SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0001 $01.00/0

ANALOGIES FROM COMPLEX ANALYSIS AND HEAT CONDUCTION FOR WAVE PROPAGATION*

EDWARD G. DUNNE† AND DALE H. MUGLER†

Abstract. D. V. Widder [Arch. Rational Mech. Anal., 21 (1966), pp. 108–119] showed that solutions of the heat equation share many of the properties of solutions of Laplace's equation, i.e., of analytic functions. In this paper, we show that those analogues extend to solutions of the wave equation. They include representation of a solution as a convolution transform or in a series of polynomials, and extend Widder's list of properties to include the wave equation with remarkable similarity.

1. Introduction. The theory of analytic functions is so closely connected with Laplace's equation that it has been likened to the theory of properties of solutions of that partial differential equation. In 1975, D. V. Widder [8] connected that theory with the theory of heat conduction by refining a set of analogues (which he had listed earlier in [7]) between analytic functions and solutions of the heat equation. That book is the primary source for this paper, in which we extend those analogies to include solutions of the third classical "differential equation of physics"—the wave equation.

The wave equation in one space variable,

$$\frac{\partial^2}{\partial x^2} u(x, t) = \frac{\partial^2}{\partial t^2} u(x, t),$$

is the form considered in this paper. As a hyperbolic differential equation, it falls into a different class from the equations considered by D. V. Widder, yet we shall show that the analogies may be extended to the solutions of this equation in a fairly natural way. We will refer to the entire class of solutions by means of the following definition.

DEFINITION. Let S be an arbitrary region of the (x, t)-plane. Then

$$u(x, t) \in W \quad \text{in } S \quad \leftrightarrow \quad u_{xx}(x, t) = u_{tt}(x, t) \quad \text{in } S. \tag{1}$$

We shall use Widder's symbolism as much as possible, to make it easier to recognize analogies. For example, the topic in the second section is the representation of solutions in terms of a series of polynomials, and all the polynomials used will be connected with the letter v. The analogue of the source solution defined in § 3 will similarly be referred to as $k(x, t)$. The Appell transformation and the resulting associated functions used in the inverse expansion described in § 4, will again be connected with the letter w. The connection between those associated functions and the distributions labeled with an ω will be fully explained at the end of § 5, where the topic of generating functions leads to a natural question of how the two forms are related. Any part of each analogy may be found in the final summary section which lists the table used by D. V. Widder along with an extra column for the set of analogues for the wave equation.

There are a variety of transformations which map solutions to solutions in class W. We list a few here that are used in what follows.

1) *Integration with respect to a parameter,*

$$u(x, t, y) \in W \quad \text{for } a \leqq y \leqq b \quad \rightarrow \quad \int_a^b u(x, t, y)\, dy \in W.$$

* Received by the editors November 11, 1979, and in revised form April 20, 1981. Supported in part by National Science Foundation-U.R.P. research grant SPI-7827194.

† Department of Mathematics, University of Santa Clara, Santa Clara, California 95053.

2) *Finite convolution,*

$$u(x, t) \in W \quad \rightarrow \quad \int_0^t u(x, t-y)g(y)\, dy \in W \quad \text{if } u(x, 0) \equiv 0 \text{ and } u_t(x, 0) \equiv 0.$$

Both 1) and 2) follow from Leibniz's rule.

3) *Component multiplication,*

$$u(x, t) = \phi(x+t) + \psi(x-t) \in W,\ v(x, t) = \alpha(x+t) + \beta(x-t) \in W$$
$$\rightarrow \quad w(x, t) = \alpha(x+t)\phi(x+t) + \beta(x-t)\psi(x-t) \in W,$$

which can be verified by direct evaluation of the derivatives of $w(x, t)$.

2. Series expansion in wave polynomials. It is a classical result that any analytic function $f(x)$ is expressible as a power series of the monomials x^n with coefficients related to the derivatives of the function. Widder has shown that solutions of the heat equation can be expressed in terms of a set of polynomials $v_n(x, t)$ generated by the exponential solution of the heat equation, e^{xz+tz^2}. We seek to extend the analogy to solutions of the wave equation.

We begin by considering the polynomials generated by the exponential solutions to the wave equation

$$e^{\alpha(x+t)} \quad \text{and} \quad e^{\alpha(x-t)},$$

where α is an arbitrary complex parameter. If, instead, we consider the parameter as a complex variable z and expand about the point $z = 0$

$$e^{(x+t)z} = \sum_{n=0}^{\infty} (x+t)^n \frac{z^n}{n!}, \tag{2}$$

we generate a set of polynomials $v_n(x, t) = (x+t)^n$. Similarly, expansion of $e^{(x-t)z}$ yields a second set of polynomials $\bar{v}_n(x, t) = (x-t)^n$. Note that $\bar{v}_n(x, t) = v_n(x, -t)$. We refer to this whole class of polynomials as *wave polynomials.*

To see that these polynomials are solutions of the wave equation, we differentiate directly. The first partials with respect to x,

$$\frac{\partial}{\partial x} v_n(x, t) = n(x+t)^{n-1} = nv_{n-1}(x, t) \quad \text{and} \quad \frac{\partial}{\partial x} \bar{v}_n = n\bar{v}_{n-1},$$

exhibit the same pattern as do the first partials for the monomials and the heat polynomials. Further derivatives are

$$\frac{\partial}{\partial t} v_n = nv_{n-1} \quad \text{and} \quad \frac{\partial^2}{\partial x^2} v_n = n(n-1)v_{n-2} = \frac{\partial^2}{\partial t^2} v_n.$$

The second set of polynomials differs only in that $(\partial/\partial t)\bar{v}_n = -n\bar{v}_{n-1}$. It is easily shown that they too are solutions of the wave equation.

The most important analogy that can be made using wave polynomials concerns their use in series expansions of functions in class W. Such solutions may be rather arbitrary functions, and need not have the differentiability properties possessed by a series of wave polynomials. Thus the representable functions necessarily form a subclass of the entire set of solutions.

We next demonstrate that fairly arbitrary functions can be used to construct solutions of the wave equation. We first transform the equation

$$u_{xx}(x, t) = u_{tt}(x, t) \quad \text{into} \quad \Omega_{\xi\eta}(\xi, \eta) = 0, \tag{3}$$

by the standard change of variables $\xi = x + t$ and $\eta = x - t$ with $\Omega(\xi, \eta) = u((\xi+\eta)/2, (\xi-\eta)/2)$. Solving the transformed equation yields

$$\Omega(\xi, \eta) = \phi(\xi) + \psi(\eta)$$

where ϕ and ψ are arbitrary functions of ξ and η respectively. Then

$$u(x, t) = \phi(x + t) + \psi(x - t). \tag{4}$$

The functions ϕ and ψ need only be twice differentiable, so that $u(x, t)$ will be a solution to the wave equation.

In order to make a polynomial series expansion we restrict our class of solutions. The additional hypothesis that will be necessary here is a property of *all* solutions of the heat and Laplace equations, where solutions are actually analytic.

DEFINITION. Let T be a region of the (x, t)-plane which contains an open segment of the x-axis including $(0, 0)$. We say $u(x, t) \in W^0$ in T if:

(1) $u(x, t) \in W$ in T;

(2) $f(x) = u(x, 0)$ and $g(x) = u_t(x, 0)$ are analytic for $|x| < m$, where $m = \max\{|x \pm t|: (x, t) \in T\}$.

THEOREM 1. *Let* $T = T(\rho) = \{(x, t): |x \pm t| < \rho, t \geqq 0\}$. *Then*

$$\begin{aligned} &u(x, t) \in W^0 \text{ in } T \text{ if and only if} \\ &\quad u(x, t) = \textstyle\sum_{n=0}^{\infty} \{a_n v_n(x, t) + b_n \bar{v}_n(x, t)\} \text{ for } (x, t) \in T. \end{aligned} \tag{5}$$

Proof. If $u(x, t) \in W^0$, we use the analysis above to say that

$$u(x, t) = \phi(x + t) + \psi(x - t).$$

These two functions can be expressed in terms of u and u_t at $t = 0$. We get

$$u(x, 0) = f(x) = \phi(x) + \psi(x), \qquad u_t(x, 0) = g(x) = \phi'(x) - \psi'(x).$$

The second condition implies that

$$\phi(x) - \psi(x) = \int_0^x g(\xi)\, d\xi + K,$$

so that

$$\phi(x) = \frac{1}{2} f(x) + \frac{1}{2} \int_0^x g(\xi)\, d\xi + \frac{1}{2} K \quad \text{and} \quad \psi(x) = \frac{1}{2} f(x) - \frac{1}{2} \int_0^x g(\xi)\, d\xi - \frac{1}{2} K. \tag{6}$$

Using these representations for ϕ and ψ in (4) gives the d'Alembert formula

$$u(x, t) = \frac{1}{2}\{f(x + t) + f(x - t)\} + \frac{1}{2} \int_{x-t}^{x+t} g(\xi)\, d\xi. \tag{7}$$

The forms for ϕ and ψ in (6) show that if f and g are analytic in a neighborhood of the origin, then so are ϕ and ψ. Expanding $\phi(v)$ in a Maclaurin series for the variable $v = x + t$ and $\psi(\bar{v})$ for $\bar{v} = x - t$ gives

$$u(x, t) = \sum_{n=0} a_n v^n + \sum_{n=0} b_n \bar{v}^n.$$

Each series exists since ϕ and ψ are analytic. Writing these series in terms of x and t proves the sufficiency.

Since the series converges uniformly in the indicated region, the necessity is a result of the wave polynomials all being in W^0.

Now that we have shown existence of a series expansion in wave polynomials for a function $u(x, t)$ in class W^0, we give an explicit formula for the coefficients in terms of the derivatives of $u(x, t)$. First, we consider the d'Alembert formula (7) for the wave equation with Cauchy data $u(x, 0) = f(x)$ and $u_t(x, 0) = g(x)$. From (4) we can write $u(x, t)$ as $\phi(x+t)+\psi(x-t)$. Examination of (6) and (7) gives for ϕ and ψ

$$\begin{aligned}\phi(x+t) &= \frac{1}{2}f(x+t)+\frac{1}{2}\int_0^{x+t} g(\xi)\,d\xi = \phi(v),\\ \psi(x-t) &= \frac{1}{2}f(x-t)-\frac{1}{2}\int_0^{x-t} g(\xi)\,d\xi = \psi(\bar{v}).\end{aligned} \tag{8}$$

Since both f and g are analytic about the origin, we know that ϕ and ψ also have series expansions:

$$\phi(v) = \sum_{n=0}^{\infty} \phi^{(n)}(0)\frac{v^n}{n!} \quad \text{and} \quad \psi(\bar{v}) = \sum_{n=0}^{\infty} \psi^{(n)}(0)\frac{\bar{v}^n}{n!}.$$

Then, by direct evaluation of the derivatives from (6), we see that

$$\phi'(v) = \frac{1}{2}\{u_x(v, 0)+u_t(v, 0)\}, \qquad \phi''(v) = \frac{1}{2}\{u_{xx}(v, 0)+u_{tx}(v, 0)\} \quad \text{and so on,}$$

$$\psi'(\bar{v}) = \frac{1}{2}\{u_x(\bar{v}, 0)-u_t(\bar{v}, 0)\}, \qquad \psi''(\bar{v}) = \frac{1}{2}\{u_{xx}(\bar{v}, 0)-u_{tx}(\bar{v}, 0)\} \quad \text{and so on.}$$

Hence, we can express $\phi(v)$ and $\psi(\bar{v})$ as

$$\begin{aligned}\phi(v) &= \frac{1}{2}u(0, 0)+\frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{n!}\left(\frac{\partial^n}{\partial x^n}u(0, 0)+\frac{\partial^n}{\partial x^{n-1}\partial t}u(0, 0)\right)v^n,\\ \psi(\bar{v}) &= \frac{1}{2}u(0, 0)+\frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{n!}\left(\frac{\partial^n}{\partial x^n}u(0, 0)-\frac{\partial^n}{\partial x^{n-1}\partial t}u(0, 0)\right)\bar{v}^n.\end{aligned} \tag{9}$$

It then follows directly that $u(x, t)$ has the series expansion

$$u(x, t) = u(0, 0)+\sum_{n=1}^{\infty} a_n v_n + \sum_{n=1}^{\infty} b_n \bar{v}_n, \tag{10}$$

where

$$\begin{aligned}a_n &= \frac{1}{2\cdot n!}\left(\frac{\partial^n}{\partial x^n}u(0, 0)+\frac{\partial^n}{\partial x^{n-1}\partial t}u(0, 0)\right),\\ b_n &= \frac{1}{2\cdot n!}\left(\frac{\partial^n}{\partial x^n}u(0, 0)-\frac{\partial^n}{\partial x^{n-1}\partial t}u(0, 0)\right),\end{aligned} \qquad n \geqq 1.$$

One should note that $a_n = b_n$ for all $n \geqq 1$ if and only if $g(x) = u_t(x, 0)$ is identically zero.

The region of convergence for the wave polynomial series is determined by the strips of convergence for $\phi(x+t)$ and $\psi(x-t)$. The expansion for the first converges for $|x+t| < \rho_1$, where

$$\frac{1}{\rho_1} = \overline{\lim_{n\to\infty}}\,|a_n|^{1/n},$$

with a_n as defined above. Similarly, $\psi(x-t)$ converges in $|x-t|<\rho_2$ with

$$\frac{1}{\rho_2}=\overline{\lim_{n\to\infty}}\,|b_n|^{1/n}.$$

The region of convergence for the series expansion of $u(x, t)$ is the intersection of these two strips. Also, since there is only physical significance for $t\geqq 0$, only the upper half-plane is usually considered. That leaves

$$T(\rho_1, \rho_2)=|x+t|<\rho_1\cap|x-t|<\rho_2\cap t\geqq 0$$

for the region of convergence. Graphically the region is as shown in Fig. 1.

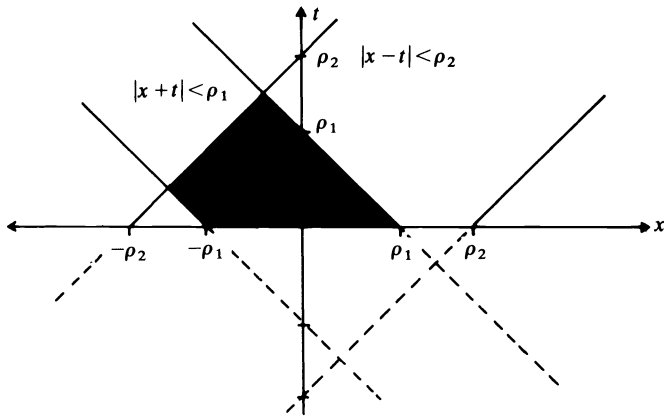

FIG. 1

If t is not restricted to nonnegative values, the region is a rectangle. If $u_t(x, 0)$ is identically zero, or if $\rho_1=\rho_2$, then the region is a triangle in the upper half-plane (see Fig. 2),

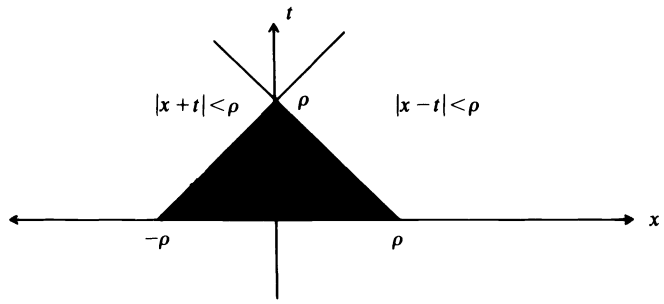

FIG. 2.

Note that the boundaries $x+t=c_1$ and $x-t=c_2$ are characteristics of the wave equation (1).

Example. It is instructive to work through an example which will demonstrate some of the properties given above. Start with Cauchy data $u(x, 0)=2\sin x$, $u_t(x, 0)=6x^2$. By the d'Alembert formula (7), when we separate the solution into its two parts, we are left with

(A) $$\phi(v)=\sin v+v^3 \quad\text{and}\quad \psi(\bar{v})=\sin\bar{v}+\bar{v}^3.$$

We may differentiate directly and use formulas (A) as well as $u(0, 0) = 0$ to find $a_1 = 1$, $a_3 = 5/3!$, and $a_{2n+1} = (-1)^n/(2n+1)!$ for $n > 1$. Evaluating for the other coefficients yields a series representation for the solution as

$$u(x, t) = v_1(x, t) + \frac{5}{6} v_3(x, t) + \sum_{n=2}^{\infty} \frac{(-1)^n}{(2n+1)!} v_{2n+1}(x, t)$$
$$+ \bar{v}_1(x, t) - \frac{7}{6} \bar{v}_3(x, t) + \sum_{n=2}^{\infty} \frac{(-1)^n}{(2n+1)!} \bar{v}_{2n+1}(x, t).$$

This representation may also be found by expanding ϕ and ψ in Maclaurin series and rearranging terms. The formulas for the region of convergence confirm that the representation is valid for all $t \geqq 0$.

3. The source solution. For the heat and wave equations, the source solutions as used below model various properties of the singular function $f(x) = 1/x$, especially with respect to their use in convolution integrals. Widder [8, p. 31] has given some "essential" properties of the "source" solution $k(x, t)$ of the heat equation. They are:

(A) $k(x, t) > 0, t > 0$;
(B) $\lim_{t \to 0} k(x, t) = 0, x \neq 0$;
(C) $\lim_{t \to 0+} k(0, t) = \infty$;
(D) $\int_{-\infty}^{+\infty} k(x, t)\, dx = 1, t > 0$;
(E) $\lim_{t \to 0+} \int_{-\delta}^{+\delta} k(x, t)\, dx = 1, \delta > 0$.

The solution he defines for the heat equation as a source is $k(x, t) = (1/\sqrt{4\pi t})\, e^{-x^2/(4t)}$ if $t > 0$, and 0 if $t \leqq 0$.

The physical interpretation of such a solution of the heat equation is as an impulse of heat energy (of magnitude one) applied to an infinite bar at time $t = 0$. The temperature at $(x, t) = (0, 0)$ is instantaneously infinite, but dissipates as time goes by. The total amount of heat in the bar remains 1.

An analogous "source" for the wave equation would represent an instantaneous point displacement at the origin of the infinite string. Rather than diffusing, however, the displacement is propagated along the string in both directions.

One of the important applications of the source solution of the heat equation is its use as the kernel in the Poisson transform. That is, in a convolution integral with the boundary values of a heat function, it produces the values of that function for positive t. An analogous "source" for the wave equation would initially need to be such that its convolution with a rather arbitrary function would be defined, since the boundary values of a solution to the wave equation need only be in C^2. Thus we would not expect a source for the wave equation to be as nice a function as the one for the heat equation.

To construct a source solution of the wave equation, we will employ the Dirac delta "function" $\delta(x)$, which is actually a distribution. We will discuss later how it may be considered as a limit of a sequence of wave functions. Recall that it may be thought of as a function which is identically zero except at the origin, where it is infinite in such a way that $\int_{-\infty}^{\infty} \delta(x)\, dx = 1$. We define

$$k(x, t) = \frac{1}{2}[\delta(x + t) + \delta(x - t)] \tag{11}$$

as the wave source solution. This models the physical situation that half the displacement travels down the positive x-axis and half down the negative x-axis, both with speed of propagation equal to one. For example, at time $t = 1$, $x + t$ and $x - t$ are zero at -1 and $+1$, respectively.

As defined above, $k(x, t)$ satisfies all the requirements for a source, although it must be interpreted as a distribution or measure. Concerning Widder's properties (A)–(E) listed earlier, the first condition must be modified slightly to be (A) $k(x, t) \geqq 0$, $t > 0$. But properties (B) and (C) follow from the definition of the delta function, and properties (D) and (E) follow from the theory of distributions.

The use of the source as a kernel in a convolution representation integral can be immediately recognized for this $k(x, t)$, although we shall shortly expand and justify this representation using another approach. The property of the delta function that is used is that

$$\int_{-\infty}^{\infty} \delta(x_0 - x) f(x)\, dx = f(x_0).$$

Thus, for a rather arbitrary function $f(x)$,

$$k(x, t) * f(x) = \int_{-\infty}^{\infty} k(x - y, t) f(y)\, dy = \frac{1}{2} \int_{-\infty}^{\infty} [\delta(x - y + t) + \delta(x - y - t)] f(y)\, dy$$

$$= \frac{1}{2}[f(x + t) + f(x - t)],$$

which is known to be the general form of a wave function when $u_t(x, 0) \equiv 0$.

To expand the use of $k(x, t)$ to a representation integral for the general solution of the wave equation, we begin with a rather classical method called the "method of descent". The general form of the solution to the wave equation in *two* space variables is [4, p. 205]

$$u(t, x_1, x_2) = \frac{1}{2\pi} \frac{\partial}{\partial t} \int_R \int \frac{f(x_1 + \xi_1, x_2 + \xi_2)}{\sqrt{t^2 - \xi_1^2 - \xi_2^2}}\, d\xi_1\, d\xi_2 + \frac{1}{2\pi} \int_R \int \frac{g(x_1 + \xi_1, x_2 + \xi_2)}{\sqrt{t^2 - \xi_1^2 - \xi_2^2}}\, d\xi_1\, d\xi_2, \tag{12}$$

where R is $\xi_1^2 + \xi_2^2 \leqq t^2$. Consider the second integral alone, and suppose that g depends only on its first argument. The integral then becomes

$$\frac{1}{2\pi} \int_{-t}^{t} g(x_1 + \xi_1) \left\{ \int_{-\sqrt{t^2 - \xi_1^2}}^{\sqrt{t^2 - \xi_1^2}} \frac{d\xi_2}{\sqrt{t^2 - \xi_1^2 - \xi_2^2}} \right\} d\xi_1.$$

Let $z = \xi_2$ and $y = x + \xi_1$. The limits $-t < \xi_1^1 < t$ become $x - t < y < x + t$, and the integral becomes

$$\frac{1}{2\pi} \int_{x-t}^{x+t} g(y) \left\{ \int_{-\sqrt{t^2 - (y-x)^2}}^{\sqrt{t^2 - (y-x)^2}} \frac{dz}{\sqrt{t^2 - (y - x)^2 - z^2}} \right\} dy. \tag{13}$$

Define

$$K(x, t) = \begin{cases} \dfrac{1}{2\pi} \displaystyle\int_{-\sqrt{t^2 - x^2}}^{\sqrt{t^2 - x^2}} \frac{dz}{\sqrt{t^2 - x^2 - z^2}} & \text{for } |x| < t, \\ 0 & \text{for } |x| > t. \end{cases} \tag{14}$$

Then the integral (13) may be written as a convolution integral, $\int_{x-t}^{x+t} g(y) K(y - x, t)\, dy$, where $g(y) = u_t(y, 0)$. Since $-t < y - x < t$, and $K(y - x, t) = 0$ for y outside of this domain, the above integral may be seen as a convolution integral over the real

line,

$$\int_{-\infty}^{\infty} g(y)K(y-x,t)\,dy.$$

Since the first integral may be rewritten in exactly the same way, it follows from (12) that

$$u(x,t)=\frac{\partial}{\partial t}\int_{-\infty}^{\infty} f(y)K(y-x,t)\,dy+\int_{-\infty}^{\infty} g(y)K(y-x,t)\,dy. \tag{15}$$

The definition of $K(x,t)$ in (14) is as an improper integral, which may be evaluated to show that in actuality $K(x,t)$ is 1 for $|x|<t$, but is 0 otherwise. This evaluation may be used to write $K(x,t)$ in terms of the Heaviside function $Y(x)$, which is defined by

$$Y(x)=\begin{cases}1 & \text{if } x\leqq 0,\\ 0 & \text{if } x>0.\end{cases}$$

The resulting expression for $K(x,t)$ is

$$K(x,t)=\frac{1}{2}[Y(x+t)-Y(x-t)]. \tag{16}$$

As a derivative distribution, it is well known that $Y'(x)=\delta(x)$. Note now that the derivative of the first integral in (15) is with respect to t, and that if we differentiate (16) we obtain

$$\frac{\partial}{\partial t}K(x,t)=\frac{\partial}{\partial t}\frac{1}{2}[Y(x+t)-Y(x-t)]=\frac{1}{2}[\delta(x+t)+\delta(x-t)]=k(x,t).$$

The representation for $u(x,t)$ now obtained from (15) is a convolution form of the d'Alembert formula (7),

$$u(x,t)=f(x)*k(x,t)+g(x)*K(x,t). \tag{17}$$

As before, the functions f and g are initial values of the wave function $u(x,t)$, as $f(x)=u(x,0)$ and $g(x)=u_t(x,0)$.

In contrast to the Poisson transform for the heat equation, this representation is a sum of two terms. This should not be surprising, since solutions of the heat equation are unique given only the functional values on the boundary, while solutions of the wave equation also require the values of the derivative at $t=0$ for uniqueness.

Finally, we return to the definition of $k(x,t)$ as a distribution. Using the property of the delta function that $\int_{-\infty}^{\infty} f(x)\delta(x)\,dx=f(0)$, $\delta(x)$ may be identified with the sequences of C^∞ functions $\{f_n(x)\}$ such that

$$\lim_{n\to\infty}\int_{-\infty}^{\infty} f(x)f_n(x)\,dx=f(0).$$

The sequence $\{\sqrt{n/(2\pi)}\,e^{-nx^2}\}$ which is frequently identified as such a sequence is also connected with Widder's heat source solution. Although not a sequence, that heat source function is such that, for fixed x, the limit as $t\to 0$ is nearly identical to that of the above sequence as $n\to\infty$.

We note that the wave source (11) may be identified with a different such sequence, each function of which is itself a C^∞ wave function. One example is the sequence of

functions $k_n(x, t)$, where

$$k_n(x, t) = \frac{1}{2\pi}\left\{\frac{n}{1+n^2(x+t)^2} + \frac{n}{1+n^2(x-t)^2}\right\}.$$

It may easily be checked that each $k_n(x, t)$ is a solution of the wave equation and that this sequence may be identified with $\frac{1}{2}[\delta(x+t)+\delta(x-t)] = k(x, t)$.

4. The Appell transformation and associated expansions. A function $f(x)$ which is analytic in $|x| < \rho$ can be transformed into a function analytic in $|x| > \rho$ by the inversion operator

$$I(f(x)) = \frac{1}{x} f\left(\frac{1}{x}\right).$$

For the solution of the heat equation, Widder has shown that the analogous operation is the Appell transformation, $\mathrm{Ap}\,(u(x, t)) = k(x, t) \cdot u(x/t, -1/t)$, where $k(x, t)$ is the source solution. If $u(x, t)$ is a solution of the heat equation in the half-plane $-\infty < t < 0$, then $\mathrm{Ap}\,(u(x, t))$ is a solution in the half-plane $0 < t < \infty$.

Our analogous transform for solutions of the wave equation maps solutions from the triangle $|x \pm t| < \rho$, $t > 0$ to the infinite wedges $|x \pm t| > \rho$, $t > 0$. We define the Appell transformation for functions of class W by

$$\mathrm{Ap}\,(u(x, t)) = \mathrm{Ap}\,(\phi(x+t)+\psi(x-t)) = \frac{1}{x+t}\phi\left(\frac{1}{x+t}\right) + \frac{1}{x-t}\psi\left(\frac{1}{x-t}\right).$$

It is readily checked that $\mathrm{Ap}\,(u(x, t)) \in W$, since it is the sum of adequately differentiable functions of $x+t$ and $x-t$. Furthermore, $\phi(x+t)$ valid for $|x+t| < \rho$ is mapped to $1/(x+t)\phi(1/(x+t))$ valid for $|x+t| > \rho$ in the same way as the inversion operator for analytic functions. Similarly for the region of validity for ψ, $|x-t| < \rho$ is moved to $|x-t| > \rho$. Thus, $\mathrm{Ap}\,(u(x, t))$ is valid in the intersection of these two regions, $|x+t| > \rho \cap |x-t| > \rho = |x \pm t| > \rho = T^{-1}(\rho)$.

When applied to the wave polynomials, the Appell transformation generates a set of associated functions which can be used in series expansions valid outside the region of convergence for a series of wave polynomials. It is immediate that

$$w_n(x, t) = \mathrm{Ap}\,(v_n(x, t)) = \frac{1}{(x+t)^{n+1}} \tag{18}$$

and

$$\bar{w}_n(x, t) = \mathrm{Ap}\,(\bar{v}_n(x, t)) = \frac{1}{(x-t)^{n+1}}.$$

As before, we note $\bar{w}_n(x, t) = w_n(x, -t)$. We expect an associated expansion to have a general form

$$u(x, t) = \sum_{n=0}^{\infty} c_n w_n(x, t) + \sum_{n=0}^{\infty} d_n \bar{w}_n(x, t).$$

We next give a criterion for $u(x, t)$ to have an associated expansion.

THEOREM 2.

$$u(x, t) = \sum_{n=0}^{\infty} c_n w_n(x, t) + \sum_{n=0}^{\infty} d_n \bar{w}_n(x, t) \quad \text{in } x+t > \rho_1 \cap x-t > \rho_2 \tag{19}$$

if and only if

$$u(x,t)=\int_0^\infty e^{-(x+t)y}\theta(y)\,dy+\int_0^\infty e^{-(x-t)y}\lambda(y)\,dy, \tag{20}$$

where $\theta(y)\in\{1,\rho_1\}$ *and* $\lambda(y)\in\{1,\rho_2\}$.

Proof. It is easily verified that

$$w_n(x,t)=\frac{1}{(x+t)^{n+1}}=\frac{1}{n!}\int_0^\infty e^{-(x+t)y}y^n\,dy, \tag{21}$$

with a similar formula for $\bar{w}_n(x,t)$. Also, since $\theta(y)$ and $\lambda(y)$ are analytic, we assume that they have Maclaurin expansions

$$\theta(y)=\sum\frac{c_n}{n!}y^n \quad\text{and}\quad \lambda(y)=\sum\frac{d_n}{n!}y^n,$$

where $c_n=D^n\theta(0)$ and $d_n=D^n\lambda(0)$. We then proceed as follows:

$$u(x,t)=\sum_{n=0}^{\infty}c_nw_n(x,t)+\sum_{n=0}^{\infty}d_n\bar{w}_n(x,t) \tag{22}$$

if and only if

$$u(x,t)=\sum_{n=0}^{\infty}c_n\frac{1}{n!}\int_0^\infty e^{-(x+t)y}y^n\,dy+\sum_{n=0}^{\infty}d_n\frac{1}{n!}\int_0^\infty e^{-(x-t)y}y^n\,dy$$

if and only if

$$u(x,t)=\int_0^\infty e^{-(x+t)y}\left(\sum_{n=0}^{\infty}\frac{c_n}{n!}y^n\right)dy+\int_0^\infty e^{-(x-t)y}\left(\sum_{n=0}^{\infty}\frac{d_n}{n!}y^n\right)dy.$$

We justify the interchange of operations by the growth restrictions on $\theta(y)$ and $\lambda(y)$. We proceed by considering the series for c_n and for d_n separately. The interchange of operations

$$\int_0^\infty e^{-(x+t)y}\left(\sum\frac{c_n}{n!}y^n\right)dy=\sum c_nw_n(x,t) \tag{23}$$

is valid if

$$\int_0^\infty e^{-(x+t)y}\left(\sum\left|\frac{c_n}{n!}\right||y|^n\right)dy<\infty.$$

Now, for *sufficiency* we assume $\theta(y)\in\{1,\rho_1\}$, so that

$$\sum_{n=0}^{\infty}\left|\frac{c_n}{n!}\right|y^n=O(e^{(\rho_1+\varepsilon)|y|}),\qquad |y|\to\infty.$$

That is, the integral in (23) is dominated by

$$M\int_0^\infty e^{-(x+t)y}e^{(\rho_1+\varepsilon)|y|}\,dy,\qquad |y|=y,\quad 0<y<\infty$$

for some M. The integral converges for $x+t>\rho_1+\varepsilon$, hence the first half of (19) holds for $x+t>\rho_1$. Similarly, the second half holds for $x-t>\rho_2$. We have, then, that equation (19) holds for $x+t>\rho_1\cap x-t>\rho_2$.

For *necessity* we again work in two parts. We express $u(x, t) = \phi(x+t) + \psi(x-t)$. We assume that $\phi(x+t) = \sum c_n w_n(x, t)$ converges for $x+t > \rho_1$. When this is so, we know that

$$\overline{\lim_{n\to\infty}} |c_n|^{1/n} \leqq \rho_1. \tag{24}$$

Since

$$\lim_{n\to\infty} \frac{n}{e \cdot (n^n e^{-n}\sqrt{2\pi n})^{1/n}} = 1,$$

the inequality (24) yields

$$\overline{\lim_{n\to\infty}} \frac{n|c_n|^{1/n}}{e \cdot (n^n e^{-n}\sqrt{2\pi n})^{1/n}} \leqq \rho_1 \quad \leftrightarrow \quad \overline{\lim_{n\to\infty}} \frac{n|c_n|^{1/n}}{e \cdot (n!)^{1/n}} \leqq \rho_1,$$

by Stirling's formula for $n!$. An entire function $f(z) = \sum a_n z^n$ has growth $\{1, \rho_1\}$ if and only if $\overline{\lim}_{n\to\infty} n|a_n|^{1/n} \leqq e\rho_1$. Hence, the series

$$\sum_{n=0}^{\infty} \frac{c_n}{n!} y^n \tag{25}$$

has growth $\{1, \rho_1\}$. But, we have defined series (25) to be $\theta(y)$. Therefore, $\theta(y) \in \{1, \rho_1\}$. Applying the same arguments for $\lambda(y) = \sum (d_n/n!) y^n$ we see that $\lambda(y) \in \{1, \rho_2\}$. Thus, the theorem is proved.

Alternatively, the associated functions can be related to the derivatives of the source solution. The analogy from analytic functions is that inverse monomials can be expressed in terms of derivatives of the singular function by

$$w_n(x) = (-1)^n \frac{D^n(1/x)}{n!}. \tag{26}$$

Also, the associated functions for the heat equation are related to the derivatives of the source solution by

$$w_n(x, t) = (-2)^n \frac{\partial^n}{\partial x^n} k(x, t), \tag{27}$$

as Widder has shown [8, p. 168]. We find that we may express associated functions in terms of the source solution for the wave equation. However, to do so involves derivatives of the Dirac δ-function, which is actually a distribution. As these derivatives will also be distributions, we cannot expect them to coincide with the associated functions from the Appell transformation, but they are related in another way, as will be shown later.

The expression for the associated functions in terms of the source solution originates from the biorthogonality property for the v_n and w_n. It is seen that, for analytic functions and for solutions of the heat equation respectively, we have

$$\frac{1}{2\pi i} \oint_{\Gamma} v_m(z) w_n(z)\, dz = \delta_{m,n},$$

$$\frac{1}{n!2^n} \int_{-\infty}^{\infty} v_m(x, t) w_n(x, t)\, dx = \delta_{m,n},$$

where $\delta_{m,n}$ is the Kronecker delta. A natural choice for $\omega_n(x, t)$ as an associated function for the wave equation is

$$\omega_n(x, t) = (-1)^n \frac{D^n\delta(x+t)}{n!}, \qquad \bar{\omega}_n(x, t) = (-1)^n \frac{D^n\delta(x-t)}{n!}. \tag{28}$$

Following Lighthill [5, pp. 16–21] and Arsac [1, p. 80], we see that $D^n\delta(x)$ is defined by the convolution integral

$$\int_{-\infty}^{\infty} \phi(x) D^n\delta(x)\, dx = \int_{-\infty}^{\infty} (-1)^n \delta(x)\phi^{(n)}(x)\, dx. \tag{29}$$

Hence, for the biorthogonality property we get

$$\begin{aligned}
\int_0^{\infty} v_m(x, t)\omega_n(x, t)\, dx &= \int_{-\infty}^{\infty} (x+t)^m \frac{D^n\delta(x+t)}{n!}(-1)^n\, dx \\
&= \int_{-\infty}^{\infty} \frac{(-1)^n}{n!}\delta(x+t)(-1)^n \frac{\partial^n}{\partial x^n}(x+t)^m\, dx \\
&= \begin{cases} 0, & m > n, \text{ since the derivative is identically zero,} \\ 1, & m = n, \text{ since the integral of the } \delta\text{-function over the real axis is defined to be 1,} \\ 0, & m < n, \text{ since the integral is zero by convolution with } \delta(x-(-t)). \end{cases} \qquad \text{(30a)} \\
&= \delta_{m,n}
\end{aligned}$$

The same arguments hold for the biorthogonality of $\bar{v}_m(x, t)$ and $\bar{\omega}_n(x, t)$ so that

$$\int_{-\infty}^{\infty} \bar{v}_m(x, t)\bar{\omega}_n(x, t)\, dx = \delta_{m,n}. \tag{30b}$$

5. Generating functions. For analytic functions, the "singular function" provides a generating function for two series involving the monomials and the inverse monomials. For example, the series involving only the inverse monomials has the form

$$f(x-r) = \frac{1}{x-r} = \sum_{n=0}^{\infty} w_n(x) r^n = \sum_{n=0}^{\infty} \frac{r^n}{x^{n+1}}.$$

Widder has shown that the heat polynomials and the associated functions are generated by the source solution in much the same way (see § 6, Table 1, no. 8).

In the case of the wave equation, our source solution is a distribution which does not easily generate a series of ordinary functions. However, we shall now exhibit a wave function which is actually a close analogue of the generating function used by Widder, and show that it generates series involving the wave polynomials and associated functions in an analogous way. At the start, we note that by associated functions in this case we mean the functions such as $w_n(x, t) = (x+t)^{-(n+1)}$ and not the distributions $\omega_n(x, t)$. We will shortly return to these two forms and relate each distribution to its respective function via the source solution.

For the wave equation, we use the function

$$W_0(x, t) = w_0(x, t) + \bar{w}_0(x, t)$$

as a generating function. This is analogous to the generating function used by Widder, since the source solution $k(x, t)$ for the heat equation actually equals $w_0(x, t)$. In our case, the series involving the associated functions has the form

$$W_0(x-r, t) = \sum_{n=0}^{\infty} w_n(x, t) r^n + \sum_{n=0}^{\infty} \bar{w}_n(x, t) r^n,$$

which is valid in the region $|r| < |x \pm t|$. The second series which also involves the wave polynomials has the form

$$W_0(y-x, s-t) = \sum_{n=0}^{\infty} v_n(x, t) w_n(y, s) + \sum_{n=0}^{\infty} \bar{v}_n(x, t) \bar{w}_n(y, s),$$

convergent for $|x+t| < |y+s|$, $|x-t| < |y-s|$.

We now wish to relate the two separate forms of the associated functions. One is derived from the Appell transform

(a) $$w_n(x, t) = \frac{1}{(x+t)^{n+1}}, \qquad \bar{w}_n(x, t) = \frac{1}{(x-t)^{n+1}},$$

and was used in § 4 to provide a series expansion of a function in W^0. The other is analogous to derivatives of the source solution

(b) $$\omega_n(x, t) = \frac{(-1)^n D^n \delta(x+t)}{n!}, \qquad \bar{\omega}_n(x, t) = \frac{(-1)^n D^n \delta(x-t)}{n!},$$

and was used in the previous section for the biorthogonality property.

First note that the monomials and the heat polynomials are given by $v_n(x) = 1/z * z^n$ and $v_n(x, t) = k(x, t) * x^n$, respectively. The wave polynomials also satisfy this kind of relation, in the form $v_n(x, t) + \bar{v}_n(x, t) = 2k(x, t) * x^n$. In our case, $k(x, t)$ may be convoluted with any (not necessarily C^∞) function so that the associated functions may also be represented in this form:

$$w_n(x, t) = \frac{1}{(x+t)^{n+1}} = \delta(x+t) * \frac{1}{x^{n+1}}.$$

But $1/x^{n+1} = (-1)^n/n! D^n(1/x)$ so that

$$w_n(x, t) = \delta(x+t) * \left\{ \frac{(-1)^n}{n!} D^n\left(\frac{1}{x}\right) \right\}.$$

However, such distributions satisfy $D^n(u * v) = (D^n u) * v = u * (D^n v)$ [6, p. 166], so that

$$w_n(x, t) = \left\{ \frac{(-1)^n D^n \delta(x+t)}{n!} \right\} * \frac{1}{x}.$$

But this connects the two forms (a) and (b):

$$w_n(x, t) = \omega_n(x, t) * \frac{1}{x}.$$

The arguments follow identically for $\bar{\omega}_n(x, t)$.

In particular, the $W_0(x, t)$ function used as generating function above is related to the source solution directly by

$$W_0(x, t) = 2k(x, t) * \frac{1}{x}.$$

TABLE 1

Analytic functions	*Solutions of the heat equation*	*Solutions of the wave equation*
1. Singular function $f(x)=\frac{1}{x}$	Source solution $k(x,t)=\frac{e^{-x^2/(4t)}}{\sqrt{4\pi t}}$	Source solution $k(x,t)=\frac{1}{2}[\delta(x+t)+\delta(x-t)]$, where δ is the Dirac delta
2. Inversion $1[f(x)]=\frac{1}{x}f\left(\frac{1}{x}\right)$	Appell transformation $\mathrm{Ap}[u(x,t)]=k(x,t)u\left(\frac{x}{t},-\frac{1}{t}\right)$	Appell transformation $\mathrm{Ap}[u(x,t)]=\frac{1}{x+t}\phi\left(\frac{1}{x+t}\right)+\frac{1}{x-t}\psi\left(\frac{1}{x-t}\right)$ $=w_0(x,t)\phi\left(\frac{1}{x+t}\right)+\bar{w}_0(x,t)\psi\left(\frac{1}{x-t}\right)$, with ϕ and ψ as in (2)
3. Monomials $v_n(x)=x^n$	Heat polynomials $v_n(x,t)=n!\sum_{k=0}^{[n/2]}\frac{x^{n-2k}}{(n-2k)!}\frac{t^k}{k!}$	Wave polynomials $v_n(x,t)=(x+t)^n$ and $\bar{v}_n(x,t)=(x-t)^n$
4. Inverse monomials $w_n(x)=I[x^n]=\frac{1}{x^{n+1}}$	Associated functions $w_n(x,t)=\mathrm{Ap}[v_n(x,t)]$ $=k(x,t)v_n(x,-t)t^{-n}$	Associated functions (a) $w_n(x,t)=\frac{1}{(x+t)^{n+1}}=\mathrm{Ap}[v_n(x,t)]$ and $\bar{w}_n(x,t)=\frac{1}{(x-t)^{n+1}}=\mathrm{Ap}[\bar{v}_n(x,t)]$ (b) $\omega_n(x,t)=\frac{(-1)^n}{n!}D^n\delta(x+t)$ and $\bar{\omega}_n(x,t)=\frac{(-1)^n}{n!}D^n\delta(x-t)$
5. Restricted analyticity, $f\in A^0$ $f(x)=\frac{1}{2\pi i}\int_\Gamma\frac{f(z)}{x-z}dz$, Γ, a restricted circle of the complex z-plane	Huygens property, $u\in H^0$ $u(x,t)=\int_{-\infty}^{\infty}k(x-y,t-s)u(y,s)\,dy,\quad s<t$	Poisson transform, $u\in W$ $u(x,t)=k(x,t)*u(x,0)+K(x,t)*u_t(x,0)$, with $K(x,t)$ as in (14).
7. Biorthogonality $\frac{1}{2\pi i}\int_\Gamma v_m(z)w_n(z)\,dz=\delta_{m,n}$	Biorthogonality $\frac{1}{n!2^n}\int_{-\infty}^{\infty}v_m(x,-t)w_n(x,t)\,dx=\delta_{m,n}$	Biorthogonality $\int_{-\infty}^{+\infty}v_m(x,t)\omega_n(x,t)\,dx$ $=\delta_{m,n}=\int_{-\infty}^{+\infty}\bar{v}_m(x,t)\bar{\omega}_n(x,t)\,dx$
8. Generating function $\frac{1}{x-r}=\sum_{n=0}^{\infty}w_n(x)r^n,\quad \lvert r\rvert<\lvert x\rvert$ $\frac{1}{t-x}=\sum_{n=0}^{\infty}v_n(x)w_n(t),\quad \lvert x\rvert<\lvert t\rvert$	Generating function $k(x-r,t)=\sum_{n=0}^{\infty}\frac{w_n(x,t)r^n}{2^n n!}$ $k(x-y,t+s)=\sum_{n=0}^{\infty}\frac{v_n(x,t)w_n(y,s)}{2^n n!}$	Generating function $W_0(x-r,t)=\sum_{n=0}^{\infty}w_n(x,t)r^n+\sum_{n=0}^{\infty}\bar{w}_n(x,t)r^n$ $W_0(y-x,s-t)=\sum_{n=0}^{\infty}v_n(x,t)w_n(y,s)$ $+\sum_{n=0}^{\infty}\bar{v}_n(x,t)\bar{w}_n(y,s)$, where $W_0(x,t)=k(x,t)*1/x$.
9. Maclaurin expansion $f(x)=\sum_{n=0}^{\infty}a_nv_n(x)$	Polynomial expansion $u(x,t)=\sum_{n=0}^{\infty}a_nv_n(x,t)$	Polynomial expansion $u(x,t)=\sum_{n=0}^{\infty}\{a_nv_n(x,t)+b_n\bar{v}_n(x,t)\}$
10. Inverse expansion $f(x)=\sum_{n=0}^{\infty}b_nw_n(x)$	Associated expansion $u(x,t)=\sum_{n=0}^{\infty}b_nw_n(x,t)$	Associated expansion $u(x,t)=\sum_{n=0}^{\infty}\{c_nw_n(x,t)+d_n\bar{w}_n(x,t)\}$
11. Criterion for polynomial expansion In largest interval $\lvert x\rvert<\rho$, where $f\in A^0$	Criterion for polynomial expansion In widest strip $\lvert t\rvert<\sigma$ where $u\in H^0$	Criterion for polynomial expansion In largest region $T(\rho)$ where $u\in W^0$
12. Criterion for inverse expansion A. Valid for $\lvert x\rvert>\rho$ if $f=I[g]$, $g\in A^0$ for $\lvert x\rvert<\frac{1}{\rho}$ B. Valid for $x>\rho$ if $f(x)=\int_0^{\infty}e^{-xr}\phi(r)\,dr,\phi\in(1,\rho)$	Criterion for associated expansion A. Valid for $t>\sigma$ if $u=\mathrm{Ap}[g]$, $g\in H^0$ for $\lvert t\rvert<\frac{1}{\sigma}$ B. Valid for $x>\sigma$ if $u(x,t)=\int_{-\infty}^{\infty}e^{ixr-tr^2}\phi(r)\,dr,\ \phi\in\{2,\sigma\}$	Criterion for associated expansion A. Valid in $T^{-1}(\rho)$ if $u=\mathrm{Ap}[g]$, $g\in W^0$ in $T\left(\frac{1}{\rho}\right)$ B. Valid for $x+t>\rho_1$, $x-t>\rho_2$ if $u(x,t)=\int_0^{\infty}\bar{e}^{(x+t)y}\theta(y)\,dy+\int_0^{\infty}\bar{e}^{(x-t)y}\lambda(y)\,dy$, with $\theta\in\{1,\rho_1\}$ and $\lambda\in\{1,\rho_2\}$.

