Q)$

$$\int_\Omega \big(U(x), \Phi(x)\big)_{L(S)}\,dx \equiv \int_\Omega \left\{\int_S U(x,z)\Phi(x,z)\,dz\right\}dx = \iint_Q U\Phi.$$

Thus $L^q(Q)$ is naturally identified with $L^q(\Omega, L^q(S))$, the Bochner qth integrable (equivalence classes of) functions from Ω to $L^q(S)$.

In order to prescribe a measurable family of cells $\{\Omega_x, x \in \Omega\}$, set $S = \mathbb{R}^n$, let $Q \subset \Omega \times \mathbb{R}^n$ be a given measurable set for which each section $\Omega_x = \{y \in \mathbb{R}^n : (x,y) \in Q\}$ is a bounded domain in $\mathbb{R}^n$. By zero-extension we identify $L^q(Q) \hookrightarrow L^q(\Omega \times \mathbb{R}^n)$ and each $L^q(\Omega_x) \hookrightarrow L^q(\mathbb{R}^n)$. Thus we obtain from above

$$L^q(Q) \cong \Big\{U \in L^q\big(\Omega, L^q(\mathbb{R}^n)\big) : U(x) \in L^q(\Omega_x) \text{ a.e. } x \in \Omega\Big\}.$$

We shall denote the duality on this Banach space by

$$(U, \Phi)_{L(Q)} = \int_\Omega \left\{ \int_{\Omega_x} U(x,y)\Phi(x,y)\, dy \right\} dx \ ,$$
$$U \in L^q(Q), \ \Phi \in L^{q'}(Q).$$

The *state space* for our problems will be the product $L^1(\Omega) \times L^1(Q)$.

Note that $W^{1,q}(\Omega_x)$ is continuously imbedded in $L^q(\Omega_x)$, uniformly for $x \in \Omega$. It follows that the direct sum

$$\mathcal{W}_q \equiv L^q\big(\Omega, W^{1,q}(\Omega_x)\big) \equiv \Big\{U \in L^q(Q) : U(x) \in W^{1,q}(\Omega_x) \text{ a.e. } x \in \Omega \ ,$$
$$\text{and} \quad \int_\Omega \|U(x)\|^q_{W^{1,q}}\, dx < \infty\Big\}$$

is a Banach space. We shall use a variety of such spaces which can be constructed in this manner. Moreover, we shall assume that each Ω_x lies locally on one side of its boundary Γ_x, and Γ_x is a C^2-manifold of dimension $n-1$. We assume the trace maps $\gamma_x : W^{1,q}(\Omega_x) \to L^q(\Gamma_x)$ are *uniformly* bounded. Thus for each $U \in \mathcal{W}_q$ it follows that the *distributed trace* $\gamma(U)$ defined by $\gamma(U)(x,s) \equiv \gamma_x(U(x))(s)$, $s \in \Gamma_x$, $x \in \Omega$, belongs to $L^q(\Omega, L^q(\Gamma_x))$. The distributed trace γ maps $\mathcal{W}_q$ onto $\mathcal{T}_q \equiv L^q(\Omega, T_x) \hookrightarrow L^q(\Omega, L^q(\Gamma_x))$.

Next consider the collection $\{W^{1,q}_x(\Omega_x) : x \in \Omega\}$ of Sobolev spaces given above and denote by $\mathcal{W}_1 \equiv L^q(\Omega, W^{1,q}_x(\Omega_x))$ the corresponding direct sum. Thus for each $U \in \mathcal{W}_1$ it follows that the distributed trace $\gamma(U)$ belongs to $L^q(\Omega)$. We define $\mathcal{W}^{1,p}_0$ to be the subspace of those $U \in \mathcal{W}_1$ for which $\gamma(U) \in W^{1,p}_0(\Omega)$. Since $\gamma : \mathcal{W}_1 \to L^q(\Omega)$ is continuous, $\mathcal{W}^{1,p}_0$ is complete with the norm

$$\|U\|_{\mathcal{W}^{1,p}_0} \equiv \|U\|_{\mathcal{W}_q} + \|\gamma U\|_{W^{1,p}_0}.$$

This Banach space $W^{1,p}_0(\Omega) \times \mathcal{W}_q$ will be the *energy space* for the regularized problem (1.1) and $\mathcal{W}^{1,p}_0$ will be the energy space for the constrained problem in which (1.1c) is replaced by the Dirichlet condition (1.1c′). Note that $\mathcal{W}^{1,p}_0$ is identified with the closed subspace $\{[\gamma U, U] : U \in \mathcal{W}^{1,p}_0\}$ of $W^{1,p}_0(\Omega) \times \mathcal{W}_q$. Finally, we shall let $\mathcal{W}_0$ denote the kernel of γ, $\mathcal{W}_0 = \{U \in \mathcal{W}_q : \gamma U = 0 \text{ in } \mathcal{T}_q\}$.

We have defined $W^{1,q}_x(\Omega_x)$ to be the set of $w \in W^{1,q}(\Omega_x)$ for which $\gamma_x w$ is a constant multiple of $\mathbf{1}_x$, the constant function equal to one on Γ_x. Thus $W^{1,q}_x(\Omega_x)$ is the pre-image by γ_x of the subspace $\mathbb{R} \cdot \mathbf{1}_x$ of T_x. We specified the subspace $\mathcal{W}_1$ similarly as the subspace of $\mathcal{W}_q$ obtained as the pre-image by γ of the subspace $L^q(\Omega)$ of $\mathcal{T}_q$. To be precise, we denote by λ the map of $L^q(\Omega)$ into $\mathcal{T}_q$ given by $\lambda v(x) = v(x) \cdot \mathbf{1}_x$,

almost everywhere $x \in \Omega$, $v \in L^q(\Omega)$; λ is an isomorphism of $L^q(\Omega)$ onto a closed subspace of $\mathcal{T}_q$. The dual map λ' taking $\mathcal{T}_q'$ into $L^{q'}(\Omega)$ is given by

$$\lambda' g(v) = g(\lambda v) = \int_\Omega g_x(\mathbf{1}_x) \cdot v(x)\, dx, \qquad g \in \mathcal{T}_q', \quad v \in L^q(\Omega),$$

so we have $\lambda' g(x) = g_x(\mathbf{1}_x)$, almost everywhere $x \in \Omega$.

Moreover, when $g_x \in L^{q'}(\Gamma_x)$ it follows that

$$g_x(\mathbf{1}_x) = \int_{\Gamma_x} g_x(y)\, dy,$$

the integral of the indicated boundary functional. Thus, for $g \in L^{q'}(\Omega, L^{q'}(\Gamma_x)) \subset \mathcal{T}_q'$, $\lambda' g \in L^{q'}(\Omega)$ is given by

$$\lambda' g(x) = \int_{\Gamma_x} g_x(y)\, dy \qquad \text{a.e. } x \in \Omega. \tag{2.1}$$

The imbedding λ of $L^q(\Omega)$ into $\mathcal{T}_q$ and its dual map λ' will play an essential role in our system below.

We consider elliptic differential operators in divergence form as realizations of monotone operators from Banach spaces to their duals. Assume we are given $\widetilde{A} : \Omega \times \mathbb{R}^n \to \mathbb{R}^n$ such that for some $1 < p < \infty$, $g_1 \in L^{p'}(\Omega)$, $g_0 \in L^1(\Omega)$, c and $c_0 > 0$

(2.2a) $\widetilde{A}(x, \vec{\xi})$ is continuous in $\vec{\xi} \in \mathbb{R}^n$ and measurable in x, and
$$|\widetilde{A}(x, \vec{\xi})| \le c|\xi|^{p-1} + g_1(x),$$

(2.2b) $$\langle \widetilde{A}(x, \vec{\xi}) - \widetilde{A}(x, \vec{\eta}),\, \vec{\xi} - \vec{\eta} \rangle \ge 0,$$

(2.2c) $$\widetilde{A}(x, \vec{\xi}) \cdot \vec{\xi} \ge c_0 |\vec{\xi}|^p - g_0(x)$$
for a.e. $x \in \Omega$ and all $\vec{\xi}, \vec{\eta} \in \mathbb{R}^n$.

Then the global diffusion operator $\mathcal{A} : W_0^{1,p}(\Omega) \to W^{-1,p'}(\Omega)$ is given by

$$\mathcal{A}u(v) = \int_\Omega \widetilde{A}\big(x, \vec{\nabla} u(x)\big) \vec{\nabla} v(x)\, dx, \qquad u, v \in W_0^{1,p}(\Omega).$$

Thus, each $\mathcal{A}u$ is equivalent to its restriction to $C_0^\infty(\Omega)$, the distribution

$$Au \equiv \mathcal{A}u\big|_{C_0^\infty(\Omega)} = -\vec{\nabla} \cdot \widetilde{A}(\cdot, \vec{\nabla} u),$$

which specifies the value of this nonlinear elliptic divergence operator.

In order to specify a collection of local diffusion operators, $\mathcal{B}_x : W^{1,q}(\Omega_x) \to W^{1,q}(\Omega_x)'$, assume we are given $\widetilde{B} : Q \times \mathbb{R}^n \to \mathbb{R}^n$ such that for some $1 < q < \infty$, $h_1 \in L^{q'}(Q)$, $h_0 \in L^1(Q)$, c and $c_0 > 0$

(2.3a) $\widetilde{B}(x, y, \vec{\xi})$ is continuous in $\vec{\xi} \in \mathbb{R}^n$ and measurable in $(x, y) \in Q$, and
$$|\widetilde{B}(x, y, \vec{\xi})| \le c|\vec{\xi}|^{q-1} + h_1(x, y),$$

(2.3b) $$\langle \widetilde{B}(x, y, \vec{\xi}) - \widetilde{B}(x, y, \vec{\eta}),\, \vec{\xi} - \vec{\eta} \rangle \ge 0,$$

(2.3c) $$\widetilde{B}(x, y, \vec{\xi}) \cdot \vec{\xi} \ge c_0 |\vec{\xi}|^q - h_0(x, y)$$
for a.e. $(x, y) \in Q$ and all $\vec{\xi}, \vec{\eta} \in \mathbb{R}^n$.

Then define for each $x \in \Omega$

$$\mathcal{B}_x w(v) = \int_{\Omega_x} \widetilde{B}\big(x, y, \vec{\nabla}_y w(y)\big)\vec{\nabla}_y v(y)\, dy\,, \qquad w, v \in W^{1,q}(\Omega_x).$$

The elliptic differential operator on Ω_x is given by the *formal part* of $\mathcal{B}_x$, the distribution

$$B_x w \equiv \mathcal{B}_x w\big|_{C_0^\infty(\Omega_x)} = -\vec{\nabla}_y \cdot \widetilde{B}(x, \cdot, \vec{\nabla}_y w)$$

in $W_0^{1,q}(\Omega_x)'$. Also, we shall denote by $\mathcal{B} : \mathcal{W}_q \to \mathcal{W}_q'$ the *distributed* operator constructed from the collection $\{\mathcal{B}_x : x \in \Omega\}$ by

$$\mathcal{B}U(x) = \mathcal{B}_x\big(U(x)\big) \qquad \text{a.e. } x \in \Omega,\ U \in \mathcal{W}_q,$$

and we note that this is equivalent to

$$\mathcal{B}U(V) \equiv \int_\Omega \mathcal{B}_x\big(U(x)\big)V(x)\, dx, \qquad U, V \in \mathcal{W}_q.$$

The coupling term in our system will be given as a monotone graph which is a subgradient operator. Thus, assume $m : \mathbb{R} \to \mathbb{R}^+$ is convex and bounded by

$$m(s) \le C(|s|^q + 1), \qquad s \in \mathbb{R}, \tag{2.4}$$

hence, continuous. Then by

$$\tilde{m}(g) \equiv \int_\Omega \int_{\Gamma_x} m\big(g(x,s)\big)\, ds\, dx, \qquad g \in L^q\big(\Omega, L^q(\Gamma_x)\big),$$

we obtain the convex, continuous $\tilde{m} : L^q(\Omega, L^q(\Gamma_x)) \to \mathbb{R}^+$. Assume $\frac{1}{q} + \frac{1}{n} \ge \frac{1}{p}$ so that $W_0^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$, and consider the linear continuous maps

$$\lambda : W_0^{1,p}(\Omega) \to L^q\big(\Omega, L^q(\Gamma_x)\big), \qquad \gamma : \mathcal{W}_q \to L^q\big(\Omega, L^q(\Gamma_x)\big).$$

Then the composite function

$$M[u, U] \equiv \tilde{m}(\gamma U - \lambda u), \qquad u \in W_0^{1,p}(\Omega)\,,\ U \in \mathcal{W}_q,$$

is convex and continuous on $W_0^{1,p}(\Omega) \times \mathcal{W}_q$. The subgradients are directly computed by standard results [12]. Specifically, we have $\hat{g} \in \partial\tilde{m}(g)$ if and only if

$$\hat{g}(x,s) \in \partial m\big(g(x,s)\big) \quad \text{a.e. } s \in \Gamma_x, \quad \text{a.e. } x \in \Omega,$$

and we have $[f, F] \in \partial M[u, U]$ if and only if $f = -\lambda'(\mu)$ in $W^{-1,p'}(\Omega)$ and $F = \gamma'(\mu)$ in $\mathcal{W}_q'$ for some $\mu \in \partial\tilde{m}(\gamma U - \lambda u)$.

The following result gives sufficient conditions for the *stationary regularized problem* to be well posed.

PROPOSITION 1. *Assume* $1 < p, q$, $\frac{1}{q} + \frac{1}{n} \ge \frac{1}{p}$, *and define the spaces and operators* λ, γ *as above. Specifically, the sets* $\{\Omega_x : x \in \Omega\}$ *are uniformly bounded with smooth*

boundaries, and the trace maps $\{\gamma_x\}$ *are uniformly bounded. Let the functions* $\widetilde{A}, \widetilde{B}$, *and* m *satisfy* (2.2)–(2.4), *and assume in addition that*

$$m(s) \geq c_0|s|^q, \qquad s \in \mathbb{R}. \tag{2.5}$$

Then for each pair $f \in W^{-1,p'}(\Omega)$, $F \in \mathcal{W}_q'$ *there exists a solution of*

$$u \in W_0^{1,p}(\Omega) : A(u) - \lambda'(\mu) = f \quad \text{in } W^{-1,p'}(\Omega), \tag{2.6a}$$

$$U \in \mathcal{W}_q : \mathcal{B}(U) + \gamma'(\mu) = F \quad \text{in } \mathcal{W}_q', \tag{2.6b}$$

$$\mu \in L^{q'}\big(\Omega, L^{q'}(\Gamma_x)\big) : \mu \in \partial\tilde{m}(\gamma U - \lambda u). \tag{2.6c}$$

For any such solution we have

$$\int_{\Gamma_x} \mu(x,s)\,ds = \big\langle F(x), \mathbf{1}_x \big\rangle \qquad \text{a.e. } x \in \Omega, \tag{2.7}$$

where $\mathbf{1}_x$ *denotes the constant unit function in* $W^{1,q}(\Omega_x)$.

Proof. The system (2.6) is a "pseudo-monotone plus subgradient" operator equation of the form

$$\begin{aligned} &[u,U] \in W_0^{1,p}(\Omega) \times \mathcal{W}_q : \quad \text{for all } [v,V] \in W_0^{1,p}(\Omega) \times \mathcal{W}_q, \\ &\mathcal{A}u(v) + \mathcal{B}U(V) + \partial M[u,U]\big([v,V]\big) \ni f(v) + F(V). \end{aligned} \tag{2.6$'$}$$

It remains only to verify a coercivity condition, namely,

$$\frac{\mathcal{A}u(u) + \mathcal{B}U(U) + \tilde{m}(\gamma U - \lambda u)}{\|u\|_{W_0^{1,p}(\Omega)} + \|U\|_{\mathcal{W}_q}} \to +\infty \tag{2.8}$$

$$\text{as } \|u\|_{W_0^{1,p}(\Omega)} + \|U\|_{\mathcal{W}_q} \to +\infty.$$

Choose $k = \max\{|y_n| : y \in \Omega_x,\ x \in \Omega\}$ and let $\nu_x = (\nu_x^1, \cdots, \nu_x^n)$ be the unit normal on Γ_x. For $v \in W^{1,q}(\Omega_x)$ we have by Gauss' theorem

$$\begin{aligned} \int_{\Omega_x} \big(|v|^q + y_n q|v|^{q-1}\partial_n v\big) &= \int_{\Omega_x} \partial_n\big(y_n|v(y)|^q\big)\,dy \\ &= \int_{\Gamma_x} \nu_x^n(s) s_n |\gamma_x v(s)|^q\,ds. \end{aligned}$$

Hölder's inequality then shows

$$\|v\|_{L^q(\Omega_x)}^q \leq k\|\gamma_x v\|_{L^q(\Gamma_x)}^q + qk\|v\|_{L^q(\Omega_x)}^{q-1}\|\partial_n v\|_{L^q(\Omega_x)},$$

and from this follows

$$\|v\|_{L^q(\Omega_x)}^q \leq 2k\|\gamma_x v\|_{L^q(\Gamma_x)}^q + (2k)^q(q-1)^{q-1}\|\partial_n v\|_{L^q(\Omega_x)}^q$$

by Young's inequality. From here we obtain

$$c_0\|V\|_{L^q(Q)}^q \leq \|\gamma V\|_{L^q(\Omega, L^q(\Gamma_x))}^q + \|\nabla_y V\|_{L^q(Q)}^q, \qquad V \in \mathcal{W}_q. \tag{2.9}$$

Thus from the a priori estimate

$$\begin{aligned} &\mathcal{A}u(u) + \mathcal{B}U(U) + M(\gamma U - \lambda u) \\ &\geq c_0\|\vec{\nabla}u\|^p_{L^p(\Omega)} - \|g_0\|_{L^1(\Omega)} + c_0\|\vec{\nabla}_y U\|^q_{L^q(Q)} - \|h_0\|_{L^1(Q)} \\ &\quad + c_0\|\gamma U - \lambda u\|^q_{L^q(\Omega, L^q(\Gamma_x))}, \qquad u \in W_0^{1,p}(\Omega),\ U \in \mathcal{W}_q, \end{aligned} \tag{2.10}$$

the Poincaré-type inequality (2.9) and the equivalence of $\|\nabla u\|_{L^p(\Omega)}$ with the norm on $W_0^{1,p}(\Omega)$, we can obtain the coercivity condition (2.8). Specifically, if (2.8) is bounded by K, then (2.10) is bounded above by

$$\begin{aligned} &K\big(\|u\|_{W_0^{1,p}(\Omega)} + \|\nabla_y U\|_{L^q(Q)} + \|\gamma U\|_{L^q(\Omega, L^q(\Gamma_x))}\big) \\ &\leq K\big(\|u\|_{W_0^{1,p}(\Omega)} + \|\nabla_y U\|_{L^q(Q)} + \|\gamma U \\ &\quad - \lambda\varphi\|_{L^q(\Omega, L^q(\Gamma_x))} + \|\lambda u\|_{L^q(\Omega)}\big), \end{aligned}$$

and the last term is dominated by the first. This gives an explicit bound on each of these terms and, hence, on $\|u\|_{W_0^{1,p}(\Omega)} + \|U\|_{\mathcal{W}_q}$.

Finally, we apply (2.6b) to the function $V \in \mathcal{W}_q$ given by $V(x,y) = v(x)$ for some $v \in L^q(\Omega)$, and this shows

$$\mu(\gamma v) = \langle F, v\rangle$$

since $\mathcal{B}U(V) = 0$, and thus

$$\int_\Omega \lambda'\mu(x)v(x)\,dx = \mu(\lambda v) = \mu(\gamma v) = \int_\Omega \langle F(x), 1\rangle v(x)\,dx.$$

The identity (2.7) now follows from (2.1).

For the more general case of the *degenerate stationary problem* corresponding to (1.1), we obtain the following result.

COROLLARY 1. *Let $\varphi : \mathbb{R} \to \mathbb{R}^+$ and $\Phi : \mathbb{R} \to \mathbb{R}^+$ be convex and continuous, with $\varphi(0) = \Phi(0) = 0$, and assume*

$$\varphi(s) \leq C\big(|s|^q + 1\big), \qquad \Phi(s) \leq C\big(|s|^q + 1\big), \qquad s \in \mathbb{R}. \tag{2.11}$$

For each pair $f \in W^{-1,p'}(\Omega)$, $F \in \mathcal{W}_q'$, there exists a solution of

$$u \in W_0^{1,p}(\Omega) : a + A(u) - \lambda'(\mu) = f \quad \text{in } W^{-1,p'}(\Omega), \tag{2.12a}$$

$$U \in \mathcal{W}_q : b + \mathcal{B}(U) + \gamma'(\mu) = F \quad \text{in } \mathcal{W}_q', \tag{2.12b}$$

$$\mu \in \partial\tilde{m}(\gamma U - \lambda u) \quad \text{in } L^{q'}\big(\Omega, L^{q'}(\Gamma_x)\big), \tag{2.12c}$$

$$a \in \partial\varphi(u) \quad \text{in } L^{q'}(\Omega), \qquad b \in \partial\Phi(U) \quad \text{in } L^{q'}(Q). \tag{2.12d}$$

For any such solution we have

$$\int_{\Omega_x} b(x,y)\,dy + \int_{\Gamma_x} \mu(x,s)\,ds = \langle F(x), \mathbf{1}_x\rangle \quad \text{a.e. } x \in \Omega. \tag{2.13}$$

Proof. This follows as above but with the continuous convex function

$$\Psi[u,U] = \int_\Omega \varphi\big(u(x)\big)\,dx + \int_\Omega \int_{\Omega_x} \Phi\big(U(x,y)\big)\,dy\,dx + \tilde{m}(\gamma U - \lambda u), \qquad [u,U] \in W_0^{1,p}(\Omega) \times \mathcal{W}_q.$$

The subgradient can be computed termwise because the three terms are continuous on $L^q(\Omega)$, $L^q(Q)$, and $L^q(\Omega, L^q(\Gamma_x))$, respectively.

Remark. The lower bound (2.5) on $m(\cdot)$ may be deleted in Corollary 1 if such a lower estimate is known to hold for Φ. It is also unnecessary in the matched microstructure model; see below.

In order to prescribe the boundary condition (1.1c) explicitly, we develop an appropriate Green's formula for the operators $\mathcal{B}_x$.

Note that we can identify $L^{q'}(\Omega_x) \subset W^{-1,q'}(\Omega_x)$ since $W_0^{1,q}(\Omega_x)$ is dense in $L^q(\Omega_x)$, so it is meaningful to define

$$D_x \equiv \big\{w \in W^{1,q}(\Omega_x) : B_x w \in L^{q'}(\Omega_x)\big\}.$$

This is the domain for the *abstract Green's theorem.*

LEMMA 1. *There is a unique operator $\partial_x : D_x \to T_x'$ for which $\mathcal{B}_x w = B_x w + \gamma_x' \partial_x w$ for all $w \in D_x$. That is, we have*

$$\mathcal{B}_x w(v) = (B_x w, v)_{L(\Omega_x)} + \langle \partial_x w, \gamma_x v\rangle, \qquad v \in W^{1,q}(\Omega_x), \tag{2.14}$$

for every $w \in D_x$.

Proof. The strict morphism γ_x of $W^{1,q}(\Omega_x)$ onto T_x has a dual γ_x' which is an isomorphism of T_x' onto $W_0^{1,q}(\Omega_x)^\perp$, the annihilator in $W^{1,q}(\Omega_x)'$ of the kernel of γ_x. For each $w \in D_x$, the difference $\mathcal{B}_x w - B_x w$ is in $W_0^{1,q}(\Omega_x)^\perp$, so it is equal to $\gamma_x'(\partial_x w)$ for a unique element $\partial_x w \in T_x'$.

Remark. The identity (2.14) is a generalized decomposition of $\mathcal{B}_x$ into a partial differential operator on Ω_x and a boundary condition on Γ_x. If Γ_x is smooth, ν_x denotes the unit outward normal on Γ_x, and if $\tilde{B}(x,\cdot,\vec{\nabla}_y w) \in [W^{1,q'}(\Omega_x)]^n$, then $w \in D_x$ and from the classical Green's theorem we obtain

$$\mathcal{B}_x w(v) - (B_x w, v)_{L(\Omega_x)} = \int_{\Gamma_x} \widetilde{B}(x,s,\vec{\nabla}_y w)\vec{\nu}_x(s)\gamma v(s)\,ds, \qquad v \in W^{1,q}(\Omega_x).$$

Thus, $\partial_x w = \widetilde{B}(x,\cdot,\vec{\nabla}_y w)\cdot \vec{\nu}_x$ is the indicated normal derivative in $L^{q'}(\Gamma_x)$ when $\widetilde{B}(x,\cdot,\vec{\nabla}_y w)$ is as smooth as above, and so we can regard $\partial_x w$ in general as an extension of this nonlinear differential operator on the boundary.

The *formal part* of $\mathcal{B} : \mathcal{W}_q \to \mathcal{W}_q'$ is the operator $B : \mathcal{W}_q \to \mathcal{W}_0'$ given by the restriction $B(U) \equiv \mathcal{B}U|_{\mathcal{W}_0}$. Since $\mathcal{W}_0$ is dense in $L^q(Q)$ we can specify the domain

$$D \equiv \big\{U \in \mathcal{W}_q : B(U) \in L^{q'}(Q)\big\}$$

on which we obtain as before a distributed form of Green's theorem.

LEMMA 2. *There is a unique operator $\partial : D \to \mathcal{T}_q'$ such that*

$$\mathcal{B}(U)(V) = \big(B(U), V\big)_{L(Q)} + \langle \partial U, \gamma V\rangle, \qquad U \in D,\ V \in \mathcal{W}_q.$$

PROPOSITION 2. *Let the Sobolev spaces and trace operators be given as above. We summarize them in the following diagrams*:

$$\begin{array}{ccccc}
L^q(\Omega_x) & & L^q(\Gamma_x) & L^q(Q) & L^q(\Omega, L^q(\Gamma_x)) \\
\cup & & \cup & \cup & \cup \\
W^{1,q}(\Omega_x) & \xrightarrow{\gamma_x} & T_x & \mathcal{W}_q & \xrightarrow{\gamma} \quad \mathcal{T}_q \\
\cup & & \lambda_x \uparrow\downarrow & \cup & \lambda \uparrow\downarrow \\
W_x^{1,q}(\Omega_x) & \longrightarrow & \mathbb{R}\cdot\mathbf{1}_x & \mathcal{W}_1 & \xrightarrow{\gamma_1} \quad L^q(\Omega) \\
\cup & & \cup & \cup & \uparrow \\
W_0^{1,q}(\Omega_x) & \longrightarrow & \{0\} & \mathcal{W}_0 & \longrightarrow \quad \{0\}
\end{array}$$

in which γ_1 is the restriction of γ to $\mathcal{W}_1$. $W_0^{1,q}(\Omega_x)$, $\mathcal{W}_0$ are dense in $L^q(\Omega_x)$, $L^q(Q)$, respectively. Let operators $\mathcal{B}_x$, $x \in \Omega$, and $\mathcal{B}$ be given and define their formal parts B_x, B as above. Then construct the domains D_x, D and boundary operators ∂_x, ∂ as in Lemmas 1 *and* 2, *respectively. It follows that for any $U \in \mathcal{W}_q$,*

(a) *$BU(x) = B_x(U(x))$ in $W_0^{1,q}(\Omega_x)'$ for a.e. $x \in \Omega$, and $U \in D$ if and only if $U(x) \in D_x$ for a.e. $x \in \Omega$ and $x \mapsto B_xU(x)$ belongs to $L^{q'}(Q)$;*

(b) *for each $U \in D$,*

$$\partial U(x) = \partial_x\big(U(x)\big) \quad \text{in } T_x' \text{ for a.e. } x \in \Omega$$

and

$$\mathcal{B}U = BU + \gamma_1'(\lambda'\partial U) \quad \text{in } \mathcal{W}_1',$$

and for each $V \in \mathcal{W}_1$ we have

$$\int_\Omega \mathcal{B}_xU(x)\big(V(x)\big)\,dx = \int_Q B_xU(x)V(x)\,dy\,dx + \int_\Omega \big\langle \partial_xU(x), \mathbf{1}_x\big\rangle(\gamma_1V)(x)\,dx.$$

Proof. (a) For $V \in \mathcal{W}_0$ we obtain from the definitions of B, $\mathcal{B}$, and B_x, respectively,

$$\int_\Omega BU(x)V(x)\,dx = \int_\Omega \mathcal{B}U(V)\,dx = \int_\Omega \mathcal{B}_xU(x)\big(V(x)\big)\,dx = \int_\Omega B_xU(x)V(x)\,dx,$$

and so the first equality holds since $\mathcal{W}_0' = L^{q'}(\Omega, W_0^{1,q}(\Omega_x)')$. The characterization of D is immediate now.

(b) For $V \in \mathcal{W}_q$ we obtain from the definitions of $\gamma, \partial, \partial_x$, respectively, and (a)

$$\int_\Omega \partial U\big(\gamma_xV(x)\big)\,dx = \int_\Omega \partial U(\gamma V)\,dx = \int_\Omega (\mathcal{B}U - BU)(x)V(x)\,dx = \int_\Omega \partial_x\big(U(x)\big)\gamma_xV(x)\,dx.$$

Since the range of γ is $\mathcal{T}_q' = L^{q'}(\Omega, T_x')$, the first equality follows. The second is immediate from Lemma 2 since on $\mathcal{W}_1$, $\gamma = \lambda \circ \gamma_1$ and $\gamma' = \gamma_1' \lambda'$, and the third follows from the preceding remarks.

COROLLARY 2. *In the situation of Corollary* 1, $f \in L^{q'}(\Omega)$ *and* $F \in L^{q'}(Q)$ *if and only if* $Au \in L^{q'}(\Omega)$ *and* $B(U) \in L^{q'}(Q)$, *and in that case the solution satisfies almost everywhere*

$$
\begin{aligned}
&a(x) \in \partial\varphi\big(u(x)\big),\\
&a(x) + Au(x) + \int_{\Omega_x} b(x,y)\,dy = f(x) + \int_{\Omega_x} F(x,y)\,dy, \qquad x \in \Omega,\\
&u(s) = 0, \qquad s \in \Gamma,\\
&b(x,y) \in \partial\Phi\big(U(x,y)\big), \quad b(x,y) + BU(x,y) = F(x,y), \qquad y \in \Omega_x,\\
&\mu(x,s) \in \partial m\big(\gamma U(x,s) - u(x)\big), \qquad \partial_x\big(U(x)\big)(s) + \mu(x,s) = 0, \quad s \in \Gamma_x.
\end{aligned}
$$

Finally, we note that corresponding results for the *stationary matched microstructure model* are obtained directly by specializing the system (2.6′) to the space $\mathcal{W}_0^{1,p}$. This is identified with $\{[\gamma U, U] : U \in \mathcal{W}_0^{1,p}\}$ as a subspace of $W_0^{1,p}(\Omega) \times \mathcal{W}_q$, and we need only to restrict the solution $[u, U]$ and the test functions $[v, V]$, $v = \gamma V$, to this subspace to resolve the matched model. Then the coupling term M does not occur in the system; see the proof of Proposition 1, especially for the coercivity. These observations yield the following analogous results for the matched microstructure model.

PROPOSITION 1′. *Assume* $1 < p, q, \frac{1}{q} + \frac{1}{n} \geq \frac{1}{p}$, *and define the spaces and operators* λ, γ *as before. Let the functions* $\widetilde{A}, \widetilde{B}$, *and* m *satisfy* (2.2)–(2.4). *Then for each pair* $f \in W^{-1,p'}(\Omega)$, $F \in \mathcal{W}_1'$ *there exists a unique solution of*

$$
\begin{aligned}
&u \in W_0^{1,p}(\Omega) : A(u) = f + \langle F, \mathbf{1}\rangle \quad \text{in } W^{-1,p'}(\Omega), && (2.15a)\\
&U \in \mathcal{W}_1 : B(U) = F \quad \text{in } \mathcal{W}_0', && (2.15b)\\
&\gamma U = \lambda u \quad \text{in } L^q(\Omega) \subset \mathcal{T}_q. && (2.15c)
\end{aligned}
$$

COROLLARY 1′. *Suppose* φ, Φ *are given as before and assume* (2.11). *For* f, F *as above there exists a unique solution of*

$$
\begin{aligned}
&u \in W_0^{1,p}(\Omega) : a + \langle b, \mathbf{1}\rangle + A(u) = f + \langle F, \mathbf{1}\rangle \quad \text{in } W^{-1,p'}(\Omega), && (2.16a)\\
&U \in \mathcal{W}_1 : b + B(U) = F \quad \text{in } \mathcal{W}_0', && (2.16b)\\
&\gamma U = \lambda u \quad \text{in } L^q(\Omega) \subset \mathcal{T}_q, && (2.16c)\\
&a \in \partial\varphi(u) \ \text{ in } L^{q'}(\Omega), \qquad b \in \partial\Phi(U) \quad \text{in } L^{q'}(Q). && (2.16d)
\end{aligned}
$$

In addition, $f \in L^{q'}(\Omega)$ and $F \in L^{q'}(Q)$ if and only if $Au \in L^{q'}(\Omega)$ and $B(U) \in L^{q'}(Q)$, and in that case the solution satisfies almost everywhere

$$a(x) \in \partial\varphi\big(u(x)\big),$$

$$a(x) + Au(x) + \int_{\Omega_x} b(x,y)\,dy = f(x) + \int_{\Omega_x} F(x,y)\,dy\,, \qquad x \in \Omega,$$

$$u(s) = 0, \qquad s \in \Gamma\,,$$

$$b(x,y) \in \partial\Phi\big(U(x,y)\big), \quad b(x,y) + BU(x,y) = F(x,y), \qquad y \in \Omega_x,$$

$$U(x,s) = u(x), \qquad s \in \Gamma_x\,.$$

Remark. For the very special case of $p = q \geq 2$ and $a(u) = u$, $b(U) = U$ in the situation of Proposition 1 it follows from [7] or [20] that the Cauchy–Dirichlet problem for (1.1) is well posed in the space $L^p(0,T; W_0^{1,p}(\Omega) \times \mathcal{W}_p)$ with appropriate initial data $u(x,0)$, $U(x,y,0)$ and source functions $f(x,t)$, $F(x,y,t)$. A similar remark holds in the case of Proposition 1′ for the matched model with (1.1c′). These restrictive assumptions will be substantially relaxed in the next section.

Furthermore, variational inequalities may be resolved for problems corresponding to either the regularized or the matched microstructure model by adding the indicator function of a convex constraint set to the convex function Ψ. Thus such problems can be handled with constraints on the global variable u, the local variables U, or their difference $\lambda u - \gamma U$ on the interface.

3. The L^r-operators. Assume we are in the situation of Proposition 1. We define a relation or multi-valued operator $\mathbb{C}_2$ on the Hilbert space $L^2(\Omega) \times L^2(Q)$ as follows: $\mathbb{C}_2[u,U] \ni [f,F]$ if and only if

$$u \in L^2(\Omega) \cap W_0^{1,p}(\Omega) : \mathcal{A}(u) - \lambda'\mu = f \in L^2(\Omega), \tag{3.1a}$$

$$U \in L^2(Q) \cap \mathcal{W}_q : \mathcal{B}(U) + \gamma'\mu = F \in L^2(Q) \tag{3.1b}$$

$$\text{for some } \mu \in \partial\tilde{m}(\gamma U - \lambda u) \quad \text{in } L^{q'}(\Omega, L^{q'}(\Gamma_x)).$$

Thus, $\mathbb{C}_2$ is the restriction of (2.6) to $L^2(\Omega) \times L^2(Q)$. Note that $\lambda'\mu \in L^2(\Omega)$ by (2.7).

LEMMA 3. *If $\sigma : \mathbb{R} \to \mathbb{R}$ is monotone, Lipschitz, and $\sigma(0) = 0$, then for each pair*

$$\mathbb{C}_2[u_j, U_j] \ni [f_j, F_j]\,, \qquad j = 1,2,$$

there follows

$$\big(f_1 - f_2,\, \sigma(u_1 - u_2)\big)_{L^2(\Omega)} + \big(F_1 - F_2,\, \sigma(U_1 - U_2)\big)_{L^2(Q)} \geq 0.$$

Proof. Since σ is Lipschitz and $\sigma(0) = 0$, we have $\sigma(u_1 - u_2) \in W_0^{1,p}(\Omega)$ and $\sigma(U_1 - U_2) \in \mathcal{W}_q$. Also the chain rule applies to these functions, so we compute

$$\big\langle \mathcal{A}u_1 - \mathcal{A}u_2,\, \sigma(u_1, u_2)\big\rangle$$

$$= \int_\Omega \big(\widetilde{A}(x, \vec{\nabla} u_1) - \widetilde{A}(x, \nabla u_2)\big)\vec{\nabla}(u_1 - u_2)\sigma'(u_1 - u_2)\,dx,$$

$$\big\langle \mathcal{B}U_1 - \mathcal{B}U_2,\, \sigma(U_1 - U_2)\big\rangle$$

$$= \int_\Omega \int_{\Omega_x} \big(\widetilde{B}(x,y,\vec{\nabla}_y U_1) - \widetilde{B}(x,y,\vec{\nabla}_y U_2)\big)$$

$$\vec{\nabla}_y(U_1 - U_2)\sigma'(U_1 - U_2)\,dy\,dx\,.$$

Both of these are nonnegative because of (2.1b), (2.2b), and $\sigma' \geq 0$. The remaining term to check is

$$\big(-\lambda'(\mu_1-\mu_2),\, \sigma(u_1-u_2)\big)_{L^2(\Omega)} + \big\langle \gamma'(\mu_1-\mu_2),\, \sigma(U_1-U_2)\big\rangle$$

$$= \int_\Omega \int_{\Gamma_x} \big(\mu_1(x,s)-\mu_2(x,s)\big)\big(\sigma(\gamma U_1-\gamma U_2)-\sigma(\lambda u_1-\lambda u_2)\big)\, ds\, dx.$$

Since σ is a monotone function and ∂m is a monotone graph, this integrand is nonnegative and the result follows.

As a consequence of Lemma 3 with $\sigma(s)=s$, the operator $\mathbb{C}_2$ is monotone on the Hilbert space $L^2(\Omega)\times L^2(Q)$. Moreover, we obtain the following.

PROPOSITION 3. *The operator $\mathbb{C}_2$ is maximal monotone on $L^2(\Omega)\times L^2(Q)$. Let $j:\mathbb{R}\to\mathbb{R}^+$ be convex, lower-semicontinuous, and $j(0)=0$. If ∂m is a function, then $\mathbb{C}_2$ is also single valued and*

$$\big(\mathbb{C}_2[u_1,U_1]-\mathbb{C}_2[u_2,U_2],[\sigma_1,\sigma_2]\big)_{L^2(\Omega)\times L^2(Q)} \geq 0 \tag{3.2}$$

for any selections $\sigma_1\in\partial j(u_1-u_2)$ in $L^2(\Omega)$ and $\sigma_2\in\partial j(U_1-U_2)$ in $L^2(Q)$.

Proof. To show $\mathbb{C}_2$ is maximal monotone it suffices to show that for any pair $[f,F]\in L^2(\Omega)\times L^2(Q)$ there is a solution of

$$u\in L^2(\Omega)\cap W_0^{1,p}(\Omega): u+\mathcal{A}(u)-\lambda'(\mu)=f \ \text{ in } W^{-1,p'}(\Omega), \tag{3.3a}$$

$$U\in L^2(Q)\cap \mathcal{W}_q: U+\mathcal{B}(U)+\gamma'(\mu)=F \ \text{ in } \mathcal{W}_q' \tag{3.3b}$$

$$\mu\in L^{q'}\big(\Omega, L^{q'}(\Gamma_x)\big): \mu\in\partial\tilde{m}(\gamma U-\lambda u). \tag{3.3c}$$

The existence of a (unique) solution of (3.3) follows as in Proposition 1, but by considering the pseudomonotone operator $[\mathcal{A},\mathcal{B}]$ on the product space $L^2(\Omega)\cap W_0^{1,p}(\Omega)\times L^2(Q)\cap\mathcal{W}_q$ and the convex function, $\frac{1}{2}\|u\|^2_{L^2(\Omega)}+\frac{1}{2}\|U\|^2_{L^2(Q)}+\tilde{m}(\gamma U-\lambda u)$, on that space.

To establish the estimate (3.2), we consider the lower-semicontinuous convex function

$$\tilde{j}[u,U]=\int_\Omega\Big(j\big(u(x)\big)+\int_{\Omega_x} j\big(U(x,y)\big)\,dy\Big)\,dx, \tag{3.4}$$
$$[u,U]\in L^2(\Omega)\times L^2(Q).$$

The subgradient of $\tilde{j}$ is given on this product space by

$$\tilde{\sigma}=[\sigma_1,\sigma_2]\in\partial\tilde{j}[u,U] \ \text{ if and only if}$$

$$\tilde{\sigma}[v,V]=\int_\Omega\Big(\sigma_1(x)v(x)+\int_{\Omega_x}\sigma_2(x,y)V(x,y)\,dy\Big)\,dx,$$
$$[v,V]\in L^2(\Omega)\times L^2(Q),$$

where

$$\sigma_1(x)\in\partial j\big(u(x)\big), \quad \text{a.e. } x\in\Omega,$$
$$\sigma_2(x,y)\in\partial j\big(U(x,y)\big), \quad \text{a.e. } (x,y)\in Q.$$

The Yoshida approximation $\tilde{j}_\varepsilon$ of $\tilde{j}$ is given as in (3.4) but with j replaced by j_ε. Since the derivative of j_ε is Lipschitz, monotone, and contains the origin, it follows by

Lemma 3 that the special case of (3.2) with j_ε is true. Thus, $\mathbb{C}_2$ is $\partial\tilde{j}$-monotone [8] and the desired result follows, since the single-valued $\mathbb{C}_2$ equals its minimal section.

We define the realization of (2.6) in $L^r(\Omega) \times L^r(Q)$, $1 \le r < \infty$, as follows. For $r \ge 2$, $\mathbb{C}_r$ is the restriction of $\mathbb{C}_2$ to $L^r(\Omega) \times L^r(Q)$, and for $1 \le r < 2$, $\mathbb{C}_r$ is the closure in $L^r(\Omega) \times L^r(Q)$ of $\mathbb{C}_2$.

COROLLARY 3. *The operator $\mathbb{C}_r$ is m-accretive in $L^r(\Omega) \times L^r(Q)$ for $1 \le r < \infty$.*

Proof. Let $(I + \varepsilon\mathbb{C}_2)([u_j, U_j]) \ni [f_j, F_j]$, $j = 1, 2$, and assume $[f_j, F_j] \in L^r(\Omega) \times L^r(Q)$ if $r \ge 2$. Set $j(s) = |s|^r$, $s \in \mathbb{R}$. From Proposition 4.7 of [8] it follows that

$$\|[u_1 - u_2, U_1 - U_2]\|_{L^r(\Omega)\times L^r(Q)} \le \|[f_1 - f_2, F_1 - F_2]\|_{L^r(\Omega)\times L^r(Q)}.$$

Taking $[f_2, F_2] = [0, 0]$, we see that $L^r(\Omega) \times L^r(Q)$ is invariant under $(I + \varepsilon\mathbb{C}_2)^{-1}$, and then the estimate shows this operator is a contraction on that space. We have $Rg(I + \varepsilon\mathbb{C}_r) = L^r(\Omega) \times L^r(Q)$ directly from the definition for $r \ge 2$, and for $1 \le r < 2$, $Rg(I + \varepsilon\mathbb{C}_r) \supset L^2(\Omega) \times L^2(Q)$, which is dense, so the result follows easily.

Remarks. The Cauchy–Dirichlet problem for the regularized model (1.1) is well posed in $L^r(\Omega) \times L^r(Q)$ when $a(u) = u$, $b(U) = U$, and $r > 1$. This follows from Corollary 3 and the theory of evolution equations generated by m-accretive operators in a uniformly convex Banach space. For example, from [19] we recall the following:

If $\mathfrak{f} \in W^{1,1}(0, T; X)$ and $w_0 \in D(\mathbb{C}_r)$, where $\mathbb{C}_r$ is m-accretive on the uniformly convex Banach space X, then there exists a unique Lipschitz function $w : [0, T] \to X$ for which

$$w'(t) + \mathbb{C}_r\big(w(t)\big) \ni \mathfrak{f}(t), \quad \text{a.e. } t \in (0, T),$$

$$w(t) \in D(\mathbb{C}_r) \quad \text{for all } t \in [0, T], \quad \text{and}$$

$$w(0) = w_0.$$

See [4] for details (Theorem III.2.3) and references. By applying this result to the operator $\mathbb{C}_r$ given in $X \equiv L^r(\Omega) \times L^r(Q)$, $1 < r < \infty$, we obtain a generalized strong solution $w(t) = [u(t), U(t)]$ of the system

$$\frac{\partial u(x,t)}{\partial t} + Au(x,t) + \int_{\Omega_x} \frac{\partial U(x,y,t)}{\partial t}\, dy = f(x,t) + \int_{\Omega_x} F(x,y,t)\, dy, \qquad x \in \Omega,\ t \in (0,T),$$

$$u(s,t) = 0, \qquad s \in \Gamma,$$

$$\frac{\partial U(x,y,t)}{\partial t} + BU(x,y,t) = F(x,y,t), \qquad y \in \Omega_x,$$

$$\mu(x,s,t) \in \partial m\big(U(x,s,t) - u(x,t)\big), \qquad \partial_x U(x,s,t) + \mu(x,s,t) = 0, \quad s \in \Gamma_x,$$

$$u(x,0) = u_0(x), \quad U(x,y,0) = U_0(x,y).$$

The restrictions on the data $\mathfrak{f}(t) = [f(t), F(t)]$ and $w_0 = [u_0, U_0]$ can be considerably relaxed in the Hilbert space case $r = 2$ [8].