6. Summary. In order to provide a convenient summary of the analogies that have been listed in the previous sections, we conclude with Table 1. The first two columns and the form of the table are drawn directly from [8, p. 196]. The corresponding analogies for the wave equation in each case are given in the third column.

Since the numbering of the analogies was kept in accordance with that in [8], it is easy to see that one analogy (no. 6) was omitted from Table 1. That particular analogy involves operational calculus, and the precise analogue for the wave equation does not appear to have a convenient form. However, the symbolic form does provide a nice analogue of the Poisson transform.

The integral representation of a solution of the heat equation as the Poisson transform can be written [8, p. 155] as

$$u(x, t) = e^{tD^2} u(x, 0).$$

The corresponding analogue for the wave equation is a form of the d'Alembert formula (7), and is

$$u(x, t) = \cosh(tD)u(x, 0) + \sinh(tD)D^{-1}u_t(x, 0). \tag{31}$$

The symbolic use of e^{tD} as $e^{tD}f(x) = f(x+t)$ is used to see that the first term is indeed $\frac{1}{2}[u(x+t, 0) + u(x-t, 0)]$. If

$$D^{-1}u_t(x, 0) = \int_0^x u_t(\xi, 0)\, d\xi,$$

the second term can be seen to match that in (7) in a similar manner.

REFERENCES

[1] J. ARSAC, *Fourier Transforms and the Theory of Distributions*, Prentice-Hall, Englewood Cliffs, NJ, 1966.

[2] R. P. BOAS, JR., *Entire Functions*, Academic Press, New York, 1954.

[3] G. F. CARRIER AND C. E. PEARSON, *Partial Differential Equations*, Academic Press, New York, 1976.

[4] P. R. GARABEDIAN, *Partial Differential Equations*, John Wiley, New York, 1964.

[5] M. J. LIGHTHILL, *Introduction to Fourier Analysis and Generalised Functions*, Cambridge University Press, London, 1958.

[6] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.

[7] D. V. WIDDER, *Some analogies from classical analysis in the theory of heat conduction*, Arch. Rational Mech. Anal., 21 (1966), pp. 108–119.

[8] ———, *The Heat Equation*, Academic Press, New York, 1975.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0002 $01.00/0

A NONLINEAR EVOLUTION PROBLEM ARISING IN THE PHYSICS OF IONIZED GASES*

D. HILHORST†

Abstract. We consider a Coulomb gas in a special experimental situation: the pre-breakdown gas discharge between two electrodes. The equation for the negative charge density can be formulated as a nonlinear parabolic equation degenerate at the origin. We prove the existence and uniqueness of the solution as well as the asymptotic stability of its unique steady state. Also some results are given about the rate of convergence.

1. Introduction.

In this paper we study the nonlinear evolution problem

$$\text{P}\qquad \begin{aligned} &u_t = \varepsilon x u_{xx} + (g(x)-u)u_x && \text{on } D=(0,\infty)\times(0,T),\\ &u(0,t)=0 && \text{for } t\in[0,T],\\ &u(x,0)=\psi(x) && \text{for } x\in(0,\infty), \end{aligned}$$

where ε is a positive constant, g is a given function which satisfies the hypothesis $H_g: g\in C^2([0,\infty))$; $g(0)=0$; $g'(x)>0$ and $g''(x)<0$ for all $x\geqq 0$ and the initial function ψ satisfies the hypothesis H_ψ:

(i) ψ is continuous, with piecewise continuous derivative on $[0,\infty)$;

(ii) $\psi(0)=0$ and $\psi(\infty)=K\in(0,g(\infty))$;

(iii) there exists a constant $M_\psi\geqq g'(0)$ such that $0\leqq\psi'(x)\leqq M_\psi$ at all points x where ψ' is defined.

In § 2 we briefly describe how the problem arises in physics and give the derivation of the equations.

In § 3 we present maximum principles for certain linear and nonlinear problems related to P; the uniqueness of the solution of P follows directly from those principles.

In § 4 we prove that P has a classical solution which satisfies furthermore the condition

$$(*)\qquad u(\infty,t)=K \quad \text{for } t\in[0,T], \quad T<\infty.$$

The methods used here are inspired by those of van Duyn [7], [8] and Gilding and Peletier [13]. We also consider the limit case $\varepsilon\downarrow 0$ and prove that u tends to the generalized solution of the corresponding hyperbolic problem.

We then investigate the behavior of u as $t\to\infty$ and prove that it converges towards the unique solution Φ of the problem P_0 defined as follows

$$\text{P}_0\qquad \begin{aligned} &\varepsilon x\Phi''+(g(x)-\Phi)\Phi'=0,\\ &\Phi(0)=0, \qquad \Phi(\infty)=\lambda_0=:\min(\max(g(\infty)-\varepsilon,0),K). \end{aligned}$$

Qualitative properties of Φ have been extensively studied by Diekmann, Hilhorst and Peletier [6]. Here we analyze its stability. In § 5, following a method of Aronson and Weinberger [2] based on the knowledge about lower and upper solutions for the steady state problem P_0, we prove that Φ is asymptotically stable.

In § 6 we investigate the rate of convergence of u towards its steady state. The function Φ turns out to be exponentially stable when the function g grows fast enough to infinity as $x\to\infty$; the proof, based on constructing upper and lower solutions for the function $u-\Phi$, follows the same lines as that of Fife and Peletier [10]. We also

* Received by the editors December 23, 1980.

† Stichting Mathematisch Centrum, Kruislaan 413, 1098 SJ Amsterdam, the Netherlands.

consider the case when g increases less fast and show that provided $\varepsilon < g(\infty) - K$ and Φ converges algebraically fast to K as $x \to \infty$ the function $u - \Phi$ decays algebraically fast; this is done by obtaining first that property for a weighted integral of $u - \Phi$ according to a method of Il'in and Oleinik [14] and van Duyn and Peletier [9]. Finally we consider the corresponding hyperbolic problem and obtain a similar result of algebraic convergence.

2. Physical derivation of the equations. The physical context of the present problem has been described in some detail by Diekmann, Hilhorst and Peletier [6]. Here we shall summarize it again and explain how one can obtain the time evolution problem P.

One considers an ionized gas between two electrodes in which the ions and electrons are present with densities $n_i(\mathbf{r})$ and $n_e(\mathbf{r}, t)$ respectively, where $\mathbf{r} = (x_1, x_2, x_3)$. The ions are heavy and slow and the density $n_i(\mathbf{r})$ may therefore be regarded as fixed. The electrons are highly mobile. The problem is then to find $n_e(\mathbf{r}, t)$ for given $n_i(\mathbf{r})$ and in particular to find out whether given an initial electron distribution, the electrons stabilize and if so to evaluate the time needed for such a stabilization.

A special situation of practical interest is a so-called pre-breakdown discharge which spreads out in filamentary form (cf. Marode [17] and Marode, Bastien and Bakker [18]). In this situation there is cylindrical symmetry about the x_3-axis and the particle densities depend on $r = (x_1^2 + x_2^2)^{1/2}$ only. We thus have effectively a two-dimensional Coulomb gas with circular symmetry. The starting equations are

(i) Coulomb's law for the electric field E,

$$\frac{1}{r}\frac{\partial}{\partial r} rE = -C_d(n_e - n_i), \tag{2.1}$$

where C_d is a fixed constant;

(ii) a constitutive equation for the electric current j,

$$j = n_e \mu E + kT\frac{\partial n_e}{\partial r}, \tag{2.2}$$

in which the first term represents Ohm's law and the second term is due to thermal diffusion, μ being the mobility, k Boltzmann's constant and T the temperature; and

(iii) the continuity equation for the electron density,

$$\frac{\partial n_e}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r} rj. \tag{2.3}$$

If we set

$$u(x, t) = \int_0^{\sqrt{x}} n_e(r, t) r\, dr$$

and

$$g(x) = \int_0^{\sqrt{x}} n_i(r) r\, dr,$$

we obtain, after redefining the constants, the equation

$$u_t = \varepsilon x u_{xx} + (g(x) - u)u_x, \tag{2.4}$$

where $\varepsilon = 2kT/(\mu C_d)$, and the boundary condition

$$u(0, t) = 0. \tag{2.5}$$

Furthermore one makes the hypothesis that the total charge is positive and fixed, that is

$$\int_0^\infty (n_i(r) - n_e(r, t))r\, dr = N > 0,$$

from which we deduce the boundary condition at infinity;

$$u(\infty, t) = K := g(\infty) - N. \tag{2.6}$$

Clearly $K \in (0, g(\infty))$.

Equations (2.4) and (2.5) together with the initial condition

$$u(x, 0) = \psi(x) \tag{2.7}$$

constitute the mathematical formulation of the problem which we propose to study in this paper. Furthermore the condition (2.6) will turn out to be satisfied at all finite times t and also, for low enough values of the small parameter ε, at the time $t = \infty$. This latter property expresses the fact that all the electrons stay attached to the ions at low enough temperature; we shall also see that if the temperature rises above a critical value then some of the electrons escape to infinity, and if it rises even further above a second critical value then all the electrons escape to infinity.

3. Maximum principles for some degenerate parabolic operators–uniqueness theorem. In this section we prove maximum principles for some linear and nonlinear operators which have a degeneracy at the origin; these principles hold for functions $u \in C^{2,1}(D) \cap C(\bar{D})$, where $C^{2,1}(D)$ is the set of continuous functions on D with two continuous x-derivatives and one continuous t-derivative. It will follow easily from those maximum principles that P can have at most one solution $u \in C^{2,1}(D) \cap C(\bar{D})$ such that u_x is bounded in $\bar{D}$.

We begin by defining a linear operator L as follows

$$Lu = \varepsilon x u_{xx} + b(x, t)u_x + c(x, t)u - u_t, \tag{3.1}$$

where the functions b and c are continuous on D and such that the quantities $b/(1+x)$ and c are bounded on $\bar{D}$. First we consider the bounded domain $D_R := (0, R) \times (0, T)$, where R is a positive constant. In the same way as for a uniformly parabolic operator one can prove the following maximum principle which holds in fact for a much wider class of degenerate parabolic operators (see, for example, Ippolito [15] or Cosner [4])

THEOREM 3.1. *Suppose $c \leqq 0$. Let $u \in C^{2,1}(D_R) \cap C(\bar{D}_R)$ satisfy $Lu \geqq 0$ on $(0, R) \times (0, T]$. Then if u has a positive maximum in $\bar{D}_R$, that maximum is attained on $((0, R) \times \{0\}) \cup (\{0, R\} \times [0, T])$.*

Next, following a method due to Aronson and Weinberger [2], we derive a comparison theorem for a class of nonlinear evolution problems.

THEOREM 3.2. *Let u and $v \in C^{2,1}(D_R) \cap C(\bar{D}_R)$ and suppose that either u_x or v_x is bounded on $\bar{D}_R$. Let u and v satisfy*

$$Lv - vv_x \geqq Lu - uu_x \quad \text{on } (0, R) \times (0, T],$$

and let

$$0 \leqq v \leqq u \leqq K \quad \text{on } (0, R) \times \{0\} \text{ and } \{0, R\} \times [0, T].$$

Then $v \leqq u$ *in* $(0, R)\times(0, T]$.

Proof. Let

$$w = (v-u)\, e^{-\alpha t},$$

where

$$\alpha = \max_{(x,t)\in \bar{D}} (c(x, t) - u_x(x, t))$$

(in the case where u_x is bounded). Then w satisfies

$$\varepsilon x w_{xx} + (b(x, t) - v) w_x + (c(x, t) - u_x - \alpha) w - w_t \geqq 0$$

and

$$w \leqq 0 \quad \text{on } (0, R)\times\{0\} \text{ and } \{0, R\}\times[0, T].$$

Thus we deduce from Theorem 3.1 that

$$w \leqq 0 \quad \text{in } (0, R)\times(0, T],$$

which completes the proof of Theorem 3.2. □

Now let us consider the unbounded domain D. To begin with we present a Phragmèn–Lindelöf principle which is a special case of a theorem due to Cosner [4].

THEOREM 3.3. *Suppose that* $b/(1+x)$ *and* c *are continuous and bounded in* $\bar{D}$. *Let* $u \in C^{2,1}(D) \cap C(\bar{D})$ *satisfy* $Lu \geqq 0$ *on* $(0, \infty)\times(0, T]$ *and the growth condition*

$$\liminf_{\mathscr{R}\to\infty} e^{-B\mathscr{R}} [\max_{0\leqq t\leqq T} u(\mathscr{R}, t)] \leqq 0 \tag{3.2}$$

for some positive constant B. *If* $u \leqq 0$ *for* $t = 0$ *and on* $\{0\}\times[0, T]$ *then* $u \leqq 0$ *in* $(0, \infty)\times(0, T]$.

Making use of Theorem 3.3 one can prove a comparison theorem on the unbounded domain D.

THEOREM 3.4. *Let* u *and* $v \in C^{2,1}(D) \cap C(\bar{D})$ *be such that either* u_x *and* v *or* u *and* v_x *are bounded on* $\bar{D}$ *and that*

$$|u(x, t)|, |v(x, t)| \leqq C e^{B_1 x}$$

for some positive constants C *and* B_1 *and uniformly in* $t \in [0, T]$. *Suppose that*

$$Lv - v v_x \geqq Lu - u u_x \quad \text{on } (0, \infty)\times(0, T]$$

and that

$$0 \leqq v \leqq u \leqq K \quad \text{on } (0, \infty)\times\{0\} \text{ and } \{0\}\times[0, T].$$

Then $v \leqq u$ *in* $(0, \infty)\times(0, T]$.

Finally let us come to the question of uniqueness of the solution of problem P.

DEFINITION. We shall say that u is a classical solution of problem P if it is such that (i) $u \in C^{2,1}(D) \cap C(\bar{D})$, (ii) u and u_x are bounded in $\bar{D}$, (iii) u satisfies the equation in D, (iv) u satisfies the initial and boundary conditions.

THEOREM 3.5. *Problem* P *can have at most one solution.*

Proof. Apply Theorem 3.4 twice to deduce that if u and v are two such solutions then their difference $w = u - v$ satisfies $w \geqq 0$ and $w \leqq 0$ and thus $w \equiv 0$. □

4. Existence and regularity of the solution. In order to be able to prove the existence of a solution of the nonlinear degenerate parabolic problem P, we consider

certain related nonlinear uniformly parabolic problems on bounded domains and observe that they have a unique solution; we then deduce that P has a generalized solution, in a certain sense. It finally turns out that this solution is in fact a classical solution of P and thus the unique solution of P and that it also satisfies condition (*). Finally we consider its limiting behavior as $\varepsilon \downarrow 0$.

4.1. Existence. Let us first introduce some notation. Let $D_n := (0, n) \times (0, T)$. We denote by $C_{2+\alpha}([0, n])$ the space of functions v which are twice differentiable and such that v'' is Hölder continuous on $[0, n]$ with exponent α. We also use the spaces $\bar{C}_\alpha(D_n)$, $\overline{C_{2+\alpha}}(D_n)$ and $C_{2+\alpha}(D_n)$, defined in Friedman [11, pp. 62, 63].

Consider the problem

$$\mathrm{P}_n \qquad \begin{aligned} &u_t = \varepsilon(x+1/n)u_{xx} + (g(x)-u)u_x && \text{in } D_n, \\ &u(0,t) = 0, \qquad u(n,t) = K, && t \in [0, T], \\ &u(x,0) = \psi_n(x), && x \in (0, n), \end{aligned}$$

with $n \geqq g^{-1}(K)$ and where ψ_n is such that

(i) $\psi_n \in C^\infty([0, \infty])$;
(ii) ψ_n satisfies H_ψ;
(iii) $\psi_n''(0) = 0$ and $\psi_n(x) = K$ for $x \in [n-1, \infty)$.

In what follows we shall denote by H_n properties (i) – (iii). The following theorem holds:

THEOREM 4.1. *There exists a unique solution $u_n \in \overline{C_{2+\alpha}}(D_n)$ of P_n for any $\alpha \in (0, 1)$; furthermore u_n satisfies the inequalities*

$$0 \leqq u_n(x,t) \leqq \min(M_{\psi_n} x, K), \tag{4.1}$$

$$0 \leqq u_{nx}(x,t) \leqq M_{\psi_n}, \tag{4.2}$$

for all $(x, t) \in \bar{D}_n$.

Proof. The existence and uniqueness of $u_n \in \overline{C_{2+\alpha}}(D_n)$ is a consequence of Theorem 5.2 of Ladyženskaja [16, pp. 564–565]. The inequalities in (4.1) can be deduced by means of a comparison theorem analogous to Theorem 3.2. From the linear theory (Friedman [11, p. 72]) we deduce that the function $w := u_{nx} \in C_{2+\alpha}(D_n)$; thus $w \in C^{2,1}(D_n) \cap C(\bar{D}_n)$. Furthermore w satisfies

$$\begin{aligned} &w_t = \varepsilon(x+1/n)w_{xx} + (g(x) - u_n + \varepsilon)w_x + (g'(x) - w)w, \\ &0 \leqq w(0,t) \leqq M_{\psi_n}, \qquad 0 \leqq w(n,t) \leqq M_{\psi_n}, \\ &w(x,0) = \psi_n'(x). \end{aligned} \tag{4.3}$$

The bounds on the function $w(n, t)$ follow from the fact that the function $\max(0, M_{\psi_n}(x-n) + K)$ is a lower solution of the boundary value problem

$$\varepsilon\left(x + \frac{1}{n}\right)\phi'' + (g(x) - \phi)\phi' = 0, \qquad \phi(0) = 0, \quad \phi(n) = K$$

and consequently a lower bound for u_n. Clearly the set

$$\{w \in C([0, n]) \text{ such that } 0 \leqq w(x) \leqq M_{\psi_n}\}$$

is invariant with respect to the problem (4.3), and thus the inequalities (4.2) are satisfied.

Next we deduce, from Theorem 4.1, the existence of solution of P. We begin by approximating the initial function ψ by a sequence of smooth functions $\{\psi_n\}$.

LEMMA 4.2. *Let the function ψ satisfy* H_ψ. *Then there exists a sequence* $\{\psi_n\}$ *which satisfies the properties* H_n *given at the beginning of this section with* $M_{\psi_n} = M_\psi$ *for all* n, *such that* $\psi_n \to \psi$ *as* $n \to \infty$, *uniformly on* $[0, \infty)$.

Proof. Let $n_0 \geqq g^{-1}(K)$ be such that for all $n \geqq n_0$ the point x_{1n} defined by $M_\psi(x_{1n} - 1/n) = \psi(x_{1n})$ is such that $1/n < x_{1n} \leqq n-2$ and the point x_{2n} defined by $x_{2n} = n - 2 + (K - \psi(n-2))/M_\psi$ satisfies $n-2 < x_{2n} < n-1$. Also define

$$\psi_n^*(x) = \begin{cases} 0, & -\infty < x \leqq \dfrac{1}{n}, \\ M_\psi\left(x - \dfrac{1}{n}\right), & \dfrac{1}{n} < x \leqq x_{1n}, \\ \psi(x), & x_{1n} < x \leqq n-2, \\ M_\psi(x - n + 2) + \psi(n-2), & n-2 < x \leqq x_{2n}, \\ K, & x_{2n} < x < +\infty. \end{cases}$$

Note that, for all x,

$$|\psi_n^*(x) - \psi(x)| \leqq \max\left(\frac{M_\psi}{n}, K - \psi(n-2)\right).$$

Next introduce the function

$$\rho(x) = \begin{cases} 0 & \text{if } |x| \geqq 1, \\ C \exp\left(\dfrac{1}{|x|^2 - 1}\right) & \text{if } |x| < 1, \end{cases}$$

where the constant C is such that $\int_{\mathbb{R}} \rho \, dx = 1$, and let

$$\rho_\delta(x) = \frac{\rho(x/\delta)}{\delta}.$$

Finally define

$$\psi_n(x) = \int_{\mathbb{R}} \rho_{\delta_n}(x - y)\psi_n^*(y) \, dy, \qquad x \in [0, n],$$

with $\delta_n = \min(1/n, x_{1n} - 1/n, n - 2 - x_{1n}, x_{2n} - n + 2, n - 1 - x_{2n})/10$. We now show that ψ_n has the desired properties. Firstly $\psi_n \in C^\infty([0, n])$. The uniform convergence of $\{\psi_n\}$ to ψ follows from the continuity of ψ_n^*, uniformly in n and in x and the uniform convergence of ψ_n^* to ψ as $n \to \infty$. Finally properties (ii) and (iii) of H_n can be deduced for ψ_n from the fact that ψ also satisfies them.

Next we prove the following theorem.

THEOREM 4.3. P *has a unique classical solution. Furthermore this solution also satisfies condition* (*):

$$(*) \qquad \lim_{x \to \infty} u(x, t) = K \quad \textit{for each } t \in (0, T].$$

Proof. We rewrite the parabolic equation of problem P_n as

$$(4.4) \qquad u_t = \varepsilon(x + 1/n)u_{xx} + c(x, t)u_x,$$

where

$$c(x, t) = g(x) - u_n(x, t).$$

From Theorem 4.1 we know that for all (x', t), $(x'', t) \in \bar{D}_n$ and for all $n \geqq n_0$

$$|u_n(x', t) - u_n(x'', t)| \leqq M_\psi |x' - x''|. \tag{4.5}$$

Now fix $I \geqq n_0$; (4.4) and (4.5) enable us to apply a theorem of Gilding [12] about the Hölder continuity of solutions of parabolic equations, and we obtain

$$|u_n(x, t') - u_n(x, t'')| \leqq C|t' - t''|^{1/2}$$

for all $n \geqq I$ and for all (x, t'), $(x, t'') \in \bar{D}_I$, with $|t' - t''| \leqq 1$. Here the constant C depends on I but not on n. The set $\{u_n(x, t)\}_{n=I}^{\infty}$ is bounded and equicontinuous in D_I, and thus there exists a continuous function $u_I(x, t)$ and a convergent subsequence $\{u_{n_k}(x, t)\}$ with $n_k \geqq I$ such that $u_{n_k}(x, t) \to u_I(x, t)$ as $n_k \to \infty$, uniformly on $\bar{D}_I$. Then, by a diagonal process, it follows that there exists a function $u(x, t)$ defined on $\bar{D}$ and a convergent subsequence, denoted by $\{u_j(x, t)\}$ such that $u_j(x, t) \to u(x, t)$ as $j \to \infty$, pointwise on $\bar{D}$. Since this convergence is uniform on any bounded subset of $\bar{D}$, the limit function u is continuous on $\bar{D}$.

It remains to show that u is a solution of P; to that purpose we shall proceed in two steps: firstly we show that u is a generalized solution of P in a certain sense and then we conclude that it is in fact a classical solution. We shall say that u is a generalized solution of P if it has the following properties:

(i) u is continuous and uniformly bounded in $\bar{D}$;
(ii) $u(0, t) = 0$ for all $t \in [0, T]$;
(iii) u has a bounded generalized derivative with respect to x in D;
(iv) u satisfies the identity

$$\iint_D [u\phi_t - \varepsilon(xu_x - u)\phi_x - (g - u/2)u\phi_x - ug'\phi]\, dx\, dt + \int_0^\infty \psi(x)\phi(x, 0)\, dx = 0 \tag{4.6}$$

for all $\phi \in C^1(\bar{D})$ which vanish for $x = 0$, large x and $t = T$.

Let us check that u satisfies those properties.

(i) We already know that u is continuous on $\bar{D}$ and furthermore, since $u(x, t) = \lim_{j\to\infty} u_j(x, t)$, we have that $0 \leqq u \leqq K$.

(ii) This property follows from a similar boundary condition in P_n.

(iii) Let ϕ be an admissible test function and let $L \geqq n_0$ be such that supp $\phi \subset D_L$. Since $|u_{jx}|$ is uniformly bounded with respect to $j \geqq L$ for all $(x, t) \in D_L$, it follows that there exists a subsequence $\{(u_{j_k})_x\}$ and a bounded function $p \in L^2(D_L)$ such that

$$(u_{j_k})_x \rightharpoonup p \quad \text{in } L^2(D_L) \text{ as } j_k \to \infty.$$

Now let $\zeta \in C_0^1(\bar{D}_L)$. Then

$$((u_{j_k})_x, \zeta) \to (p, \zeta) \quad \text{as } j_k \to \infty, \tag{4.7}$$

where $(\cdot, \cdot)$ denotes the inner product in $L^2(D_L)$. But since $u_{j_k} \to u$ as $j_k \to \infty$, uniformly on $\bar{D}_L$, we have

$$(u_{j_k}, \zeta_x) \to (u, \zeta_x) \quad \text{as } j_k \to \infty. \tag{4.8}$$

Hence, combining (4.7) and (4.8), we find that p is the generalized derivative of u.

(iv) Since u_{j_k} is a classical solution of P_n it follows that

$$\iint_{D_L} \left[u_{j_k}\phi_t - \varepsilon\left(\left(x+\frac{1}{j_k}\right)(u_{j_k})_x - u_{j_k}\right)\phi_x - \left(g - \frac{u_{j_k}}{2}\right)u_{j_k}\phi_x - u_{j_k}g'\phi \right] dx\,dt + \int_0^L \psi_{j_k}(x)\phi(x,0)\,dx = 0. \tag{4.9}$$

The sequences $\{u_{j_k}\}$ and $\{u_{j_k}^2\}$ converge to u and u^2, respectively, strongly in $L^2(D_L)$ as $j_k \to \infty$. Furthermore since $(u_{j_k})_x$ is uniformly bounded we have

$$\iint_{D_L} \frac{1}{j_k}(u_{j_k})_x\phi_x\,dx\,dt \to 0 \quad \text{as } j_k \to \infty.$$

Thus letting $j_k \to \infty$ we obtain (4.6). Because ϕ has been chosen arbitrarily, we may conclude that u is indeed a generalized solution of P.

It remains to show that u is a classical solution of P. One can do it by using a classical bootstrap argument (see, for example, Gilding and Peletier [13]) to show that for whatever $\eta, L > 0$ there exists $\alpha(\eta, L) \in (0, 1)$ such that

$$u \in \overline{C_{2+\alpha}}((\eta, L)\times(\eta, T)), \tag{4.10}$$

where α and $\|u\|_{C_{2+\alpha}}$ may be estimated independently of T. In particular,

$$u \in C^{2,1}(D) \cap C(\bar{D}).$$

Since furthermore u and u_x are uniformly bounded u is a classical solution of problem P and by Theorem 3.5 it is the unique solution of P.

Finally let us analyze the behavior of u for large x; since we have $0 \leqq u \leqq K$ and $u_x \geqq 0$, $u(\infty, t) = \lim_{x\to\infty} u(x, t)$ is well defined for all $t \in [0, T]$ and such that $0 \leqq u(\infty, t) \leqq K$. Next we show that $u(\infty, t) \equiv K$ by constructing a time dependent lower solution for P. Consider the problem

$$\begin{aligned} &u_t = \varepsilon x u_{xx} + (K-u)u_x, \\ &u(x_0, t) = 0, \qquad x_0 \geqq g^{-1}(K), \\ &u(x, 0) = \psi(x). \end{aligned} \tag{4.11}$$

Since $u_x \geqq 0$ we have that

$$\begin{aligned} \varepsilon x u_{xx} + (g(x)-u)u_x - u_t &= \varepsilon x u_{xx} + (K-u)u_x - u_t + (g(x)-K)u_x \\ &\geqq \varepsilon x u_{xx} + (K-u)u_x - u_t \quad \text{for all } x \geqq g^{-1}(K). \end{aligned}$$

Thus a lower solution $\hat{u}$ of (4.11) with $\hat{u}_x \geqq 0$ is also a lower solution of P on $[x_0, \infty)\times[0, T]$. We search such functions $\hat{u}_k$ which satisfy furthermore

$$\hat{u}_k(\infty, t) = K - k \quad \text{for all } t \in [0, T] \text{ and with } k \in (0, K).$$

Writing

$$\hat{v} = K - \hat{u},$$

reduces this to finding an upper solution $\hat{v}_k$ of

$$\begin{aligned} &v_t = \varepsilon x v_{xx} + v v_x, \\ &v(x_0, t) = K, \qquad v(\infty, t) = 0. \end{aligned}$$

Next we look for such a function $\hat{v}_k$, also requiring that

$$\hat{v}_k(x,t) = \hat{f}_k\left(\frac{x}{t+1}\right).$$

Setting

$$\eta = \frac{x}{t+1},$$

one can easily derive that $\hat{f}_k$ should be an upper solution for the boundary value problem

π

$$\varepsilon\eta f'' + (f+\eta)f' = 0,$$
$$f(x_0) = K, \qquad f(\infty) = 0.$$

Let $x_0 > \max(\varepsilon, g^{-1}(K))$, and take

$$\hat{f}_k(\eta) = k + (K-k)\left(\frac{\eta}{x_0}\right)^{1-x_0/\varepsilon}.$$

One can check that indeed $\hat{f}_k$ is an upper solution for problem π and consequently that $\hat{u}_k(x,t) = K - \hat{f}_k(x/(t+1))$ is a lower solution for problem P on the sector $\{t \geqq 0, x \geqq x_0(t+1)\}$ provided that x_0 is large enough. Since k can be chosen arbitrarily in $(0, K)$ it follows that $u(\infty, t) = K$ for all $t < \infty$. □

4.2. The limiting behavior as $\varepsilon \downarrow 0$. In this section we study the limiting behavior of the solution u of P as $\varepsilon \downarrow 0$. To begin with, we consider the following hyperbolic problem:

H

$$u_t = (g(x) - u)u_x \quad \text{in } D,$$
$$u(x,0) = \psi(x) \quad \text{for all } x \in (0,\infty),$$

and make some heuristic considerations about the solution $\bar{u}$ of problem H; they are due to Wilders [23]. One possible configuration of g and ψ is drawn in Fig. 1; the corresponding characteristics are represented in Fig. 2. Their equations are

$$\frac{dx}{dt} = -(g(x) - \psi(x(0))).$$

Along those characteristics $\bar{u}$ is constant, i.e., $\bar{u} = \psi(x(0))$. Also, since $\psi(0) = 0$ it follows that the line $x = 0$ is the characteristic passing through the point $(0,0)$ and

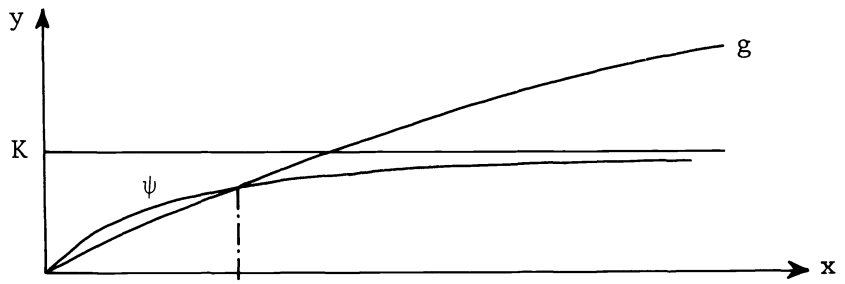

FIG. 1

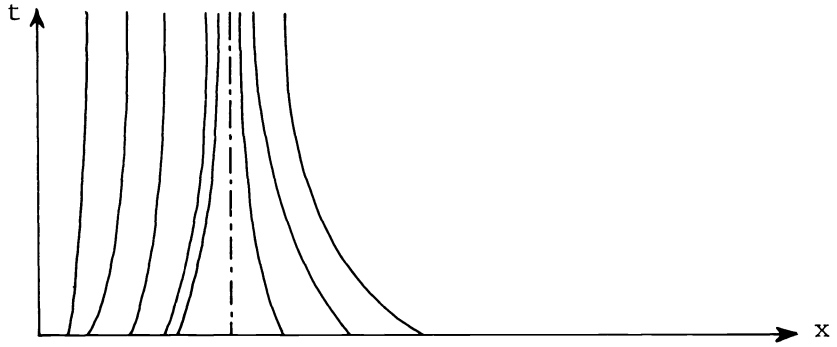

FIG. 2

consequently that $\bar{u}$ automatically satisfies a boundary condition of the form $\bar{u}(0, t) = 0$. Next we deduce from the fact that ψ is nondecreasing that two characteristics do not intersect. Suppose that there exist two characteristics issuing from the points $x = a$ and $x = b (a < b)$ on the initial line, intersecting each other at the point $(x, t) = (x^*, t^*)$. Then if they would intersect transversally we would have $-(g(x^*) - \psi(a)) > -(g(x^*) - \psi(b))$ and hence $\psi(a) > \psi(b)$, which is impossible. Now if the characteristics would be tangent to each other at the point (x^*, t^*) we would have $-(g(x^*) - \psi(a)) = -(g(x^*) - \psi(b))$ and consequently $\psi(a) = \psi(b)$; both characteristics would then be described by the same differential equation $dx/dt = -(g(x) - \psi(a))$, which, by the standard uniqueness theorem for ordinary differential equations, implies $a = b$. Finally we conclude that since the initial condition ψ is continuous and nondecreasing, no shock wave can occur and $\bar{u}(\cdot, t)$ is continuous at all times.

In [19] Oleinik proved existence and uniqueness of the generalized solution of Cauchy problems and boundary value problems related to problem H but since the boundary line $x = 0$ is a characteristic for H (which is reflected in the relation $g(0) - \bar{u}(0, 0) = 0$), problem H does not satisfy all the assumptions made in [19]. This leads us to give here a proof of the existence of a solution of problem H, by showing that the solution u of problem P tends to a limit as $\varepsilon \downarrow 0$; the uniqueness is a consequence of [19]. Following [19, Lemmas 18 and 19], we say that $\bar{u}$ is a generalized solution of H if it satisfies

(i) $\bar{u}$ is bounded and measurable in $\bar{D}$;

(ii) $\dfrac{\bar{u}(x_1, t) - \bar{u}(x_2, t)}{x_1 - x_2} \leqq M_\psi$ for all points $(x_1, t), (x_2, t) \in \bar{D}$;

(iii) $\bar{u}$ satisfies the identity

$$\iint_D \left[\bar{u}\phi_t - \left(g - \frac{\bar{u}}{2} \right) \bar{u}\phi_x - \bar{u}g'\phi \right] dx\, dt + \int_0^\infty \psi(x)\phi(x, 0)\, dx = 0 \tag{4.12}$$

for all $\phi \in C^1(\bar{D})$ which vanish for large x and $t = T$.

Next we shall prove the following theorem.

THEOREM 4.4. *The solution $u(x, t)$ of* P *tends uniformly on all compact subdomains of D to a limit $\bar{u}$ as $\varepsilon \downarrow 0$, where $\bar{u}$ is the unique generalized solution of* H. *The function $\bar{u}$ is furthermore continuous, nondecreasing in x at all times $t \in [0, T]$ and satisfies the boundary conditions $\bar{u}(0, t) = 0$ and $\bar{u}(\infty, t) = K$.*

Before proving Theorem 4.4, let us introduce a class of upper and lower solutions for problem P which depend neither on ε nor on time. They will turn out to be very useful both to prove that $\bar{u}(\infty, t) = K$ in Theorem 4.4 and to study the asymptotic behavior of u as $t \to \infty$ in the next sections. Next we define

$$s^+(x) := \min(M_\psi x, K)$$

and

$$s^-(x, \lambda, x_1, \nu) := \max\left(0, \lambda\left(1 - \left(\frac{x}{x_1}\right)^{-\nu}\right)\right),$$

where the constants $\lambda \in [0, K]$, $\nu > 0$ and $x_1 > 0$ are chosen in the following manner:

(a) If $\varepsilon < g(\infty)$, we choose $x_1 > 0$ so that $g(x_1) > \varepsilon$, then $\lambda > 0$ so that $\lambda < g(x_1) - \varepsilon$ and finally $\nu > 0$ so that

$$\nu \leqq \varepsilon^{-1}(g(x_1) - \lambda) - 1. \tag{4.13}$$

(b) If $\varepsilon \geqq g(\infty)$, we set $\lambda = 0$, which amounts to setting $s^- \equiv 0$.

It is easily seen that s^- satisfies the inequality

$$\hat{\varepsilon} x (s^-)'' + (g - s^-)(s^-)' \geqq 0 \quad \text{for all } x \in [0, \infty) \backslash \{x_1\},\ \hat{\varepsilon} \in (0, \varepsilon).$$

Thus if $\varepsilon < g(\infty)$, given any $\hat{\lambda} < \lambda_0 = \min(g(\infty) - \varepsilon, K)$, one can find $\hat{x}_1$ and $\hat{\nu}$ satisfying (4.13) and such that $s^-(\cdot, \hat{\lambda}, \hat{x}_1, \hat{\nu}) \leqq \psi$. Applying the comparison Theorem 3.4 we deduce that $s^-(\cdot, \hat{\lambda}, \hat{x}_1, \hat{\nu}) \leqq u$ (and thus that $\lambda_0 \leqq u(\infty, t)$ for all $t \leqq \infty$). Similarly one can check that $u \leqq s^+$.

Proof of Theorem 4.4. The uniqueness of the solution of problem H can be proven along the same lines as in the proof of [19, Thm. 1, Lemma 21]. Next we show its existence. Fix $I \geqq 1$. Since u and u_x are bounded uniformly in ε we deduce from Gilding [12] that u is equicontinuous on $\bar{D}_I$; thus, there exists a subsequence $\{u_{\varepsilon_n}\}_{n=I}^{\infty}$ of u and a function $\bar{u}_I \in C(\bar{D}_I)$, such that $u_{\varepsilon_n} \to \bar{u}_I$ as $\varepsilon_n \downarrow 0$ uniformly in $\bar{D}_I$ and such that for all $\lambda < K$, one can find x_1 and ν satisfying (4.13) and $s^-(\cdot, \lambda, x_1, \nu) \leqq \bar{u}_I(\cdot, t) \leqq s^+(\cdot)$. Then by a diagonal process, it follows that there exists a bounded continuous function $\bar{u}$ and a converging subsequence denoted by $\{u_{\varepsilon_k}\}$ such that $u_{\varepsilon_k} \to \bar{u}$ as $\varepsilon_k \downarrow 0$, pointwise on D and uniformly on all compact subsets of D. Since $0 \leqq (u_{\varepsilon_k})_x \leqq M_\psi$, $\bar{u}$ is nondecreasing in the x-direction and satisfies (ii); $u_{\varepsilon_k}(0) = 0$ implies the same property for $\bar{u}$. The boundary condition $\bar{u}(\infty, t) = K$ follows from the inequalities $s^-(\cdot, \lambda, x_1, \nu) \leqq \bar{u}(\cdot, t) \leqq s^+(\cdot)$ for all $\lambda < K$.

It remains to show that $\bar{u}$ is a generalized solution of H. Let $\phi \in C^1(\bar{D})$ vanish for large x and $t = T$, and let $L \geqq 1$ be such that ϕ vanishes in the neighborhood of $x = L$ and for $x > L$. Because the functions u_{ε_k} are classical solutions of P, we have

$$\iint_{D_L} \left[u_{\varepsilon_k} \phi_t - \varepsilon_k (x u_{\varepsilon_k x} - u_{\varepsilon_k}) \phi_x - \left(g - \frac{u_{\varepsilon_k}}{2}\right) u_{\varepsilon_k} \phi_x - u_{\varepsilon_k} g' \phi \right] dx\, dt + \int_0^L \psi(x) \phi(x, 0)\, dx = 0.$$

Now letting $\varepsilon_k \downarrow 0$ we deduce that $\bar{u}$ satisfies (4.12); because ϕ has been chosen arbitrarily we conclude that $\bar{u}$ is indeed the generalized solution of H and that $\{u_\varepsilon\}$ converges to $\bar{u}$ as $\varepsilon \downarrow 0$. □

5. Asymptotic stability of the steady state. Adapting a method due to Aronson and Weinberger [2] we investigate the stability of the solution Φ of problem P_0. To that purpose we consider the solution u of the corresponding evolution problem P; since its dependence on ψ plays a central role in what follows, we denote this solution by $u(x, t, \psi)$. We show that for all the functions ψ satisfying the hypothesis H_ψ given in the introduction we have that

$$u(x, t, \psi) \to \Phi(x) \quad \text{as } t \to \infty.$$

To begin with we prove two auxiliary lemmas.

LEMMA 5.1. (i) *Let* $\varepsilon < g(\infty)$ *and* $\hat{\lambda}, \hat{x}_1, \hat{\nu}$ *satisfy* (4.13). *The function* $u(x, t, s^-(\cdot, \hat{\lambda}, \hat{x}_1, \hat{\nu}))$ *is nondecreasing in time and such that*

$$\lim_{t\to\infty} u(x, t, s^-(\cdot, \hat{\lambda}, \hat{x}_1, \hat{\nu})) = \phi_{\hat{\lambda}}(x), \tag{5.1}$$

where $\phi_{\hat{\lambda}}$ *is the unique solution of*

$$\begin{aligned} &\varepsilon x\phi'' + (g(x) - \phi)\phi' = 0, \\ &\phi(0) = 0, \qquad \phi(\infty) = \hat{\lambda}. \end{aligned} \tag{5.2}$$

(ii) *The function* $u(x, t, s^+)$ *is nonincreasing in time. Furthermore*

$$\lim_{t\to\infty} u(x, t, s^+) = \Phi. \tag{5.3}$$

Proof. First note that it follows from the proofs in § 4 that problem P with initial value $s^-(x, \hat{\lambda}, \hat{x}_1, \hat{\nu})$ has a unique classical solution $u(x, t, s^-(\cdot, \hat{\lambda}, \hat{x}_1, \hat{\nu}))$ with $u(\infty, t) = \hat{\lambda}$ for all $t \leqq \infty$. Applying repeatedly Theorem 3.4, one can show that $u(x, t, s^-(\cdot, \hat{\lambda}, \hat{x}_1, \hat{\nu}))$ is nondecreasing in time and that $u(x, t, s^+)$ is nonincreasing in time; it also follows from Theorem 3.4 that

$$u(x, t, s^-(\cdot, \hat{\lambda}, \hat{x}_1, \hat{\nu})) \leqq \phi_{\hat{\lambda}}(x),$$

and that

$$u(x, t, s^+) \geqq \Phi(x).$$

Now for each x, $u(x, t, s^-(\cdot, \hat{\lambda}, \hat{x}_1, \hat{\nu}))$ is nondecreasing in t and bounded from above. Therefore it has a limit $\tau^-(x)$ as $t \to \infty$ and one can use standard arguments (see for example Aronson and Weinberger [2]) to show that $\tau^- \in C_{2+\alpha}((0, \infty)) \cap C([0, \infty))$ and satisfies the differential equation in (5.2) and the boundary conditions $\tau^-(0) = 0$ and $\tau^-(\infty) = \hat{\lambda}$. Finally since $\phi_{\hat{\lambda}}$ is the unique solution of problem (5.2) we have that $\tau^- = \phi_{\hat{\lambda}}$. Similarly one can show that $u(x, t, s^+)$ converges to a function $\tau^+ \in C_{2+\alpha}((0, \infty)) \cap C([0, \infty))$ which satisfies the steady state equation, the boundary condition $\tau^+(0) = 0$ and the condition $\Phi(\infty) \leqq \tau^+(\infty) \leqq K$. The fact that $\tau^+(\infty) = \Phi(\infty)$ follows from [6, Lemma 5.1]. Consequently $\tau^+ = \Phi$.

LEMMA 5.2. $\phi_{\hat{\lambda}}$ *is an increasing and continuous function of* $\hat{\lambda}$. *More precisely if* $\hat{\lambda}_1 \geqq \hat{\lambda}_2$ *we have*

$$0 \leqq \phi_{\hat{\lambda}_1} - \phi_{\hat{\lambda}_2} \leqq \hat{\lambda}_1 - \hat{\lambda}_2.$$

Proof. Let $m = \phi_{\hat{\lambda}_1} - \phi_{\hat{\lambda}_2}$. It satisfies the differential equation

$$\varepsilon x m'' + (g - \phi_{\hat{\lambda}_1})m' - \phi'_{\hat{\lambda}_2} m = 0$$

and the boundary conditions $m(0)=0$ and $m(\infty)=\hat{\lambda}_1-\hat{\lambda}_2\geqq 0$. Suppose that m attains a negative minimum at a certain point $\xi\in(0,\infty)$; then $m(\xi)<0$, $m'(\xi)=0$ and $m''(\xi)\geqq 0$ which is in contradiction with $\varepsilon\xi m''(\xi)=\phi'_{\hat{\lambda}_2}(\xi)m(\xi)$. Thus $m\geqq 0$. In the same way one can show that m cannot attain a positive maximum, which implies $m\leqq\hat{\lambda}_1-\hat{\lambda}_2$.

Finally we are in a position to prove the following theorem.

THEOREM 5.3. *Let* $\Phi(x)$ *be the solution of problem* P_0. *Suppose* ψ *satisfies the hypothesis* H_ψ, *then for each* $x\geqq 0$

$$\lim_{t\to\infty} u(x,t,\psi)=\Phi(x).$$

If $\varepsilon\leqq g(\infty)-K$ *the convergence is uniform on* $[0,\infty)$; *if* $\varepsilon>g(\infty)-K$ *it is uniform on all compact intervals of* $[0,\infty)$.

Proof. Since the functions u and u_x are bounded uniformly in t, we apply the Arzela-Ascoli theorem and a diagonal process to deduce that there exists a function $\tau\in C([0,\infty))$ and a sequence $\{u(t_n)\}$ with $u(t_n)=u(\cdot,t_n,\psi)$ such that $u(t_n)\to\tau$ as $t_n\to\infty$, uniformly on all compact subsets of $[0,\infty)$. Let $\varepsilon<g(\infty)$; then for each $\hat{\lambda}<\lambda_0=\min(g(\infty)-\varepsilon,K)$ one can find $\hat{\nu}$ and $\hat{x}_1$ satisfying (4.13) and such that $s^-(\cdot,\hat{\lambda},\hat{x}_1,\hat{\nu})\leqq\psi$. Applying Theorem 3.4 we obtain

$$u(x,t,s^-(\cdot,\hat{\lambda},\hat{x}_1,\hat{\nu}))\leqq u(x,t,\psi)\leqq u(x,t,s^+). \tag{5.4}$$

Letting $t\to\infty$ in (5.4) and applying Lemma 5.1 we obtain

$$\phi_{\hat{\lambda}}\leqq\tau\leqq\Phi \quad \text{for all } \hat{\lambda}<\lambda_0.$$

Next we deduce from Lemma 5.2 that

$$\Phi-\tau<\lambda_0-\hat{\lambda} \quad \text{for all } \hat{\lambda}<\lambda_0$$

and thus, that $\tau=\Phi$. If $\varepsilon\geqq g(\infty)$, then the inequalities

$$0\leqq u(x,t,\psi)\leqq u(x,t,s^+)$$

imply

$$0\leqq\tau\leqq\Phi=0.$$

Thus also in this case we have that $\tau=\Phi$. Finally we conclude that as $t\to\infty$, $u(\cdot,t,\psi)$ converges to Φ, uniformly on all compact intervals of $[0,\infty)$. This convergence result can be made slightly stronger in the case that $\varepsilon\leqq g(\infty)-K$: since then $\Phi(\infty)=K$ and since u is nondecreasing in x one can apply Diekmann [5, Lemma 2.4] to deduce that the convergence is uniform on $[0,\infty)$. □

6. Rate of convergence of the solution towards the steady state. In this section we analyze the rate of convergence of the solution u of P towards its steady state Φ. The results which we are able to derive depend strongly on the behavior of g as $x\to\infty$. If g tends to infinity fast enough, we can prove exponential convergence with a certain weighted norm. In the more general case, when $\varepsilon<g(\infty)-K$ we find that the solution converges algebraically fast towards its steady state on all finite x-intervals. No results are available in the case $\varepsilon\geqq g(\infty)-K$, which coincides with the physical situation when some (or all the) electrons escape to infinity.

We write

$$u(x,t,\psi)=\Phi(x)+v(x,t).$$

Then v satisfies the problem

$$\begin{aligned} &v_t = \varepsilon x v_{xx} + (g-\Phi)v_x - \Phi' v - v v_x, \\ &v(0,t) = 0, \\ &v(x,0) = \psi(x) - \Phi(x). \end{aligned} \tag{6.1}$$

Now let us make the change of function

$$v(x,t) = \exp\left(-\int_0^x \frac{g(\zeta)-\Phi(\zeta)}{2\varepsilon\zeta}\, d\zeta\right) \tilde{v}(x,t).$$

Problem (6.1) becomes

$$\begin{aligned} &\tilde{v}_t = \varepsilon x \tilde{v}_{xx} - q(x)\tilde{v} + h(x, \tilde{v}, \tilde{v}_x), \\ &\tilde{v}(0,t) = 0, \\ &\tilde{v}(x,0) = \exp\left(\int_0^x \frac{g(\zeta)-\Phi(\zeta)}{2\varepsilon\zeta}\, d\zeta\right)(\psi(x)-\Phi(x)), \end{aligned} \tag{6.2}$$

where

$$q(x) = \frac{(g(x)-\Phi(x))^2}{4\varepsilon x} + \frac{g'(x)+\Phi'(x)}{2} - \frac{g(x)-\Phi(x)}{2x}$$

and

$$h(x, \tilde{v}, \tilde{v}_x) = -\exp\left(-\int_0^x \frac{g(\zeta)-\Phi(\zeta)}{2\varepsilon\zeta}\, d\zeta\right)\tilde{v}\left(\tilde{v}_x - \frac{g(x)-\Phi(x)}{2\varepsilon x}\,\tilde{v}\right).$$

In particular, there exists $M > 0$ such that

$$|h(x, \tilde{v}, \tilde{v}_x)| \leqq M(\|\tilde{v}\|^2 + \|\tilde{v}_x\|^2), \qquad 0 < x < \infty,$$

where the notation $\|\cdot\|$ indicates the sup-norm.

In what follows we shall distinguish two cases: (i) the case when $\liminf_{x\to\infty} q(x) = \delta > 0$: this is so if $g(x) \geqq C_0\sqrt{x}$ for all $x \geqq x_2$ for some positive constants C_0 and x_2; (ii) the case when $\liminf_{x\to\infty} q(x) = 0$.

6.1. Case when g tends to infinity at least as fast as $\sqrt{x}$ for $x \to \infty$. The theorem we give next is very similar in its form and in its proof to a theorem of Fife and Peletier [10].

THEOREM 6.1. *Suppose that there exist constants x_2, $C_0 \geqq 0$ such that*

$$g(x) \geqq C_0\sqrt{x} \quad \textit{for all } x \geqq x_2. \tag{6.3}$$

Then there exist positive constants δ, μ, C such that if

$$\left\| \exp\left(\int_0^{\cdot} \frac{g(\zeta)-\Phi(\zeta)}{2\varepsilon\zeta}\, d\zeta\right)(\psi-\Phi) \right\| \leqq \delta$$

then

$$\left\| \exp\left(\int_0^{\cdot} \frac{g(\zeta)-\Phi(\zeta)}{2\varepsilon\zeta}\, d\zeta\right)(u(\cdot, t, \psi)-\Phi) \right\| \leqq C e^{-\mu t}, \qquad t \geqq 0,$$

where the notation $\|\cdot\|$ indicates the sup-norm.

Proof. To begin with we note that with the hypothesis of Theorem 6.1 we have that $v(\infty, t) = 0$ (since $\varepsilon < g(\infty) - K$) or equivalently

$$\lim_{x\to\infty} \exp\left(-\int_0^x \frac{g(\zeta)-\Phi(\zeta)}{2\varepsilon\zeta}\, d\zeta\right)\tilde{v}(x,t) = 0.$$

Next let us consider the boundary value problem

$$\varepsilon x w'' - (q(x)+\lambda)w = -\theta(\Phi'(\mathscr{R})+\lambda)\min\left(\tilde{\Phi}(x), (x/\mathscr{R})^{-\nu_0}\tilde{\Phi}(\mathscr{R})\right), w(0) = 0, \tag{6.4}$$

where

$$\tilde{\Phi}(x) = \exp\left(\int_0^x \frac{g(\zeta)-\Phi(\zeta)}{2\varepsilon\zeta}\, d\zeta\right)\Phi(x).$$

The right-hand side of the differential equation in (6.4) has been chosen in a special manner so that one can exhibit upper and lower solutions for a problem closely related to (6.4); more precisely we shall prove in the appendix that this problem has at least one solution $w \in C^2([0,\infty))$ with w, w' and w'' bounded such that

$$0 < w(x) \leqq \min\left(\tilde{\Phi}(x), \left(\frac{x}{\mathscr{R}}\right)^{-\nu_0}\tilde{\Phi}(\mathscr{R})\right)$$

for all constants $\nu_0 > 1$ provided that the constants $\theta \in (0,1)$, $\mathscr{R} > 0$ and $\lambda < 0$ satisfy certain conditions. We adjust θ such that $\|w\| + \|w'\| \leqq 1$.

We are now in a position to prove Theorem 6.1. Let

$$z(x,t) = \beta(w(x)+\gamma)\, e^{-\mu t},$$

in which β, γ and μ are positive constants still to be determined, and let

$$\mathscr{M}z = \varepsilon x z_{xx} - q(x)z + h(x, z, z_x) - z_t.$$

(i) The function q is positive for x near zero and, because of condition (6.3), also for large x; thus there exists $\bar{q}_0 > 0$ and ζ_1, $\zeta_2 \in (0,\infty)$ such that $\bar{q}_0 = \min\{q(x): x \in [0,\zeta_1]\cup[\zeta_2,\infty)\}$ is positive; therefore

$$\mathscr{M}z \leqq \beta\, e^{-\mu t}((\lambda+\mu)w + \gamma(-\bar{q}_0+\mu) + M\beta(1+\gamma)^2).$$

Choose

$$0 < \mu < \min(-\lambda, \bar{q}_0);$$

assume that γ is known (we shall specify it later), and choose

$$\beta = \frac{\gamma(\bar{q}_0-\mu)}{M(1+\gamma)^2}.$$

Then $\mathscr{M}z \leqq 0$ for all $x \in [0,\zeta_1]\cup[\zeta_2,\infty)$ and $t \geqq 0$,

(ii) Let $\zeta_1 \leqq x \leqq \zeta_2$; since $w(x) > 0$ on $(0,\infty)$, and since w is continuous we have

$$m = \min\{w(x): \zeta_1 \leqq x \leqq \zeta_2\} > 0.$$

Therefore

$$\mathscr{M}z \leqq \beta\, e^{-\mu t}((\lambda+\mu)m + \gamma(-\bar{q}+\mu) + M\beta(1+\gamma)^2),$$

where $\bar{q}$ is an arbitrary constant such that

$$\bar{q} < \min\{q(x): x \in [0,\infty)\}.$$

Hence

$$\mathcal{M}z \leqq \beta e^{-\mu t}((\lambda+\mu)m+\gamma(-\bar{q}+\bar{q}_0)).$$

Therefore if we choose

$$\gamma = -\frac{\lambda+\mu}{-\bar{q}+\bar{q}_0}m$$

we have

$$\mathcal{M}z \leqq 0 \quad \text{for } \zeta_1 \leqq x \leqq \zeta_2 \text{ and } t \geqq 0.$$

Thus for the above choice of β, γ and μ the function z is an upper solution of the equation $\mathcal{M}\tilde{v} = 0$. Let

$$\sup_{[0,\infty)} \tilde{v}(x, 0) \leqq \delta,$$

where $\delta = \beta\gamma$. Then

$$\tilde{v}(x, 0) \leqq z(x, 0) \quad \text{for all } x \in [0, \infty),$$

and hence by Theorem 3.4

$$\tilde{v}(x, t) \leqq z(x, t) \quad \text{for all } x \in [0, \infty), \quad t \geqq 0.$$

In a similar manner one can show that if

$$\inf_{[0,\infty)} \tilde{v}(x, 0) \geqq -\delta$$

then

$$\tilde{v}(x, t) \geqq -z(x, t) \quad \text{for all } x \in [0, \infty), \quad t \geqq 0.$$

Hence if $\|\tilde{v}(\cdot, 0)\| \leqq \delta$ then $\|\tilde{v}(\cdot, t)\| \leqq C e^{-\mu t}$ where we define

$$C = \beta(1+\gamma) = (1+1/\gamma)\delta. \qquad \square$$

6.2. Algebraic decay rate in the case that $\varepsilon < g(\infty) - K$. Provided that $\varepsilon < g(\infty) - K$ and that the initial function ψ converges algebraically fast to K as $x \to \infty$, we prove that the solution u of P converges algebraically fast to the steady state solution Φ for all finite values of x. To that purpose we show that a certain weighted space integral of the function $|u - \Phi|^p$, for some integer $p \geqq 1$, decays algebraically in time; a similar proof, with exponent $p = 1$, has been given, for example, by van Duyn and Peletier [9].

THEOREM 6.2. *Provided that $\varepsilon < g(\infty) - K$ and that $\psi \geqq s^-(\cdot, K, \bar{x}_1, \bar{\nu})$ for some $\bar{x}_1$, $\bar{\nu}$ satisfying* (4.13) *with $\lambda = K$, we have that*

$$\int_0^\infty (g'(x) + (p-1)\Phi'(x))|u(x, t, \psi) - \Phi(x)|^p \, dx$$

$$\leqq \left[\int_0^\infty ((s^+ - \Phi)^p + (\Phi - s^-)^p) \, dx\right] \Big/ t \tag{6.5}$$

for all $t > 0$ and $p = [1/\bar{\nu}] + 1$.

Proof. Since $|v(x, t)|^p \leqq (s^+(x) - s^-(x, K, \bar{x}_1, \bar{\nu}))^p$ it follows that $\int_0^\infty (v(x, t))^p \, dx$ is defined for all $t \geqq 0$. If $p \geqq 2$ let us multiply the differential equation in (6.1) by v^{p-1}

and integrate with respect to x; we obtain

$$\frac{d}{dt}\int_0^\infty \frac{v^p}{p}\,dx = [\varepsilon x v_x v^{p-1}]_0^\infty - \left[\varepsilon\frac{v^p}{p}\right]_0^\infty - \varepsilon(p-1)\int_0^\infty x v^{p-2}(v_x)^2\,dx$$
$$+\left[g\frac{v^p}{p}\right]_0^\infty - \int_0^\infty (g'+\Phi'(p-1))\frac{v^p}{p}\,dx - \left[\Phi\frac{v^p}{p}+\frac{v^{p+1}}{p+1}\right]_0^\infty$$

Since v tends to zero at least as fast as $x^{-\bar\nu}$ as $x\to\infty$, the equation above can be written in the simpler form

$$\frac{d}{dt}\int_0^\infty \frac{v^p}{p}\,dx = \left[\varepsilon x v_x v^{p-1}\right]_0^\infty - \varepsilon(p-1)\int_0^\infty x v^{p-2}(v_x)^2\,dx$$
$$-\int_0^\infty (g'+\Phi'(p-1))\frac{v^p}{p}\,dx. \tag{6.6}$$

Now let us define the functions v^+ and v^- as the solutions of (6.1) with initial values $v^+(x,0)=s^+(x)-\Phi(x)$ and $v^-(x,0)=s^-(x,K,\bar x_1,\bar\nu)-\Phi(x)$, respectively. By Theorem 3.4 we know that $v^+\geqq 0$ and $v^-\leqq 0$. Furthermore, it follows from Lemma 5.1 that v^+ is nonincreasing in time and v^- nondecreasing. Of course both v^+ and v^- satisfy (6.6) and in order to simplify this expression we use the following lemma which we shall prove later.

LEMMA 6.3. *Let* $\varepsilon < g(\infty)-K$. *Then* $\lim_{x\to\infty} x\Phi'(x)=0$. *If furthermore* $\psi\geqq s^-(\cdot,K,\bar x_1,\bar\nu)$ *for some* $\bar x_1$, $\bar\nu$ *satisfying* (4.13) *with* $\lambda=K$ (*we suppose furthermore that* $\bar\nu>1$ *if* $\varepsilon<(g(\infty)-K)/2$) *and* $\psi\in C_{1,\alpha}([x_3,\infty))$ *for some* $\alpha, x_3>0$, *then* $\lim_{x\to\infty} x u_x(x,t)=0$ *for all* $t\in(0,\infty)$.

From Lemma 6.3 and formula (6.6) we deduce that v^+ satisfies

$$\frac{d}{dt}\int_0^\infty \frac{(v^+)^p}{p}\,dx = -\varepsilon(p-1)\int_0^\infty x(v^+)^{p-2}(v_x^+)^2\,dx - \int_0^\infty (g'+\Phi'(p-1))\frac{(v^+)^p}{p}\,dx.$$

If $p=1$, similar calculations yield

$$\frac{d}{dt}\int_0^\infty v^+\,dx = -\int_0^\infty g' v^+\,dx.$$

Since $0<g'(x)<g'(0)$ and $0<\Phi'(x)<\Phi'(0)$, we have for all $p\geqq 1$

$$\int_0^\infty (v^+(x,t))^p\,dx \geqq \frac{1}{g'(0)+(p-1)\Phi'(0)}\int_0^\infty (g'(x)+(p-1)\Phi'(x))(v^+(x,t))^p\,dx,$$

and thus

$$\int_0^\infty (g'(x)+(p-1)\Phi'(x))(v^+(x,t))^p\,dx$$
$$\leqq (g'(0)+(p-1)\Phi'(0))\int_0^\infty (v^+(x,0))^p\,dx$$
$$-(g'(0)+(p-1)\Phi'(0))\int_0^t d\tau\int_0^\infty (g'(x)+(p-1)\Phi'(x))(v^+(x,\tau))^p\,dx.$$

In what follows we apply the following lemma that we shall prove later.

LEMMA 6.4. *Let* $y \in C([0, \infty))$ *with* $y' \in L^1((0, \infty))$ *and* $y' \leqq 0$ *such that*

$$0 \leqq y(t) \leqq N - M \int_0^t y(\tau)\, d\tau \tag{6.7}$$

for some constants $N \geqq 0$, $M > 0$. *Then*

$$y(t) \leqq \frac{N}{Mt}. \tag{6.8}$$

Since the function $\int_0^\infty (g'(x) + (p-1)\Phi'(x))(v^+(x, t))^p\, dx$ is continuous and nonincreasing (because v^+ is nonincreasing), we deduce from Lemma 6.4 that

$$\int_0^\infty (g'(x) + (p-1)\Phi'(x))(v^+(x, t))^p\, dx \leqq \left(\int_0^\infty (v^+(x, 0))^p\, dx \right) \Big/ t.$$

Similarly one can show that

$$\int_0^\infty (g'(x) + (p-1)\Phi'(x))(-v^-(x, t))^p\, dx \leqq \left(\int_0^\infty (-v^-(x, 0))^p\, dx \right) \Big/ t.$$

Formula (6.5) is then deduced from the fact that

$$|v(x, t)|^p \leqq \max\left((v^+(x, t))^p, (-v^-(x, t))^p\right) \leqq (v^+(x, t))^p + (-v^-(x, t))^p. \qquad \square$$

Proof of Lemma 6.3. We first show that $\lim_{x \to \infty} x\Phi'(x) = 0$. Since

$$\varepsilon x \Phi'(x) = \varepsilon \Phi(x) - \int_0^x (g(\zeta) - \Phi(\zeta))\Phi'(\zeta)\, d\zeta \leqq \varepsilon K,$$

we have

$$0 \leqq x\Phi'(x) \leqq K.$$

Furthermore

$$(x\Phi')' = x\Phi'' + \Phi' = -\frac{g - \Phi - \varepsilon}{\varepsilon}\Phi' \leqq 0 \quad \text{for } x \text{ large enough.}$$

Since the function $x\Phi'$ is bounded and decreasing for large x, we deduce that there exists $E \in [0, K]$ such that

$$\lim_{x \to \infty} x\Phi'(x) = E,$$

which implies

$$\Phi(x) \sim E \ln x + C \quad \text{as } x \to \infty.$$

Since

$$\lim_{x \to \infty} \Phi(x) = K,$$

we deduce that $E = 0$.

Next we show that $\lim_{x \to \infty} x u_x = 0$ by making use of Bernstein's argument, in a similar way as in Aronson [1] and Peletier and Serrin [21].
Let

$$R_n = \left(\frac{n}{2}, \frac{3n}{2}\right) \times (0, T], \qquad n > 3x_3$$

and let

$$\phi(r)=\frac{Nr(4-r)}{3},$$

where $N=\sup_{\bar{R}_n}u-\inf_{\bar{R}_n}u$. The function ϕ increases from 0 to N as r increases from 0 to 1. Note that $\phi'(r)=2N(2-r)/3>0$ and $\phi''(r)=-2N/3<0$ and define a new function w such that

$$u=\inf_{\bar{R}_n}u+\phi(w).$$

Then w satisfies the differential equation

$$w_t=\varepsilon x w_{xx}+\varepsilon x\frac{\phi''(w)}{\phi'(w)}(w_x)^2+(g-\phi(w)-\inf_{\bar{R}_n}u)w_x.$$

Set $p=w_x$ and differentiate the last equation with respect to x; we get

$$p_t=\varepsilon x p_{xx}+\varepsilon p_x+\varepsilon\frac{\phi''}{\phi'}p^2+2\varepsilon x\frac{\phi''}{\phi'}pp_x+\varepsilon x\left(\frac{\phi''}{\phi'}\right)'p^3$$
$$+(g-\phi-\inf_{\bar{R}_n}u)p_x+(g'-\phi'p)p,$$

and thus

$$\frac{1}{2}(p^2)_t-\varepsilon x p p_{xx}=\varepsilon x\left(\frac{\phi''}{\phi'}\right)'p^4+\varepsilon\left(\frac{\phi''}{\phi'}-\phi'\right)p^3$$
$$+2\varepsilon x\frac{\phi''}{\phi'}p^2p_x+(g-\phi-\inf_{\bar{R}_n}u+\varepsilon)pp_x+g'p^2. \tag{6.9}$$

Let $R_n^*=(3n/4,\,5n/4)\times(0,\,T]$, and let $\zeta=1-4(x-n)^2/n^2$. Set $z=\zeta^2p^2$.

(i) If z attains its maximum value at the lower boundary of R_n we have

$$\sup_{\bar{R}_n^*}z\leqq z(\tilde{x},0)\quad\text{where }\tilde{x}\in\left[\frac{n}{2},\frac{3n}{2}\right].$$

Hence,

$$\sup_{\bar{R}_n^*}\zeta|w_x|\leqq\zeta(\tilde{x})|w_x(\tilde{x},0)|.$$

Since $\zeta\geqq 3/4$ in $(3n/4,\,5n/4)$ and since $u_x=\phi'(w)w_x$ we find

$$\sup_{\bar{R}_n^*}|u_x|\leqq\frac{4}{3}\frac{\sup\phi'}{\inf\phi'}|\psi'(\tilde{x})|\leqq\frac{8M_\psi}{3}.$$

(ii) If z attains its maximum value at an interior point $(\tilde{x},\tilde{t})$ of R_n we have at that point

$$z_x=2\zeta\zeta'p^2+2\zeta^2pp_x=0,$$
$$\varepsilon x z_{xx}-z_t\leqq 0. \tag{6.10}$$

The last inequality can be cast in the more explicit form

$$\zeta^2(\tfrac{1}{2}(p^2)_t-\varepsilon x p p_{xx})\geqq\varepsilon x(\zeta'^2p^2+\zeta\zeta''p^2+4\zeta\zeta'pp_x+\zeta^2p_x^2).$$

Using (6.9), (6.10) and the inequality

$$|4\zeta\zeta' pp_x| \leqq \zeta^2 p_x^2 + 4\zeta'^2 p^2$$

we obtain

$$-\zeta^2 \varepsilon \left(\frac{\phi''}{\phi'}\right)' p^4 \leqq \left(-2\varepsilon\zeta\zeta' \frac{\phi''}{\phi'} + \varepsilon \frac{\zeta^2}{x}\frac{\phi''}{\phi'} - \frac{\zeta^2}{x}\phi'\right) p^3$$
$$+ \left(\zeta^2 \frac{g'}{x} + 3\varepsilon\zeta'^2 - \varepsilon\zeta\zeta'' - \frac{g - \phi - \inf_{\bar{R}_n} u + \varepsilon}{x}\zeta\zeta'\right) p^2.$$

Since $(\phi''/\phi')' \leqq -\frac{1}{4}$, this implies

$$2\zeta^2 p^4 \leqq \mathscr{C}_1 p^2 + \zeta \mathscr{C}_2 |p|^3,$$

where the $\mathscr{C}_i$'s are positive and depend only on N and n. Since

$$\zeta \mathscr{C}_2 |p|^3 \leqq \zeta^2 p^4 + \frac{\mathscr{C}_2^2}{4} p^2$$

it follows that

$$z(x, t) \leqq \max_{\bar{R}_n} (z(x, t)) \leqq \mathscr{C}_1 + \frac{\mathscr{C}_2^2}{4} \equiv \mathscr{C}_3.$$

Therefore

$$\max_{\bar{R}_n^*} |w_x| \leqq \frac{4\mathscr{C}_3^{1/2}}{3}.$$

Finally $u_x = \phi'(w) w_x$ and $\phi' \leqq 4N/3$ imply that

$$\max_{\bar{R}_n^*} |u_x| \leqq 16N\mathscr{C}_3^{1/2}/9.$$

Note that $N \leqq \sup_{\bar{R}_n}(K - s^-(x, K, \bar{x}_1, \bar{\nu}))$ (which behaves as $x^{-\bar{\nu}}$, where $\bar{\nu} > 0$) is furthermore such that $\bar{\nu} > 1$ if $\varepsilon < (g(\infty) - K)/2$.
Thus

$$\max_{\bar{R}_n^*} |u_x| \leqq 16\mathscr{C}_3^{1/2} \sup_{\bar{R}_n} \frac{(K - s^-(x, K, \bar{x}_1, \bar{\nu}))}{9}. \tag{6.11}$$

If $\varepsilon < (g(\infty) - K)/2$ $\mathscr{C}_3$ is bounded uniformly in n, and we deduce that xu_x tends to zero as $x \to \infty$. If on the other hand $(g(\infty) - K)/2 \leqq \varepsilon < g(\infty) - K$, then we only have that $\bar{\nu} > 0$ in (6.11) and $\sup_{\bar{R}_n}(K - s^-(x, K, \bar{x}_1, \bar{\nu}))$ tends to zero as $x \to \infty$. However $\mathscr{C}_3^{1/2}$ tends to zero as $1/x$ when $x \to \infty$, which also yields the result. □

Proof of Lemma 6.4. Integrating by parts we get

$$\int_0^t y(\tau)\, d\tau = ty(t) - \int_0^t \tau y'(\tau)\, d\tau \geqq ty(t).$$

Also we deduce from (6.7) that

$$\int_0^t y(\tau)\, d\tau \leqq \frac{N}{M},$$

and thus (6.8) follows. □

Next we deduce from Theorem 6.2 that there is also pointwise convergence. More precisely we prove the following theorem.

THEOREM 6.5. *Provided that* $\varepsilon < g(\infty) - K$ *and that* $\psi \geqq s^-(\cdot, K, \bar{x}_1, \bar{\nu})$ *for some* $\bar{x}_1, \bar{\nu}$ *satisfying* (4.13) *with* $\lambda = K$, *we have that*

$$\|(g'(\cdot) + (p-1)\Phi'(\cdot))^{1/p}(u(\cdot, t, \psi) - \Phi)\| \leqq \frac{C_\varepsilon}{t^{1/2p}} \quad \text{for all } t > 0, \tag{6.12}$$

and $p = [1/\bar{\nu}] + 1$, *where*

$$C_\varepsilon = \left[2\left(\left(K^{p-1}p^2 + K^p \frac{p-1}{\varepsilon}\right)(g'(0))^2 + K^p \sup_{x\in[0,\infty)} |g''(x)|\right) \cdot \int_0^\infty ((s^+ - \Phi)^p + (\Phi - s^-)^p)\, dx\right]^{1/2p}. \tag{6.13}$$

In particular, if $\varepsilon < (g(\infty) - K)/2$ and $\bar{\nu} > 1$, then $p = 1$ and formulas (6.12) and (6.13) simplify as follows

$$\|g'(\cdot)(u(\cdot, t, \psi) - \Phi)\| \leqq \frac{C}{\sqrt{t}} \quad \text{for all } t > 0, \tag{6.14}$$

where

$$C = \left[2\left((g'(0))^2 + K \sup_{x\in[0,\infty)} |g''(x)|\right)\int_0^\infty (s^+(x) - s^-(x, K, \bar{x}_1, \bar{\nu}))\, dx\right]^{1/2}.$$

Proof. To prove Theorem 6.5 we need the following auxiliary lemma.

LEMMA 6.6. *Let* ϕ *be defined for* $0 \leqq x < \infty$ *and satisfy the conditions*

(i) $\phi(x) \geqq 0$ *and* $\phi(0) = 0$;
(ii) ϕ *is Lipschitz continuous with constant* l;
(iii) $\int_0^\infty \phi(x)\, dx \leqq \mathcal{N}$.

Then

$$\sup_{0\leqq x<\infty} |\phi(x)| \leqq \sqrt{2\mathcal{N}l}.$$

We omit here the demonstration of this lemma since the main ideas of the proof are given in the proof of Peletier [20, Lemma 3].

Now let us apply Lemma 6.6 to the function $(g' + (p-1)\Phi')|u - \Phi|^p$; it is non-negative, equal to zero at the origin and its derivative is continuous by parts and bounded by

$$\left\{\left(K^{p-1}p^2 + K^p \frac{p-1}{\varepsilon}\right)(g'(0))^2 + K^p \sup_{x\in[0,\infty)} |g''(x)|\right\}$$

at all points where it is defined. Finally the bound on its integral is given in Theorem 6.2. Inequality (6.12) follows. □

6.3. Asymptotic behavior of the solution $\bar{u}$ of the hyperbolic problem H as $t \to \infty$.

THEOREM 6.7. *Let* ψ *satisfy* H_ψ *and be such that* $\psi \geqq s^-(\cdot, K, \bar{x}_1, \bar{\nu})$ *for some* $\bar{x}_1 > 0$, $\bar{\nu} > 1$ *satisfying* (4.13) *with* $\lambda = K$ *and define* $\bar{\Phi}(x) = \min(g(x), K)$. *Then*

$$\|g'(\cdot)(\bar{u}(\cdot, t, \psi) - \bar{\Phi})\| \leqq \frac{C}{\sqrt{t}} \quad \text{for all } t > 0,$$

where C is the constant defined in Theorem 6.5.

Proof. Let $\varepsilon \in (0, (g(\infty)-K)/2)\downarrow 0$ in inequality (6.14), note that the constant C does not depend in ε, and use the fact that Φ converges to $\bar{\Phi}$ uniformly on $[0, \infty)$ as $\varepsilon \downarrow 0$ (see [6]). □

Appendix. In what follows we shall prove the following theorem:

THEOREM A1. *Suppose that there exist constants x_2, $C_0 > 0$ such that the condition* (6.3) *is satisfied. There exist $\theta \in (0, 1)$, $\mathcal{R} > 0$ and $\lambda < 0$ such that the Cauchy–Dirichlet problem* (6.4) *has at least one solution $w \in C^2([0, \infty))$ with w, w', w'' bounded and*

$$0 < w(x) \leqq \min\,(\tilde{\Phi}(x), (x/\mathcal{R})^{-\nu_0}\tilde{\Phi}(\mathcal{R})) \quad \textit{for all } x \in (0, \infty).$$

Proof. Let $n \geqq 1$, and consider the boundary value problem

$$\text{(A1)}\quad \varepsilon\Big(x+\frac{1}{n}\Big)w'' - (q_n(x)+\lambda)w = -\theta(\Phi'(\mathcal{R})+\lambda)\min\,(\tilde{\Phi}_n(x), (x/\mathcal{R})^{-\nu_0}\tilde{\Phi}_n(\mathcal{R})),$$

$$\text{(A2)}\quad w(0) = 0,$$

where

$$\tilde{\Phi}_n(x) = \exp\Big(\int_0^x \frac{g(\zeta)-\Phi(\zeta)}{2\varepsilon(\zeta+1/n)}\,d\zeta\Big)\Phi(x),$$

and

$$q_n(x) = \frac{(g(x)-\Phi(x))^2}{4\varepsilon(x+1/n)} + \frac{g'(x)+\Phi'(x)}{2} - \frac{g(x)-\Phi(x)}{2(x+1/n)},$$

$\nu_0 > 1$ is arbitrary and where the constants $\theta \in (0, 1)$, $\mathcal{R} > 0$ and $\lambda \in (-\Phi'(\mathcal{R}), 0)$ satisfy some additional conditions which will be given later. Obviously zero is a lower solution for the differential equation in (A1). We shall now construct an upper solution. Firstly we deduce from the asymptotic behavior of g that there exists $\mathcal{R}_1 \geqq 1$ and $q_0 > 0$ such that $q_n(x) \geqq 2q_0$ for $x \geqq \mathcal{R}_1$. Also if $\lambda > \max\,(-q_0, -\Phi'(\mathcal{R}))$ and $\theta < (q_0+\lambda)/(\Phi'(\mathcal{R})+\lambda)$, then the function $(x/\mathcal{R})^{-\nu_0}\tilde{\Phi}_n(\mathcal{R})$ is an upper solution of the differential equation (A1) for $x \geqq \mathcal{R} := \max\,(\mathcal{R}_1, 2\varepsilon\nu_0(\nu_0+1)/q_0)$. Next we note that $\tilde{\Phi}_n$ is an upper solution of (A1) on $[0, \mathcal{R}]$ and thus that $\min\,(\tilde{\Phi}_n(x), (x/\mathcal{R})^{-\nu_0}\tilde{\Phi}_n(\mathcal{R}))$ is an upper solution of (A1) on $[0, \infty)$. Finally we conclude that there exists at least one solution $w_n \in C^2([0, \infty))$ of (A1), (A2) [3, Thm. 1.7.1], such that

$$0 \leqq w_n(x) \leqq \min\Big(\tilde{\Phi}_n(x), \Big(\frac{x}{\mathcal{R}}\Big)^{-\nu_0}\tilde{\Phi}_n(\mathcal{R})\Big),$$

which, since $\tilde{\Phi}_n \leqq \tilde{\Phi}$, implies that

$$\text{(A3)}\qquad 0 \leqq w_n(x) \leqq \min\Big(\tilde{\Phi}(x), \Big(\frac{x}{\mathcal{R}}\Big)^{-\nu_0}\tilde{\Phi}(\mathcal{R})\Big).$$

Furthermore, the inequalities (A3) and

$$\text{(A4)}\qquad |q_n(x)| \leqq \frac{(g-\Phi)^2}{4\varepsilon x} + \frac{g'+\Phi'}{2}$$

yield, together with (A1),

$$|w_n''(x)| \leqq C \quad \text{for all } x \in [0, \infty),$$

where $C>0$ is independent of n. Now let us integrate (A1); we get

$$w_n'(x) = w_n'(0) + \int_0^x \frac{(q_n(\zeta)+\lambda)w_n(\zeta) - \theta(\Phi'(\mathscr{R})+\lambda)\min(\tilde{\Phi}_n(\zeta), (\zeta/\mathscr{R})^{-\nu_0}\tilde{\Phi}_n(\mathscr{R}))\,d\zeta}{\varepsilon(\zeta+1/n)}, \tag{A5}$$

and again using (A3) and (A4) we obtain

$$|w_n'(x)| \leqq C \quad \text{for all } x \in [0,\infty].$$

Using the Arzela–Ascoli theorem and a diagonal process, we deduce that there exist a function $w \in C^1([0,\infty))$ and a subsequence $\{w_{n_k}\}$ of $\{w_n\}$ such that $w_{n_k} \to w$ as $n_k \to \infty$ uniformly in $C^1([0,\infty))$ on all compact subsets of $[0,\infty)$. Also setting $n = n_k$ in (A5) and letting $n_k \to \infty$, we deduce that w satisfies the differential equation

$$\varepsilon x w'' - (q(x)+\lambda)w = -\theta(\Phi'(\mathscr{R})+\lambda)\min(\tilde{\Phi}(x), (x/\mathscr{R})^{-\nu_0}\tilde{\Phi}(\mathscr{R})) \tag{A6}$$

and the boundary condition

$$w(0) = 0.$$

It follows from (A6) that $w \in C^2((0,\infty))$, and since

$$\lim_{x\to\infty} w''(x) = [(\Phi'(0)+\lambda)w'(0) - \theta(\Phi'(\mathscr{R})+\lambda)\tilde{\Phi}'(0)]/\varepsilon,$$

we deduce that in fact $w \in C^2([0,\infty))$. Finally the strict inequality $w>0$ is proven by means of a maximum principle argument. □

Acknowledgment. The author wishes to express her thanks to Professor L. A. Peletier whose advice has been invaluable for the completion of this work. It is a pleasure to acknowledge discussions with O. Diekmann and conversations with P. Wilders and A. Y. le Roux concerning the limit $\varepsilon \downarrow 0$.

REFERENCES

[1] D. G. ARONSON, *Regularity properties of flows through porous media*, SIAM J. Appl. Math., 17 (1969), pp. 461–467.

[2] D. G. ARONSON AND H. F. WEINBERGER, *Nonlinear diffusion in population genetics, combustion and nerve impulse propagation*, Lecture Notes in Mathematics 446, Springer, New York, 1975, pp. 5–49.

[3] S. R. BERNFELD AND V. LAKSHMIKANTHAM, *An Introduction to Nonlinear Boundary Value Problems*, Academic Press, New York, 1974.

[4] C. COSNER, *Asymptotic behavior of solutions of second order parabolic partial differential equations with unbounded coefficients*, J. Differential Equations, 35 (1980), pp. 407–428.

[5] O. Diekmann, *Limiting behavior in an epidemic model*, Nonlinear analysis TMA, 1 (1977), pp. 459–470.

[6] O. DIEKMANN, D. HILHORST AND L. A. PELETIER, *A singular boundary value problem arising in a prebreakdown gas discharge*, SIAM J. Appl. Math., 39 (1980), 48–66.

[7] C. J. VAN DUYN, *Regularity properties of solutions of an equation arising in the theory of turbulence*, J. Differential Equations, 33 (1979), pp. 226–238.

[8] ———, *On the diffusion of immiscible fluids in porous media*, this Journal, 10 (1979), pp. 486–497.

[9] C. J. VAN DUYN, AND L. A. PELETIER, *Asymptotic behaviour of solutions of a nonlinear diffusion equation*, Arch. Rat. Mech. Anal., 65 (1977), pp. 363–377.

[10] P. C. FIFE AND L. A. PELETIER, *Nonlinear diffusion in population genetics*, Arch. Rat. Mech. Anal., 64 (1977), pp. 93–109.

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

[12] B. J. GILDING, *Hölder continuity of solutions of parabolic equations*, J. London Math. Soc., 13 (1976), pp. 103–106.

[13] B. J. GILDING AND L. A. PELETIER, *The Cauchy problem for an equation in the theory of infiltration*, Arch. Rat. Mech. Anal., 61 (1976), pp. 127–140.

[14] A. M. IL'IN AND O. A. OLEINIK, *Asymptotic behaviour of solutions of the Cauchy problem for certain quasilinear equations for large values of time*, Mat. Sb., 51 (1960), pp. 191–216.

[15] P. M. IPPOLITO, *Maximum principles and classical solutions for degenerate parabolic equations*, J. Math. Anal. Appl. 64 (1978), pp. 530–561.

[16] O. A. LADYŽENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasi-linear Equations of Parabolic Type*, Trans. Math. Monographs 23, American Mathematical Society, Providence, RI, 1968.

[17] E. MARODE, *The mechanism of spark breakdown in air at atmospheric pressure between a positive point and a plane*; I: *Experimental: nature of the streamer track*; II: *Theoretical: computer simulation of the streamer track*, J. Appl. Phys., 46 (1975), pp. 2005–2015, pp. 2016–2020.

[18] E. MARODE, F. BASTIEN AND M. BAKKER, *A model of the streamer-induced spark formation based on neutral dynamics*, J. Appl. Phys., 50 (1979), pp. 140–146.

[19] O. A. OLEINIK, *Discontinuous solutions of non-linear differential equations*, AMS Transl., 26 (1963), pp. 95–172.

[20] L. A. PELETIER, *Asymptotic behavior of solutions of the porous media equation*, SIAM J. Appl. Math., 21 (1971), pp. 542–551.

[21] L. A. PELETIER AND J. SERRIN, *Gradient bounds and Liouville theorems for quasilinear elliptic equations*, Ann. Scuola Norm. Sup. Pisa Série IV, V (1978), pp. 65–104.

[22] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[23] P. WILDERS, private communication.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0003 $01.00/0

BOUNDED POSITIVE SOLUTIONS OF SEMILINEAR SCHRÖDINGER EQUATIONS*

C. A. SWANSON†

Abstract. The Schrödinger equation (1) $\Delta u + f(x, u) = 0$ is considered in an exterior domain Ω in R^n, $n \geqq 2$, where f is Hölder continuous and nonnegative and $f(x, u)/u$ is majorized above and below by nonnegative functions $g(|x|, u)$ which are monotone in u for $u > 0$, $|x| \geqq 0$. Conditions on f are found which are necessary and sufficient for (1) to have a uniformly positive bounded solution in $\Omega \subset R^2$, and corresponding results in $\Omega \subset R^n$, $n \geqq 3$. Such theorems constitute the only characterizations discovered to date of partial differential equations possessing positive solutions with specified behavior at ∞.

1. Introduction. Positive solutions $u(x)$ of the semilinear Schrödinger equation

$$Lu \equiv \Delta u + f(x, u) = 0, \qquad x \in \Omega \tag{1}$$

will be considered in exterior domains Ω of n-dimensional Euclidean space R^n, $n \geqq 2$, where f is Hölder continuous and nonnegative, and $f(x, u)/u$ is monotone in u for $u > 0$; detailed hypotheses are listed in § 2. Our main objective is to extend the known characterization of the existence of a bounded positive solution of the ordinary differential equation

$$\frac{d^2u}{dx^2} + ug(x, u) = 0 \tag{2}$$

to R^n, where $g(x, u)$ is either nondecreasing in u (superlinear case) or nonincreasing in u (sublinear case) for each $x \geqq 0$. Equation (2) was first studied in the case $ug(x, u) = p(x)u^\gamma$, $\gamma > 0$, when (2) is usually called the Emden–Fowler equation. The historical origin dates back to Lane [15], Emden [9] and Fowler [10], [11]. Excellent summaries of the applications to gas dynamics, fluid dynamics, astrophysics, relativistic mechanics, particle physics and chemistry have been given by Bellman [4], Conti, Graffi and Sansonne [8], Wong [24], with many additional references contained therein.

The one-dimensional theorem below, in the form given by Coffman and Wong [7], will be extended to n dimensions. Earlier similar results were proved by Atkinson [3], Belohorec [5], [6], Izyumova [13], Moore and Nehari [16] and Nehari [18].

THEOREM 1.1. *Let $f(x, u) = ug(x, u)$ be continuous and nonnegative for $0 < x < \infty$, $0 < u < \infty$, and suppose that $g(x, u)$ is either nonincreasing or nondecreasing in u for each x. Then* (2) *has a bounded positive solution in some interval (x_0, ∞), $x_0 > 0$ if and only if*

$$\int^\infty xg(x, c)\, dx < \infty \tag{3}$$

for some $c > 0$.

A positive solution $u(x)$ of (1) in $\Omega \subset R^2$ is always *uniformly positive* (see Lemma 4.2), but is not necessarily bounded; for example, the radial equation

$$\Delta u + r^{-2}(\log r)^{-3}u^3 = 0, \qquad r = |x| > 1$$

has the unbounded positive solution $u(x) = \frac{1}{2}(\log r)^{1/2}$, $r \geqq r_0 > 1$. It is therefore a natural question to ask for conditions analogous to (3) which are necessary and

* Received by the editors June 27, 1980, and in revised form March 17, 1981. This work was supported in part by the Natural Sciences and Engineering Research Council Canada under grant A3105.

† University of British Columbia, Vancouver, Canada V6T 1Y4.

sufficient for (1) to have a bounded positive solution in an exterior domain of R^2. The conclusions are given in §§ 3 and 4, also containing corresponding results in R^n, $n \geqq 3$. Our principal results, contained in Theorems 3.1, 4.3, and 4.4, are summarized in Theorem 1.2.