By applying Proposition 1′ similarly, it follows that corresponding results for the matched model are obtained. Thus we obtain a generalized strong solution in $L^r(\Omega) \times L^r(Q)$, $1 < r < \infty$, of the system

$$\frac{\partial u(x,t)}{\partial t} + Au(x,t) + \int_{\Omega_x} \frac{\partial U(x,y,t)}{\partial t}\, dy = f(x,t) + \int_{\Omega_x} F(x,y,t)\, dt, \qquad x \in \Omega, \quad t \in (0,T),$$

$$u(s,t) = 0, \qquad s \in \Gamma,$$

$$\frac{\partial U(x,y,t)}{\partial t} + BU(x,y,t) = F(x,y,t), \qquad y \in \Omega_x,$$

$$U(x,s,t) = u(x,t), \qquad s \in \Gamma_x,$$

$$u(x,0) = u_0(x), \qquad U(x,y,0) = U_0(x,y).$$

This follows as above from the analogue of Proposition 3 and Corollary 3.

We return to consider the fully nonlinear model (1.1). The generator of this evolution system will be obtained by closing up the composition of $\mathbb{C}_2$ with the inverse of $[\partial\varphi, \partial\Phi]$ in $L^1(\Omega) \times L^1(Q)$. Thus, we begin with the following.

DEFINITION. $\mathbb{C}[a,b] \ni [f,F]$ if $\mathbb{C}_2[u,U] \ni [f,F]$ and $a \in \partial\varphi(u)$ in $L^2(\Omega)$, $b \in \partial\Phi(U)$ in $L^2(Q)$ for some pair $[u,U]$ as in (3.1).

LEMMA 4. *The operator $\mathbb{C}$ is accretive on $L^1(\Omega)\times L^1(Q)$ if either ∂m is a function or if both $\partial\varphi$ and $\partial\Phi$ are functions.*

Proof. Let $\varepsilon > 0$ and suppose that $(I + \varepsilon\mathbb{C})[a_j, b_j] \ni [f_j, F_j]$ for $j = 1, 2$. Thus we have $\varepsilon\mathbb{C}_2[u_j, U_j] \ni [f_j - a_j, F_j - b_j]$, $a_j \in \partial\varphi(u_j)$, $b_j \in \partial\Phi(U_j)$ as above. First we choose $\sigma(s) = \operatorname{sgn}_\delta^+(s)$, the Yoshida approximation of the maximal monotone sgn^+, apply Lemma 3 and obtain

$$\big(a_1 - a_2, \operatorname{sgn}_\delta^+(u_1 - u_2)\big)_{L^2(\Omega)} + \big(b_1 - b_2, \operatorname{sgn}_\delta^+(U_1 - U_2)\big)_{L^2(Q)} \le \|(f_1 - f_2)^+\|_{L^1(\Omega)} + \|(F_1 - F_2)^+\|_{L^1(Q)}.$$

If $\partial\varphi$ and $\partial\Phi$ are functions, then

$$(a_1 - a_2)\operatorname{sgn}_0^+(u_1 - u_2) = (a_1 - a_2)^+,$$
$$(b_1 - b_2)\operatorname{sgn}_0^+(U_1 - U_2) = (b_1 - b_2)^+,$$

so letting $\delta \to 0$ gives

$$\|(a_1 - a_2)^+\|_{L^1(\Omega)} + \|(b_1 - b_2)^+\|_{L^1(Q)} \le \|(f_1 - f_2)^+\|_{L^1(\Omega)} + \|(F_1 - F_2)^+\|_{L^1(Q)}. \tag{3.5}$$

The same holds for negative parts, so it follows that $(I + \varepsilon\mathbb{C})^{-1}$ is an order-preserving contraction with respect to $L^1(\Omega) \times L^1(Q)$ for each $\varepsilon > 0$.

Next we suppose ∂m is a function. Choose $j(s) = s^+$, so that $\partial j = \operatorname{sgn}^+$, and then set

$$\sigma_1(x) = \operatorname{sgn}_0^+(u_1 - u_2 + a_1 - a_2) \in \operatorname{sgn}^+(u_1 - u_2) \cap \operatorname{sgn}^+(a_1 - a_2),$$
$$\sigma_2(x,y) = \operatorname{sgn}_0^+(U_1 - U_2 + b_1 - b_2) \in \operatorname{sgn}^+(U_1 - U_2) \cap \operatorname{sgn}^+(b_1 - b_2).$$

Proposition 3 applies here to give (3.5). A similar estimate for negative parts yields the result.

Although $\mathbb{C}$ is not accretive on L^r for $1 < r$, we can obtain L^∞ estimates when the graphs $\partial\varphi$, $\partial\Phi$ are not too dissimilar.

COROLLARY 4. *If* $(I + \varepsilon\mathbb{C})[a, b] \ni [f, F]$ *with* $\varepsilon > 0$, *then*

$$\begin{aligned} \|a^+\|_{L^\infty(\Omega)} &\le \max\big(a_0(k), \|f^+\|_{L^\infty(\Omega)}\big), \\ \|b^+\|_{L^\infty(Q)} &\le \max\big(b_0(k), \|F^+\|_{L^\infty(Q)}\big), \end{aligned} \tag{3.6}$$

where $k \equiv \max(a_0^{-1}(\|f^+\|_{L^\infty}), b_0^{-1}(\|F^+\|))$.

Remarks. Here a_0 is the *minimal section* $(\partial\varphi)_0$, a_0^{-1} is the minimal section of $(\partial\varphi)^{-1}$, and b_0, b_0^{-1} are defined similarly from $\partial\Phi$. Specifically, we obtain an explicit a priori bound on $\|a^+\|_{L^\infty(\Omega)}$ and $\|b^+\|_{L^\infty(Q)}$ when $\|f^+\|_{L^\infty(\Omega)} in Rg(\partial\varphi)$ and $\|F^+\|_{L^\infty(Q)} \in Rg(\partial\Phi)$. By similar estimates for negative parts, we obtain explicit estimates on $\|a\|_{L^\infty(\Omega)}$ and $\|b\|_{L^\infty(Q)}$ for *any* pair $f \in L^\infty(\Omega)$, $F \in L^\infty(\Omega)$ if $Rg(\partial\varphi) = \mathbb{R}$ and $Rg(\partial\Phi) = \mathbb{R}$ or (trivially) if both $Rg(\partial\varphi)$ and $Rg(\partial\Phi)$ are bounded in $\mathbb{R}$. Finally, we note that in the special case $\varphi = \Phi$, we obtain

$$\max\big(\|a^+\|_{L^\infty(\Omega)}, \|b^+\|_{L^\infty(Q)}\big) \le \max\big(\|f^+\|_{L^\infty(\Omega)}, \|F^+\|_{L^\infty(Q)}\big).$$

Proof. By the choice of $k \ge 0$ we have

$$\partial\varphi(k) \ni \ell_1 \ge \|f^+\|_{L^\infty}, \qquad \partial\Phi(k) \ni \ell_2 \ge \|F^+\|_{L^\infty}$$

for some pair ℓ_1, ℓ_2. Subtract these from the operator equation, multiply by either

$$\operatorname{sgn}_\delta^+(u - k), \qquad \operatorname{sgn}_\delta^+(U - k)$$

or by

$$\operatorname{sgn}_0^+(a - \ell_1 + u - k), \qquad \operatorname{sgn}_0^+(b - \ell_2 + U - k)\ ,$$

depending on whether $\partial\varphi$ and $\partial\Phi$ are functions or ∂m is a function, respectively. Apply Lemma 3 and let $\delta \to 0$ or apply Proposition 3, respectively, to obtain

$$\begin{aligned} &\|(a - \ell_1)^+\|_{L^1(\Omega)} + \|(b - \ell_2)^+\|_{L^1(Q)} \\ &\quad \le \|(f^+ - \ell_1)^+\|_{L^1(\Omega)} + \|(F^+ - \ell_2)^+\|_{L^1(Q)}. \end{aligned}$$

The right side is zero, so the result follows.

PROPOSITION 4 (Moser). *Let* $(u, U) \in W_0^{1,p}(\Omega) \times \mathcal{W}_q$ *be a solution to*

$$\begin{aligned} &\mathcal{A}(u) - \lambda'\mu \ni f \quad \text{in } W^{-1,p'}(\Omega), \\ &\mathcal{B}(U) + \gamma'\mu \ni F \quad \text{in } \mathcal{W}_q', \\ &\qquad \mu \in \partial m(\gamma U - \lambda u). \end{aligned}$$

(a) *If* $(f, F) \in L^{\tau'}(\Omega) \times L^{\tau'}(Q)$ *with* $\tau' > \frac{n}{p}$, *and*

$$\widetilde{A}(x, \vec{\xi}) \cdot \vec{\xi} \ge c_0|\xi|^p - g_0(x) \tag{2.2c$'$}$$

where $g_0 \in L^{\tau'}(\Omega)$, *then* $u \in L^\infty(\Omega)$.

(b) *If, additionally,* $F \in L^\infty[\Omega; L^{t'}(\Omega_x)]$ *with* $t' > \frac{n}{q}$,

$$\widetilde{B}(x, y, \vec{\xi}) \cdot \vec{\xi} \geq c_0|\xi|^q - h_0(x, y), \tag{2.3c'}$$

where $h_0 \in L^\infty[\Omega; L^{t'}(\Omega_x)]$, *and* m *satisfies the growth condition* (2.5) *and* $m(0) = 0$, *then* $U \in L^\infty(Q)$.

Proof. (a) Estimate (2.7) of Proposition 1 shows that $\lambda'\mu \in L^{\tau'}(\Omega)$, so that

$$\mathcal{A}(u) = f - \lambda'\mu = \tilde{f} \in L^{\tau'}(\Omega).$$

Lemma 3 of [23] can now be used to conclude $u \in L^\infty(\Omega)$.

(2) Define $\widetilde{U} = U - u\mathbf{1}$. Since $\mathcal{B}(\widetilde{U}) = B(U)$, it follows that

$$\mathcal{B}(\widetilde{U}) + \gamma'\mu = F \text{ in } \mathcal{W}_q', \qquad \mu \in \partial m(\gamma\widetilde{U}),$$

and for almost every $x \in \Omega$, and every $V \in \mathcal{W}_q$

$$(*) \qquad \int_{\Omega_x} \widetilde{B}\big(x, \cdot, \nabla_y \widetilde{U}(x)\big) \cdot \nabla_y V(x) + \int_{\Gamma_x} \mu(x)\gamma V(x) = \int_{\Omega_x} F(x, \cdot)V(x),$$

with $\mu(x) \in \partial m(\gamma\widetilde{U}(x))$. We will now use Moser iteration with $(*)$ to conclude $\|\widetilde{U}(x)\|_{L^\infty(\Omega_x)} \leq C$, where C is to be chosen independently of $x \in \Omega$.

If $\widetilde{U}(x) \in L^r(\Omega_x)$ ($r = q$ suffices for the first iterate), define $s = 1 + (r - t/tq)$ ($\frac{1}{t} + \frac{1}{t'} = 1$). Let $H \in C^1(\mathbb{R})$ satisfy $H(s) = |s|^s$ if $|s| \leq s_0$, H affine for $|s| > s_0$, and define $G(s) = \int_0^s |H'(\xi)|^q \, d\xi$. Since H has linear growth, it follows that $G(\widetilde{U}) \in \mathcal{W}_q$. Substituting $G(\widetilde{U})$ for V in $(*)$ gives

$$\int_{\Omega_x} \widetilde{B}\big(x, \cdot, \nabla_y \widetilde{U}(x)\big) \cdot \nabla_y \widetilde{U}(x) G'\big(\widetilde{U}(x)\big) + \int_{\Gamma_x} \mu(x)\gamma G\big(\widetilde{U}(x)\big) = \int_{\Omega_x} F(x, \cdot) G\big(\widetilde{U}(x)\big).$$

The first term of the formula above is bounded below using (2.3c'). To estimate the second term, use

(i) $\mu\widetilde{U} \geq m(\widetilde{U})$ (as $m(0) = 0$), and

(ii) $\operatorname{sgn}(\widetilde{U}) = \operatorname{sgn}\big(G(\widetilde{U})\big)$ (so that $G(\widetilde{U})/\widetilde{U} \geq 0$ when $\widetilde{U} \neq 0$)

to get

$$\begin{aligned} \mu G(\widetilde{U}) = \mu\widetilde{U}\, G(\widetilde{U})/\widetilde{U} &\geq m(\widetilde{U})\, G(\widetilde{U})/\widetilde{U} \\ &\geq c_0|\widetilde{U}|^q\, G(\widetilde{U})/\widetilde{U} = c_0|\widetilde{U}|^{q-1}|G(\widetilde{U})|, \end{aligned}$$

$$c_0 \int_{\Omega_x} |\nabla_y \widetilde{U}|^q G'(\widetilde{U}) + c_0 \int_{\Gamma_x} |\widetilde{U}|^{q-1}|G(\widetilde{U})| \leq \int_{\Omega_x} FG(\widetilde{U}) + h_0 G'(\widetilde{U}).$$

The first term may be written as $|\nabla_y H(\widetilde{U})|^q$ which, using the Sobolev embedding theorem, is bounded below by

$$c(\varepsilon)\|H(\widetilde{U})\|^q_{L^{\frac{nq}{n-q}}(\Omega_x)} - \varepsilon \int_{\Gamma_x} |H(\widetilde{U})|^q,$$

where $\varepsilon > 0$ can be chosen arbitrarily small (see (2.9)). The right-hand side is bounded using Hölder's inequality.

$$c(\varepsilon)\|H(\widetilde{U})\|^q_{L^{nq/(n-q)}(\Omega_x)} + \int_{\Gamma_x} |\widetilde{U}|^{q-1}|G(\widetilde{U})| - \varepsilon|H(\widetilde{U})|^q$$

$$\leq \frac{1}{c_0}\left(\|F\|_{L^{t'}(\Omega_x)}\|G(\widetilde{U})\|_{L^t(\Omega_x)} + \|h_0\|_{L^{t'}(\Omega_x)}\|G'(\widetilde{U})\|_{L^t(\Omega_x)}\right)$$

when $s_0 \to \infty$, $H(\widetilde{U}) \to |\widetilde{U}|^s$, and $|\widetilde{U}|^{q-1}|G(\widetilde{U})| \to \eta(r)|\widetilde{U}|^{sq}$, where $\eta(r) = \frac{t}{r}s^q = \frac{t}{r}(1 + (r - t/qt))^q$. If ε is chosen as $\varepsilon = \min_{q \leq r < \infty} \eta(r)$, it follows that

$$\|\widetilde{U}\|_{L^{sq(n/n-q)}(\Omega_x)} \leq (cs)^{1/s} \max\left[1, \|\widetilde{U}\|_{L^r(\Omega_x)}\right].$$

The result now follows by iteration of the above estimate.

THEOREM 1. *Assume the hypotheses of Proposition* 1, *Corollary* 1, *Lemma* 4, *and Proposition* 4. *Also, assume that* $Rg(\partial\varphi)$ *and* $Rg(\partial\Phi)$ *are both bounded or that both are equal to* $\mathbb{R}$. *Then* $\overline{\mathbb{C}}$, *the closure of* $\mathbb{C}$ *in* $L^1(\Omega) \times L^1(Q)$, *is m-accretive.*

Proof. Let $f \in L^\infty(\Omega)$ and $F \in L^\infty(Q)$. Corollary 1 asserts there is a solution of (2.12). If the graphs $\partial\varphi$ and $\partial\Phi$ have bounded range, then $a \in L^\infty(\Omega)$, $b \in L^\infty(Q)$, and it follows from Proposition 4 that $u \in L^2(\Omega)$ and $U \in L^2(Q)$. This shows $\mathbb{C}_2[u, U] \ni [a - f, b - F]$, so $(I + \mathbb{C})([a, b]) \ni [f, F]$. Thus, $Rg(I + \overline{\mathbb{C}})$ is dense in and, hence, equal to $L^1(\Omega) \times L^1(Q)$.

If the ranges of $\partial\varphi$ and $\partial\Phi$ equal $\mathbb{R}$, then by Corollary 4 any solution satisfies

$$\|a\|_{L^\infty(\Omega)} \leq K, \qquad \|b\|_{L^\infty(Q)} \leq K,$$

where K depends on f and F. Replace $\partial\varphi, \partial\Phi$ by the appropriately truncated $\partial\varphi_K, \partial\Phi_K$.

The solution with these truncated graphs, then, is a solution of the equation with the original graphs, so we are done.

COROLLARY 5. *Under the hypotheses of Theorem* 1, *problem* (1.1) *has a unique generalized solution* $(a, b) \in C[0, T; L^1(\Omega) \times \overline{L^1(Q)}]$, *provided the data satisfy* $(f, F) \in L^1[0, T; L^1(\Omega) \times L^1(Q)]$, *and* $(a(0), b(0)) \in \overline{D(\mathbb{C})}$.

This follows from the Crandall–Liggett theorem [9], which is proved by showing that the step functions (a^N, b^N), constructed from solutions to the differencing scheme

$$(a^n, b^n) - (a^{n-1}, b^{n-1}) + \tau\mathbb{C}(a^n, b^n) \ni \tau(f^n, F^n) \tag{3.7}$$

($\tau = \frac{T}{N}$), converge uniformly when the operator $\mathbb{C}$ is m-accretive. Benilan [6] proves that these generalized solutions are unique.

All of our results hold for the matched microstructure model problem. Specifically, Lemma 4 and Corollary 4 are obtained from Proposition 3, and Proposition 4 is actually simpler for the matched problem. The analogues of Theorem 1 and Corollary 5 show that the matched problem (1.1a), (1.1b), (1.1c′) has a unique generalized solution $(a, b) \in C[0, T; L^1(\Omega) \times L^1(Q)]$.

The next theorem shows that if the data is further restricted, the generalized solutions will satisfy the partial differential equation (1.1). The following notation is used:

$$L^r(T) = L^r\left[0, T; L^r(\Omega) \times L^r(Q)\right], \qquad 1 \leq r \leq \infty,$$
$$V = W_0^{1,p}(\Omega) \times \mathcal{W}_q,$$
$$\mathcal{V}(T) = L^p\left[0, T; W_0^{1,p}(\Omega)\right] \times L^q[0, T; \mathcal{W}_q],$$
$$\widehat{\mathcal{V}}(T) = W^{1,p'}\left[0, T; W^{-1,p'}(\Omega)\right] \times W^{1,q'}[0, T; \mathcal{W}_q'].$$

THEOREM 2. *Assume the hypotheses of Theorem* 1 *hold and in addition that* $(f,F) \in L^1(T) \cap \mathcal{V}(T)'$ *and* $(a(0),b(0)) \in \overline{D(\mathbb{C})} \cap V'$. *Then the generalized solutions of Corollary* 5 *satisfy*

$$(a,b) \in \widehat{\mathcal{V}}(T), \qquad (u,U) \in \mathcal{V}(T), \tag{3.8a}$$

$$\frac{\partial}{\partial t}(a,b) + \big(\mathcal{A}(u) - \lambda'\mu,\ \mathcal{B}(U) + \gamma'\mu\big) = (f,F) \text{ in } \mathcal{V}(T)', \tag{3.8b}$$

$$(a,b) \in \big(\partial\phi(u), \partial\Phi(U)\big), \qquad \mu \in \partial\tilde{m}(\lambda u - \gamma U). \tag{3.8c}$$

Proof. The results of Grange and Mignot [15] show that the step functions (a^N, b^N) and (u^N, U^N) generated from the differencing scheme (3.7) converge weakly in $\widehat{\mathcal{V}}(T)$ and $\mathcal{V}(T)$, respectively. Moreover, equation (3.8) will be satisfied in the limit, provided the weak limits (a,b) and (u,U) satisfy $(a,b) \in (\partial\phi(u), \partial\Phi(U))$. To establish this inclusion, let $(v,V) \in \mathcal{V}(T)$ and $(\tilde{a},\tilde{b}) \in (\partial\phi(v), \partial\Phi(V))$. The growth conditions on ϕ and Φ guarantee that (a^N, b^N) and $(\tilde{a},\tilde{b}) \in \mathcal{V}(T)'$ are functions, so it is possible to define $(a^N - \tilde{a},\ b^N - \tilde{b})_s$ to be the pair of functions truncated above and below by $\pm s$ $(s > 0)$. This pair of functions is bounded in $L^\infty(T)$ and converges in $L^1(T)$ to $(a - \tilde{a},\ b - \tilde{b})_s$, and so converges in $L^r(T)$ for $1 \le r < \infty$. If $r \ge \max(p', q')$, it follows that $L^r(T) \subset \mathcal{V}(T)'$, so the sequence $(a^N - \tilde{a},\ b^N - \tilde{b})_s$ converges strongly in $\mathcal{V}(T)'$. The monotonicity of $\partial\phi$ and $\partial\Phi$ imply

$$0 \le \Big\langle (a^N - \tilde{a},\ b^N - \tilde{b})_s,\ (u^N - v,\ U^n - V) \Big\rangle.$$

Passing to the limit as $N \to \infty$ and then letting $s \to \infty$ yields

$$0 \le \Big\langle (a - \tilde{a},\ b - \tilde{b})\ ,\ (u - v,\ U - V) \Big\rangle, \qquad (\tilde{a},\tilde{b}) \in \big(\partial\phi(v), \partial\Phi(V)\big).$$

Since $(\partial\phi(\cdot), \partial\Phi(\cdot))$ is maximally monotone, it follows that $(a,b) \in (\partial\phi(u), \partial\Phi(U))$.

Finally, we note that the corresponding solution of the matched problem satisfies

$$(a,b) \in \widetilde{\mathcal{V}}(T), \qquad (\gamma U, U) \in \mathcal{V}(T), \tag{3.8a$'$}$$

$$\frac{\partial}{\partial t}(a,b) + \big(\mathcal{A}(\gamma U), \mathcal{B}(U)\big) = (f,F) \text{ in } \mathcal{V}_0(T)', \tag{3.8b$'$}$$

$$(a,b) \in \big(\partial\varphi(\gamma U), \partial\Phi(U)\big), \qquad U \in \mathcal{W}_0, \tag{3.8c$'$}$$

where the space $\mathcal{V}_0(T)$ is given by

$$\mathcal{V}_0(T) \equiv \Big\{ U \in L^q[0,T;\mathcal{W}_1] : \gamma(U) \in L^p\big[0,T;W_0^{1,p}(\Omega)\big] \Big\}$$

with the appropriate norm for which $(\gamma(U), U) \in \mathcal{V}(T)$ for each $U \in \mathcal{V}_0(T)$.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] T. ARBOGAST, J. DOUGLAS, AND U. HORNUNG, *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal., 21 (1990), pp. 823–836.

[3] ———, *Modeling of naturally fractured reservoirs by formal homogenization techniques*, to appear.

[4] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff Groningen, the Netherlands, 1976.

[5] J. A. BARKER, *Block-geometry functions characterizing transport in densely fissured media*, J. Hydrology, 77 (1985), pp. 263–279.

[6] PH. BENILAN, *Équations d'évolution dans un espace de Banach quelconque et applications*, Thesis, Orsay, 1972.

[7] H. BREZIS, *Problemes unilateraux*, J. Math. Pures Appl., 51 (1972), pp. 1–164.

[8] ———, *Opérateurs Maximaux Monotones et Semigroupes de Contractions dans les Espaces de Hilbert*, North Holland, Amsterdam, 1973.

[9] M. G. CRANDALL, *An introduction to evolution governed by accretive operators*, in Dynamical Systems—An International Symposium, L. Cesari, J. Hale, and J. LaSalle, eds., Academic Press, New York, 1976, pp. 131–165.

[10] P. F. DIESLER, JR. AND R.H. WILHELM, *Diffusion in beds of porous solids: measurement by frequency response techniques*, Indust. Engrg. Chem., 45 (1953), 1219–1227.

[11] J. DOUGLAS, P. J. PAES LEME, T. ARBOGAST, AND T. SCHMITT, *Simulation of flow in naturally fractured reservoirs*, in Proceedings, Ninth SPE Symposium on Reservoir Simulation, Society of Petroleum Engineers, Dallas, TX, 1987, pp. 271–279, Paper SPE 16019.

[12] I. EKLAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North Holland, Amsterdam, 1976.

[13] A. FRIEDMAN AND A. TZAVARAS, *A quasilinear parabolic system arising in modelling of catalytic reactors*, J. Differential Equations, 70 (1987), pp. 167–196.

[14] M. TH. VAN GENUCHTEN AND F. N. DALTON, *Models for simulating salt movement in aggregated field soils*, Geoderma, 38 (1986), pp. 165–183.

[15] O. GRANGE AND F. MIGNOT, *Sur la resolution d'une équation et d'une inéquation paraboliques non-lineares*, J. Funct. Anal., 11 (1972), pp. 77–92.

[16] U. HORNUNG, *Miscible displacement in porous media influenced by mobile and immobile water*, in Nonlinear Partial Differential Equations, P. Fife and P. Bates, eds., Springer, New York, 1988.

[17] U. HORNUNG AND W. JÄGER, *Diffusion, convection, adsorption, and reaction of chemicals in porous media*, preprint 522, Sonderforschungs bereich 123, Heidelberg, FRG, 1982.

[18] U. HORNUNG AND R. E. SHOWALTER, *Diffusion models for fractured media*, J. Math. Anal. Appl., 147 (1990), pp. 69–80.

[19] T. KATO, *Accretive operators and nonlinear evolution equations in Banach spaces*, Non-Linear Functional Analysis, Proc. Sympos. Pure Math., Vol. 18, American Mathematical Society, Providence, RI, 1970.

[20] J. L. LIONS, *Quelques Méthods de Resolutions des Problèmes aux Limites Non Linéares*, Dunod, Paris, 1969.

[21] J. B. ROSEN, *Kinetics of a fixed bed system for solid diffusion into spherical particles*, J. Chem. Phys., 20 (1952), pp. 387–394.

[22] J.B. ROSEN AND W.E. WINSHE, *The admittance concept in the kinetics of chromatography*, J. Chem. Phys., 18 (1950), 1587–1592.

[23] R. E. SHOWALTER AND N.J. WALKINGTON, *A diffusion system for fluid in fractured media*, Differential and Integral Equations, 3 (1990), pp. 219–236.

[24] ———, *Micro-structure models of diffusion in fissured media*, J. Math. Anal. Appl., 155 (1991), pp. 1–20.

[25] CH. VOGT, *A homogenization theorem leading to a Volterra integro-differential equation for permeation chromotography*, preprint 155, Sonderforschungs bereich 123, Heidelberg, FRG, 1982.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1723–1737, November 1991

011

ON PARABOLIC VOLTERRA EQUATIONS IN SEVERAL SPACE DIMENSIONS*

HONG-MING YIN†

Abstract. In this paper some parabolic integrodifferential equations in n-space dimensions are studied. For the solution of such a linear equation, the classical Schauder and $L_p(Q_T)$ estimates are derived. As a direct corollary, the continuous dependence and the uniqueness of the solution for the full nonlinear integrodifferential equation are obtained. Then the global solvability for a class of quasilinear integrodifferential equations is considered. Using the method of energy estimates and the integral iteration technique along with the results of linear equations, an a priori estimate in the classical space $C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)$ is deduced and the global solution of our problem is established by the continuation argument similar to the case of a parabolic equation.

Key words. nonlinearity, a priori estimates, global existence

AMS(MOS) subject classifications. 35Q25, 45K05

1. Introduction.

Let $T > 0$ and $Q_T = \Omega \times (0, T]$, $T > 0$, where Ω is an open bounded region in R^n with a smooth boundary $\partial\Omega$. In this paper we first study the linear equation:

$$u_t - L_1 u = \int_0^t L_2 u d\tau + f(x,t), \qquad \text{in } Q_T \tag{1.1}$$

where L_1 is a linear elliptic operator:

$$L_1 u = a_{ij}(x,t)u_{x_i x_j} + b_i(x,t)u_{x_i} + c(x,t)u$$

and L_2 is an arbitrary differential operator of the second order:

$$L_2 u = c_{ij}(x,t,\tau)u_{x_i x_j} + d_i(x,t,\tau)u_{x_i} + e(x,t,\tau)u.$$

For the solution of this linear equation, we shall derive a priori Schauder and $W_p^{2,1}(Q_T)$ estimates which are well known for solutions of parabolic partial differential equations. Using these results, we then consider a class of nonlinear equations:

$$\begin{aligned} u_t = {} & \frac{\partial}{\partial x_i}\left[a_{ij}\left(x,t,u,\int_0^t u d\tau\right)\frac{\partial u}{\partial x_j}\right] \\ & + \int_0^t \left[\frac{\partial}{\partial x_i} b_i(x,t,\tau,u) + c(x,t,\tau,u)\right] d\tau \quad \text{in } Q_T \end{aligned} \tag{1.2}$$

and establish the global solvability. Here and throughout the paper we shall use the standard notation: $u_x = (u_{x_1}, \cdots, u_{x_n}) \in R^n, u_{xx} = (u_{x_i x_j}) \in R^{n^2}$ and the repeated subscript implies summation from 1 to n.

The equation (1.1) or its more general form

$$u_t = A(x,t,u,u_x,u_{xx}) + \int_0^t B(x,t,\tau,u,u_x,u_{xx})d\tau \quad \text{in } Q_T, \tag{1.3}$$

* Received by the editors November 26, 1990; accepted for publication January 28, 1991. This work is partially funded by the National Science and Engineering Research Council of Canada.

† Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada L8S 4K1. Present address, Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 1A1.

where the ellipticity condition

$$\frac{\partial A(x,t,u,p,r)}{\partial r_{ij}}\xi_i\xi_j \geq A_0|\xi|^2 \qquad (A_0 > 0) \tag{1.4}$$

holds, is often called a Volterra integrodifferential equation of parabolic type. This kind of equation represents many mathematical models in physical and engineering fields such as heat transfer in a material with memory, the propagation of disturbances in viscous media, and so on (cf. [19]). The nonlinear equation (1.2) can be regarded as the generalization of some practical models in physics and biology. As an example, we consider the nonlinear version of the nuclear reactor model (cf. [21, p. 172]

$$u_t - \frac{\partial}{\partial x_i}[a_{ij}(x,t,u,v)u_{x_j}] = u(\lambda - bv), \tag{1.5}$$

$$v_t = -cv + au, \tag{1.6}$$

where u represents the fast neutron density and v the fuel temperature. It is easy to see by (1.6) that

$$v(x,t) = v_0(x) + a\int_0^t e^{-c(t-\tau)}u(x,\tau)d\tau.$$

Hence, the system (1.5)–(1.6) can be written as the form which is similar to the equation (1.2). Similar models can be found in [21] such as the Fitzhugh–Nagumo system and the degenerate Volterra–Lotka model in which only one species diffuses is included. Recently, much attention has been received in the study of the well-posedness of the problem as well as its numerical solution. A large group of people write the equation (1.3) as a Volterra type integrodifferential equation in an abstract Banach space:

$$\frac{du(t)}{dt} = A(t,u(t)) + \int_0^t B(s,u(s))ds.$$

They employ the analytic semigroup theory to study the solvability of the problem and other properties (cf., e.g., [4], [7], [9], [12], [13], [16], [18]). There are also a number of authors who take the derivative with respect to t in the equation (1.3) and obtain a third-order partial differential equation which is of mixed type (cf. [3], [5], [6], [8], [20], [22], etc.). For such an equation, they use the Galerkin approach and some special treatments to investigate the local and global solvability of the problem when the principal part $A(x,t,u,u_x,u_{xx})$ is a linear elliptic operator. These methods are powerful and many good results have been carried out previously under certain conditions. For a nonlinear elliptic operator A, the global solvability of the problem becomes much more difficult. Some effort has been devoted to this investigation (cf., e.g., [4], [16], [18], etc.). More recently, Yin [24] and [25] considers the problem from a rather different viewpoint. He regards the equation (1.1) as a parabolic one with a perturbation (the integral term). Hence, the theory of parabolic equations can be applied to the study of the problem. Indeed, as will be seen in §2 of this paper, the solution of the linear equation possesses many features such as Schauder and $W_p^{2,1}(Q_T)$ estimates which play important roles when studying nonlinear equations. In [24] and [25], the author establishes the global existence of the solutions for two classes of nonlinear integrodifferential equations in one space dimension. In the present

paper we will show global solvability for the nonlinear equation (1.2). We first derive a uniform bound by applying the integral iteration technique. Then we employ an analogous approach as in [2] to derive the $W_p^{2,1}$-estimate. Finally, the results for the solutions of linear problems are used to deduce an a priori estimate in $C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)$ and then the global solution is obtained by the method of continuity.

Remark. The argument presented in this paper can be used to deal with a more general equation than (1.2). However, these generalizations are left to the reader.

In §2 the Schauder and $L_p(Q_T)$ estimates are proved, and then as a direct corollary the continuous dependence and uniqueness of the solution for the full nonlinear equation (1.3) are obtained. In §3 we establish the global existence of the solution for the equation (1.3) associated with the suitable initial and boundary conditions.

For the reader's convenience, we list the following notation: Let $Q_t = \Omega \times (0,t]$ for $t \in (0,T]$, $S = \partial\Omega$, $S_T = S \times [0,T]$, $P_T = Q_T \times [0,T]$.

$$
\begin{aligned}
&||u||_{C(\bar{Q}_T)} = \max_{\bar{Q}_T} |u(x,t)|,\\
&[u]_{C^{\alpha,\alpha/2}(Q_t)} = \sup_{(x,s)\neq(x',s'),0\leq s,s'\leq t} \frac{|u(x,s)-u(x',s')|}{|x-x'|^\alpha + |s-s'|^{\alpha/2}},\\
&||u||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)} = ||u||_{C(\bar{Q}_T)} + [u]_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)},\\
&||u||_{C^{1+\alpha,\frac{1+\alpha}{2}}(\bar{Q}_T)} = ||u||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)} + ||u_x||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)},\\
&||u||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)} = ||u||_{C^{1+\alpha,\frac{1+\alpha}{2}}(\bar{Q}_T)} + ||u_{xx}||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)} + ||u_t||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)},\\
&||u||^p_{W^1_p(\Omega)} = \int_\Omega (u^p + u_x^p)dx,\\
&||w||^p_{W_p^{2,1}(Q_T)} = \int\int_{Q_T} [|u|^p + |u_x|^p + |u_{xx}|^p + |u_t|^p]dxdt
\end{aligned}
$$

for $p > 1$ arbitrary. Moreover, on $P_T = Q_T \times (0,T]$,

(1.7)
$$
[u]^*_{C^{\alpha,\alpha/2,\alpha/2}(P_T)} = \sup_{(x,t,s)\neq(x',t',s'),0\leq t,t',s,s'\leq T} \frac{|u(x,t,s)-u(x',t',s')|}{|x-x'|^\alpha + |t-t'|^{\alpha/2} + |s-s'|^{\alpha/2}}.
$$

The norms $||\cdot||^*_{C^{\alpha,\alpha/2,\alpha/2}(P_T)}, \cdots, ||\cdot||^*_{C^{2+\alpha,1+\alpha/2,1+\alpha/2}(P_T)}$ can be defined similarly.

2. Schauder and $L_P(Q_T)$ estimates. It is well known that the theory of Schauder and $L_p(Q_T)$ estimates are important in the study of parabolic equations. In this section we shall derive such a priori estimates for the solutions of linear integrodifferential equations. We begin with the following Gronwall's inequality without its proof (cf. [14, Lemma 7.1.1].

LEMMA 2.1. *Assume that the function $g(t)$ is a nonnegative, nondecreasing, and integrable function on $[0,T]$. The function $f(t)$ satisfies*

$$
0 \leq f(t) \leq g(t) + \int_0^t \frac{f(\tau)}{(t-\tau)^\alpha} d\tau, \qquad 0 \leq t \leq T, \tag{2.1}
$$

where $0 \leq \alpha < 1$. Then

$$
f(t) \leq C(\alpha,T)g(t), \qquad t \in [0,T],
$$

where $C(\alpha, T)$ depends only on α and the upper bound of T.

The following lemma is elementary.

LEMMA 2.2. *Let $g(x,t)$ be a function in $C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)(0 \le \alpha < 1)$ and $G(x,t)$ be a function defined as $\int_0^t h(x,t,\tau)g(x,\tau)d\tau$, where $h(x,t,s)$ is defined on $Q_T \times [0,T]$ and Hölder continuous with respect to x,t and s with the exponent α, $\frac{\alpha}{2}$ and $\frac{\alpha}{2}$, respectively. Then*

$$||G||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_t)} \le C\int_0^t \left[2 + \frac{1}{(t-\tau)^\alpha}\right] ||g(x,s)||_{C^{\alpha,\alpha/2}(\bar{Q}_\tau)}d\tau, \tag{2.2}$$

where C depends on the Hölder norm of $h(x,t,s)$ and T.

Proof. By the definition, we have

$$||G(x,\tau)||_{C(\bar{Q}_t)} \le C\int_0^t ||g(x,s)||_{C(\bar{Q}_\tau)}d\tau.$$

For any two points (x,s) and (x',s') in Q_t, we assume $s' > s$ without loss of generality.

$$\begin{aligned}
&\frac{|G(x,s) - G(x',s')|}{|x-x'|^\alpha + |s-s'|^{\frac{\alpha}{2}}} \\
&= \frac{|\int_0^s h(x,s,z)g(x,z)dz - \int_0^{s'} h(x',s',z)g(x',z)dz|}{|x-x'|^\alpha + |s-s'|^{\frac{\alpha}{2}}} \\
&\le \frac{\int_0^s [h(x,s,z) - h(x',s',z)]g(x,z)dz + \int_0^s |h||g(x,z) - g(x',z)|dz + \int_s^{s'} |h||g(x',z)|dz}{|x-x'|^\alpha + |s-s'|^{\frac{\alpha}{2}}}.
\end{aligned}$$

We split the above fraction into three terms denoted by K_1, K_2, and K_3. It is clear by the Hölder continuity of $h(x,t,s)$ that

$$K_1 \le C\int_0^s ||g(x,s)||_{C(\bar{Q}_\tau)}d\tau$$

and

$$\begin{aligned}
K_2 &\le C\int_0^s \frac{|g(x,z) - g(x',z)|}{|x-x'|^\alpha}dz \\
&\le C\int_0^{s'} ||g(x,\tau)||_{C^{\alpha,\alpha/2}(\bar{Q}_z)}dz.
\end{aligned}$$

To estimate K_3, noting that $s \le z \le s'$, we have

$$\frac{1}{|s'-s|} \le \frac{1}{|s'-z|}.$$

Therefore,

$$\begin{aligned}
K_3 &\le C\int_s^{s'} \frac{|g(x',z)|}{|s'-z|^{\frac{\alpha}{2}}}dz \\
&\le C\int_0^{s'} \frac{|g(x',z)|}{|s'-z|^{\frac{\alpha}{2}}}dz \\
&\le C\int_0^{s'} \frac{1}{|s'-z|^{\frac{\alpha}{2}}}||g||_{C(Q_z)}dz.
\end{aligned}$$

By combining the estimates for K_1, K_2, and K_3 and then taking the supremum over $\bar{Q}_t$ on both sides of the above inequality in conjunction with the maximum norm of $G(x,t)$, we obtain

$$||G||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_t)} \leq C \sup_{0\leq s'\leq t} \int_0^{s'} \left[2 + \frac{1}{(s'-\tau)^\alpha}\right] ||g(x,s)||_{C^{\alpha,\alpha/2}(\bar{Q}_\tau)} d\tau.$$

Since $||g(x,s)||_{C^{\alpha,\alpha/2}(\bar{Q}_\tau)}$ is a monotone increasing function of τ, by the integration by parts, we see that the function

$$\int_0^{s'} \left[2 + \frac{1}{(s'-\tau)^\alpha}\right] ||g(x,s)||_{C^{\alpha,\alpha/2}(\bar{Q}_\tau)} d\tau$$

is monotone increasing. Thus, we end our proof.

Now we consider the following linear initial boundary value problem:

$$u_t - L_1 u = \int_0^t L_2 u dx + f(x,t), \qquad Q_T, \tag{2.3}$$

$$u(x,t) = g(x,t), \qquad (x,t) \in S_T, \tag{2.4}$$

$$u(x,0) = u_0(x), \qquad x \in \bar{\Omega}, \tag{2.5}$$

where the operators L_1 and L_2 are the same as in (1.1).

To obtain a Schauder type estimate, we assume the following conditions hold.

H(2.1). The coefficients of the operators L_1 and L_2 are in the Banach spaces $C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)$ and $C^{\alpha,\frac{\alpha}{2},\frac{\alpha}{2}}(\bar{P}_T)$, respectively. Moreover, there exists a constant A_0 which may depend on Q_T such that

$$\sum_{i,j=1}^n [||a_{ij}||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)} + ||c_{ij}||^*_{C^{\alpha,\alpha/2,\alpha/2}(P_T)}] + \sum_{i=1}^n [||b_i||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)} + ||d_i||^*_{C^{\alpha,\alpha/2,\alpha/2}(P_T)}]$$
$$+ ||c||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)} + ||e||^*_{C^{\alpha,\alpha/2,\alpha/2}(P_T)} \leq A_0.$$

The function $f(x,t) \in C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)$. Moreover, the ellipticity condition

$$a_{ij}(x,t)\xi_i\xi_j \geq a_0|\xi|^2 \qquad (a_0 > 0),$$

for $\xi \in R^n$ holds.

H(2.2). There exists a function $\Phi(x,t) \in C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)$ such that $\Phi(x,t) = g(x,t)$ on S_T and $\Phi(x,0) = u_0(x)$ on $\bar{\Omega}$.

Now we state the Schauder estimate.

THEOREM 2.1. *Under the conditions* H(2.1), (2.2), *the solution* $u(x,t)$ *of the problem* (2.3)–(2.5) *satisfies*

$$||u||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)} \leq C[||\Phi||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)} + ||f||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)}]. \tag{2.6}$$

Proof. Let $u(x,t)$ be an arbitrary classical solution of the problem (2.3)–(2.5). Regarding the right side of the equation (2.3) as an inhomogeneous term $F(x,t)$, we apply the Schauder estimate (cf. [10], Thm. 6, Chap. 3) for a parabolic equation to obtain

$$||u||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_t)} \leq C[||F||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_t)} + ||\Phi||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_t)}],$$

where the constant C depends only on a_0, A_0, and Q_T.

Since the coefficients of L_2 are Hölder continuous over P_T, by a direct calculation and using Lemma 2.2, we find that

$$||F||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_t)} \le C\int_0^t \left[2+\frac{1}{(t-\tau)^{\frac{\alpha}{2}}}\right]||u||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_\tau)}d\tau + ||f||_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_t)},$$

where the constant C depends only on the Hölder norms of the coefficients of L_2. Thus, the desired result follows from Gronwall's inequality (Lemma 2.1).

From the above result, we immediately have the following corollaries.

COROLLARY 2.1. *Under conditions* H(2.1), (2.2), *the problem* (2.3)–(2.5) *has a unique classical solution.*

COROLLARY 2.2. *Assume that the functions A and B in* (1.3) *are smooth with respect to all of their arguments and the ellipticity condition* (1.4) *is satisfied. Let $u_1(x,t)$ and $u_2(x,t)$ be two classical solutions of the* (1.3) *corresponding to the initial-boundary data $(u_{01}(x), g_1(x,t))$ and $(u_{02}(x), g_2(x,t))$, respectively. Then*

$$||u_1-u_2||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)} \le C||\Phi||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)},$$

where $\Phi(x,t) \in C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)$ coincides with $g_1(x,t)-g_2(x,t)$ on S_T and $u_{01}(x)-u_{02}(x)$ on $\{(x,t): x\in\bar{\Omega}, t=0\}$.

In fact, the existence of the solution for (2.3)–(2.5) can be obtained by the method of continuity or the bootstrap argument (see the proof of Theorem 3.1). To prove the result of Corollary 2.2, let $w(x,t) = u_1(x,t) - u_2(x,t)$. Then $w(x,t)$ satisfies the linear equation (2.3) with the Hölder coefficients of L_1 and L_2. Hence, the Schauder estimate (2.6) implies the desired result.

To study nonlinear parabolic integrodifferential equations, we often need the following $W_p^{2,1}$-estimate.