MAIN THEOREM 1.2. *Suppose that the function f in* (1) *satisfies the positivity, monotony and regularity conditions* (A), (B), (C) *and* (D) *listed in* § 2, *in an exterior domain $\Omega \subset R^n$. If $n = 2$, condition* (6) *is sufficient and condition* (12) *is necessary for the existence of a bounded uniformly positive solution $u(x)$ of equation* (1) *in Ω. If $n \geqq 3$, condition* (8) *is sufficient and* (16) *is necessary for* (1) *to have a positive solution $u(x)$ with $|x|^{n-2}u(x)$ bounded in Ω.*

In the case of Emden–Fowler type equations (21), this theorem generates necessary and sufficient conditions (22), (23) for the existence of a bounded positive solution in Ω (see Theorem 4.5). These are the only known characterizations, as far as we are aware, of partial differential equations possessing a positive solution with specified behavior at infinity.

2. Preliminaries. Points in Euclidean n-space R^n are denoted by $x = (x_1, x_2, \cdots, x_n)$ and the Euclidean norm of x is written $|x|$. The notation below will be used throughout:

$$S_a = \{x \in R^n : |x| = a\}, \qquad a > 0,$$

$$G_a = \{x \in R^n : |x| > a\}.$$

The measure on S_r and S_1 will be denoted by s and ω, respectively: $ds = r^{n-1}\, d\omega$.

For a bounded domain $M \subset R^n$, let $C^\alpha(\bar{M})$, $C^{\alpha+n}(\bar{M})$ denote the usual Hölder spaces, $0 < \alpha < 1$, $n = 1, 2, \cdots$ [14], [20].

An *exterior* domain Ω in R^n, $n \geqq 2$ has the property that $G_a \subset \Omega$ for some positive number a. For convenience, a function $g : [0, \infty) \times (0, \infty) \to [0, \infty)$ is called *monotone* in the second variable when $g(r, u)$ is either nondecreasing in u for each fixed r or nonincreasing in u for each fixed r. The Schrödinger equation (1) is to be considered in an exterior domain Ω under assumptions selected from the list below.

Assumptions.

(A) $f \in C^\alpha(\bar{M} \times \bar{J})$ for some α in $0 < \alpha < 1$, for every bounded domain $M \subset \Omega$ and for every bounded positive interval J.

(B) $f(x, u) \geqq 0$ for all $x \in \Omega$ and for all $u \geqq 0$.

(C) $f(x, u) \geqq u g_1(|x|, u)$ for all $x \in \Omega$, $u \geqq 0$, where $g_1(r, u)$ is continuous and nonnegative for $0 \leqq r < \infty$, $0 < u < \infty$ and monotone in u for each r.

(D) $f(x, u) \leqq u g_2(|x|, u)$ for all $x \in \Omega$, $u \geqq 0$, where $g_2 \in C^\alpha(\bar{I} \times \bar{J})$ for all bounded intervals $I = [0, b]$, $J = [a, b]$, $0 < a < b$, $0 < \alpha < 1$, and $g_2(r, u)$ is monotone in u for each $r \geqq 0$.

For example, all the assumptions (A)–(D) hold in the Emden–Fowler prototype $f(x, u) = p(x)u^\gamma$, $\gamma > 0$, where p is nonnegative in Ω and $p \in C^\alpha(\bar{M})$ for every bounded domain $M \subset \Omega$. In this case, we can take

$$g_1(r, u) = [\min_{|x|=r} p(x)]u^{\gamma-1},$$

$$g_2(r, u) = [\max_{|x|=r} p(x)]u^{\gamma-1}.$$

Then each $g_i(r, u)$, $i = 1, 2$ is increasing in u if $\gamma > 1$ (superlinear case) and decreasing in u if $0 < \gamma < 1$ (sublinear case). The existence theorem below was proved recently by E. S. Noussair and the author [21, p. 125].

THEOREM 2.1. *Let L, Ω and α be as above, and let a be a positive number such that $G_a \subset \Omega$. Under assumptions* (A) *and* (B), *if there exist positive solutions v, w of $Lv \leqq 0, Lw \geqq 0$, respectively in G_a, $v, w \in C^{2+\alpha}(\bar{M})$ for every bounded subdomain $M \subset G_a$ such that $w(x) \leqq v(x)$ throughout $G_a \cup S_a$, then* (1) *has at least one solution u satisfying $w(x) \leqq u(x) \leqq v(x)$ throughout $G_a \cup S_a$, $u \in C^{2+\alpha}(\bar{M})$ for every bounded $M \subset G_a$.*

Versions of this theorem for *bounded* domains had been given much earlier by Nagumo [17], Amann [2] and others.

In particular, let v be the function defined in G_a by $v(x) = \zeta(r)$, $r = |x| \geqq a$, where ζ is assumed to be a positive solution in the space $C^{2+\alpha}[a, b]$ for all $b > a$ of the ordinary differential equation

$$\frac{d}{dr}\left(r^{n-1}\frac{d\zeta}{dr}\right) + r^{n-1}\zeta(r)g_2(r, \zeta(r)) = 0. \tag{4}$$

COROLLARY 2.2. *Suppose that* (4) *has a positive solution $\zeta \in C^{2+\alpha}[a, b]$ for some $a > 0$ and for all $b > a$ such that $\zeta(r) \geqq Kr^{2-n}$, where K is some positive constant. Then, under assumptions* (A), (B) *and* (D), (1) *has a positive solution $u(x)$ satisfying $K|x|^{2-n} \leqq u(x) \leqq \zeta(|x|)$ for all $x \in G_a \cup S_a$, $n \geqq 2$.*

In fact, computation shows that

$$\begin{aligned} r^{n-1}Lv &= \frac{d}{dr}\left(r^{n-1}\frac{d\zeta}{dr}\right) + r^{n-1}f(x, v) \\ &\leqq \frac{d}{dr}\left(r^{n-1}\frac{d\zeta}{dr}\right) + r^{n-1}\zeta(r)g_2(r, \zeta(r)), \end{aligned}$$

and hence $Lv \leqq 0$ for all $x \in G_a$ by (4). Since $w(x) = Kr^{2-n}$, $r = |x|$ satisfies $Lw \geqq 0$, Corollary 2.2 follows from Theorem 2.1.

3. Sufficient conditions for bounded positive solutions. If $n = 2$, Liouville's change of variable $r = e^s$, $h(s) = \zeta(e^s)$ transforms (4) into the canonical form

$$h''(s) + e^{2s}h(s)g_2(e^s, h(s)) = 0. \tag{5}$$

By Theorem 1.1, (5) has a bounded positive solution $h(s)$ in some interval (s_0, ∞) if and only if

$$\int^{\infty} s\, e^{2s} g_2(e^s, c)\, ds < \infty$$

for some $c > 0$, which is equivalent to

$$\int^{\infty} r \log r\, g_2(r, c)\, dr < \infty \tag{6}$$

for some $c > 0$. Furthermore, standard regularity theory [14] shows that $h \in C^{2+\alpha}[s_0, s]$ for all $s > s_0$ if $g_2 \in C^{\alpha}$, as in assumption (D). If (6) holds, then a bounded positive solution $h(s)$ of (5) in (s_0, ∞) satisfies $h''(s) < 0$ by (5), from which $h'(s) > 0$ throughout (s_0, ∞) by a standard obvious argument. Consequently, there exist positive constants K_1 and K_2 such that $K_1 \leqq h(s) = \zeta(r) \leqq K_2$ for all r in $[a, \infty)$, where $a = \exp s_0$, and $\zeta \in C^{2+\alpha}[a, b]$ for all $b > a$. By Corollary 2.2, condition (6) is then sufficient for (1) to have a positive solution $u(x)$ in $G_a \subset R^2$ satisfying $K_1 \leqq u(x) \leqq K_2$ for some constants K_1 and K_2, $0 < K_1 < K_2$.

If $n \geqq 3$, the change of variables

$$r = \beta(s) = (\mu s)^{\mu}, \qquad h(s) = s\zeta(\beta(s)),$$

where $\mu = (n-2)^{-1}$, transforms (4) into

$$h''(s) + s^{-4}[\beta(s)]^{2n-2} h(s) g_2\left(\beta(s), \frac{h(s)}{s}\right) = 0. \tag{7}$$

Theorem 1.1 shows that (7) has a bounded positive solution $h(s)$ in some interval (s_0, ∞) if and only if

$$\int^{\infty} s^{-3}[\beta(s)]^{2n-2} g_2\left(\beta(s), \frac{c}{s}\right) ds < \infty,$$

for some $c > 0$, which is equivalent to

$$\int^{\infty} r g_2(r, cr^{2-n})\, dr < \infty, \qquad n \geqq 3 \tag{8}$$

for some $c > 0$. If (8) is satisfied, then, by exactly the same argument given above in the case $n = 2$, there exists a solution $\zeta(r)$ of (4) such that $K_1 r^{2-n} \leqq \zeta(r) \leqq K_2 r^{2-n}$ for all $r \geqq a$, where a, K_1 and K_2 are positive constants. Corollary 2.2 therefore shows that (8) is sufficient for (1) to have a positive solution $u(x)$ satisfying $K_1 r^{2-n} \leqq u(x) \leqq K_2 r^{2-n}$ for all $x \in G_a \cup S_a$, $n \geqq 3$, $0 < K_1 < K_2$. This establishes the theorem below.

THEOREM 3.1. *Under assumptions* (A), (B) *and* (D), (1) *has a bounded positive solution* $u(x)$ *in an exterior domain* $G_a \subset R^n$, *for some* $a > 0$, *with* $|x|^{n-2} u(x)$ *uniformly positive and bounded if* (6), (8) *hold for* $n = 2$, $n \geqq 3$, *respectively.*

4. Necessary conditions for bounded positive solutions. The spherical mean of a function $u: R^n \to R^1$ over the sphere S_r of radius r is defined by

$$U(r) = \frac{1}{s(S_r)} \int_{S_r} u(x)\, ds = \frac{1}{\omega(S_1)} \int_{S_1} u(x)\, d\omega. \tag{9}$$

A proof of the lemma below is given in [19, p. 70].

LEMMA 4.1. *If assumption* (C) *holds, then the spherical mean* $U(r)$ *of a positive-valued solution* $u(x)$ *of* (1) *in* G_a *satisfies the ordinary differential inequality*

$$-\frac{d}{dr}\left(r^{n-1} \frac{dU}{dr}\right) \geqq \frac{r^{n-1}}{\omega(S_1)} \int_{S_1} u(x) g_1(r, u(x))\, d\omega, \tag{10}$$

for $a < r < \infty$.

The next lemma is a special case of a result in [22, p. 917]; see also [1, p. 935], [23 p. 147].

LEMMA 4.2. *Every positive solution of the differential inequality* $\Delta u \leqq 0$ *in* $G_a \cup S_a$ *satisfies the inequality*

$$u(x) \geqq \left(\frac{a}{|x|}\right)^{n-2} \inf_{|x|=a} u(x), \qquad |x| \geqq a, \quad n \geqq 2. \tag{11}$$

THEOREM 4.3. *Under assumptions* (A), (B), (C), *a necessary condition for* (1) *to have a bounded positive solution in an exterior domain* $G_b \subset R^2$, $b > 0$, *is*

$$\int^{\infty} r \log r\, g_1(r, c)\, dr < \infty \tag{12}$$

for some positive number c.

Proof. If $u(x)$ is positive in $G_a \cup S_a$, $a > b$, Lemma 4.2 shows that $u(x)$ is uniformly positive for $|x| \geqq a$, $n = 2$, and hence, there are positive constants K_1 and K_2 such that $K_1 \leqq u(x) \leqq K_2$ for all $|x| \geqq a$. In the superlinear case, i.e., $g_1(r, u)$ nondecreasing in u, Lemma 4.1 then yields the inequality

$$-\frac{d}{dr}\left(r\frac{dU}{dr}\right) \geqq K_1 r g_1(r, K_1), \qquad r \geqq a. \tag{13}$$

In the sublinear case, i.e., $g_1(r, u)$ nonincreasing in u, (13) is replaced by

$$-\frac{d}{dr}\left(r\frac{dU}{dr}\right) \geqq K_1 r g_1(r, K_2), \qquad r \geqq a.$$

In both cases there exists a positive number c such that

$$-\frac{d}{dr}\left(r\frac{dU}{dr}\right) \geqq K_1 r g_1(r, c), \qquad r \geqq a. \tag{14}$$

Multiplication by $\log r$ and integration by parts over (a, r) gives

$$-r \log r\, U'(r) + a \log a\, U'(a) + U(r) - U(a) \geqq K_1 \int_a^r t \log t\, g_1(t, c)\, dt. \tag{15}$$

However, since $V(r) = rU'(r)$ is nonincreasing by (10), it follows that $U'(r) > 0$ for all $r > a$; in fact, if $V(R) < 0$ for some $R > a$, then

$$U(r) - U(R) = \int_R^r \frac{V(t)}{t}\, dt \leqq V(R) \log \frac{r}{R},$$

contradicting the positivity of $U(r)$ for all $r > a$. Since $U(r)$ is bounded, (15) implies the conclusion (12) of Theorem 4.3.

THEOREM 4.4. *Under assumptions* (A), (B), (C), *a necessary condition for* (1) *to have a positive solution* $u(x)$ *with* $|x|^{n-2}u(x)$ *bounded in an exterior domain* $G_b \subset R^n$, $n \geqq 3$, $b > 0$, *is*

$$\int^{\infty} r g_1(r, cr^{2-n})\, dr < \infty, \tag{16}$$

for some positive number c.

Proof. By Lemma 4.2, there exist positive constants K_1 and K_2 such that

$$K_1 |x|^{2-n} \leqq u(x) \leqq K_2 |x|^{2-n} \tag{17}$$

for all $|x| \geqq a$, where $a > b$. Then similarly to (14), assumption (C), (10) and (17) show that

$$-\frac{d}{dr}\left(r^{n-1}\frac{dU}{dr}\right) \geqq K_1 r g_1(r, cr^{2-n}), \qquad r \geqq a, \tag{18}$$

where $c = K_1$ or $c = K_2$ according as $g(r, u)$ is nondecreasing or nonincreasing in u, respectively. The change of variables

$$r = \beta(s) = (\mu s)^{\mu}, \qquad h(s) = sU(\beta(s)),$$

where $\mu = (n-2)^{-1}$, transforms (18) into

$$-h''(s) \geqq K_1 s^{-3} [\beta(s)]^n g_1(\beta(s), c[\beta(s)]^{2-n}), \tag{19}$$

$n \geqq 3$. We mulitiply (19) by s and integrate by parts over (A, s) to obtain

$$-sh'(s)+Ah'(A)+h(s)-h(A) \geqq K_1 \int_A^s s^{-2}[\beta(s)]^n g_1(\beta(s), c[\beta(s)]^{2-n})\, ds$$
$$= K_0 \int_a^r r g_1(r, cr^{2-n})\, dr, \tag{20}$$

where $a=\beta(A)$, $r=\beta(s)$ and K_0 is another positive constant. A standard argument similar to that used in the proof of Theorem 4.3 shows that $sh'(s)$ is bounded if $h(s)$ is positive and $h''(s)<0$. However, $h(s)=(n-2)r^{n-2}U(r)$ is positive and bounded by hypothesis, and therefore (20) implies (16). This completes the proof of Theorem 4.4.

In the case that (1) is the Emden–Fowler equation

$$\Delta u + p(x)u^\gamma = 0, \qquad \gamma > 0, \tag{21}$$

where $p(x)$ is nonnegative in Ω and $p \in C^\alpha(\bar{M})$ for every bounded domain $M \subset \Omega$, we define

$$P_1(r) = \min_{|x|=r} p(x), \qquad P_2(r) = \max_{|x|=r} p(x).$$

Then assumptions (C) and (D) hold, where

$$g_1(r, u) = P_1(r)u^{\gamma-1}, \qquad g_2(r, u) = P_2(r)u^{\gamma-1}.$$

Conditions (6), (8), (12), (16) of Theorems 3.1, 4.3 and 4.4 reduce to, respectively

$$\int^\infty r \log r P_2(r)\, dr < \infty, \qquad n = 2, \tag{22}$$

$$\int^\infty r^\sigma P_2(r)\, dr < \infty, \qquad \sigma = n-1-\gamma(n-2), \qquad n \geqq 3, \tag{23}$$

$$\int^\infty r \log r P_1(r)\, dr < \infty, \qquad n = 2, \tag{24}$$

$$\int^\infty r^\sigma P_1(r)\, dr < \infty, \qquad n \geqq 3. \tag{25}$$

Necessary and sufficient conditions for (21) to have a positive solution $u(x)$ with $|x|^{n-2}u(x)$ bounded in exterior domain are obtained under the additional hypothesis

$$\limsup_{r\to\infty} \frac{P_2(r)}{P_1(r)} < \infty. \tag{26}$$

THEOREM 4.5. *If* (26) *and the hypotheses accompanying* (21) *are fulfilled, a necessary and sufficient condition for* (21) *to have a positive solution* $u(x)$ *with* $|x|^{n-2}u(x)$ *bounded in an exterior domain* G_b *in* R^n *for some* $b>0$ *is* (22) *in the case* $n=2$, *or* (23) *in the case* $n \geqq 3$.

Proof. The sufficiency of (22), (23) is the content of Theorem 3.1. By assumption (26), (24) implies (22), and (25) implies (23). Hence, Theorems 4.3 and 4.4 establish the necessity of (22) and (23), respectively.

5. Concluding remarks. In the *superlinear* case $\gamma > 1$, conditions (22) and (23) also characterize nonoscillatory equations (21), i.e., equations (21) for which there exist positive solutions in some exterior domain [20, pp. 1001–1002]. This is not true

in the *sublinear* case $0<\gamma<1$; in fact it was shown recently by E. S. Noussair and the author [21, p. 132] that nonoscillatory sublinear equations (21) are characterized by the condition

$$\int^{\infty} rP_2(r)\,dr<\infty, \qquad n\geqq 3. \tag{27}$$

Since $\sigma=1+(n-2)(1-\gamma)>1$ in the sublinear case, condition (23) is not necessary for nonoscillation of (21). An analogue of (27) is not known if $n=2$, but a sufficient condition for nonoscillation of (21) is [21, p. 130], [23, p. 152],

$$\int^{\infty} r(\log r)^{\gamma}P_2(r)\,dr<\infty, \qquad n=2. \tag{28}$$

The necessary condition [22, p. 920], [23, p. 152]

$$\int^{\infty} rP_2(r)\,dr<\infty, \qquad n=2$$

can easily be improved, but it is not clear at this writing that (28) is necessary for nonoscillation (sublinear case).

Recently Gidas, Ni and Nirenberg [12] have given sufficient conditions for certain positive solutions of the autonomous equation $\Delta u+f(u)=0$ in R^n to be radially symmetric about a point. An explicit solution is demonstrated in the case $f(u)=u^{\gamma}$, $\gamma=(n+2)/(n-2)$, $n\geqq 3$. However, since $f(u)$ is independent of x in [12], the problem studied here does not arise; for example, in the case of (21) with $p(x)$ identically equal to 1, condition (22) is never satisfied and (23) is satisfied if and only if $\gamma>n/(n-2)$.

REFERENCES

[1] W. ALLEGRETTO, *Oscillation criteria for quasilinear equations*, Canad. J. Math., 26 (1974), pp. 931–947.

[2] H. AMANN, *On the existence of positive solutions of nonlinear elliptic boundary value problems*, Indiana Univ. Math. J., 21 (1971), pp. 125–146.

[3] F. V. ATKINSON, *On second order nonlinear oscillations*, Pacific J. Math., 5 (1955), pp. 643–647.

[4] R. BELLMAN, *Stability Theory of Differential Equations*, McGraw-Hill, New York, 1953.

[5] S. BELOHOREC, *Oscillatory solutions of certain nonlinear differential equations of second order*, Mat.-Fyz. Casopis Sloven. Akad. Vied., 11 (1961), pp. 250–255.

[6] ———, *Monotone and oscillatory solutions of a class of nonlinear differential equations*, Ibid., 19 (1969), pp. 169–187.

[7] C. V. COFFMAN AND J. S. W. WONG, *Oscillation and nonoscillation theorems for second order ordinary differential equations*, Funkcial. Ekvac., 15 (1972), pp. 119–130.

[8] R. CONTI, D. GRAFFI AND G. SANSONE, *The Italian contribution to the theory of nonlinear ordinary differential equations and to nonlinear mechanics during the years* 1951–1961, Qualitative Methods in the Theory of Nonlinear Vibrations, Proc. Internat. Sympos. Nonlinear Vibrations, vol. II, 1961, pp. 172–189.

[9] R. EMDEN, *Gaskugeln, Anwendungen der mechanischen Warmen-theorie auf Kosmologie und meterologische Probleme*, B. G. Teubner, Leipzig, 1907, Chap. XII.

[10] R. H. FOWLER, *The form near infinity of real, continuous solutions of a certain differential equation of the second order*, Quart. J. Math. 45 (1914), pp. 289–350.

[11] ———, *The solution of Emden's and similar differential equations*, Monthly Notices Roy. Astro. Soc., 91 (1930), pp. 63–91.

[12] B. GIDAS, W-M. NI AND L. NIRENBURG, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[13] D. V. IZYUMOVA, *On the conditions for oscillation and nonoscillation of solutions of nonlinear second order differential equations*, Differential Equations 2 (1966), pp. 814–821. (=Differencial'nye Uravnenija, 2 (1966), pp. 1572–1586.)

[14] O. A. LADYZHENSKAYA AND N. N. URAL'TSEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[15] I. J. HOMER LANE, *On the theoretical temperature of the sun under the hypothesis of a gaseous mass maintaining its volume by its internal heat and depending on the laws of gases known to terrestrial experiment*, Amer. J. Sci. and Arts, 4 (1869–70), pp. 57–74.

[16] R. A. MOORE AND Z. NEHARI, *Nonoscillation theorems for a class of nonlinear differential equations*, Trans. Amer. Math. Soc., 93 (1959), pp. 30–52.

[17] M. NAGUMO, *On principally linear elliptic differential equations of the second order*, Osaka Math. J., 6 (1954), pp. 207–229.

[18] Z. NEHARI, *On a class of nonlinear second order differential equations*, Trans. Amer. Math. Soc., 95 (1960), pp. 101–123.

[19] E. S. NOUSSAIR AND C. A. SWANSON, *Oscillation theory for semilinear Schrödinger equations and inequalities*, Proc. Roy. Soc. Edinburgh, *A*, 75 (1975/76), pp. 67–81.

[20] ———, *Positive solutions of semilinear Schrödinger equations in exterior domains*, Indiana Univ. Math. J., 28 (1979), pp. 993–1003.

[21] ———, *Positive solutions of quasilinear elliptic equations in exterior domains*, J. Math. Anal. Appl., 75 (1980), pp. 121–133.

[22] ———, *Oscillation of semilinear elliptic inequalities by Riccati transformations*, Canad. J. Math., 32 (1980), pp. 908–923.

[23] C. A. SWANSON, *Semilinear second order elliptic oscillation*, Canad. Math. Bull., 22 (1979), pp. 139–157.

[24] JAMES S. W. WONG, *On the generalized Emden-Fowler equation*, SIAM Rev., 17 (1975), pp. 339–360.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0004 $01.00/0

A NONLINEAR-BOUNDARY VALUE PROBLEM SUGGESTED BY THE LAPLACE* EQUATION FOR AN ELASTIC AND AXISYMMETRIC MEMBRANE

ANNE-MARIE LEFEVERE†

Abstract. This paper concerns the existence of solutions for a nonlinear boundary value problem related to the Laplace equation

$$A\theta+\frac{1}{\theta(1+\theta'^2)^{1/2}}=\frac{f}{T},\quad \text{where } A\theta=-\frac{d}{dx}\frac{\theta'}{(1+\theta'^2)^{1/2}}$$

and equations $A\theta+F(\theta+\theta_0)=g+F(\theta_0)$, related to the precedent one. These results are obtained by a regularization and the use of Galerkin and monotonicity methods. Maximal, minimal and periodic solutions are also studied.

1. Introduction. The starting point of this paper concerns the equilibrium shape of an elastic and axisymmetric membrane, governed by the Laplace equation

$$-\frac{d}{dx}\frac{\theta'}{(1+\theta'^2)^{1/2}}+\frac{1}{\theta(1+\theta'^2)^{1/2}}=\frac{f}{T},$$

subject to appropriate boundary conditions. Here, $\theta: x\to\theta(x)$ is the equation of the meridian, f is the jump of pressure and T is the surface tension. Such an equation arises in problems of crystal growth, for instance in Czochralski growth [10] and in floating zone melting [12]. In the present paper, we mainly study the existence and uniqueness of solutions θ for the following problem, related with the first one

$$A\theta+\frac{1}{\theta}-\frac{1}{\theta_0}=g,\qquad 0<x<L,$$

$$\theta(0)=\theta(L)=\theta_0>0,$$

where $A\theta=-(d/dx)\theta'/(1+\theta'^2)^{1/2}$, and g belongs to $L^q(]0,L[)$, $q>1$. For our study, there is no extra difficulty in considering a more general case, replacing $1/\theta-1/\theta_0$ by $F(\theta)-F(\theta_0)$, where F is a continuous decreasing and positive function on $]0,\infty[$.

Let $u=\theta-\theta_0$. We obtain the problem called $P(\theta_0, L, g)$,

$$Au+F(u+\theta_0)=g+F(\theta_0),$$

$$u(0)=u(L)=0.$$

However, the problem $P(\theta_0, L, g)$ in its variational formulation is not coercive on $W_0^{1,p}(\Omega)$ for $p>1$. (It is coercive on $W_0^{1,1}(\Omega)$, but $W_0^{1,1}(\Omega)$ is not a reflexive Banach space.) Therefore, we introduce the operator Δ_p such that

$$\Delta_p u=-\frac{d}{dx}(|u'|^{p-2}u'),$$

where $1<p<\infty$ and $p\geqq q'$, and we replace A by $A_\varepsilon=A+\varepsilon\Delta_p$, $\varepsilon>0$. We note that A_ε is coercive on $W_0^{1,p}(\Omega)$.

The regularized problem, which we call $P_\varepsilon(\theta_0, L, g)$, will be treated by a Galerkin method (applicable to the Laplace equation) and by a monotonicity method.

* Received by the editors February 8, 1980 and in final revised form March 25, 1981.

† Université de Pau, Faculté des Sciences Exactes, Dept. de Mathématiques, Avenue Philippon, 64000 Pau, France.

In the two cases, upper and lower solutions are used. However the existence of such functions is not sufficient to guarantee the existence of solutions for $P(\theta_0, L, g)$. They must satisfy additional conditions which allow to get estimates, independent of ε, on the derivatives of the approximate solutions, and then to pass to the limit when $\varepsilon \to 0$.

The organization of the paper is as follows.

Section 2: Notations and basic lemmas.

Section 3: Positive solutions by a Galerkin method.

Section 4: Maximal and minimal solutions.

In the two following sections one is interested in the case $F(u) = 1/u$.

Section 5: Some comments on the existence of upper and lower solutions. Conditions on the data L, θ_0, g, insuring the existence of upper and lower solutions which satisfy the hypotheses of the existence theorems for $P(\theta_0, L, g)$, are investigated.

Section 6: Periodic solutions for $A\theta + 1/\theta = f$. This section is an application of the results of § 4.

2. Notations and basic lemmas. $C_0^\infty(\Omega)$ is the vector space of C^∞ functions with compact support in Ω. We use classical notation and results concerning Sobolev spaces. Let $\Omega =]0, L[$. For $p \geqq 1$, $W^{m,p}(\Omega)$ denotes the Sobolev space of functions $u \in L^p(\Omega)$ such that the distributional derivatives $d^k u/dx^k \in L^p(\Omega)$, $1 \leqq k \leqq m$.

If $p = 2$, we write $W^{m,2}(\Omega) = H^m(\Omega)$. We recall that $W_0^{1,p}(\Omega) = \{u | u \in W^{1,p}(\Omega), u(0) = u(L) = 0\}$ and by the Poincaré inequality $\|u\|_{W_0^{1,p}(\Omega)} = (\int_\Omega |u'|^p \, dx)^{1/p}$ is an equivalent norm to $\|u\|_{W^{1,p}(\Omega)}$ on $W_0^{1,p}(\Omega)$. For $\infty > p \geqq 1$ the space $W^{-1,p'}$, $1/p + 1/p' = 1$, denotes the dual of $W_0^{1,p}(\Omega)$.

Note that, since $n = 1$, we have by the Sobolev embedding theorem

$$W^{m,p}(\Omega) \subset C^{m-1,1/p'}(\bar{\Omega}) \quad \text{(with continuous embedding)}. \tag{2.1}$$

For any pair of functions u, v in $L^p(\Omega)$, such that $u \leqq v$ a.e. in Ω, we define the convex set

$$[u, v] = \{z | z \in L^p(\Omega), u \leqq z \leqq v \text{ a.e. in } \Omega\}.$$

We now point out our definition of upper and lower solutions.

DEFINITION 2.1. A function $w \in W^{1,p}(\Omega)$, $p > 1$, is called a *lower solution* of $P(\theta_0, L, g)$ if $Aw \in L^{p'}(\Omega)$, $w + \theta_0 > 0$ and if

$$\begin{aligned} &Aw + F(w + \theta_0) \leqq g + F(\theta_0) \quad \text{a.e. in } \Omega, \\ &w(0) \leqq 0, \qquad w(L) \leqq 0. \end{aligned} \tag{2.2}$$

An upper solution is defined by reversing the above inequality signs. Observe that this definition does not imply that $w \in C^1(\bar{\Omega})$. Now we give some lemmas which will be used in the next sections.

LEMMA 2.1.

(i) *Let $V = W_0^{1,p}(\Omega)$, $1 < p < \infty$ and let $f \in V'$. Then there is a unique $u_\varepsilon \in V$ such that $A_\varepsilon u_\varepsilon = f$.*

(ii) *If further, $f \in L^q(\Omega)$, where $q \geqq p'$, then $u_\varepsilon \in W^{2,q}(\Omega)$, $Au_\varepsilon \in L^q(\Omega)$, $\Delta_p u_\varepsilon \in L^q(\Omega)$.*

Proof.

(i) This follows directly from [9, Thm 1.2, Chap. II].

Indeed A_ε, which maps V into V', is bounded. More precisely,

$$\|A_\varepsilon v\|_{V'} \leqq \varepsilon \|v\|_V^{p-1} + L^{1/p'},$$

$$\frac{(A_\varepsilon v, v)}{\|v\|_V} \to \infty \quad \text{if } \|v\|_V \to \infty, \tag{2.3}$$

A_ε is hemicontinuous, and since the application $h : X \to \varepsilon |X|^{p-2}X + X/(1+X^2)^{1/2}$ is strictly monotone on $\mathbb{R}$, A_ε is strictly monotone.

(ii) Since $h(u'_\varepsilon) \in L^{p'}(\Omega)$ and $A_\varepsilon u_\varepsilon \in L^q(\Omega)$, we have $h(u'_\varepsilon) \in W^{1,q}(\Omega)$ and $h(u'_\varepsilon) \in C^0(\bar{\Omega})$. Since h^{-1} exists, we write $u'_\varepsilon = h^{-1}[h(u'_\varepsilon)]$. h^{-1} is uniformly Lipschitzian on any bounded set of $\mathbb{R}$. Now using Stampacchia's results [15], we get $u'_\varepsilon \in W^{1,q}(\Omega)$. Therefore $u_\varepsilon \in W^{2,q}(\Omega)$ and so by (2.1) $u_\varepsilon \in C^1(\bar{\Omega})$. Next, again from the results of [15], it follows that $u'_\varepsilon/(1+u'^2_\varepsilon)^{1/2} \in W^{1,q}(\Omega)$. Thus $Au_\varepsilon \in L^q(\Omega)$ and $\Delta_p u_\varepsilon = A_\varepsilon u_\varepsilon - Au_\varepsilon \in L^q(\Omega)$.

Remark 2.1 If $f \in L^p(\Omega)$ and $f \geqq 0$, then $\Delta_p u_\varepsilon \geqq 0$ and $Au_\varepsilon \geqq 0$ a.e. in Ω.

The two following remarks follow from Stampacchia's results [15].

Remark 2.2. For a function $u \in W^{2,q}(\Omega)$, $q \geqq 1$, $Au = -u''/(1+u'^2)^{3/2}$.

Remark 2.3. If $u \in W^{1,\infty}(\Omega)$ and $Au \in L^q(\Omega)$, $q \geqq 1$, then $u \in W^{2,q}(\Omega)$.

Lemma 2.2 is a consequence of the weak maximum principle.

LEMMA 2.2. *Let* $1 < p < \infty$. *Let* $u, v \in W^{1,p}(\Omega)$ *satisfy* $A_\varepsilon u, A_\varepsilon v \in L^{p'}(\Omega)$,

$$A_\varepsilon v - A_\varepsilon u \geqq 0 \quad \text{a.e. in } \Omega,$$

$$v(0) \geqq u(0), \qquad v(L) \geqq u(L). \tag{2.4}$$

Then $v \geqq u$ (*we have the same result if* A_ε *is replaced by* A).

Proof. Since $v - u \in W^{1,p}(\Omega)$, $(v-u)^- \in W^{1,p}(\Omega)$ (cf. [15]) and

$$\frac{d}{dx}(v-u)^- = \begin{cases} -d/dx(v-u) & \text{if } v-u < 0 \\ 0 & \text{if } v-u \geqq 0 \end{cases} \quad \text{a.e. in } \Omega.$$

Multiplying (2.4) by $(v-u)^- \in W_0^{1,p}(\Omega)$, and integrating over Ω, we get

$$-\int_{\Omega^-} (h(v'(x)) - h(u'(x)))(v'(x) - u'(x))\, dx \geqq 0,$$

where $\Omega^- = \{x \in \Omega / v(x) - u(x) < 0\}$.

Hence, on Ω^-, $u - v$ is constant (locally) and this constant is necessarily 0.

The following lemma will supply a priori bounds on the derivatives of the approximate solutions.

LEMMA 2.3. *Let* $\infty > p > 1$, *and let* $u, v \in W_0^{1,p}(\Omega)$ *be such that* $\Delta_p u \in L^{p'}(\Omega)$, $\Delta_p u \geqq 0$ a.e. *in* Ω (resp. $\Delta_p u \leqq 0$) *and*

$$0 \leqq u \leqq v \quad (\text{resp. } v \leqq u \leqq 0). \tag{2.5}$$

Then $|u'|_{L^p(\Omega)} \leqq |v'|_{L^p(\Omega)}$.

Proof. Multiplication of (2.5) by $\Delta_p u$, integration by parts and the Hölder inequality yield

$$\int_\Omega |u'|^p\, dx \leqq \int_\Omega |u'|^{p-2} u'v'\, dx \leqq |u'|^{p-1}_{L^p(\Omega)} |v'|_{L^p(\Omega)}.$$

Then, using the Young inequality, we get

$$|u'|^p_{L^p(\Omega)} \leqq \frac{1}{p} |v'|^p_{L^p(\Omega)} + \frac{p-1}{p} |u'|^p_{L^p(\Omega)}.$$

Lemma 2.3 follows. (This result remains valid for $p = \infty$.)

For each function $z \in L^p(\Omega)$ such that $z + \theta_0 \geqq \alpha > 0$, let $B_\varepsilon z$ be the solution (existence and uniqueness are provided by Lemma 2.1) of the nonlinear boundary value problem

$$A_\varepsilon y_\varepsilon = g + F(\theta_0) - F(z + \theta_0) \quad \text{a.e. in } \Omega, \qquad y_\varepsilon(0) = y_\varepsilon(L) = 0. \tag{2.6}$$

The mapping B_ε will be used to define the sequences converging to the minimal and maximal solutions of $P(\theta_0, L, g)$.

LEMMA 2.4. *Let z_i be in $L^p(\Omega)$ and $z_i + \theta_0 \geqq \alpha > 0$, $i = 1, 2$. If $y_{i\varepsilon} = B_\varepsilon z_i$, $i = 1, 2$ and $z_1 \leqq z_2$ a.e. in Ω, then $y_{1\varepsilon} \leqq y_{2\varepsilon}$.*

Proof. We have

$$A_\varepsilon y_{1\varepsilon} - A_\varepsilon y_{2\varepsilon} = -F(\theta_0 + z_1) + F(\theta_0 + z_2), \qquad y_{i\varepsilon}(0) = y_{i\varepsilon}(L) = 0, \quad i = 1, 2.$$

Since F is decreasing and $z_1 \leqq z_2$,

$$A_\varepsilon y_{1\varepsilon} - A_\varepsilon y_{2\varepsilon} \leqq 0, \qquad y_{i\varepsilon}(0) = y_{i\varepsilon}(L) = 0, \qquad i = 1, 2.$$

Therefore by Lemma 2.2, $y_{1\varepsilon} \leqq y_{2\varepsilon}$.

3. Positive solutions by a Galerkin method. Consider the problem

$$Au + F(|u| + \theta_0) = g + F(\theta_0), \tag{3.1}$$

$$u(0) = u(L) = 0. \tag{3.2}$$

Any positive solution of (3.1), (3.2) is solution of $P(\theta_0, L, g)$ and conversely. Let $g \in L^q(\Omega)$, where $q > 1$. Let $p \geqq q'$ and $\infty > p > 1$. We study problem (3.1), (3.2) as the limit of the regularized problem $P'_\varepsilon(\theta_0, L, g)$

$$\varepsilon \Delta_p u_\varepsilon + Au_\varepsilon + F(|u_\varepsilon| + \theta_0) = g + F(\theta_0), \tag{3.3}$$

$$u_\varepsilon(0) = u_\varepsilon(L) = 0. \tag{3.4}$$

Now, suppose that u_ε is a smooth solution of (3.3), (3.4), then u_ε is a solution of the following variational problem.

Problem. Find $u_\varepsilon \in W_0^{1,p}(\Omega)$ such that

$$\varepsilon \int_\Omega |u'_\varepsilon|^{p-2} u'_\varepsilon v' \, dx + \int_\Omega \frac{u'_\varepsilon v'}{(1 + u'^2_\varepsilon)^{1/2}} \, dx + \int_\Omega F(|u_\varepsilon| + \theta_0) v \, dx$$

$$= \int_\Omega (g + F(\theta_0)) v \, dx \quad \forall v \in W_0^{1,p}(\Omega).$$

Conversely, by a standard argument, a solution of this problem satisfies (3.3) in the distribution sense in Ω and so is a weak solution.

In the sequel we shall solve $P'_\varepsilon(\theta_0, L, g)$ in its weak form.

THEOREM 3.1. *Let $g \in L^q(\Omega)$, where $q > 1$, and $g \geqq 0$. Then $P'_\varepsilon(\theta_0, L, g)$ has a positive solution $u_\varepsilon \in W_0^{1,p}(\Omega) \cap W^{2,q}(\Omega)$.*

Proof. We use the Galerkin method. Let $\{w_i\}$, a sequence of functions in $W_0^{1,p}(\Omega)$ such that for each m, $w_1, w_2, \cdots, w_m$ are linearly independent, and the linear combinations of the w_i are dense in $W_0^{1,p}(\Omega)$.

Next we seek $u_{\varepsilon m}$ such that

$$u_{\varepsilon m} = \sum_{i=1}^{m} \xi_{im} w_i,$$

$$(A_\varepsilon u_{\varepsilon m}, w_i) + (F(|u_{\varepsilon m}| + \theta_0), w_i) = (g + F(\theta_0), w_i), \qquad 1 \leqq i \leqq m. \tag{3.5}$$

The existence of $u_{\varepsilon m}$ follows from a fixed point theorem [9, Lemma 4.3 chap I]; further, there is a constant C_1, independent of m, such that for an "approximating solution" $u_{\varepsilon m}$ there holds

$$|u'_{\varepsilon m}|_{L^p(\Omega)} \leqq \frac{C_1}{\varepsilon^{1/p-1}}. \tag{3.6}$$

It follows that the sequence $u_{\varepsilon m}$ is bounded in $W_0^{1,p}(\Omega)$. Therefore, by (2.3)

$$\|A_\varepsilon u_{\varepsilon m}\|_{W^{-1,p'}(\Omega)} \leqq C_2. \tag{3.7}$$

By (3.6), (3.7) and by the compact embedding theorem of $W^{1,p}(\Omega)$ into $L^p(\Omega)$, there exists a subsequence (still denoted $u_{\varepsilon m}$) such that as $m \to \infty$

$$u_{\varepsilon m} \to u_\varepsilon \quad \text{weakly in } W_0^{1,p}(\Omega), \text{ strongly in } L^p(\Omega) \text{ and a.e. in } \Omega, \tag{3.8}$$

$$A_\varepsilon u_{\varepsilon m} \to \chi \quad \text{weakly in } W^{-1,p'}(\Omega). \tag{3.9}$$

(Indeed, here, since $n = 1$, (3.6) implies that $u_{\varepsilon m}$ is equicontinuous and there exists a subsequence which converges uniformly to u_ε.)

Passing to the limit $m \to \infty$ is now a standard matter. Let i be fixed and $m > i$. Letting $m \to \infty$ in (3.5) and using (3.8), (3.9) and the Lebesgue dominated convergence theorem we get

$$(\chi, w_i) = (g + F(\theta_0)) - (F(|u_\varepsilon| + \theta_0), w_i) \quad \forall i.$$

The finite linear combinations of the w_i are dense in $W_0^{1,p}(\Omega)$ so that

$$(\chi, v) = (g + F(\theta_0)) - (F(|u_\varepsilon| + \theta_0), v) \quad \forall v \in W_0^{1,p}(\Omega). \tag{3.10}$$

It remains to show that $\chi = A_\varepsilon u_\varepsilon$. This follows essentially from the monotonicity and hemicontinuity of A_ε (cf. [9, chap. II]). Then by restriction of (3.10) to functions v in $\mathscr{C}_0^\infty(\Omega)$, we find that

$$A_\varepsilon u_\varepsilon = g + F(\theta_0) - F(|u_\varepsilon| + \theta_0) \tag{3.11}$$

in the distributional sense in Ω. Since $g \in L^q(\Omega)$ and $F(|u_\varepsilon| + \theta_0) \in L^q(\Omega)$, equality (3.11) holds a.e. in Ω. $A_\varepsilon u_\varepsilon \in L^q(\Omega)$ and by our assumptions on g and F

$$A_\varepsilon u_\varepsilon \geqq 0 \quad \text{a.e. in } \Omega. \tag{3.12}$$

Then Lemma 2.2 yields $u_\varepsilon \geqq 0$.

Remark. We can give another proof of Theorem 3.1 by proving that $u \to A_\varepsilon u + F(|u| + \theta_0)$ is a pseudomonotone operator.

Now we study the behavior of u_ε, when $\varepsilon \to 0$. The a priori estimates for $P'_\varepsilon(\theta_0, L, g)$ do not allow the passage to the limit when $\varepsilon \to 0$. Under stronger assumptions than in Theorem 3.1, we establish the existence of a positive solution for $P(\theta_0, L, g)$.

THEOREM 3.2. *Let $g \in L^q(\Omega)$, $q > 1$, $g \geqq 0$. Suppose there exists $v \in W_0^{1,p}(\Omega)$, where $\infty > p > 1$ and $p \geqq q'$, such that $Av \in L^{p'}(\Omega)$ and*

$$Av \geqq g + F(\theta_0) \quad \text{a.e. } in\ \Omega. \tag{3.13}$$

Then $P(\theta_0, L, g)$ has a positive solution $u \in W_0^{1,p}(\Omega)$.

Proof. It follows from Theorem 3.1 that $P'_\varepsilon(\theta_0, L, g)$ has a positive solution u_ε such that $Au_\varepsilon \in L^q(\Omega)$, $\Delta_p u_\varepsilon L^q(\Omega)$. By (3.12) and Remark 2.

$$\Delta_p u_\varepsilon \geqq 0 \quad \text{a.e. in } \Omega. \tag{3.14}$$

By (3.11) and (3.14), we have

$$Au_\varepsilon \leqq g + F(\theta_0) - F(|u_\varepsilon| + \theta_0). \tag{3.15}$$

(3.13) and (3.15) yield

$$Av - Au_\varepsilon \geqq F(|u_\varepsilon| + \theta_0) \geqq 0.$$

Thus, by Lemma 2.2, $v \geqq u_\varepsilon$. On the other hand, since $u_\varepsilon \geqq 0$, $\Delta_p u_\varepsilon \in L^q(\Omega)$ and $\Delta_p u_\varepsilon \geqq 0$, we can apply Lemma 2.3 so that

$$|u'_\varepsilon|_{L^p(\Omega)} \leqq |v'|_{L^p(\Omega)}. \tag{3.16}$$

It follows from (3.16) that the u_ε solution of $P'_\varepsilon(\theta_0, L, g)$ is bounded uniformly in $W_0^{1,p}(\Omega)$ and that $|\varepsilon(\Delta_p u_\varepsilon, y)|$ is bounded by $\varepsilon |y'|_{L^p(\Omega)} |v'|^{p/p'}_{L^p(\Omega)}$, $\forall y \in W_0^{1,p}(\Omega)$.