THEOREM 2.2. *Assume that $a_{ij}(x,t)$ is continuous on $\bar{Q}_T$ and*

$$a_{ij}(x,t)\xi_i\xi_j \ge a_0|\xi|^2 \qquad (a_0>0).$$

Moreover, the coefficients of L_1 and L_2 are in $L^\infty(Q_T)$ and $L^\infty(P_T)$ $f(x,t)\in L^p(Q_T)$. There exists a function $\Phi(x,t)\in W_p^{2,1}(Q_T)$ with $\Phi(x,0)=u_0(x)$ on $\bar{\Omega}$ and $\Phi(x,t)=g(x,t)$ on S_T. Then

$$||u||_{W_p^{2,1}(Q_T)} \le C[||\Phi(x,t)||_{W_p^{2,1}(Q_T)} + ||f(x,t)||_{L^p(Q_T)}], \tag{2.7}$$

where C depends only on the continuity modulus of $a_{ij}(x,t)$ and the bounds of the coefficients of L_1 and L_2.

Proof. We regard the right side of (2.3) as $F(x,t)$ again and apply $W_p^{2,1}(Q_T)$-estimate for parabolic equations (cf. [17, Thm. 9.1, p. 341]) to obtain

$$\begin{aligned}||u||_{W_p^{2,1}(Q_T)} &\le C[||\Phi||_{W_p^{2,1}(Q_T)} + ||F||_{L_p(Q_T)}]\\ &\le C[||\Phi||_{W_p^{2,1}(Q_T)} + \int_0^T ||u||_{W_p^{2,1}(Q_t)}dt].\end{aligned}$$

Since T can be arbitrary, an application of Gronwall's inequality implies our result.

3. Global solvability. In this section we are concerned with the global existence of the solution for the equation

$$u_t = \frac{\partial}{\partial x_i}\left[a_{ij}\left(x,t,u,\int_0^t u d\tau\right)\frac{\partial u}{\partial x_j}\right] + \int_0^t \left[\frac{\partial}{\partial x_i} b_i(x,t,\tau,u) + c(x,t,\tau,u)\right] d\tau \quad \text{in } Q_T \tag{3.1}$$

subject to the initial and boundary conditions

$$u(x,t) = 0, \qquad (x,t) \in S_T, \tag{3.2}$$

$$u(x,0) = u_0(x), \qquad x \in \bar{\Omega}, . \tag{3.3}$$

Throughout this section, the following conditions are assumed.

H(3.1). The function $a_{ij}(x,t,u,s) \in C^{2,2,2,3}(\bar{Q}_T \times R^2)$ and $a_{ij}(x,t,u,s)\xi_i\xi_j \geq a_0|\xi|^2$ $(a_0 > 0)$ for any $\xi \in R^n$. The functions $b_i(x,t,s,u)$ and $c(x,t,s,u)$ are in $C^{2,1,1,2}(\bar{Q}_T \times [0,T] \times R)$. Moreover, there exists a constant A_0 such that

$$\sum_{i=1}^{n} |b_i(x,t,s,u)| + |c(x,t,s,u)| \leq A_0[1+|u|],$$

where a_0 and A_0 may depend on Q_T.

The initial data $u_0(x) \in C^{2+\alpha}(\bar{\Omega})$ and the consistency conditions

$$u_0(x) = 0, \qquad \frac{\partial}{\partial x_i}[a_{ij}(x,0,u_0(x),0)u_{x_j}] = 0 \quad \text{on } \partial\Omega$$

hold. To obtain the global existence, we shall derive an a priori estimate in the Banach space $C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)$. Since the classical maximum principle is no longer valid, we will employ energy estimates to deduce such a bound step-by-step.

LEMMA 3.1. *The solution $u(x,t)$ of (3.1)–(3.3) satisfies*

$$\int_\Omega u^2 dx + \int_0^t \int_\Omega u_x^2 dx \leq C_1, \tag{3.4}$$

where C_1 depends only on the known data and the upper bound of T.

Proof. If the equation (3.1) is multiplied by $u(x,t)$ and integrated over Q_t, we have

$$\frac{1}{2}\int_\Omega u(x,t)^2 dx + a_0 \int_0^t \int_\Omega u_x^2 dx dt \leq \frac{1}{2}\int_\Omega u_0^2 dx + \left| \int_0^t \int_\Omega u \left\{ \int_0^\tau \left[\frac{\partial}{\partial x_i} b_i(\cdots) + c(\cdots)\right] d\xi \right\} dx d\tau \right|.$$

We perform the integration by parts to the second term of the right side of the above inequality and apply condition H(3.1). It follows by Cauchy's inequality that

$$\begin{aligned} &\left| \int_0^t \int_\Omega u \left\{ \int_0^\tau \left[\frac{\partial}{\partial x_i} b_i(\cdots) + c(\cdots)\right] d\xi \right\} dx d\tau \right| \\ &= \left| \int_0^t \int_\Omega \left\{ -u_{x_i} \left[\int_0^\tau b_i(\cdots) d\xi\right] + \left[\int_0^\tau c(\cdots)\right] d\xi \right\} dx d\tau \right| \\ &\leq C(T) + \varepsilon \int\int_{Q_t} u_x^2 dx + C(\varepsilon) \int\int_{Q_t} u^2 dx dt. \end{aligned}$$

By taking $\varepsilon = a_0/2$ and applying Gronwall's inequality, we have the desired result.

LEMMA 3.2. *There exists a constant C_2 which depends only on the known data and the upper bound of T such that*

$$\sup_{0\le t\le T} ||u(\cdot,t)||_{L^\infty(\Omega)} \le C_2. \tag{3.5}$$

Proof. The proof is based on the integral iteration technique similar to Alikakos [1]. We do the following calculation for any $p = 2^k$ with $k \ge 1$:

$$\begin{aligned}
&\int_0^t \frac{d}{d\tau}\left[\int_\Omega u^p dx\right] d\tau \\
&\quad = \int_0^t \int_\Omega p u^{p-1} u_\tau dx d\tau \\
&\quad = \int_0^t \int_\Omega p u^{p-1}\left[\frac{\partial}{\partial x_i} a_{ij}(\cdots)\frac{\partial u}{\partial x_j}\right] dx d\tau \\
&\qquad + \int_0^t \int_\Omega p u^{p-1}\left\{\int_0^\tau \left[\frac{\partial}{\partial x_i} b_i(\cdots) + c(\cdots)\right] d\xi\right\} dx d\tau \\
&\quad = -\int_0^t \int_\Omega p(p-1)u^{p-2} u_{x_i} u_{x_j} a_{ij}(\cdots) dx d\tau \\
&\qquad - \int_0^t \int_\Omega p(p-1)u^{p-2} u_{x_i}\left[\int_0^\tau b_i(\cdots) d\xi\right] dx d\tau + \int_0^t \int_\Omega p u^{p-1}\left[\int_0^\tau c(\cdots) d\xi\right] dx d\tau \\
&\quad = I_1 + I_2 + I_3.
\end{aligned}$$

It is clear that the elliptic condition implies

$$I_1 \le -a_0 \int_0^t \int_\Omega p(p-1)u^{p-2} u_x^2 dx dt.$$

By Cauchy's inequality and condition H(3.1), we have

$$\begin{aligned}
|I_2| &\le \frac{a_0}{2} p(p-1)\int\int_{Q_t} u^{p-2} u_x^2 dx dt + Cp(p-1)\int\int_{Q_t} u^{p-2}\left[\int_0^\tau (1+u^2) d\xi\right] dx d\tau \\
&\le \frac{a_0}{2} p(p-1)\int\int_{Q_t} u^{p-2} u_x^2 dx dt + C(T)p(p-1)\int\int_{Q_t} [u^{p-2} + u^p] dx dt.
\end{aligned}$$

Here at the final step we have used Young's inequality:

$$ab \le \eta \frac{a^r}{r} + \eta^{-\frac{s}{r}} \frac{b^s}{s}$$

for $\frac{1}{r} + \frac{1}{s} = 1$ $(a, b \ge 0,\ r, s > 1)$.

Similarly,

$$|I_3| \le C(T)p \int\int_{Q_t} [|u|^{p-1} + u^p] dx d\tau.$$

We may assume that

$$\sup_{0\le t\le T} ||u(\cdot,t)||_{L^\infty(\Omega)} \ge 1.$$

Otherwise, we have the result. We also assume that there exists an integer p_0 such that

$$\int_\Omega u_0^p dx \le \frac{1}{2} \sup_{0\le t\le T} \int_\Omega u^p dx \tag{3.6}$$

holds for $p > p_0$. Otherwise, there exists a sequence p_i which approaches ∞ as $i \to \infty$ such that

$$\int_\Omega u_0^{p_i} dx \ge \frac{1}{2} \sup_{0\le t\le T} \int_\Omega u^{p_i} dx.$$

If p_ith root is taken on both sides of the above inequality, the desired result is obtained. Let p_0 be the smallest integer such that the inequality (3.6) holds for $p > p_0$. Hence we combine the above estimates to obtain for $p > p_0$

$$\begin{aligned} &\int_\Omega u(x,t)^p dx + p(p-1)\int_0^t \int_\Omega u^{p-2} u_x^2 dx dt \\ &\qquad \le Cp(p-1) \sup_{0\le t\le T} \int_\Omega u^p dx, \end{aligned} \tag{3.7}$$

where C depends only on the known data and the upper bound of T. Note that Gagliardo–Nirenberg's inequality (cf. [17, p. 62]) for $v(x) \in W_2^1(\Omega)$ with $v(x) = 0$ on $\partial\Omega$ implies:

$$||v(x)||_{L_2(\Omega)} \le \varepsilon ||v_x(x)||_{L_2(\Omega)} + \varepsilon^{-\frac{n}{2}} C ||v(x)||_{L_1(\Omega)}.$$

Observe that

$$\int_\Omega p(p-1) u^{p-2} u_x^2 dx \ge 2 \int_\Omega [(u^{\frac{p}{2}})_x]^2 dx.$$

We apply the above interpolation inequality for $v(x) = u(x,t)^{\frac{p}{2}}$ and take $\varepsilon^2 = [(p(p-1))/2]^{-1}$ in (3.7) to obtain

$$\int_\Omega u^p dx \le Cp^{n+2} \sup_{0\le t\le T} \int_\Omega u^{\frac{p}{2}} dx.$$

If we define

$$A_k = \sup_{0\le t\le T} [\int_\Omega u^p dx]^{\frac{1}{p}}$$

for $p = 2^k, k \ge 1$, we have

$$A_k \le [C2]^{\frac{(n+2)k}{2^k}} A_{k-1}.$$

Letting $k \to \infty$, we have

$$\sup_{0\le t\le T} ||u(\cdot,t)||_{L^\infty(\Omega)} \le C A_{p_0}$$

since

$$\sum_{k=1}^{\infty} \frac{(n+2)k}{2^k}$$

is convergent.

If $p_0 = 2$, then the estimate (3.4) indicates that A_0 is bounded, and so is A_∞. If $p_0 > 2$, since p_0 is the smallest integer such that the inequality (3.6) holds, then

$$A_{p_0} \leq \left[2 \int_\Omega u_0^{p_0} dx\right]^{\frac{1}{p_0}}$$

which is bounded. Therefore, we conclude the result.

By Lemma 3.2, we can obtain the Hölder estimate for the solution of the problem (3.1)–(3.3).

COROLLARY 3.1. *For the solution $u(x,t)$ of* (3.1)–(3.3), *there exist two constants $\alpha (0 < \alpha < 1)$ and C_3 which depends only on the known data and T such that*

$$||u||_{C^{\alpha, \frac{\alpha}{2}}(\bar{Q}_T)} \leq C_3. \tag{3.8}$$

Proof. To show the estimate, we rewrite (3.1) in the following form:

$$u_t = \frac{\partial}{\partial x_i}\left[a_{ij}\left(x,t,u,\int_0^t u d\tau\right)\frac{\partial u}{\partial x_j}\right] + \frac{\partial}{\partial x_i}\left[\int_0^t b_i(x,t,\tau,u)\right] + \int_0^t c(x,t,\tau,u)d\tau \quad \text{in } Q_T.$$

Since $u(x,t)$ is uniformly bounded by the estimate (3.5), it follows that the functions $\int_0^t b_i(x,t,\tau,u)d\tau$ and $\int_0^t c(x,t,\tau,u)d\tau$ are uniformly bounded. We apply the results of [17, Thm. 10.1, Chap. 3, p. 204] to conclude the result.

Next we shall modify the idea in [2] to derive a $W_p^{2,1}(Q_T)$-estimate for $p > 1$. For $\sigma \in [0,1]$, consider the following problem (P_σ) with the equation

$$u_t - \frac{\partial}{\partial x_i}\left[a_{ij}\left(x,t,\sigma u,\int_0^t u d\tau\right)\frac{\partial u}{\partial x_j}\right] = \int_0^t \left\{\frac{\partial}{\partial x_i}[b_i(x,t,\tau,u)] + c(x,t,\tau,u)\right\} d\tau \tag{3.9}$$

subject to the initial-boundary conditions (3.2), (3.3).

It is easy to see by Lemmas 3.1 and 3.2 that any solution $u_\sigma(x,t)$ of the problem (P_σ) satisfies the estimates (3.4), (3.5), and (3.8), which we will use frequently in the sequel without explanation.

We first need a $W_p^{2,1}(Q_T)$-estimate for the case of $\sigma = 0$.

LEMMA 3.3. *Assume that $u(x,t)$ is a classical solution of (P_0). Then*

$$||u||_{W_p^{2,1}(Q_T)} \leq C, \tag{3.10}$$

where C depends only on the known data and the upper bound of T.

Proof. To apply the $W_p^{2,1}(Q_T)$-estimate, we need to rewrite (3.9) in the form:

$$u_t - a_{ij}\left(x,t,0,\int_0^t u d\tau\right)u_{x_i x_j} = a_{ijx_i}u_{x_j} + a_{ijs}\int_0^t u_{x_i}d\tau u_{x_j} + \int_0^t \left\{\frac{\partial}{\partial x_i}[b_i(x,t,\tau,u)] + c(x,t,\tau,u)\right\} d\tau.$$

We denote the right-hand side of the above equation by $G(x,t,u) = J_1 + J_2 + J_3$. Since we have obtained an a priori bound of the Hölder norm for $u(x,t)$ by Corollary 3.1, the continuity moduli of the coefficients $a_{ij}(x,t,0,\int_0^t u d\tau)$ are known. Therefore, for any solution $u(x,t)$ of (P_0), the $W_p^{2,1}(Q_T)$ estimate for parabolic equations (cf. [17 p. 341]) can be applied to obtain

$$||u||_{W_p^{2,1}(Q_T)} \leq C||G||_{L^p(Q_T)}, \qquad p > 1,$$

where C depends only on C_2 and C_3.

It is easy to see that

$$||J_1||_{L^p(Q_T)} + ||J_3||_{L^p(Q_T)} \leq C[1 + ||u_x||_{L^p(Q_T)}]$$

which can be dominated by means of the interpolation inequality (cf. [11, Thm. 1.10.1]) by

$$\varepsilon||u||_{W_p^{2,1}(Q_T)} + C(\varepsilon).$$

To estimate J_2, we use Cauchy's inequality to get

$$\begin{aligned} ||J_2||^p_{L^p(Q_T)} &\leq \int_0^T \int_\Omega \left[\int_0^t |u_x| d\tau \cdot |u_x|\right]^p dx \\ &\leq \varepsilon \int_0^T \int_\Omega u_x^{2p} dx + C(\varepsilon) T \int_0^T \int_\Omega \int_0^t u_x^{2p} dx \\ &\equiv J_{21} + J_{22}. \end{aligned}$$

Now for any fixed $t \in [0,T]$ we use Gagliardo–Nirenberg's inequality (cf. [11, Thm. 1.10.1]):

$$||u(\cdot,t)_x^2||_{L^p(\Omega)} \leq C||u(\cdot,t)||_{L^\infty(\Omega)}||u(\cdot,t)||_{W_p^2(\Omega)}.$$

It follows that

$$\int_\Omega u_x^{2p} dx \leq C + C\int_\Omega u_{xx}^p dx.$$

Consequently,

$$\int_0^T \int_\Omega u_x^{2p} dx dt \leq C + C\int_0^T \int_\Omega u_{xx}^p dx dt$$

and

$$\int_0^T \int_0^t \int_0^\Omega u_x^{2p} dx d\tau dt \leq C + C\int_0^T \int_0^t \int_\Omega u_{xx}^p dx d\tau dt.$$

Finally, we first take ε to be small enough such that $C\varepsilon < \frac{1}{2}$ and then employ Gronwall's inequality to conclude the desired estimate.

We now intend to estimate $u_\sigma(x,t)$ in the norm of $W_p^{2,1}(Q_T)$. Let u_1 and u_2 be two classical solutions of (P_σ) corresponding to the parameters σ_1 and σ_2.

Define $w(x,t) = u_1(x,t) - u_2(x,t)$, $(x,t) \in \bar{Q}_T$. It is easy to see that $w(x,t)$ satisfies the following equations:

$$
\begin{aligned}
&w_t - \frac{\partial}{\partial x_i}\left[a_{ij}\left(x,t,\sigma_1 u_1, \int_0^t u_1 d\tau\right)\frac{\partial w}{\partial x_j}\right] \\
&\quad = \frac{\partial}{\partial x_i}\left\{\left[A_{1ij}(x,t)\sigma_1 w + A_{2ij}(x,t)\int_0^t w d\tau + A_{3ij}(x,t)(\sigma_1 - \sigma_2)\right]\frac{\partial u_2}{\partial x_j}\right\} \\
&\quad + \int_0^t \left\{\frac{\partial}{\partial x_i}[B_i(x,t,\tau)w(x,\tau)] + C(x,t,\tau)w(x,\tau)\right\} d\tau, \qquad (3.11)
\end{aligned}
$$

$$
w(x,t) = 0, \qquad (x,t) \in S_T, \qquad (3.12)
$$

$$
w(x,0) = 0, \qquad x \in \bar{\Omega}, \qquad (3.13)
$$

where

$$
\begin{aligned}
A_{1ij}(x,t) &= \int_0^1 a_{iju}\left(x,t,z(\sigma_1 u_1) + (1-z)(\sigma_1 u_2), \int_0^t u_1 d\tau\right) dz, \\
A_{2ij}(x,t) &= \int_0^1 a_{ijs}\left(x,t,\sigma_2 u_2, z\int_0^t u_1 d\tau + (1-z)\int_0^t u_2 d\tau\right) dz, \\
A_{3ij}(x,t) &= \int_0^1 a_{iju}\left(x,t,z(\sigma_1 u_2) + (1-z)(\sigma_2 u_2), \int_0^t u_1 d\tau\right) dz, \\
B_i(x,t,\tau) &= \int_0^1 b_{iu}(x,t,\tau, zu_1 + (1-z)u_2) dz, \\
C(x,t,\tau) &= \int_0^1 c_u(x,t,\tau, zu_1 + (1-z)u_2) dz,
\end{aligned}
$$

which are uniformly bounded $(i,j = 1,\cdots,n)$.

We claim that

$$
\sup_{0\le t\le T} ||w(\cdot,t)||_{L^\infty(\Omega)} \le C(\sigma_1 - \sigma_2), \qquad (3.14)
$$

where the constant C depends only on the known data and the upper bound of T.

Indeed, if we multiply the equation (3.11) by $w(x,t)$ and integrate it over Q_T, by a similar calculation to that of Lemma 3.1 we find

$$
\sup_{0\le t\le T}\int_\Omega w^2 dx + \int_0^T\int_\Omega w_x^2 dx dt \le C[\sigma_1 - \sigma_2]^2.
$$

By an analogous computation to that in the proof of Lemma 3.2 and noting the uniform boundedness of $w(x,t)$, we can deduce our assertion (3.14).

With the above auxiliary estimates in hand we are ready to show the following result.

LEMMA 3.4. *Let $u(x,t)$ be a solution of the problem* (3.1)–(3.3). *There exists a constant C_4 which depends only on the known data and the upper bound of T such that*

$$
||u||_{W_p^{2,1}(Q_T)} \le C_4. \qquad (3.15)
$$

Proof. We assert that the following estimate holds:

$$||w||^p_{w_p^{2,1}(Q_T)} \leq C + C\int_0^T \int_\Omega w_x^{2p} dxdt.$$

In fact, we write (3.11) into a nondivergence form and note that

$$\left\| \frac{\partial A_{kij}}{\partial x_i} \right\|_{L^p(Q_T)} \leq C[||u_{1x}||_{L^p(Q_T)} + ||u_{2x}||_{L^p(Q_T)}].$$

Then we perform the same calculation for the solution $w(x,t)$ of the problem (3.11)–(3.13) as for $u(x,t)$ of (P_0) in Lemma 3.2 and then employ $W_p^{2,1}$-estimate as well as Cauchy's inequality to get

$$||w||^p_{W_p^{2,1}(Q_T)} \leq C[1 + ||u_{1x}^2||^p_{L^p(Q_T)} + ||u_{2x}^2||^p_{L^p(Q_T)} + ||w_x^2||^p_{L^p(Q_T)}].$$

Hence, the relationship $u_2 = w - u_1$ yields

$$||w||^p_{w_p^{2,1}(Q_T)} \leq C\left[1 + \int_0^T \int_\Omega u_{1x}^{2p} dxdt + \int_0^T \int_\Omega w_x^{2p}\right] dxdt.$$

Again, we apply Gagliardo–Nirenberg's inequality

$$||w_x^2||_{L^p(Q_T)} \leq C||w||_{L^\infty(\Omega)}||w||_{W_p^2(\Omega)}$$

and the estimate (3.14) to arrive at

$$||w||^p_{W_p^{2,1}(Q_T)} \leq C\left[1 + \int_0^T \int_\Omega u_{1x}^{2p} dxdt + |\sigma_1 - \sigma_2|||w||^p_{W_p^{2,1}(Q_T)}\right].$$

Now we take $\sigma_1 = 0$ and $u_1(x,t)$ is the corresponding solution of (P_0). Moreover, we restrict that $\sigma_2 \in [0, \frac{1}{N}]$, where N is large enough such that $\frac{C}{N} \leq \frac{1}{2}$. Thus, we obtain the estimate (3.15) for u_{σ_2} with $\sigma_2 \in [0, \frac{1}{N}]$ since u_1 satisfies (3.10). We can repeat the above procedure and derive the estimate (3.15) for the solution u_σ with $\sigma \in [\frac{1}{N}, \frac{2}{N}]$. After N steps, we finally reach the estimate (3.15) for $\sigma = 1$.

By Lemma 3.3, we immediately have the following corollary.

COROLLARY 3.2. *For any solution $u(x,t)$ of* (3.1)–(3.3), *we have*

$$||u||_{C^{1+\alpha,\frac{1+\alpha}{2}}(\bar{Q}_T)} \leq C_6. \tag{3.16}$$

Proof. This follows from Lemma 3.4 and the interpolation inequality (cf. [17, Cor., p. 342]) if $p > \frac{n+2}{2}$.

LEMMA 3.5. *There exists a constant C_7 which depends only on the known data and T such that*

$$||u||_{C^{2+\alpha,1+\frac{\alpha}{2}}(\bar{Q}_T)} \leq C_7. \tag{3.17}$$

Proof. We rewrite (3.1) in the following form:

$$u_t - a_{ij}\left(x,t,u,\int_0^t u d\tau\right)u_{x_ix_j} = a_{iju}u_{x_i}u_{x_j} + \left[a_{ijx_i} + a_{ijs}\cdot\int_0^t u_{x_i}d\tau\right]u_{x_j} + \int_0^t\left[\frac{\partial}{\partial x_i}b_i(x,t,\tau,u) + c(x,t,\tau,u)\right]d\tau.$$

From Corollary 3.2, the right side of the above equality is uniformly bounded in the norm of $C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)$. Moreover, the coefficient $a_{ij}(\cdots)$ is Hölder continuous on $\bar{Q}_T$. By the Schauder estimate (2.6), we conclude the result.

THEOREM 3.1. *The problem* (3.1)–(3.3) *has a unique global solution.*

Proof. The result can be demonstrated by the continuation argument which is similar to that of parabolic equations (cf. [10, Chap. 3]). For $\lambda \in [0,1]$, we consider the system (P_λ)

$$u_t - \frac{\partial}{\partial x_i}\left[a_{ij}(x,t,u,\lambda\int_0^t u d\tau)\frac{\partial u}{\partial x_j}\right] = \lambda\int_0^t\left[\frac{\partial}{\partial x_i}b_i(x,t,\tau,u) + c(x,t,\tau,u)\right]d\tau \quad \text{in } Q_T$$

subject to the initial-boundary conditions (3.2), (3.3), where the initial data $u_0(x)$ is replaced by $\lambda u_0(x)$. Clearly, the estimate in Lemma 3.5 is also true for any solution u_λ of (P_λ). Let $\Sigma = \{\lambda :$ the problem (P_λ) is solvable on $[0,T]\}$. It is easy to see (cf., e.g., [17, Thm. 6, p. 452]) that $0 \in \Sigma$. By applying the a priori estimate (3.17) we can show that the set Σ is open as well as closed. It follows that $\Sigma = [0,1]$. When $\lambda = 1$, it follows that the problem (3.1)–(3.3) is solvable on $[0,T]$ for any $T > 0$.

Remark. With the a priori estimate (3.17), Theorem 3.1 can alternatively be shown by the bootstrap argument (cf. [21]).

Acknowledgments. The author thanks the referees for their many valuable comments.

REFERENCES

[1] N. D. ALIKAKOS, *An application of the invariant principle in reaction-diffusion equations*, J. Differential Equations, 33 (1979), pp. 201-225.

[2] H. AMANN AND M. G. CRANDALL, *On some existence theorems for semilinear elliptic equations*, Indiana Univ. Math. J., 27 (1978), pp. 779–790.

[3] G. ANDREWS, *On existence of solutions to the equation* $u_{tt} = u_{xxt} + \sigma(u_x)_x$, J. Differential Equations, 35 (1980), pp. 200–231.

[4] M. G. CRANDALL, S. O. LONDEN AND J. A. NOHEL, *An abstract nonlinear Volterra integrodifferential equation*, J. Math. Anal. Appl., 64 (1978), pp. 701–735.

[5] T. K. CAUGHEY AND J. ELLISON, *Existence, uniqueness and stability of solutions of a class of nonlinear partial differential equations*, J. Math. Anal. Appl., 51 (1975), pp. 1–32.

[6a] Y. EBIHARA, *On some nonlinear evolution equations with the strong dissipation* I, J. Differential Equations, 30 (1978), pp. 149–164.

[6b] ———, *On some nonlinear evolution equations with the strong dissipation* II, J. Differential Equations, 34 (1979), pp. 339–352.

[7] G. DA PRATO AND M. IANNELLI, *Existence and regularity for a class of integrodifferential equations of parabolic type*, J. Math. Anal. Appl., 112 (1985), pp. 36–55.

[8] C. M. DAFERMOS, *The mixed initial boundary value problem for the equations of nonlinear one dimensional visco-elasticity*, J. Differential Equations, 6 (1969), pp. 71–86.

[9] W. E. FITZGIBBON, *Semilinear integrodifferential equations in Banach space*, Nonlinear Anal., Theory Meth. Appl., 4 (1980), pp. 745–760.

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

[11] ———, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[12] M. L. HEARD, *An abstract parabolic Volterra integrodifferential equation*, SIAM J. Math. Anal., 31 (1982), pp. 81–105.

[13] M. L. HEARD AND S. M. RANKIN III, *A semilinear parabolic Volterra integrodifferential equation*, J. Differential Equations, 71 (1988), pp. 201–233.

[14] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture. Notes in Math., 840, Springer-Verlag, Berlin, New York, 1984.

[15] M. ISMATOV, *A mixed problem for the equation describing sound propagation in a viscous gas*, Differential Equations, 20 (1984), pp. 1023–1035.

[16] P. M. D. KHAN AND B. G. PACHPATTE, *On quasilinear Volterra integrodifferential equations in a Banach space*, Indian J. Pure Appl. Math., 18 (1987), pp. 32–49.

[17] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasi-linear Equations of Parabolic Type*, Amer. Math. Soc. Trans., Vol. 23, Providence, RI, 1968.

[18] A. LUNARDI AND E. SINESTRARI, *Fully nonlinear integrodifferential equations in general Banach spaces*, Math. Z., 190 (1985), pp. 225–248.

[19] J. A. NOHEL, *Nonlinear Volterra equations for heat flow in materials with memory, Integral and functional differential equations*, in Lecture Notes in Pure and Applied Mathematics, T. L. Herdman, S. M. Rankin III, and H. W. Stech, eds., Marcel Dekker, New York, 1981.

[20] H. PECHER, *On global solutions of third order partial differential equations*, J. Math. Anal. Appl., 73 (1980), pp. 278–299.

[21] F. ROTHE, *Global Solutions of Reaction-Diffusion Systems*, Springer-Verlag, New York, 1984.

[22] I. V. SUVEIKA, *Mixed problems for an equation describing the propagation of disturbances in viscous media*, Differential Equations, 19 (1984), pp. 337–347.

[23] I. I. VRABIE, *Compactness methods for an abstract nonlinear Volterra integrodifferential equation*, Nonlinear Anal. Theory Meth. Appl., 5 (1981), pp. 355–371.

[24] H. M. YIN, *The classical solutions for nonlinear integrodifferential equations*, J. Integral Equations Appl., 1 (1988), pp. 249–263.

[25] ———, *The solvability for a class of nonlinear integrodifferential equations of parabolic type*, J. Partial Differential Equations, to appear.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1738-1743, November 1991

012

ANALYTICITY OF SOLUTIONS OF THE KORTEWEG–DE VRIES EQUATION*

NAKAO HAYASHI†

Abstract. It is proven that if the initial function of the Korteweg-de Vries equation is analytic and has an analytic continuation to a strip containing the real axis, then the local in time solution has the same property, although the width of the strip might decrease with time. The result contains the case of the complex-valued initial function.

Key words. analyticity of solutions, Korteweg-de Vries equation

AMS(MOS) subject classification. 35Q20

1. Introduction.

We consider the following equation:

$$\text{(KdV)}\qquad \begin{aligned} &\partial_t u+\partial_x^3 u+a(u)\partial_x u=0 \qquad (t,x)\in \mathbb{R}^+\times\mathbb{R},\\ &u(0,x)=\phi(x), \qquad x\in\mathbb{R}, \end{aligned}$$

where $a(\lambda)$ is a polynomial. We show that if ϕ is analytic and has an analytic continuation to a strip containing the real axis, then the local in time solution of (KdV) has the same property for space variable, although the width of the strip might decrease with time.

We shall establish existence and analyticity simultaneously. This is in contrast to the work of Kato and Masuda [5]. They proved analyticity of the solution whose existence is guaranteed by $W^{s,2}$-theory (see [2]-[6]). Existence theorems in $W^{s,2}$-theory are based essentially on the following estimate: If u is real valued, then

$$|(a(u)\partial_x u, u)_{W^{s',2}}|\leqq \tilde{a}(\|u\|_{W^{s,2}})\|u\|^2_{W^{s',2}},$$

where $\frac{3}{2}<s\leqq s'\leqq s+1$ and $\tilde{a}(\lambda)$ is a certain polynomial depending only on a and s (see [4]). Therefore the arguments in [5] seem to work only for the case where ϕ is real valued. Our method in this paper is different from [5] and is a modification of the methods in [1]. Of course, our arguments work for complex valued ϕ. It should be noted that in [5] the function $a(\lambda)$ is only assumed to be real analytic, while we assume that $a(\lambda)$ is a polynomial.

We state notation and introduce function spaces. We let $L^p=\{f(x)$ is measurable on $\mathbb{R}$, $\|f\|_p<\infty\}$, where $\|f\|_p=(\int|f(x)|^p\,dx)^{1/p}$ if $1\leqq p<\infty$ and $\|f\|_\infty=\sup\{|f(x)|;\ x\in\mathbb{R}\}$. We denote by $(\cdot,\cdot)$ the inner product in L^2. We let $W^{m,p}(\mathbb{R})=\{f\in L^p;\ \|f\|_{m,p}<\infty\}$, where $\|f\|^p_{m,p}=\sum_{j=0}^m\|\partial_x^j f\|_p^p$ if $1\leqq p<\infty$ and $\|f\|_{m,\infty}=\sum_{j=0}^m\|\partial_x^j f\|_\infty$. We denote by $[\mathcal{F}f](\xi)$ (or $\hat{f}(\xi)$) the Fourier transform $\int e^{-ix\xi}f(x)\,dx$ and by $[\mathcal{F}^{-1}g](x)$ the inverse Fourier transform of $g(\xi)$. For $r>0$ we define

$$L_r=\{f\in L^2:\ \|f\|^2_{L_r}=(f,f)_{L_r}=(\cosh(2r\xi)\hat{f},\hat{f})<\infty\},$$

$$X_r=\left\{f\in L^2;\ \|f\|^2_{X_r}=(f,f)_{X_r}=\sum_{j=0}^1(\xi^{2j}(\cosh(2r\xi)+\xi\sinh(2r\xi))\hat{f},\hat{f})<\infty\right\},$$

$$Y_r=\left\{f,\partial_x f\in L^2;\ \|f\|^2_{Y_r}=(f,f)_{Y_r}=\sum_{j=0}^1(\xi^{2j+2}\cosh(2r\xi)\hat{f},\hat{f})<\infty\right\}.$$

* Received by the editors January 17, 1990; accepted for publication (in revised form) December 14, 1990.
† Department of Mathematics, Faculty of Engineering, Gunma University, Kiryu 376, Japan.

Obviously, $X_r \subset L_r$; if $\partial_x f \in X_r$, then $f \in Y_r$. We let $z = x + iy$ $(x, y \in \mathbb{R})$ and let $S(r)$ be the strip $\{z = x + iy; |y| < r\}$. Note that the above function spaces are closely related to the analytic Hardy space $H^p(r) = \{F: F$ is analytic on $S(r)$, $\|F\|_{H^p(r)} = \sup_{|y|<r} \|F(\cdot + iy)\|_p < \infty\}$ on the strip $S(r)$ in the sense of Stein-Weiss [7]. For a nonnegative integer m we let

$$H^{m,p}(r) = \left\{ F \in H^p(r): \|F\|^p_{H^{m,p}(r)} = \sum_{j=0}^{m} \|\partial_z^j F\|^p_{H^p(r)} < \infty \right\}.$$

Obviously, $H^{0,p}(r) = H^p(r)$; $\|F\|_{H^{m,p}(r)} = \sup_{|y|<r} \|F(\cdot + iy)\|_{m,p}$. We have observed the following [1].

THEOREM 1 [1]. *We assume that $F \in H^2(r)$ and let f be the trace of F on the real axis. Then $f \in L_r$ and*

$$\|f\|_{L_r} \leqq \sqrt{2} \|F\|_{H^2(r)}.$$

Conversely, we assume that $f \in L_r$. Then f has an analytic extension $F \in H^2(r)$ and

$$\|F\|_{H^2(r)} \leqq \sqrt{2} \|f\|_{L_r}.$$

For the convenience of the reader we shall give a proof of Theorem 1 in § 2. For an interval I or $\mathbb{R}$ and a Banach space B with norm $\|\cdot\|_B$, we let $C(I; B) = \{f(t): f(t)$ is continuous from I to B, $\sup\{\|f(t)\|_B; t \in I\} < \infty\}$.

Our main result is the following theorem.

THEOREM 2. *If $\phi \in X_{\sigma_0}$ for $\sigma_0 > 0$, then there exist a positive constant $T = T(\|\phi\|_{X_{\sigma_0}}, a)$ and a positive monotone decreasing function $\sigma(t)$ satisfying $\sigma(0) = \sigma_0$ such that* (KdV) *has a unique solution $u(t, x) \in C([0, T]; X_{\sigma(T)}) \cap X_{\sigma(t)}$ for $0 \leqq t \leqq T$.*

From Theorems 1 and 2, we have Corollary 1.

COROLLARY 1. *Let $u(t, x)$ be the solution constructed in Theorem 2. Then $u(t, x)$ has an analytic continuation to $S(\sigma(t))$ for $0 \leqq t \leqq T$.*

2. Preliminary estimates. In this section we collect some preliminary estimates. We first prove Theorem 1.

Proof of Theorem 1. For $F \in H^2(r)$ and $|y| < r$ we let $f_y(x) = F(x + iy)$. We note that the trace f of F on the real axis is equal to f_0. It is well known that

$$\hat{f}_y(\xi) = e^{-y\xi} \hat{f}(\xi) \qquad \text{(see [7, p. 99])}. \tag{2.1}$$

The Plancherel theorem says

$$\|f_y\|_2^2 = \int e^{-2y\xi} |\hat{f}(\xi)|^2 \, d\xi \leqq 2 \int \cosh(2r\xi) |\hat{f}(\xi)|^2 \, d\xi$$

for $|y| < r$. This yields the second inequality of the theorem. Conversely, observe from the monotone convergence theorem that

$$\|f_y\|_2^2 \geqq \int_0^\infty e^{-2y\xi} |\hat{f}(\xi)|^2 \, d\xi \to \int_0^\infty e^{2r\xi} |\hat{f}(\xi)|^2 \, d\xi$$

as $y \to -r$. Hence

$$\|F\|^2_{H^2(r)} \geqq \int_0^\infty \cosh(2r\xi) |\hat{f}(\xi)|^2 \, d\xi,$$

and similarly

$$\|F\|^2_{H^2(r)} \geqq \int_{-\infty}^0 \cosh(2r\xi) |\hat{f}(\xi)|^2 \, d\xi.$$

Thus we have the first inequality of the theorem.

As a result we have the following lemma.

LEMMA 2.1. *Let $r>0$.*

(a) *If $f\in X_r$, then f has an analytic continuation F on $S(r)$ and*

$$\|F\|_{H^{1,2}(r)}\leqq\sqrt{2}\|f\|_{X_r}.$$

(b) *If $f\in Y_r$, then f has an analytic continuation F on $S(r)$ and*

$$\|\partial_z F\|_{H^{1,2}(r)}\leqq\sqrt{2}\|f\|_{Y_r}.$$

Proof. Let $f\in X_r$. By Theorem 1 f has an analytic continuation $F\in H^2(r)$. Let $f_y=F(\cdot+iy)$ for $|y|<r$. Then by (2.1) $f_y=\mathscr{F}^{-1}(\cosh(y\xi)\hat{f})+\mathscr{F}^{-1}(\sinh(-y\xi)\hat{f})$. Hence the Schwarz inequality yields

$$\|f_y\|_2^2+\|\partial_x f_y\|_2^2\leqq 2\sum_{j=0}^{1}(\|\partial_x^j\mathscr{F}^{-1}\cosh(y\xi)\hat{f}\|_2^2+\|\partial_x^j\mathscr{F}^{-1}\sinh(-y\xi)\hat{f}\|_2^2).$$

By a direct calculation we see that the right-hand side of the above inequality is equal to

$$2\sum_{j=0}^{1}((\xi^{2j}\cosh(y\xi)\hat{f},\cosh(y\xi)\hat{f})+(\xi^{2j}\sinh(y\xi)\hat{f},\sinh(y\xi)\hat{f}))$$

$$=2\sum_{j=0}^{1}(\xi^{2j}\cosh(2y\xi)\hat{f},\hat{f})\leqq 2\sum_{j=0}^{1}(\xi^{2j}\cosh(2r\xi)\hat{f},\hat{f}).$$

This is less than or equal to $2\|f\|_{X_r}^2$. Thus we obtain (a). The assertion (b) is proved in the same way as in the proof of (a) and so we omit it.

Conversely, the inner product $(\cdot,\cdot)_{X_r}$ is controlled by the Hardy norm in the following way.

LEMMA 2.2. *Let $r>0$. Suppose $F\in H^{1,2}(r)$ and $V\in H^{2,2}(r)$. If f and v are the traces of F and V on the real axis, then*

$$|(f,v)_{X_r}|\leqq C_1\|F\|_{H^{1,2}(r)}\|V\|_{H^{2,2}(r)}.$$

Proof. Let $0<\rho<r$. By a simple calculation

$$2(f,v)_{X_\rho}=2\sum_{j=0}^{1}(\xi^j(\cosh(2\rho\xi)+\xi\sinh(2\rho\xi))\hat{f},\xi^j\hat{v})$$

$$=\sum_{j=0}^{1}\{(\xi^j e^{\rho\xi}\hat{f},\xi^j(1+\xi)e^{\rho\xi}\hat{v})+(\xi^j e^{-\rho\xi}\hat{f},\xi^j(1-\xi)e^{-\rho\xi}\hat{v})\}$$

$$=\sum_{j=0}^{1}\{(\partial_x^j f_{-\rho},\partial_x^j(1-i\partial_x)v_{-\rho})+(\partial_x^j f_\rho,\partial_x^j(1+i\partial_x)v_\rho)\},$$

where $f_{\pm\rho}=F(\cdot\pm i\rho)$ and $v_{\pm\rho}=V(\cdot\pm i\rho)$. Therefore we have from the Schwarz inequality

$$|(f,v)_{X_\rho}\|\leqq C_2\cdot(\|f_\rho\|_{1,2}\|v_\rho\|_{2,2}+\|f_{-\rho}\|_{1,2}\|v_{-\rho}\|_{2,2})$$

$$\leqq 2C_2\|F\|_{H^{1,2}(r)}\|V\|_{H^{2,2}(r)}.$$

The monotone convergence theorem completes the proof.

Using Sobolev's inequality for $F(\cdot+iy)$, we obtain a multiplicative property of $\|F\|_{H^{1,2}(r)}$.

LEMMA 2.3. *Let $r>0$. Suppose $F, G\in H^{1,2}(r)$. Then $FG\in H^{1,2}(r)$ and*

$$\|FG\|_{H^{1,2}(r)}\leqq C_3\|F\|_{H^{1,2}(r)}\|G\|_{H^{1,2}(r)}.$$

Proof. Sobolev's inequality shows

$$\|F\|_{H^\infty(r)} \leqq C_4 \|F\|_{H^{1,2}(r)}.$$

Hence

$$\|FG\|_{H^2(r)} \leqq \|F\|_{H^\infty(r)} \|G\|_{H^2(r)} \leqq C_4 \|F\|_{H^{1,2}(r)} \|G\|_{H^{1,2}(r)},$$

$$\begin{aligned}\|\partial_z(FG)\|_{H^2(r)} &\leqq \|\partial_z F\|_{H^2(r)} \|G\|_{H^\infty(r)} + \|F\|_{H^\infty(r)} \|\partial_z G\|_{H^2(r)} \\ &\leqq 2C_4 \|F\|_{H^{1,2}(r)} \|G\|_{H^{1,2}(r)}.\end{aligned}$$

The lemma follows.

We are now in a position to prove the main estimate.

LEMMA 2.4. *There exists a polynomial $\tilde{a}$ of which coefficients are all nonnegative with the following property: If $r>0$ and $f, g, v \in X_r \cap Y_r$, then*

$$\begin{aligned}&|(a(f)\partial_x f - a(g)\partial_x g, v)_{X_r}| \\ &\quad \leqq \tilde{a}(\|f\|_{X_r}, \|g\|_{X_r})\{\|f\|_{Y_r}\|f-g\|_{X_r} + \|f-g\|_{Y_r}\}(\|v\|_{X_r} + \|v\|_{Y_r}).\end{aligned}$$

COROLLARY. *There exists a polynomial $\tilde{a}_1$ with nonnegative coefficient such that if $f, v \in X_r \cap Y_r$, then*

$$|(a(f)\partial_x f, v)_{X_r}| \leqq \tilde{a}_1(\|f\|_{X_r})\|f\|_{Y_r}(\|v\|_{X_r} + \|v\|_{Y_r}).$$

Proof of Lemma 2.4. We have

$$a(f)\partial_x f - a(g)\partial_x g = b(f, g)(f-g)\partial_x f + a(g)\partial_x(f-g),$$

where $b(f, g)$ is a polynomial. Let us estimate $(b(f, g)(f-g)\partial_x f, v)_{X_r}$ and $(a(g)\partial_x(f-g), v)_{X_r}$ separately. By Lemma 2.1 we see that f, g, and v gave analytic continuations F, G, and V on $S(r)$. Obviously, $b(F, G)$ is the analytic continuation of $b(f, g)$. It follows from Lemmas 2.2 and 2.3 that

$$\begin{aligned}|(b(f, g)(f-g)\partial_x f, v)_{X_r}| &\leqq C_1 \|b(F, G)(F-G)\partial_x F\|_{H^{1,2}(r)} \|V\|_{H^{2,2}(r)} \\ &\leqq C_5 \|b(F, G)\|_{H^{1,2}(r)} \|F-G\|_{H^{1,2}(r)} \|\partial_x F\|_{H^{1,2}(r)} \|V\|_{H^{2,2}(r)} \\ &\leqq a_1(\|F\|_{H^{1,2}(r)}, \|G\|_{H^{1,2}(r)}) \|F-G\|_{H^{1,2}(r)} \|\partial_x F\|_{H^{1,2}(r)} \|V\|_{H^{2,2}(r)},\end{aligned}$$

where a_1 is a polynomial with nonnegative coefficients. Hence Lemma 2.1 yields

$$|(b(f, g)(f-g)\partial_x f, v)_{X_r}| \leqq C_6 a_1(\|f\|_{X_r}, \|g\|_{X_r}) \|f-g\|_{X_r} \|f\|_{Y_r}(\|v\|_{X_r} + \|v\|_{Y_r}).$$

In the same way,

$$|(a(g)\partial_x(f-g), v)_{X_r}| \leqq a_2(\|g\|_{X_r}) \|f-g\|_{Y_r}(\|v\|_{X_r} + \|v\|_{Y_r}),$$

where a_2 is a polynomial with nonnegative coefficients. The lemma follows.

3. Proof of Theorem 2. For any fixed $\sigma_0 > 0$, we let $\sigma(t) = \sigma_0 e^{-At/\sigma_0}$ with a positive constant A to be determined later. For $T>0$ we define

$$B(T) = \left\{ v(t, x);\ \|v\|^2_{B(T)} = \sup_{0 \leqq t \leqq T} \|v(t)\|^2_{X_{\sigma(t)}} + A \int_0^T \|v(t)\|^2_{Y_{\sigma(t)}}\, dt < \infty \right\}$$

and

$$B_\rho(T) = \{v \in B(T);\ \|v\|_{B(T)} \leqq \rho\}.$$

We shall show that there exist $t>0$ and $A>0$ depending only on $\|\phi\|_{X_{\sigma_0}}$ such that (KdV) has a unique solution $u(t,x)\in B(T)$. Let $\rho=2\|\phi\|_{X_{\sigma_0}}$. For $v\in B_\rho(T)$ we define the mapping M by $u=Mv$, where u is the solution of the following linearized equation:

$$\partial_t u+\partial_x^3 u=-a(v)\partial_x v, \qquad (t,x)\in\mathbb{R}^+\times\mathbb{R},$$
$$u(0,x)=\phi(x), \qquad x\in\mathbb{R}.$$

It is sufficient to show that M is a contraction mapping from $B_\rho(T)$ to itself for suitable chosen T and A. Hereafter we let

$$e^{-AT/\sigma_0}>\tfrac{1}{2}. \tag{3.1}$$

Take $v\in B_\rho(T)$. By the Fourier transform

$$\partial_t\hat{u}-i\xi^3\hat{u}=-\mathcal{F}[a(v)\partial_x v]. \tag{3.2}$$

We multiply both sides of (3.2) by $\xi^{2j}(\cosh(2\sigma(t)\xi)+\xi\sinh(2\sigma(t)\xi)\hat{u}$, integrate in ξ and take the real part to obtain

$$\begin{aligned}&\frac{d}{dt}(\|u(t)\|^2_{X_{\sigma(t)}})-2\sigma'(t)(\xi^{2j}(\xi\sinh(2\sigma(t)\xi)+\xi^2\cosh(2\sigma(t)\xi))\hat{u}(t),\hat{u}(t))\\ &\quad=-2\,\mathrm{Re}\,(a(v)\partial_x v,u)_{X_{\sigma(t)}}.\end{aligned} \tag{3.3}$$

For the sake of brevity we suppress the subscript $\sigma(t)$ of $X_{\sigma(t)}$ and $Y_{\sigma(t)}$. From (3.3) and the corollary to Lemma 2.4 it follows that

$$\begin{aligned}&\frac{d}{dt}(\|u(t)\|_X^2)+2Ae^{-At/\sigma_0}\|u(t)\|_Y^2\\ &\quad\le 2\tilde{a}_1(\|v(t)\|_X)\|v(t)\|_Y(\|u(t)\|_X+\|u(t)\|_Y)\\ &\quad\le 2\tilde{a}_1(\rho)\|v(t)\|_Y(\|u(t)\|_X+\|u(t)\|_Y).\end{aligned}$$

Integrating in t and using the Schwarz inequality, we obtain from (3.1)

$$\begin{aligned}&\sup_{0\le t\le T}\|u(t)\|_X^2+A\int_0^T\|u(t)\|_Y^2\,dt\\ &\quad\le\|\phi\|^2_{X_{\sigma_0}}+2\tilde{a}_1(\rho)\left(\int_0^T\|v(t)\|_Y^2\,dt\right)^{1/2}\left\{\left(\int_0^T\|u(t)\|_X^2\,dt\right)^{1/2}\right.\\ &\qquad\left.+\left(\int_0^T\|u(t)\|_Y^2\,dt\right)^{1/2}\right\}.\end{aligned}$$

Hence we have

$$\begin{aligned}\|u\|^2_{B(T)}&\le\frac{\rho^2}{4}+\frac{\tilde{a}_1(\rho)}{\sqrt{A}}4\rho\left(\sqrt{T}+\frac{1}{\sqrt{A}}\right)\|u\|_{B(T)}\\ &\le\frac{\rho^2}{4}+\frac{1}{2}\left\{\frac{\tilde{a}_1(\rho)}{\sqrt{A}}4\rho\left(\sqrt{T}+\frac{1}{\sqrt{A}}\right)\right\}^2+\frac{1}{2}\|u\|^2_{B(T)}.\end{aligned}$$

This implies

$$\|u\|^2_{B(T)}\le 2\left\{\frac{1}{4}+\frac{8\tilde{a}_1(\rho)^2}{A}\left(\sqrt{T}+\frac{1}{\sqrt{A}}\right)^2\right\}\rho^2.$$

Therefore, if

$$\frac{\tilde{a}_1(\rho)}{\sqrt{A}}\left(\sqrt{T}+\frac{1}{\sqrt{A}}\right)<\frac{1}{4\sqrt{2}}, \tag{3.4}$$

then M is a mapping from $B_\rho(T)$ into itself.

Take $v_1, v_2 \in B_\rho(T)$. Let $w = Mv_1 - Mv_2$. In the same way as in the proof of (3.3) we have

$$\frac{d}{dt}\|w(t)\|_X^2 + 2Ae^{-At/\sigma_0}\|w(t)\|_Y^2 \leqq -2\,\mathrm{Re}\,(a(v_1)\partial_x v_1(t) - a(v_2)\partial_x v_2(t), w(t))_X.$$

From Lemma 2.4 the right-hand side is estimated by

$$\begin{aligned}&2\tilde{a}(\rho,\rho)\{\|v_1(t)\|_Y\|v_1(t)-v_2(t)\|_X + \|v_1(t)-v_2(t)\|_Y\}(\|w(t)\|_X + \|w(t)\|_Y)\\ &\quad\leqq 2\tilde{a}(\rho,\rho)\{\|v_1(t)\|_Y\|v_1-v_2\|_{B(T)} + \|v_1(t)-v_2(t)\|_Y\}(\|w(t)\|_X + \|w(t)\|_Y).\end{aligned}$$

Integrating in t and using the Schwarz inequality, we obtain from (3.1)

$$\begin{aligned}\|w\|_{B(T)}^2 &\leqq 2\tilde{a}(\rho,\rho)\left\{\|v_1-v_2\|_{B(T)}\left(\int_0^T \|v_1(t)\|_Y^2\,dt\right)^{1/2} + \left(\int_0^T \|v_1(t)-v_2(t)\|_Y^2\,dt\right)^{1/2}\right\}\\ &\quad\times\left\{\left(\int_0^T \|w(t)\|_X^2\,dt\right)^{1/2} + \left(\int_0^T \|w(t)\|_Y^2\,dt\right)^{1/2}\right\}\\ &\leqq 2\tilde{a}(\rho,\rho)\frac{1}{\sqrt{A}}(\|v_1\|_{B(T)}+1)\|v_1-v_2\|_{B(T)}\left(\sqrt{T}+\frac{1}{\sqrt{A}}\right)\|w\|_{B(T)}.\end{aligned}$$

Hence

$$\|w\|_{B(T)} \leqq \tilde{a}(\rho,\rho)\frac{2}{\sqrt{A}}(\rho+1)\left(\sqrt{T}+\frac{1}{\sqrt{A}}\right)\|v_1-v_2\|_{B(T)}.$$

Therefore, if

$$\tilde{a}(\rho,\rho)\frac{2}{\sqrt{A}}(\rho+1)\left(\sqrt{T}+\frac{1}{\sqrt{A}}\right) < 1, \tag{3.5}$$

then M is a contraction mapping. We can, in fact, choose $T>0$, $A>0$ satisfying (3.1), (3.4), and (3.5) altogether. For these T and A, M is a contraction mapping from $B_\rho(T)$ to itself and has a unique fixed point. This fixed point is the solution of (KdV). The property that $u(t,x) \in C([0,T]; X_{\sigma(T)})$ follows from $\sup_{0\leqq t\leqq T}\|u(t)\|_{X_{\sigma(T)}}^2 \leqq \|u\|_{B(T)}^2$ immediately.

Acknowledgments. The author is grateful to Professor H. Aikawa for his careful reading of the first manuscript and many valuable comments. The author also thanks the referee for important remarks.

REFERENCES

[1] H. Aikawa, N. Hayashi, and S. Saitoh, *Analyticity of solutions for semi-linear heat equations in one space dimension* II, preprint.

[2] J. L. Bona and L. R. Scott, *Solutions of the Korteweg-de Vries equation in fractional order Sobolev spaces*, Duke Math. J., 43 (1976), pp. 87-99.

[3] J. L. Bona and R. Smith, *The initial value problem for the Korteweg-de Vries equation*, Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 555-601.

[4] T. Kato, *On the Cauchy problem for the (generalized) Korteweg-de Vries equation*, in Studies in Applied Mathematics, V. Guillemin, ed., Adv. in Math. Suppl. Stud., 18, Academic Press, New York, 1983, pp. 93-128.

[5] T. Kato and K. Masuda, *Nonlinear evolution equations and analyticity* I, Ann. Inst. Henri Poincaré, Anal. non linéaire, 3 (1986), pp. 455-467.

[6] J. C. Saut and R. Temam, *Remark on the Korteweg-de Vries equation*, Israel J. Math., 24 (1976), pp. 78-87.

[7] E. M. Stein and G. Weiss, *Introduction to Fourier analysis on Euclidean spaces*, Princeton University Press, Princeton, NJ, 1971.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1744-1754, November 1991

013

GEOMETRY OF RATIONAL FUNCTIONS AND NONLINEAR INTEGRABLE SYSTEMS*

YOSHIMASA NAKAMURA†

Abstract. A complete parametrization of the space $\mathrm{Rat}_p(n)$ of rational functions of degree n and fixed denominator $p(z)$ in terms of a nonlinear integrable system is established. The space $\mathrm{Rat}_p(n)$ can be viewed as the moduli space of certain controllable and observable linear dynamical systems. It is proved that $\mathrm{Rat}_p(n)$ is diffeomorphic to the moduli (or parameter) space of solutions of a natural generalization of the celebrated finite Toda equation called the cyclic-Toda hierarchy. The original Toda flow is identified with one of the connected components of $\mathrm{Rat}_p(n)$, where $p(z)$ is the characteristic polynomial of a Jacobi matrix. To prove the correspondence we use an exponential of cyclic matrix polynomials and its QR factorization which induces isospectral deformations.

Key words. space of rational functions, nonlinear integrable systems, scaling theory for linear systems, cyclic-Toda equation hierarchy

AMS(MOS) subject classifications. 93B27, 58F07, 34A05

1. Introduction. There have been some systematic efforts toward a theory of dynamical systems with parameters. Brockett and Krishnaprasad [3] presented an analysis of the action of certain one-parameter groups on the space rat (n) of strictly proper rational functions of (McMillan) degree n with real coefficients and discussed its application to the identification problem for linear dynamical systems. We denote each function of rat (n) by the form

$$f(z)=\frac{q_{n-1}z^{n-1}+\cdots+q_0}{z^n+p_{n-1}z^{n-1}+\cdots+p_0}=\frac{q(z)}{p(z)}\in \mathrm{rat}\,(n), \tag{1a}$$

where $p(z)$ and $q(z)$ do not have any common factor. They considered five types of one-parameter groups due to (1) frequency scaling, (2) shift of imaginary axis, (3) amplitude scaling, (4) output feedback, and (5) shift of time axis, and showed that these group actions leave invariant connected components rat $(n-m, m)$, $0\leqq m\leqq n$, of rat (n). Here the connected components rat $(n-m, m)$ are distinguished by the Cauchy index $2m-n$ viewed as a continuous map from rat (n) into the set $\{-n, -n+2, \cdots, n-2, n\}$ [2]. In the geometry of rational functions it is natural to introduce the equivalence relation $f(z)\sim e^{\theta}f(z)$, $\theta\in\mathbb{R}$, under the third one-parameter group action. We set

$$\mathrm{Rat}\,(n)=\mathrm{rat}\,(n)/\sim. \tag{1b}$$

Let us restrict ourselves to the fifth one-parameter group taking the form

$$f(z)=C^T(zI-A)^{-1}B\to C^T(zI-A)^{-1}\,e^{\sigma A}B, \tag{2}$$

for $\sigma\in\mathbb{R}$, where I is the $n\times n$ unit matrix and C^T denotes the transpose of C. The function $f(z)\in\mathrm{rat}\,(n)$ is regarded as the transfer function of the controllable and observable linear dynamical system

$$\frac{d}{dt}x(t)=Ax(t)+Bu(t),\qquad y(t)=C^Tx(t), \tag{3}$$

* Received by the editors August 16, 1989; accepted for publication (in revised form) November 14, 1990.
† Department of Mathematics, Gifu University, Yanagido, Gifu 501-11, Japan.

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$, $C \in \mathbb{R}^{n\times 1}$, and $x(t) \in \mathbb{R}^{n\times 1}$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$, and $t \in \mathbb{R}$. The flow (2) on the space rat (n) is given as the Laplace transform of time shift $C^T e^{tA} B \to C^T e^{(t+\sigma)A} B$ for the weight function of (3). Byrnes [4] discussed the action of time shift on the space of symmetric matrix-valued rational functions and its relationship to a network synthesis problem for RC and RLC networks with parameters. Fuhrmann and Krishnaprasad [6] pointed out that only the fifth action (2) does not leave invariant the continued fraction cell decomposition of the space rat (n). It is to be noted that the differential equation $(d/ds) \log F(s) = f(s)$ defines a generalization of Pearson's family of probability distributions [18], where $F(s)$ denotes the density function and $s \in \mathbb{R}$. It seems that in these results we see increasing significance of group actions, especially of the time shift, in the geometry of rational functions and its applications.

The importance of nonlinear integrable systems (or equations) in control theory lies in the fact that they have rich information about the moduli (or parameter) space of a special class of solutions and they describe isospectral deformations of some linear operator or matrix. Because of this importance, various applications of nonlinear integrable systems have been proposed. We recall the early work by Hermann [9]. Helmke [8] and Sontag [22] studied moduli spaces of multimode and bilinear systems, respectively, in terms of the moduli space of instanton solutions of the self-dual Yang-Mills equations of gauge theory. Nakamura [14] introduced a fractional transformation group acting on the space of linear predictors (solutions of linear prediction problems for stochastic processes). It was shown in [14] and [15] that a special member of such a transformation group gives solutions of certain nonlinear integrable systems and describes a discrete isospectral deformation of a given matrix. Recently, Nakamura and Duncan [17] observed an equivalence between the moduli space of n-monopole solutions of the SU(2) Yang-Mills-Higgs equations and the space rat$^{\mathbb{C}}$ (n) of rational functions of degree n over $\mathbb{C}$.

Krishnaprasad, in his remarkable paper [11], considered the action of time shift (2) on the connected component rat $(n, 0)$ of rat (n) and showed that the action is equivalent to the flow of nonlinear integrable equation of Lax type called the finite nonperiodic Toda lattice [13]

$$\frac{d}{d\sigma} A = [A_L^T - A_L, A], \qquad A = \begin{pmatrix} b_1 & a_1 & \cdots & 0 \\ a_1 & & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1} \\ 0 & \cdots & a_{n-1} & b_n \end{pmatrix}, \tag{4}$$

where A_L denotes the strictly lower triangular part of A, $A_L^T = (A_L)^T$, $[M, N] = MN - NM$ and $a_j \neq 0$. Note that $A = A(\sigma)$ is a Jacobi (or real tridiagonal) matrix having n real distinct eigenvalues. In Moser's coordinates $\{\alpha_j, \zeta_j\}_{1\leqq j\leqq n}$ [13] defined by

$$e_n^T (zI - A)^{-1} e_n = \frac{1}{\sum_{j=1}^n \exp(\alpha_j)} \sum_{j=1}^n \frac{1}{z - \zeta_j} \exp(\alpha_j) \in \text{rat}(n, 0), \tag{5}$$

where $\alpha_j \in \mathbb{R}$, $\zeta_j \in \mathbb{R}$, $\zeta_k \neq \zeta_l$ for $k \neq l$ and $e_n = (0 \cdots 0\ 1)^T$, the Toda equation (4) becomes the system $(1 \leqq j \leqq n)$ of linear equations

$$\frac{d}{d\sigma} \alpha_j = \zeta_j, \qquad \frac{d}{d\sigma} \zeta_j = 0. \tag{6}$$

Hence α_j and ζ_j are regarded as generalized coordinates and momentums, respectively, for the dynamical system with Hamiltonian $H = \frac{1}{2} \text{Tr}(A^2) = \frac{1}{2} \sum_{j=0}^n \zeta_j^2$. Krishnaprasad

[11] proved that for $f(z)=e_n^T(zI-A)^{-1}e_n$ the action (2) takes the form

$$(7)\qquad f(z)\to e_n^T(zI-A)^{-1}e^{\sigma A}e_n=\frac{1}{\sum_{j=1}^n \exp(\alpha_j+\sigma\zeta_j)}\sum_{j=1}^n \frac{1}{z-\zeta_j}\exp(\alpha_j+\sigma\zeta_j)$$

by using the Laurent expansion $f(z)=\sum_{k=0}^{\infty} h_k z^{-k-1}$ and observed that the expression (7) is precisely the linearlized Toda flow (6). Thus it is concluded that the Toda equation (4) can be reduced to a linear flow on the connected component rat $(n, 0)$ induced by the time shift. We remark here that the Toda flow leaves invariant the poles of $f(z)$ and rat $(n, 0)$ is diffeomorphic to the phase space $\mathbb{R}^n\times\mathbb{R}^n$ of (4) via Moser's coordinates $\{\alpha_j, \zeta_j\}_{1\leq j\leq n}$. Inspired by this pioneer work, Byrnes [4] considered a periodic Toda flow on a space of matrix-valued rational functions. Let us note that the Toda flow also arises in closely related problems, for example, the topology of the space of symmetric matrices with fixed eigenvalues [24] and the existence of Morse-Smale diffeomorphism on some quotient space induced by linear maps with fixed eigenvalues [21].

It will be well worth stating how the invariant tori/cylinder theorem in classical mechanics [1, p. 395] can be used in the geometry of rational functions. Krishnaprasad and Martin [12] made clear the importance of the concept of families of linear systems and their parameter variations. They considered the family of controllable and observable linear systems and the corresponding space of rational functions

$$(8a)\qquad \mathrm{rat}_p(n)=\left\{f(z)\,\middle|\, f(z)=\frac{q(z)}{p(z)}\in \mathrm{rat}(n),\ p(z):\text{fixed}\right\},$$

where $p=\{p_0,\cdots,p_{n-1}\}\in\mathbb{R}^n$. It was known (cf. [12. Thm. 2]) that the upper bound of the number of connected components of $\mathrm{rat}_p(n)$ is 2^r, where r is the number of real distinct roots of $p(z)$. Krishnaprasad [11] also observed that each connected component of $\mathrm{rat}_p(n)$ is diffeomorphic to $T^l\times\mathbb{R}^{n-l}$, where T^l denotes the l-torus and $l\in\{0,\cdots,n-1\}$. The integer l is constant on an open subset of rat $(n-m, m)$. This implies that rat $(n-m, m)$ has an n-dimensional foliation whose leaves are diffeomorphic to $T^l\times\mathbb{R}^{n-l}$. For example, on the connected component rat $(n, 0)$, l is equal to zero. Let us fix denominators as some $p(z)$. Then the connected component $\mathrm{rat}_p(n, 0)$ of $\mathrm{rat}_p(n)$ is diffeomorphic to $\mathbb{R}^n$ and the foliation is a trivial fibration $\mathbb{R}^n\times\mathbb{R}^n$. Recall that $\mathbb{R}^n\times\mathbb{R}^n$ is the phase space of the Toda equation (4) via $\{\alpha_j, \zeta_j\}_{1\leq j\leq n}$, where ζ_j are real distinct roots of $p(z)=\det(zI-A)$ and A is Jacobi. Since ζ_j are conserved, we see that the Toda flow gives a parametrization (of an open dense subset) of $\mathrm{rat}_p(n, 0)$ (Moser and Krishnaprasad's theorem [11], [13]). However, it is not known what nonlinear integrable system completely parametrizes the whole space of rational functions with any fixed denominator.

What we shall do in this paper is to solve this open problem and to show that the same sort of analysis used in various problems [3], [4], [6], [11]-[13] can be performed in more general situations. We consider the space of equivalence classes

$$(8b)\qquad \mathrm{Rat}_p(n)=\mathrm{rat}_p(n)/\sim,$$

where $f(z)\sim e^{\theta}f(z)$ for $\theta\in\mathbb{R}$. The space $\mathrm{Rat}_p(n)$ is characterized by n real parameters and the first one can be limited to ± 1. The number of connected components of $\mathrm{Rat}_p(n)$ is equal to that of $\mathrm{rat}_p(n)$. It will be proved that the moduli space of solutions of a nonlinear integrable system is *diffeomorphic* to the space $\mathrm{Rat}_p(n)$ of rational functions. The nonlinear system we use is regarded as a natural generalization of the original Toda equation (4) and will be called the *cyclic-Toda hierarchy.*

2. The cyclic-Toda hierarchy. Let $f(z)$ be an element of $\mathrm{rat}_p(n)$ with fixed denominator $p(z)$. According to the state space realization theory, $f(z)$ admits the unique factorization

$$f(z) = C_0^T (zI - A_0)^{-1} B_0,$$

$$A_0 = \begin{pmatrix} 0 & \cdots & 0 & -p_0 \\ 1 & & \vdots & \vdots \\ 0 & \ddots & & \\ \vdots & & 0 & -p_{n-2} \\ 0 & \cdots & 1 & -p_{n-1} \end{pmatrix}, \tag{9}$$

$$B_0 = \begin{pmatrix} q_0 \\ \vdots \\ q_{n-2} \\ q_{n-1} \end{pmatrix}, \qquad C_0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix},$$

where the characteristic polynomial of A_0 is equal to the given $p(z)$,

$$p(z) = \det(zI - A_0). \tag{10}$$

The expression (9) of $f(z)$ is the observable canonical form, namely,

$$\operatorname{rank}(C_0 \; A_0^T C_0 \; \cdots \; (A_0^T)^{n-1} C_0) = n. \tag{11a}$$

Since the numerator $q(z)$ and the denominator $p(z)$ of $f(z)$ have no common factor, it follows that

$$\operatorname{rank}(B_0 \; A_0 B_0 \; \cdots \; A_0^{n-1} B_0) = n. \tag{11b}$$

The proof can be found in [10, p. 127]. The vectors B_0 and C_0 are called cyclic vectors of A_0. In this paper, we call $\{A_0, B_0, C_0\}$ a *cyclic triplet* if it satisfies (11a) and (11b). In linear systems theory $\{A_0, B_0, C_0\}$ is called a minimal realization of $f(z)$. We can associate each $f(z) = C_0^T (zI - A_0)^{-1} B_0 \in \mathrm{rat}_p(n)$ with the controllable and observable linear dynamical system

$$\frac{d}{dt} x(t) = A_0 x(t) + B_0 u(t), \qquad y(t) = C_0^T x(t). \tag{12}$$

A nonlinear (completely) integrable system is always related to some kind of group factorization. For example, the Toda equation (4) can be solved by the Bruhat decomposition [5], and the self-dual Yang-Mills equations are solved by the Riemann-Hilbert factorization [25]. The hierarchy point of view has also played quite an important role in the study of nonlinear integrable systems (see [19], [20] on soliton equations and [16] on gauge field equations). Here the terminology "hierarchy" indicates a system of differential equations that describes infinite (or finite) sets of commutative flows on the solution space of original equation. Let us show that if we go further in this direction toward a characterization of the space $\mathrm{Rat}_p(n)$ in terms of a nonlinear intgrable system, then we shall encounter a finite hierarchy of cyclic-Toda equations in a very natural way.

Every real nonsingular matrix M has a unique QR factorization $M = Q^{-1}R$, where Q is orthogonal and R is upper triangular with positive diagonal entries. Proof is given by the Gramm-Schmid orthogonalization (see [26, pp. 233-244]). Let τ be a set of n real parameters,

$$\tau = \{\tau_0, \cdots, \tau_{n-1}\} \in \mathbb{R}^n. \tag{13}$$

Let us consider the QR factorization with parameters

$$\exp\left(\sum_{k=0}^{n-1}\tau_k A_0^k\right)=Q^{-1}(\tau)R(\tau),\quad Q(\mathbf{0})=I,\quad R(\mathbf{0})=I, \tag{14}$$

where $\mathbf{0}=\{0,\cdots,0\}\in\mathbb{R}^n$. The parameter τ_0 appears only in $R(\tau)$ as a diagonal factor $\exp(\tau_0 I)$. Taking derivatives on both sides, we have

$$QA_0^kQ^{-1}=-\frac{\partial}{\partial\tau_k}Q\cdot Q^{-1}+\frac{\partial}{\partial\tau_k}R\cdot R^{-1} \tag{15}$$

for $0\leqq k\leqq n-1$. Since the first and the second terms of the right-hand side are skew-symmetric and upper triangular, it follows that

$$\begin{aligned}\frac{\partial}{\partial\tau_k}Q\cdot Q^{-1}&=(QA_0^kQ^{-1})_L^T-(QA_0^kQ^{-1})_L,\\ \frac{\partial}{\partial\tau_k}R\cdot R^{-1}&=(QA_0^kQ^{-1})_U+(QA_0^kQ^{-1})_L^T,\end{aligned} \tag{16}$$

where $M_U=M-M_L$. We have proved Lemma 2.1.

LEMMA 2.1. *If the nonsingular matrix* $\exp(\sum_{k=0}^{n-1}\tau_k A_0^k)$ *has the* QR *factorization* (14), *then the resulting* C^∞*-factors* $Q(\tau)$ *and* $R(\tau)$ *solve the initial value problem* (16) *with* $Q(\mathbf{0})=I$ *and* $R(\mathbf{0})=I$.

It is to be noted that $QA_0^kQ^{-1}$ in the right-hand side of (16) is the kth power of the similarity transform QA_0Q^{-1} of A_0 by the factor Q. The next lemma shows that $Q(\tau)$ induces a flow on $\mathbb{R}^{n\times n}$ characterized by a system of nonlinear equations.

LEMMA 2.2. *Let* $A(\tau)$ *be a* C^∞*-matrix defined by*

$$A(\tau)=Q(\tau)A_0Q^{-1}(\tau); \tag{17}$$

then $A(\tau)$ *solves the initial value problem for the system* $(0\leqq k\leqq n-1)$

$$\begin{aligned}\frac{\partial}{\partial\tau_k}A(\tau)&=[A^k(\tau)_L^T-A^k(\tau)_L,A(\tau)]\\&=[A^k(\tau)_U+A^k(\tau)_L^T,A(\tau)],\qquad A(\mathbf{0})=A_0,\end{aligned} \tag{18}$$

where $A_L^{kT}=(A^k)_L^T$.

Taking derivatives of (17), we have $(\partial/\partial\tau_k)A=[(\partial/\partial\tau_k)Q\cdot Q^{-1},A]$. By (16) with (15) and (17), the assertion of Lemma 2.2 is proved. The solution $A(\tau)$ clearly does not depend on the parameter τ_0; however, τ_0 is useful in our parametrization of $\mathrm{Rat}_p(n)$. Lemmata 2.1 and 2.2 imply that the maximal interval of existence for initial value problem (18) is $(-\infty,\infty)$ for each τ_k. Furthermore, since $A(\tau)$ is always orthogonally similar to A_0 for $\tau\in\mathbb{R}^n$, the flow $A(\tau)$ is *isospectral*, namely, $\mathrm{Spec}\,A(\tau)=\mathrm{Spec}\,A_0$. Indeed, the trace of $A^k(\tau)$ is an integral of motion of (18) expressed by the eigenvalues of A_0. Since A_0 commutes with $Q^{-1}(\tau)R(\tau)$, we obtain $A(\tau)=R(\tau)A_0R^{-1}(\tau)$. This implies that $A(\tau)$ is upper-Hessenberg. Set $\mathbb{1}=\{1,\cdots,1\}\in\mathbb{R}^n$. From (14), $\exp(\sum_{k=0}^{n-1}A_0^k)=Q^{-1}(\mathbb{1})R(\mathbb{1})$. On the other hand, since

$$\begin{aligned}\exp\left(\sum_{k=0}^{n-1}\tau_k A^k(\tau)\right)&=Q(\tau)\exp\left(\sum_{k=0}^{n-1}\tau_k A_0^k\right)Q^{-1}(\tau)\\&=R(\tau)Q^{-1}(\tau),\end{aligned}$$

we derive $\exp(\sum_{k=0}^{n-1}A^k(\mathbb{1}))=R(\mathbb{1})Q^{-1}(\mathbb{1})$. Thus the time evolution (17) from $\tau=\mathbf{0}$ to $\tau=\mathbb{1}$ performs the QR iteration [26, pp. 515–521] for the exponential of matrix

$\exp(\sum_{k=0}^{n-1} A_0^k)$. If the eigenvalues of A_0 satisfy $|\lambda_1|<|\lambda_2|<\cdots<|\lambda_n|$ and A_0 can be diagonalized, $MA_0M^{-1}=\operatorname{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n)$, by a nonsingular matrix M which admits an LU factorization, then $A(\tau)$ approaches an upper triangular matrix as $|\tau|\to\infty$.

Next, we introduce a time evolution of the vectors B_0 and C_0.

LEMMA 2.3. *Let $B(\tau)$ and $C(\tau)$ be vectors defined by*

$$B(\tau)=R(\tau)B_0,\qquad C(\tau)=Q(\tau)C_0; \tag{19}$$

then they satisfy the system of linear equations $(0\leqq k\leqq n-1)$

$$\begin{aligned}\frac{\partial}{\partial\tau_k}B(\tau)&=(A^k(\tau)+A^k(\tau)_L^T-A^k(\tau)_L)B(\tau),\qquad B(\mathbf{0})=B_0,\\ \frac{\partial}{\partial\tau_k}C(\tau)&=(A^k(\tau)_L^T-A^k(\tau)_L)C(\tau),\qquad C(\mathbf{0})=C_0.\end{aligned} \tag{20}$$

The proof is given by (16) with (15) and (17). Thus we obtain the nontrivial flow $\{A_0,B_0,C_0\}\to\{A(\tau),B(\tau),C(\tau)\}$ on $\mathbb{R}^{n\times n+2n}$. The following proposition is the key connection between the system (18) with (20) and the space $\operatorname{rat}_p(n)$ of rational functions.

PROPOSITION 2.4. *If $\{A_0,B_0,C_0\}$ is a cyclic triplet, then $\{A(\tau),B(\tau),C(\tau)\}$ is also a cyclic triplet for every $\tau\in\mathbb{R}^n$.*

Proof. It is known from (14), (17), and (19) that

$$\begin{aligned}(B\,AB\cdots A^{n-1}B)&=(RB_0\;QA_0Q^{-1}RB_0\cdots QA_0^{n-1}Q^{-1}RB_0)\\&=Q(\exp(H_\tau)B_0\;A_0\exp(H_\tau)B_0\cdots A_0^{n-1}\exp(H_\tau)B_0)\\&=Q\exp(H_\tau)(B_0\;A_0B_0\cdots A_0^{n-1}B_0),\\(C\,A^TC\cdots(A^T)^{n-1}C)&=Q(C_0\;A_0^TC_0\cdots(A_0^T)^{n-1}C_0),\end{aligned}$$

where $H_\tau=\sum_{k=0}^{n-1}\tau_kA_0^k$. This implies that $\operatorname{rank}(B\,AB\cdots A^{n-1}B)=n$ and $\operatorname{rank}(C\,A^TC\cdots(A^T)^{n-1}C)=n$. □

Since $\operatorname{Spec}A(\tau)=\operatorname{Spec}A_0$, it is easy to prove Proposition 2.5.

PROPOSITION 2.5. *If the controllable and observable linear system* (12) *is stable* (*respectively, unstable*), *then the parametric linear systems*

$$\begin{aligned}\frac{d}{dt}x(t;\tau)&=A(\tau)x(t;\tau)+B(\tau)u(t;\tau),\\y(t;\tau)&=C(\tau)^Tx(t;\tau)\end{aligned} \tag{21}$$

are stable (*respectively, unstable*) *as well as controllable and observable.*

The hierarchy (18) for $0\leqq k\leqq n-1$ is a generalization of the system of nonlinear equations discussed by Moser [13], $(d/dt)L=[B_p,L]$ for $1\leqq p\leqq n-1$ in his notation, where L is *Jacobi* and $(d/dt)L=[B_1,L]$ is the original Toda equation (4). Indeed, when $f(z)$ is in rat $(n,0)$, we can always transform $A(\tau)$ into a Jacobi matrix via Cauer's canonical form (5). Moreover, the hierarchy is clearly different from the known hierarchy of finite Toda equations [23] which is a finite truncated form of the infinite Toda hierarchy. The system (20) describing the time evolution of cyclic vectors is essentially new.

Finally in this section, we discuss the compatibility of (18) and (20) in the sense in which $A(\tau)$, $B(\tau)$, and $C(\tau)$ satisfy the integrability condition with respect to τ. Set

$$F_k(A)=A^k(\tau)_L^T-A^k(\tau)_L, \tag{22}$$

for simplicity. By a standard theory of Toda equation, we see that each equation of (18) is equivalent to the Lax pair $AY=\zeta Y$ and $(\partial/\partial\tau_k)Y=F_k(A)Y$, where $Y=Y(\tau;\zeta)\in\mathbb{C}^{n\times 1}$. The integrability for Y leads to the system of zero curvature equations

$$\frac{\partial}{\partial\tau_l}F_k(A)-\frac{\partial}{\partial\tau_k}F_l(A)+[F_k(A),F_l(A)]=0 \tag{23}$$

for $0\leqq k,\ l\leqq n-1$. Thus if $F_k(A)$ solve (23), then the system (18) is clearly integrable. It is to be noted that the skew symmetric matrices $F_k(A)$ given by $F_k(A)=(\partial/\partial\tau_k)Q\cdot Q^{-1}$ obviously satisfy (23). See (16), (17), and (22). Conversely, let $A(\tau)$ solve (18). We have

$$\begin{aligned}0&=\frac{\partial}{\partial\tau_l}A^k-[F_l,A^k]\\&=\left(\frac{\partial}{\partial\tau_l}G_k-\frac{\partial}{\partial\tau_k}F_l-[F_l,G_k]\right)+\left(\frac{\partial}{\partial\tau_k}F_l-\frac{\partial}{\partial\tau_l}F_k+[F_l,F_k]\right),\end{aligned}$$

where $G_k=A^k+F_k$. Noting that $A(\tau)$ is upper-Hessenberg and G_k is upper triangular, we can prove that $F_k=F_k(A)$ solves (23). Thus (18) is equivalent to (23). Equations (20) are expressed as $(\partial/\partial\tau_k)B=(A^k+F_k(A))B$ and $(\partial/\partial\tau_k)C=F_k(A)C$. Since

$$\left(\frac{\partial}{\partial\tau_l}\frac{\partial}{\partial\tau_k}-\frac{\partial}{\partial\tau_k}\frac{\partial}{\partial\tau_l}\right)H=\left(\frac{\partial}{\partial\tau_l}F_k(A)-\frac{\partial}{\partial\tau_k}F_l(A)+[F_k(A),F_l(A)]\right)H$$

for $H=B$ and C, the system (23) guarantees the integrability of linear system (20). We conclude that the hierarchy (18) with (20) is compatible. Since (18) satisfies the Cauchy-Kovalevskaya condition, the solution given by (17) is unique. In the next section, we shall show that the triplet $\{A(\tau),B(\tau),C(\tau)\}$ is quite important in the parametrization of $\mathrm{Rat}_p(n)$.

DEFINITION 2.6. The hierarchy of nonlinear integrable equations (18) with the supplementary system (20), where $0\leqq k\leqq n-1$, is called the *cyclic-Toda hierarchy*. The space of C^∞-solutions $A(\tau)$ of (18) for *any* equivalence class $[\{A_0,B_0,C_0\}]$ of initial cyclic triplets satisfying (9)-(11b) under the relation $f(z)\sim e^\theta f(z)$, $\theta\in\mathbb{R}$, is called the *moduli* (or *parameter*) *space* of the cyclic-Toda hierarchy. We denote the moduli space as $\mathcal{M}_p$.

3. Parametrization of the space of rational functions: main result. First we show that the cyclic-Toda hierarchy induces a set of n flows parametrized by $\tau\in\mathbb{R}^n$ on the space $\mathrm{rat}_p(n)$ of rational functions of degree n and the fixed denominator $p(z)$. In the previous section, we obtained the nontrivial flow $\{A_0,B_0,C_0\}\to\{A(\tau),B(\tau),C(\tau)\}$ on $\mathbb{R}^{n\times n+2n}$. This flow can be projected onto the space $\mathrm{rat}_p(n)$. For any $f(z)=C_0^T(zI-A_0)^{-1}B_0\in\mathrm{rat}_p(n)$, let us define

$$f(z;\tau)=C(\tau)^T(zI-A(\tau))^{-1}B(\tau). \tag{24}$$

From Proposition 2.4 and $\det(zI-A(\tau))=p(z)$, it is not hard to see $f(z;\tau)\in\mathrm{rat}_p(n)$ for every $\tau\in\mathbb{R}^n$. Furthermore, we prove Proposition 3.1.

PROPOSITION 3.1. *The parameter τ_0 and τ_1 of τ describe the amplitude scaling and the shift on time axis, respectively, for the controllable and observable linear system* (12).

Proof. From the definition of the flow and $Q^TQ=I$, we have

$$\begin{aligned}f(z;\tau)&=(QC_0)^T(zI-QA_0Q^{-1})^{-1}RB_0\\&=C_0^T(zI-A_0)^{-1}Q^{-1}RB_0\\&=C_0^T(zI-A_0)^{-1}\exp(H_\tau)B_0,\end{aligned}$$

where $H_\tau = \sum_{k=0}^{n-1} \tau_k A_0^k$. Thus $f(z;\tau)$ is the Laplace transform of the weight function $W(t;\tau) = C_0^T \exp(tA_0 + H_\tau)B_0$. The parameter τ_0 clearly describes the amplitude scaling $f(z) \to \exp(\tau_0) f(z)$. Setting $\tau = \{0, \tau_1, 0, \cdots, 0\}$, we obtain the expression

$$f(z) \to f(z;\tau) = C_0^T (zI - A_0)^{-1} \exp(\tau_1 A_0) B_0$$

of the action of time shift for the linear system (12). □

This proposition is a straightforward generalization of the result in [3] and [11] (see (2)-(6)). But the proof has been carried out without using the Laurent expansion of $f(z)$ and without supposing that A_0 is Jacobi and $B_0 = C_0 = e_n$. The remaining parameters $\{\tau_2, \cdots, \tau_{n-1}\}$ give actions of one-parameter groups being outside the known ones.

Now we shall discuss the correspondence between $\mathrm{Rat}_p(n)$ and the moduli space $\mathcal{M}_p$ of the cyclic-Toda hierarchy. First we shall construct a one-to-one mapping from $\mathrm{Rat}_p(n)$ to $\mathcal{M}_p$. Let $A(\tau)$ be an arbitrary C^∞-solution of the cyclic-Toda hierarchy for an initial value $\{A_0, B_0, C_0\}$ satisfying (9)-(11b). Recall that $A(\tau)$ is an isospectral deformation of A_0. By a theorem [7, p. 219] there are nonsingular upper triangular matrices $R(\tau)$ such that

$$A(\tau)R(\tau) = R(\tau)A_0, \qquad R(\mathbf{0}) = I. \tag{25}$$

We note that each $R(\tau)$ with B_0 gives a cyclic vector $B(\tau)$ via (19) which solves the supplementary system (20). Thus $(\partial/\partial\tau_k)R \cdot R^{-1} = A^k + F_k(A)$. Let us consider

$$\frac{\partial}{\partial \tau_k} Q(\tau) = F_k(A)Q(\tau), \qquad Q(\mathbf{0}) = I. \tag{26}$$

There is a unique C^∞-solution of (26). This follows from the compatibility proved in § 2. The resulting $Q(\tau)$ and C_0 give $C(\tau)$, which solves (20). Since $F_k(A)$ is skew-symmetric, $Q(\tau)$ is an orthogonal matrix. Multiplying $A(\tau)$ to (26) from the left and using $(\partial/\partial\tau_k)A = [F_k(A), A]$ derived from (18) and (22), we obtain $(\partial/\partial\tau_k)(AQ) = F_k(A)AQ$, where $A(\mathbf{0})Q(\mathbf{0}) = A_0$. On the other hand, $(\partial/\partial\tau_k)QA_0 = F_k(A)QA_0$ and $Q(\mathbf{0})A_0 = A_0$. Because of the uniqueness of solution, $A(\tau)Q(\tau) = Q(\tau)A_0$. Thus we obtain $Q(\tau)A_0Q^{-1}(\tau) = A(\tau)$. Set $U(\tau) = Q^{-1}(\tau)R(\tau)$. By differentiating this, we have

$$\begin{aligned} Q(\tau)\frac{\partial}{\partial \tau_k} U(\tau) \cdot R^{-1}(\tau) &= -\frac{\partial}{\partial \tau_k} Q(\tau) \cdot Q^{-1}(\tau) + \frac{\partial}{\partial \tau_k} R(\tau) \cdot R^{-1}(\tau) \\ &= A^k(\tau) \\ &= Q(\tau) A_0^k Q^{-1}(\tau) \end{aligned}$$

and consequently, we see that $U(\tau)$ satisfies $(\partial/\partial\tau_k)U = A_0^k U$, $U(\mathbf{0}) = I$. Thus

$$U(\tau) = \exp\left(\sum_{k=0}^{n-1} \tau_k A_0^k\right). \tag{27}$$

Let $f(z;\tau)$ be a rational function defined by $f(z;\tau) = C(\tau)^T (zI - A(\tau))^{-1} B(\tau)$. Since $f(z;\tau) = C_0^T (zI - A_0)^{-1} U(\tau) B_0$, we conclude from (27) and Proposition 2.4 that $f(z;\tau) \in \mathrm{rat}_p(n)$ for any $\tau \in \mathbb{R}^n$. Note that the solution of (25) is not unique even if we suppose that the diagonal entries of $R(\tau)$ are positive. Indeed, if $R(\tau)$ satisfies (25), then $R(\tau')$ does also, where $\tau' = \{\tau_0 + \theta, \tau_1, \cdots, \tau_{n-1}\}$ for $\theta \in \mathbb{R}$. By identifying $R(\tau)$ with $R(\tau')$, we can obtain a unique rational function of $\mathrm{Rat}_p(n)$. Note that this identification amounts to the equivalence class $[\{A_0, B_0, C_0\}]$ of initial values. Namely, each point on the cyclic-Toda flow $A(\tau)$ for $[\{A_0, B_0, C_0\}]$ corresponds to a unqiue rational function of $\mathrm{Rat}_p(n)$. Thus we have obtained a one-to-one mapping $\alpha: \mathrm{Rat}_p(n) \to \mathcal{M}_p$.

Conversely, let $f(z)$ be an arbitrary rational function of $\mathrm{rat}_p(n)$. Recall that $f(z)$ admits a unique factorization (9). As was shown in the previous section, there always exists a flow of the cyclic-Toda hierarchy $\{A(\tau), B(\tau), C(\tau)\}$ for $\tau \in \mathbb{R}^n$. We see from (24) that this flow induces a flow $f(z;\tau)$ on $\mathrm{rat}_p(n)$ which passes $f(z)$ at $\tau = \mathbf{0}$. Furthermore, by introducing the equivalence relation $f(z) \sim e^{\theta} f(z)$, $\theta \in \mathbb{R}$, we obtain a unique cyclic-Toda flow for the equivalence class $[\{A_0, B_0, C_0\}]$ of initial triplets. Thus α is an onto mapping from $\mathrm{Rat}_p(n)$ to $\mathcal{M}_p$. We have established the following theorem.

THEOREM 3.2. *There is a one-to-one correspondence between* (a) *the space* $\mathrm{Rat}_p(n)$ *of rational functions of degree n and fixed denominator* $p(z)$ *and* (b) *the moduli space* $\mathcal{M}_p$ *of the cyclic-Toda hierarchy for any equivalence class* $[\{A_0, B_0, C_0\}]$ *of initial cyclic triplets.*

Furthermore, the mapping $\alpha: \mathrm{Rat}_p(n) \to \mathcal{M}_p$ clearly depends differentiably on each parameter of $\tau \in \mathbb{R}^n$. Recall that solutions $Q(\tau)$ and $R(\tau)$ of (25) and (26) are of class C^{∞}. This guarantees the differentiability of α^{-1}. Combining this with Theorem 3.2, we have proved a stronger result.

THEOREM 3.3. *The bijection* $\alpha: \mathrm{Rat}_p(n) \to \mathcal{M}_p$ *is a diffeomorphism with the natural topologies on* $\mathrm{Rat}_p(n)$ *and* $\mathcal{M}_p$ *induced by* $\mathbb{R}^n$.