There exists a subsequence (still denoted u_ε) such that u_ε converges to u weakly in $W_0^{1,p}(\Omega)$, uniformly on $\bar{\Omega}$, and Au_ε converges weakly star to χ in $W^{1,\infty}(\Omega)$. Passing to the limit in the variational formulation, as previously, we get

$$(Au, y) + F(|u| + \theta_0, y) = (g + F(\theta_0), y) \quad \forall y \in W_0^{1,p}(\Omega);$$

$u \in W_0^{1,p}(\Omega)$ and $u \geqq 0$. Then u is a solution of $P(\theta_0, L, g)$.

Remark. v being a particular upper solution of $P(\theta_0, L, g)$, we shall see that the previous result is contained in Theorem 4.1.

So this method can appear to be less interesting than the following one in § 4. However it applies, under slight modifications, to a fairly general problem involving the Laplace equation. Thus, we get the following.

THEOREM 3.3. *Let $g \in L^2(\Omega)$, $g \geqq 0$. Suppose there is $v \in H_0^1(\Omega) \cap H^2(\Omega)$ such that $Av \geqq g + 1/\theta_0$ a.e. in Ω. Then the problem*

$$Au + \frac{1}{(u + \theta_0)(1 + u'^2)^{1/2}} = g + \frac{1}{\theta_0}, \qquad u(0) = u(L) = 0$$

has a positive solution $u \in H^2(\Omega)$.

We must use an additional a priori estimate on $u''_{\varepsilon m}$ (obtained with a special basis for the Galerkin method) and on u''_ε and the rest of the proof is similar to the given one.

4. Maximal and minimal solutions. The construction of solutions by monotonicity methods has been used by many authors, but for problems involving a second order uniformly elliptic operator (cf. [1], [4], [5], [6], [13]). Further, here, under suitable assumptions, we can also obtain properties of monotonicity for the derivatives of the iterates, which provides upper bounds for them.

We first give an existence theorem for g positive or negative on Ω.

THEOREM 4.1. *Suppose that*

(i) $g \in L^q(\Omega)$, $q > 1$, $g \geqq 0$ (resp. $g \leqq 0$).

(ii) *There exists an upper solution v* (resp. *a lower solution w*) *of $P(\theta_0, L, g)$ such that v* (resp. w) *belongs to $W_0^{1,p}(\Omega)$, where $p > 1$ and $p \geqq q'$, and $v \geqq 0$ ($w \leqq 0$).*

Then $P(\theta_0, L, g)$ has a minimal solution $u_{\min}$ and a maximal solution $u_{\max}$ which belong to $W_0^{1,p}(\Omega) \cap [0, v]$ (resp. $W_0^{1,p}(\Omega) \cap [w, 0]$). *That is, if z is any solution of $P(\theta_0, L, g)$ such that $0 \leqq z \leqq v$ ($w \leqq z \leqq 0$) then $u_{\min} \leqq z \leqq u_{\max}$.*

Proof. We treat the case $g \geqq 0$. For $g \leqq 0$ the proof is identical with obvious reversals of inequalities. Using the mapping B_ε defined by (2.6), we introduce the

sequences $u_{\varepsilon n}$ and $v_{\varepsilon n}$ by means of

$$u_{\varepsilon n} = B_\varepsilon u_{\varepsilon n-1}, \quad \text{where } u_{\varepsilon 0} = 0,$$

$$v_{\varepsilon n} = B_\varepsilon v_{\varepsilon n-1}, \quad \text{where } v_{\varepsilon 0} = v.$$

Let us show that $u_{\varepsilon 1} \geqq u_{\varepsilon 0}$ and $v_{\varepsilon 1} \leqq v_{\varepsilon 0}$. We have

$$Av \geqq g + F(\theta_0) - F(v + \theta_0) = A_\varepsilon v_{\varepsilon 1} \quad \text{a.e. in } \Omega.$$

From hypotheses (i), (ii), we get $A_\varepsilon v_{\varepsilon 1} \geqq 0$. This yields $\Delta_p v_{\varepsilon 1} \geqq 0$, and therefore $A_\varepsilon v_{\varepsilon 1} \geqq A v_{\varepsilon 1}$. Then by Lemma 2.2, it follows that $v \geqq v_{\varepsilon 1}$ and since $A_\varepsilon u_{\varepsilon 1} \geqq 0$, $u_{\varepsilon 1} \geqq 0$.

Now suppose $v_{\varepsilon n} \leqq v_{\varepsilon n-1}$. Then by Lemma 2.4 $v_{\varepsilon n+1} = B_\varepsilon v_{\varepsilon n} \leqq B_\varepsilon v_{\varepsilon n-1} = v_{\varepsilon n}$. So by induction the sequence $v_{\varepsilon n}$, ε being fixed, is monotone decreasing. Similarly $u_{\varepsilon n}$, ε being fixed, defines a monotone increasing sequence. Further, since $u_{\varepsilon 0} \leqq v_{\varepsilon 0}$, $u_{\varepsilon n} \leqq v_{\varepsilon n}$ by the same arguments. Then

$$0 \leqq u_{\varepsilon 1} \leqq \cdots \leqq u_{\varepsilon n-1} \leqq u_{\varepsilon n} \leqq \cdots \leqq v_{\varepsilon n} \leqq v_{\varepsilon n-1} \leqq \cdots \leqq v. \tag{4.1}$$

Thus, by Lemma 2.3, the derivatives of the iterates verify here

$$|u'_{\varepsilon n-1}|_{L^p(\Omega)} \leqq |u'_{\varepsilon n}|_{L^p(\Omega)} \leqq \cdots \leqq |v'_{\varepsilon n}|_{L^p(\Omega)} \leqq |v'_{\varepsilon n-1}|_{L^p(\Omega)} \leqq \cdots \leqq |v'_\varepsilon|_{L^p(\Omega)}. \tag{4.2}$$

Hence,

$$\|u_{\varepsilon n}\|_{W_0^{1,p}(\Omega)} \leqq \|v_{\varepsilon n}\|_{W_0^{1,p}(\Omega)} \leqq |v'|_{L^p(\Omega)}. \tag{4.3}$$

Furthermore we have, for all n and for $\varepsilon \geqq \nu > 0$,

$$v_{\varepsilon n} \leqq v_{\nu n}, \qquad u_{\varepsilon n} \leqq u_{\nu n}. \tag{4.4}$$

In fact, $v_{\varepsilon 0} = v_{\nu 0}$. Suppose $u_{\varepsilon n-1} \leqq u_{\nu n-1}$, then $A_\varepsilon u_{\varepsilon n} - A_\nu u_{\nu n} \leqq 0$. Since $A_\varepsilon u_{\varepsilon n} \geqq 0$, $\Delta_p u_{\varepsilon n} \geqq 0$. Therefore $A_\varepsilon u_{\varepsilon n} \geqq A_\nu u_{\varepsilon n}$. So (4.4) follows from Lemma 2.2.

Now, n being fixed, let $\varepsilon \to 0$. The sequence $v_{\varepsilon n}$ is uniformly bounded and equicontinuous because of (4.1), (4.3). Since it is monotone by (4.4), the full sequence $v_{\varepsilon n}$ (not merely a subsequence) converges uniformly on $\bar{\Omega}$ by the Ascoli–Arzela theorem, and by (4.3) it converges weakly in $W_0^{1,p}(\Omega)$. Let $v_n = \lim_{\varepsilon \to 0} v_{\varepsilon n}$. $v_n \in W_0^{1,p}(\Omega)$ and, by the same arguments as in § 3, v_n satisfies

$$Av_n = g + F(\theta_0) - F(v_{n-1} + \theta_0) \quad \text{a.e. in } \Omega,$$

$$v_n(0) = v_n(L) = 0.$$

From (4.1) we deduce that the sequence v_n is still monotone. In the same way the full sequence v_n converges to $u_{\max}$, where $u_{\max} \in W_0^{1,p}(\Omega)$, and is a solution of $P(\theta_0, L, g)$. (Similarly $u_n = \lim_{\varepsilon \to 0} u_{\varepsilon n}$ converges to $u_{\min} \in W_0^{1,p}(\Omega)$ and is a solution of $P(\theta_0, L, g)$).

Now, let z be a solution of $P(\theta_0, L, g)$ in $W_0^{1,p}(\Omega)$ such that

$$u_0 = 0 \leqq z \leqq v = v_0.$$

An induction yields $u_n \leqq z \leqq v_n$. Hence, $u_{\min} \leqq z \leqq u_{\max}$. Observe that we cannot establish $z \leqq v_{\varepsilon n}$ for all $\varepsilon > 0$ and all n.

If we suppose only $g \in L^q(\Omega)$, where $q > 1$, an existence theorem can be proved under appropriate conditions on the upper and lower solutions.

THEOREM 4.2. *Let $g \in L^q(\Omega)$, where $q > 1$. Let v and w be upper and lower solutions of $P(\theta_0, L, g)$ in $W^{1,p}(\Omega)$, where $\infty > p > 1$ and $p \geqq q'$, such that*

$$\Delta_p v, \Delta_p w \in L^{p'}(\Omega), \quad \Delta_p w \leqq 0, \quad \Delta_p v \geqq 0 \quad \text{a.e. in } \Omega. \tag{4.5}$$

$$\max\left(\left|\int_\Omega Aw\,dx\right|, \int_\Omega Av\,dx\right) \leqq k < 1. \tag{4.6}$$

Then $P(\theta_0, L, g)$ has a maximal solution and a minimal solution in $W_0^{1,\infty}(\Omega) \cap [w, v]$.

Proof. As in Theorem 4.1, we define two sequences $u_{\varepsilon n}$ and $v_{\varepsilon n}$ by $u_{\varepsilon n} = B_\varepsilon u_{\varepsilon n-1}$ and $v_{\varepsilon n} = B_\varepsilon v_{\varepsilon n-1}$, where $u_{\varepsilon 0} = w$ and $v_{\varepsilon 0} = v$. Using (4.5), we see that v and w are upper and lower solutions in $C^1(\bar{\Omega})$ for $P_\varepsilon(\theta_0, L, g)$, for any $\varepsilon > 0$, and that $w \leqq 0 \leqq v$. Then by induction

$$w \leqq u_{\varepsilon 1} \leqq \cdots \leqq u_{\varepsilon n-1} \leqq u_{\varepsilon n} \leqq \cdots \leqq v_{\varepsilon n} \leqq v_{\varepsilon n-1} \leqq \cdots \leqq v_{\varepsilon 1} \leqq v. \tag{4.7}$$

In this case the a priori bounds on the derivatives of the iterates do not result of monotonicity properties for these last.

Since $u_{\varepsilon n} = B_\varepsilon u_{\varepsilon n-1}$, $u_{\varepsilon n} \in C^1(\bar{\Omega})$ by Lemma 2.1. Then there exists $\lambda_{\varepsilon n} \in \Omega$ such that $u'_{\varepsilon n}(\lambda_{\varepsilon n}) = 0$. Integrating (2.6) over $[x, \lambda_{\varepsilon n}]$ (with $u_{\varepsilon n}$ and $u_{\varepsilon n-1}$ replacing y_ε and z) we get

$$\varepsilon u'_{\varepsilon n}(x)|u'_{\varepsilon n}(x)|^{p-2} + \frac{u'_{\varepsilon n}(x)}{(1+u'^2_{\varepsilon n}(x))^{1/2}} = \int_x^{\lambda_{\varepsilon n}} (g + F(\theta_0) - F(u_{\varepsilon n-1} + \theta_0))\,dt. \tag{4.8}$$

By the hypotheses on F, (4.7) and Definition 2.1

$$\begin{aligned} Aw &\leqq g + F(\theta_0) - F(w + \theta_0) \leqq g + F(\theta_0) - F(u_{\varepsilon n+1} + \theta_0) \\ &\leqq g + F(\theta_0) - F(v + \theta_0) \leqq Av. \end{aligned} \tag{4.9}$$

Then from (4.6), (4.8), (4.9), we finally get

$$|u'_{\varepsilon n}(x)| \leqq \frac{k}{(1-k^2)^{1/2}}. \tag{4.10}$$

Because of (4.10) $u_{\varepsilon n}$ and $v_{\varepsilon n}$ are uniformly bounded in $W_0^{1,\infty}(\Omega)$. Let $\varepsilon \to 0$. By the classical method of the diagonal, we construct a sequence u_n. From the sequence $u_{\varepsilon 1}$, we can extract a subsequence $u_{\varepsilon_1,1}$ which converges uniformly to u_1. From the sequence $u_{\varepsilon_1,2}$, we extract a subsequence $u_{\varepsilon_2,2}$ converging to u_2. By repeating this method, the sequence $u_{\varepsilon_n,n}$ extracted from $u_{\varepsilon_{n-1},n}$ converges uniformly to u_n. $u_n \in W_0^{1,\infty}(\Omega)$ and verifies

$$Au_n = g + F(\theta_0) - F(u_{n-1} + \theta_0).$$

(Since A is strictly monotone, any converging subsequence extracted from $u_{\varepsilon_{n-1},n}$ converges to the same function u_n.) Similarly we construct a sequence v_n. These two sequences are still monotone. Letting $n \to \infty$, they converge uniformly to $u_{\min}$ and $u_{\max}$ which belong to $W_0^{1,\infty}(\Omega)$ and are solutions of $P(\theta_0, L, g)$. By Remark 2.3, this yields $u_{\min}, u_{\max} \in C^1(\bar{\Omega})$.

As in Theorem 4.1, if z is a solution of $P(\theta_0, L, g)$ such that $w \leqq z \leqq v$, then $u_{\min} \leqq z \leqq u_{\max}$.

Under somewhat restrictive assumptions we give a second method of monotone sequences. In the present scheme our hypotheses are as follows

$$g \in L^2(\Omega), \quad g \geqq 0 \text{ (resp. } g \leqq 0), \quad g(x) = g(L-x) \quad \text{a.e. in } \Omega. \tag{4.11}$$

(4.12) There is an upper solution v (a lower solution w) of $P(\theta_0, L, g)$ such that $v \in H^2(\Omega)$, $v \geqq 0$ (resp. $w \leqq 0$) and $v(x) = v(L-x)$ $(w(x) = w(L-x))$.

For each function $z \in C^1(\bar{\Omega}) \cap [0, v]$ $(C^1(\bar{\Omega}) \cap [w, 0])$, define the image y of the mapping T to be the solution of the Dirichlet problem

$$-y'' = (g + F(\theta_0) - F(z + \theta_0))(1 + z'^2)^{3/2} \quad \text{a.e. in } \Omega, \qquad y \in H^2(\Omega) \cap H_0^1(\Omega).$$

We first establish

LEMMA 4.1. *Suppose* $g \in L^2(\Omega)$, $g \geqq 0$ *and* $g(x) = g(L-x)$. *Let* z_i *be in* $C^1(\bar{\Omega}) \cap [0, v]$, *and let* $y_i = Tz_i$, $i = 1, 2$. *If* $z_1(x) \leqq z_2(x)$, $|z_1'(x)| \leqq |z_2'(x)|$ *and* $z_i(x) = z_i(L-x)$, $x \in \bar{\Omega}$, *then* $y_1(x) \leqq y_2(x)$, $|y_1'(x)| \leqq |y_2'(x)|$ *and* $y_i(x) = y_i(L-x)$ *for* $x \in \bar{\Omega}$. (*For* $g \leqq 0$ *the inequalities relative to the derivatives are reversed.*)

Proof. Let $\phi = y_1 - y_2$. By our assumptions on g, y_i and y_i', $i = 1, 2$. ϕ satisfies

$$-\phi'' \leqq 0 \quad \text{a.e. in } \Omega, \qquad \phi(0) = \phi(L) = 0. \tag{4.13}$$

Therefore, $y_1 \leqq y_2$. Since $g(x) = g(L-x)$ and $z_i(x) = z_i(L-x)$, it is easily seen that $y_i(x) = y_i(L-x)$ (which yields $y_i'(x) = -y_i'(L-x)$). Thus by integrating (4.13) over $[x, L/2]$ and since $-y_i'' \geqq 0$, we get $|y_1'(x)| \leqq |y_2'(x)|$. Now we state the following result.

THEOREM 4.3. *Let* (4.11) *and* (4.12) *be satisfied. Let* u_n *and* v_n *defined by* $u_n = Tu_{n-1}$, $v_n = Tv_{n-1}$ *where* $u_0 = 0$ (resp. $u_0 = w$) *and* $v_0 = v$ (resp. $v_0 = 0$). *Then the sequences* u_n *and* v_n *converge uniformly and monotonically to* $u_{\min}$ *and* $u_{\max}$, *which belong to* $H^2(\Omega) \cap [0, v]$ (resp. $H^2(\Omega) \cap [w, 0]$) *and are solutions of* $P(\theta_0, L, g)$, *giving*

$$u_0 \leqq u_1 \leqq \cdots \leqq u_n \leqq \cdots \leqq u_{\min} \leqq u_{\max} \leqq \cdots \leqq v_n \leqq \cdots \leqq v_1 \leqq v_0. \tag{4.14}$$

Further,

$$|'_0| \leqq |u_1'| \leqq \cdots \leqq |u_n'| \leqq \cdots \leqq |u_{\min}'| \leqq |u_{\max}'| \leqq \cdots \leqq |v_n'| \leqq \cdots \leqq |v_1'| \leqq v_0'|. \tag{4.15}$$

(*The inequalities of* (4.15) *are reversed for* $g \leqq 0$.)

Proof. It is easily seen that the pair of functions v_1, v_0 satisfies the hypotheses of Lemma 4.1 (v_1, v_0 replacing z_1, z_2), as well as u_0, u_1 and u_0, v_0. We have (4.14), (4.15) by an induction argument using Lemma 4.1. Since the sequences u_n and u_n' are uniformly bounded, the sequence u_n is equicontinuous and $-u_n'' = (g + F\theta_0 - F(u_{n-1} + \theta_0))(1 + u_{n-1}'^2)^{3/2}$ is bounded in $L^2(\Omega)$. Hence the sequence u_n' is also equicontinuous. An application of the Ascoli–Arzela theorem shows that the full sequence u_n converges uniformly on $\bar{\Omega}$ to $u_{\min}$ and that the full sequence of the derivatives converges uniformly on $\bar{\Omega}$ to $u_{\min}'$.

Next we examine the integral equation which is equivalent to $u_n = Tu_{n-1}$,

$$u_n(x) = \int_\Omega G(x, t)(g + F(\theta_0) - F(u_{n-1}(t) + \theta_0))(1 + u_{n-1}'^2(t))^{3/2}\, dt,$$

where G is the Green's function of the Laplace operator relative to the boundary conditions $u_n(0) = u_n(L) = 0$. Letting $n \to \infty$ and applying the Lebesgue dominated convergence theorem, we see that $u_{\min} = \lim u_n$ is solution of $P(\theta_0, L, g)$. As previously, we prove by induction that $u_{\min}$ and $u_{\max}$ are minimal and maximal solutions of $P(\theta_0, L, g)$ in $[u_0, v_0]$.

This monotonicity method can be adapted to the Laplace equation. In the definition of the mapping T, $F(z+\theta_0)$ is obviously replaced by $1/(z+\theta_0)(1+z'^2)^{1/2}$ and Lemma 4.1 remains valid for $g \geqq 0$ or $g \leqq -1/\theta_0$. We get

THEOREM 4.4. *Let $g \in L^2(\Omega)$ be such that $g \geqq 0$ (resp. $g \leqq -1/\theta_0$) and $g(x) = g(L-x)$ a.e. in Ω. Suppose there is $v \in H^2(\Omega)$, $v \geqq 0$ (resp. $v \leqq 0$) such that*

$$Av + \frac{1}{(v+\theta_0)(1+v'^2)^{1/2}} \geqq g + \frac{1}{\theta_0} \quad \text{(resp. } \textit{inequality reversed}\text{)}.$$

Then the problem

$$Au + \frac{1}{(u+\theta_0)(1+u'^2)^{1/2}} = g + \frac{1}{\theta_0}, \qquad u(0) = u(L) = 0$$

has a minimal solution and a maximal solution in $H^2(\Omega) \cap [0, v]$ (resp. $H^2(\Omega) \cap [v, 0]$).

We conclude this section with a uniqueness result.

THEOREM 4.5. *Let the hypotheses* (4.5), (4.6) *hold. Suppose that the derivative of F exists and is bounded for $u \in [w, v]$. Assume furthermore*

$$L \leqq (2/M)^{1/2}(1-k^2)^{3/4}, \quad \textit{where } M = \sup_{w \leqq t \leqq v} |F'(t+\theta_0)|. \tag{4.16}$$

Then there exists a unique solution of $P(\theta_0, L, g)$ in $[w, v] \cap C^1(\bar{\Omega})$.

Proof. Theorem 4.2 insures the existence of at least a solution. Let u_1 and u_2 be two solutions of $P(\theta_0, L, g)$ in $[w, v] \cap C^1(\bar{\Omega})$. Thus,

$$Au_1 - Au_2 + F(u_1+\theta_0) - F(u_2+\theta_0) = 0. \tag{4.17}$$

By similar arguments to those of Theorem 4.2, we see that

$$|u_i'(x)| \leqq \frac{k}{(1-k^2)^{1/2}}, \qquad i = 1, 2. \tag{4.18}$$

We multiply (4.17) by $z = u_1 - u_2$, and integrate over Ω

$$\int_\Omega \left(\frac{u_1'}{(1+u_1'^2)^{1/2}} - \frac{u_2'}{(1+u_2'^2)^{1/2}} \right) z' \, dx = \int_\Omega (F(u_2+\theta_0) - F(u_1+\theta_0)) z \, dx. \tag{4.19}$$

We use the mean theorem and (4.18) to bound from below the left-hand side of (4.19). Then, by the Poincaré inequality, we get

$$(1-k^2)^{3/2} \int_\Omega z'^2 \, dx \leqq M \int_\Omega z^2 \, dx \leqq \frac{ML^2}{2} \int_\Omega z'^2 \, dx.$$

Hence, we conclude that under the condition (4.16), $z \equiv 0$ in Ω.

5. Some comments on the existence of upper and lower solutions. The previous results show the importance of finding upper and lower solutions for $P(\theta_0, L, g)$. The purpose of this section is to show that, under conditions on the data L, θ_0, g, there exists such functions.

A first result is contained in

THEOREM 5.1. *Suppose $g \in L^q(\Omega)$, $q > 1$, satisfies $|g + F(\theta_0)|_{L^1(\Omega)} < 1$. Then there exists an upper solution $v \geqq 0$, in $W_0^{1,\infty}(\Omega)$, for $P(\theta_0, L, g)$.*

Proof. Let $G = (g + F(\theta_0))^+$ and let $v_\varepsilon \in W_0^{1,p}(\Omega) \cap W^{2,q}(\Omega)$ be the solution of

$$\varepsilon \Delta_p v_\varepsilon + A v_\varepsilon = G \quad \text{a.e. in } \Omega, \qquad v_\varepsilon(0) = v_\varepsilon(L) = 0$$

(where $p \geqq q'$ and $\infty > p > 1$). Since $|G|_{L^1(\Omega)} < 1$, it follows that $|v_\varepsilon'|_{L^\infty(\Omega)} \leqq C$. Then

there exists a subsequence, still denoted v_ε, which converges to $v \in W_0^{1,\infty}(\Omega)$ such that $Av = G$. Thus, it is easily seen that v is an upper solution for $P(\theta_0, L, g)$ and since $G \geqq 0$, $v \geqq 0$ by Lemma 2.2.

Now we study more precisely the case $F(u) = 1/u$ and we suppose that $g \in L^\infty(\Omega)$. Let g and θ_0 be given, we can show that, provided L is sufficiently small, there exist upper and lower solutions of $P(\theta_0, L, g)$ in a simple family naturally related to the problem: the family of functions z associated to arcs of circle and verifying $z(0) = z(L) = 0$.

Let $\phi \in [-\pi/2, \pi/2]$. We note z_ϕ, the one of the above functions such that $z'_\phi(0) = tg\phi$. Thus, $Az_\phi = 2 \sin \phi/L$, $z_\phi \in W_0^{1,p}(\Omega)$, where $p < 2$ if $\phi = \mp\pi/2$, otherwise $z_\phi \in C^\infty(\bar{\Omega})$.

Suppose $\sup_\Omega g > 0$. By using Definition 2.1 and an upper bound of z_ϕ, we find that z_ϕ, $\phi \in]0, \pi/2]$ is an upper solution if L satisfies

$$\frac{1-\cos\phi}{\sin\phi}\left(\sup_\Omega g + \frac{1}{\theta_0}\right)L^2 + 2(\theta_0 \sup_\Omega g - 1 + \cos\phi)L - 4\theta_0 \sin\phi \leqq 0. \tag{5.1}$$

Suppose $\inf_\Omega g < 0$. In the same way, z_ϕ, $\phi \in [-\pi/2, 0[$ verifies (2.2) if L satisfies

$$\frac{1-\cos\phi}{\sin\phi}\left(\inf_\Omega g + \frac{1}{\theta_0}\right)L^2 + 2(\theta_0 \inf_\Omega g - 1 + \cos\phi)L - 4\theta_0 \sin\phi \geqq 0. \tag{5.2}$$

For z_ϕ, $\phi \in [-\pi/2, 0[$, to be a lower solution we add the condition

$$\min_\Omega z_\phi > -\theta_0. \tag{5.3}$$

Relations (5.1), (5.2), (5.3) yield the following results z_ϕ; $\phi \in]0, \pi/2]$ is an upper solution of $P(\theta_0, L, g)$ if

$$L \leqq h_1(\theta_0, \phi, \sup_\Omega g), \tag{5.4}$$

where

$$h_1(\alpha, \phi, \mu) = \frac{4\alpha \sin\phi}{\alpha\mu + \cos\phi - 1 + ((\alpha\mu + 1 - \cos\varphi)^2 + 4(1-\cos\varphi))^{1/2}}$$

z_ϕ, $\phi \in [-\pi/2, 0[$ is a lower solution of $P(\theta_0, L, g)$ if

$$L \leqq h_2(\theta_0, \phi, \inf_\Omega g), \tag{5.5}$$

where

$$h_2(\alpha, \phi, \mu) = \frac{4\alpha \sin\phi}{\alpha\mu + \cos\phi - 1 - ((\alpha\mu + 1 - \cos\varphi)^2 + 4(1-\cos\varphi))^{1/2}}.$$

Under the condition (5.4) or (5.5), Theorem 4.1 applies. When g has not a constant sign, we have the following analogous result.

THEOREM 5.2. *Let $g \in L^\infty(\Omega)$. Then there exists a mapping $k : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ such that, for $L < k(\theta_0, |g|_{L^\infty(\Omega)})$, the hypotheses of Theorem 4.2 are satisfied.*

Proof. Consider in the family z_ϕ, $\phi \in [0, \pi/2]$ an upper solution z_α and in the family z_ϕ, $\phi \in [-\pi/2, 0]$ a lower solution z_β of $P(\theta_0, L, g)$ which verify (4.6). This yields the condition

$$\frac{z'_\alpha(0)}{(1+z'_\alpha(0)^2)^{1/2}} - \frac{z'_\alpha(L)}{(1+z'_\alpha(L)^2)^{1/2}} = 2 \sin\alpha < 1.$$

Then $\alpha < \pi/6$ and similarly we find $\beta > -\pi/6$.

Now by (5.4) and (5.5), we conclude that the condition

$$L < \max\left[\max_{\phi\in[0,\pi/6]} h_1(\theta_0, \phi, \sup_\Omega g), \max_{\phi\in[-\pi/6,0]} h_2(\theta_0, \phi, \inf_\Omega g)\right] \tag{5.6}$$

insures the existence of upper and lower solutions of $P(\theta_0, L, g)$ which satisfy the hypotheses of Theorem 4.2, and thus the existence of a solution. Since h_1 is a decreasing function of $\sup_\Omega g$ (for $\phi \geqq 0$) and h_2 an increasing function of $\inf_\Omega g$ (for $\phi \leqq 0$), we can replace (5.6) by the sufficient condition

$$L < \max\left[\max_{\phi\in[0,\pi/6]} h_1(\theta_0, \phi, |g|_{L^\infty(\Omega)}), \max_{\phi\in[-\pi/6,0]} h_2(\theta_0, \phi, -|g|_{L^\infty(\Omega)})\right]$$

Theorem 5.2 follows and we note that k is a decreasing function of $|g|_{L^\infty(\Omega)}$.

6. Periodic solutions for $Au + 1/u = f$. In this section we study the existence of periodic solutions for

$$Au + \frac{1}{u} = f, \tag{6.1}$$

i.e., solutions which satisfy the periodic boundary conditions

$$u(0) = u(L), \qquad u'(0) = u'(L). \tag{6.2}$$

We first give a nonexistence result.

THEOREM 6.1. *Let $f \in L^1(\Omega)$ satisfy $|f|_{L^1(\Omega)} \leqq 1$ and $\int_\Omega f(x)\,dx = 0$. Then the problem* (6.1), (6.2) *has no solution in $W^{2,1}(\Omega)$.*

Proof. Let us assume the existence of such a solution for problem (6.1), (6.2); then $Au \in L^1(\Omega)$ and $1/u \in L^1(\Omega)$.

We multiply (6.1) by u and integrate by parts over Ω to find

$$\int_\Omega \frac{u'^2}{(1+u'^2)^{1/2}}\,dx + L = \int_\Omega fu\,dx = \int_\Omega f(x) \int_0^x u'(t)\,dt\,dx. \tag{6.3}$$

It follows easily from (6.3) that

$$|u'|_{L^1(\Omega)} < |f|_{L^1(\Omega)} |u'|_{L^1(\Omega)}.$$

But this contradicts our assumption and Theorem 6.1 is therefore established. We now give conditions under which the unique solvability of problems $P(\theta_0, L, g)$ implies the existence of a periodic solution for (6.1). Our assumptions are as follows.

(i) $f \in L^\infty(\Omega)$, $f > 0$ a.e. in Ω. We set

$$\lambda = \frac{1}{\sup_\Omega f}, \qquad \mu = \frac{1}{\inf_\Omega f}.$$

(ii) There exists an upper solution v of $P(\mu, L, f - 1/\mu)$ satisfying (4.5).

(iii) There exists a lower solution w of $P(\lambda, L, f - 1/\lambda)$ satisfying (4.5).

(iv) v and w verify the hypotheses (4.6) and (4.16) (with λ replacing θ_0).

Remark. If problem (6.1), (6.2) associated to f has a solution u, then problem (6.1), (6.2) associated to $-f$ admits $-u$ as solution.

THEOREM 6.2. *Under the assumptions* (i)–(iv) *stated above, problem* (6.1), (6.2) *has a solution $y \in C^1(\bar\Omega)$ such that $\lambda \leqq y(0) \leqq \mu$.*

Proof. For all $\theta_0 \in [\lambda, \mu]$, we observe that v and w are respectively upper and lower solutions for $P(\theta_0, L, f - 1/\theta_0)$, and it follows from the § 4 that $P(\theta_0, L, f - 1/\theta_0)$ has a unique solution u_{θ_0} in $C^1(\bar\Omega) \cap [w, v]$ such that $|u'_{\theta_0}(x)| \leqq k/(1-k^2)^{1/2}$.

Therefore, u''_{θ_0} is uniformly bounded. Then, by an application of the Ascoli–Arzela theorem, $u'_{\theta_0}(x)$ is a continuous function in θ_0, on $[\lambda, \mu]$, for all $x \in [0, L]$.

Thus, the function U defined on $[\lambda, \mu]$ by $U(\theta_0) = u'_{\theta_0}(0) - u'_{\theta_0}(L)$ is a continuous function on $[\lambda, \mu]$.

On the other hand, $P(\mu, L, f - 1/\mu)$ has a unique solution in $[w, v]$. It follows from Theorem 4.1 and (i) that this solution u_μ is in fact in $[0, v]$. We also have $u''_\mu \leqq 0$ a.e. in Ω, which implies that $U(\mu) \geqq 0$. Similarly, the solution u_λ of $P(\lambda, L, f - 1/\lambda)$ is in $[w, 0]$. Since $u''_\lambda \geqq 0$ a.e. in Ω, we have $U(\lambda) \leqq 0$.

The intermediate value theorem for continuous functions now implies that there exists α, $\lambda \leqq \alpha \leqq \mu$, such that $U(\alpha) = 0$.

Let u_α be the solution of $P(\alpha, L, f - 1/\alpha)$ in $[w, v]$ and let $y = u_\alpha + \alpha$, then y is a solution of (6.1), (6.2).

Remark. Using § 5 and making some computations, we find that for $L \leqq 0.7\lambda$, the hypotheses of Theorem 6.2 are satisfied.

REFERENCES

[1] H. AMANN, *Supersolutions, monotone iterations and stability*, J. Differential Equations, 21 (1976), pp. 363–377.

[2] A. J. CALLEGARI AND E. L. REISS, *Non-linear boundary value problems for the circular membrane*, Arch. Rational Mech. Anal., 31 (1968), pp. 390–400.

[3] A. J. CALLEGARI, E. L. REISS AND H. B. KELLER, *Membrane buckling: A study of solution multiplicity*, Comm. Pure Appl. Math., 24 (1971), pp. 499–527.

[4] J. CHANDRA AND P. W. DAVIS, *A monotone method for quasilinear boundary value problems*, Arch. Rational Mech. Anal., 54 (1974), pp. 257–266.

[5] D. S. COHEN AND H. B. KELLER, *Some positone problems suggested by nonlinear heat generation*, J. Math. Mech., 16 (1967), pp. 1361–1376.

[6] H. B. KELLER, *Elliptic boundary value problems suggested by nonlinear diffusion processes*, Arch. Rational Mech. Anal., 35 (1969), pp. 363–381.

[7] J. P. KERNEVEZ, Thèse, Université Paris VI, 1972.

[8] A. M. LEFEVERE, Thèse 3$^{\text{emè}}$ cycle, Orsay, Université Paris XI, 1976.

[9] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[10] K. MIKA AND W. UELHOFF, *Shape and stability of menisci in Czochralski growth and comparison with analytical approximations*, Crystal Growth, 30 (1975), pp. 9–20.

[11] J. P. PUEL, *Existence, comportement à l'infini et stabilité dans certains problèmes quasi-linéaires elliptiques et paraboliques d'ordre 2*, Paris, Publ. L.A. C.N.R.S. 189, 1974.

[12] J. M. QUENISSET, Thèse, Bordeaux, Labo. C.N.R.S. Chimie du Solide, 1980.

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

[14] K. SCHMITT, *Periodic solutions of systems of second order differential equations*, J. Differential Equations, 11 (1972), pp. 180–192.

[15] G. STAMPACCHIA, *Equations elliptiques du second ordre à coefficients discontinus*, Séminaire de mathématiques supérieures, Université de Montréal, 1965.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0005 $01.00/0

DIFFERENTIAL INEQUALITIES OF HIGHER ORDER AND THE ASYMPTOTIC SOLUTION OF NONLINEAR BOUNDARY VALUE PROBLEMS*

F. A. HOWES†

Abstract. Two differential inequality results of Nagumo on initial and boundary value problems for systems [Proc. Phys. Math. Soc. Japan, 19 (1937), pp. 861–866; 21 (1939), pp. 529–534] are combined to yield existence and comparison results on certain boundary value problems for nth order scalar nonlinear differential equations and their system analogues. This theory is then applied to several classes of singularly perturbed boundary value problems of higher order. Many examples are discussed in order to motivate the theory and indicate avenues of further study.

1. Introduction. We consider here some differential inequality theorems for boundary value problems involving systems which, in particular, simplify considerably several recent results for scalar boundary value problems of higher order. At the same time, we use these results to extend the theory of certain singularly perturbed nonlinear second-order boundary value problems to higher-order differential equations. Our treatment of such phenomena is motivated, on the one hand, by the simplicity and directness of our approach. On the other hand, there does not appear to be any general theory whatsoever for higher-order perturbed nonlinear boundary value problems. This paper constitutes, then, a first attempt to develop such a theory along the lines of the author's previous work [11], [15].

Before discussing the most general results, we study a third-order problem in order to illustrate our ideas in the simplest setting.

2. The scalar problem. The third-order scalar boundary value problem

$$\begin{aligned} &y''' = f(t, y, y', y''), \qquad a < t < b, \\ &y(a) = A_0, \quad y'(a) = A_1, \quad y'(b) = B_1 \end{aligned} \tag{2.1}$$

and various generalizations have been studied by several authors using the method of comparison functions (cf., for example, [23], [21], [19] and the references contained therein). The basic idea is to employ the solutions of certain differential inequalities, in conjunction with a growth estimate on f as a function of y'', to obtain a priori bounds on solutions of (2.1), and to then apply a fixed point or continuation argument. An important consequence of this approach is that in the course of proving the existence of a solution of the boundary value problem, one obtains simultaneously an estimate of this solution in terms of the solutions of the differential inequalities.

Suppose now that we rewrite the boundary value problem (2.1) as a system consisting of a first-order initial value problem and a second-order boundary value problem, namely

$$\begin{aligned} &y' = z, \qquad y(a) = A_0, \\ &z'' = f(t, y, z, z'), \quad z(a) = A_1, \quad z(b) = B_1. \end{aligned} \tag{2.2}$$

* Received by the editors July 17, 1980, and in revised form March 20, 1981.

† Department of Mathematics, University of California at Davis, Davis, California 95616. This work was supported in part by the National Science Foundation under grants MCS 78-00907 at the University of Minnesota and MCS 80-01615 at the University of California at Davis.

This suggests that it might be profitable to examine existence and comparison results for the more general system on (a, b)

$$\begin{aligned} y' &= g(t, y, z), \qquad y(a) = A_0, \\ z'' &= f(t, y, z, z'), \quad z(a) = A_1, \quad z(b) = B_1. \end{aligned} \tag{2.3}$$

Indeed, earlier results of Nagumo [26], [27] imply that if there exist two pairs of comparison functions (u, v) and (α, β) satisfying the appropriate inequalities, then the boundary value problem (2.3) has a solution $(y, z) = (y(t), z(t))$ with $u(t) \leqq y(t) \leqq v(t)$ and $\alpha(t) \leqq z(t) \leqq \beta(t)$ on $[a, b]$ provided that $f(t, y, z, w) = O(|w|^2)$ as $|w| \to \infty$ uniformly for y in $[u, v]$ $(=\{x: u(t) \leqq x \leqq v(t)\})$. The functions u and v are required to satisfy the inequalities

$$u \leqq v, \qquad u(a) \leqq A_0 \leqq v(a), \tag{2.4}$$

and for t in $(a, b]$

$$u'(t) \leqq g(t, u(t), z), \qquad v'(t) \geqq g(t, v(t), z) \tag{2.5}$$

for any z in $[\alpha, \beta]$. Similarly, the functions α and β are required to satisfy the corresponding second-order inequalities

$$\alpha \leqq \beta, \quad \alpha(a) \leqq A_1 \leqq \beta(a), \quad \alpha(b) \leqq B_1 \leqq \beta(b), \tag{2.6}$$

and for t in (a, b)

$$\alpha''(t) \geqq f(t, y, \alpha(t), \alpha'(t)), \qquad \beta''(t) \leqq f(t, y, \beta(t), \beta'(t)) \tag{2.7}$$

for any y in $[u, v]$. The precise result is contained in the following theorem.

THEOREM 2.1. *Suppose that*

(1) *the function $g = g(t, y, z)$ is continuous in the domain $\mathcal{D} = [a, b] \times [u, v] \times [\alpha, \beta]$ and the function $f = f(t, y, z, w)$ is continuous in the domain $\mathcal{D} \times \mathbb{R}^1$;*

(2) *the comparison functions $u, v (\alpha, \beta)$ are of class $C^{(1)}[a, b] (C^{(2)}[a, b])$ and satisfy the inequalities* (2.4), (2.5) ((2.6), (2.7));

(3) (*Nagumo condition*) *for (t, y, z) in $\mathcal{D}$,*

$$f(t, y, z, w) = O(|w|^2) \quad \text{as } |w| \to \infty.$$

Then the boundary value problem (2.3) *has a solution $(y, z) = (y(t), z(t))$ of class $C^{(1)}[a, b] \times C^{(2)}[a, b]$ such that*

$$u(t) \leqq y(t) \leqq v(t) \quad \text{and} \quad \alpha(t) \leqq z(t) \leqq \beta(t)$$

for $a \leqq t \leqq b$.

Proof. The existence of the functions (u, v) and (α, β) together with the assumption of the Nagumo condition allow us to find bounds for y, z and z' and apply Schauder's fixed point theorem as in [9, Chap. 12, Pt. 2], [7] and [22]. The details are straightforward and are omitted. □

Before turning to the general nth order scalar problem, we make two remarks.

First of all, the differentiability assumptions on the functions (u, v) and (α, β) can be weakened in the following sense (cf. [3] and [20]). The functions u and v need only be piecewise continuously differentiable on $(a, b]$ provided that $u'(v')$ is replaced by $u'_-(v'_+)$ at a point of nondifferentiability. (Here $w'_+(w'_-)$ denotes the right-hand (left-hand) derivative.) Similarly, the functions α and β need only be piecewise twice continuously differentiable provided that $\alpha'_- \leqq \alpha'_+$ and $\beta'_+ \leqq \beta'_-$ at a point of nondifferentiability, and that $\alpha''(\beta'')$ is replaced by $\alpha''_-(\beta''_+)$ at such a point.

Secondly, we note that Theorem 2.1 applies to the special case (2.1) if we assume, in addition to the Nagumo condition, the existence of functions u and v of class $C^{(3)}[a, b]$ such that

$$u \leqq v, \quad u' \leqq v', \quad u(a) \leqq A_0 \leqq v(a),$$
$$u'(a) \leqq A_1 \leqq v'(a), \qquad u'(b) \leqq B_1 \leqq v'(b),$$

and for t in (a, b)

$$u'''(t) \geqq f(t, y, u'(t), u''(t)), \qquad v'''(t) \leqq f(t, y, v'(t), v''(t))$$

for any y in $[u, v]$. Indeed, we can apply this theorem to the system (2.2) (with $\alpha = u', \beta = v'$) and deduce the existence of a solution $(y(t), y'(t))$ of (2.2) which satisfies $u(t) \leqq y(t) \leqq v(t)$ and $u'(t) \leqq y'(t) \leqq v'(t)$ for $a \leqq t \leqq b$.

Consider now the general nth order scalar boundary value problem

$$\begin{aligned} &y^{(n)} = f(t, y, y', \cdots, y^{(n-1)}), \qquad a < t < b, \\ &y^{(j)}(a) = A_j, \quad 0 \leqq j \leqq n-2, \qquad y^{(n-2)}(b) = B_{n-2}. \end{aligned} \tag{2.8}$$

One advantage of our approach is that it is just as easy to treat (2.8) as it was to treat (2.1) since Nagumo's lemma [27] applies equally well to systems.

We begin by rewriting (2.8) as an $(n-2)$nd-order initial value problem together with a second-order boundary value problem, namely

$$\begin{aligned} &y_i' = y_{i+1}, \qquad y_i(a) = A_{i-1}, \qquad i = 1, \cdots, n-3, \\ &y_{n-2}' = z, \qquad y_{n-2}(a) = A_{n-3}, \\ &z'' = f(t, y_1, \cdots y_{n-2}, z, z'), \quad z(a) = A_{n-2}, \quad z(b) = B_{n-2}. \end{aligned} \tag{2.9}$$

This leads us to consider an existence and comparison result for the more general system on (a, b)

$$\begin{aligned} &\mathbf{y}' = \mathbf{g}(t, \mathbf{y}, z), \qquad \mathbf{y}(a) = \mathbf{A}, \\ &z'' = f(t, \mathbf{y}, z, z'), \quad z(a) = \xi, \quad z(b) = \eta, \end{aligned} \tag{2.10}$$

where $\mathbf{y}, \mathbf{g}, \mathbf{A}$ are in $\mathbb{R}^m$. The same two results of Nagumo [26], [27] suggest that if there exist functions $\mathbf{u} = \mathbf{u}(t), \mathbf{v} = \mathbf{v}(t)$ in $\mathbb{R}^m$ and scalar functions α, β which satisfy inequalities analogous to (2.4), (2.5) and (2.6), (2.7), respectively, then the problem (2.10) has a solution $(\mathbf{y}, z) = (\mathbf{y}(t), z(t))$ with $\mathbf{u}(t) \leqq \mathbf{y}(t) \leqq \mathbf{v}(t)$ and $\alpha(t) \leqq z(t) \leqq \beta(t)$ for $a \leqq t \leqq b$ provided that f satisfies a Nagumo condition uniformly in $\mathbf{y}$ in $[\mathbf{u}, \mathbf{v}]$. (Here the inequality $\leqq$ for vectors is to be interpreted as stating that the corresponding scalar inequality holds for respective components of the vectors and $[\mathbf{u}, \mathbf{v}] = \{\mathbf{w} \text{ in } \mathbb{R}^m : \mathbf{u}(t) \leqq \mathbf{w} \leqq \mathbf{v}(t)\}$.) Indeed, as regards $\mathbf{u}$ and $\mathbf{v}$ we need only require that

$$\mathbf{u} \leqq \mathbf{v}, \qquad \mathbf{u}(a) \leqq \mathbf{A} \leqq \mathbf{v}(a), \tag{2.4'}$$

and for t in $(a, b]$ and $i = 1, \cdots, m$

$$u_i'(t) \leqq g_i(t, \hat{\mathbf{u}}_i(t), z), \qquad v_i'(t) \geqq g_i(t, \hat{\mathbf{v}}_i(t), z) \tag{2.5'}$$

for any z in $[\alpha, \beta]$, where

$$\hat{\mathbf{u}}_i = (y_1, \cdots, y_{i-1}, u_i, y_{i+1}, \cdots, y_m)$$

and

$$\hat{\mathbf{v}}_i = (y_1, \cdots, y_{i-1}, v_i, v_{i+1}, \cdots, y_m)$$

for y_j in $[u_j, v_j](j \neq i)$; cf. [27]. Then we have the following result for (2.10) which is proved by mimicking arguments in [7], [22] and [9, Chap. 12, Pt. 2].