To compute the number of connected components of $\mathrm{Rat}_p(n)$, it is important to recall that only real zeros of $f(z)$ create obstructions to deformations on $\mathrm{rat}_p(n)$ [12]. This fact was first noted by Brockett [2]. The position of zeros of (9) depends on A_0 and B_0. The cyclic-Toda hierarchy for the initial value $\{A_0, B_0, C_0\}$ induces a set of flows parametrized by $\tau \in \mathbb{R}^n$ on one of the connected components of $\mathrm{rat}_p(n)$. Since C_0 is fixed, the choice of the connected components is determined by that of A_0 and B_0. Let $f(z) = q(z)/p(z) \in \mathrm{rat}_p(n)$ and ζ_j be roots of $p(z) = \det(zI - A_0)$. By the spectral mapping theorem we see a cyclic matrix polynomial $\sum_{k=0}^{n-1} q_k A_0^k$ is nonsingular if and only if $q(\zeta_j) \neq 0$ for any ζ_j. Thus any rational function of $\mathrm{Rat}_p(n)$ uniquely determines a nonsingular matrix H_q of the space $\mathcal{H}_{A_0}$ defined by

$$\mathcal{H}_{A_0} = \mathcal{h}_{A_0}/\sim, \qquad \mathcal{h}_{A_0} = \{H_q \mid H_q = \textstyle\sum_{k=0}^{n-1} q_k A_0^k : \text{nonsingular}, q \in \mathbb{R}^n\}, \tag{28}$$

where $H_q \sim e^{\theta} H_q$ for $\theta \in \mathbb{R}$ and vice versa. The mapping $\beta: \mathrm{Rat}_p(n) \to \mathcal{H}_{A_0}$ and its inverse are clearly differentiable with respect to $q = \{q_0, \cdots, q_{n-1}\} \in \mathbb{R}^n$. Thus $\mathcal{H}_{A_0}$ is diffeomorphic to $\mathrm{Rat}_p(n)$. Define the spaces of cyclic vectors

$$\mathcal{W}_{A_0} = \omega_{A_0}/\sim, \qquad \omega_{A_0} = \{W \mid W \in \mathbb{R}^{n\times 1}, \operatorname{rank}(W A_0 W \cdots A_0^{n-1} W) = n\}, \tag{29}$$

where $W \sim e^{\theta} W$ for $\theta \in \mathbb{R}$. We see that $\mathcal{H}_{A_0}$ acts freely on $\mathcal{W}_{A_0}$, $\mathcal{H}_{A_0} \times \mathcal{W}_{A_0} \to \mathcal{W}_{A_0}$ by $(H_q, W) \to H_q W$. Thus $\mathcal{W}_{A_0}$ is diffeomorphic to $\mathcal{H}_{A_0}$ and the number of connected components of $\mathcal{W}_{A_0}$ is equal to that of $\mathrm{Rat}_p(n)$.

Suppose that $p(z)$ admits r real distinct zeros ζ_j, where $r \geqq 1$ and $1 \leqq j \leqq r$. We write ζ_j in order, $\zeta_1 < \zeta_2 < \cdots < \zeta_r$. Set $s = \{s_0, s_1, \cdots, s_{r-1}\}$, where $s_j = 0$ or 1. If an odd number of zeros of $f(z)$ are on the interval (ζ_j, ζ_{j+1}), then we assign 1 to s_j for $1 \leqq j \leqq r-1$. Otherwise, we assign zero to s_j. Set $s_0 = 1$ (respectively, zero) if the coefficient of the highest power of z in $q(z)$ is positive (respectively, negative). Thus the multi-index s labels the connected components of $\mathrm{rat}_p(n)$. The maximum number of such components is 2^r. We express each component as $\mathrm{rat}_p^s(n)$. If there is no real root of $p(z)$, $\mathrm{rat}_p(n)$ itself is connected and denoted by $\mathrm{rat}_p^{\phi}(n)$. Note that $\mathrm{Rat}_p^s(n) = \mathrm{rat}_p^s(n)/\sim$ is exactly one of the connected components of $\mathrm{Rat}_p(n)$. From this result and Theorem 3.3, we prove Corollary 3.4.

COROLLARY 3.4. *Let* $\{A_0, B_0, C_0\}$ *be a cyclic triplet such that* $C_0^T (zI - A_0)^{-1} B_0 \in \mathrm{rat}_p^s(n)$. *Then the cyclic-Toda flow* $A(\tau)$, $\tau \in \mathbb{R}^n$, *for the initial value* $\{A_0, B_0, C_0\}$ *is identified with the connected component* $\mathrm{Rat}_p^s(n)$.

This corollary hints that the space $\mathrm{Rat}_p(n)$ is the primary object, rather than $\mathrm{rat}_p(n)$ that was used to prove Theorem 3.2.

Finally, we shall give simple but conspicuous examples. Let

$$\left\{A_0=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, B_0=\begin{pmatrix}\alpha\\ \beta\end{pmatrix}, C_0=\begin{pmatrix}0\\ 1\end{pmatrix}\right\}$$

be a cyclic triplet. We obtain $f(z)=(\alpha+\beta)/2(z-1)+(\beta-\alpha)/2(z+1)$ and

$$\exp\left(\sum_{k=0}^{1}\tau_k A_0^k\right)=\exp(\tau_0)\cdot\begin{pmatrix}\cosh(\tau_1) & \sinh(\tau_1)\\ \sinh(\tau_1) & \cosh(\tau_1)\end{pmatrix}.$$

If $|\beta|>|\alpha|$, then the zero of $f(z)$ is on the interval $(-1, 1)$, and then the flow is identified with $\mathrm{Rat}_p^s(n)$, where $p=\{0,-1\}$ and $s=\{1,1\}$ for $\beta>0$ or $s=\{0,1\}$ for $\beta<0$. Let us note that the Cauchy index of $f(z;\tau)$ is equal to ± 2, namely, $f(z;\tau)\in \mathrm{rat}_p(2,0)\approx\mathbb{R}^2$ or $\mathrm{rat}_p(0,2)\approx\mathbb{R}^2$. Setting $\alpha=0$, we derive a hierarchy of the usual 2×2 Toda equation. If $|\alpha|>|\beta|\neq 0$, then $f(z)$ admits a zero on $(-\infty,-1)$ or $(1,\infty)$. Therefore, the resulting flow is identified with $\mathrm{Rat}_p^{s'}(n)$, where $s'=\{1,0\}$ for $\beta>0$ or $s'=\{0,0\}$ for $\beta<0$. Here the Cauchy index of $f(z;\tau)$ is always zero, namely, $f(z;\tau)\in\mathrm{rat}_p(1,1)\approx\mathbb{R}^1\times S^1$. If $\beta=0$, then $f(z)$ has no zero. The flow is identified with $\mathrm{Rat}_p^{s''}(n)$, where $s''=\{1,0\}$ for $\alpha>0$ or $s''=\{0,0\}$ for $\alpha<0$, and then $f(z;\tau)\in\mathrm{rat}_p(1,1)$.

Let $\tilde{p}(z)=z^n+\tilde{p}_{n-1}z^{n-1}+\cdots+\tilde{p}_0$ have n real distinct roots. Any $f(z)$ of $\mathrm{rat}_{\tilde{p}}(n)$ can be factored into $f(z)=C_0^T(zI-A_0)^{-1}B_0$, where $B_0=C_0$, A_0 is Jacobi, and B_0 is cyclic. In this case, we obtain a subhierarchy of (18) called the *Jacobi-Toda hierarchy* whose moduli space is homeomorphic to $\mathrm{Rat}_{\tilde{p}}(n)$. By setting $B_0=(0\cdots 01)^T$, we obtain the Jacobi-Toda flow being identified with $\mathrm{Rat}_{\tilde{p}}^s(n)$, where $s=\{s_0,1,\cdots,1\}$. This is exactly the case analyzed by Moser [13]. The original Toda equation (4) is a special member of this flow parametrized by τ_1. Thus we can conclude that the extension to general cyclic triplet $\{A_0, B_0, C_0\}$ is not a trivial formal generalization. As we have observed, the cyclic-Toda hierarchy gives a *complete* parametrization of the space $\mathrm{Rat}_p(n)$ of controllable and observable linear systems for the fixed polynomial $p(z)=\det(zI-A_0)$. Let us recall that by taking the union with respect to $p=\{p_0,\cdots,p_{n-1}\}$, we recover the whole space $\mathrm{Rat}(n)$ of rational functions of degree n,

$$\mathrm{Rat}(n)=\bigcup_{p\in\mathbb{R}^n}\mathrm{Rat}_p(n).$$

It would be interesting to study the topology of the flows of nonlinear integrable systems on $\mathrm{Rat}(n)$ and its applications, for example, to various cellular decompositions of $\mathrm{Rat}(n)$ and to the problem of limiting linear systems.

Acknowledgments. It is a pleasure to thank T. E. Duncan, H. Fukawa, and P. S. Krishnaprasad for their helpful advice, and especially D. Hinrichsen for various fruitful discussions during my stay as a guest professor of the Forschungsschwerpunkt Dynamische Systeme, Universität Bremen.

REFERENCES

[1] R. ABRAHAM AND J. E. MARSDEN, *Foundations of Mechanics*, Benjamin/Cummings, London, 1978.

[2] R. W. BROCKETT, *Some geometrical questions in the theory of systems*, IEEE Trans. Automat. Control, 21 (1976), pp. 449-455.

[3] R. W. BROCKETT AND P. S. KRISHNAPRASAD, *A scaling theory for linear systems*, IEEE Trans. Automat. Control, 25 (1980), pp. 197-207.

[4] C. I. BYRNES, *On certain families of rational functions arising in dynamics*, in Proc. 1978 IEEE Conference on Decision and Control, San Diego, CA, 1979, pp. 1002-1006.

[5] P. DEIFT, L. C. LI, AND C. TOMEI, *Matrix factorization and integrable systems*, Comm. Pure Appl. Math., 42 (1989), pp. 443-521.

[6] P. A. FUHRMANN AND P. S. KRISHNAPRASAD, *Toward a cell decomposition for rational functions*, IMA J. Math. Control Inform., 3 (1986), pp. 137-150.

[7] F. R. GANTMACHER, *The Theory of Matrices Vol.* I, Chelsea, New York, 1959.

[8] U. HELMKE, *Linear dynamical systems and instantons in Yang-Mills theory*, IMA J. Math. Control Inform., 3 (1986), pp. 151-166.

[9] R. HERMANN, *Cartanian Geometry, Nonlinear Waves, and Control Theory Part* A, Interdisciplinary Mathematics, Vol. 20, Mathematical Science Press, Brookline, MA, 1979.

[10] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[11] P. S. KRISHNAPRASAD, *Symplectic mechanics and rational functions*, Ricerche Automat., 10 (1979), pp. 107-135.

[12] P. S. KRISHNAPRASAD AND C. F. MARTIN, *On a family of systems and deformations*, Internat. J. Control, 38 (1983), pp. 1055-1079.

[13] J. MOSER, *Finitely many mass points on the line under the influence of an exponential potential—An integrable system*, in Dynamical Systems, Theory and Applications, J. Moser, ed., Lecture Notes Physics, Vol. 38, Springer-Verlag, New York, 1975, pp. 467-497.

[14] Y. NAKAMURA, *Fractional transformation group induced by* QR *factorization and linear prediction problems*, Systems Control Lett., 10 (1988), pp. 181-184.

[15] ———, *Applications of transformation theory for nonlinear integrable systems to linear prediction problems and isospectral deformations*, in Algebraic Analysis Vol. II, M. Kashiwara and T. Kawai, eds., Academic Press, New York, 1988, pp. 505-515.

[16] ———, *Transformation group acting on a self-dual Yang-Mills hierarchy*, J. Math. Phys., 29 (1988), pp. 244-248.

[17] Y. NAKAMURA AND T. E. DUNCAN, *Remarks on the moduli space of* SU(2) *monopoles and Toda flow*, Lett. Math. Phys., 19 (1990), pp. 127-131.

[18] K. PEARSON, *Contribution to the mathematical theory of evolution* II. *Skew variation in homogeneous material*, Philos. Trans. Roy. Soc. London, 186 (1895), pp. 343-414.

[19] M. SATO AND Y. SATO, *Soliton equations as dynamical systems on infinite dimensional Grassmann manifold*, in Nonlinear PDE in Applied Science; Proc. U.S.-Japan Seminar, Tokyo, 1982, H. Fujita, P. D. Lax, and G. Strang, eds., Lecture Notes in Numerical and Applied Analysis, Vol. 5, Kinokuniya, Tokyo, 1983, pp. 259-271.

[20] G. SEGAL AND G. WILSON, *Loop groups and equations of* KdV *type*, Publ. Math. I.H.E.S., 61 (1985), pp. 5-65.

[21] M. SHUB AND A. T. VASQUEZ, *Some linearly induced Morse-Smale systems, the* QR *algorithm and the Toda lattice*, in The Legacy of Sonya Kovalevskaya, L. Keen, ed., Contemporary Mathematics, Vol. 64, American Mathematical Society, Providence, RI, 1987, pp. 181-194.

[22] E. D. SONTAG, *A remark on bilinear systems and moduli spaces of instantons*, Systems Control Lett., 9 (1987), pp. 361-367.

[23] K. TAKASAKI, *Initial value problem for the Toda lattice hierarchy*, in Group Representations and Systems of Differential Equations, K. Okamoto, ed., Advanced Studies in Pure Mathematics, Vol. 4, Kinokuniya, Tokyo, 1984, pp. 139-163.

[24] C. TOMEI, *The topology of isospectral manifolds of tridiagonal matrices*, Duke Math. J., 51 (1984), pp. 981-996.

[25] K. UENO AND Y. NAKAMURA, *Transformation theory for anti-self-dual equations*, Publ. Res. Inst. Math. Sci., Kyoto Univ., 19 (1983), pp. 519-547.

[26] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, 1965.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1755-1762, November 1991

014

THE INTERIOR TRANSMISSION PROBLEM AND INVERSE SCATTERING FROM INHOMOGENEOUS MEDIA*

B. P. RYNNE† AND B. D. SLEEMAN‡

Abstract. This paper is concerned with the class of far field patterns corresponding to the scattering of time harmonic acoustic plane waves by an inhomogeneous medium in a bounded domain B, with refractive index $n(x)$. It has previously been shown that the class of far field patterns is complete in $L_2(S^2)$ except at wavenumbers k, which are so-called transmission eigenvalues of the homogeneous interior transmission problem. In this paper the interior transmission problem is studied and, under milder conditions on n than previously used, the set of transmission eigenvalues is shown to be discrete. Also, at points other than transmission eigenvalues, it is shown that the inhomogeneous interior transmission problem is uniquely solvable. This result is of importance in certain methods for solving the inverse scattering problem of determining the function n from the scattered far fields.

Key words. far field patterns, acoustic waves, inverse scattering

AMS(MOS) subject classifications. 35P25, 76Q05

1. Introduction. This paper is concerned with the class of far field patterns corresponding to the scattering of time harmonic acoustic plane waves by an inhomogeneous medium in a bounded domain B, with refractive index $n(x)$. In [4] it was shown that the class of far field patterns is complete in the Hilbert space $L_2(S^2)$ (where S^2 is the unit sphere in $\mathbb{R}^3$) for all positive wavenumbers k, except possibly on a discrete set of points, provided that the medium is nonabsorbing and the function $m(x) = 1 - n(x)$ is smooth, is strictly positive (or strictly negative) in B, and satisfies a certain integral bound (which restricts the behavior of m near the boundary of B). Here we allow m to be nonsmooth and to change sign. In addition, the medium may be absorbent in B. However, we assume that m is bounded away from zero in B; thus if m changes sign it must do so discontinuously. Also, m must be discontinuous at the boundary ∂B of B (since by definition, $m \equiv 0$ outside B). This is in contrast with the condition imposed in [4], which requires that m be smooth and approach zero near ∂B at a certain rate. In our analysis we will employ, essentially, ideas drawn from the theory of partial differential operators, in contrast to the integral operator methods of [4]; thus this paper complements the results of [4].

Of particular interest is the study of the so-called interior transmission problem, which plays a fundamental role in solving the inverse scattering problem in [3]. This is treated in §§ 3 and 4 and offers an alternative to the treatment in [4]. Also in § 4, we consider the question of approximating the solutions of the interior transmission problem by Herglotz wave functions. This is required in the discussion of inverse scattering in [3]. In §§ 2–4 we consider the case of a nonabsorbing medium. The modifications required to deal with the case of an absorbing medium are described in § 5.

2. Wave propagation and far field patterns. Consider acoustic scattering of a plane harmonic incident wave

$$u^i(x, t) = \exp(ikx \cdot \hat{\alpha} - i\omega t), \tag{2.1}$$

* Received by the editors January 31, 1990; accepted for publication (in revised form) December 7, 1990.
† Department of Mathematics, Heriot-Watt University, Riccarton, Edinburgh EH14 4AS, Scotland.
‡ Department of Mathematics and Computer Science, University of Dundee, Dundee DD1 4HN, Scotland.

where $\omega > 0$ is the frequency, $k = \omega/c_0$ is the wavenumber, $c_0 > 0$ is the speed of sound outside the medium, and the unit vector $\hat{\alpha}$ is the direction of propagation of the wave. Let $c(x)$ be the local sound speed at any point $x \in \mathbb{R}^3$, and let $n(x) = (c_0/c(x))^2$, $m(x) = 1 - n(x)$. Letting

$$B = \{x \in \mathbb{R}^3 : n(x) \neq 1\},$$

we assume that the set B is an open, bounded, simply connected set with C^4 boundary ∂B and suppose that the function n is measurable and satisfies

$$\text{ess} \cdot \sup \{n(x) : x \in B\} < \infty, \qquad \text{ess} \cdot \inf \{|m(x)| : x \in B\} > 0 \tag{2.2}$$

(note that we are allowing m to change sign in B, but (2.2) implies that it must do so discontinuously). Without loss of generality, we will assume that the origin $x = 0$ belongs to B. If we factor out the time dependence $\exp(-i\omega t)$, then the velocity potential u of the total field belongs to $H_2(B)$ (the Sobolev space of order 2 on B; see [1]) and is a solution of the problem

$$\Delta u + k^2 n u = 0 \quad \text{in } \mathbb{R}^3, \tag{2.3}$$

$$u(x) \equiv \exp(ikx \cdot \hat{\alpha}) + u^s(x), \tag{2.4}$$

$$\lim_{r \to \infty} \left(\frac{\partial u^s}{\partial r} - iku^s \right) = 0, \tag{2.5}$$

where $u^s(x)$ denotes the scattered field, $|x| = r$, and the Sommerfeld radiation condition (2.5) is assumed to hold uniformly in $\hat{x} = x/|x| \in S^2$.

As in [4] it is easy to show that as $r \to \infty$ the scattered field $u^s(x)$ has the asymptotic behavior

$$u^s(x) = \frac{\exp(ikr)}{r} F(\hat{x}; k, \hat{\alpha}) + O(r^{-2}),$$

where the function $F(\hat{x}; k, \hat{\alpha})$ is the *far field pattern* corresponding to the incident plane wave (2.1). The following result is similar to Lemma 2 of [4].

Let $\{\hat{\alpha}_n : n = 1, \cdots, \infty\}$ be a countable dense set of vectors on the unit sphere S^2, and for each fixed k define the class $\mathbf{F}$ of far field patterns by

$$\mathbf{F} = \text{span} \{F(\hat{x}; k, \hat{\alpha}) : n = 1, 2, \cdots\}.$$

LEMMA 2.1 [4]. *The orthogonal complement of* $\mathbf{F}$ *in* $L_2(S^2)$ *consists of those functions* $g \in L_2(S^2)$ *for which there exists* $w \in H_2(B)$ *and* v *defined by*

$$v(x) = \int_{S^2} g(\hat{y}) \exp(ikx \cdot \hat{y}) \, ds(y) \tag{2.6}$$

such that the pair $\{v, w\}$ *is a solution to*

$$\Delta v + k^2 v = 0, \tag{2.7}$$

$$\Delta w + k^2 n w = 0, \tag{2.8}$$

$$v - w|_{\partial B} = \frac{\partial}{\partial \nu}(v - w)|_{\partial B} = 0, \tag{2.9}$$

where (2.7) *and* (2.8) *are to be regarded as holding in* $L_2(B)$; $\partial/\partial \nu$ *denotes differentiation along the exterior normal to* ∂B; *the notation* $\phi|_{\partial B}$ *denotes the trace of a function* ϕ *on the boundary* ∂B *in the* L_2 *sense* (*see* [1]); *equation* (2.9) *holds on* ∂B *in the* L_2 *sense. Similar interpretations apply to the equations below except where stated otherwise.*

The functions v defined by (2.6) are called *Herglotz wave functions* with *Herglotz kernel g*. The boundary value problem (2.7)-(2.9) is the *homogeneous interior transmission problem*, studied in [3], [4], and others, which, together with the inhomogeneous interior transmission problem considered later, plays a fundamental role in solving the inverse scattering problem of determining the speed of sound in an inhomogeneous medium (see [3]). For this purpose it is important to prove that the set **F** is complete in $L_2(S^2)$. It follows from Lemma 2.1 and the theory of Herglotz wave functions (see [5]) that this is so, except for those values of k for which (2.7)-(2.9) possesses a nontrivial solution $\{v, w\}$. In the next section we will show that such values form a discrete set in $\mathbb{C}$, and hence, for almost all k, **F** is complete.

3. The homogeneous interior transmission problem. The space $H_2^{\text{loc}}(B)$ is defined to be the set of measurable functions ϕ on B that have the property that $\phi \in H_2(D)$ for all open subsets D such that $\bar{D} \subset B$. We will say that the pair of functions $\{v, w\}$ is a strong solution of the homogeneous interior transmission problem if $v, w \in H_2^{\text{loc}}(B) \cap L_2(B)$, $v - w \in H_2(B)$, and equations (2.7)-(2.9) are satisfied for some $k \in \mathbb{C}$. If a nontrivial strong solution of the homogeneous interior transmission problem exists for some $k \in \mathbb{C}$, then k is said to be a transmission eigenvalue.

Suppose that k is a transmission eigenvalue and let

$$z = w - v \in H_2(B). \tag{3.1}$$

It follows from (2.9) that $z \in H_2^0(B)$, where $H_2^0(B) \subset H_2(B)$ is the set of functions $\phi \in H_2(B)$ whose derivatives up to order 1 vanish on ∂B. Also, from (2.7) and (2.8) we have

$$(\Delta + k^2)z = k^2 m w, \tag{3.2}$$

where $m(x) = 1 - n(x)$, $x \in B$. It follows from (2.2) that the function $x \to 1/m(x)$, which we denote by m^{-1}, is essentially bounded on B, so from (3.2) we have

$$m^{-1}(\Delta + k^2)z = k^2 w.$$

Now, by definition $w \in H_2^{\text{loc}}(B)$, and so we may apply $(\Delta + k^2 n)$ to both side of this equation to yield, by (2.8),

$$(\Delta + k^2 n)m^{-1}(\Delta + k^2)z = 0. \tag{3.3}$$

Thus, in order for k to be a transmission eigenvalue there must be a nontrivial function $z \in H_2^0(B)$ that satisfies (3.3).

Since we have only required that $z \in H_2^0(B)$ and (3.3) is a fourth-order equation, we will introduce a weak formulation of this equation. For any complex number k we define the sesquilinear form F_k on $H_2^0(B)$ by

$$\begin{aligned} F_k(\phi, \psi) &= \int_B m^{-1}(\Delta + k^2)\phi(\Delta + \bar{k}^2 n)\bar{\psi}\, dx \\ &= (m^{-1}(\Delta + k^2)\phi, (\Delta + \bar{k}^2 n)\psi), \qquad \phi, \psi \in H_2^0(B) \end{aligned} \tag{3.4}$$

(we let $(\cdot, \cdot)$ and $\|\cdot\|$ denote the standard $L_2(B)$ inner product and norm). Clearly, the form F_k is bounded on the space $H_2^0(B)$.

LEMMA 3.1. *A nonzero point $k \in \mathbb{C}$ is a transmission eigenvalue if and only if there exists a nonzero function $z \in H_2^0(B)$ such that*

$$F_k(z, \psi) = 0, \quad \text{all } \psi \in H_2^0(B). \tag{3.5}$$

Proof. It follows immediately from (3.3), using integration by parts, that if k is a transmission eigenvalue then the function z defined in (3.1) satisfies (3.5).

Now suppose that $z \in H_2^0(B)$ is nonzero and satisfies (3.5). Since $z \in H_2^0(B)$, we can define the functions v, $w \in L_2(B)$ by

$$w = k^{-2} m^{-1} (\Delta + k^2) z \tag{3.6}$$

and

$$v = w - z. \tag{3.7}$$

By definition we have

$$(w, (\Delta + \bar{k}^2 n)\psi) = k^{-2}(m^{-1}(\Delta + k^2)z, (\Delta + \bar{k}^2 n)\psi) = k^{-2} F_k(z, \psi) = 0, \quad \text{all } \psi \in H_2^0(B),$$

and by standard interior regularity results for elliptic operators (see [1, Thm. 6.3]) this shows that $w \in H_2^{\text{loc}}(B)$ and satisfies (2.8). Since $z \in H_2^0(B)$, we also have $v \in H_2^{\text{loc}}(B)$. Now,

$$(\Delta + k^2)v = (\Delta + k^2)w - (\Delta + k^2)z = k^2 mw - k^2 mw = 0,$$

using (3.6) and (2.8). Thus v satisfies (2.7). These results show that the nontrivial pair of functions $\{v, w\}$ is a strong solution of the homogeneous interior transmission problem, and hence k is a transmission eigenvalue. This completes the proof of the lemma.

We now define another form $\langle \cdot, \cdot \rangle$ on $H_2^0(B)$ by

$$\langle \phi, \psi \rangle = (\Delta\phi, \Delta\psi), \qquad \phi, \psi \in H_2^0(B).$$

Also, we let $(\cdot, \cdot)_j$ and $\|\cdot\|_j$ denote the inner product and norm in $H_j(B)$ for any integer $j \geqq 0$.

LEMMA 3.2. *There exists $c_1 > 0$ such that for any $\phi \in H_2^0(B)$,*

$$\langle \phi, \phi \rangle \geqq c_1 \|\phi\|_2^2. \tag{3.8}$$

Proof. The inequality follows immediately from Lemma 7.7 in [1].

For any $k \in \mathbb{C}$, we can decompose the form F_k into a sum of forms $F^0 + F_k^1 + F_k^2$, where

$$F^0(\phi, \psi) = (m^{-1}\Delta\phi, \Delta\psi), \quad F_k^1(\phi, \psi) = (m^{-1}(\Delta + k^2)\phi, \bar{k}^2 n\psi),$$
$$F_k^2(\phi, \psi) = (k^2 m^{-1}\phi, \Delta\psi),$$

where $\phi, \psi \in H_2^0(B)$. Lemma 3.2 and the Lax-Milgram theorem (see [7, p. 344]) allow us to define bounded linear operators S_k, S^0, S_k^1, S_k^2, on $H_2^0(B)$ by means of the following identities, which are supposed to hold for all $\phi, \psi \in H_2^0(B)$:

$$F_k(\phi, \psi) = \langle S_k\phi, \psi \rangle, \quad F^0(\phi, \psi) = \langle S^0\phi, \psi \rangle, \quad F_k^1(\phi, \psi) = \langle S_k^1\phi, \psi \rangle,$$
$$F_k^2(\phi, \psi) = \langle S_k^2\phi, \psi \rangle.$$

Clearly,

$$S_k = S^0 + S_k^1 + S_k^2. \tag{3.9}$$

LEMMA 3.3. *For any $k \in \mathbb{C}$, the operators S_k^1, S_k^2 are compact.*

Proof. Let $\phi \in H_2^0(B)$. Then by the definition of $\langle \cdot, \cdot \rangle$, F_k^1 and S_k^1,

$$(\Delta S_k^1\phi, \Delta\psi) = \bar{k}^2(nm^{-1}(\Delta + k^2)\phi, \psi) = \bar{k}^2(\theta, \psi), \quad \text{all } \psi \in H_2^0(B),$$

where $\theta = nm^{-1}(\Delta + k^2)\phi \in L_2(B)$. By a standard regularity result for elliptic operators (see [1, Thm. 9.8]), this implies that $S_k^1\phi \in H_4(B) \cap H_2^0(B)$ and

$$\|S_k^1\phi\|_4 \leqq c_2(\|\phi\|_0 + \|S_k^1\phi\|_0) \leqq c_3\|\phi\|_2.$$

Since this holds for all $\phi \in H_2^0(B)$ and the injection operator from $H_4(B)$ to $H_2(B)$ is compact, this shows that the operator S_k^1 is compact. To show that S_k^2 is compact we first choose a sequence of functions m_j^*, $j=1,2,\cdots$, such that $m_j^* \in C^\infty(B)$, and $m_j^* \to m^{-1}$ in $L_2(B)$ (this is possible by Theorem 1.7 of [1]), and we define the sequence of operators $S_{k,j}^2 : H_2^0(B) \to H_2^0(B)$ by the relations

$$\langle S_{k,j}^2 \phi, \psi \rangle = (k^2 m_j^* \phi, \Delta\psi), \qquad \phi, \psi \in H_2^0(B).$$

As before, for any $\phi \in H_2^0(B)$, we have

$$(\Delta S_{k,j}^2 \phi, \Delta\psi) = (k^2 m_j^* \phi, \Delta\psi), \quad \text{all } \psi \in H_2^0(B),$$

and so

$$\Delta S_{k,j}^2 \phi = k^2 m_j^* \phi \tag{3.10}$$

in $L_2(B)$. Now, the right-hand side of (3.10) belongs to $H_2^0(B)$, so by the regularity properties of the Dirichlet problem for the Laplace operator $S_{k,j}^2 \phi \in H_4(B) \cap H_2^0(B)$ and

$$\|S_{k,j}^2 \phi\|_4 \leqq c_4 \|m_j^* \phi\|_2 \leqq c_{4,j} \|\phi\|_2,$$

since $m_j^* \in C^\infty(\bar{B})$. Hence the operators $S_{k,j}^2$ are compact. Also, since $m^{-1}\phi \in L_2(B)$, we have, by a similar argument,

$$\|(S_k^2 - S_{k,j}^2)\phi\|_2 \leqq c_5 \|(m^{-1} - m_j^*)\phi\| \leqq c_5 \|m^{-1} - m_j^*\| \, |\phi| \leqq c_6 \|m^{-1} - m_j^*\| \, \|\phi\|_2,$$

by Sobolev's inequality (where $|\phi|$ is the sup norm of ϕ on B). Hence

$$\|S_k^2 - S_{k,j}^2\|_2 \leqq c_6 \|m^{-1} - m_j^*\|$$

(where the norm on the left of this inequality is the uniform operator norm on $H_2^0(B)$), and so S_k^2 is the uniform limit of a sequence of compact operators and so is compact. This completes the proof of the lemma.

We will now impose the following assumption which is, essentially, a condition on the function m.

Assumption. The operator S^0 is nonsingular.

This assumption does not seem to be unduly restrictive. We conjecture that it holds for "generic" functions m on B. If $m(x) \geqq c_7 > 0$ (or $m(x) \leqq c_7 < 0$), then the assumption is certainly true (this follows from (3.8) and [7, p. 344]).

LEMMA 3.4. *A nonzero point $k \in \mathbb{C}$ is a transmission eigenvalue if and only if the operator S_k is singular.*

Proof. It follows from Lemma 3.1 that any nonzero $k \in \mathbb{C}$ is a transmission eigenvalue if and only if

$$N(S_k) \neq 0. \tag{3.11}$$

However, by Lemma 3.3 and the Fredholm alternative for compact operators it can easily be shown that this holds if and only if S_k is singular.

We can now prove the main result of this section.

THEOREM 3.5. *The set of transmission eigenvalues is discrete.*

Proof. It is clear from the definitions of the forms F_k^1, F_k^2, $k \in \mathbb{C}$ that these are bounded-holomorphic families of forms (see [6, § VII-4.1, p. 395]). Consequently, the families of operators S_k^1, S_k^2, are bounded holomorphic (with respect to the uniform operator norm on $H_2^0(B)$). Hence it follows easily from [6, Thm. VII-1.9, p. 370] that (3.11) holds for either a discrete set of points $k \in \mathbb{C}$, or for all $k \in \mathbb{C}$. However, S^0 is nonsingular, so the latter alternative cannnot hold, which proves the theorem.

4. The inverse scattering problem and Herglotz wave functions. The inverse scattering problem is to determine the index of refraction $n(x)$ from the far field patterns $F(\hat{x}; k, \hat{\alpha})$ for a range of k, $\hat{\alpha} \in S^2$. This problem has been discussed by Colton and Monk [3] using the following approach. First of all, for a fixed k, we seek a function $g \in L_2(S^2)$ such that

$$\int_{S^2} F(\hat{x}; k, \hat{\alpha}) g(\hat{x})\, ds(\hat{x}) = 1 \tag{4.1}$$

for all $\hat{\alpha} \in S^2$. If we define the Herglotz wave function v by (2.6) then it can be shown that (4.1) holds if and only if there is a function w such that the pair $\{v, w\}$ is a solution of the following inhomogeneous interior transmission problem (the proof is similar to the proof of Lemma 2 in [4]; see also [3, § 3]). The pair of functions $\{v, w\}$ is said to be a strong solution of the inhomogeneous interior transmission problem if $v, w \in H_2^{\text{loc}}(B) \cap L_2(B)$, $v - w \in H_2(B)$, and

$$\Delta v + k^2 v = 0, \tag{4.2}$$

$$\Delta w + k^2 n w = 0, \tag{4.3}$$

$$v - w|_{\partial B} = \frac{1}{r} e^{-ikr}, \tag{4.4}$$

$$\frac{\partial}{\partial \nu}(v - w)|_{\partial B} = \frac{\partial}{\partial \nu}\left(\frac{1}{r} e^{-ikr}\right) \tag{4.5}$$

for some $k \in \mathbb{C}$. The method of Colton and Monk is to determine n from (4.1) and the interior transmission problem (see [3]). To apply their methods it is necessary to show that the homogeneous interior transmission problem has a unique solution and, in addition, it is necessary to show that the function v thus found can be approximated by Herglotz wave functions. The proofs of these results as given in [3] are not valid here, so we now proceed to prove them for the present situation.

THEOREM 4.1. *If $k \neq 0$ is not a transmission eigenvalue, then there exists a unique strong solution to the inhomogeneous interior transmission problem.*

Proof. First, note that if there were two distinct solutions to the inhomogeneous problem then their difference would satisfy the homogeneous problem, and hence k would be a transmission eigenvalue. Thus we have uniqueness whenever k is not a transmission eigenvalue. It remains to prove existence.

Choose a function $g \in C^\infty(\mathbb{R}^3)$ such that

$$g(x) = \frac{1}{r} e^{-ikr}$$

in some open neighbourhood of ∂B, which does not contain the origin, and $g(x) = 0$ elsewhere. We now define a function $f_g : H_2^0(B) \to \mathbb{C}$ by

$$f_g(\psi) = (m^{-1}(\Delta + k^2)g, (\Delta + k^2)\psi), \qquad \psi \in H_2^0(B). \tag{4.6}$$

Clearly, f_g is bounded and antilinear, so that by the Lax–Milgram theorem there exists an element $\theta_g \in H_2^0(B)$ such that

$$f_g(\psi) = \langle \theta_g, \psi \rangle, \qquad \psi \in H_2^0(B).$$

Now, by Lemma 3.4, S_k^{-1} exists and is bounded since k is not a transmission eigenvalue. Therefore, we may define $\xi_g = S_k^{-1}\theta_g \in H_2^0(B)$ and we have, by the above definitions,

$$F_k(\xi_g, \psi) = \langle \theta_g, \psi \rangle = F_k(g, \psi), \qquad \psi \in H_2^0(B)$$

(where $F_k(g, \psi)$ is defined by (4.6), even though $g \notin H_2^0(B)$). Thus, if we put $z = g - \xi_g$ we can obtain a strong solution of the inhomogeneous interior transmission problem from z in the same way that we obtained a strong solution of the homogeneous problem in the proof of Lemma 3.1.

Now let $H \subset L_2(B)$ be the linear span in $L_2(B)$ of the set of functions

$$x \to j_l(k|x|)\, Y_l(x/|x|), \quad x \in B, \quad l = 0, 1, \cdots, \quad -l \leqq m \leqq l,$$

where j_l is a spherical Bessel function and Y_l is a spherical harmonic. Let $\bar{H}$ be the closure of H in $L_2(B)$ and let $\bar{H}^{\perp}$ denote the orthogonal complement of $\bar{H}$ in $L_2(B)$.

THEOREM 4.2. *Suppose that* $v \in H_2^{\mathrm{loc}}(B) \cap L_2(B)$ *satisfies equation* (2.7). *Then* $v \in \bar{H}$.

Proof. If $h \in H$ then

$$(h, (\Delta + k^2)\phi) = 0, \quad \text{all } \phi \in C_0^{\infty}(B), \tag{4.7}$$

and, by continuity, (4.7) holds for all $h \in \bar{H}$. Putting

$$v = v_1 + v_2, \quad v_1 \in \bar{H}, \quad v_2 \in \bar{H}^{\perp},$$

then v_1 must satisfy (4.7) and, from our hypothesis that v satisfies (2.7), v also satisfies (4.7), and so v_2 must satisfy (4.7). Thus, by continuity, we have

$$(v_2, (\Delta + k^2)\phi) = 0, \quad \text{all } \phi \in H_2^0(B). \tag{4.8}$$

Now, let

$$\Phi(x, y) = \frac{\exp(ik|x-y|)}{4\pi|x-y|}, \qquad x, y \in \mathbb{R}^3,$$

and define

$$z(x) = \int_B v_2(y)\Phi(x, y)\, dy, \qquad x \in \mathbb{R}^3.$$

Then $z \in H_2(\mathbb{R}^3)$ and, since $v_2 \in \bar{H}^{\perp}$, it follows from the addition formula for Bessel functions (see, [2, eqn. (3.60), p. 94]) that $z(x) = 0$ when $x \notin B$. Thus the restriction of z to B, which we denote by z_B, belongs to $H_2^0(B)$ and

$$(\Delta + k^2) z_B = v_2. \tag{4.9}$$

Now, by (4.8),

$$\|(\Delta + k^2) z_B\|^2 = ((\Delta + k^2) z_B, (\Delta + k^2) z_B) = (v_2, (\Delta + k^2) z_B) = 0,$$

and hence z satisfies the Helmholtz equation

$$(\Delta + k^2) z = 0 \quad \text{in } \mathbb{R}^3. \tag{4.10}$$

It now follows from [1] that $z \in C^2(\mathbb{R}^3)$ and by [2] z is analytic. Hence, since z is identically zero outside B, it must be zero everywhere, and so by (4.9), $v_2 = 0$. This proves the theorem.

5. Absorbent media. In this section we will briefly describe the modifications required in the preceding analysis to deal with the case where the medium in B is absorbing. Let $a(x) \geqq 0$ denote the absorption coefficient at the point $x \in \mathbb{R}^3$ ($a(x) = 0$ for x outside B), and let $\kappa^2(x) = k^2 + ika(x)/c_0$. Then in the fundamental equation (2.3) describing the wave propagation, the term k^2 should be replaced by κ^2. Similarly, in the interior transmission problem the term k^2 in (2.8) should be replaced by κ^2. Now, we again define the function z on B by (3.1), but (3.2) now becomes

$$(\Delta + k^2) z = k^2 m w + ikanw. \tag{5.1}$$

Putting $M(x) = km(x) + ia(x)n(x)$, we obtain the following analogue of the basic equation (3.3):

$$(\Delta + \kappa^2 n)M^{-1}(\Delta + k^2)z = 0. \tag{5.2}$$

From this point onwards the analysis follows the above lines, with κ^2 replacing k^2 at appropriate points and M replacing m. One slight difference is that S^0 now depends on k (via the function M) and it is necessary to assume that S^0 is nonsingular for all k. A sufficient condition for this to be true is that $a(x) \geqq c_8 > 0$ for all $x \in B$ (to see this we observe that if this condition holds then, for each k, the numerical range of S^0 is bounded away from zero, and so the result follows from Theorem VI.3.1 of [7]).

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary-Value Problems*, Van Nostrand, Princeton, NJ, 1965.

[2] D. COLTON AND R. KRESS, *Integral Equation Methods in Scattering Theory*, Interscience, New York, 1983.

[3] D. COLTON AND P. MONK, *The inverse scattering problem for time-harmonic acoustic waves in an inhomogeneous medium*, Quart. J. Mech. Appl. Math., 40 (1987), pp. 189–212.

[4] D. COLTON, A. KIRSCH, AND L. PAIVÄRINTA, *Far field patterns for acoustic waves in an inhomogeneous medium*, SIAM J. Math. Anal., 20 (1989), pp. 1472–1483.

[5] P. HARTMAN AND C. WILCOX, *On solutions of the Helmholtz equation in exterior domains*, Math. Z., 75 (1961), pp. 229–255.

[6] T. KATO, *Perturbation Theory for Linear Operators*, Second Ed., Springer-Verlag, Berlin, New York, 1976.

[7] A. E. TAYLOR AND D. C. LAY, *Introduction to Functional Analysis*, Second Ed., John Wiley, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1763-1775, November 1991

015

ON THE INVERSE OF THE DISCRETE TWO-DIMENSIONAL WAVE OPERATOR AT CFL*

ROBERT GLASSEY†, ERNST HORST‡, ANDREW LENARD†, AND JACK SCHAEFFER§

Abstract. The initial value problem for the linear nonhomogeneous wave equation in two space dimensions is discretized in the usual way via centered second-order differences, with the timestep size chosen on the CFL "boundary:" $\Delta x = \Delta y = h$, $\Delta t = h/\sqrt{2}$. The solution of this discrete problem is explicitly given as a functional of the data.

Key words. finite differences, fundamental solution, CFL, Courant-Friedrichs-Levy

AMS(MOS) subject classifications. 35L05, 65M99

1. Introduction. In the analysis of convergence of difference schemes for hyperbolic partial differential equations, energy estimates naturally arise. For nonlinear problems in more than one space dimension, these often do not suffice and must be supplemented with additional inequalities. Even in the continuous case, higher-order estimates usually cannot be obtained directly; a uniform (L^∞) estimate is needed first. The problem studied in this paper arose from an attempt to prove convergence of a particle-finite difference method for the two-dimensional Vlasov-Maxwell equations from plasma physics. There we desire an L^∞ estimate on a solution to a discetized inhomogeneous wave equation, as is suggested by the continuous theory [3]. Of course, the standard energy estimate for the wave equation does not provide an L^∞ bound in two dimensions. Thus a *representation* for the discrete solution as an explicit functional of the data is an important first step.

Our results concern discretization of the initial value problem for the wave equation

$$\begin{aligned} &\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(t, x, y) \qquad (0 < t < \infty,\ x, y \in \mathbb{R}), \\ &u(0, x, y) = \phi(x, y), \qquad u_t(0, x, y) = \psi(x, y), \end{aligned} \tag{1}$$

where we will assume that the data ϕ, ψ, and f are smooth and have compact support in x, y. We choose a stepsize

$$\Delta x = \Delta y = h > 0$$

and then choose the maximum timestep Δt allowed by the classical Courant-Friedrichs-Levy (CFL) condition

$$\Delta t = \frac{h}{\sqrt{2}}. \tag{2}$$

Approximating (1) by standard second-order differences and writing

$$\begin{aligned} &x_k = kh, \quad y_j = jh, \quad t^n = n\Delta t \quad (k, j \in \mathbb{Z},\ n \in \mathbb{N}), \\ &u^n_{kj} \cong u(t^n, x_k, y_j) \end{aligned}$$

* Received by the editors May 15, 1989; accepted for publication November 23, 1990. This research was supported in part by National Science Foundation grants DMS8721721 and DMS8801738.

† Department of Mathematics, Indiana University, Bloomington, Indiana 47405.

‡ Fachbereich Mathematik und Informatik der GHS Paderborn, Warburgerstrasse 100, D 4790 Paderborn, Germany.

§ Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

we get

$$\frac{u_{kj}^{n+1}-2u_{kj}^{n}+u_{kj}^{n-1}}{\Delta t^2}-\left[\frac{u_{k+1,j}^{n}-2u_{kj}^{n}+u_{k-1,j}^{n}}{h^2}\right]-\left[\frac{u_{k,j+1}^{n}-2u_{kj}^{n}+u_{k,j-1}^{n}}{h^2}\right]=f_{kj}^{n} \tag{3}$$

or, in view of (2),

$$u_{kj}^{n+1}=-u_{kj}^{n-1}+\tfrac{1}{2}[u_{k+1,j}^{n}+u_{k-1,j}^{n}+u_{k,j+1}^{n}+u_{k,j-1}^{n}]+\Delta t^2 f_{kj}^{n}. \tag{4}$$

As is standard for this second-order scheme, we specify as initial conditions

$$\begin{aligned} u_{kj}^{0}&=\phi_{kj}\equiv\phi(x_k, y_j),\\ u_{kj}^{1}&=\phi_{kj}+\Delta t\psi_{kj}. \end{aligned} \tag{5}$$

For the purpose of this introduction, we will describe our result for the case $\phi\equiv 0$, $\psi\equiv 0$.

THEOREM 1. *Let $\phi=\psi=0$. Then the solution of* (4) *is given by*

$$u_{kj}^{n+1}=\Delta t^2\sum_{l=0}^{n-1}2^{-l}\sum_{p=0}^{[l/2]}\sum_{\substack{\alpha,\beta\\ |\alpha-k|+|\beta-j|=l-2p}} S_{p,k-\alpha}^{l} f_{\alpha\beta}^{n-l}, \tag{6}$$

where the kernel S arises in the form

$$S_{p,k}^{l}=\sum_{m=0}^{p}(-4)^m\binom{l-m}{m}\binom{l-2m}{p-m}\binom{l-2m}{|k|+p-m}. \tag{7}$$

We know that the fundamental solution of the wave equation is positive inside the light cone $t=|x|$, yet (7) does not "look" positive. Moreover, the discretization need not a priori preserve positivity. However, *the kernel $S_{p,k}^{l}$ is nonnegative*, as we see in Theorem 2.

THEOREM 2. *Denote by $P_n^{(\alpha,\beta)}(x)$ the* Jacobi polynomial *of degree n in x, with parameters α, β (cf.* [5]). *Then for $k\geqq 0$,*

$$S_{p,k}^{l}=\frac{\binom{l-p}{k}}{\binom{k+p}{k}}\cdot 4^p\cdot[P_p^{(k,l-2p-k)}(0)]^2. \tag{8}$$

The relations (6) and (8) then resolve the Cauchy problem for (4) with zero data.

In order to "recognize" the representation (6), we use the classical Duhamel formula to write the solution to (1) with $\phi=\psi=0$ ($x\in\mathbb{R}^2$):

$$\begin{aligned} 2\pi u(t,x)&=\int_0^t d\tau\int_{|y-x|<t-\tau}\frac{f(\tau,y)\,dy}{\sqrt{(t-\tau)^2-|y-x|^2}}\\ &=\int_0^t d\tau\int_0^{t-\tau}\frac{dr}{\sqrt{(t-\tau)^2-r^2}}\int_{|y-x|=r} f(\tau,y)\,ds_y. \end{aligned} \tag{9}$$

Replacing r by ξ through $r=t-\tau-2\xi$, we obtain

$$\begin{aligned} 2\pi u(t,x)&=\int_0^t d\tau\int_0^{(1/2)(t-\tau)}\frac{d\xi}{\sqrt{\xi}\sqrt{t-\tau-\xi}}\int_{|y-x|=t-2\xi-\tau} f(\tau,y)\,ds_y\\ &=\int_0^t d\tau\int_0^{(1/2)\tau}\frac{d\xi}{\sqrt{\xi}\sqrt{\tau-\xi}}\int_{|y-x|=\tau-2\xi} f(t-\tau,y)\,ds_y, \end{aligned} \tag{10}$$

which is the continuous analogue of (6).

2. Construction of the representation. The representation (6) will follow from the discrete Fourier transformation applied to (4). Since (4) is an explicit scheme, uniqueness of a solution follows immediately.

Consider (4) with zero initial values and $\lambda \equiv \Delta t/h = 1/\sqrt{2}$.

Given a sequence $\{u_{kj}\}_{k,j=-\infty}^{\infty}$ that is summable, we define its *Fourier transform* by

$$\hat{u}(\theta_1, \theta_2) = \sum_{k,j} e^{-i(k\theta_1 + j\theta_2)} u_{kj} = \mathcal{F}u(\theta) \qquad (\theta = (\theta_1, \theta_2)). \tag{11}$$

Given a *translation* τ_j $(j = (j_1, j_2))$ defined by

$$(\tau_j u)_k = u_{k+j} \qquad (k = (k_1, k_2)), \tag{12}$$

it is well known that

$$(\tau_j u)\hat{}\,(\theta) = e^{i(j\cdot\theta)}\hat{u}(\theta). \tag{13}$$

The scheme (4) can be written as

$$u^{n+1} = -u^{n-1} + \tfrac{1}{2}[\tau_{1,0} + \tau_{-1,0} + \tau_{0,1} + \tau_{0,-1}]u^n + \Delta t^2 \cdot f^n \tag{14}$$

and, after Fourier transformation, it becomes

$$\hat{u}^{n+1} = -\hat{u}^{n-1} + \tfrac{1}{2}[2\cos\theta_1 + 2\cos\theta_2]\hat{u}^n + \Delta t^2 \cdot \hat{f}^n \tag{15}$$

or

$$\hat{u}^{n+1} = -\hat{u}^{n-1} + \beta\hat{u}^n + \Delta t^2 \cdot \hat{f}^n, \tag{16}$$

where

$$\beta = \cos\theta_1 + \cos\theta_2. \tag{17}$$

Since $|\beta| \leqq 2$, we can define an angle ψ by

$$\cos\psi = \frac{\beta}{2}. \tag{18}$$

When $\hat{f}^n \equiv 0$, (16) is the Chebyshev difference equation that possesses the linearly independent solutions

$$T_n(\cos\psi) = \cos n\psi, \qquad U_n(\cos\psi) = \frac{\sin(n+1)\psi}{\sin\psi}. \tag{19}$$

Here $T_n(x)$ denotes the standard Chebyshev polynomial, and, with $x = \cos\psi = \beta/2$,

$$U_n(x) = \sum_{m=0}^{[n/2]} (-1)^m \binom{n-m}{m} (2x)^{n-2m} \tag{20}$$

(cf. [5, p. 257]). To represent the solution of (16), we use the classical Duhamel formula ([4, p. 409])

$$\hat{u}^{n+1}(\theta) = \sum_{l=0}^{n-1} \Delta t^2 \cdot \hat{f}^{l+1} \cdot \frac{\sin(n-l)\psi}{\sin\psi}. \tag{21}$$

In view of (21) and the convolution theorem, it will suffice to find the Fourier inverse of $\sin n\psi/\sin\psi$. For this purpose we recall that

$$2x = \beta = \cos\theta_1 + \cos\theta_2,$$

so that (20) is

$$\frac{\sin(n+1)\psi}{\sin\psi} = \sum_{m=0}^{[n/2]} (-1)^m \binom{n-m}{m} (\cos\theta_1 + \cos\theta_2)^{n-2m}.$$

Now consider the sequence

$$\chi_{kj} = \delta_{|k|+|j|,1} = \begin{cases} 1 & \text{if } |k|+|j|=1, \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

defined for $k, j \in \mathbb{Z}$. Then by an elementary computation,

$$\begin{aligned} \hat{\chi}(\theta) &= \sum_{\substack{k,j \\ (k,j)=(\pm 1,0) \\ (k,j)=(0,\pm 1)}} e^{-i(k\theta_1+j\theta_2)}\chi_{kj} \\ &= e^{-i\theta_1} + e^{i\theta_1} + e^{-i\theta_2} + e^{i\theta_2} = 2(\cos\theta_1 + \cos\theta_2). \end{aligned} \tag{23}$$

It follows that

$$\frac{\sin(n+1)\psi}{\sin\psi} = \sum_{m=0}^{[n/2]} (-1)^m \binom{n-m}{m} 2^{2m-n} (\hat{\chi}(\theta))^{n-2m}. \tag{24}$$

Now define

$$\chi^{*(N)} \equiv \underbrace{\chi * \chi * \cdots * \chi}_{N \text{ times}} \qquad (\chi^{*(1)} \equiv \chi), \tag{25}$$

so that

$$\mathcal{F}(\chi^{*(N)})(\theta) = (\hat{\chi}(\theta))^N. \tag{26}$$

Applying $\mathcal{F}^{-1}$ to (24) we then have

$$\mathcal{F}^{-1}\left(\frac{\sin(n+1)\psi}{\sin\psi}\right)_{kj} = \sum_{m=0}^{[n/2]} (-1)^m \binom{n-m}{m} 2^{2m-n} (\chi^{*(n-2m)})_{kj}. \tag{27}$$

Thus our representation will follow once we have computed the N-fold convolution $\chi^{*(N)}$.

LEMMA 1. *For $k, j \in \mathbb{Z}$ and $n \geq 2$,*

$$(\chi^{*(n)})_{kj} = \sum_{p=0}^{[n/2]} \binom{n}{p} \binom{n}{|k|+p} \delta_{|k|+|j|,n-2p}.$$

Proof. By definition,

$$\begin{aligned} (\chi^{*(n)})_{kj} = (\chi * \chi^{*(n-1)})_{kj} &= \sum_{\substack{m,l \\ (m,l)=(\pm 1,0) \\ (m,l)=(0,\pm 1)}} \chi_{ml}\chi^{*(n-1)}_{k-m,j-l} \\ &= \chi^{*(n-1)}_{k-1,j} + \chi^{*(n-1)}_{k+1,j} + \chi^{*(n-1)}_{k,j-1} + \chi^{*(n-1)}_{k,j+1}. \end{aligned}$$

(From this it follows that $\chi^{*(n)}$ is symmetric in each quadrant.) Thus for $n = 2$,

$$(\chi^{*(2)})_{kj} = \delta_{|k-1|+|j|,1} + \delta_{|k+1|+|j|,1} + \delta_{|k|+|j-1|,1} + \delta_{|k|+|j+1|,1}. \tag{28}$$

Now consider, for $N \in \mathbb{N}$, the expression

$$\begin{aligned} \delta_{|k-1|+|j|,N} + \delta_{|k+1|+|j|,N} &= \begin{cases} \delta_{|k|+|j|,N+1} + \delta_{|k|+|j|,N-1} & \text{if } |k| \geq 1, \\ 2\delta_{|k|+|j|,N-1} & \text{if } k = 0, \end{cases} \\ &= 2\delta_{k0}\delta_{|k|+|j|,N-1} + (1-\delta_{k0})(\delta_{|k|+|j|,N-1} + \delta_{|k|+|j|,N+1}) \\ &= (1+\delta_{k0})\delta_{|k|+|j|,N-1} + (1-\delta_{k0})\delta_{|k|+|j|,N+1} \end{aligned} \tag{29}$$

for $N \geq 1$.

Using (29) with $N=1$ in (28) twice, we obtain

$$\begin{aligned}(\chi^{*(2)})_{kj} &= (1+\delta_{k0})\delta_{|k|+|j|,0}+(1-\delta_{k0})\delta_{|k|+|j|,2}\\ &\quad +(1+\delta_{j0})\delta_{|k|+|j|,0}+(1-\delta_{j0})\delta_{|k|+|j|,2}\\ &= 4\delta_{|k|+|j|,0}+(2-\delta_{k0}-\delta_{|k|,2})\delta_{|k|+|j|,2}.\end{aligned}$$

The claim of Lemma 1 is that, when $n=2$,

$$\begin{aligned}(\chi^{*(2)})_{kj} &= \sum_{p=0}^{1}\binom{2}{p}\binom{2}{|k|+p}\delta_{|k|+|j|,2-2p}\\ &= \binom{2}{|k|}\delta_{|k|+|j|,2}+\binom{2}{1}\binom{2}{|k|+1}\delta_{|k|+|j|,0}\\ &= \binom{2}{|k|}\delta_{|k|+|j|,2}+4\delta_{|k|+|j|,0}.\end{aligned}$$

Thus the initial stage $n=2$ is verified if

$$\binom{2}{|k|} = 2-\delta_{k0}-\delta_{|k|,2}$$

holds for $k=0, \pm1, \pm2$, and this is obvious.

Proceeding by induction, we assume that the lemma is valid at some index $n\geqq 2$. Then by the first lines of the proof,

$$\begin{aligned}(\chi^{*(n+1)})_{kj} &= \chi^{*(n)}_{k-1,j}+\chi^{*(n)}_{k+1,j}+\chi^{*(n)}_{k,j-1}+\chi^{*(n)}_{k,j+1}\\ &= \sum_{p=0}^{[n/2]}\binom{n}{p}\left[\binom{n}{|k-1|+p}\delta_{|k-1|+|j|,n-2p}+\binom{n}{|k+1|+p}\delta_{|k+1|+|j|,n-2p}\right]\\ &\quad +\sum_{p=0}^{[n/2]}\binom{n}{p}\binom{n}{|k|+p}[\delta_{|k|+|j-1|,n-2p}+\delta_{|k|+|j+1|,n-2p}].\end{aligned}$$

Using the symmetry cited above, we first assume that $k\geqq 1$, $j\geqq 1$. Then

$$\begin{aligned}(\chi^{*(n+1)})_{kj} &= \sum_{p=0}^{[n/2]}\binom{n}{p}\left[\binom{n}{|k|+p-1}\delta_{|k|+|j|,n-2p+1}+\binom{n}{|k|+p+1}\delta_{|k|+|j|,n-2p-1}\right]\\ &\quad +\sum_{p=0}^{[n/2]}\binom{n}{p}\binom{n}{|k|+p}[\delta_{|k|+|j|,n-2p+1}+\delta_{|k|+|j|,n-2p-1}]\\ &= \sum_{p=0}^{[n/2]}\binom{n}{p}\left[\binom{n}{|k|+p-1}+\binom{n}{|k|+p}\right]\delta_{|k|+|j|,n-2p+1}\\ &\quad +\sum_{p=0}^{[n/2]}\binom{n}{p}\left[\binom{n}{|k|+p+1}+\binom{n}{|k|+p}\right]\delta_{|k|+|j|,n-2p-1}.\end{aligned}$$

Using the identity $\binom{x}{r-1}+\binom{x}{r}=\binom{x+1}{r}$, we get

$$\begin{aligned}(\chi^{*(n+1)})_{kj} &= \sum_{p=0}^{[n/2]}\binom{n}{p}\binom{n+1}{|k|+p}\delta_{|k|+|j|,n+1-2p}+\sum_{p=0}^{[n/2]}\binom{n}{p}\binom{n+1}{|k|+p+1}\delta_{|k|+|j|,n-2p-1}\\ &= \sum_{p=0}^{[n/2]}\binom{n}{p}\binom{n+1}{|k|+p}\delta_{|k|+|j|,n+1-2p}+\sum_{p=1}^{[n/2]+1}\binom{n}{p-1}\binom{n+1}{|k|+p}\delta_{|k|+|j|,n+1-2p}.\end{aligned}$$

Now we can extend the first sum to $p=[n/2]+1$ since the Kronecker delta factor then vanishes at this $p: n+1-2p \leqq 0$, while $|k|+|j|>0$ by assumption. Similarly, we can extend the second sum to $p=0$ since the factor $\binom{n}{p-1}$ then vanishes by convention. Therefore,

$$(\chi^{*(n+1)})_{kj} = \sum_{p=0}^{[n/2]+1} \binom{n+1}{|k|+p} \left[\binom{n}{p} + \binom{n}{p-1} \right] \delta_{|k|+|j|,n+1-2p}$$

and this proves the result. (The remaining cases, $(k,j)=(0,0)$, $(\pm 1,0)$, $(0,\pm 1)$, are easily treated as above.)

Now, using Lemma 1, we have from (27)

$$\mathscr{F}^{-1}\left(\frac{\sin(n+1)\psi}{\sin\psi}\right) = \sum_{m=0}^{[n/2]} (-1)^m \binom{n-m}{m} 2^{2m-n} \sum_{p=0}^{[(n-2m)/2]} \binom{n-2m}{p} \cdot \binom{n-2m}{|k|+p} \delta_{|k|+|j|,n-2m-2p}$$

and, according to (21), we have

$$u_{kj}^{n+1} = \sum_{l=0}^{n-1} \Delta t^2 \left\{ f^{l+1} * \mathscr{F}^{-1}\left(\frac{\sin(n-l)\psi}{\sin\psi}\right) \right\}_{kj}. \tag{30}$$

The indicated convolution equals

$$\begin{aligned} &\sum_{\alpha,\beta} f_{\alpha\beta}^{l+1} \sum_{m=0}^{[(n-l-1)/2]} (-1)^m \binom{n-l-1-m}{m} 2^{2m-n+l+1} \\ &\quad \cdot \sum_{p=0}^{[(n-l-1-2m)/2]} \binom{n-l-1-2m}{p} \binom{n-l-1-2m}{|k-\alpha|+p} \delta_{|k-\alpha|+|j-\beta|,n-l-1-2m-2p} \\ &= \sum_{m=0}^{[(n-l-1)/2]} (-1)^m \binom{n-l-1-m}{m} 2^{2m-n+l+1} \\ &\quad \cdot \sum_{p=0}^{[(n-l-2m-1)/2]} \binom{n-l-1-2m}{p} \sum_{\substack{\alpha,\beta \\ |\alpha-k|+|\beta-j|=n-l-1-2m-2p}} f_{\alpha\beta}^{l+1} \binom{n-l-1-2m}{|k-\alpha|+p} \end{aligned}$$

Summing this over l, $0 \leqq l \leqq n-1$ and multiplying the result by Δt^2, as (30) dictates, we find a representation for u_{kj}^{n+1}. When we replace l by $n-l-1$ there, we get

$$\begin{aligned} u_{kj}^{n+1} = \Delta t^2 \sum_{l=0}^{n-1} \sum_{m=0}^{[l/2]} (-1)^m \binom{l-m}{m} 2^{2m-l} \\ \cdot \sum_{p=0}^{[(l/2)-m]} \binom{l-2m}{p} \sum_{\substack{\alpha,\beta \\ |\alpha-k|+|\beta-j|=l-2m-2p}} f_{\alpha\beta}^{n-l} \binom{l-2m}{|k-\alpha|+p}, \end{aligned}$$

and then by replacing p by $p+m$ we have

$$\begin{aligned} u_{kj}^{n+1} = \Delta t^2 \sum_{l=0}^{n-1} \sum_{m=0}^{[l/2]} (-1)^m \binom{l-m}{m} 2^{2m-l} \\ \cdot \sum_{p=m}^{[l/2]} \binom{l-2m}{p-m} \sum_{\substack{\alpha,\beta \\ |\alpha-k|+|\beta-j|=l-2p}} f_{\alpha\beta}^{n-l} \binom{l-2m}{|k-\alpha|+p-m}. \end{aligned}$$

This sum is taken over those grid points as shown in Fig. 1.

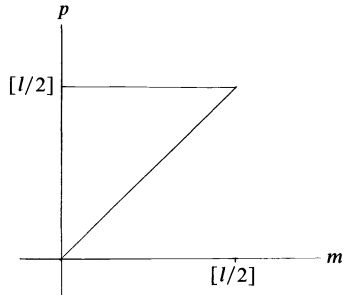


FIG. 1.

Inverting the order of the p, m summations, we can write the result above as

$$u_{kj}^{n+1} = \sum_{l=0}^{n-1} \Delta t^2 \cdot 2^{-l} \sum_{p=0}^{[l/2]} \sum_{\substack{\alpha,\beta \\ |\alpha-k|+|\beta-j|=l-2p}} f_{\alpha\beta}^{n-l} \sum_{m=0}^{p} (-4)^m \binom{l-m}{m} \binom{l-2m}{p-m} \cdot \binom{l-2m}{|k-\alpha|+p-m}. \tag{31}$$

From (7) we recognize the sum over m to be $S_{p,k-\alpha}^{l}$. Thus (31) is the same as (6), and this proves Theorem 1.

Remark. As is seen in (31), the parameters n and p, k, l in $S_{p,k}^{l}$ satisfy $n = 1, 2, \cdots$; $l = 0, 1, \cdots, n-1$; $p = 0, 1, \cdots, [l/2]$; $|k| = 0, 1, \cdots$. Under certain additional conditions on the parameters, it is known that

$$S_{p,k}^{l} = \frac{(-4)^p (l-p)!}{p!\,k!\,(l-2p-k)!}\, {}_4F_3\left(\begin{matrix} -p,\ l+1-p,\ \frac{l+1}{2}-p,\ \frac{l+2}{2}-p \\ l+1-2p,\ k+1,\ l+1-2p-k \end{matrix}; 1\right).$$

We thank R. Asley for this observation. We refer to [1], [5] for the definition of the generalized hypergeometric function ${}_4F_3$.

COROLLARY. *Consider the homogeneous wave equation*

$$u_{tt} - \Delta u = 0$$

and its discretization as in (3):

$$(Lu)_{kj}^{n} \equiv \frac{u_{kj}^{n+1} - 2u_{kj}^{n} + u_{kj}^{n-1}}{\Delta t^2} - \frac{1}{h^2}[u_{k+1,j}^{n} - 2u_{kj}^{n} + u_{k-1,j}^{n}] - \frac{1}{h^2}[u_{k,j+1}^{n} - 2u_{kj}^{n} + u_{k,j-1}^{n}] = 0, \tag{32}$$

where $\Delta x = \Delta y \equiv h$, $\Delta t/h = 1/\sqrt{2}$.

(a) *The solution of*

$$(Lv)_{kj}^{n} = 0, \quad v_{kj}^{0} = \phi_{kj}, \quad v_{kj}^{1} = 0$$

is represented by

$$v_{kj}^{2} = -\phi_{kj}, \quad v_{kj}^{n} = -2^{2-n} \sum_{p=0}^{[(n-2)/2]} \sum_{\substack{\alpha,\beta \\ |\alpha-k|+|\beta-j|=n-2-2p}} S_{p,k-\alpha}^{n-2} \phi_{\alpha\beta} \quad \textit{for } n \geqq 3.$$

(b) *The solution of*

$$(Lw)^n_{kj} = 0, \quad w^0_{kj} = 0, \quad w^1_{kj} = g_{kj}$$

is represented by

$$w^n_{kj} = 2^{1-n} \sum_{p=0}^{[(n-1)/2]} \sum_{\substack{\alpha,\beta \\ |\alpha-k|+|\beta-j|=n-1-2p}} S^{n-1}_{p,k-\alpha} g_{\alpha\beta}.$$

Proof. To prove (a) we make the special choice

$$f^n_{kj} = -(\Delta t)^{-2} \cdot \delta_{n,1} \phi_{kj}$$

in (31). Then (a) follows: the expression (a) is clearly a solution of (32), and it is easily checked that it generates the correct initial values for $n = 2, 3$.

(b) is established by taking the result of (a) and making the identifications

$$w^n_{kj} = -v^{n+1}_{kj}; \qquad \phi_{kj} = g_{kj}.$$

Again, a direct check of the initial values concludes the proof.

3. Positivity of the kernel $S^l_{p,k}$. We will sketch several proofs of Theorem 2 (which sums $S^l_{p,k}$ and establishes its positivity). The first argument cites early work of Watson [6]. Using it has two drawbacks:

(i) It renders the present work non-self-contained;

(ii) The required identification of $S^l_{p,k}$ with a multiple of a particular hypergeometric function of seven parameters holds only under additional restrictions on p, k, l.

For these reasons we also sketch two other proofs. One of these is based on the derivation of a linear recursion for S and on an unusual factorization. The proof of this factorization can be based on generating functions, and this is the content of the second proof.

3.1. Watson's proof. As is standard we denote by ${}_2F_1(a, b, c; z)$ the hypergeometric function of three parameters, and ${}_4F_3\left[\begin{smallmatrix} a,b,c,d \\ e,f,g \end{smallmatrix}; z\right]$ that of seven (cf. [1], [2], [5]). Furthermore, $F_4(\cdots)$ denotes the Appell hypergeometric function ([1, Chap. 9]). Then Watson's result [6] can be written as

$$\begin{aligned} &{}_2F_1(-n, n+a, b; x) \cdot {}_2F_1(-n, n+a, b; 1-x) \\ &= (-1)^n \frac{(a+1-b)_n}{(b)_n} F_4(-n, n+a, b, a+1-b; x(1-x), x(1-x)) \\ &= \frac{(-1)^n (a+1-b)_n}{(b)_n} {}_4F_3\left[\begin{matrix} -n, n+a, \frac{a}{2}, \frac{a+1}{2} \\ a+1-b, b, a, \end{matrix}; 4x(1-x)\right], \end{aligned} \tag{33}$$

where $(a)_n \equiv a(a+1)(a+2)\cdots(a+n-1)$.

We make the following choices:

$$b = k+1, \quad a = l-2p+1, \quad n = p, \quad x = \tfrac{1}{2}.$$

Then the ${}_4F_3(\cdots)$ appearing here is proportional to $S^l_{p,k}$, as is seen from the remark following (31). On the left side of (33) we then have the square of

$${}_2F_1\left(-p, l-p+1, k+1; \frac{1}{2}\right) = \binom{p}{p+k} P_p^{(k,l-2p-k)}(0)$$

using [5, p. 212]. Theorem 2 now follows. We thank G. Gasper for bringing Watson's result to our attention.

3.2. A recursion and factorization for S. First we symmetrize S by defining

$$\tilde{S}^l_{p,q} = \sum_{m=0}^{\infty} (-4)^m \binom{l-m}{m}\binom{l-2m}{p-m}\binom{l-2m}{q-m} \tag{34}$$

for $l = 0, 1, 2, \cdots$; $p, q, = 0, 1, \cdots, l$. Furthermore, we define $\tilde{S}^l_{p,q} = 0$ for integers $p, q, l \geqq 0$ not in the above range. Thus in our previous notation, $S^l_{p,k} = \tilde{S}^l_{p,k+p}$ for $k \geqq 0$, and $m \leqq \max\{p, q\}$ in the above sum.

LEMMA 2. *For $l = 0, 1$ set*

$$\tilde{S}^0_{p,q} = \binom{0}{p}\binom{0}{q}, \qquad \tilde{S}^1_{p,q} = \binom{1}{p}\binom{1}{q}.$$

Then for all integers $p, q, l \geqq 0$, the following recursion holds:

$$\tilde{S}^l_{p,q} = \tilde{S}^{l-1}_{p,q} + \tilde{S}^{l-1}_{p-1,q} + \tilde{S}^{l-1}_{p,q-1} + \tilde{S}^{l-1}_{p-1,q-1} - 4\tilde{S}^{l-2}_{p-1,q-1}.$$

We omit the straightforward but lengthy proof, which uses the recursion for the binomial coefficients. This recursion can be used as follows for integers $p, l = 0, 1, 2, \cdots$ and $q = 0, 1, \cdots, l$. Define

$$R^l_{p,q} = \sum_{\nu=0}^{\infty} (-2)^\nu \binom{p}{\nu}\binom{l-\nu}{q-\nu}, \tag{35}$$

and define $R^l_{p,q} = 0$ for integers $p, q, l \geqq 0$ not in the range above. Then we have the following identity.

LEMMA 3. $\tilde{S}^l_{p,q}$ *admits the factorization*

$$\tilde{S}^l_{p,q} = R^l_{p,q} \cdot R^l_{q,p}.$$

The proof is carried out by showing that both sides satisfy the recursion of Lemma 2. In order to use this result, we recall that $p \leqq q = k + p$, $k \geqq 0$. Since

$$R^l_{p,q} = \binom{l}{q} \sum_\nu (-2)^\nu \frac{\binom{p}{\nu}\binom{q}{\nu}}{\binom{l}{\nu}},$$

$$\tilde{S}^l_{p,q} = \binom{l}{p}\binom{l}{q}\left[\sum_\nu (-2)^\nu \frac{\binom{p}{\nu}\binom{q}{\nu}}{\binom{l}{\nu}}\right]^2 \geqq 0,$$

and we obtain Theorem 2 as follows: by Lemma 3,

$$\tilde{S}^l_{p,q} = R^l_{p,q} R^l_{q,p} = (-2)^{p+q} P_p^{(q-p,l-p-q)}(0) \cdot P_q^{(p-q,l-p-q)}(0),$$

where we have used the definition of the Jacobi polynomial (cf. [5]). For the second factor, we have from [5, p. 210]

$$P_q^{(p-q,l-p-q)}(0) = \frac{\binom{l-p}{q-p}}{\binom{q}{q-p}}\left(-\frac{1}{2}\right)^{q-p} P_p^{(q-p,l-p-q)}(0).$$

When we evaluate the product $\tilde{S}^l_{p,q}$ with $q = k + p$, Theorem 2 results.

Having recognized the existence of the factorization in Lemma 3, we can give a different proof of its validity using generating functions.

The first step is to determine explicitly the generating functions

$$\tilde{S}(x, y, z) = \sum_{p,q,l} \tilde{S}^l_{p,q} x^p y^q z^l,$$

$$R(x, y, z) = \sum_{p,q,l} R^l_{p,q} x^p y^q z^l.$$

Here x, y, z are independent complex variables, restricted to a suitably small neighborhood of the origin in $\mathbb{C}^3$ to make the series absolutely convergent. We find

$$\tilde{S}(x, y, z) = \frac{1}{1 - z(1+x)(1+y) + 4z^2xy},$$

$$R(x, y, z) = \frac{1}{(1 - z(1+y))(1 - zx(1-y))}.$$

To verify the first of these formulae, for instance, we may expand first the geometrical series in powers of $z(1+x)(1+y) - 4z^2xy$, then use the binomial theorem several times, and finally collect terms with equal powers of x, y, z. It is similar for the second formula. The absolute convergence condition for the series $\tilde{S}(x, y, z)$ is

$$|z(1+x)(1+y) - 4z^2xy| < 1;$$

for $R(x, y, z)$ the two conditions

$$|z(1+y)| < 1, \qquad |zx(1-y)| < 1.$$

Before proceeding further, it is useful to make a general remark. Suppose a_n and b_n are two sequences with respective generating functions $f(x) = \sum_n a_n x^n$, $g(x) = \sum_n b_n x^n$, converging absolutely and uniformly in the closed complex disk $|x| \leqq r$. Suppose that x is some complex value for which the strict inequality $|x| < r^2$ holds and we look at the product $f(\xi x) g(1/\xi)$ as a function of ξ. Formally, it is given by the double series $\sum_n \sum_m a_n b_m x^n \xi^{n-m}$, which is a Laurent series in ξ. But by our assumption about f and g it converges uniformly and absolutely in the closed circular annulus

$$\frac{1}{r} \leqq |\xi| \leqq \frac{r}{|x|}.$$

If C is a positively oriented circle around the origin lying inside this annulus we may integrate term by term and obtain

$$\frac{1}{2\pi i} \int_C \frac{d\xi}{\xi} f(x\xi) g\left(\frac{1}{\xi}\right) = \sum_n a_n b_n x^n.$$

This is the basic formula for the generating function of a product sequence $a_n b_n$ in terms of the generating functions of the component sequences a_n and b_n. We propose to apply the three-variable version of this formula to find the generating function of the triple sequence $R^l_{p,q} R^l_{q,p}$.

Accordingly, this generating function will be

$$\frac{1}{(2\pi i)^3} \int_{C_1} \frac{d\xi}{\xi} \int_{C_2} \frac{d\eta}{\eta} \int_{C_3} \frac{d\zeta}{\zeta} R(x\xi, y\eta, z\zeta) R\left(\frac{1}{\eta}, \frac{1}{\xi}, \frac{1}{\zeta}\right)$$

provided x, y, z are small enough and the radii of the circles C_1, C_2, C_3 are suitably fixed. It proves convenient to take a small positive ε (whose actual size is regulated by some detail of the calculations below) and require

$$|x|, |y|, |z| \leqq \varepsilon;$$

furthermore, we take a positive ρ such that

$$2<\rho<\frac{1}{2\varepsilon}.$$

It is then seen that all three integration circles may be taken with radius ρ, and the conditions of validity for the above integral formula are met. We may now do the repeated integrals in any order most convenient for abbreviating the calculations. It turns out that we should first integrate over η, then over ζ, and last over ξ. The integrand is a rational function of η whose polynomial denominator has three factors depending on η; but, fortunately, *only one* vanishes, and that one at a simple zero, inside the contour $|\eta|=\rho$. Thus the η-integral is easily calculated by the Residue theorem, yielding a rational function of ζ and ξ with three factors in the denominator. The story repeats itself, and only one of the three vanishes inside the next integration contour $|\zeta|=\rho$. In the last step we are left with two factors in the denominator depending on ξ, but again only one vanishes inside $|\xi|=\rho$. After the last residue calculation we are left with

$$\frac{1}{(1-zx-zxy)(1-z-zy)+(z-zy)(-zx+zxy)},$$

and a little algebra shows that this is $\tilde{S}(x, y, z)$ as required.

COROLLARY 1. *The kernel $S^{l}_{p,k}$ vanishes when the parameters satisfy any of the following conditions*:

(i) $l=2(k+p)$, *p odd*;

(ii) $l=2p$, $k=p-1, p-3, \cdots, p+1-2[p/2]$;

(iii) $l=8j+1 (j\geqq 1)$, $p=2j$, $k=p+2$, *i.e.*, $l=4p+1$, *p even*, $k=p+2$;

(iv) $l=j^2$, $j=2, 4, 6, \cdots$; $p=2$, $k=\frac{1}{2}(j^2+j-4)$.

Proof. (i) In this case the parameters in $P_p^{(\alpha,\beta)}(0)$ satisfy $\alpha=k$, $\beta\equiv l-2p-k=k=\alpha$. By the Rodrigues formula [5],

$$\begin{aligned}P_p^{(k,k)}(0)&=\frac{(-1)^p}{p!2^p}\frac{d^p}{dx^p}[(1-x^2)^{k+p}]|_{x=0}\\&=c_p\frac{d^p}{dx^p}\sum_{j=0}^{k+p}\binom{k+p}{j}(-1)^j x^{2j}|_{x=0}.\end{aligned}$$

This is an odd-order derivative of an even polynomial, evaluated at $x=0$, and hence vanishes, as claimed.

(ii) Here we have $\alpha=k$, $\beta=l-2p-k=-k=-\alpha$.
By [5, p. 210],

$$P_p^{(k,-k)}(0)=(-1)^p P_p^{(-k,k)}(0)=\frac{(-1)^p\binom{p+k}{k}}{\binom{p}{k}}\left(-\frac{1}{2}\right)^k P_{p-k}^{(k,k)}(0).$$

Applying the Rodrigues formula again, we get

$$\begin{aligned}P_{p-k}^{(k,k)}(0)&=\frac{(-1)^{p-k}}{(p-k)!2^{p-k}}\frac{d^{p-k}}{dx^{p-k}}[(1-x^2)^p]\Big|_{x=0}\\&=\frac{(-1)^{p-k}}{(p-k)!2^{p-k}}\frac{d^{p-k}}{dx^{p-k}}\sum_{\nu=0}^{p}(-1)^\nu\binom{p}{\nu}x^{2\nu}\Big|_{x=0}.\end{aligned}$$

Since $p-k$ is odd, this vanishes as in case (i), and (ii) is proved.

(iii) Here the parameters α, β satisfy

$$\alpha = k, \qquad \beta = l - 2p - k = 8j + 1 - 4j - (2j+2),$$

so $\beta = 2j - 1 = p - 1$ and $\alpha = k = p + 2 = 2j + 2$.
Thus we will show that

$$P_p^{(p+2,p-1)}(0) = 0 \qquad (p \geqq 2).$$

Using Rodrigues again, we see that this vanishes if and only if

$$R_p \equiv \frac{d^p}{dx^p}[(1-x)^{2p+2}(1+x)^{2p-1}]|_{x=0} = 0.$$

Now

$$R_p = \frac{d^p}{dx^p}[(1-x^2)^{2p-1}(1-x)^3]|_{x=0}$$
$$= \sum_{k=0}^{3} \binom{p}{k} \frac{d^k}{dx^k}(1-x)^3|_{x=0} \frac{d^{p-k}}{dx^{p-k}}(1-x^2)^{2p-1}|_{x=0}.$$

Defining $I^l = (d^l/dx^l)(1-x^2)^{2p-1}|_{x=0}$, we obtain

$$R_p = I^p - 3pI^{p-1} + 6\binom{p}{2} I^{p-2} - 6\binom{p}{3} I^{p-3}.$$

Since p is even in case (iii), the terms I^{p-1} and I^{p-3} vanish as in the arguments above. Next we expand $(1-x^2)^{2p-1}$ and differentiate; after a brief calculation we obtain

$$I^p = (-1)^{p/2}\binom{2p-1}{p/2} p!$$

and

$$I^{p-2} = (-1)^{(p-2)/2}\binom{2p-1}{\frac{p-2}{2}}(p-2)!,$$

from which it follows that $R_p = 0$.

(iv) Since $p = 2$, we have by [5]

$$P_p^{(\alpha,\beta)}(0) = P_2^{(\alpha,\beta)}(0)$$
$$= (\alpha+1)(\alpha+2) + (\beta+1)(\beta+2) - 2(\alpha+2)(\beta+2).$$

Substituting $\alpha = k$, $\beta = l - 2p - k = l - 4 - k$, and setting $P_2^{(\alpha,\beta)}(0) = 0$, we obtain the quadratic

$$4k^2 + 4(4-l)k + l^2 - 9l + 16 = 0.$$

The zeros are then written as a function of $l = j^2$, and (iv) follows.

COROLLARY 2. *Define* $\sigma^l = \sum_{p,q \geqq 0} \tilde{S}_{p,q}^l$ *for* $l = 0, 1, \cdots$ *and let* $\tilde{S}_{p,q}^l$ *have the initial values given in Lemma* 2. *Then* $\sigma^l = (1+l) \cdot 2^l$. *Thus, since* $\tilde{S} \geqq 0$ *by Theorem* 2, *the* L^1-*norm of* $S_{p,k}^l$ *is known.*

Proof. Summing the recursion of Lemma 2 over p, q, we obtain $\sigma^l = 4\sigma^{l-1} - 4\sigma^{l-2}$ whence $\sigma^l = (A + Bl)2^l$. Use of the initial values

$$\sigma^0 = 1, \qquad \sigma^1 = 4$$

then gives the result immediately.

Acknowledgment. We are grateful to Professors R. Askey and G. Gasper for sharing their special functions expertise with us.

REFERENCES

[1] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge Tracts in Math. and Math. Phys., 32, Cambridge University Press, Cambridge, UK, 1935.

[2] G. GASPER AND M. RAHMAN, *Product formulas of Watson, Bailey, and Bateman types and positivity of the Poisson kernel for q-Racah polynomials*, SIAM J. Math. Anal., 15 (1984), pp. 768–789.

[3] R. GLASSEY AND W. STRAUSS, *Singularity formation in a collisionless plasma could occur only at high velocities*, Arch. Rational Mech. Anal., 92 (1986), pp. 59–90.

[4] E. ISAACSON AND H. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.

[5] W. MAGNUS, F. OBERHETTINGER, AND R. P. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, Berlin, New York, 1966.

[6] G. N. WATSON, *The product of two hypergeometric functions*, Proc. London Math. Soc., 20 (1922), pp. 189–195.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1776–1789, November 1991

016

APPROXIMATION BY PIECEWISE EXPONENTIALS*

JUNJIANG LEI AND RONG-QING JIA†

Abstract. A function is called an exponential if it is a linear combination of products of polynomials with pure exponentials. In this paper lower and upper bounds for families of spaces of piecewise exponentials are established. In particular, the exact L_p-approximation order ($1 \leq p \leq \infty$) is found for a family $\{S_h\}_{h>0}$ of function spaces when each S_h is generated by an exponential box spline and its multi-integer translates.

Key words. multivariate approximation, order of approximation, exponentials, exponential box splines

AMS(MOS) subject classifications. 41A15, 41A25, 41A63

1. Introduction.

Spaces spanned by multi-integer translates of compactly supported functions have recently attracted much attention. The general setup can be described as follows: Let $\{S_h\}_h$ be a family of function spaces each of which is spanned by the $h\mathbb{Z}^n$-translates of one or several compactly supported functions on $\mathbb{R}^n$. We want to investigate the approximation order of $\{S_h\}_h$ and the ways to realize this approximation order. Here the approximation order of $\{S_h\}_h$ is defined to be the largest real number k for which

$$\operatorname{dist}(f, S_h) = O(h^k)$$

for all sufficiently smooth complex-valued functions on a domain $G \subseteq \mathbb{R}^n$, where dist is measured by some norm (usually an L_p-norm).

A case of particular interest is the scaling case that occurs when the refined spaces S_h are dilations of S_1, i.e.,

$$S_h = \sigma_h S_1$$

with σ_h being the scaling operator

$$\sigma_h : f \mapsto f(\cdot/h).$$

As early as 1946, Schoenberg [19] considered the scaling case where S_1 is spanned by the integer translates of a single compactly supported function ϕ on $\mathbb{R}$. In the late sixties and early seventies Schoenberg's work was extended to the case where S_1 is spanned by several compactly supported functions on $\mathbb{R}^n$. In particular, in the setting of the finite element method, Strang and Fix [20] successfully characterized the so-called controlled approximation order of $\{\sigma_h S_1\}_h$ when S_1 is spanned by a single compactly supported function. However, when S_1 is spanned by several compactly supported functions, their attempt at characterizing the controlled approximation order of $\{\sigma_h S_1\}_h$ failed, as was demonstrated by Jia's counterexample [13]. Nevertheless, the conditions formulated by them to ensure a certain approximation power of $\{\sigma_h S_1\}_h$ have been widely used, and these conditions are now called the Strang–Fix

* Received by the editors July 16, 1990; accepted for publication (in revised form) December 3, 1990.

† Department of Mathematics, University of Oregon, Eugene, Oregon 97403.

conditions. In [4] de Boor and Jia gave a characterization of the local approximation order of $\{\sigma_h S_1\}_h$ in terms of the Strang–Fix conditions.

Both the work of Strang and Fix [20] and that of de Boor and Jia [4] put some restriction on the approximation from $\{\sigma_h S_1\}_h$ (either "controlled" or "local"), and hence do not give a characterization for the (unconditional) approximation order. However, when S_1 is spanned by the multi-integer translates of a box spline, the approximation order of $\{\sigma_h S_1\}_h$ was already established by de Boor and Höllig in [3]. Their work was extended by Ron [18] to the case where S_1 is spanned by the multi-integer translates of a compactly supported function on $\mathbb{R}^n$ under an additional condition. Earlier, Jia [14] characterized the approximation order of $\{\sigma_h S_1\}_h$ when S_1 is spanned by the integer translates of several compactly supported functions on $\mathbb{R}$.

Examples of the nonscaling case were given by Dyn and Ron in [12]. In particular, they emphasized that the Strang–Fix conditions are not applicable to approximation by translates of exponential box splines, which were first introduced by Ron [17]. Using quasi-interpolant schemes based on the Neumann series approach (see [7]), Dyn and Ron [12] established a lower bound for the L_∞-approximation order, but they did *not* show that this lower bound is, in fact, the exact approximation order. This is in sharp contrast to the case where $S_h = \sigma_h S_1$ and S_1 is spanned by the multi-integer translates of a box spline. In such a case, the approximation order is relatively easy to determine (see [3]). Even though a characterization of the approximation order was given by Ron [18], it is not clear how his characterization can be applied to actually find the exact order of approximation from $\{S_h\}_h$, where S_h are spaces spanned by the translates of exponential box splines.

The primary goal of this paper is to fill this gap. For this purpose we need to introduce some terminology and notation first. We shall use the standard multi-index notation. Let $\mathbb{N}$ be the set of nonnegative integers. An element $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n$ is called a multi-index, and the length of α is defined to be $|\alpha| := \sum_{j=1}^n \alpha_j$. The factorial of α is $\alpha! := \alpha_1! \cdots \alpha_n!$. Let $\mathbb{Z}$ be the set of integers. An element of $\mathbb{Z}^n$ is called a multi-integer. A mapping from $\mathbb{Z}^n$ to $\mathbb{C}$ is called a sequence on $\mathbb{Z}^n$.

Let $\mathbb{R}^n$ be the n-dimensional real space equipped with the uniform norm, i.e.,

$$\|x\| := \max_{1 \le j \le n} |x_j|, \qquad x = (x_1, \cdots, x_n) \in \mathbb{R}^n.$$

If $\Omega \subseteq \mathbb{R}^n$ and $r \ge 0$, we denote by $B_r(\Omega)$ the *closed* ball of radius r around Ω, that is,

$$B_r(\Omega) := \{\, x \in \mathbb{R}^n : \ \operatorname{dist}(x, \Omega) \le r \,\}$$

with

$$\operatorname{dist}(x, \Omega) := \inf_{y \in \Omega} \|x - y\|.$$

When $\Omega = \{x\}$, we write $B_r(x)$ for $B_r(\Omega)$, and simply write B_r for $B_r(0)$. Evidently, if Ω is a closed set, then

$$B_r(\Omega) = \Omega + B_r := \{\, x + y : \ x \in \Omega,\, y \in B_r \,\}.$$

If f is a measurable function on a measurable subset Ω of $\mathbb{R}^n$, we denote by $\|f\|_p(\Omega)$ the quantity $(\int_\Omega |f|^p\,dx)^{1/p}$. Similarly,

$$|f|_{k,p}(\Omega) := \sum_{|\alpha|=k} \|D^\alpha f\|_p(\Omega) \quad \text{and} \quad \|f\|_{k,p}(\Omega) := \sum_{|\alpha|\le k} \|D^\alpha f\|_p(\Omega).$$

When Ω is omitted, the norm is understood to be taken over $\mathbb{R}^n$. For a subspace H of $L_p(\Omega)$ and an element $f \in L_p(\Omega)$, let

$$\operatorname{dist}_p(f, H)(\Omega) := \inf_{g\in H} \|f - g\|_p(\Omega).$$

Let $W_p^k = W_p^k(\mathbb{R}^n)$ be the usual Sobolev space equipped with the norm $\|\cdot\|_{k,p}$.