THEOREM 2.2 *Suppose that*

(1) *the function* $\mathbf{g} = \mathbf{g}(t, \mathbf{y}, z)$ *is continuous in the domain* $\mathcal{D}_1 = [a, b] \times [\mathbf{u}, \mathbf{v}] \times [\alpha, \beta]$ *and the function* $f = f(t, \mathbf{y}, z, w)$ *is continuous in the domain* $\mathcal{D}_1 \times \mathbb{R}^1$;

(2) *the comparison functions* $(\mathbf{u}, \mathbf{v})((\alpha, \beta))$ *are of class* $C^{(1)}[a, b](C^{(2)}[a, b])$ *and satisfy the inequalities* (2.4'), (2.5') ((2.6), (2.7) *with* y *replaced by* $\mathbf{y}$ *and* $[u, v]$ *replaced by* $[\mathbf{u}, \mathbf{v}]$);

(3) (*Nagumo condition*) *for* $(t, \mathbf{y}, z)$ *in* $\mathcal{D}_1$,

$$f(t, \mathbf{y}, z, w) = O(|w|^2) \quad \text{as } |w| \to \infty.$$

Then the boundary value problem (2.10) *has a solution* $(\mathbf{y}, z) = (\mathbf{y}(t), z(t))$ *of class* $C^{(1)}[a, b] \times C^{(2)}[a, b]$ *such that*

$$\mathbf{u}(t) \leqq \mathbf{y}(t) \leqq \mathbf{v}(t) \quad \text{and} \quad \alpha(t) \leqq z(t) \leqq \beta(t)$$

for $a \leqq t \leqq b$.

We note that less differentiability can be required of the comparison functions $\mathbf{u}, \mathbf{v}, \alpha$ and β (cf. our earlier remark). Also, if the function $\mathbf{g}$ is quasi-monotone nondecreasing with respect to $\mathbf{y}$ for each fixed t in $[a, b]$ and each fixed z in $[\alpha, \beta]$, then the differential inequalities in (2.5') can be replaced by the simpler ones

$$\mathbf{u}'(t) \leqq \mathbf{g}(t, \mathbf{u}(t), z), \qquad \mathbf{v}'(t) \geqq \mathbf{g}(t, \mathbf{v}(t), z);$$

cf. for example [3, Chap. 1]. As a result of this observation, we can apply Theorem 2.2 to study the scalar problem (2.8) since in the equivalent system (2.9) the corresponding function $\mathbf{g}$ is simply $(y_2, y_3, \cdots, z)$ which is certainly quasi-monotone nondecreasing for each fixed z. In fact, we need only assume that the function $f = f(t, \mathbf{y}, z, z')$ satisfies a Nagumo condition and that there exist functions u and v of class $C^{(n)}[a, b]$ such that

$$u^{(k)} \leqq v^{(k)}, \qquad u^{(k)}(a) \leqq A_k \leqq v^{(k)}(a) \qquad k = 0, \cdots, n-2,$$
$$u^{(n-2)}(b) \leqq B_{n-2} \leqq v^{(n-2)}(b),$$

and for t in (a, b)

$$u^{(n)}(t) \geqq f(t, \mathbf{y}, u^{(n-2)}(t), u^{(n-1)}(t)),$$
$$v^{(n)}(t) \leqq f(t, \mathbf{y}, v^{(n-2)}(t), v^{(n-1)}(t)),$$

for any $\mathbf{y} = (y_1, \cdots, y_{n-2})$ with y_l in $[u^{(l-1)}, v^{(l-1)}](1 \leqq l \leqq n-2)$. Then we can apply Theorem 2.2 to the problem (2.9) (with $\mathbf{u} = (u, u', \cdots, u^{(n-3)})$, $\mathbf{v} = (v, v', \cdots, v^{(n-3)})$, $\alpha = u^{(n-2)}$ and $\beta = v^{(n-2)}$) and deduce the existence of a solution $y = y(t)$ of (2.8) which is of class $C^{(n)}[a, b]$ and which satisfies for $k = 0, \cdots, n-2$,

$$u^{(k)}(t) \leqq y^{(k)}(t) \leqq v^{(k)}(t) \quad \text{on } [a, b].$$

Before discussing a more general problem in the next section we remark that this result for (2.8) is due to Kelley [21] who gave a rather complicated proof. In addition, we note that it is sometimes necessary to weaken the restriction imposed by the Nagumo condition (cf. Theorem 4.1 below). This condition assures us that a bound on z' implies a bound on z'', and so Theorem 2.2 (and Theorem 2.1 as well) are valid if we simply make this assumption which is called a *generalized* Nagumo condition (cf. [10] or [21]).

3. The general problem. We consider now the following boundary value problem on (a, b):

$$\begin{aligned} \mathbf{y}' &= \mathbf{g}(t, \mathbf{y}, \mathbf{z}), \qquad \mathbf{y}(a) = \mathbf{A}, \\ \mathbf{z}'' &= \mathbf{h}(t, \mathbf{y}, \mathbf{z}, \mathbf{z}'), \quad \mathbf{z}(a) = \boldsymbol{\xi}, \quad \mathbf{z}(b) = \boldsymbol{\eta}, \end{aligned} \tag{3.1}$$

where $\mathbf{y}, \mathbf{g}, \mathbf{A}$ belong to $\mathbb{R}^m$ and $\mathbf{z}, \mathbf{h}, \boldsymbol{\xi}, \mathbf{h}$ to $\mathbb{R}^n$. The results of § 2 suggest that in order to study this problem we can simply combine earlier theory on the properties of invariant regions for initial value problems and boundary value problems for systems of differential equations.

Let us consider first the initial value problem in $\mathbb{R}^m$ for $a < t \leqq b$

$$\mathbf{y}' = \mathbf{G}(t, \mathbf{y}), \qquad \mathbf{y}(a) = \mathbf{A}. \tag{3.2}$$

Suppose that there exist M real-valued functions $r_i = r_i(t, \mathbf{y})$ of class $C^{(1)}([a, b] \times \mathbb{R}^m)$ such that for $i = 1, \cdots, M$

$$r_i' := \frac{\partial r_i}{\partial t} + [\operatorname{grad}_{\mathbf{y}} r_i] \cdot \mathbf{G}(t, \mathbf{y}) \leqq 0 \tag{3.3}$$

when $r_i = 0$. (Here $\operatorname{grad}_{\mathbf{w}}$ denotes the gradient taken with respect to the components of $\mathbf{w}$, while $\cdot$ denotes the usual Euclidean inner product.) Then it is known (cf. [1] or [8]) that if the function $\mathbf{G}$ is continuous in the region

$$\Omega = \{(t, \mathbf{y}) \text{ in } [a, b] \times \mathbb{R}^m : r_i(t, \mathbf{y}) \leqq 0, \, i = 1, \cdots, M\}$$

and if the initial pair $(a, \mathbf{A})$ belongs to Ω, the problem (3.2) has a solution $\mathbf{y} = \mathbf{y}(t)$ of class $C^{(1)}[a, b]$ such that

$$r_i(t, \mathbf{y}(t)) \leqq 0 \quad \text{on } [a, b] \quad \text{for } i = 1, \cdots, M.$$

In other words, Ω is an invariant region for (3.2) since trajectories which originate there remain there on $[a, b]$.

Consider next the boundary value problem in $\mathbb{R}^n$ for $a < t < b$:

$$\mathbf{z}'' = \mathbf{H}(t, \mathbf{z}, \mathbf{z}'), \quad \mathbf{z}(a) = \boldsymbol{\xi}, \quad \mathbf{z}(b) = \boldsymbol{\eta}. \tag{3.4}$$

It turns out that the existence of functions similar to the r_i also allows us to deduce an existence and estimation result for solutions of (3.4) provided the right-hand side satisfies a growth condition with respect to $\mathbf{z}'$. In this vector case we say that $\mathbf{H} = (H_1, \cdots, H_n)$ satisfies a Nagumo condition if one of the following two situations obtains (cf. [22]):

(1) for $j = 1, \cdots, n$ there exist positive, nondecreasing continuous functions φ_j on $(0, \infty)$ such that

$$\int^{\infty} s/\varphi_j(s) \, ds = \infty$$

and

$$|H_j(t, \mathbf{z}, \mathbf{w})| \leqq \varphi_j(|w_j|)$$

for $(t, \mathbf{z})$ in compact subsets of $[a, b] \times \mathbb{R}^n$ and all $\mathbf{w}$ in $\mathbb{R}^n$;

(2) there exists a positive, nondecreasing continuous function φ on $(0, \infty)$ such that

$$s^2/\varphi(s) \to \infty$$

and

$$\|\mathbf{H}(t, \mathbf{z}, \mathbf{w})\| \leqq \varphi(\|\mathbf{w}\|)$$

for $(t, \mathbf{z})$ in compact subsets of $[a, b] \times \mathbb{R}^n$ and all $\mathbf{w}$ in $\mathbb{R}^n$. (Here $\|\cdot\|$ is the usual Euclidean norm.)

Suppose now that there exist N real-valued functions $\rho_l = \rho_l(t, \mathbf{z})$ of class $C^{(2)}$ $([a, b] \times \mathbb{R}^n)$ such that for $l = 1, \cdots, N$

$$\rho_l'' := \frac{\partial^2 \rho_l}{\partial t^2} + \left[2 \operatorname{grad}_{\mathbf{z}} \frac{\partial \rho_l}{\partial t}\right] \cdot \mathbf{z}' + [\mathscr{H}\mathbf{z}'] \cdot \mathbf{z}' + [\operatorname{grad}_{\mathbf{z}} \rho_l] \cdot \mathbf{H}(t, \mathbf{z}, \mathbf{z}') \geqq 0 \tag{3.5}$$

when $\rho_l = 0$ and $\rho_l' = 0$. (Here $\mathscr{H}$ is the Hessian of ρ_l with respect to $\mathbf{z}$ and $\rho_l' = \partial \rho_l / \partial t + [\operatorname{grad}_{\mathbf{z}} \rho_l] \cdot \mathbf{z}'$.) And let us define the region

$$\mathscr{E} = \{(t, \mathbf{z}) \text{ in } [a, b] \times \mathbb{R}^n : \rho_l(t, \mathbf{z}) \leqq 0, \, l = 1, \cdots, N\}.$$

Then it is known (cf. [22]) that if the function $\mathbf{H} = \mathbf{H}(t, \mathbf{z}, \mathbf{w})$ is continuous and satisfies a Nagumo condition in the domain $\mathscr{E} \times \mathscr{R}^n$, and if the initial and terminal pairs $(a, \boldsymbol{\xi})$, $(b, \boldsymbol{\eta})$ belong to $\mathscr{E}$ and can be joined by a smooth path in $\mathscr{E}$, the problem (3.4) has a solution $\mathbf{z} = \mathbf{z}(t)$ of class $C^{(2)}[a, b]$ such that

$$\rho_l(t, \mathbf{z}(t)) \leqq 0 \quad \text{on } [a, b] \quad \text{for } l = 1, \cdots, N.$$

Put succinctly, $\mathscr{E}$ is an invariant region for this boundary value problem.

We are now ready to study the problem (3.1). As a preliminary, we say that condition (3.3′) holds if there exist M functions $r_i = r_i(t, \mathbf{y})$ of class $C^{(1)}(\Omega)$ which satisfy condition (3.3), with $\mathbf{G}(t, \mathbf{y})$ replaced by $\mathbf{g}(t, \mathbf{y}, \mathbf{z})$, for all $(t, \mathbf{z})$ in $\mathscr{E}$; while we say that condition (3.5′) holds if there exist N functions $\rho_l = \rho_l(t, \mathbf{z})$ of class $C^{(2)}(\mathscr{E})$ which satisfy condition (3.5), with $\mathbf{H}(t, \mathbf{z}, \mathbf{z}')$ replaced by $\mathbf{h}(t, \mathbf{y}, \mathbf{z}, \mathbf{z}')$, for all $(t, \mathbf{y})$ in Ω. In other words, we require (3.3) to hold uniformly with respect to $\mathbf{z}$ in $\mathscr{E}$ and (3.5) to hold uniformly with respect to $\mathbf{y}$ in Ω. This uniformity allows us to prove the following result by arguing as in [7]. For ease of exposition, we define the domain

$$\mathscr{F} = \{(t, \mathbf{y}, \mathbf{z}) \text{ in } [a, b] \times \mathbb{R}^{m+n} : r_i(t, \mathbf{y}) \leqq 0, \, i = 1, \cdots, M; \, \rho_l(t, \mathbf{z}) \leqq 0, \, l = 1, \cdots, N\}.$$

THEOREM 3.1. *Suppose that*

(1) *the function* $\mathbf{g} = \mathbf{g}(t, \mathbf{y}, \mathbf{z})$ *is continuous in* $\mathscr{F}$ *and the function* $\mathbf{h} = \mathbf{h}(t, \mathbf{y}, \mathbf{z}, \mathbf{w})$ *is continuous in* $\mathscr{F} \times \mathbb{R}^n$;

(2) *conditions* (3.3′) *and* (3.5′) *hold*;

(3) *the function* $\mathbf{h}$ *satisfies a Nagumo condition uniformly with respect to* $\mathbf{y}$ *in* Ω.

Then the boundary value problem (3.1) *has a solution* $(\mathbf{y}, \mathbf{z}) = (\mathbf{y}(t), \mathbf{z}(t))$ *of class* $C^{(1)}[a, b] \times C^{(2)}[a, b]$ *such that on* $[a, b]$

$$r_i(t, \mathbf{y}(t)) \leqq 0 \quad \textit{for } i = 1, \cdots, M$$

and

$$\rho_l(t, \mathbf{z}(t)) \leqq 0 \quad \textit{for } l = 1, \cdots, N.$$

We note that the comparison functions r_i and ρ_l afford a variety of ways of describing solutions of (3.1). For example, if $M = N = 1$ and if $r(t, \mathbf{y}) = \|\mathbf{y}\| - \gamma_1(t)$ and $\rho(t, \mathbf{z}) = \|\mathbf{z}\| - \gamma_2(t)$ satisfy the appropriate inequalities, then we obtain bounds on the norm of the solution of the form $\|\mathbf{y}(t)\| \leqq \gamma_1(t)$ and $\|\mathbf{z}(t)\| \leqq \gamma_2(t)$. On the other hand, for $M = 2m$ and $N = 2n$, let us define

$$r_i(t, \mathbf{y}) = y_i - v_i(t), \qquad r_{i+m}(t, \mathbf{y}) = -y_i + u_i(t),$$

for $i = 1, \cdots, m$, and

$$\rho_l(t, \mathbf{z}) = z_l - \beta_l(t), \qquad \rho_{l+n}(t, \mathbf{z}) = -z_l + \alpha_l(t),$$

for $l = 1, \cdots, n$. If these functions satisfy the proper inequalities, then we obtain the *two-sided* bounds (cf. § 2)

$$u_i(t) \leqq y_i(t) \leqq v_i(t) \quad \text{and} \quad \alpha_l(t) \leqq z_l(t) \leqq \beta_l(t)$$

on the solution $(\mathbf{y}, \mathbf{z})$.

Some applications of the theory of this section and the previous one to singular perturbation problems are given in the next section. We close with several remarks.

Remark 3.1. An existence result for the problem (3.1) was proved by Hartman [7]; however, he did not give estimates for the solution of the generality afforded by our bounding functions r_i and ρ_l.

Remark 3.2. Under certain assumptions a result similar to Theorem 3.1 holds if the Dirichlet boundary conditions for $\mathbf{z}$ are replaced by Robin boundary conditions of the form

$$P_1\mathbf{z}(a) - Q_1\mathbf{z}'(a), \qquad P_2\mathbf{z}(b) + Q_2\mathbf{z}'(b)$$

prescribed (cf. [10], [21] and [24]). Here P_i, $Q_i (i = 1, 2)$ are nonnegative definite $(n \times n)$-matrices.

Remark 3.3. The arbitrariness of the dimensions m and n in the formulation of the problem (3.1) allows us, in particular, to treat a number of boundary value problems for scalar differential equations. As an illustration, consider the fifth-order equation $y^{(5)} = f(t, y, y', \cdots, y^{(4)})$ on (a, b). If the boundary conditions are that $y(a), y'(a), y''(a), y'''(a), y'''(b)$ are prescribed, then we would set $z = y'''$ and apply Theorem 3.1 with $m = 3$ and $n = 1$. On the other hand, if the boundary conditions are that $y(a), y'(a), y'(b), y'''(a), y'''(b)$ are prescribed, then we would set $z_1 = y', z_2 = y'''$ and apply Theorem 3.1 with $m = 1$ and $n = 2$.

4. Singular perturbation problems of higher order. We consider now some applications of the above theory to several classes of singularly perturbed boundary value problems of order three and higher. The general plan calls for us to first regard the given problem as the combination of a singularly perturbed second-order problem and an unperturbed initial value problem. Then we apply previous theory (cf., for example, [4], [11], [15], [28], [31]) to the perturbed part and evaluate the contribution of the unperturbed variables with the aid of our differential inequality results.

To fix these ideas, let us consider first a result for an nth-order scalar equation which is related to an earlier theorem of Levinson [25] for a third-order equation. The problem is

$$\begin{aligned} &\varepsilon y^{(n)} = f(t, y, y', \cdots, y^{(n-1)}), \qquad a < t < b, \quad n \geqq 3, \\ &y^{(j)}(a, \varepsilon) = A_j, \quad 0 \leqq j \leqq n-2, \quad y^{(n-2)}(b, \varepsilon) = B_{n-2}, \end{aligned} \tag{4.1}$$

where ε is a small positive parameter. Following [25] (cf. also [6], [14]), let us assume that the corresponding reduced ($\varepsilon = 0$) equation

$$0 = f(t, Y, Y', \cdots, Y^{(n-1)}), \qquad a < t < b, \tag{4.2}$$

has a solution $Y = Y_L(t)$ of class $C^{(n)}[a, t_L]$ such that

$$Y_L^{(j)}(a) = A_j \quad \text{for } j = 0, \cdots, n-2,$$

and a solution $Y = Y_R(t)$ of class $C^{(n)}[t_R, b]$ such that

$$Y_R^{(n-2)}(b) = B_{n-2},$$

with $a \leqq t_R < t_L \leqq b$. In addition, let us assume that at a point t_0 in (t_R, t_L)

$$Y_L^{(j)}(t_0) = Y_R^{(j)}(t_0)(=\sigma_j) \quad \text{for } j = 0, \cdots, n-2,$$

and that

$$\mu_L = Y_L^{(n-1)}(t_0) \neq Y_R^{(n-1)}(t_0) = \mu_R.$$

Then, if the functions Y_L, Y_R are stable in the sense that

$$f_{y^{(n-1)}}[Y_L(t)] \geqq k > 0 \quad \text{on } [a, t_0]$$

and

$$f_{y^{(n-1)}}[Y_R(t)] \leqq -k < 0 \quad \text{on } [t_0, b]$$

for $[X(t)] = (t, X(t), X'(t), \cdots, X^{(n-1)}(t))$ and some positive constant k, and if the crossing condition

$$(\mu_R - \mu_L) f(t_0, \sigma_0, \sigma_1, \cdots, \sigma_{n-2}, \lambda) > 0 \tag{4.3}$$

holds for all λ strictly between μ_L and μ_R, we anticipate that the full problem (4.1) has a solution $y = y(t, \varepsilon)$ of class $C^{(n)}[a, b]$ for each $\varepsilon > 0$ sufficiently small. Moreover, this solution should satisfy

$$\lim_{\varepsilon \to 0^+} y^{(j)}(t, \varepsilon) = Y^{(j)}(t) \quad \text{on } [a, b]$$

for $j = 0, \cdots, n-2$, and

$$\lim_{\varepsilon \to 0^+} y^{(n-1)}(t, \varepsilon) = \begin{cases} Y_L^{(n-1)}(t), & a \leqq t < t_0, \\ Y_R^{(n-1)}(t), & t_0 < t \leqq b, \end{cases}$$

where $Y(t) = Y_L(t)$ on $[a, t_0]$ and $Y(t) = Y_R(t)$ on $[t_0, b]$.

In order to verify this we rewrite (4.1) as the system

$$\begin{gathered} y_i' = y_{i+1}, \qquad y_i(a, \varepsilon) = A_{i-1}, \qquad i = 1, \cdots, n-3, \\ y_{n-2}' = z, \qquad y_{n-2}(a, \varepsilon) = A_{n-3}, \\ \varepsilon z'' = f(t, y_1, \cdots, y_{n-2}, z, z'), \quad z(a, \varepsilon) = A_{n-2}, \quad z(b, \varepsilon) = B_{n-2}, \end{gathered} \tag{4.4}$$

and apply Theorem 2.2 (with assumption (3) replaced by the generalized Nagumo condition) in conjunction with our previous scalar theory [14]. The idea is that the asymptotic behavior of the solution of (4.1) as $\varepsilon \to 0^+$ is really determined by the behavior of the solution of the last problem in (4.4) with $y_j = Y_j(t)$ for $j = 1, \cdots, n-2$, if the components $y_j = y_j(t, \varepsilon)(1 \leqq j \leqq n-2)$ of the solution $(\mathbf{y}, z)$ of (4.4) differ from the components $Y_j(t)$ of the solution

$$Y(t) = \begin{cases} Y_L(t), & a \leqq t \leqq t_0, \\ Y_R(t), & t_0 \leqq t \leqq b, \end{cases}$$

of the reduced problem

$$\begin{gathered} Y_i' = Y_{i+1}, \qquad Y_i(a) = A_{i-1}, \qquad i = 1, \cdots, n-3, \\ Y_{n-2}' = Z, \qquad Y_{n-2}(a) = A_{n-3}, \\ 0 = f(t, Y_1, \cdots, Y_{n-2}, Z, Z'), \quad Z(a) = A_{n-2}, \quad Z(b) = B_{n-2}, \end{gathered} \tag{4.5}$$

on $[a, b]$ by terms of order ε. In other words, to terms of order ε, we can analyze the solution of (4.4) by analyzing the solution of the *scalar* problem

$$\begin{aligned} &\varepsilon z'' = F(t, z, z'), \qquad a < t < b, \\ &z(a, \varepsilon) = A_{n-2}, \qquad z(b, \varepsilon) = B_{n-2}, \end{aligned} \tag{4.6}$$

where $F(t, z, z') = f(t, Y_1(t), \cdots, Y_{n-2}(t), z, z')$. When viewed in this light, the stability conditions on Y_L, Y_R, as well as the crossing condition, are nothing more than the classical assumptions on the problem (4.6) (cf. [6], [14]), that is,

$$F_{z'}[Y_L(t)] \geqq k > 0 \quad \text{on } [a, t_0], \qquad F_{z'}[Y_R(t)] \leqq -k < 0 \quad \text{on } [t_0, b],$$

and for all λ strictly between μ_L and μ_R

$$(\mu_L - \mu_R) F(t_0, \sigma_{n-2}, \lambda) > 0.$$

The precise statement is the following result. For ease of exposition, let us set

$$\mathcal{R} = [a, b] \times \Gamma \times \{w : |w - Y^{(n-1)}(t)| \leqq d(t)\},$$

where $\Gamma = \{(y_1, \cdots, y_{n-1}) : |y_j - Y^{(j-1)}(t)| \leqq \delta, 1 \leqq j \leqq n-1\}$ for a small positive constant δ, and d is a smooth positive function such that $|\mu_L - \mu_R| \leqq d(t) \leqq |\mu_L - \mu_R| + \delta$ on $[t_0 - \delta/2, t_0 + \delta/2]$ and $d(t) \leqq \delta$ on $[a, t_0 - \delta] \cup [t_0 + \delta, b]$.

THEOREM 4.1. *Suppose that the reduced equation* (4.2) *has smooth, stable solutions* $Y = Y_L(t)$ *and* $Y = Y_R(t)$ *with the above properties. Suppose also that the function* f *is continuous with respect to* $(t, y_1, \cdots, y_{n-2}, z, w)$ *and continuously differentiable with respect to* $(y_1, \cdots, y_{n-1}, w)$ *in the region* $\mathcal{R}$ *and that the crossing condition* (4.3) *holds.*

Then there exists an $\varepsilon_0 > 0$ *such that the problem* (4.1) *has a solution* $y = y(t, \varepsilon)$ *whenever* $0 < \varepsilon \leqq \varepsilon_0$. *Moreover, for* $a \leqq t \leqq b$ *we have that*

$$y^{(i)}(t, \varepsilon) = Y^{(i)}(t) + O(\varepsilon), \qquad i = 0, \cdots, n-3,$$

and

$$y^{(n-2)}(t, \varepsilon) = Y^{(n-2)}(t) + w_I(t, \varepsilon) + O(\varepsilon),$$

where $w_I(t, \varepsilon) = \varepsilon k_1^{-1} |\mu_L - \mu_R| \exp[-k_1 |t - t_0| \varepsilon^{-1}]$, $0 < k_1 < k$, *is the interior layer corrector at* t_0.

Proof. For simplicity let us consider just the case $n = 3$. Then in order to show the existence of a solution of (4.1) with the stated properties we must construct smooth functions u, v, α, β on $[a, b] \times [0, \varepsilon_0]$ such that $u \leqq v$, $\alpha \leqq \beta$, $u(a, \varepsilon) \leqq A_0 \leqq v(a, \varepsilon)$, $\alpha(a, \varepsilon) \leqq A_1 \leqq \beta(a, \varepsilon)$, $\alpha(b, \varepsilon) \leqq B_1 \leqq \beta(b, \varepsilon)$, and for t in $(a, b]$

$$u' \leqq z \leqq v' \quad \text{for all } z \text{ in } [\alpha, \beta], \tag{4.7}$$

while for t in (a, b)

$$\varepsilon \alpha'' \geqq f(t, y, \alpha, \alpha') \quad \text{and} \quad \varepsilon \beta'' \leqq f(t, y, \beta, \beta') \tag{4.8}$$

for all y in $[u, v]$. To this end, we define for $\varepsilon > 0$ the functions α and β by

$$\alpha(t, \varepsilon) = \begin{cases} Y_L'(t) - w_I(t, \varepsilon) - \varepsilon \gamma_1 l^{-1}(e^{\lambda(t-a)} - 1), & a \leqq t \leqq t_0, \\ Y_R'(t) - w_I(t, \varepsilon) - \varepsilon \gamma_2 l^{-1}(e^{\lambda(b-t)} - 1), & t_0 \leqq t \leqq b, \end{cases}$$

and

$$\beta(t, \varepsilon) = \begin{cases} Y_L'(t) + w_I(t, \varepsilon) + \varepsilon \gamma_1 l^{-1}(e^{\lambda(t-a)} - 1), & a \leqq t \leqq t_0, \\ Y_R'(t) + w_I(t, \varepsilon) + \varepsilon \gamma_2 l^{-1}(e^{\lambda(b-t)} - 1), & t_0 \leqq t \leqq b, \end{cases}$$

for appropriate positive constants γ_1, γ_2 and l, where $|f_{y'}| \leqq l$ in $\mathscr{R}$, and for $\lambda = -lk_1^{-1} + O(\varepsilon) < 0$ a root of $\varepsilon\lambda^2 + k_1\lambda + l$. Having defined α and β, we choose u and v as appropriate solutions of the initial value problems

$$u' = \alpha(t, \varepsilon), \qquad u(a, \varepsilon) = A_0,$$

$$v' = \beta(t, \varepsilon), \qquad v(a, \varepsilon) = A_0,$$

and thereby satisfy the inequalities in (4.7). It is now a straightforward exercise to verify that α and β satisfy the required algebraic inequalities and the differential inequalities in (4.8) for ε sufficiently small, say $0 < \varepsilon \leqq \varepsilon_0$ (cf. the argument in [14]). Thus, by virtue of Theorem 2.1, we conclude that the problem (4.1) (with $n = 3$) has a solution $y = y(t, \varepsilon)$ for $0 < \varepsilon \leqq \varepsilon_0$ satisfying on $[a, b]$

$$u(t, \varepsilon) \leqq y(t, \varepsilon) \leqq v(t, \varepsilon)$$

and

$$\alpha(t, \varepsilon) \leqq y'(t, \varepsilon) \leqq \beta(t, \varepsilon);$$

that is, the conclusion of Theorem 4.1 obtains. □

A similar argument establishes this result for general $n \geqq 3$.

In the same vein, we can extend the interior crossing results of [13] to the problem (4.1) with f independent of $y^{(n-1)}$, namely the problem

$$\begin{aligned} &\varepsilon y^{(n)} = h(t, y, y', \cdots, y^{(n-2)}), \qquad a < t < b, \quad n \geqq 3, \\ &y^{(j)}(a, \varepsilon) = A_j, \quad 0 \leqq j \leqq n-2, \qquad y^{(n-2)}(b, \varepsilon) = B_{n-2}. \end{aligned} \tag{4.9}$$

To this end, suppose that the corresponding reduced equation

$$0 = h(t, Y, Y', \cdots, Y^{(n-2)}), \qquad a < t < b, \tag{4.10}$$

has solutions $Y = Y_L(t)$ and $Y = Y_R(t)$ with the same properties as the solutions of (4.2), and that the smooth function h satisfies, for example,

$$h_{y^{(n-2)}}[Y(t)] \geqq m > 0 \quad \text{on } [a, b]$$

for

$$Y(t) = \begin{cases} Y_L(t), & a \leqq t \leqq t_0, \\ Y_R(t), & t_0 \leqq t \leqq b, \end{cases}$$

and a positive constant m (cf. [13]). Then we can use Theorem 2.2 to show that the problem (4.9) has a smooth solution $y = y(t, \varepsilon)$ for each $\varepsilon > 0$ sufficiently small which satisfies on $[a, b]$

$$y^{(i)}(t, \varepsilon) = Y^{(i)}(t) + O(\varepsilon), \qquad 0 \leqq i \leqq n-3,$$

and

$$y^{(n-2)}(t, \varepsilon) = Y^{(n-2)}(t) + v_I(t, \varepsilon) + O(\varepsilon).$$

Here $v_I(t, \varepsilon) = (\varepsilon m_1^{-1})^{1/2} |\mu_L - \mu_R| \exp[-(m_1\varepsilon^{-1})^{1/2}|t - t_0|]$, $0 < m_1 < m$, is the interior layer corrector at $t = t_0$.

Up to now we have considered interior layer phenomena for the problem (4.1); however, we can treat boundary layer phenomena just as easily.

For example, let us assume that the solution $Y = Y_L(t)$ of the reduced equation (4.2) exists on all of $[a, b]$ (and, of course, satisfies $Y_L^{(j)}(a) = A_j$ for $j = 0, \cdots, n-2$).

Consideration of linear problems of this form (cf. [28] or [31]) suggests that if there exists a positive constant k such that

$$f_{y^{(n-1)}}[Y_L(t)] \geqq k > 0 \quad \text{on } [a, b],$$

and if f is further restricted, then the problem (4.1) has a solution $y = y(t, \varepsilon)$ for each sufficiently small $\varepsilon > 0$ such that

$$\lim_{\varepsilon \to 0^+} y^{(i)}(t, \varepsilon) = Y_L^{(i)}(t), \qquad 0 \leqq i \leqq n-3, \quad \text{for } a \leqq t \leqq b,$$

while

$$\lim_{\varepsilon \to 0^+} y^{(n-2)}(t, \varepsilon) = Y_L^{(n-2)}(t) \quad \text{for } a \leqq t < b,$$

if $Y_L^{(n-2)}(b) \neq B_{n-2}$. In other words, the $(n-2)$nd derivative of the solution exhibits boundary layer behavior at $t = b$ due to the lower order of the reduced equation.

To make these ideas precise, let us first define the regions $\mathscr{D}$ and $\mathscr{D}_{\pm}$ by

$$\mathscr{D} = [a, b] \times \Omega \times \{z : |z - Y_L^{(n-2)}(t)| \leqq d_1(t)\} \times \{w : |w - Y_L^{(n-1)}(t)| \leqq d_2(t)\},$$

$$\mathscr{D}_+ = \{(t, y_1, \cdots, y_{n-2}, z, w) \text{ in } \mathscr{D} : w - Y_L^{(n-1)}(t) \geqq 0 \text{ on } [b-\delta, b]\},$$

and

$$\mathscr{D}_- = \{(t, y_1, \cdots, y_{n-2}, z, w) \text{ in } \mathscr{D} : w - Y_L^{(n-1)}(t) \leqq 0 \text{ on } [b-\delta, b]\}.$$

Here $\Omega = \{(y_1, \cdots, y_{n-2}) : |y_i - Y_L^{(i-1)}(t)| \leqq \delta, i = 1, \cdots, n-2\}$ for a small positive constant δ, d_1 is a smooth positive function such that $(\sigma =) |B_{n-2} - Y_L^{(n-2)}(b)| \leqq d_1(t) \leqq \sigma + \delta$ on $[b - \delta/2, b]$ and $d_1(t) \leqq \delta$ on $[a, b-\delta]$, and d_2 is a smooth positive function such that $\varepsilon^{-1}\sigma \exp[-k(b-t)\varepsilon^{-1}] \leqq d_2(t) \leqq \varepsilon^{-1}\sigma \exp[-k(b-t)\varepsilon^{-1}] + \delta$ on $[b-\delta/2, b]$ and $d_2(t) \leqq \delta$ on $[a, b-\delta]$. The functions d_1 and d_2 have the properties near b which reflect the anticipated boundary layer behavior of the $(n-2)$nd derivative of the solution there.

The following two results are extensions of a recent result of Goecke [5] on the third-order problem (4.1). In the first one, we assume that f grows at most linearly as a function of $y^{(n-1)}$.

THEOREM 4.2. *Suppose that*

(1) *reduced equation* (4.2) *has a solution* $Y = Y_L(t)$ *of class* $C^{(n)}[a, b]$ *such that* $Y_L^{(j)}(a) = A_j$ *for* $j = 0, \cdots, n-2$;

(2) *the function* f *is continuous with respect to* $(t, y_1, \cdots, y_{n-2}, z, w)$ *and continuously differentiable with respect to* $(y_1, \cdots, y_{n-2}, z, w)$ *in* $\mathscr{D}$;

(3) *there exists a positive constant* k *such that* $f_w \geqq k > 0$ *in* $\mathscr{D}$;

(4) *for* $(t, y_1, \cdots, y_{n-2}, z)$ *in compact subsets of* $[a, b] \times \mathbb{R}^{n-1}$

$$f(t, y_1, \cdots, y_{n-2}, z, w) = O(|w|) \quad \text{as } |w| \to \infty.$$

Then there exists an $\varepsilon_0 > 0$ *such that the problem* (4.1) *has a solution* $y = y(t, \varepsilon)$ *whenever* $0 < \varepsilon \leqq \varepsilon_0$. *Moreover, for* t *in* $[a, b]$ *we have that*

$$y^{(i)}(t, \varepsilon) = Y_L^{(i)}(t) + O(\varepsilon) \quad \text{for } i = 0, \cdots, n-3,$$

and

$$y^{(n-2)}(t, \varepsilon) = Y_L^{(n-2)}(t) + w_R(t, \varepsilon) + O(\varepsilon).$$

Here $w_R(t, \varepsilon) = |B_{n-2} - Y_L^{(n-2)}(b)| \exp[ct] \exp[k(t-b)\varepsilon^{-1}]$ *(for* c *a known positive constant) is the boundary layer corrector at* $t = b$.

The analogous result for a right-hand side f which depends quadratically on $y^{(n-1)}$ is given by the next theorem.

THEOREM 4.3. *Suppose that*

(1) *the reduced equation* (4.2) *has a solution* $Y = Y_L(t)$ *of class* $C^{(n)}[a, b]$ *such that* $Y_L^{(j)}(a) = A_j$ *for* $j = 0, \cdots, n-2$, *and* $Y_L^{(n-2)}(b) < B_{n-2} (Y_L^{(n-2)}(b) > B_{n-2})$;

(2) *the function* f *is continuous with respect to* $(t, y_1, \cdots, y_{n-2}, z, w)$ *and twice continuously differentiable with respect to* $(y_1, \cdots, y_{n-2}, z, w)$ *in* $\mathscr{D}_+(\mathscr{D}_-)$;

(3) $f_w \geqq 0$ *in* $\mathscr{D}_+(\mathscr{D}_-)$;

(4) *there exists a positive constant* k *such that* $f_{ww} \geqq k > 0$ $(f_{ww} \leqq -k < 0)$ *in* $\mathscr{D}_+(\mathscr{D}_-)$;

(5) *for* $(t, y_1, \cdots, y_{n-2}, z)$ *in compact subsets of* $[a, b] \times \mathbb{R}^{n-1}$

$$f(t, y_1, \cdots, y_{n-2}, z, w) = O(|w|^2) \quad \text{as } |w| \to \infty.$$

Then the conclusion of Theorem 4.2 *is valid with* $w_R(t, \varepsilon)$ *replaced by* $v_R(t, \varepsilon) = k_1 \varepsilon \ln [(b-a)^{-1}(b-t+\{t-a\} \exp [-(k_1\varepsilon)^{-1}|B_{n-2} - Y_L^{(n-2)}(b)|])]$ *for* $0 < k_1 < k$, *where* v_R *is the decaying solution of* $\varepsilon v'' = k_1 v'^2$, $v(b, \varepsilon) = -|B_{n-2} - Y_L^{(n-2)}(b)|$.

Theorems 4.2 and 4.3 are proved by applying the theory of [11] and [15] to the z-equation of system (4.4) and noting that the components $(y_1, \cdots, y_{n-2})$ of the solutions are $O(\varepsilon)$-perturbations of the corresponding derivatives $(Y_L, \cdots, Y_L^{(n-3)})$ of the reduced solution Y_L. The details are straightforward and are omitted.

These two results are distinguished by the assumption that the partial derivatives f_w or f_{ww} have a certain sign along the reduced solution Y_L and inside of the boundary layer at $t = b$. Suppose however that f is independent of $y^{(n-1)}$ and that the reduced equation (4.10) has a solution $Y = Y(t)$ satisfying $Y^{(i)}(a) = A_i$ for $i = 0, \cdots, n-3$. If Y is stable in the sense that $h_{y^{(n-2)}} > 0$ in the region $\mathscr{G} = \mathscr{D} \cap \mathbb{R}^n$, then we expect (cf. [13] or [15]) that the problem (4.9) has a solution $y = y(t, \varepsilon)$ for each $\varepsilon > 0$ sufficiently small such that

$$\lim_{\varepsilon \to 0^+} y^{(i)}(t, \varepsilon) = Y^{(i)}(t), \qquad 0 \leqq i \leqq n-3, \quad \text{for } a \leqq t \leqq b,$$

and

$$\lim_{\varepsilon \to 0^+} y^{(n-2)}(t, \varepsilon) = Y^{(n-2)}(t) \quad \text{for } a < t < b.$$

The precise statement is the next theorem which contains a more general stability criterion than a corresponding result of Goecke [5] for a third-order problem. It is proved by converting the problem to a system and applying results in [13] or [15] to the second-order problem for $z = y^{(n-2)}$.

THEOREM 4.4. *Suppose that*

(1) *the reduced equation* (4.10) *has a solution* $Y = Y(t)$ *of class* $C^{(n)}[a, b]$ *such that* $Y^{(i)}(a) = A_i$ *for* $i = 0, \cdots, n-3$;

(2) *the function* $h = h(t, y_1, \cdots, y_{n-2}, z)$ *is continuous with respect to* $(t, y_1, \cdots, y_{n-2}, z)$ *and continuously differentiable with respect to* $(y_1, \cdots, y_{n-2}, z)$ *in* $\mathscr{G}$;

(3) *there exists a positive constant* m *such that* $h_z[Y(t)] \geqq m > 0$ *for* $a \leqq t \leqq b$;

(4) *if* $\mu = Y^{(n-2)}(a) \neq A_{n-2}$, *then*

$$[A_{n-2} - \mu] \int_{\mu}^{\xi} h(a, A_0, \cdots, A_{n-3}, s)\, ds > 0$$

for $\mu < \xi \leqq A_{n-2}$ *or* $A_{n-2} \leqq \xi < \mu$; *while if* $\nu = Y^{(n-2)}(b) \neq B_{n-2}$, *then*

$$[B_{n-2} - \nu] \int_{\nu}^{\eta} h(b, Y(b), \cdots, Y^{(n-3)}(b), s)\, ds > 0$$

for $\nu < \eta \leqq B_{n-2}$ *or* $B_{n-2} \leqq \eta < \nu$.

Then there exists an $\varepsilon_0 > 0$ *such that the problem* (4.9) *has a solution* $y = y(t, \varepsilon)$ *whenever* $0 < \varepsilon \leqq \varepsilon_0$. *Moreover, for* t *in* $[a, b]$ *we have that*

$$y^{(i)}(t, \varepsilon) = Y^{(i)}(t) + O(\varepsilon), \qquad i = 0, \cdots, n-4,$$
$$y^{(n-3)}(t, \varepsilon) = Y^{(n-3)}(t) + O(\varepsilon^{1/2} w_L(t, \varepsilon)) + O(\varepsilon^{1/2} w_R(t, \varepsilon)) + O(\varepsilon),$$

and

$$y^{(n-2)}(t, \varepsilon) = Y^{(n-2)}(t) + O(w_L(t, \varepsilon)) + O(w_R(t, \varepsilon)) + O(\varepsilon),$$

where $w_L(t, \varepsilon) = |A_{n-2} - Y^{(n-2)}(a)| \exp[-(m_1 \varepsilon^{-1})^{1/2}(t-a)]$ *and* $w_R(t, \varepsilon) = |B_{n-2} - Y^{(n-2)}(b)| \exp[-(m_1 \varepsilon^{-1})^{1/2}(b-t)]$ *for* $0 < m_1 < m$.

The previous theorems suggest that if a problem of the form (4.1) is expressed as the system (4.4) and if the corresponding reduced problem has a smooth solution (Y, Z), then one should be able to adapt theorems for perturbed second-order boundary value problems to this more general context. The idea is that if the $(n-2)$nd derivative of the solution y (that is, z) exhibits nonuniform behavior in $[a, b]$ of the type described above, then the lower order derivatives $y^{(i)} (i = 0, \cdots, n-3)$ should be close to the corresponding derivatives $Y^{(i)}$ of the solution of the reduced problem in the sense that $y^{(i)}(t, \varepsilon) = Y^{(i)}(t) + o(1)$ on $[a, b]$ for $i = 0, \cdots, n-3$. Consequently, these variables in the right-hand side of (4.1) can be replaced (to terms of order $o(1)$) by the derivatives of the known reduced solution. One now has an algorithm for constructing asymptotic approximations to the solutions of many singularly perturbed systems of higher order.

We discuss next an application of Theorem 3.1 to the problem on (a, b)

$$\begin{aligned} &\mathbf{y}' = \mathbf{z}, \qquad \mathbf{y}(a, \varepsilon) = \mathbf{A}, \\ &\varepsilon \mathbf{z}'' = \mathbf{H}(\mathbf{t}, \mathbf{y}, \mathbf{z}), \quad \mathbf{z}(a, \varepsilon) = \boldsymbol{\xi}, \quad \mathbf{z}(b, \varepsilon) = \boldsymbol{\eta}. \end{aligned} \tag{4.11}$$

Here $\mathbf{y}$, $\mathbf{z}$, $\mathbf{A}$, $\mathbf{H}$, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are in $\mathbb{R}^n$, and the corresponding reduced problem is

$$\begin{aligned} &\mathbf{Y}' = \mathbf{Z}, \qquad \mathbf{Y}(a) = \mathbf{A}, \\ &\mathbf{0} = \mathbf{H}(t, \mathbf{Y}, \mathbf{Z}). \end{aligned} \tag{4.12}$$

For simplicity, let us assume that $\mathbf{H}(t, \mathbf{A}, \mathbf{0}) \equiv \mathbf{0}$, that is, we will study the behavior of solutions of (4.11) relative to the solution $(\mathbf{Y}, \mathbf{Z}) = (\mathbf{A}, 0)$ of (4.12). And let us define the region $\mathcal{K}$ by $\mathcal{K} = \{(t, \mathbf{y}, \mathbf{z}): a \leqq t \leqq b, \|\mathbf{y} - \mathbf{A}\| \leqq \delta, \|\mathbf{z}\| \leqq d(t)\}$, where $\delta > 0$ is a small constant and $d > 0$ is a smooth function such that $\|\boldsymbol{\xi}\| \leqq d(t) \leqq \|\boldsymbol{\xi}\| + \delta$ on $[a, a + \delta/2]$, $d(t) \leqq \delta$ on $[a + \delta, b - \delta]$, and $\|\boldsymbol{\eta}\| \leqq d(t) \leqq \|\boldsymbol{\eta}\| + \delta$ on $[b - \delta/2, b]$. Then the results of [17] lead us to assume, for example, that the Jacobian matrix $J = \partial \mathbf{H}/\partial \mathbf{z}$ satisfies

$$\mathbf{z} \cdot J\mathbf{z} \geqq m\|\mathbf{z}\|^2 \quad \text{in } \mathcal{K} \tag{4.13}$$

for a positive constant m. We expect that under this assumption the full problem (4.11) has a solution $(\mathbf{y}, \mathbf{z}) = (\mathbf{y}(t, \varepsilon), \mathbf{z}(t, \varepsilon))$ for each $\varepsilon > 0$ sufficiently small such that

$$\lim_{\varepsilon \to 0^+} \|\mathbf{y}(t, \varepsilon) - \mathbf{A}\| = 0 \quad \text{for } a \leqq t \leqq b,$$

and

$$\lim_{\varepsilon \to 0^+} \mathbf{z}(t, \varepsilon) = \mathbf{0} \quad \text{for } a < t < b.$$

The precise result is the next theorem whose proof is left to the reader.