We denote by $\Pi = \Pi(\mathbb{R}^n)$ the linear space of all polynomials on $\mathbb{R}^n$. For a nonnegative integer k, we denote by Π_k (respectively, $\Pi_{<k}$) its subspace of all polynomials of (total) degree at most k (respectively, less than k). In particular, the monomials given by

$$(\ )^\alpha : x \mapsto x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}, \qquad x \in \mathbb{R}^n,$$

are elements of Π. If $p(x) = \sum a_\alpha x^\alpha$ is a polynomial, then $p(D)$ denotes the differential operator induced by p, i.e., $p(D) = \sum a_\alpha D^\alpha$. In particular, $D^\alpha = D_1^{\alpha_1} \cdots D_n^{\alpha_n}$ with D_j being the jth partial derivative operator, $j = 1, \cdots, n$.

Following de Boor and Ron [6], we call a function on $\mathbb{R}^n$ an exponential if it is a linear combination of products of polynomials with the pure exponentials

$$e_\theta : x \mapsto e^{\theta\cdot x}, \qquad \theta \in \mathbb{C}^n,$$

where $\theta \cdot x$ denotes the inner product of θ and x. Note that any finite-dimensional D-invariant (i.e., invariant under differentiation) space of distributions is a space of exponentials (see [1]).

Let A_0 be the linear space of all functions analytic at the origin. An element $f \in A_0$ can be expanded into a power series in a neighborhood of the origin:

$$f(x) = \sum_{\alpha\in\mathbb{N}^n} D^\alpha f(0) x^\alpha/\alpha!.$$

For $j \in \mathbb{N}$, let f_j be the jth homogeneous part of f, i.e.,

$$f_j = \sum_{|\alpha|=j} D^\alpha f(0)(\ )^\alpha/\alpha!.$$

The least term of f, denoted by $f_\downarrow$, is defined as f_j with j being the smallest integer for which $f_j \neq 0$ (see [5]). For a subspace H of A_0 we denote by $H_\downarrow$ the space spanned by all $f_\downarrow$ for $f \in H$.

Now let H be a finite-dimensional space of exponentials. In §3, we will prove our main result, which states that if each S_h consists of piecewise H-functions only, then

the approximation order of $\{S_h\}_h$ does not exceed any integer k for which $\Pi_k \not\subseteq H_\downarrow$. This result was already obtained by de Boor and Höllig [3] for the case where H itself is a space of polynomials.

Lower bounds for the approximation order of $\{S_h\}_h$ are usually established by a concrete approximation scheme using quasi-interpolant methods. In the scaling case, the Strang–Fix conditions and their various equivalent forms are the core of such a quasi-interpolant scheme. As pointed out by Dyn and Ron in [12], the original form of the Strang–Fix conditions is not applicable to the nonscaling case. However, a modified version of the Strang–Fix conditions is still available if we are content with spaces of exponentials, as demonstrated by Jia in [15]. In §2, based on de Boor's survey paper [2], and on our recent work [16], we shall give a construction of L_p-approximation ($1 \le p \le \infty$) from $\{S_h\}_h$ when each S_h is spanned by $h\mathbb{Z}^n$-translates of a compactly supported function on $\mathbb{R}^n$.

In [6], [12], and [18], only L_∞-approximation was considered. In this paper we deal not only with L_∞-approximation, but also L_p-approximation ($1 \le p < \infty$). We contend that L_p-approximation is important. Indeed, Strang and Fix concentrated on L_2-approximation, since their main concern was the finite element method. In the work of DeVore and Popov [11] on approximation by multivariate splines with free knots, L_p-approximation ($0 < p < \infty$) played an essential role. Furthermore, L_p-approximation ($1 \le p < \infty$) has a nature different from that of L_∞-approximation. We can say that L_∞-approximation is essentially local, while L_p-approximation ($1 \le p < \infty$) is global. This point will be made clear in the following sections.

2. Lower bounds for the approximation order. In this section, using a quasi-interpolation scheme, we provide lower bounds for the L_p-approximation order ($1 \le p < \infty$) of a family $\{S_h\}_h$ of approximating spaces, each of which is spanned by the $h\mathbb{Z}^n$-translates of a single compactly supported function.

Let ϕ be a complex-valued Lebesgue-measurable function on $\mathbb{R}^n$. We say that ϕ is a normal function (see [15]), if for any $x \in \mathbb{R}^n$

$$\phi(x) = \lim_{\varepsilon \to 0} \frac{1}{m(B_\varepsilon(x))} \int_{B_\varepsilon(x)} \phi(y)\, dy,$$

where m denotes the Lebesgue measure.

Let ϕ be a compactly supported normal function. For a sequence b on $\mathbb{Z}^n$, the semidiscrete convolution product $\phi *' b$ is the function given by

$$\phi *' b := \sum_{\nu \in \mathbb{Z}^n} \phi(\cdot - \nu) b(\nu).$$

More generally, for a sequence b on $h\mathbb{Z}^n$, the h-scaled semidiscrete convolution product $\phi *'_h b$ is given by

$$\phi *'_h b := \sum_{\nu \in \mathbb{Z}^n} \phi(\cdot - h\nu) b(h\nu).$$

In this section, we assume that $\{\phi_h\}_h$ is a collection of normal functions on $\mathbb{R}^n$ satisfying the following three conditions:

(i) $\operatorname{supp} \phi_h \subseteq hB_r$,

(ii) $\|\phi_h\|_\infty \le c$, and

(iii) $h^{-n}|\hat{\phi}_h(0)| \ge b$,

where r, b and c are positive constants independent of h. In the third condition above, $\hat{\phi}$ denotes the Fourier transform of ϕ:

$$\hat{\phi}(\xi) := \int_{\mathbb{R}^n} \phi(x) e^{-i\xi\cdot x}\, dx.$$

These conditions were first formulated in [6].

Let H be a finite-dimensional D-invariant space of exponentials. From [1] we know that H has the form $\sum_{\theta\in T} e_\theta P_\theta$, where T is a finite subset of $\mathbf{C}^n$ and each P_θ is a D-invariant polynomial space.

THEOREM 2.1. *Let $H = \sum_{\theta\in T} e_\theta P_\theta$ be a finite-dimensional D-invariant exponential space such that $H_\downarrow \supseteq \Pi_{<k}$ for some positive integer k. Assume that $\{\phi_h\}_{h>0}$ is a collection of normal functions on $\mathbb{R}^n$ satisfying the above conditions* (i), (ii), *and* (iii). *If, in addition, each ϕ_h satisfies the Strang–Fix conditions for H, i.e., for each $\theta \in T$,*

$$q(-iD)\hat{\phi}_h(2\pi\nu/h - i\theta) = 0 \quad \text{for all } q \in P_\theta \text{ and } \nu \in \mathbb{Z}^n\backslash\{0\}, \tag{2.1}$$

then with $S_h := \operatorname{range}(\phi_h *'_h)$, $\{S_h\}_{h>0}$ provides L_p-approximation of order at least k $(1 \le p \le \infty)$.

Proof. This theorem is a consequence of Lemma 2.2 and Theorem 2.4, which will be proved later. □

Remark. The Strang–Fix conditions as given in (2.1) are equivalent to the statement that $\phi_h *'_h$ maps $e_\theta P_\theta$ to itself for each $\theta \in T$ (see [15]). Obviously, the latter implies that $\phi_h *'_h$ maps H to itself, while the converse is true if $h(T-T)\cap 2\pi i\mathbb{Z}^n = \{0\}$ (see [6]). Since T is fixed (independent of h), this condition is satisfied for sufficiently small h. Also see [8] for some related results.

LEMMA 2.2. *Let $\{\phi_h\}_{h>0}$ be a collection of compactly supported normal functions on $\mathbb{R}^n$ satisfying the Strang–Fix conditions for $H = \sum_{\theta\in T} e_\theta P_\theta$ as given in* (2.1). *Then for each $\theta \in T$,*

$$\phi_h *'_h (e_\theta p) = h^{-n} e_\theta \big(p(\cdot - iD)\hat{\phi}_h\big)(-i\theta) \quad \text{for all } p \in P_\theta. \tag{2.2}$$

If, in addition, $\{\phi_h\}_{h>0}$ satisfies the condition (i), (ii), *and* (iii), *then for each $h > 0$ there exists a linear combination ψ_h of ϕ_h and its $h\mathbb{Z}^n$-translates such that $\psi_h *'_h$ is an identity on H, and the family $\{\psi_h\}_{h>0}$ satisfies the same conditions* (i) *and* (ii) *with possibly different constants r and c.*

Proof. The proof of Theorem 3.2 in [15] can be carried over verbatim to prove the first statement of this lemma. The second statement was proved by de Boor and Ron in [6]. Here we sketch a proof, taking the Neumann series approach as introduced by Chui and Diamond in [7], and developed by Dyn and Ron in [12]. In the following we denote by 1 the identity mapping.

Conditions (i) and (ii) imply that for $\xi \in \mathbf{C}^n$,

$$\begin{aligned} h^{-n}|\hat{\phi}_h(\xi) - \hat{\phi}_h(0)| &= h^{-n}\left|\int_{B_{rh}} \phi_h(x)(e^{-i\xi\cdot x} - 1)\,dx\right| \\ &\le h^{-n}c\int_{B_{rh}} |e^{-i\xi\cdot x} - 1|\,dx \le h^{-n}c(2rh)^n \max_{\|x\|\le rh} |e^{-i\xi\cdot x} - 1|; \end{aligned}$$

hence $h^{-n}|\hat{\phi}_h(\xi) - \hat{\phi}_h(0)| \to 0$ as $h \to 0$. Thus, if we set $b_{\theta,h} := h^{-n}\hat{\phi}_h(-i\theta)$ for $\theta \in T$ and $h > 0$, then by condition (iii), $|b_{\theta,h}| \ge b/2$ for all $\theta \in T$ and sufficiently small $h > 0$. It follows from (2.2) that for any $\theta \in T$, $p \in P_\theta$, and sufficiently small $h > 0$,

$$(1 - \phi_h *_h' / b_{\theta,h})(e_\theta p) = e_\theta q$$

for some polynomial $q \in P_\theta$ of degree less than $\deg p$. Let $d_\theta := \max\{\deg p : p \in P_\theta\}$ for each $\theta \in T$. We see that the operator $(1 - \phi_h *_h' / b_{\theta,h})^{d_\theta+1}$ annihilates $e_\theta P_\theta$. For each $h > 0$ let V_h be the polynomial in one variable given by

$$V_h(t) := \left\{1 - \prod_{\theta\in T}(1 - t/b_{\theta,h})^{d_\theta+1}\right\}/t, \qquad t \in \mathbb{R},$$

and set

$$\psi_h := V_h(\phi_h *_h')\phi_h.$$

Then $\psi_h *_h'$ is an identity on H, since $1 - \psi_h *_h' = \prod_{\theta\in T}(1 - \phi_h *_h' / b_{\theta,h})^{d_\theta+1}$ annihilates H. Furthermore, we see from the construction of ψ_h that the family $\{\psi_h\}_{h>0}$ satisfies conditions (i) and (ii) with possibly different constants r and c. □

Now let $\{\phi_h\}_{h>0}$ be a collection of normal functions satisfying the conditions (i) and (ii). Suppose that H is a finite-dimensional D-invariant space of exponentials such that $H_\downarrow \supseteq \Pi_{<k}$ and $\phi_h *_h'$ is the identity on H for every $h > 0$. Let $S_h := \text{range}(\phi_h *_h')$. Given $f \in W_p^k(\mathbb{R}^n)$ $(1 \le p \le \infty)$, we want to construct an L_p-approximation scheme from $\{S_h\}_{h>0}$. In the case $p = \infty$, this was done by de Boor and Ron in [6]. For $f \in W_\infty^k$ and $h > 0$, let

$$s_h(x) := \sum_{\nu\in\mathbb{Z}^n} \phi_h(x - \nu h) f(\nu h), \qquad x \in \mathbb{R}^n. \tag{2.3}$$

Then there exists a positive constant C_1 independent of f and h such that

$$|f(x) - s_h(x)| \le C_1 \text{dist}_\infty(f, H)(B_{rh}(x)) \quad \text{for all } x \in \mathbb{R}^n,$$

where r is the constant appearing in the condition (i). Since $H_\downarrow \supseteq \Pi_{<k}$, for given $f \in W_\infty^k$ there exists a constant C_2 independent of h and f such that

$$\text{dist}_\infty(f, H)(B_{rh}(x)) \le C_2 h^k \|f\|_{k,\infty}(B_{rh}(x))$$

(see [12, Thm. 3.1]). Thus with $C = C_1C_2$ we have

$$|f(x) - s_h(x)| \le Ch^k \|f\|_{k,\infty}(B_{rh}(x)) \quad \text{for } x \in \mathbb{R}^n. \tag{2.4}$$

To deal with L_p-approximation ($1 \le p < \infty$) we shall apply a smoothing technique as employed in [16]. Choose a function $\chi \in C_c^\infty(\mathbb{R}^n)$ such that $\operatorname{supp} \chi \subseteq B_1(0)$, $\chi \ge 0$ and $\int \chi = 1$. Set

$$\chi_h := \chi(\cdot/h)/h^n, \qquad h > 0.$$

For a given function $f \in W_p^k$ and $h > 0$, consider the following function:

$$f_h(x) := \int_{\mathbb{R}^n} (f - \nabla_u^k f)(x)\chi_h(u)\, du, \qquad x \in \mathbb{R}^n, \tag{2.5}$$

where ∇_u denotes the difference operator given by

$$\nabla_u f := f - f(\cdot - u).$$

In the univariate case, such a smoothing technique was first introduced by DeVore [10].

The following lemma was proved in [16].

LEMMA 2.3. *The functions f_h are C^∞-smooth. Moreover, there exists a constant C depending only on k such that for any measurable set $\Omega \subseteq \mathbb{R}^n$*

(a) $\|f_h\|_p(\Omega) \le C\|f\|_p(B_{kh}(\Omega))$;

(b) $\|f_h\|_\infty(\Omega) \le Ch^{-n/p}\|f\|_p(B_{kh}(\Omega))$;

(c) $\|f - f_h\|_p(\Omega) \le Ch^k|f|_{k,p}(B_{kh}(\Omega))$.

We are now in a position to construct an L_p-approximation scheme for $1 \le p < \infty$.

THEOREM 2.4. *Let $\{\phi_h\}_{h>0}$ be a collection of normal functions satisfying the conditions* (i) *and* (ii). *Suppose that H is a finite-dimensional D-invariant space of exponentials such that $H_\downarrow \supseteq \Pi_{<k}$ and $\phi_h *_h' $ is the identity on H for every $h > 0$. For a given function $f \in W_p^k$ $(1 \le p \le \infty)$ and any $h > 0$, set*

$$s_h(x) := \sum_{\nu \in \mathbb{Z}^n} \phi_h(x - h\nu) f_h(h\nu), \qquad x \in \mathbb{R}^n,$$

where f_h are given by (2.5). *Then for sufficiently small $h > 0$,*

$$\|f - s_h\|_p \le C\|f\|_{k,p} h^k,$$

where C is a constant independent of f, h, and p.

Proof. By Lemma 2.3(c),

$$\|f - f_h\|_p \le Ch^k|f|_{k,p}.$$

Hence it remains to show that

$$\|f_h - s_h\|_p \le Ch^k \|f\|_{k,p}.$$

When $p = \infty$, this follows from (2.4), if we replace f by f_h in (2.3). In what follows we assume that $1 \le p < \infty$. Our argument is motivated by the work of Dahmen and Micchelli [9]. For $\alpha \in \mathbb{Z}^n$ and $h > 0$, let

$$G_{\alpha,h} := (\alpha + [0,1]^n)h.$$

By (2.4), there exists a constant C_1 such that

$$|f_h(x) - s_h(x)| \le C_1 \|f_h\|_{k,\infty}(B_{rh}(x)) \qquad \text{for all } x \in \mathbb{R}^n.$$

Since the volume of each $G_{\alpha,h}$ is h^n, it follows that

$$\begin{aligned} \int_{G_{\alpha,h}} |f_h(x) - s_h(x)|^p \, dx &\le h^n \sup_{x \in G_{\alpha,h}} |f_h(x) - s_h(x)|^p \\ &\le C_1^p h^n \|f_h\|_{k,\infty}^p (G_{\alpha,h} + B_{rh}). \end{aligned} \tag{2.6}$$

By Lemma 2.3(b), there exists a constant C_2 such that

$$\|f_h\|_{k,\infty}(G_{\alpha,h} + B_{rh}) \le C_2 h^{-n/p} \|f\|_{k,p}(G_{\alpha,h} + B_{(r+k)h}). \tag{2.7}$$

Now, (2.6) and (2.7) together imply

$$\begin{aligned} \|f_h - s_h\|_p^p &= \sum_{\alpha \in \mathbb{Z}^n} \int_{G_{\alpha,h}} |f_h(x) - s_h(x)|^p \, dx \\ &\le (C_1 C_2)^p \sum_{\alpha \in \mathbb{Z}^n} \|f\|_{k,p}^p (G_{\alpha,h} + B_{(r+k)h}). \end{aligned} \tag{2.8}$$

Let $\beta \in \mathbb{N}^n$, $|\beta| \le k$. Then

$$\sum_{\alpha \in \mathbb{Z}^n} \|D^\beta f\|_p^p (G_{\alpha,h} + B_{(r+k)h}) = \int_{\mathbb{R}^n} |D^\beta f(x)|^p \sum_{\alpha \in \mathbb{Z}^n} \rho_\alpha(x) \, dx, \tag{2.9}$$

where ρ_α is the characteristic function of the set $G_{\alpha,h} + B_{(r+k)h}$. For any fixed $x \in \mathbb{R}^n$, the number of $\alpha \in \mathbb{Z}^n$ such that $\rho_\alpha(x) \ne 0$ does not exceed $(2r + 2k + 2)^n$; hence

$$\int_{\mathbb{R}^n} |D^\beta f(x)|^p \sum_{\alpha \in \mathbb{Z}^n} \rho_\alpha(x) \, dx \le (2r + 2k + 2)^n \|D^\beta f\|_p^p. \tag{2.10}$$

Since

$$\|f\|_{k,p}^p = \sum_{|\beta| \le k} \|D^\beta f\|_p^p,$$

it follows from (2.8)–(2.10) that there is a constant C independent of f, h, and p such that

$$\|f_h - s_h\|_p^p \le C^p \|f\|_{k,p}^p,$$

as desired. □

3. Upper bounds for the approximation order. In this section we provide upper bounds for the L_p-approximation order ($1 \le p \le \infty$) of a family $\{S_h\}_{h>0}$ of approximating spaces, each of which is $h\mathbb{Z}^n$-translation invariant and consists of piecewise exponentials.

THEOREM 3.1. *Let H be a finite-dimensional D-invariant space of exponentials, and let Ω be a nonempty open subset of the open unit cube $(0,1)^n \subset \mathbb{R}^n$. Assume that $\{S_h\}_{h>0}$ is a family of linear spaces of functions such that*

$$S_h|_{h\Omega+h\nu} \subseteq H|_{h\Omega+h\nu} \quad \text{for all } \nu \in \mathbb{Z}^n. \tag{3.1}$$

Then for $1 \le p \le \infty$ the L_p-approximation order of $\{S_h\}_{h>0}$ on any domain $G \subseteq \mathbb{R}^n$ does not exceed k, where k is the largest integer such that $\Pi_{<k} \subseteq H_\downarrow$.

Proof. Let $\{q_1, \cdots, q_N\}$ be a basis of homogeneous polynomials for $H_\downarrow$. There exists a basis $\{f_1, \cdots, f_N\}$ for H such that $q_j = f_{j\downarrow}$, $j = 1, \cdots, N$ (see [5]). Moreover, since $\Pi_{<k} \subseteq H_\downarrow$, we can choose f_j $(j = 1, \cdots, N)$ so that

$$f_j(x) - q_j(x) = O(x^k) \quad \text{as } x \to 0.$$

Let

$$d_j := \max\{k-1, \deg q_j\}.$$

Since $f_1, \cdots, f_N$ are exponentials, there exists a positive constant M such that for all $\|x\| \le 1$,

$$|f_j(x) - q_j(x)| \le M\|x\|^{d_j+1}, \qquad j = 1, \cdots, N. \tag{3.2}$$

Since $\Pi_k \not\subseteq H_\downarrow$, there exists $\beta \in \mathbb{N}^n$, $|\beta| = k$, such that $q := (\ )^\beta \notin H_\downarrow$. Let G be a bounded domain in $\mathbb{R}^n$. Our goal is to estimate $\operatorname{dist}_p(q, S_h)(G)$ from below, i.e., to prove that for some positive constant C independent of h,

$$\operatorname{dist}_p(q, S_h)(G) \ge Ch^k. \tag{3.3}$$

Without loss of generality, we may assume that G has Lebesgue measure less than 1. Then by Hölder's inequality,

$$\operatorname{dist}_p(q, S_h)(G) \ge \operatorname{dist}_1(q, S_h)(G) \quad \text{for all } p \ge 1.$$

Hence it suffices to prove (3.3) for $p = 1$.

Note that $q, q_1, \cdots, q_N$ are linearly independent, hence the function given by

$$(a_0, a_1, \cdots, a_N) \mapsto \left\| a_0 q + \sum_{j=1}^{N} a_j q_j \right\|_1 (\Omega) \quad \text{for } (a_0, a_1, \cdots, a_N) \in \mathbf{C}^{N+1}$$

induces a norm on $\mathbf{C}^{N+1}$. Since any two norms on $\mathbf{C}^{N+1}$ are equivalent, there exists a positive constant C_1 such that

$$\left\| a_0 q + \sum_{j=1}^{N} a_j q_j \right\|_1 (\Omega) \geq C_1 \sum_{j=0}^{N} |a_j| \quad \text{for all } a_0, a_1, \cdots, a_N \in \mathbf{C}. \tag{3.4}$$

For any $w \in \mathbb{R}^n$, $q(\cdot + w) - q \in \Pi_{<k} \subseteq H_\downarrow$; hence it is a linear combination of $q_1, \cdots, q_N$:

$$q(\cdot + w) - q = \sum_{j=1}^{N} b_j(w) q_j,$$

where b_j are continuous functions on $\mathbb{R}^n$ with $b_j(0) = 0$ $(j = 1, \cdots, N)$.

In what follows we assume that $0 < h < 1$. For any $a_1, \cdots, a_N \in \mathbf{C}$, we have

$$\begin{aligned} \left\| q(\cdot + h\nu) - \sum_{j=1}^{N} a_j f_j \right\|_1 (h\Omega) &= \left\| q + \sum_{j=1}^{N} b_j(h\nu) q_j - \sum_{j=1}^{N} a_j f_j \right\|_1 (h\Omega) \\ &\geq \left\| q + \sum_{j=1}^{N} (b_j(h\nu) - a_j) q_j \right\|_1 (h\Omega) \\ &\quad - \left\| \sum_{j=1}^{N} a_j (f_j - q_j) \right\|_1 (h\Omega). \end{aligned} \tag{3.5}$$

To estimate the first term of the far right side of (3.5) we note that

$$\|g\|_1(h\Omega) = h^n \|g(h\cdot)\|_1(\Omega).$$

This together with (3.4) implies that

$$\begin{aligned} J_1 := \left\| q + \sum_{j=1}^{N} (b_j(h\nu) - a_j) q_j \right\|_1 (h\Omega) \\ = h^n \left\| h^k q + \sum_{j=1}^{N} (b_j(h\nu) - a_j) h^{\deg q_j} q_j \right\|_1 (\Omega) \\ \geq C_1 h^n \left(h^k + \sum_{j=1}^{N} |b_j(h\nu) - a_j| h^{d_j} \right). \end{aligned} \tag{3.6}$$

To estimate the second term of the far right side of (3.5) we note that for any $g \in L_1(h\Omega)$

$$\|g\|_1(h\Omega) \le h^n \sup_{x\in h\Omega} |g(x)|, \tag{3.7}$$

since $h\Omega \subset h(0,1)^n$. It follows from (3.2) and (3.7) that

$$J_2 := \left\| \sum_{j=1}^N a_j(f_j - q_j) \right\|_1 (h\Omega) \le \sum_{j=1}^N |a_j| \|f_j - q_j\|_1(h\Omega) \le Mh^n \sum_{j=1}^N |a_j| h^{d_j+1}. \tag{3.8}$$

For sufficiently small h, say $h \le C_1/M$, we have

$$C_1|b_j(h\nu) - a_j| \ge Mh\left(|a_j| - |b_j(h\nu)|\right).$$

This together with (3.6) and (3.8) yields

$$J_1 - J_2 \ge h^n \left(C_1 h^k - M \sum_{j=1}^N |b_j(h\nu)| h^{d_j+1} \right).$$

Since $d_j + 1 \ge k$, while $b_j(h\nu)$ can be made small if $h\nu$ is small, we conclude that there exists $\delta > 0$ such that

$$\left\| q(\cdot + h\nu) - \sum_{j=1}^N a_j f_j \right\|_1 (h\Omega) \ge J_1 - J_2 \ge (C_1/2)h^{k+n} \quad \text{for } \|\nu\| \le \delta/h.$$

The above estimate holds for any $a_1, \cdots, a_N \in \mathbf{C}$; hence

$$\operatorname{dist}_1\left(q(\cdot + h\nu), H\right)(h\Omega) \ge (C_1/2)h^{k+n} \quad \text{for } \|\nu\| \le \delta/h.$$

Since H is translation invariant, it follows that

$$\begin{aligned} \operatorname{dist}_1(q, H)(h\Omega + h\nu) &= \operatorname{dist}_1(q(\cdot + h\nu), H)(h\Omega) \\ &\ge (C_1/2)h^{k+n} \quad \text{for } \|\nu\| \le \delta/h. \end{aligned} \tag{3.9}$$

We are now ready to estimate $\operatorname{dist}_1(q, S_h)(B_\delta) =: \varepsilon(h)$. Observe that

$$B_\delta \supseteq \bigcup_{\|\nu\| \le \delta/h - 1} (h\Omega + h\nu).$$

It follows that

$$\varepsilon(h) \ge \sum_{\|\nu\| \le \delta/h - 1} \operatorname{dist}_1(q, S_h)(h\Omega + h\nu). \tag{3.10}$$

But by (3.1), $S_h|_{h\Omega+h\nu} \subseteq H|_{h\Omega+h\nu}$; hence by (3.9) we have

$$\begin{aligned} \operatorname{dist}_1(q, S_h)(h\Omega + h\nu) &\geq \operatorname{dist}_1(q, H)(h\Omega + h\nu) \\ &\geq (C_1/2)h^{k+n} \quad \text{for } \|\nu\| \leq \delta/h. \end{aligned} \tag{3.11}$$

For sufficiently small $h > 0$, the set $\{\nu \in \mathbb{Z}^n : \|\nu\| \leq \delta/h - 1\}$ has cardinality $\geq (\delta/h)^n$; therefore (3.10) and (3.11) yield

$$\varepsilon(h) \geq (\delta/h)^n (C_1/2)h^{n+k} = (C_1/2)\delta^n h^k.$$

This shows that for sufficiently small $h > 0$,

$$\operatorname{dist}_1(q, S_h)(B_\delta) \geq Ch^k,$$

where $C = (C_1/2)\delta^n$. This proves (3.3) for $p = 1$, as desired. □

Finally, we apply the previous results to the problem of approximation by exponential box splines. Following Dyn and Ron [12], we introduce the h-scaled EB-splines as follows. Let Γ be a finite set of pairs (not necessarily distinct) of the form

$$\gamma = (x_\gamma, \lambda_\gamma), \qquad x_\gamma \in \mathbb{R}^n \backslash \{0\}, \quad \lambda_\gamma \in \mathbb{C}.$$

Hereafter we always assume that $X := \{x_\gamma : \gamma \in \Gamma\} \subseteq \mathbb{Z}^n$ and that X spans $\mathbb{R}^n$. The h-scaled EB-spline $B_h(\Gamma \mid \cdot)$ is defined by the equation

$$\int_{\mathbb{R}^n} B_h(\Gamma \mid x)\psi(x)\,dx = h^{n-\#\Gamma} \int_{[0,h]^\Gamma} \left(\prod_{\gamma\in\Gamma} e^{\lambda_\gamma t_\gamma} \right) \psi\left(\sum_{\gamma\in\Gamma} t_\gamma x_\gamma \right) dt,$$

where ψ is taken from a suitable space of test functions. It is known from [12] that the collection $\{B_h(\Gamma \mid \cdot)\}_{h>0}$ satisfies conditions (i), (ii), and (iii).

Let $\mathcal{D}'(\mathbb{R}^n)$ denote the space of all n-dimensional complex-valued distributions. For $K \subseteq \Gamma$, let $p_K(D)$ be the differential operator induced by the polynomial

$$p_K(x) := \prod_{\gamma\in K} (x \cdot x_\gamma - \lambda_\gamma).$$

Let

$$K(\Gamma) := \{K \subseteq \Gamma : \ \operatorname{span}\{x_\gamma\}_{\gamma\in\Gamma\backslash K} \neq \mathbb{R}^n\},$$

and

$$k(X) := \min\{\#K : \ K \in K(\Gamma)\}.$$

Define

$$H := \{f \in \mathcal{D}'(\mathbb{R}^n) : \ p_K(D)f = 0 \text{ for all } K \in K(\Gamma)\}.$$

Then H is a finite-dimensional D-invariant space of exponentials (see, e.g., [1]). Let S_h be the range of $B_h(\Gamma \mid \cdot)*'_h$. Then S_h is $h\mathbb{Z}^n$-translation invariant, $S_h \supset H$, and $S_h|_{(0,h)^n}$ consists of piecewise H-functions. It is also known from [12] that with $k = k(X)$, $H_\downarrow \supseteq \Pi_{<k}$ but $H_\downarrow \not\supseteq \Pi_k$. Dyn and Ron [12] proved that k is a lower bound for the L_∞-approximation order of $\{S_h\}_{h>0}$. Now that $\{S_h\}_{h>0}$ and H satisfy the conditions of Theorems 2.1 and 3.1, we have the following concluding result.

THEOREM 3.2. *Let S_h be the range of $B_h(\Gamma \mid \cdot)*'_h$. Then the exact L_p-approximation order of $\{S_h\}_{h>0}$ is $k(X)$ $(1 \le p \le \infty)$.*

Acknowledgment. The authors thank Professor Carl de Boor for his valuable suggestions and comments on this paper.

REFERENCES

[1] A. BEN-ARTZI AND A. RON, *Translates of exponential box splines and their related spaces*, Trans. Amer. Math. Soc., 309 (1988), pp. 683–710.

[2] C. DE BOOR, *Quasiinterpolants and approximation power of multivariate splines*, in Computation of Curves and Surfaces, W. Dahmen, M. Gasca and C. A. Micchelli, eds., Kluwer, Dordrecht, the Netherlands, 1990, pp. 313–345.

[3] C. DE BOOR AND K. HÖLLIG, *B-splines from parallelepipeds*, J. Analyse Math., 42 (1982/3), pp. 99–115.

[4] C. DE BOOR AND R. Q. JIA, *Controlled approximation and a characterization of the local approximation order*, Proc. Amer. Math. Soc., 95 (1985), pp. 547–553.

[5] C. DE BOOR AND A. RON, *On multivariate polynomial interpolation*, Constr. Approx., 6 (1990), pp. 287–302.

[6] _____, *The exponentials in the span of the integer translates of a compactly supported function*, Computer Sciences Tech. Report 887, Univ. of Wisconsin, Madison, WI, 1989.

[7] C. K. CHUI AND H. DIAMOND, *A natural formulation of quasi-interpolation by multivariate splines*, Proc. Amer. Math. Soc., 99 (1987), pp. 643–646.

[8] C. K. CHUI AND J. Z. WANG, *Quasi-interpolation functionals on the space of EP splines*, preprint.

[9] W. DAHMEN AND C. A. MICCHELLI, *On the approximation order from certain multivariate spline spaces*, J. Austral. Math. Soc. Ser B, 26 (1984), pp. 233–246.

[10] R. A. DEVORE, *Degree of approximation*, in Approximation II, G. G. Lorentz, C. K. Chui, and L. L. Schumaker, eds., Academic Press, New York, 1976, pp. 117–162.

[11] R. A. DEVORE AND V. A. POPOV, *Free multivariate splines*, Constr. Approx., 3 (1987), pp. 239–248.

[12] N. DYN AND A. RON, *Local approximation by certain spaces of exponential polynomials, approximation order of exponential box splines, and related interpolation problems*, Trans. Amer. Math. Soc., 319 (1990), pp. 381–404.

[13] R. Q. JIA, *A counterexample to a result concerning controlled approximation*, Proc. Amer. Math. Soc., 97 (1986), pp. 647–654.

[14] _____, *A characterization of the approximation order of translation invariant spaces of functions*, Proc. Amer. Math. Soc., 111 (1991), pp. 61–70.

[15] _____, *A dual basis for the integer translates of an exponential box spline*, Rocky Mountain J. Math., to appear.

[16] R. Q. JIA AND J. J. LEI, *Approximation by multi-integer translates of functions having global support*, J. Approx. Theory, to appear.

[17] A. RON, *Exponential box splines*, Constr. Approx., 4 (1988), pp. 357–378.

[18] _____, *A characterization of the approximation order of multivariate spline spaces*, Computer Sciences Tech. Report 885, Univ. of Wisconsin, Madison, WI, 1989.

[19] I. J. SCHOENBERG, *Contributions to the problem of approximation of equidistant data by analytic functions*, A, B, Quart. Appl. Math., 4 (1946), pp. 45–99, 112–141.

[20] G. STRANG AND G. FIX, *A Fourier analysis of the finite-element variational method*, in Constructive Aspects of Functional Analysis, G. Geymonat, ed., C.I.M.E., Rome, 1973, pp. 793–840.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, pp. 1790-1802, November 1991

017

ON THE DETERMINATION OF ZIGLIN MONODROMY GROUPS*

RICHARD C. CHURCHILL† AND DAVID L. ROD‡

Abstract. The monodromy group of a second-order linear differential equation with rational coefficients is called Ziglin if it preserves a nonconstant rational function. The determination of which monodromy groups are Ziglin is essential in integrability questions for complex analytic Hamiltonian systems. In this paper the problem is solved completely for the Fuchsian case by using the Kovacic algorithm to determine the differential Galois group of that second-order equation and then relating this to the monodromy group. Applications are given to Hamiltonian systems.

Key words. Ziglin group, Hamiltonian system, nonintegrable, monodromy group, differential Galois group, linear differential equation GL(2, $\mathbb{C}$), SL(2, $\mathbb{C}$), Kovacic algorithm, Fuchsian equation

AMS(MOS) subject classifications. 58F05, 34A20, 34A30, 20G20, 13N05, 13B10

Introduction. Consider a linear differential equation $y''+p(x)y'+q(x)y=0$, $'=(d/dx)$, with rational coefficients on the Riemann sphere $\mathbb{P}^1$ or, more generally, a holomorphic (flat) connection on a rank 2 complex vector bundle over a Riemann surface (see [1]). The monodromy group can be viewed as an automorphism group of some distinguished fiber, and is called Ziglin if it preserves a nonconstant rational function on this vector space (see [20]). The determination of which monodromy groups are Ziglin is crucial in integrability questions for complex analytic Hamiltonian systems. Here we solve the problem completely for the Fuchsian case on $\mathbb{P}^1$, and in more general bundle contexts where symmetries allow reduction to this first case. The key elements are a classification of Ziglin subgroups of GL(2, $\mathbb{C}$) and an algorithm of Kovacic [10] that determines the nature of the differential Galois group of a second-order equation (as above) on $\mathbb{P}^1$.

The necessary background in differential Galois theory appears in § 2, and its application to determining the Ziglin subgroups of GL(2, $\mathbb{C}$) in § 3. The Kovacic algorithm is outlined for the Fuchsian case in § 4, with applications to Hamiltonian systems in § 5. We note that, although the Kovacic algorithm works for any second-order linear differential equation, we consider the Fuchsian case because it is only in that case the Zariski closure of the monodromy group is equal to the differential Galois group. (In the irregular case it is necessary to add the Stokes multipliers.) For background in these group theoretical properties of differential equations we refer to [7], [9], and [11], and especially the surveys [16] and [17]. After this paper was completed the authors became aware of the recent thesis of Morales [21], which also considers Ziglin analysis in terms of differential Galois theory, but with a different emphasis and without Kovacic's algorithm. We refer the reader to this work and [22] for their many nice examples and as complementary reading.

1. Motivation. Let X be a Riemann surface and ∇ a connection on a rank 2 complex vector bundle $\pi_E: E \to X$. To each loop γ in X based at $x_0 \in X$ assign the

* Received by the editors April 16, 1990; accepted for publication (in revised form) November 6, 1990. This work was supported in part by the Institute for Mathematics and Its Applications with funds provided by the National Science Foundation.

† Department of Mathematics, Hunter College, 695 Park Avenue, New York, New York 10021. The research of this author was supported in part by National Science Foundation grant DMS 8802911.

‡ Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta T2N 1N4, Canada. The research of this author was supported in part by Natural Sciences and Engineering Research Council of Canada grant OGP0008507.

automorphism of the fiber $E_0 = \pi_E^{-1}(\{x_0\})$ which results on sending $v \in E_0$ to the endpoint of the horizontal lift of γ^{-1} issuing from v. Since ∇ is flat, this lift depends only on the homotopy class of γ in $\pi_1(X, x_0)$, and defines the *monodromy representation* $\rho : \pi_1(X, x_0) \to \mathrm{Aut}(E_0)$. The image $G_0 = \rho(\pi_1(X, x_0))$ is the *monodromy group* of ∇ at x_0. Different choices of basepoint give isomorphic monodromy groups, and given a basis of E_0 we can identify G_0 with a subgroup of $\mathrm{GL}(2, \mathbb{C})$. In the classical case $X = \mathbb{C}\backslash\{\text{finite set}\}$ and ∇ can be viewed as a linear ordinary differential equation on the complex plane with possible singularities on this finite set as well as at ∞ (see [4, p. 93]). Then E_0 is identified with the initial values (and hence germs at x_0) of solutions, and ρ, and hence G_0, are defined by analytic continuation of these solutions along γ^{-1}.

Throughout this paper any function said to be an "integral" will be assumed nonconstant. Thus G_0 has an *integral* f, or is a *Ziglin group*, if there is a nonconstant rational function f on E_0 that is preserved by the action; that is, $g^* f = f$ for all $g \in G_0$. An equivalent condition is the existence of an *integral* F for ∇, that is, a meromorphic function F on E that is rational on fibers and constant along horizontal lifts of curves in X. Indeed, $f = F \mid E_0$ will be preserved by G_0, hence G_0 will be Ziglin, and any such f can be uniquely extended to a corresponding integral F of ∇ through "parallel transport."

Example 1.1 (Ziglin theory for two degree-of-freedom Hamiltonian systems). Let M be a complex symplectic 4-manifold and Γ a nonequilibrium phase curve, within an energy surface Σ, of a holomorphic Hamiltonian vector field X_H on M. Linearization along Γ induces a holomorphic (hence flat) connection ∇ on the normal bundle $N = (T\Sigma \mid \Gamma)/T\Gamma$ of Γ in Σ, called the *normal variational equation* (NVE). By a result of Ziglin [20] any meromorphic integral of X_H independent on a neighborhood (in M) of Γ (but not necessarily on Γ itself) will induce an integral for ∇; hence G_0 will be a Ziglin group. Therefore, if G_0 is not a Ziglin group, X_H cannot be integrable by such meromorphic functions on a neighborhood of the phase curve (although X_H could still be integrable, say, by differentiable functions).

When reasonable symmetries are present a connection can be reduced. Specifically, suppose a finite group G acts π_E-equivariantly on $E \to X$, freely and properly discontinuously on X, and linearly on fibers. Then X/G is again a Riemann surface and the connection $\tilde{\nabla} = (1/|G|) \sum_{g \in G} g^* \nabla$ is preserved by G, thus inducing a connection ∇_G on $E/G \to X/G$. We say ∇ is *symmetric* (with respect to G) if $\nabla = \tilde{\nabla}$.

THEOREM 1.2. *Let ∇ be symmetric with respect to the finite group G which acts on $E \to X$ as above. Then ∇ admits an integral if and only if this is the case for ∇_G. In particular, the monodromy group of ∇ is Ziglin if and only if the monodromy group of ∇_G is Ziglin.*

Proof. (a) Assume that ∇ admits an integral. It suffices to show that the existence of an integral F for ∇ implies the existence of a G-invariant integral. We do this by adapting an argument from [20, p. 186], which deals with a related situation.

Let $G = \{g_j\}_{j=1}^m$ with $g_1 = \mathrm{id}$, and set $f_j = g_j^* F$. Assume $\{f_1, \cdots, f_k\}$ are the distinct f_j, and let m_j be the number of occurrences of f_j in $\{f_1, \cdots, f_m\}$. Now consider the G-invariant integrals $\psi_i = \sum_{j=1}^m (f_j)^i$ of ∇. Since

$$\det \frac{\partial(\psi_1, \cdots, \psi_k)}{\partial(f_1, \cdots, f_k)} = (k!)\left(\prod_{j=1}^k m_j\right) \prod_{1 \le i < j \le k} (f_j - f_i) \neq 0$$

(the last term is a standard Vandermonde determinant), we can solve locally for the f_i as analytic functions of the $\{\psi_j\}$. In particular, we can write $F = f_1$ locally as an analytic function of the $\{\psi_j\}$, and when restricted to a fiber we then have $0 \neq dF = \sum a_j \, d\psi_j$

for appropriate a_j. We conclude that some ψ_j must be nonconstant and hence a symmetric integral for ∇.

(b) The converse of (a) follows from the fact that the monodromy group of ∇ can be identified with a subgroup of the monodromy group of ∇_G (see [4, Prop. 1.1, p. 94]) and the comments preceding Example 1.1 on the relation between integrals of the monodromy group and integrals of the connection. □

COROLLARY 1.3. *Assume the connection ∇ on $N \to \Gamma$ of Example* 1.1 *is symmetric with respect to a finite group G acting as in Theorem* 1.2. *If the monodromy group of the reduced connection ∇_G is not Ziglin, then X_H has no meromorphic integral independent of H.*

We will see an example of Corollary 1.3 in § 5. We refer to [3] and [4] for examples of symmetric connections. In particular, [3, § 4] presents a general theory concerning the case when a given connection is symmetric. In examples the group G is often given as an action on the base space X; [4, Prop. 2.3] shows when this can be lifted to a π_E-equivariant action on $E \to X$ under which ∇ is symmetric (see also [3, § 4]).

In the next few sections we will assume that reduction has transformed ∇ to a connection ∇_G that can be viewed as a linear ordinary differential equation on $\mathbb{P}^1$. When ∇_G is Fuchsian we show that the monodromy group of ∇_G is Ziglin if and only if the differential Galois group of ∇_G is Ziglin. We then adapt an algorithm of Kovacic [10] (see also [5]) in § 4 to determine whether the differential Galois group has this property.

2. Preliminaries on differential Galois theory. In this section we collect some standard results on differential Galois theory and algebraic subgroups of $\mathrm{SL}(2,\mathbb{C})$ needed in later sections. Let $\mathbb{C}(x)$ denote the field of rational functions in x with coefficients in $\mathbb{C}$ and consider the (normalized) linear ordinary differential equation

$$y'' = r(x)y, \qquad ' = (d/dx) \tag{2.1}$$

on the Riemann sphere $\mathbb{P}^1$, where $r(x) \in \mathbb{C}(x)$. Let Λ denote the set of poles of $r(x)$, set $X = \mathbb{P}^1 \setminus (\Lambda \cup \{\infty\})$, and fix $x_0 \in X$.

$\mathcal{M}$ will denote the field of germs of meromorphic functions at x_0, which we view as an extension of $\mathbb{C}(x)$ by identifying the latter with the germs of such functions at x_0. $V \subset \mathcal{M}$ will denote the linear space of germs of solutions of (2.1) at x_0, $V' \subset \mathcal{M}$ the associated derivatives, and $E \subset \mathcal{M}$ the extension of $\mathbb{C}(x)$ generated by $V \cup V'$. E is the *Picard-Vessiot extension* of $\mathbb{C}(x)$ associated to (2.1), and the *differential Galois group* $G_D = G(E/\mathbb{C}(x))$ of that equation is the group of automorphisms of E that fix $\mathbb{C}(x)$ and commute with differentiation. Elements of G_D are determined by their action on V; hence G_D may be viewed as a subgroup of $\mathrm{Aut}(V)$.

A subgroup of $\mathrm{Aut}(V) \cong \mathrm{GL}(2,\mathbb{C})$ is (1) *reducible* (or *triangulizable*) if it is conjugate to a (lower) triangular subgroup of $\mathrm{GL}(2,\mathbb{C})$ and (2) a DP-*group* if it is conjugate to a subgroup of $\{\left(\begin{smallmatrix}\alpha & 0\\ 0 & \beta\end{smallmatrix}\right)\} \cup \{\left(\begin{smallmatrix}0 & \gamma\\ \delta & 0\end{smallmatrix}\right)\} \subset \mathrm{GL}(2,\mathbb{C})$. (This terminology for "diagonal permutation" (DP) is from [2].) Note that a diagonalizable group is both reducible and a DP-group.