THEOREM 4.5. *Suppose that*

(1) *the reduced problem* (4.12) *has a solution* $(\mathbf{Y}, \mathbf{Z}) \equiv (\mathbf{A}, \mathbf{0})$;

(2) *the function* $\mathbf{H}$ *is continuous with respect to* $(t, \mathbf{y}, \mathbf{z})$ *and continuously differentiable with respect to* $(\mathbf{y}, \mathbf{z})$ *in the region* $\mathscr{K}$;

(3) *there exists a positive constant* m *such that the inequality* (4.13) *holds.*

Then there exists an $\varepsilon_0 > 0$ *such that the problem* (4.11) *has a solution* $(\mathbf{y}, \mathbf{z}) = (\mathbf{y}(t, \varepsilon), \mathbf{z}(t, \varepsilon))$ *whenever* $0 < \varepsilon \leqq \varepsilon_0$. *Moreover, for* t *in* $[a, b]$ *we have that*

$$\|\mathbf{y}(t, \varepsilon) - \mathbf{A}\| = O(\varepsilon^{1/2} w_L(t, \varepsilon)) + O(\varepsilon^{1/2} w_R(t, \varepsilon))$$

and

$$\|\mathbf{z}(t, \varepsilon)\| = O(w_L(t, \varepsilon)) + O(w_R(t, \varepsilon)),$$

where $w_L(w_R)$ *is as in the conclusion of Theorem* 4.4 *with the pre-exponential factor replaced by* $\|\boldsymbol{\xi}\|(\|\boldsymbol{\eta}\|)$.

Before turning to a discussion of some examples in the next section, we make several remarks.

Remark 4.1. Theorems 4.2 and 4.3 describe the occurrence of boundary layer behavior at $t = b$ under the assumption that the reduced equation has a stable solution satisfying the $(n-1)$ conditions at $t = a$. However, suppose that the reduced equation has a smooth solution $Y = Y_R(t)$ on $[a, b]$ such that

$$Y_R^{(i)}(a) = A_i \quad \text{for } i = 0, \cdots, n-3 \tag{4.14}$$

and

$$Y_R^{(n-2)}(b) = B_{n-2}, \tag{4.15}$$

and that Y_R is stable in the sense that for $w = y^{(n-1)}$

$$f_w[Y_R(t)] < 0 \quad \text{on } [a, b]$$

and

$$f_w < 0$$

in the boundary layer at $t = a$. Then we can show that the problem (4.1) has a solution $y = y(t, \varepsilon)$ for each $\varepsilon > 0$ sufficiently small such that

$$\lim_{\varepsilon \to 0^+} y^{(i)}(t, \varepsilon) = Y_R^{(i)}(t), \qquad 0 \leqq i \leqq n-3, \quad \text{for } a \leqq t \leqq b$$

and

$$\lim_{\varepsilon \to 0^+} y^{(n-2)}(t, \varepsilon) = Y_R^{(n-2)}(t) \quad \text{for } a < t \leqq b.$$

Similarly, the analogue of Theorem 4.4 can be established for a reduced solution $Y = Y_R(t)$ satisfying (4.14) and (4.15) (cf. Example 5.1 below).

Remark 4.2. The results of this section can be veiwed as providing simple illustrations of what could be called "nonlinear cancellation laws." Namely, for higher-order singularly perturbed *linear* boundary value problems one studies the asymptotic behavior of solutions relative to the solution of the reduced equation which satisfies a lesser number of the original boundary conditions. The choice of these "reduced" boundary conditions is dictated by rather general cancellation laws (cf. [31], [28] and [29]). However, a problem of current interest in higher-order nonlinear singular perturbation theory is the discovery of the appropriate analogues of these

linear cancellation results. The above theory suggests one possible approach to this problem.

Remark 4.3. In addition to the boundary layer and interior crossing layer behavior discussed here, solutions of (4.1) may also exhibit shock layer behavior in the sense that the $(n-2)$nd derivative may transfer from one reduced solution to another discontinuously in the limit of $\varepsilon = 0$. The study of such phenomena proceeds as in the above analysis by applying second-order results on shock layer behavior contained, for example, in [15] (cf. Example 5.5 below).

Remark 4.4. The boundary conditions for the z-parts of our problems have been of Dirichlet type. However, there is now a fairly well-developed theory for perturbed scalar second-order differential equations whose solutions satisfy boundary conditions of Robin or Neumann type (cf. [16], [18]). In terms of the differential equation (4.1) this means that we could impose boundary conditions of the form

$$y^{(i)}(a, \varepsilon) = A_i, \qquad i = 0, \cdots, n-3,$$
$$p_1 y^{(n-2)}(a, \varepsilon) - q_1 y^{(n-1)}(a, \varepsilon) = A_{n-2},$$
$$p_2 y^{(n-2)}(b, \varepsilon) + q_2 y^{(n-1)}(b, \varepsilon) = B_{n-2},$$

for nonnegative constants p_k, q_k such that $p_k + q_k > 0$, $k = 1, 2$. The interested reader should have little difficulty applying the results of [16] and [18] to perturbed problems of higher order.

5. Illustrations of the theory. In this section we present several examples whose solutions display some of the behavior outlined above.

Example 5.1. Let us consider first the problem

$$\text{(E1)} \qquad \begin{aligned} \varepsilon y''' &= 1 - (y'')^2 = f(y''), \qquad 0 < t < 1, \\ y(0, \varepsilon) &= A_0, \quad y'(0, \varepsilon) = A_1, \quad y'(1, \varepsilon) = B_1, \end{aligned}$$

which is uniquely solvable for all $\varepsilon > 0$. (Existence follows from Theorem 2.1, while uniqueness follows from the maximum principle [30].) For various choices of A_1 and B_1 solutions of (E1) exhibit interior and boundary layer behavior of the types described by Theorems 4.1 and 4.3, respectively.

To see this, we first examine the reduced equation

$$\text{(*)} \qquad 0 = 1 - (u'')^2,$$

and find that $u = u_L(t) = -\frac{1}{2}t^2 + A_1 t + A_0$ and $u = \tilde{u}_L(t) = \frac{1}{2}t^2 + A_1 t + A_0$ are solutions of (*) which satisfy the left-hand boundary conditions. Similarly, we find that $u = u_R(t) = \frac{1}{2}t^2 + (B_1 - 1)t + c_1$ and $u = \tilde{u}_R(t) = -\frac{1}{2}t^2 + (B_1 - 1)t + c_2$ are one-parameter families of solutions of (*) which satisfy the right-hand boundary condition. Since $f'(\tilde{u}_L) < 0$ and $f'(\tilde{u}_R) > 0$, while $f'(u_L) > 0$ and $f'(u_R) < 0$, we can reject $\tilde{u}_L$ and $\tilde{u}_R$ as candidates for limiting solutions as $\varepsilon \to 0^+$, and thus concentrate on the stable functions u_L and u_R. We note that u_L is uniquely determined at this stage, while u_R still contains the free constant c_1.

Let us now attempt to apply Theorem 4.1. We see that $u_L' = u_R'$ at the point $t_0 = \frac{1}{2}(A_1 - B_1 + 1)$ and that t_0 belongs to $(0, 1)$ only if $|A_1 - B_1| < 1$. Next, we also require that $u_L(t_0) = u_R(t_0)$, and in order to achieve this, we choose the constant c_1 as $A_0 + \frac{1}{4}(A_1 - B_1 + 1)^2$. Finally, we note that $f(\lambda) = 1 - \lambda^2 > 0$ for $|\lambda| < 1$. Thus for $|A_1 - B_1| < 1$ and $u_R(t) = \frac{1}{2}t^2 + (B_1 - 1)t + A_0 + \frac{1}{4}(A_1 - B_1 + 1)^2$, all of the assumptions of Theorem 4.1 obtain, and we conclude that the solution $y = y(t, \varepsilon)$ of (E1) satisfies as

$\varepsilon \to 0^+$

$$y(t, \varepsilon) \to u(t) \quad \text{and} \quad y'(t, \varepsilon) \to u'(t) \quad \text{on } [0, 1],$$

while

$$y''(t, \varepsilon) \to u''(t) \quad \text{on } [0, 1] \backslash \{t_0\}.$$

Here $u = u(t)$ is the composite path defined by $u(t) = u_L(t)$ on $[0, t_0]$ and $u(t) = u_R(t)$ on $[t_0, 1]$.

Suppose next that $A_1 - B_1 \geqq 1$. If $A_1 - B_1 = 1$ then by inspection $y(t, \varepsilon) = u_L(t)$ is the solution of (E1). However, if $A_1 - B_1 > 1$, then $u_L'(1) = A_1 - 1 > B_1$ and we deduce from Theorem 4.3 since $f'' < 0$ that the solution of (E1) satisfies as $\varepsilon \to 0^+$

$$y(t, \varepsilon) \to u_L(t) \quad \text{on } [0, 1]$$

and

$$y'(t, \varepsilon) \to u_L'(t) \quad \text{on } [0, 1).$$

Finally, if $A_1 - B_1 \leqq -1$, then we anticipate using the right-hand reduced solution $u_R^0(t) = \frac{1}{2}t^2 + (B_1 - 1)t + A_0$ which now satisfies $u_R^0(0) = A_0$. This is obvious in the case that $A_1 - B_1 = -1$ since $y(t, \varepsilon) = u_R^0(t)$ is the solution of (E1). More generally, if $A_1 - B_1 < -1$, then $u_R^{0\prime}(0) = B_1 - 1 > A_1$ and we deduce from the analogue of Theorem 4.3 (cf. Remark 4.1) that the solution of (E1) satisfies as $\varepsilon \to 0^+$

$$y(t, \varepsilon) \to u_R^0(t) \quad \text{on } [0, 1]$$

and

$$y'(t, \varepsilon) \to u_R^{0\prime}(t) \quad \text{on } (0, 1].$$

This example serves to illustrate the important fact that for higher-order singular perturbation phenomena different members of a family of stable reduced solutions often must be used to describe the asymptotic behavior of solutions of the full problem for different choices of the boundary values (cf. Remark 4.2).

Example 5.2. In this example we illustrate our remarks on the problem (4.9) which followed the proof of Theorem 4.1. Consider then the problem

$$\text{(E2)} \qquad \begin{aligned} \varepsilon y''' &= y' - 2|t| = h(t, y'), \qquad -1 < t < 1, \\ y(-1, \varepsilon) &= A_0, \quad y'(-1, \varepsilon) = 2, \quad y'(1, \varepsilon) = 2. \end{aligned}$$

As with the previous example, the existence and uniqueness of solutions of (E2) are guaranteed by Theorem 2.1 and the maximum principle. However, to understand their asymptotic behavior as $\varepsilon \to 0^+$ we must apply the theory of § 4.

Let us begin by noting that

$$u = u(t) = \begin{cases} A_0 + 1 - t^2, & -1 \leqq t \leqq 0, \\ A_0 + 1 + t^2, & 0 \leqq t \leqq 1 \end{cases}$$

is the solution of the reduced equation $u' = 2|t|$ which satisfies all three boundary conditions. Moreover, it is stable in the sense that $h_{y'} \equiv 1 > 0$. From our remarks on (4.9) we conclude therefore that the solution $y = y(t, \varepsilon)$ of (E2) satisfies as $\varepsilon \to 0^+$

$$y(t, \varepsilon) \to u(t) \quad \text{and} \quad y'(t, \varepsilon) \to u'(t) \quad \text{on } [-1, 1].$$

We note that if $y'(-1, \varepsilon) \neq 2$ and/or $y'(1, \varepsilon) \neq 2$, then we can combine our interior layer theory with the boundary layer theory of Theorem 4.4 to show that

$$y(t, \varepsilon) \to u(t) \quad \text{on } [-1, 1]$$

but

$$y'(t, \varepsilon) \to u'(t) \quad \text{on } (-1, 1).$$

The latter limit is attained, of course, at $t = -1 (t = 1)$ if $y'(-1, \varepsilon) = 2 (y'(1, \varepsilon) = 2)$.

Example 5.3. We consider now an illustration of the integral conditions in Theorem 4.4. The problem is

$$\begin{aligned} &\varepsilon y''' = y'^2 - (t+1)^2 = h(t, y'), \qquad 0 < t < 1, \\ &y(0, \varepsilon) = A_0, \quad y'(0, \varepsilon) = A_1, \quad y'(1, \varepsilon) = B_1, \end{aligned} \tag{E3}$$

and its reduced equation $u'^2 = (t+1)^2$ has the two solutions $u = u(t) = A_0 + \frac{1}{2}t^2 + t$ and $u = \tilde{u}(t) = A_0 - \frac{1}{2}t^2 - t$ satisfying $u(0) = \tilde{u}(0) = A_0$. However, since $h_{y'}(t, u(t)) > 0$ while $h_{y'}(t, \tilde{u}(t)) < 0$, we select $u(t)$ as our candidate for an approximate solution of (E3) on $(0, 1)$. In order to apply Theorem 4.4 we must determine for what range of A_1 and B_1 this function u can attract the solution of (E3) in the boundary layers at $t = 0$ and $t = 1$. First of all, if $A_1 > 1 = u'(0)$, then

$$\int_1^{A_1} h(0, s)\, ds = \frac{1}{3}A_1^3 - A_1 + \frac{2}{3} > 0$$

for all such A_1, while if $A_1 < 1$ then

$$\int_{A_1}^{1} h(0, s)\, ds = -\left(\frac{1}{3}A_1^3 - A_1 + \frac{2}{3}\right) < 0$$

for $A_1 > -2$. Similarly, if $B_1 > 2 = u'(1)$, then

$$\int_2^{B_1} h(1, s)\, ds = \frac{1}{3}B_1^3 - 4B_1 + \frac{16}{3} > 0$$

for all such B_1, while if $B_1 < 2$ then

$$\int_{B_1}^{2} h(1, s)\, ds = -\left(\frac{1}{3}B_1^3 - 4B_1 + \frac{16}{3}\right) < 0$$

for $B_1 > -4$. Therefore we deduce from Theorem 4.4 that for $A_1 > -2$ and $B_1 > -4$ the problem (E3) has a solution $y = y(t, \varepsilon)$ for each $\varepsilon > 0$ sufficiently small such that as $\varepsilon \to 0^+$

$$y(t, \varepsilon) \to u(t) \quad \text{on } [0, 1]$$

and

$$y'(t, \varepsilon) \to u'(t) \quad \text{on } (0, 1).$$

The latter limit is attained, of course, at $t = 0$ $(t = 1)$ if $A_1 = 1$ $(B_1 = 2)$.

Example 5.4. It is often necessary to combine the reasoning used in formulating results like Theorem 4.2 and Theorem 4.4 in order to study certain asymptotic phenomena. An illustration is afforded by the problem

$$\begin{aligned} &\varepsilon y''' = y' - (\tfrac{1}{2} - t)yy'' = f(t, y, y', y''), \qquad 0 < t < 1, \\ &y(0, \varepsilon) = A_0, \quad y'(0, \varepsilon) = A_1, \quad y'(1, \varepsilon) = B_1, \end{aligned} \tag{E4}$$

where the data A_0, A_1 and B_1 are positive constants. Since $f_{y'} \equiv 1 > 0$ we are assured of the existence and uniqueness of the solution for each $\varepsilon > 0$.

Let us consider just the solutions $u \equiv$ const. of the reduced equation $u' = (\frac{1}{2} - t)uu''$. If we ask that $u(0) = A_0$, then $u(t) \equiv A_0 > 0$, and we must see whether u is stable in any of our previous senses. First of all,

$$f_{y''}(t, A_0, 0, 0) = A_0(t - \tfrac{1}{2})$$

is negative (positive) on $[0, \frac{1}{2})((\frac{1}{2}, 1])$ and zero at $t = \frac{1}{2}$, and so we might be tempted to reject it out of hand as an approximate solution. Nevertheless, $f_{y'} \equiv 1$ and so $u \equiv A_0$ does possess the type of stability contained in Theorem 4.4. If u were to approximate the solution of (E4) on $[0, 1]$ with the exception of boundary layer regions at 0 and 1 (since $A_1, B_1 > 0$), then we would have to check whether the second derivative term in f could disturb the boundary layer structure. Fortunately, this term enhances the boundary layer behavior of y in the sense that $f_{y''} < 0$ near 0 and $f_{y''} > 0$ near 1 provided only that $y > 0$. Indeed, by arguing as in the previous section, we can show that the solution $y = y(t, \varepsilon)$ of (E4) satisfies on $[0, 1]$

$$y(t, \varepsilon) = A_0 + O(\varepsilon)$$

and

$$y'(t, \varepsilon) = O(v_L(t, \varepsilon)) + O(v_R(t, \varepsilon)) + O(\varepsilon).$$

Here $v_L(t, \varepsilon) = A_1 \exp[-\sigma t \varepsilon^{-1}]$ and $v_R(t, \varepsilon) = B_1 \exp[-\sigma(1-t)\varepsilon^{-1}]$ for $0 < \sigma < \frac{1}{2}A_0$ are the boundary layer correctors at $t = 0$ and $t = 1$, respectively. A general result of this kind can be formulated and proved using the theorems of § 2 together with Theorem 5.7 of [15].

Example 5.5. In order to illustrate Remark 4.3 let us now consider the problem

$$\text{(E5)} \qquad \begin{aligned} &\varepsilon y''' = y' - y'y'', \qquad 0 < t < 1, \\ &y(0, \varepsilon) = A_0, \quad y'(0, \varepsilon) = A_1, \quad y'(1, \varepsilon) = B_1, \end{aligned}$$

for data A_1, B_1 such that

$$B_1 > A_1 + 1 \quad \text{and} \quad -(B_1 + 1) < A_1 < 1 - B_1.$$

This example can be viewed as a higher-order problem of the type first considered by Lagerstrom and Cole [2] (cf. also [15]) if we write (E5) as the system

$$\text{(E5')} \qquad \begin{aligned} &y' = z, \qquad y(0, \varepsilon) = A_0, \\ &\varepsilon z'' = z - zz', \quad z(0, \varepsilon) = A_1, \quad z(1, \varepsilon) = B_1. \end{aligned}$$

Indeed, for our choice of A_1 and B_1 we know (cf. [15] or [4]) that the z-problem in (E5') has a solution $z = z(t, \varepsilon)$ such that

$$\lim_{\varepsilon \to 0^+} z(t, \varepsilon) = \begin{cases} t + A_1, & 0 \leqq t < t^*, \\ t + B_1 - 1, & t^* < t \leqq 1, \end{cases}$$

for $t^* = \frac{1}{2}(1 - A_1 - B_1)$, that is, z exhibits shock layer behavior at $t = t^*$. The function $y = y(t, \varepsilon)$ is then seen to satisfy the asymptotic relation

$$y(t, \varepsilon) \sim \begin{cases} A_0 + \frac{1}{2}t^2 + A_1 t, & 0 \leqq t \leqq t^*, \\ A_0 + \frac{1}{2}t^2 + (B_1 - 1)t + t^*(A_1 - B_1 + 1), & t^* \leqq t \leqq 1, \end{cases}$$

since $y' = z$.

Example 5.6. We consider finally an application of Theorem 4.5 to the system on (0, 1)

$$\mathbf{y}' = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \mathbf{z},$$

$$\varepsilon \mathbf{z}'' = \begin{pmatrix} z_1 - z_1^3 + y_1 y_2 z_2 \\ z_2 - z_2^3 - y_1 y_2 z_1 \end{pmatrix} = \mathbf{H}(\mathbf{y}, \mathbf{z}), \tag{E6}$$

$$\mathbf{y}(0, \varepsilon) = \mathbf{A}_0, \quad \mathbf{z}(0, \varepsilon) = \boldsymbol{\xi}, \quad \mathbf{z}(1, \varepsilon) = \boldsymbol{\eta}.$$

The corresponding reduced ($\varepsilon = 0$) system has the solution $(\mathbf{A}_0, \mathbf{0})$ and we must see if this function is stable in the sense of relation (4.13). However,

$$\mathbf{z} \cdot \mathbf{H} = z_1^2 + z_2^2 - (z_1^4 + z_2^4) \geqq \|\mathbf{z}\|^2 (1 - \|\mathbf{z}\|^2)$$

since $z_1^4 + z_2^4 \leqq (z_1^2 + z_2^2)^2$. As a result, Theorem 4.5 allows us to deduce the existence of a solution $(\mathbf{y}, \mathbf{z}) = (\mathbf{y}(t, \varepsilon), \mathbf{z}(t, \varepsilon))$ of (E6) for each $\varepsilon > 0$ sufficiently small provided $\|\boldsymbol{\xi}\| < 1$ and $\|\boldsymbol{\eta}\| < 1$. Moreover, for t in [0, 1] we have that

$$\|\mathbf{y}(t, \varepsilon) - \mathbf{A}_0\| = O(\varepsilon^{1/2})$$

and

$$\|\mathbf{z}(t, \varepsilon)\| = O(\|\boldsymbol{\xi}\| \exp[-\gamma t \varepsilon^{-1/2}]) + O(\|\boldsymbol{\eta}\| \exp[-\gamma(1-t)\varepsilon^{-1/2}]) + O(\varepsilon),$$

for $\gamma = \max\{1 - \|\boldsymbol{\xi}\|^2, 1 - \|\boldsymbol{\eta}\|^2\}$.

We note that if we argue as in [17], then we can show that the solution of (E6) actually exists for $\|\boldsymbol{\xi}\| < \sqrt{2}$ and $\|\boldsymbol{\eta}\| < \sqrt{2}$. and satisfies as $\varepsilon \to 0^+$

$$\|\mathbf{y}(t, \varepsilon) - \mathbf{A}_0\| \to 0 \quad \text{on } [0, 1]$$

and

$$\|\mathbf{z}(t, \varepsilon)\| \to 0 \quad \text{on } (0, 1).$$

Acknowledgments. The author gratefully acknowledges the support of the National Science Foundation. He also wishes to thank Professor L. E. Scriven who suggested looking into perturbed problems of higher order, and Mary Hall who read the original manuscript.

REFERENCES

[1] H. BREZIS, *On a characterization of flow-invariant sets*, Comm. Pure Appl. Math., 23 (1970), pp. 261–263.
[2] J. D. COLE, *Perturbation Methods in Applied Mathematics*, Ginn-Blaisdell, Waltham, MA, 1968.
[3] W. A. COPPEL, *Stability and Asymptotic Behavior of Differential Equations*, Heath, Boston, 1965.
[4] F. W. DORR, S. V. PARTER AND L. F. SHAMPINE, *Application of the maximum principle to singular perturbation problems*, SIAM Rev., 15 (1973), pp. 43–88.
[5] D. M. GOECKE, *Third-order differential inequalities and singular perturbations*, Doctoral dissertation, Univ. of Oklahoma, Norman, 1979.
[6] S. HABER AND N. LEVINSON, *A boundary value problem for a singularly perturbed differential equation*, Proc. Amer. Math. Soc., 6 (1955), pp. 866–872.
[7] P. HARTMAN, *On boundary value problems for systems of ordinary, nonlinear, second order differential equations*, Trans. Amer. Math. Soc., 96 (1960), pp. 493–509.
[8] ———, *On invariant sets and a theorem of Ważewski*, Proc. Amer. Math. Soc., 32 (1972), pp. 511–520.
[9] ———, *Ordinary Differential Equations*, Hartman, Baltimore, 1973.

[10] J. W. HEIDEL, *A second-order nonlinear boundary value problem*, J. Math. Anal. Appl., 48 (1974), pp. 493–503.
[11] F. A. HOWES, *Singular perturbations and differential inequalities*, Memoirs Amer. Math. Soc., 168, 1976.
[12] ———, *An application of Nagumo's lemma to some singularly perturbed systems*, Int. J. Nonlinear Mech., 10 (1975), pp. 315–324.
[13] ———, *A class of boundary value problems whose solutions possess angular limiting behavior*, Rocky Mtn. J. Math., 4 (1976), pp. 591–607.
[14] ———, *Singularly perturbed boundary value problems with angular limiting solutions*, Trans. Amer. Math. Soc., 241 (1978), pp. 155–182.
[15] ———, *Boundary-interior layer interactions in nonlinear singular perturbation theory*, Memoirs Amer. Math. Soc., 203, 1978.
[16] ———, *An asymptotic theory for a class of nonlinear Robin problems*, J. Differential Equations, 30 (1978), pp. 192–234.
[17] ———, *Singularly perturbed semilinear systems*, Studies in Appl. Math., 61 (1979), pp. 185–209.
[18] ———, *An asymptotic theory for a class of nonlinear Robin problems: II*, Trans. Amer. Math. Soc., 260 (1980), pp. 527–552.
[19] J. E. INNES AND L. K. JACKSON, *Nagumo conditions for ordinary differential equations*, in International Conf. on Differential Equations, H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 385–398.
[20] L. K. JACKSON, *Subfunctions and second-order ordinary differential inequalities*, Advances in Math., 2 (1968), pp. 307–363.
[21] W. G. KELLEY, *Some existence theorems for n-th order boundary value problems*, J. Differential Equations, 18 (1975), pp. 158–169.
[22] ———, *A geometric method of studying two point boundary value problems for second order systems*, Rocky Mtn. J. Math., 7 (1977), pp. 251–263.
[23] G. A. KLAASEN, *Differential inequalities and existence theorems for second and third order boundary value problems*, J. Differential Equations, 10 (1971), pp. 529–537.
[24] A. LASOTA AND J. A. YORKE, *Existence of solutions of two point boundary value problems for nonlinear systems*, ibid, 11 (1972), pp. 509–518.
[25] N. LEVINSON, *A boundary value problem for a singularly perturbed differential equation*, Duke Math. J., 25 (1958), pp. 331–343.
[26] M. NAGUMO, *Über die Differentialgleichung* $y'' = f(x, y, y')$, Proc. Phys. Math. Soc. Japan, 19 (1937), pp. 861–866.
[27] ———, *Über das Verhalten der Integrale von* $\lambda y'' + f(x, y, y', \lambda) = 0$ *für* $\lambda \to 0$, ibid., 21 (1939), pp. 529–534.
[28] R. E. O'MALLEY, JR., *Introduction to Singular Perturbations*, Academic Press, New York, 1974.
[29] R. E. O'MALLEY, JR. AND J. B. KELLER, *Loss of boundary conditions in the asymptotic solution of linear ordinary differential equations, II: boundary value problems*, Comm. Pure Appl. Math., 21 (1968), pp. 263–270.
[30] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.
[31] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, Wiley-Interscience, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0006 $01.00/0

AN ABSTRACT PARABOLIC VOLTERRA INTEGRODIFFERENTIAL EQUATION*

MELVIN L. HEARD†

Abstract. We consider semilinear integrodifferential equations of the form

$$u'(t)+A(t)u(t)=\int_0^t [a(t,s)g_0(s,u(s))+g_1(t,s,u(s))]\,ds+f_0(t)+f_1(t,u(t)),$$

$$u(0)=u_0.$$

For each $t\geqq 0$, the operator $A(t)$ ia assumed to be the negative generator of an analytic semigroup in a Banach space X. Thus, our models are Volterra integrodifferential equations of parabolic type. These types of equations arise naturally in the study of heat flow in materials with memory. Our main results are the proofs of local and global existence, uniqueness, continuous dependence and differentiability of solutions.

1. Introduction. We consider the abstract Cauchy problem for the Volterra integrodifferential equation

$$u'(t)+A(t)u(t)=\int_0^t [a(t,s)g_0(s,u(s))+g_1(t,s,u(s))]\,ds + f_0(t)+f_1(t,u(t)), \qquad t\geqq 0, \tag{1.1}$$

$$u(0)=u_0. \tag{1.2}$$

We assume that for each $t\geqq 0$ the operator $A(t)$ is the negative generator of an analytic semigroup in a Banach space X. Thus we regard (1.1), (1.2) as a Volterra integrodifferential equation of parabolic type. These equations arise in problems concerned with heat flow in materials with memory (see [5], [15] and references listed there).

Our formulation of (1.1), (1.2) is a direct attempt to generalized some results of G. F. Webb [21], [22] who studied problems similar to (1.1), (1.2) for the case when $A=A(t)$ does not depend on t. We were also influenced by the work of Friedman and Shinbrot [10] and W. E. Fitzgibbon [8].

Highly nonlinear versions of (1.1) have been considered by many authors, viz. Barbu [2], [3], Barbu and Malik [4], Crandall, Londen and Nohel [7], R. C. MacCamy [13], Rennolet [17] and Vrabie [20]. In [7] a complete existence theory (as well as boundedness and asymptotic behavior) is developed for the Volterra integrodifferential equation

$$u'(t)+Bu(t)+\int_0^t a(t-s)Au(s)\,ds \ni F(t), \qquad t\geqq 0, \tag{1.3}$$

$$u(0)=u_0.$$

The operators A and B in (1.3) are taken to be subdifferentials of proper, lower semicontinuous convex functions on a Hilbert space. It is shown that for an appropriate initial function u_0 and forcing term $F(t)$, (1.3) possesses a global strong solution provided the kernel $a(t)$ satisfies some general conditions broad enough to cover two

* Received by the editors December 10, 1979, and in final revised form March 17, 1981.
† Department of Mathematics, University of Illinois at Chicago Circle, Chicago, Illinois 60680.

physically important classes:

(a_1) $a(0)>0$, $a(t)$ is locally absolutely continuous on $[0, \infty)$, $a'(t)$ is locally of bounded variation on $[0, \infty)$;

(a_2) $a(0)>0$, $a \in C[0, \infty) \cap C^2(0, \infty)$ and $a(t)$ is nonnegative, nonincreasing and convex on $[0, \infty)$.

In [17] these conditions are generalized to the case of nonconvolution kernels $a(t, s)$.

In our work the assumptions on the kernel are of a different type (see hypothesis (A5)). We must assume continuity of $a(t, s)$ as a function of (t, s) plus a Hölder continuity assumption in the first variable. A similar hypothesis is made on $f_0(t)$ (see (A4)). These assumptions are related to the method of approach we use in the study of (1.1), (1.2). We treat this problem as a perturbation problem for the linear equation

$$u'(t)+A(t)u(t)=f(t), \qquad t \geqq 0, \tag{1.4}$$

$$u(0)=u_0. \tag{1.5}$$

For basic results on the Cauchy problem (1.4), (1.5) we rely on the operator-theoretic methods of Sobolevsky [18] and Tanabe [19]. Thus our Hölder conditions cannot be appreciably relaxed. The nonlinear term $f_1(t, u)$ is assumed to be of the same type as in [22].

In [7] (and also in [17]) the operators A and B are related by means of assumption (1.7). This essentially says that B dominates A in a certain sense. In our work the same type of assumption is made relating $A(t)$ to the nonlinear operator $g_0(t, u)$ (see, for example, Corollary 2, § 3). The operator $g_1(t, s, u)$ in (1.1) should be regarded as a Lipschitz perturbation of $g_0(t, u)$. It is displayed separately because it has a different range than $g_0(t, u)$, which allows weaker continuity assumptions in the variables (t, s) (see hypothesis (A6)). It is deliberately included in order to have a broader physical application for (1.1). Lipschitz perturbations are also allowed in [7] for (1.3). However, we are able to obtain both existence and uniqueness for problem (1.1), (1.2). Furthermore, by making use of the special nature of (1.1) (i.e., $-A(t)$ generates an analytic semigroup) we are able to use compactness arguments to study (1.1) for certain nonlinear operators $g_0(t, u)$ which are neither monotone nor of Lipschitz type.

The paper is organized as follows. In § 2 we present the basic assumptions and preliminary lemmas. In § 3 we prove existence, uniqueness and continuation theorems for (1.1), (1.2). The basic assumption in this part is that the nonlinear operator $g_0(t, u)$ must satisfy a type of local Lipschitz condition. Section 4 deals with continuous dependence and differentiability (in time) of solutions of (1.1), (1.2). In § 5, by making compactness assumptions, we generalize the results obtained in § 3 to allow for greater nonlinearities in the operator $g_0(t, u)$. In § 6 we briefly indicate how Volterra integrodifferential equations of the type (1.1), (1.2) arise in heat flow problems. We then discuss two examples illustrating the theory of §§ 3 and 5. In § 7 the problem of regularity is considered.

2. Preliminaries. Let X be a Banach space over the complex numbers with norm $\|\cdot\|$. Let $\{A(t): 0 \leqq t \leqq T\}$ be a family of closed linear operators in X satisfying the assumptions:

(A1) The domain $D(A)$ of $A(t)$ is dense in X and does not depend on t.

(A2) For each t in $[0, T]$ the resolvent $R(\lambda; A(t))$ exists for all $\operatorname{Re} \lambda \leqq 0$ and

there is a constant $C>0$ such that

$$\|R(\lambda; A(t))\| \leqq \frac{C}{|\lambda|+1}, \qquad \operatorname{Re}\lambda \leqq 0, \quad 0 \leqq t \leqq T.$$

(A3) There is a constant $C>0$ and a number α, $0<\alpha<1$, such that if t, s, τ belong to $[0, T]$ then

$$\|[A(t)-A(\tau)]A^{-1}(s)\| \leqq C|t-\tau|^{\alpha}.$$

We denote by $\mathscr{L}(X)$ the Banach algebra of all bounded linear operators on X. Then $-A(s)$ is the infinitesimal generator of an analytic semigroup $\{e^{-tA(s)}: t \geqq 0\}$ in $\mathscr{L}(X)$ for each $0 \leqq s \leqq T$ (see [9]). Moreover, there exist positive constants C and δ such that

$$\|e^{-tA(s)}\| \leqq C e^{-\delta t}, \qquad t \geqq 0,$$

$$\|A(s)\, e^{-tA(s)}\| \leqq \frac{C e^{-\delta t}}{t}, \qquad t>0$$

for all $0 \leqq s \leqq T$. For each $\mu>0$ the fractional power $A^{-\mu}(t)$ exists and is given by

$$A^{-\mu}(t) = \frac{1}{\Gamma(\mu)} \int_0^{\infty} e^{-sA(t)} s^{\mu-1}\, ds, \qquad 0 \leqq t \leqq T.$$

It is known that $A^{-\mu}(t)$ is a one-to-one bounded linear operator on X. We define positive fractional powers of $A(t)$ by $A^{\mu}(t) \equiv [A^{-\mu}(t)]^{-1}$. Then $A^{\mu}(t)$ is a closed operator with dense domain $D(A^{\mu}(t))$ (which may depend on t) in X and $D(A^{\mu}(t)) \subset D(A^{\nu}(t))$ if $\mu>\nu$. Furthermore, for all real μ, ν we have

$$A^{\mu}(t)A^{\nu}(t)x = A^{\nu}(t)A^{\mu}(t)x = A^{\nu+\mu}(t)x,$$

if $x \in D(A^{\gamma}(t))$, where $\gamma = \min(\mu, \nu, \mu+\nu)$. Put $A \equiv A(0)$ and for each $y \in D(A^{\mu}(0))$ define $\|y\|_{\mu} = \|A^{\mu}y\|$. Equipped with this norm, $D(A^{\mu})$ is a Banach space which we denote by X_{μ}. For each $0<\mu \leqq 1$ the embedding $X_{\mu} \to X$ is dense and continuous, and by the closed graph theorem, $A(t)A^{-1}(s)$ belongs to $\mathscr{L}(X)$ for all $0 \leqq s, t \leqq T$. So the functional $y \to \|A(t)y\|$ defines for each t an equivalent norm on $D(A)$ and the mapping $t \to A(t)$ is uniformly Hölder continuous from $[0, T]$ into $\mathscr{L}(X_1, X)$. We make the following additional assumptions:

(A4) The function $f_0(t)$ is uniformly Hölder continuous from $[0, T]$ into X with exponent β.

(A5) The function $a(t, s)$ is a continuous complex-valued function on $[0, T] \times [0, T]$ and satisfies a uniform Hölder condition with exponent ρ in the first place; i.e., there exists a constant $a_0>0$ such that

(2.1) $$|a(t, s)-a(\tau, s)| \leqq a_0|t-\tau|^{\rho}$$

for all $0 \leqq t, \tau, s \leqq T$.

(A6) Let W be a nonempty open subset of X_1 and for each $y \in W$ let $B(y, r) = \{z \in X_1: \|y-z\|_1 \leqq r\}$. Let $0<\mu \leqq 1$ be a constant and let

(2.2) $$g_0: [0, T] \times W \to X,$$

(2.3) $$g_1: [0, T] \times [0, T] \times W \to X_1,$$

(2.4) $$f_1: [0, T] \times W \to X_{\mu}$$

be three continuous operators having the property that for each $y \in W$ there exist $r > 0$ and positive continuous functions b_0, b_1, c such that $B(y; r) \subset W$ and

$$\|g_0(t, y_1) - g_0(t, y_2)\| \leqq b_0(t)\|y_1 - y_2\|_1, \tag{2.5}$$

$$\|g_1(t, s, y_1) - g_1(t, s, y_2\|_1 \leqq b_1(t, s)\|y_1 - y_2\|_1, \tag{2.6}$$

$$\|f_1(t, y_1) - f_1(t, y_2)\|_\mu \leqq c(t)\|y_1 - y_2\|_1 \tag{2.7}$$

for all $y_1, y_2 \in B(y; r)$ and $0 \leqq t, s \leqq T$.

Assumptions (A1), (A2), (A3) imply that the Sobolevsky–Tanabe theory of parabolic equations in a Banach space is applicable. This means that if $\Delta = \{(t, s): 0 \leqq s \leqq t \leqq T\}$ then there is a unique evolution operator $\{U(t, s): (t, s) \in \Delta\}$ having the following properties:

(i) $U(t, s) \in \mathscr{L}(X)$ for all $(t, s) \in \Delta$ and $U: \Delta \to \mathscr{L}(X)$ is strongly continuous.

(ii) For each $x \in X$ we have $U(t, s)x \in D(A)$ for all $0 \leqq s < t \leqq T$.

(iii) $U(t, r)U(r, s) = U(t, s)$ for all $0 \leqq s \leqq r \leqq t \leqq T$.

(iv) $U(t, s) \in \mathscr{L}(X_1)$ for all $(t, s) \in \Delta$ and $U: \Delta \to \mathscr{L}(X_1)$ is strongly continuous.

(v) The derivative $(\partial U/\partial t)(t, s)$ exists in the strong topology on $\mathscr{L}(X)$ and belongs to $\mathscr{L}(X)$ for $0 \leqq s < t \leqq T$. It is also strongly continuous in (t, s) for $0 \leqq s < t \leqq T$ and satisfies

$$\frac{\partial U}{\partial t}(t, s) + A(t)U(t, s) = 0, \qquad s < t \leqq T,$$

$$U(s, s) = I.$$

Throughout the sequel, unless otherwise stated, the letter C will always denote the universal constant appearing in [18] (and also in [9, Part II]) which occurs in all of the various estimates on term such as $e^{-tA(s)}$, $A(s)\, e^{-tA(s)}$, $A(t)A^{-1}(s)$, $U(t, s)$, $A(t)U(t, s)$, etc.

Now consider the nonhomogeneous Cauchy problem

$$\frac{du}{dt} + A(t)u = h(t), \qquad u(t_0) = u_0. \tag{2.8}$$

Let $C^\beta([t_0, T]; X)$ denote the space of all X-valued functions $h(t)$ which are uniformly Hölder continuous on $[t_0, T]$ with exponent β. Define

$$[h]_\beta = \sup_{t_0 \leqq t, s \leqq T} \frac{\|h(t) - h(s)\|}{|t - s|^\beta}.$$

Then $C^\beta([t_0, T]; X)$ is a Banach space with respect to the norm

$$\|h\|_{C^\beta([t_0; T]; X)} = \sup_{t_0 \leqq t \leqq T} \|h(t)\| + [h]_\beta.$$

It is well known that if $h \in C^\beta([t_0, T]; X)$ then the function

$$u(t) = U(t, t_0)u_0 + \int_{t_0}^{t} U(t, s)h(s)\, ds, \qquad t_0 \leqq t \leqq T,$$

is continuous from $[t_0, T]$ to X, continuously differentiable from $[t_0, T]$ to X and is the unique solution of (2.8) on $t_0 < t \leqq T$. Moreover, if $u_0 \in D(A)$ then $u(t)$ is continuously differentiable on $[t_0, T]$ and satisfies (2.8) on $t_0 \leqq t \leqq T$.

We now present some additional preliminary results which will be useful later on.

LEMMA 1. ([10, Lemma I.1]). *For each* $h \in C^{\beta}([t_0, T]; X)$ *define*

$$(Lh)(t) = \int_{t_0}^{t} U(t, s)h(s)\, ds, \qquad t_0 \leqq t \leqq T.$$

Then $L : C^{\beta}([t_0, T]; X) \to C([t_0; T]; X_1)$ *is a bounded mapping and*

$$\|Lh\|_{C([t_0,T];X_1)} \leqq C\|h\|_{C^{\beta}([t_0,T];X)}.$$

COROLLARY 1. *Define*

$$P(y; h) = U(t, 0)y + \int_{0}^{t} U(t, s)h(s)\, ds, \qquad 0 \leqq t \leqq T.$$

The P *is a continuous linear mapping from* $X_1 \times C^{\beta}([0, T]; X)$ *into* $C([0, T]; X_1)$.

LEMMA 2. *Let* $0 < \mu \leqq 1$ *and* $f \in C([t_0, T]; X_{\mu})$. *Define*

$$w(t) = \int_{t_0}^{t} U(t, s)f(s)\, ds, \qquad t_0 \leqq t \leqq T.$$

Then $w \in C([t_0; T]; X_1) \cap C^1([t_0, T]; X)$ *and* $w'(t) + A(t)w(t) = f(t)$, $t_0 \leqq t \leqq T$.

Proof. For $\mu = 1$ this result is due to Kato [12]. By standard arguments it follows that $w(t) \in D(A)$ and

$$A(t)w(t) = \int_{t_0}^{t} A(t)U(t, s)f(s)\, ds, \qquad t_0 \leqq t \leqq T.$$

It is easy to see that it suffices to prove that $A(t)w(t)$ is continuous from $[t_0, T]$ to X. Let $t_0 \leqq \tau < t \leqq T$, then

$$A(t)w(t) - A(\tau)w(\tau) = \int_{\tau}^{t} A(t)U(t, s)f(s)\, ds + \int_{t_0}^{\tau} [A(t)U(t, s) - A(\tau)U(\tau, s)]f(s)\, ds.$$

Let $0 < \eta < \mu$ and choose $0 < \varepsilon < \min(\eta, \alpha)$. Let $\gamma = 1 + \varepsilon$. Then by [18, (1.69)]

$$\int_{t_0}^{\tau} \|[A(t)U(t, s) - A(\tau)U(\tau, s)]f(s)\|\, ds \leqq C(t - \tau)^{\varepsilon}\|f\|_{\mu,\infty} \frac{(\tau - t_0)^{\eta-\varepsilon}}{\eta - \varepsilon},$$

where $\|f\|_{\mu,\infty} = \sup\{\|f(s)\|_{\mu} : t_0 \leqq s \leqq T\}$. Similarly

$$\int_{\tau}^{t} \|A(t)U(t, s)f(s)\|\, ds \leqq C\|f\|_{\mu,\infty} \frac{(t - \tau)^{\eta}}{\eta}.$$

Therefore $\|A(t)w(t) - A(\tau)w(\tau)\| \leqq C\|f\|_{\mu,\infty}(t - \tau)^{\varepsilon}$, and this proves continuity. □

The next result is based on a simple compactness argument and its proof will be omitted.

LEMMA 3. *Let* $\{E(t, s) : 0 \leqq s, t \leqq T\}$ *be a family of strongly continuous operators in* $\mathcal{L}(X)$. *Then for each* $x \in X$ *we have*

$$\lim_{(t,s)\to(t_0,s_0)} E(t, s)x = E(t_0, s_0)x$$

uniformly for (t_0, s_0) *in* $[0, T] \times [0, T]$ *and also uniformly for* x *in compact subsets of* X.