PROPOSITION 2.2. *Let G_N be the monodromy group of* (2.1) *and G_D its differential Galois group. Then*

(a) *G_D is an algebraic unimodular subgroup of* $\mathrm{Aut}(V)$. *In particular, it is Zariski closed in this space.*

(b) *Any element of E fixed by all elements of G_D must be in $\mathbb{C}(x)$.*

(c) *$G_N \subset G_D$.*

Moreover, if (2.1) *is Fuchsian we have*

(d) *$\bar{G}_N = G_D$, i.e., G_N is Zariski dense in G_D;*

(e) *G_N is reducible, diagonalizable, or a* DP*-group if and only if G_D is such*; *and*

(f) *G_N is finite if and only if G_D is such, in which case $G_N = G_D$.*

Proof. (a) and (b) are standard (e.g., see [9, pp. 36, 38, and 41]). (c) is adapted from [18]: analytic continuation along the inverse of any loop in X based at x_0 defines an element of G_N and one of G_D, and these are identical when viewed in Aut (V). (d) is Proposition III of [18]. (e) and (f) are immediate from (c) and (d). □

PROPOSITION 2.3. *An algebraic subgroup of* SL $(2, \mathbb{C})$ *is either*

Case I: *reducible*;

Case II: *a DP-group*;

Case III: *a finite group which, if not of Case* I *or* II, *must be projectively* (*i.e.*, mod ±id) *the tetrahedral, octahedral, or icosahedral group*; *or*

Case IV: SL $(2, \mathbb{C})$.

In particular, the differential Galois group G_D of (2.1) must have one of these forms.

Proof. See [9, pp. 31, 32] or [10, pp. 7, 27]. □

PROPOSITION 2.4. *If every element of an algebraic subgroup of* GL $(n, \mathbb{C})$ *has finite order, then that group must be finite.*

Proof. This is a lemma in [18, p. 1328]. □

3. Ziglin subgroups of GL (2, C). Application of Corollary 1.3 when the action of the group G is not symplectic may yield a reduced connection ∇_G that is Fuchsian on $\mathbb{P}^1$ but not of the form (2.1). The monodromy of ∇_G will then be a subgroup of GL $(2, \mathbb{C})$ rather than SL $(2, \mathbb{C})$. The purpose of this section is to develop relationships (see Theorem 3.5 and Corollary 3.6 below) that allow us to exploit the classification of algebraic subgroups of SL $(2, \mathbb{C})$ given in Proposition 2.3 (which should be compared with Corollary 3.4 below).

Recall from § 1 that a subgroup $Z \subset$ GL $(2, \mathbb{C})$ is *Ziglin*, or a *Ziglin group*, if there is a nonconstant rational function $f: \mathbb{C}^2 \to \mathbb{C}$ preserved by Z; i.e., such that $g^* f = f$ for all $g \in Z$.

PROPOSITION 3.1. (a) *Any conjugate of a Ziglin group Z is again Ziglin.*

(b) *Any subgroup of a Ziglin groupis again Ziglin.*

(c) *The Zariski closure $\bar{Z}$ of any Ziglin subgroup Z preserves any rational function f preserved by Z. In particular, $\bar{Z}$ must be Ziglin.*

(d) *A subgroup $Z \subset$ GL $(2, \mathbb{C})$ is Ziglin if and only if the Zariski closure $\bar{Z}$ is Ziglin.*

Proof. For (a), $g^* f = f$ implies $(h^{-1}gh)^* h^* f = h^* f$ for any $h \in$ GL $(2, \mathbb{C})$. (b) is obvious. For (c), write $f = (p/q)$ where p and q are polynomials and fix $w \in \mathbb{C}^2$. Then $P_w(g) = (g^* p)(w) \cdot q(w) - p(w) \cdot (g^* q)(w)$ is a polynomial vanishing on Z; hence $\bar{Z}$ is contained in the zero set of P_w. Since w was arbitrary the result follows. We obtain (d) by (b) and (c). □

Now let $Z \subset$ GL $(2, \mathbb{C})$ be Ziglin and let $f = (p/q)$ be a nonconstant rational function preserved by Z. Since $g^* f = f$ for $g \in Z$ if and only if $(g^* p) \cdot q = p \cdot (g^* q)$, by comparing lowest- (or highest-) order terms in this last expression, we see that p and q may be assumed homogeneous. Then by factoring and, if necessary, conjugating by some suitable rotation h, we can assume

$$f(x, y) = \prod_{j=1}^{r} (x - \lambda_j y)^{m_j}, \tag{3.2}$$

where the $\{\lambda_j\}$ are distinct and $m_j \in \mathbb{Z} \setminus \{0\}$. Note that such a factorization does not generally hold for homogeneous polynomials in three or more variables (e.g., see [6, pp. 50, 51]).

THEOREM 3.3. *The Ziglin subgroups of* $\mathrm{GL}(2,\mathbb{C})$ *are precisely those that can be conjugated to a subgroup of one of the following groups*:

(1) $T(n)=\left\{\begin{pmatrix}\alpha & 0\\ \delta & \beta\end{pmatrix}\mid \alpha^n=1\right\}$,

(2) $T(m,n)=\left\{\begin{pmatrix}\alpha & 0\\ 0 & \beta\end{pmatrix}\mid \alpha^m\beta^n=1\right\}$,

(3) $D(n)=\left\{\begin{pmatrix}\alpha & 0\\ 0 & \beta\end{pmatrix}\mid (\alpha\beta)^n=1\right\}\cup\left\{\begin{pmatrix}0 & \gamma\\ \delta & 0\end{pmatrix}\mid (\delta\gamma)^n=1\right\}$,

(4) *The projectively finite groups.*

Proof. Let Z be a Ziglin subgroup of $\mathrm{GL}(2,\mathbb{C})$ that fixes the rational function (3.2). We have three cases.

Case I ($r=1$). Z then fixes the line $x=\lambda_1 y$, and hence there is a conjugacy h so that $(h^*f)(x,y)=x^n$ (on setting $m_1=n$) and $(h^{-1}Zh)\subset T(n)$. Conversely, $(h^{-1}Z_0h)\subset T(n)$ implies that the subgroup Z_0 preserves $(h^{-1})^*(x^n)$.

Case II ($r=2$). Each $g\in Z$ will preserve or permute the two lines $x=\lambda_j y$, $j=1,2$. There is then a conjugacy h so that $(h^*f)(x,y)=x^m y^n$ (on setting $m_1=m$, $m_2=n$) and $(h^{-1}Zh)\subset\{\left(\begin{smallmatrix}\alpha & 0\\ 0 & \beta\end{smallmatrix}\right)\}\cup\{\left(\begin{smallmatrix}0 & \gamma\\ \delta & 0\end{smallmatrix}\right)\}$. If there are no elements of the form $g=\left(\begin{smallmatrix}0 & \gamma\\ \delta & 0\end{smallmatrix}\right)$ in $(h^{-1}Zh)$, then $\alpha^m\beta^n=1$ implies $(h^{-1}Zh)\subset T(m,n)$. If there is such an element g then $(g^*f)(x,y)=\gamma^m\delta^n x^n y^m=x^m y^n$ implies $m=n$ and $(h^{-1}Zh)\subset D(n)$. Conversely, if $(h^{-1}Z_0h)$ is a subgroup of (2) or (3) above, then Z_0 preserves $(h^{-1})^*(x^m y^n)$, with $m=n$ for case (3).

Case III ($r\geqq 3$). Let $K=\{c\cdot \mathrm{id}\mid c\in\mathbb{C}\backslash\{0\}\}$. We must show that the projectivization $PZ=Z/(Z\cap K)$ is finite. There is a positive integer n so that for all $g\in Z$ the element g^n fixes each of the lines $x=\lambda_j y$ ($j=1,\cdots,r$). Since $r\geqq 3$, this forces $g^n=c(g)\cdot\mathrm{id}$, where $c(g)$ is a constant dependent on g.

(a) If in (3.2) we have $\sum_{j=1}^r m_j\neq 0$, then $(g^n)^*f=f$ implies each $c(g)$ is a root of unity. By Proposition 3.1(c) we may replace Z by the algebraic group $\bar{Z}$; hence Z itself must be finite by Proposition 2.4.

(b) Now assume $\sum_{j=1}^r m_j=0$ in (3.2). Then PZ preserves (3.2), and applying the last part of the argument in (a) above to the Zariski closure of $[PZ\cup(-1)PZ]$ (which we can think of as a subgroup of $\mathrm{SL}(2,\mathbb{C})$ and which also preserves f), we see that PZ is finite.

Conversely, any subgroup $Z_0\subset\mathrm{GL}(2,\mathbb{C})$, for which the projectivization PZ_0 is finite, preserves the rational function $f_\lambda(x,y)$ where $f_\lambda=(\prod g^*x)/(\prod g^*(x-\lambda y))$ and the products are taken over $g\in PZ_0$. We then choose the parameter $\lambda\in\mathbb{C}\backslash\{0\}$ so that f_λ is nontrivial (we must be careful on this point since, for example, $(\prod g^*x)/(\prod g^*y)=1$ for g in the two-element group $\{\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right),\left(\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right)\}$). □

COROLLARY 3.4 (Baider). *The Ziglin subgroups of* $\mathrm{SL}(2,\mathbb{C})$ *are precisely those that can be conjugated to a subgroup of one of the following groups*:

(1) $T(n)\cap\mathrm{SL}(2,\mathbb{C})$,

(2) $D(n)\cap\mathrm{SL}(2,\mathbb{C})$,

(3) *The finite groups.*

Proof. $T(m, n) \cap \mathrm{SL}(2, \mathbb{C}) \subset T(m-n) \cap \mathrm{SL}(2, \mathbb{C})$ since $\alpha\beta = 1$. □

The format of the next theorem follows that in the algorithm due to Kovacic [10] that we explain in § 4. Moreover, the groups G_M and G_N will be the respective monodromy groups of the differential equations (4.1) and (4.2) in that section.

THEOREM 3.5. *For $\theta_j \in \mathbb{C}\backslash\{0\}$ let $N = \{N_j\}_{j=1}^k \subset \mathrm{GL}(2, \mathbb{C})$ and $M = \{M_j = \theta_j N_j\}_{j=1}^k$ generate the respective groups G_N and G_M. Then G_N and G_M are simultaneously reducible, diagonalizable, or* DP*-groups. Moreover, they are simultaneously finite if and only if all θ_j are roots of unity. Now assume that $G_N \subset \mathrm{SL}(2, \mathbb{C})$. Then*

(a) *If G_N is reducible but not diagonalizable, under a conjugacy h for which $h^{-1}N_j h = \left(\begin{smallmatrix} \lambda_j & 0 \\ * & \lambda_j^{-1} \end{smallmatrix}\right)$, then G_M is Ziglin if and only if all $\theta_j\lambda_j$ are roots of unity.*

(b) *If G_N is diagonalizable with $h^{-1}N_j h = \left(\begin{smallmatrix} \lambda_j & 0 \\ 0 & \lambda_j^{-1} \end{smallmatrix}\right)$, then G_M is Ziglin if and only if all $\theta_j\lambda_j$ are roots of unity or there are integers m and n such that $(\theta_j)^{m+n} \cdot (\lambda_j)^{m-n} = 1$ for all j.*

(c) *If G_N is a* DP*-group but is not reducible, with $h^{-1}N_j h = \left(\begin{smallmatrix} \lambda_j & 0 \\ 0 & \lambda_j^{-1} \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} 0 & \rho_j \\ -\rho_j^{-1} & 0 \end{smallmatrix}\right)$, then G_M is Ziglin if and only if all θ_j are roots of unity.*

(d) *If G_N is finite then G_M is Ziglin.*

(e) *If none of the cases above hold for G_N, then G_M is not Ziglin.*

Proof. The initial statements concerning reducibility, diagonalizability, the DP-structure, and finiteness are clear. Statements (a)-(d) follow from the statement and proof of Theorem 3.3 (in (c) we must use the presence of an element $\left(\begin{smallmatrix} 0 & \rho_j \\ -\rho_j^{-1} & 0 \end{smallmatrix}\right)$).

For (e) assume that G_M is Ziglin. Then the assumptions on G_N imply by Proposition 2.3 that the Zariski closure $\bar{G}_N = \mathrm{SL}(2, \mathbb{C})$ and by Theorem 3.3 that the projectivization PG_M is finite. We can then readily construct a homogeneous polynomial that vanishes on PG_M and hence on G_M and G_N but not on $\mathrm{SL}(2, \mathbb{C})$, contradicting the fact that $\bar{G}_N = \mathrm{SL}(2, \mathbb{C})$. □

PROPOSITION 3.6. *If all $\{\theta_j\}_{j=1}^k$ are roots of unity, then G_N is Ziglin if and only if G_M is Ziglin.*

Proof. There is an integer d so that $(\theta_j)^d = 1$ for all j. Now raise the respective polynomials in the proof of Theorem 3.3 to the d-th power. □

For many applications the assumption on the $\{\theta_j\}$ in Proposition 3.6 is natural.

4. The algorithm. Here we show how to determine whether the monodromy group $G_M \subset \mathrm{GL}(2, \mathbb{C})$ of a second-order Fuchsian equation

$$z'' + p(x)z' + q(x)z = 0, \qquad ' = \frac{d}{dx}, \tag{4.1}$$

on $\mathbb{P}^1$ is a Ziglin group. This is done with the aid of an algorithm due to Kovacic [10] that decides which of the cases (a)-(e) of Theorem 3.5 holds for (4.1) in its normal form,

$$y'' = r(x)y, \quad r(x) = -[q(x) - (\tfrac{1}{4})p^2(x) - (\tfrac{1}{2})p'(x)], \quad ' = (d/dx), \tag{4.2}$$

which, we note, is also Fuchsian. We need to establish some notation.

In (4.1) we have

$$p(x) = \sum_{j=1}^k \frac{A_j}{(x-a_j)}, \quad q(x) = \sum_{j=1}^k \frac{B_j}{(x-a_j)^2} + \sum_{j=1}^k \frac{C_j}{(x-a_j)}, \quad \sum_{j=1}^k C_j = 0. \tag{4.3}$$

This implies that in (4.2) we have

$$r(x) = \sum_{j=1}^k \frac{\beta_j}{(x-a_j)^2} + \sum_{j=1}^k \frac{\delta_j}{(x-a_j)}, \qquad \sum_{j=1}^k \delta_j = 0, \tag{4.4}$$

where, on setting $A_\infty = \sum_{j=1}^k A_j$, $B_\infty = \sum_{j=1}^k (B_j + C_j a_j)$, and $\beta_\infty = \sum_{j=1}^k (\beta_j + \delta_j a_j)$,

$$\begin{aligned} \beta_j &= (\tfrac{1}{4})[(1-A_j)^2 - 4B_j - 1], \\ \delta_j &= -C_j + (\tfrac{1}{2})A_j\left[\sum_{i\neq j}^{k} \frac{A_i}{(a_j - a_i)}\right], \\ \beta_\infty &= (\tfrac{1}{4})[(1-A_\infty)^2 - 4B_\infty - 1]. \end{aligned} \tag{4.5}$$

Note that the characteristic exponents of (4.2) are

$$\begin{aligned} &\tau_j^{\pm} = \tfrac{1}{2}[1 \pm \{1 + 4\beta_j\}^{1/2}] = \tfrac{1}{2}[1 \pm \{(1-A_j)^2 - 4B_j\}^{1/2}] \quad \text{at } a_j, \\ &\tau_\infty^{\pm} = \tfrac{1}{2}[-1 \pm \{1 + 4\beta_\infty\}^{1/2}] = \tfrac{1}{2}[1 \pm \{(1-A_\infty)^2 - 4B_\infty\}^{1/2}] \quad \text{at } \infty. \end{aligned} \tag{4.6}$$

Let $\Lambda = \{a_1, \cdots, a_k\}$ be the finite poles of $p(x)$ and $q(x)$ as in (4.3), set $X = \mathbb{P}^1 \backslash (\Lambda \cup \{\infty\})$, and fix $x_0 \in X$. For each point $a_j \in \Lambda$ let γ_j be a positively oriented loop in X based at x_0 that encircles only a_j; then $\pi_1(X, x_0)$ is freely generated by $\{\gamma_j\}_{j=1}^k$. Also, let γ_∞ be a corresponding loop around ∞ satisfying $(\prod_{j=1}^k \gamma_j)\gamma_\infty = I$. Then the monodromy representation $\rho_M: \pi_1(X, x_0) \to \mathrm{GL}(2, \mathbb{C})$ of (4.1) with $M_j = \rho_M(\gamma_j)$, $M_\infty = \rho_M(\gamma_\infty)$, satisfies $(\prod_{j=1}^k M_j)M_\infty = I$, and $\{M_j\}_{j=1}^k$ generates the monodromy group G_M of (4.1).

The poles of $r(x)$ in (4.4) are a subset of $\Lambda \cup \{\infty\}$. The monodromy representation ρ_N of (4.2) has range $G_N \subset \mathrm{SL}(2,\mathbb{C})$ and is generated by $\{N_j = \rho_N(\gamma_j)\}_{j=1}^k$. Moreover (see [3, § 6]),

$$M_j = \theta_j N_j, \quad \text{where } \theta_j = \exp\left[-\frac{1}{2}\int_{\gamma_j^{-1}} p\right] = \exp(\pi i A_j). \tag{4.7}$$

The algorithm below consists of three successive cases. Each case is examined in turn, and lack of success in determining a solution in all three cases will correspond to (e) in Theorem 3.5 (with the first case covering both (a) and (b) of that theorem). Throughout we let G_D denote the differential Galois group of (4.2).

Case I (The reducible-diagonalizable case). The algorithm is phrased in terms of the "modified" characteristic exponents

$$\begin{aligned} &\alpha_j^{\pm} = \tau_j^{\pm} \quad \text{if } \beta_j \neq 0; \quad \alpha_j^{\pm} = 1 \quad \text{if } \beta_j = 0,\ \delta_j \neq 0; \quad \alpha_j^{\pm} = 0 \quad \text{if } \beta_j = 0 = \delta_j, \\ &\alpha_\infty^{\pm} = \tau_\infty^{\pm} + 1 \quad \text{if } \beta_\infty \neq 0; \quad \alpha_\infty^{+} = 0 \quad \text{and} \quad \alpha_\infty^{-} = 1 \quad \text{if } \beta_\infty = 0. \end{aligned} \tag{4.8}$$

THEOREM 4.9. *The following two statements are equivalent*:

(a) G_M, G_N, *and* G_D *are simultaneously reducible.*

(b) *There is a solution of* (4.2) *of the form* $y = \exp(\int \theta)$, *with* $\theta \in \mathbb{C}(x)$, *which is necessarily a common eigenvector for* G_N.

Moreover, there is a solution as in (b) *if and only if*

(1) *There is a choice* $s(j)$ *and* $s(\infty)$ *of a plus or minus sign so that*

$$d = \left[\alpha_\infty^{s(\infty)} - \sum_{j=1}^{k} \alpha_j^{s(j)}\right] \text{ is a nonnegative integer;}$$

(2) *There is a unique monic, degree* d *polynomial* P (*which can be found by the method of undetermined coefficients*) *satisfying*

$$P'' + 2\omega P' + (\omega' + \omega^2 - r)P = 0, \quad \text{where } \omega = \omega(x) = \sum_{j=1}^{k} \frac{\alpha_j^{s(j)}}{(x - a_j)},$$

and

(3) $\theta = \omega + (P'/P)$.

The algorithm above will generate two distinct θ's if and only if the three groups in (a) above are simultaneously diagonalizable. If there are no such solutions these groups are irreducible.

Proof. The statement in (a) is just a combination of Proposition 2.2(e) and the first part of Theorem 3.5, and similarly for simultaneous diagonalizability of these groups. For the equivalence of (a) and (b) see [10, pp. 7, 8]. The algorithm given in steps (1)-(3) for finding solutions of the form $y = \exp(\int \theta)$ with $\theta \in \mathbb{C}(x)$ is a restatement of [10, pp. 11, 12] for the Fuchsian case. The remainder of the theorem can be found in the proof in Kovacic [10, pp. 15-17], with the uniqueness of P in (2) being shown by calculating the Wronskian of the two solutions. □

Such reducibility criteria have a long history; for example, see [15, pp. 176-178], which was published in 1895. For the Fuchsian case with three regular singular points (e.g., the hypergeometric equation), simpler formulations are available (see [2, Thm. 2.24]). It should be noted that Kovacic's algorithm is not restricted to the Fuchsian case of (4.2).

COROLLARY 4.10. *Assume the reducibility algorithm implicit in steps* (1)-(3) *of Theorem* 4.9 *has yielded* (1) *only one solution, or* (2) *two independent solutions of* (4.2) *with the form* $y = \exp(\int \theta)$, *where*

$$\theta = \sum_{j=1}^{k} \frac{\alpha_j^{s(j)}}{(x - a_j)} + (P'/P).$$

Then:

(1) *In case* (1) *the monodromy group* G_M *of* (4.1) *is a Ziglin group if and only if each of the numbers* $[A_j + 2\alpha_j^{s(j)}]$, $j = 1, 2, \cdots, k$, *is rational.*

(b) *In case* (2) G_M *is a Ziglin group if and only if there are integers* m *and* n *so that* $[(m+n)A_j + 2(m-n)\alpha_j^{s(j)}]$ *is an even integer for* $j = 1, 2, \cdots, k$. *In particular,* G_M *is Ziglin if each of the numbers* $[A_j \pm 2\alpha_j^{s(j)}]$, $j = 1, 2, \cdots, k$, *is rational.*

Proof. In terms of a basis $\{*, y\}$ of germs of solutions of (4.2) at x_0 we have

$$N_j = \begin{bmatrix} \exp(2\pi i \alpha_j^{s(j)}) & 0 \\ * & \exp(-2\pi i \alpha_j^{s(j)}) \end{bmatrix}.$$

The result then follows from (4.7) and Theorem 3.5(a) and (b), respectively. □

Case II (The DP-case). The algorithm is stated in terms of the following sets:

$$\begin{aligned} &E_j = \{2 + e(1 + 4\beta_j)^{1/2} \mid e = 0, \pm 2\} \cap \mathbb{Z} \quad \text{if } \beta_j \neq 0, \\ &E_j = \{4\} \quad \text{if } \beta_j = 0, \quad \delta_j \neq 0, \\ &E_j = \{0\} \quad \text{if } \beta_j = 0 = \delta_j, \end{aligned} \tag{4.11}$$

and

$$\begin{aligned} &E_\infty = \{2 + e(1 + 4\beta_\infty)^{1/2} \mid e = 0, \pm 2\} \cap \mathbb{Z} \quad \text{if } \beta_\infty \neq 0, \\ &E_\infty = \{0, 2, 4\} \quad \text{if } \beta_\infty = 0. \end{aligned} \tag{4.12}$$

THEOREM 4.13. *The following two statements are equivalent*:

(a) G_M, G_N, *and* G_D *are irreducible* (*i.e., Case* I *does not hold*) *but are simultaneously* DP-*groups.*

(b) *There is a solution of* (4.2) *of the form* $y = \exp(\int \omega)$, *where* ω *is algebraic over* $\mathbb{C}(x)$ *of degree* 2, *and Case* I *does not hold.*

Moreover, there is a solution as in (b) *if and only if*

(1) *There is a choice of* $e_j \in E_j$ *and* $e_\infty \in E_\infty$ *which are* not *all even integers so that* $d = \frac{1}{2}[e_\infty - \sum_{j=1}^{k} e_j]$ *is a nonnegative integer*; *and*

(2) *There is a monic, degree d polynomial P (which can be found by the method of undetermined coefficients) satisfying*

$$P'''+3\theta P''+(3\theta^2+3\theta'-4r)P'+(\theta''+3\theta\theta'+\theta^3-4r\theta-2r')P=0,$$

where $\theta=\frac{1}{2}\sum_{j=1}^{k}(e_j/(x-a_j))$.

Specifically, for P as in (2) let $\phi=\theta+(P'/P)$ and choose a solution ω of $\omega^2+\phi\omega+[\frac{1}{2}\phi'+\frac{1}{2}\phi^2-r]=0$; then $y=\exp(\int\omega)$ will be a solution of (4.2) as in (b) above.

Proof. For the equivalence of (a) and (b) see [10, pp. 7, 8]. The remainder is a restatement of [10, p. 18] for the Fuchsian case. □

COROLLARY 4.14. *Assume that* G_D *is irreducible but that the* DP*-algorithm implicit in* (1) *and* (2) *and the final statement of Theorem* 4.13 *results in a solution of the required form. Then the monodromy group* G_M *of* (4.1) *is Ziglin if and only if all* A_j *are rational.*

Proof. This follows from (4.7) and Theorem 3.5(c). □

Remark 4.15. A necessary condition for the algorithm of Theorem 4.13 to give a solution of the required form is that in (4.4) some $\beta_j\neq 0$ (see [10, p. 8]). Thus, if all $\beta_j=0$ we need to examine only Cases I and III.

Case III (The finite case). The algorithm is stated in terms of the following sets, where $n=4$, 6, or 12:

$$F_j(n)=\left\{6+\frac{12e}{n}(1+4\beta_j)^{1/2}\,\middle|\,e=0,\pm1,\cdots,\pm\frac{n}{2}\right\}\cap\mathbb{Z}\quad\text{if }\beta_j\neq 0,$$

$$F_j(n)=\{12\}\quad\text{if }\beta_j=0,\ \delta_j\neq 0, \tag{4.16}$$

$$F_j(n)=\{0\}\quad\text{if }\beta_j=0=\delta_j,$$

and

$$F_\infty(n)=\left\{6+\frac{12e}{n}(1+4\beta_\infty)^{1/2}\,\middle|\,e=0,\pm1,\cdots,\pm\frac{n}{2}\right\}\cap\mathbb{Z}, \tag{4.17}$$

regardless of whether or not $\beta_\infty=0$.

THEOREM 4.18. *Assume that* G_D *is not reducible and not a* DP*-group (i.e., Cases* I *and* II *do not hold). Then the following procedure will determine if* G_D *is finite with all solutions of* (4.2) *being algebraic over* $\mathbb{C}(x)$.

(1) *Let* $n=4$ *and write down all choices of* $f_j\in F_j(n)$ *and* $f_\infty\in F_\infty(n)$ *for which* $d=(n/12)[f_\infty-\sum_{j=1}^{k}f_j]$ *is a nonnegative integer;*

(2) *For each such choice set* $\theta=(n/12)\sum_{j=1}^{k}(f_j/(x-a_j))$, *and with* $S=\Pi(x-a_j)$ (*where the product is taken over only those* a_j *which are poles of* $r(x)$) *determine* (*e.g., by the method of undetermined coefficients*) *if there is a monic, degree d polynomial P such that if we set* $P_n=-P$ *and recursively define*

$$P_{i-1}=-SP_i'+[(n-i)S'-S\theta]P_i-(n-i)(i+1)S^2rP_{i+1}$$

for $i=n,\ n-1,\cdots,0$, *then* $P_{-1}\equiv 0$;

(3) *Repeat, if necessary, steps* (1) *and* (2) *with* $n=6$ *and then with* $n=12$; *and*

(4) *If such a P is found in* (2) *for* $n=4$, 6, *or* 12, *then* G_D *is finite. Moreover, a solution* ω *to the equation* $\sum_{i=0}^{n}((S^iP_i)/(n-i)!)\omega^i=0$ *will give a solution* $y=\exp(\int\omega)$ *to* (4.2).

Proof. See [10, pp. 7, 8 and pp. 22, 23] and recall that (4.2) is Fuchsian. □

COROLLARY 4.19. *Assume* G_D *is irreducible and not a* DP*-group, but is finite. Then* G_M *is Ziglin.*

Proof. The proof is obtained by Theorem 3.5(d). □

Remarks 4.20. A success in the algorithm (1)-(4) of Theorem 4.19 implies G_D is finite and projectively the tetrahedral ($n=4$), octahedral ($n=6$), or iscosahedral

($n=12$) group (see [10, p. 27]). A failure implies that $G_D = \mathrm{SL}(2,\mathbb{C})$ (see [10, p. 7]). Necessary conditions for a success in the algorithm of Theorem 4.19 are that all the characteristic exponents $\tau_j^{\pm}$ and $\tau_\infty^{\pm}$ of (4.6) be rational; that is, all $(1+4\beta_j)^{1/2}$ and $(1+4\beta_\infty)^{1/2}$ are rational (see [10, p. 8]). If this does not hold, only Cases I and II need to be examined.

5. Applications.

Example (a). Let $a(x)$, $b(x)$, and $c(x,y)$ be arbitrary meromorphic functions on $\mathbb{C}^2=\{(x,y)\}$, let $h\in\mathbb{C}$, and consider the analytic set

$$\gamma=\{(x,y)\in\mathbb{C}^2 \,|\, b(x)c(x,y)=h\}. \tag{5.1}$$

As an example, if $b(x)$ is a separable polynomial of positive degree $(2g+1)$ or $(2g+2)$, $h=1$, and $c(x,y)=y^{-2}$, then γ is a punctured algebraic curve of genus g. Returning to generalities, suppose $\Gamma\subset\gamma$ is a Riemann surface on which $a(x)/b(x)$ is finite, and having the property that the projection $\pi(x,y)=x$ of Γ into $\mathbb{C}$ is unbounded. Via the embedding $(x,y)\to(x,0,y,0)$ we may view $E=\{(x,\xi_2,y,\eta_2)\,|\,(x,y)\in\Gamma\}\subset\mathbb{C}^4$ with projection $(x,\xi_2,y,\eta_2)\to(x,y)$ as a rank 2 complex vector bundle over Γ. A holomorphic connection ∇ can then be defined on E through the local coordinate representation

$$d\begin{pmatrix}\xi_2\\ \eta_2\end{pmatrix}=\begin{pmatrix}0 & 1\\ \dfrac{-a(x)}{b(x)} & 0\end{pmatrix}\begin{pmatrix}\xi_2\\ \eta_2\end{pmatrix}dx. \tag{5.2}$$

∇ is designed so as to be the pullback of

$$\xi''+\frac{a(x)}{b(x)}\xi=0,\quad \xi=\xi_2,\quad '=\frac{d}{dx}, \tag{5.3}$$

under the projection $\pi:\Gamma\to\mathbb{C}$ (see [2, § 4]). As a consequence, the monodromy group G_0 of ∇ embeds into the monodromy group G_N of (5.3) (see [4, Prop. 1.1]). If G_N is Ziglin, then Proposition 3.1(b) implies that G_0 is Ziglin.

There are simple instances in which we can view the projection $\pi:\Gamma\to\mathbb{C}$ as a reduction with respect to a finite symmetry group G. For example, if Γ is invariant under each mapping

$$(\exp(2\pi ij/n),(x,y))\to(x,\exp(2\pi ij/n)\cdot y),\qquad j=1,2,\cdots,n,$$

as would be the case if $c(x,y)=y^{\pm n}$, then Theorem 1.2 can be applied (the group action on the fibers of $E\to\Gamma$ being the identity map in the (ξ_2,η_2)-coordinates of (5.2)).

The algorithm of § 4 applies directly to (5.3) when $r(x)=-[a(x)/b(x)]$ is of the form given in (4.4). For example, suppose

$$r(x)=\left[\frac{1}{x^2}+\frac{1}{(x-1)^2}+\frac{1}{(x-2)^2}-\frac{1}{x}+\frac{1}{(x-1)}\right]. \tag{5.4}$$

Then (I) there is no choice of the $\alpha_j^{\pm}=\frac{1}{2}[1\pm\sqrt{5}]$ and $\alpha_\infty^{\pm}=\frac{1}{2}[1\pm\sqrt{17}]$ that makes the d of Theorem 4.9(1) a nonnegative integer; (II) all $E_j=\{2\}=E_\infty$ so that there are no choices of the e_j and e_∞, not all even, with which to construct a nonnegative d as in Theorem 4.13(1); and (III) $\beta_1=1$ implies $(1+4\beta_1)^{1/2}=\sqrt{5}$ is not rational so that Case III does not apply (see the necessary conditions in Remark 4.20). The differential Galois group G_D of (5.3)-(5.4) is then $\mathrm{SL}(2,\mathbb{C})$. Recalling Proposition 2.2(d), we see that by Proposition 3.1(d) and Corollary 3.4 the monodromy group $G_M=G_N$ of (5.3)-(5.4) cannot be Ziglin. If (5.3)-(5.4) is achieved from the ∇ of (5.2) by reduction using a finite group (see the previous paragraph for an example), then by Theorem 1.2 the monodromy group G_0 of ∇ is also not Ziglin.

To fit this into a Hamiltonian context, give $\mathbb{C}^4=\{(x, x_2, y, y_2)\}$ the standard symplectic structure $\omega=dx\wedge dy+dx_2\wedge dy_2$, let $p(x, y)=a(x)(\partial c/\partial y)(x, y)$ and $q(x, y)=b(x)c(x, y)$, and consider the Hamiltonian

$$H(x, x_2, y, y_2)=q(x, y)+\tfrac{1}{2}p(x, y)x_2^2+\tfrac{1}{2}(\partial q/\partial y)(x, y)y_2^2+\mathcal{O}_3(x_2, y_2). \tag{5.5}$$

The associated vector field X_H is tangent to the (x, y)-plane, a phase curve Γ in that plane of energy h is contained in the set γ of (5.1), and the normal variational equation (NVE) along Γ may be identified with (5.2) where (ξ_2, η_2) are the (global) linearized variables associated to (x_2, y_2) (see [2, § 4]). In the context of the previous paragraphs, conclusions about the nonintegrability of X_H can be drawn from Corollary 1.3.

Example (b). For our second example we illustrate how reduction can be combined with Kovacic's algorithm to explicitly compute the monodromy of a holomorphic connection that arises in a Hamiltonian system.

First consider the Fuchsian equation

$$y''+p(x)y'+q(x)y=0, \qquad '=\frac{d}{dx}, \tag{5.6}$$

on $\mathbb{C}$ with

$$\begin{aligned} p(x)&=\left[\frac{1}{x}+\frac{(1/2)}{(x-1)}\right], \\ q(x)&=\left[-\frac{1}{x^2}+\frac{(1/2)}{x}-\frac{(1/2)}{(x-1)}\right], \end{aligned} \tag{5.7}$$

and the associated normal form

$$y''=r(x)y, \quad r(x)=\left[\frac{(3/4)}{x^2}-\frac{(3/16)}{(x-1)^2}+\frac{(3/4)}{x(x-1)}\right]. \tag{5.8}$$

Applying Case I of § 4, we find only two choices of the $\alpha^{\pm}$ for which the d in Theorem 4.9(1) is a nonnegative integer:

$$\begin{aligned} \{\alpha_\infty^+, \alpha_1^+, \alpha_2^+\}&=\{\tfrac{7}{4}, \tfrac{3}{2}, \tfrac{1}{4}\} \quad \text{with } d=0, \\ \{\alpha_\infty^+, \alpha_1^-, \alpha_2^-\}&=\{\tfrac{7}{4}, -\tfrac{1}{2}, \tfrac{1}{4}\} \quad \text{with } d=2. \end{aligned}$$

Both cases lead to the single solution

$$y(x)=x^{3/2}(x-1)^{1/4} \tag{5.9}$$

of (5.8).

From the proof of Corollary 4.10 we see that the monodromy group G_N of (5.8) is generated by $N_1=\left(\begin{smallmatrix}-1 & 0\\ -a & -1\end{smallmatrix}\right)$ and $N_2=\left(\begin{smallmatrix}i & 0\\ b & -i\end{smallmatrix}\right)$. Note that $a\neq 0$ in N_1, since otherwise the distinct eigenvalues of N_2 would imply that G_N was diagonalizable, contrary to the algorithm giving us only the one solution (5.9) in Case I. Now recall from (4.7) that the monodromy group G_M of (5.6) is generated by $M_j=\theta_j N_j$, where $\theta_j=\exp(\pi i A_j)$, $j=1, 2$, with the A_j defined as in (4.3). From $p(x)$ in (5.7) we see that $A_1=1$ and $A_2=\frac{1}{2}$, and so $M_1=-N_1=\left(\begin{smallmatrix}1 & 0\\ a & 1\end{smallmatrix}\right)$ and $M_2=iN_2=\left(\begin{smallmatrix}-1 & 0\\ ib & 1\end{smallmatrix}\right)$.

Equation (5.6) with (5.7) occurs in a Hamiltonian context somewhat as in Example (a). Using the standard symplectic structure on $\mathbb{C}^4$ and

$$H(x_1, x_2, y_1, y_2)=\tfrac{1}{2}(y_1^2+y_2^2)+\tfrac{1}{2}(x_1^2+x_2^2)+\tfrac{2}{3}x_1^3+x_1x_2^2, \tag{5.10}$$

the associated vector field X_H is tangent to the (x_1, y_1)-plane, and there is a phase curve Γ within this plane at energy $h=0$ contained in the algebraic curve defined by

$$y_1^2=x_1^2[-\tfrac{4}{3}x_1-1]. \tag{5.11}$$

The $\mathbb{Z}_2$-action $(x_1, y_1) \to (x_1, -y_1)$ preserves (5.11) and lifts to a symmetry of the NVE ∇ along Γ, where ∇ is the pullback of (5.6)-(5.7) under the reduction mapping $(x_1, y_1) \to -(4/3)x_1 = x$. (Analogous computations are done in [4, Ex. C, pp. 110-112].) This mapping is unbranched over $x = 0$ and branched with order 2 over $x = 1$. Since $a \neq 0$ and $M_2^2 = id$, we see that the monodromy group of ∇ is isomorphic to the infinite cyclic group generated by M_1.

The Hamiltonian (5.10) is but one member of a family that has been extensively studied using Ziglin analysis (see [8, Thm. 4, p. 472] and [13, Cor. 1, p. 266]), which in this case fails to detect the nonexistence of a second independent integral. We might have anticipated some degeneracy in the monodromy of ∇ from the fact that the above Hamiltonian (5.10) and one which is known to be completely integrable have the same linearized equations about solutions in the (x_1, y_1)-plane (see [12, § 5]). However, the integrability status of (5.10) is unknown. For another example of a Hamiltonian system that is completely integrable with infinite cyclic monodromy group for its NVE, see [3, § 5]. For other examples in Ziglin analysis we refer to [3] and [4] and the surveys [14] and [19].

Acknowledgments. The authors thank Michael Singer for lengthy discussions on differential Galois theory, for suggesting that such methods would prove useful in Ziglin analysis, and for communicating a preliminary result which is now subsumed by Theorem 3.5(e). We also thank Alberto Baider for extensive discussions on the Ziglin group classification problem, and for communicating Corollary 3.4, which led directly to Theorem 3.3. Comments on the second example of § 5 by Haruo Yoshida are also gratefully acknowledged. Finally, we thank the Director and staff of the Institute for Mathematics and Its Applications at the University of Minnesota, Minneapolis, for their hospitality while this work was carried out, and also the anonymous referee of this paper for many helpful comments.

REFERENCES

[1] D. G. BABBITT AND V. S. VARADARAJAN, *Local Moduli for Meromorphic Differential Equations*, Astérisque, Vols. 169 and 170, 1989.

[2] A. BAIDER AND R. C. CHURCHILL, *On monodromy groups of second-order Fuchsian equations*, SIAM J. Math. Anal., 21 (1990), pp. 1642-1654.

[3] A. BAIDER, R. C. CHURCHILL, AND D. L. ROD, *Monodromy and non-integrability in complex Hamiltonian systems*, J. Dynamics Differential Equations, 2 (1990), pp. 451-481.

[4] R. C. CHURCHILL AND D. L. ROD, *Geometrical aspects of Ziglin's non-integrability theorem for complex Hamiltonian systems*, J. Differential Equations, 76 (1988), pp. 91-114.

[5] A. DUVAL AND M. LODAY-RICHAUD, *A propos de l'algorithme de Kovacic*, preprint 89-12, Université de Paris-Sud, Paris, 1989.

[6] W. FULTON, *Algebraic Curves*, Benjamin/Cummings, Reading, MA, 1969.

[7] J. GRAY, *Linear Differential Equations and Group Theory from Riemann to Poincaré*, Birkhäuser, Boston, MA, 1986.

[8] H. ITO, *A criterion for non-integrability of Hamiltonian systems with non-homogeneous potentials*, Z. Angew. Math. Phys., 38 (1987), pp. 459-476.

[9] I. KAPLANSKY, *An Introduction to Differential Algebra*, Second ed., Hermann, Paris, 1976.

[10] J. KOVACIC, *An algorithm for solving second order linear homogeneous differential equations*, J. Symbolic Comput., 2 (1986), pp. 3-43.

[11] L. MARKUS, *Group theory and differential equations*, Lecture Notes at the University of Minnesota, Minneapolis, MN, 1959-1960.

[12] A. RAMANI, B. GRAMMATICOS, AND H. YOSHIDA, *Rigorous non-integrability results related to singularity analysis*, in Proceedings of Conference at Como, 1988, A. P. Fordy, ed., Manchester University Press, Manchester, UK, 1990.

[13] D. L. ROD, *On a theorem of Ziglin in Hamiltonian dynamics*, in Hamitonian Dynamical Systems, K. R. Meyer and D. E. Saari, eds., Contemp. Math., 81, American Mathematical Society, Providence, RI, 1988, pp. 259–270.

[14] D. L. ROD AND R. C. CHURCHILL, *On the applicability of Ziglin's non-integrability theorem*, in Proc. of the Workshop on Finite Dimensional Integrable Nonlinear Dynamical Systems, (P. G. L. Leach and W. H. Steeb, eds., World Scientific, Singapore, 1988, pp. 94–109.

[15] L. SCHLESINGER, *Handbuch der Theorie der linearen Differentialgleichungen*, Teubner, Leipzig, 1895.

[16] M. SINGER, *An outline of differential Galois theory*, in Computer Algebra and Differential Equations, E. Tournier, ed., Academic Press, New York, 1990, pp. 3–57.

[17] ———, *Formal Solutions of Differential Equations*, J. Symbolic Comput., 10 (1990), pp. 59–94.

[18] C. TRETKOFF AND M. TRETKOFF, *Solution of the inverse problem of differential Galois theory in the classical case*, Amer. J. Math., 101 (1979), pp. 1327–1332.

[19] H. YOSHIDA, *Ziglin analysis for proving non-integrability of Hamiltonian systems*, in Proc. of the Workshop on Finite Dimensional Integrable Nonlinear Dynamical Systems, P. G. L. Leach and W. H. Steeb, eds., World Scientific, Singapore, 1988, pp. 74–93.

[20a] S. L. ZIGLIN, *Branching of solutions and non-existence of first integrals in Hamiltonian mechanics* I, Funct. Anal. Appl., 16 (1982) , pp. 181–189.

[20b] ———, *Branching of solutions and non-existence of first integrals in Hamiltonian mechanics* II, Funct., Anal. Appl., 17 (1983), pp. 6–17.

[21] J. J. MORALES, *Tecnicas algebraicas para el estudio de la integrabilidad de sistemas Hamiltonianos*, Ph.D. thesis, University of Barcelona, Barcelona, Spain, 1989.

[22] J. J. MORALES AND C. SIMÓ, *Picard-Vessiot theory and Ziglin's theorem*, preprint.

SIAM J. MATH. ANAL.
Vol. 22, No. 6, p. 1803, November 1991

018

ADDENDUM: HYPERGEOMETRIC EXPANSIONS OF HEUN POLYNOMIALS*

E. G. KALNINS† AND W. MILLER, JR.‡

We should have pointed out the earlier papers on Heun functions by Sleeman [3] and Schmidt and Wolf [2]. These authors take a somewhat similar point of view to ours and use the simultaneous separability of a generalized Schrödinger equation in several cooordinate systems to derive integral relations for Heun functions. In [4] and in the present paper we are making clear the geometrical setting of their results and ours: polynomial orthogonal bases on the n-sphere characterized as eigenfunctions of commuting sets of self-adjoint symmetry operators.

REFERENCES

[1] E. G. KALNINS AND W. MILLER, JR., *Hypergeometric expansions of Heun polynomials*, SIAM J. Math. Anal., 22 (1991), pp. 1450–1459.

[2] D. SCHMIDT AND G. WOLF, *A method of generating integral relations by the simultaneous separability of generalized Schrödinger equations*, SIAM J. Math. Anal., 10 (1979), pp. 823–838.

[3] B. D. SLEEMAN, *Non-linear integral equations for Heun functions*, Proc. Edinburgh Math. Soc., 16 (1969), pp. 281–289.

[4] E. G. KALNINS, W. MILLER, AND M. V. TRATNICK, *Families of orthogonal and biorthogonal polynomials on the n-sphere*, SIAM J. Math. Anal., 22 (1991), pp. 272–294.

* Received by the editors August 19, 1991; accepted for publication on August 20, 1991.

† Department of Mathematics and Statistics, University of Waikato, Hamilton, New Zealand.

‡ School of Mathematics, 127 Vincent Hall, University of Minnesota, Minneapolis, Minnesota 55455.