The last result that we need is a Gronwall lemma. Its proof can be accomplished by an argument similar to that of [22, Lemma 2.1] plus the use of Lemma 1.

LEMMA 4. *Let* $0 \leqq t_0 < t_1 \leqq T$ *and let* $R : [t_0, t_1] \to X_1$ *be a continuous function. Let* $0 < \mu \leqq 1$ *be a constant and* W *be an open subset of* X_1. *Suppose there are positive*

constants C_i, $1 \leqq i \leqq 6$ *and continuous functions*

$$H_0 : [t_0, t_1] \times [t_0, t_1] \times W \to X,$$

$$H_1 : [t_0, t_1] \times [t_0, t_1] \times W \to X_1,$$

$$f : [t_0, t_1] \times W \to X_\mu,$$

such that

$$\|H_0(t, s, y)\| \leqq C_1 \|y\|_1 + C_2, \tag{2.9}$$

$$\|H_1(t, s, y)\|_1 \leqq C_3 \|y\|_1 + C_4, \tag{2.10}$$

$$\|f(t, y)\|_\mu \leqq C_5 \|y\|_1 + C_6, \tag{2.11}$$

$$\|H_0(t, s, y) - H_0(\tau, s, y)\| \leqq |t - \tau|^\beta (C_1 \|y\|_1 + C_2) \tag{2.12}$$

for all t, τ, s *in* $[t_0, t_1]$ *and* $y \in W$. *Suppose* $u : [t_0, t_1] \to W$ *is a continuous solution of the integral equation*

$$u(t) = R(t) + \int_{t_0}^{t} U(t, s) \int_{t_0}^{s} [H_0(s, \tau, u(\tau)) + H_1(s, \tau, u(\tau))]\, d\tau\, ds$$

$$+ \int_{t_0}^{t} U(t, s) f(s, u(s))\, ds,$$

for $t_0 \leqq t \leqq t_1$. *Then there are constants* $C_i \geqq 0$ *for* $i = 7, 8, 9$ *such that if* $0 < \eta < \mu$ *and if* $\gamma > 0$ *is chosen so that* $\kappa \equiv C_7 \gamma^{-\eta} \Gamma(\eta) + C_8 \gamma^{-1} < 1$ *then*

$$\|u(t)\|_1 \leqq (1 - \kappa)^{-1} \{2 \lim_{t_0 \leqq s \leqq t} \|R(s)\|_1 + C_9\}\, e^{\gamma(t - t_0)}$$

for all $t_0 \leqq t \leqq t_1$. *If* $C_2 = C_4 = C_6 = 0$, *we can choose* $C_9 = 0$.

3. Existence, uniqueness and continuation. Our first result concerns the local problem

$$u'(t) + A(t)u(t) = \int_{t_0}^{t} [a(t, s) g_0(s, u(s)) + g_1(t, s, u(s))]\, ds + f_0(t) + f_1(t, u(t)), \qquad t \geqq t_0, \tag{3.1}$$

$$u(t_0) = u_0. \tag{3.2}$$

Given $u_0 \in W$ we shall say that $u(t)$ is a *strong solution of* (3.1), (3.2) on an interval $[t_0, t_0 + \delta]$ if $u(t) \in W$ for all $t_0 \leqq t \leqq t_0 + \delta$, $u(t)$ is continuously differentiable from $[t_0, t_0 + \delta]$ to X, $u(t_0) = u_0$ and $u(t)$ satisfies (3.1) on $[t_0, t_0 + \delta]$.

THEOREM 1. *Assume* (A1)–(A6) *hold and let* $u_0 \in W$, $t_0 \in [0, T)$ *be given. Then there exists a positive number* $\delta = \delta(t_0, u_0)$ *and a unique strong solution* $u(t)$ *of* (3.1), (3.2) *on the interval* $[t_0, t_0 + \delta]$ *which also belongs to* $C([t_0, t_0 + \delta]; X_1)$.

Proof. Given $u_0 \in W$, let $B = B(u_0; r) \subset W$ be determined by (A6). Let $0 < \delta \leqq T - t_0$ be a positive number to be specified later and put $I = [t_0, t_0 + \delta]$. For each function

$v \in C(I; B)$ we define

$$(G_0 v)(t) = \int_{t_0}^{t} a(t, \tau) g_0(\tau, v(\tau))\, d\tau, \tag{3.3}$$

$$(G_1 v)(t) = \int_{t_0}^{t} g_1(t, \tau, v(\tau))\, d\tau \tag{3.4}$$

for $t \in I$. Let

$$a_\infty = \sup_{0 \leqq s, t \leqq T} |a(t, s)|, \qquad b_{0,\infty} = \sup_{0 \leqq t \leqq T} |b_0(t)|,$$

$$b_{1,\infty} = \sup_{0 \leqq s, t \leqq T} |b_1(t, s)|, \qquad c_\infty = \sup_{0 \leqq t \leqq T} |c(t)|.$$

Then by (2.6)

$$\sup_{t \in I} \|(G_0 v)(t)\| \leqq a_\infty \delta C_{g_0}, \qquad v \in C(I; B),$$

where

$$C_{g_0} = r b_{0,\infty} + \sup_{0 \leqq t \leqq T} \|g_0(t, u_0)\|. \tag{3.5}$$

Assuming $\rho = \beta$, we have by (2.1), (3.5)

$$\frac{\|G_0 v(t) - G_0 v(s)\|}{|t - s|^\beta} \leqq (a_\infty |t - s|^{1-\beta} + |s - t_0| a_0) C_{g_0}.$$

It follows that $G_0 v \in C^\beta(I; X)$ and

$$\|G_0 v\|_{C^\beta(I;X)} \leqq [(a_0 + a_\infty)\delta + a_\infty \delta^{1-\beta}] C_{g_0}. \tag{3.6}$$

From (2.1), (2.5) we have

$$\sup_{t \in I} \|G_0 u(t) - G_0 v(t)\| \leqq a_\infty b_{0,\infty} \delta \sup_{t \in I} \|u(t) - v(t)\|_1,$$

and

$$[G_0 u - G_0 v]_\beta \leqq (a_\infty \delta^{1-\beta} + a_0 \delta) b_{0,\infty} \sup_{t \in I} \|u(t) - v(t)\|_1.$$

Therefore the mapping $v \to G_0 v$ is Lipschitz continuous from $C(I; B)$ into $C^\beta(I; X)$ and

$$\|G_0 u - G_0 v\|_{C^\beta(I;X)} \leqq [(a_0 + a_\infty)\delta + a_\infty \delta^{1-\beta}] b_{0,\infty} \|u - v\|_{C(I;X_1)}. \tag{3.7}$$

Similarly, from Property (iv), (2.6) and (3.4) we have $G_1 v \in C(I; X_1)$ and

$$\|G_1 v\|_{C(I;X_1)} \leqq \delta C_{g_1}, \tag{3.8}$$

$$\|G_1 u - G_1 v\|_{C(I;X_1)} \leqq \delta b_{1,\infty} \|u - v\|_{C(I;X_1)}, \tag{3.9}$$

where

$$C_{g_1} = r b_{1,\infty} + \sup_{0 \leqq t, s \leqq T} \|g_1(t, s, u_0)\|_1.$$

Thus mapping $v \to G_1 v$ is Lipschitz continuous from $C(I; B)$ into $C(I; X_1)$.

Now for each $v \in C(I; B)$ we define

$$(\Phi v)(t) = U(t, t_0)u_0 + \int_{t_0}^{t} U(t, s)\{G_0 v(s) + G_1 v(s) + f_0(s)\}\, ds$$

$$+ \int_{t_0}^{t} U(t, s) f_1(s, v(s))\, ds, \qquad t \in I. \tag{3.10}$$

By Lemmas 1 and 2 it is clear that Φ maps $C(I; B)$ into $C(I; X_1)$. We show that for δ sufficiently small, Φ maps $C(I; B)$ into itself and is a contraction.

Let $0 < \eta < \mu$, then by (2.7) and [9, (II.14.12), (II.14.14)], we have for each $v \in C(I; B)$,

$$\left\| \int_{t_0}^{t} U(t, s) f(s, v(s))\, ds \right\|_1 \leqq C C_{f_1} \frac{\delta^h}{\eta}, \qquad t \in I, \tag{3.11}$$

where

$$C_{f_1} = r c_\infty + \sup_{0 \leqq t \leqq T} \|f(t, u_0)\|_\mu.$$

Now let

$$\varphi(\delta) = (1 + a_0 + a_\infty)\delta + a_\infty \delta^{1-\beta} + \frac{\delta^\eta}{\eta},$$

$$w(t) = U(t, t_0)u_0 + \int_{t_0}^{t} U(t, s) f_0(s)\, ds, \qquad t \in I.$$

Then by Lemma 1, (3.6), (3.8), (3.11) we have

$$\|\Phi v(t) - u_0\|_1 \leqq \|w(t) - u_0\|_1 + C\varphi(\delta)[C_{g_0} + C_{g_1} + C_{f_1}]$$

for all $t \in I$, $v \in C(I; B)$. Since $u_0 \in D(A)$ we have $w \in C(I; X_1)$ and $w(t_0) = u_0$. So there is a $\delta_1 > 0$ such that

$$\sup_{t_0 \leqq t \leqq t_0 + \delta_1} \|\Phi v(t) - u_0\|_1 \leqq r, \qquad v \in C(I; B).$$

Furthermore, if $u, v \in C(I; B)$ then by (2.7), (3.7), (3.9)

$$\|\Phi u - \Phi v\|_{C(I; X_1)} \leqq C\varphi(\delta)(b_{0,\infty} + b_{1,\infty} + c_\infty)\|u - v\|_{C(I; X_1)}. \tag{3.12}$$

Choose $\delta_2 > 0$ so that $C\varphi(\delta_2)(b_{0,\infty} + b_{1,\infty} + c_\infty) < 1$. Then if $\delta = \min(\delta_1, \delta_2)$ it follows that Φ is a contraction mapping of $C(I; B)$ into itself. Hence there is a unique fixed point u of Φ in $C(I; B)$ and $u(t)$ is a local strong solution of (3.1), (3.2).

To prove uniqueness, suppose we have two solutions $u_1(t)$, $u_2(t)$ of (3.1), (3.2) on $I = [t_0, t_0 + \delta]$ which belong to $C(I; X_1) \cap C^1(I; X)$. Let

$$t_1 = \sup\{t \in I : u_1(s) = u_2(s) \text{ for all } t_0 \leqq s \leqq t\},$$

and suppose that $t_1 < t_0 + \delta$. Then both $u_1(t)$ and $u_2(t)$ are solutions of the

integrodifferential equation

$$u'(t)+A(t)u(t)=\int_{t_1}^{t}[a(t,s)g_0(s,u(s))+g_1(t,s,u(s))]\,ds+\tilde{f}_0(t)+f_1(t,u(t)),$$
$$t_1\leqq t\leqq t_0+\delta,$$

$$u(t_1)=u_1,$$

where $u_1=u_1(t_1)=u_2(t_1)$ and

$$\tilde{f}_0(t)=\int_{t_0}^{t_1}[a(t,s)g_0(s,u_1(s))+g_1(t,s,u_1(s))]\,ds+f_0(t).$$

Since $u\in W$, there is a closed ball $B_1=B(u_1;r_1)\subset W$ such that (2.5), (2.6), (2.7) hold. For each $v\in C([t_1,t_0+\delta];B_1)$ define

$$(\Phi_1 v)(t)=U(t,t_1)u_1+\int_{t_1}^{t}U(t,s)f_1(s,v(s))\,ds$$
$$+\int_{t_1}^{t}U(t,s)\left\{\int_{t_1}^{s}[a(s,\tau)g_0(\tau,v(\tau))+g_1(s,\tau,v(\tau))]\,d\tau+\tilde{f}_0(s)\right\}ds,$$
$$t_1\leqq t\leqq t_0+\delta.$$

Then by previous arguments there exists $\delta_1>0$ such that Φ_1 maps $C([t_1,t_1+\delta_1];B_1)$ into itself and is a contraction. But by continuity we have $u_1(t)\in B_1$ and $u_2(t)\in B_1$ for all $t_1\leqq t\leqq t_1+\delta_1$ if δ_1 is sufficiently small. Thus $u_1(t)$ and $u_2(t)$ are both fixed points of Φ_1 and we get $u_1(t)=u_2(t)$ for $t_1\leqq t\leqq t_1+\delta_1$. This contradicts the definition of t_1 and proves uniqueness. □

We now discuss noncontinuable solutions of (1.1), (1.2). It is expected that an a priori estimate in the X_1-norm should produce a global solution defined on $[0,T]$. We show that this is in fact the case under slightly stronger assumptions on g_0, g_1 and f_1:

(A6)′ Let W, μ, g_0, g_1 and f_1 be given by (2.2), (2.3), (2.4). For each closed bounded set $B\subset W$ there are positive continuous functions b_0, b_1 and c such that (2.5), (2.6), (2.7) hold for all y_1, $y_2\in B$ and $0\leqq t,s\leqq T$.

THEOREM 2. *Assume* (A1)–(A5) *and* (A6)′ *hold. Let* $\tilde{u}(t)$ *be the solution of* (1.1), (1.2) *on a maximal interval of existence* $[0,d)$ *where* $d<T$. *Then for each closed bounded set* $B\subset W$ *with nonempty interior* Int B, *there is a sequence* $\{t_n\}$ *such that* $t_n\to d^-$ *as* $n\to\infty$ *and* $\tilde{u}(t_n)\notin$ Int B *for all* $n\geqq 1$.

Proof. We argue by contradiction and suppose $\tilde{u}(t)\in$ Int B for all $0\leqq t<d$. We shall produce a positive number δ, $0<\delta\leqq T-d$, such that for any t_0 in $[0,d)$ the integrodifferential equation

$$u'(t)+A(t)u(t)=\int_{t_0}^{t}[a(t,s)g_0(s,u(s))+g_1(t,s,u(s))]\,ds+\tilde{f}_0(t)+f_1(t,u(t)),\tag{3.13}$$

$$u(t_0)=\tilde{u}_0,\tag{3.14}$$

where $\tilde{u}_0=\tilde{u}(t_0)$ and

$$\tilde{f}_0(t)=\int_{0}^{t_0}[a(t,s)g_0(s,\tilde{u}(s))+g_1(t,s,\tilde{u}(s))]\,ds+f_0(t)$$

has a solution $u(t)$ on $[t_0,t_0+\delta]$. By our uniqueness result, for t_0 sufficiently close

to d, $u(t)$ represents a continuation of $\tilde{u}(t)$ to the right of d. This contradicts the definition of d.

First we choose an arbitrary point $t_0 \in [0, d)$ and keep it fixed. Then we define

$$M_{g_0} = \sup\{\|g_0(t, y)\| : 0 \leqq t \leqq T,\ y \in B\},$$

$$M_{g_1} = \sup\{\|g_1(t, s, y)\|_1 : 0 \leqq s, t \leqq T,\ y \in B\},$$

$$M_{f_1} = \sup\{\|f(t, y)\|_\mu : 0 \leqq t \leqq T,\ y \in B\},$$

and let $I = [t_0, t_0 + \delta]$, where δ is some positive number less than or equal to $T - d$ which is to be determined later (depending only on B). For each $v \in C(I; B)$ define $G_0 v$, $G_1 v$ by (3.3), (3.4) and Φv by (3.10) with u_0, f_0 replaced by $\tilde{u}_0$, $\tilde{f}_0$, respectively. Then from (3.12) it is clear that we can choose $\delta > 0$, depending only on B, such that Φ is a contraction mapping of $C(I; B)$ into $C(I; X_1)$.

For $v \in C(I; B)$ we have

$$\begin{aligned}(\Phi v)(t) - \tilde{u}_0 &= U(t, t_0)\tilde{u}_0 - \tilde{u}_0 + \int_{t_0}^{t} U(t, s) f_0(s)\, ds \\ &\quad + \int_{t_0}^{t} U(t, s)\left\{\int_0^{t_0} a(s, \tau) g_0(\tau, \tilde{u}(\tau))\, d\tau\right\} ds \\ &\quad + \int_{t_0}^{t} U(t, s)\left\{\int_0^{t_0} g_1(s, \tau, \tilde{u}(\tau))\, d\tau\right\} ds \\ &\quad + \int_{t_0}^{t} U(t, s) G_0 v(s)\, ds + \int_{t_0}^{t} U(t, s) G_1 v(s)\, ds \\ &\quad + \int_{t_0}^{t} U(t, s) f_1(s, v(s))\, ds \\ &= \sum_{i=1}^{7} J_i.\end{aligned}$$

Consider J_5, J_6 and J_7. From (3.6), (3.8), (3.11) we have

$$\|J_5\|_1 \leqq C[(a_0 + a_\infty)\delta + a_\infty \delta^{1-\beta}] M_{g_0},$$

$$\|J_6\|_1 \leqq C\delta^2 M_{g_1}, \qquad \|J_7\|_1 \leqq C M_{f_1} \frac{\delta^\eta}{\eta}.$$

Similarly for J_4 we have

$$\|J_4\|_1 \leqq CTM_{g_1}^{\delta}.$$

Consider J_2. Following Sobolevsky [18, p. 32] we have

$$\|A(t)J_2\| \leqq C[f_0]_\beta \frac{(t - t_0)^\beta}{\beta} + C\|f_0(t)\| \frac{(t - t_0)^\beta}{\beta} + E(t, t - t_0) f_0(t),$$

where $E(t, s) = I - e^{-sA(t)}$. The function $f_0(t)$ has compact range in X and $E(t, s)$ is strongly continuous for $0 \leqq s, t \leqq T$ (see [9, Lemma II.4.4]). So by Lemma 3

$$\lim_{t \to t_0} E(t, t - t_0) f_0(t) = 0,$$

uniformly in t_0. Thus $\|J_2\|_1 \to 0$ as $t \to t_0$, uniformly in t_0 and a similar argument works for J_3. For the remaining term J_1 we use the integral equation to write

$$\begin{aligned} J_1 &= [U(t,0) - U(t_0,0)]u_0 \\ &\quad + \int_0^{t_0} [U(t,s) - U(t_0,s)] \left\{ \int_0^s a(s,\tau) g_0(\tau, \tilde{u}(\tau))\, d\tau \right\} ds \\ &\quad + \int_0^{t_0} [U(t,s) - U(t_0,s)] \left\{ \int_0^s g_1(s,\tau,\tilde{u}(\tau))\, d\tau \right\} ds \\ &\quad + \int_0^{t_0} [U(t,s) - U(t_0,s)] f_1(s, \tilde{u}(s))\, ds \\ &\quad + \int_0^{t_0} [U(t,s) - U(t_0,s)] f_0(s)\, ds, \end{aligned}$$

where $u_0 = \tilde{u}(0)$. By Property (iv) and Lemmas 1, 2 and 3 each term above goes to zero in the X_1-norm as $t \to t_0$ uniformly in t_0.

Therefore given $\varepsilon > 0$ we can choose $\delta > 0$, independent of t_0, such that

$$\|(\Phi v)(t) - \tilde{u}_0\|_1 < \varepsilon, \qquad t_0 \leqq t \leqq t_0 + \delta, \qquad v \in C(I; B).$$

Since $\tilde{u}_0 \in \operatorname{Int} B$, we can choose $\varepsilon > 0$ such that the ball $B_\varepsilon = \{y \in X_1 : \|y - \tilde{u}_0\|_1 < \varepsilon\}$ is contained in B. Then $\Phi : C(I; B) \to C(I; B_\varepsilon) \subset C(I; B)$ and by the contraction mapping principle there is a unique solution $u(t)$ of (3.13), (3.14) on $[t_0, t_0 + \delta]$. □

COROLLARY 2. *Assume* (A1)–(A5) *hold and let* (A6)′ *hold with* $W = X_1$. *Suppose there are positive constants* C_i, $1 \leqq i \leqq 6$, *such that*

$$\|g_0(t,y)\| \leqq C_1\|y\|_1 + C_2, \tag{3.15}$$

$$\|g_1(t,s,y)\|_1 \leqq C_3\|y\|_1 + C_4,$$

$$\|f_1(t,y)\|_\mu \leqq C_5\|y\|_1 + C_6 \tag{3.16}$$

for all $0 \leqq s, t \leqq T$ *and* $y \in X_1$. *Then for each* $u_0 \in X_1$ *there exists a unique global solution* $u(t)$ *of* (1.1), (1.2) *on* $[0, T]$.

Proof. By Theorem 2, it suffices to show that each solution of (1.1), (1.2) is a priori bounded in the X_1-norm. From Theorem 1 with $t_0 = 0$ we have

$$\begin{aligned} u(t) &= R(t) + \int_0^t U(t,s) \int_0^s [a(s,\tau) g_0(\tau, u(\tau)) + g_1(s,\tau,u(\tau))]\, d\tau\, ds \\ &\quad + \int_0^t U(t,s) f_1(s, u(s))\, ds, \end{aligned}$$

where

$$R(t) = U(t,0)u_0 + \int_0^t U(t,s) f_0(s)\, ds.$$

Let $H_0(t,s,y) = a(t,s)g_0(s,y)$ and $H_1(t,s,y) = g_1(t,s,y)$. Then H_0 and H_1 satisfy (2.9), (2.10), (2.11). So by Lemma 4 we obtain the desired bound. □

4. Continuous dependence and differentiability of solutions. We begin our discussion of continuous dependence by considering a sequence of integrodifferential

equations

$$u'(t)+A(t)u(t)=\int_0^t [a_n(t,s)g_{0,n}(s,u(s))+g_{1,n}(t,s,u(s))]\,ds$$

$$+f_{0,n}(t)+f_{1,n}(t,u(t)), \qquad t\geqq 0, \tag{4.1}$$

$$u(0)=u_{0,n}. \tag{4.2}$$

We assume that (A1)–(A6) hold and let $u(t)$ denote the solution of (1.1), (1.2). We suppose $u(t)$ is defined on some closed interval $[0, d]$ $(d \leqq T)$ and we seek conditions on (4.1), (4.2) which guarantee that for n sufficiently large, its solutions will also exist on $[0, d]$ and converge to $u(t)$ as $n\to\infty$. We shall assume that each $a_n(t, s)$ satisfies (A5) with constants a_0 and ρ independent of n. Similarly, we assume that each $g_{0,n}$, $g_{1,n}$ and $f_{1,n}$ satisfies (A6) with the constant r and the functions $b_0(t)$, $b_1(t, s)$, $c(t)$ independent of n. We assume that $f_{0,n}\in C^{\beta}([0, T]; X)$ for all $n\geqq 1$ and we shall use the notation

$$[a(\cdot,\tau)]_{\rho}=\sup_{0\leqq t,s\leqq T}\frac{|a(t,\tau)-a(s,\tau)|}{|t-s|^{\rho}}.$$

We note that, by Theorem 1, for each $u_{0,n}\in W$, there is a unique solution $u_n(t)$ of (4.1), (4.2).

THEOREM 3. *In addition to the above assumptions, suppose that*:

(i) $\lim_{n\to\infty}\sup_{0\leqq t,\tau\leqq T}|a_n(t,r)-a(t,\tau)|=0;$

(ii) $\lim_{n\to\infty}\sup_{0\leqq\tau\leqq T}[a_n(\cdot,\tau)-a(\cdot,\tau)]_{\rho}=0;$

(iii) $\lim_{n\to\infty}\|g_{0,n}(t,y)-g_0(t,t)\|=0$, *uniformly on bounded subsets of* $[0, T]\times W$;

(iv) $\lim_{n\to\infty}\|g_{1,n}(t,s,y)-g_1(t,s,y)\|_1=0$, *uniformly on bounded subsets of* $[0, T]\times[0, T]\times W$;

(v) $\lim_{n\to\infty}\|f_{0,n}-f_0\|_{C^{\beta}([0,T];X)}=0;$

(vi) $\lim_{n\to\infty}\|f_{1,n}(t,y)-f_1(t,y)\|_{\mu}=0$, *uniformly on bounded subsets of* $[0, T]\times W$;

(vii) $\lim_{n\to\infty}\|u_{0,n}-u_0\|_1=0.$

Then there is an integer $N\geqq 1$ *such that for each* $n\geqq N$ *the solution* $u_n(t)$ *of* (4.1), (4.2) *is defined on* $[0, d]$ *and*

$$\lim_{n\to\infty}\sup_{0\leqq t\leqq d}\|u_n(t)-u(t)\|_1=0,$$

$$\lim_{n\to\infty}\sup_{0\leqq t\leqq d}\|u_n'(t)-u'(t)\|=0.$$

Since the argument used to prove Theorem 3 is fairly standard, we omit the proof.

We now turn to the question of differentiability of solutions of the Cauchy problem

$$u'(t)+A(t)u(t)=\int_0^t [a(t,s)g_0(s,u(s))+g_1(t,s,u(s))]\,ds+f_0(t), \tag{4.3}$$

$$u(0)=u_0, \tag{4.4}$$

We always assume (A1)–(A6) hold and we let $u(t)$ be the solution of (4.3), (4.4) on an interval $[0, d]$ ($d \leqq T$), which belongs to $C^1([0, d]; X) \cap C([0, d]; X_1)$. Define

$$h_0(t)=\int_0^t a(t,s)g_0(s,u(s))\,ds+f_0(t), \qquad 0 \leqq t \leqq d,$$

$$h_1(t)=\int_0^t g_1(t,s,u(s))\,ds, \qquad 0 \leqq t \leqq d.$$

In § 3 we showed that $h_0 \in C^\beta([0, d]; X)$, $h_1 \in C([0, d]; X_1)$ and

$$u(t)=U(t,0)u_0+\int_0^t U(t,s)[h_0(s)+h_1(s)]\,ds, \qquad 0 \leqq t \leqq d. \tag{4.5}$$

We make the following additional assumptions:

(C_1) The mapping $t \to A(t)$ is strongly continuously differentiable from $[0, T]$ to $\mathscr{L}(X_1, X)$ and we let

$$\frac{d}{dt}A(t)y \equiv A^{(1)}(t)y, \qquad y \in X_1.$$

We assume that $A^{(1)}(t)A^{-1}(0)$ is uniformly bounded for $0 \leqq t \leqq T$ and

$$\|[A^{(1)}(t)-A^{(1)}(s)]A^{-1}(0)\| \leqq C|t-s|^\alpha, \qquad 0 \leqq s, t \leqq T.$$

(f_0) The function $f_0(t)$ has a derivative $f_0'(t)$ which is uniformly Hölder continuous on $[0, T]$ with exponent β.

(a_1) The function $a(t, s)$ is uniformly Hölder continuous along the diagonal of $[0, T]\times[0, T]$ with exponent ρ:

$$|a(t,t)-a(s,s)| \leqq a_0|t-s|^\rho, \qquad 0 \leqq s, t \leqq T.$$

Furthermore, the partial derivative $(\partial a/\partial t)(t, s)$ is continuous on $[0, T]\times[0, T]$ and uniformly Hölder continuous on $[0, T]$ in the first place with exponent ρ:

$$\left|\frac{\partial a}{\partial t}(t,s)-\frac{\partial a}{\partial t}(\tau,s)\right| \leqq a_0|t-\tau|^\rho, \qquad 0 \leqq t, \tau, s \leqq T.$$

(g_0) For each compact set $K \subset W$ there are positive constants $\mathscr{J}_0$, γ, $\mu \leqq 1$ such that

$$\|g_0(t,y_1)-g_0(s,y_2)\| \leqq \mathscr{J}_0[|t-s|^\mu+\|y_1-y_2\|_1^\gamma]$$

for all $y_1, y_2 \in K$ and $0 \leqq s, t \leqq T$.

(g_1) The function $(\partial/\partial t)\,g_1(t, s, y):[0, T]\times[0, T]\times W \to X_1$ is continuous.

LEMMA 5. *Let $0<\omega<\frac{1}{2}\min(\alpha, \beta)$. Then for each $\varepsilon>0$, $u(t)$ is uniformly Hölder continuous in the X_1-norm on $[\varepsilon, d]$ with exponent ω. That is, there exists a constant $C(\varepsilon)>0$ such that*

$$\|u(t)-u(s)\|_1 \leqq C(\varepsilon)|t-s|^\omega, \qquad \varepsilon \leqq t, s \leqq d. \tag{4.6}$$

Using Lemma 5 and the well-known differentiability result [9, Thm. II.8.1] we have the following theorem.

THEOREM 4. *Let the assumption* (A1)–(A6) *hold and let* $u(t)$ *be given by* (4.5). *Suppose that* (C_1), (f_0), (a_1), (g_0) *and* (g_1) *are satisfied. Then* $u(t)$ *is twice strongly continuously differentiable in the* X*-norm on the interval* $(0, d]$ *and satisfies the integrodifferential equation*

$$\begin{aligned} &u''(t)+A(t)u'(t)+A^{(1)}(t)u(t) \\ &\quad=\int_0^t\left[\frac{\partial a}{\partial t}(t,s)g_0(s,u(s))+\frac{\partial}{\partial t}g_1(t,s,u(s))\right]ds \\ &\qquad+a(t,t)g_0(t,u(t))+g_1(t,t,u(t))+f_0'(t), \qquad 0<t\leqq d. \end{aligned} \tag{4.7}$$

If, in addition, $u'(0)\in D(A)$ then $u(t)$ belongs to $C^2([0,d];X)\cap C^1([0,d];X_1)$ and satisfies (4.7) on $[0,d]$.

5. Generalizations. In this section we consider the abstract Cauchy problem (1.1), (1.2) for more general nonlinear operators $g_0(t,u)$. We assume that

$$A^{-1}(0) \quad \text{is completely continuous on } X. \tag{5.1}$$

It follows (see [9, pp. 169–170]) that $A^{-\mu}(t)$ is completely continuous on X for all $0<\mu\leqq 1$, $0\leqq t\leqq T$. We denote by $X_\mu(t)$ the domain $D(A^\mu(t))$ with norm $\|x\|_\mu = \|A^\mu(t)x\|$. Then the embeddings $X_1\to X_\mu(t)\to X$ are compact for all $0<\mu<1$. We assume there is a Banach space E and a number $\omega\in(0,1)$ such that

$$X_\omega(t)\to E \quad \text{is a continuous embedding} \tag{5.2}$$

for all $0\leqq t\leqq T$ with embedding contant independent of t.

Let W be a nonempty open subset of X_1 and assume there is a continuous function

$$q:[0,T]\times E\times W\to X$$

which satisfies

(B1) For each $y\in W$ there are a constant $r>0$ and a positive continuous functional b_0 defined on $[0,T]\times E$, which is bounded on bounded subsets of $[0,T]\times E$, such that the closed ball $B(y;r)=\{z\in X_1:\|y-z\|_1\leqq r\}$ belongs to W and

$$\|q(t,x,y_1)-q(t,x,y_2)\|\leqq b_0(t,x)\,\|y_1-y_2\|_1$$

for all $0\leqq t\leqq T$, $x\in E$ and $y_1, y_2\in B(y;r)$.

(B2) For each compact set $K\subset[0,T]\times W$ the mapping $x\to q(t,x,y)$ is continuous from E to X uniformly for (t,y) in K.

(B3) For each $y\in W$ the mapping $(t,x)\to q(t,x,y)$ is bounded on bounded subsets of $[0,T]\times E$.

We now define

$$g_0(t,y)=q(t,y,y) \quad \text{for } 0\leqq t\leqq T, \quad y\in W. \tag{5.3}$$

Our first result concerns local existence for solutions of the abstract Cauchy problem (3.1), (3.2). The main ideas for the proof are taken from Sobolevsky [18] and Fitzgibbon [8]. Since the proof is essentially the same as for Theorem 6, the details will be omitted.

THEOREM 5. *Assume* (A1)–(A5), (B1), (B2), (B3) *hold and let* $g_0(t,y)$ *be defined by* (5.3). *Let* f_1, g_1 *be defined as in* (A6). *Then given* $u_0\in W$ *and* $t_0\in[0,T)$ *there is a* $\delta=\delta(u_0,t_0)>0$ *and a function* $u(t)$ *belonging to* $C([t_0,t_0+\delta];X_1)\cap C^1([t_0,t_0+\delta];X)$ *and satisfying* (3.1), (3.2) *on* $[t_0,t_0+\delta]$.

We now consider global solutions of (1.1), (1.2) in the present situation, where g_0 is defined by (5.3). First, we strengthen our assumptions in the nonlinear operator q.

(B1)′ q is continuous from $[0, T]\times E\times X_1$ to X and for each bounded set $B\subset X_1$ there is a positive continuous functional b_0 defined on $[0, T]\times E$, which is bounded on bounded subsets of $[0, T]\times E$, such that

$$\|q(t, x, y_1)-q(t, x, y_2)\|\leqq b_0(t, x)\|y_2-y_1\|_1$$

for all $0\leqq t\leqq T$, $x\in E$ and $y_1, y_2\in B$.

THEOREM 6. *Assume* (A1)–(A5), (B1)′, (B2) *with* $W=X_1$, *are true. Let* f_1 *and* g_1 *satisfy the hypotheses of Corollary* 2. *Assume also that there are constants* $C_0>0$, $C_1>0$ *such that*

$$\|q(t, x, y)\|\leqq C_1(\|x\|_E+\|y\|_1)+C_0 \tag{5.4}$$

for all (t, x, y) *in* $[0, T]\times E\times X$. *Let* g_0 *be defined by* (5.3). *Then for each* $u_0\in X_1$ *there is a global solution* $u(t)$ *of* (1.1), (1.2) *on* $[0, T]$.

Proof. Let $u_0\in X_1$. Then by Theorem 5 there exists a local solution $u(t)$ of (1.1), (1.2). Let t_0 be a point where $u(t_0)$ is defined and let $I=[t_0, t_0+\delta]$. Let $0<\eta<1-\omega$, where ω is determined from (5.2). In what follows, we use the notation $C_i(i\geqq 0)$ to denote universal positive constants which are independent of t_0. For fixed $M>0$ let S be the class of all functions $v: I\to X$ which satisfy

$$\|v(t)-v(\tau)\|\leqq M|t-\tau|^{\eta},\qquad t_0\leqq t, \tau\leqq t_0+\delta_0,\qquad v(t_0)=A^{\omega}(t_0)u(t_0).$$

By (5.3) and (5.4) there is a positive constant C_2 such that

$$\|g_0(t, y)\|\leqq C_2\|y\|_1+C_0,\qquad (t, y)\in[0, T]\times X_1. \tag{5.5}$$

So by Lemma 4 there is $C_3>0$, $C_4>0$ such that

$$\|u(t)\|_1\leqq C_3 \tag{5.6}$$

and

$$\|A(t)u(t)\|\leqq C_4 \tag{5.7}$$

as long as $u(t)$ exists. By interpolation we have

$$\|A^{\omega}(t_0)u(t_0)\|\leqq C_5,$$

so that

$$\|v(t)\|\leqq C_6,\qquad \|A^{-\omega}(t_0)v(t)\|\leqq C_7 \tag{5.8}$$

for all $t\in I$, $v\in S$.

Now consider the integrodifferential equation

$$\begin{aligned} w'(t)+A(t)w(t)=&\int_{t_0}^{t}[a(t, s)g_0(s, w(s))+g_1(t, s, w(s))]\,ds\\ &+h_0(t)+h_1(t)+f_1(t, w(t)),\qquad t\geqq t_0, \end{aligned} \tag{5.9}$$

$$w(t_0)=u(t_0), \tag{5.10}$$

where

$$h_0(t)=\int_0^{t_0}a(t, s)g_0(s, u(s))\,ds+f_0(t), \tag{5.11}$$

$$h_1(t) = \int_0^{t_0} g_1(t, s, u(s))\, ds. \tag{5.12}$$

We shall show that there is a positive constant $\delta > 0$, independent of t_0, such that if $\delta_0 = \min(\delta, T - t_0)$ then (5.9), (5.10) has a solution on $[t_0, t_0 + \delta_0]$. By stepping off intervals of length δ_0, we construct a global solution of (1.1), (1.2).

For each $v \in S$, define

$$g_v(t, y) = q(t, A^{-\omega}(t_0)v(t), y), \qquad (t, y) \in I \times X_1.$$

Let $B \subset X_1$ be any bounded set. Then by (B1)′ we have

$$\|g_v(t, y_1) - g_v(t, y_2)\| \leqq b_0(t, A^{-\omega}(t_0)v(t))\|y_1 - y_2\|_1$$

for all $t \in I$, $v \in S$ and $y_1, y_2 \in B$. By (5.8) there is a constant $b_\infty > 0$ (independent of v and t_0) such that

$$\sup_{t \in I} b_0(t, A^{-\omega}(t_0)v(t)) \leqq b_\infty < \infty.$$

Therefore,

$$\|g_v(t, y_1) - g_v(t, y_2)\| \leqq b_\infty \|y_1 - y_2\|_1$$

for all $t \in I$, $y_1, y_2 \in B$ and $v \in S$. Similarly by (5.4) there is $C_8 > 0$ such that

$$\|g_v(t, y)\| \leqq C_1 \|y\|_1 + C_8, \qquad (t, y) \in I \times X_1. \tag{5.13}$$

Applying Corollary 2, it follows that for each $v \in S$ there exists a unique global solution $w(t)$ of the integrodifferential equation

$$w'(t) + A(t)w(t) = \int_{t_0}^{t} [a(t, s)g_v(s, w(s)) + g_1(t, s, w(s))]\, ds + h_0(t) + h_1(t) + f_1(t, w(t)), \qquad t \in I,$$

$$w(t_0) = u(t_0).$$

By Lemma 4, there is $C_9 \geqq 0$ such that

$$\|w(t)\|_1 \leqq (1 - \kappa)^{-1} \{2 \sup_{s \in I} \|R_0(s)\|_1 + C_9\} e^{\gamma T}, \qquad t \in I, \tag{5.14}$$

where

$$R_0(t) = U(t, t_0)u(t_0) + \int_{t_0}^{t} U(t, s)[h_0(s) + h_1(s)]\, ds.$$

Now from (3.15), (5.5), (5.6), (5.11), (5.12), (5.14) and Lemma 1 we have

$$\|w(t)\|_1 \leqq C_{10}, \qquad t \in I. \tag{5.15}$$

Hence from (3.16), (5.13), (5.15) it follows that

$$\|g_v(t, w(t))\| \leqq C_{11}, \qquad t \in I \tag{5.16}$$

and

$$\|f_1(t, w(t))\|_\mu \leqq C_{12}, \qquad t \in I, \tag{5.17}$$

with C_{11}, C_{12} independent of both v and t_0.

Define a mapping θ on S by

$$(\theta v)(t) = A^{\omega}(t_0)w(t), \qquad t \in I,$$

where v and w are related by means of the integral equation

$$w(t) = U(t, t_0)u(t_0) + \int_{t_0}^{t} U(t, s)[\Psi_v(w)(s) + f_1(s, w(s))]\, ds \tag{5.18}$$

with

$$\Psi_v(w)(t) = \int_{t_0}^{t} [a(t, s)g_v(s, w(s)) + g_1(t, s, w(s))]\, ds + h_0(t) + h_1(t)$$

for $t \in I$. From (3.15), (5.6), (5.16) there is $C_{13} > 0$ such that

$$\|\Psi_v(w)(t)\| \leqq C_{13}, \qquad t \in I, \quad v \in S. \tag{5.19}$$

We wish to show that $\theta v \in S$ if δ is sufficiently small. It is clear that $\theta v \in C(I; X)$. So let $t_0 \leqq t \leqq t + \Delta t \leqq t_0 + \delta$ and set $A_0 \equiv A(t_0)$. Then from (5.18)

$$\begin{aligned}\theta v(t+\Delta t) - \theta v(t) &= A_0^{\omega}[U(t+\Delta t, t_0) - U(t, t_0)]u(t_0) \\ &\quad + A_0^{\omega}\left\{\int_{t_0}^{t+\Delta t} U(t+\Delta t, s)[\Psi_v(w)(s) + f_1(s, w(s))]\, ds \right. \\ &\quad \left. - \int_{t_0}^{t} U(t, s)[\Psi_v(w)(s) + f_1(s, w(s))]\, ds\right\} \\ &= J_1 + J_2.\end{aligned}$$

By [9, Lemma II.14.4], (5.17), (5.19)

$$\|J_1\| \leqq C(\Delta t)^{1-\omega} C_4$$

and

$$\|J_2\| \leqq C(\Delta t)^{1-\omega}(|\log \Delta t| + 1)(C_{12} + C_{13}).$$

Choose $0 < \varepsilon_0 < 1 - \omega - \eta$ and define δ to be the minimum of 1 and the expression

$$\left\{\frac{M}{C[C_4 + (C_{12} + C_{13}) \sup_{0<s\leq 1} s^{1-\omega-\eta-\varepsilon_0}(|\log s| + 1)]}\right\}^{1/\varepsilon_0}.$$

Then δ does not depend on t_0 and

$$\|\theta v(t+\Delta t) - \theta v(t)\| \leqq M(\Delta t)^{\eta}, \qquad t \in I, \quad v \in S.$$

Since $\theta v(t_0) = A_0^{\omega} u(t_0)$, it follows that $\theta v \in S$ for all $v \in S$.

We consider S as a closed, bounded, convex subset of the Banach space $C(I; X)$. Then θ maps S into itself and we next show that θ is continuous. Choose a sequence $\{v_n\}$ in S such that $v_n \to v$ in $C(I; X)$ and let

$$\theta v_n(t) = A_0^{\omega} w_n(t).$$

Then from (5.18)

$$\begin{aligned}w_n(t) - w(t) &= \int_{t_0}^{t} U(t, s)[\Psi_{v_n}(w_n)(s) - \Psi_v(w)(s)]\, ds \\ &\quad + \int_{t_0}^{t} U(t, s)[f_1(s, w_n(s)) - f_1(s, w(s))]\, ds.\end{aligned}$$

Applying Lemma 4 (with $C_2 = C_4 = C_6 = 0$) we obtain constants κ and γ, independent of n, such that

$$\|w_n(t) - w(t)\|_1 \leqq (1-\kappa)^{-1}\{2 \sup_{t_0 \leqq \tau \leqq t} \|R_n(s)\|_1\} e^{\gamma(t-t_0)},$$

where

$$R_n(t) = \int_{t_0}^{t} U(t, s) \int_{t_0}^{s} a(s, \tau)[g_{v_n}(\tau, w(\tau)) - g_v(\tau, w(\tau))]\, d\tau\, ds.$$

We now show that $R_n \to 0$ in $C(I; X_1)$ as $n \to \infty$. By Lemma 1, it suffices to show that

$$\lim_{n\to\infty} \int_{t_0}^{t} a(t, \tau)[g_{v_n}(\tau, w(\tau)) - g_v(\tau, \omega(\tau))]\, d\tau = 0$$

in the space $C^{\beta}(I; X)$. But this follows easily from (B2). Hence $w_n \to w$ in $C(I; X_1)$ so that $\theta v_n \to \theta v$ in $C(I; X)$. This proves continuity if θ. Standard arguments may be used to show that θ is also compact. Thus by the Schauder fixed point theorem there is an element $v \in S$ such that $\theta v = v$. Consequently $v(t) = A^{\omega}(t_0)w(t)$ on I and $w(t)$ satisfies

$$w'(t) + A(t)w(t) = \int_{t_0}^{t} [a(t, s)q(s, w(s), w(s)) + g_1(t, s, w(s))]\, ds + h_0(t) + h_1(t) + f_1(t, w(t)), \qquad t \in I,$$

with $w(t_0) = u(t_0)$. If we set $u(t) \equiv w(t)$, we obtain the desired solution of (3.1), (3.2). □

6. Applications. We first give a very brief description as to how equations of the type (1.1), (1.2) can arise in applications to problems in heat flow. A thorough account of these types of problems is given in [15]. Consider a rigid heat conductor in which heat flows in only one direction. Let $u(t, x)$, $\varepsilon(t, x)$, $q(t, x)$ and $h(t, x)$ denote the temperature, internal energy, heat flux and heat supply, respectively, where t denotes time and x denotes the position in the body. The energy balance equation is

$$\varepsilon_t(t, x) = -q_x(t, x) + h(t, x). \tag{6.1}$$

The classical linear theory for heat flow in a homogeneous isotropic material is obtained from (6.1) by assuming that the internal energy depends linearly on the temperature

$$\varepsilon(t, x) = \varepsilon_0 + b_0 u(t, x), \qquad b_0 > 0, \tag{6.2}$$

and the heat flux is related to the temperature by Fourier's law:

$$q(t, x) = -c_0 u_x(t, x), \qquad c_0 > 0. \tag{6.3}$$

However, in materials of fading memory type (Coleman and Mizel [6]) the classical theory of heat flow is inadequate. In these types of materials assumptions (6.2), (6.3) are replaced by more general relations which assume that the internal energy and heat flux are *functionals* (rather than functions) of the temperature and the gradient of the temperature, respectively. In the linear theory (Nunziato [16]) the functionals $\varepsilon(t, x)$ and $q(t, x)$ are taken, respectively, as

$$\varepsilon(t, x) = \varepsilon_0 + b_0 u(t, x) + \int_0^t b(t-s)u(s, x)\, ds, \qquad t \geqq 0, \tag{6.4}$$

$$q(t, x) = -c_0 u_x(t, x) + \int_0^t a(t-s)u_x(s, x)\, ds, \qquad t \geqq 0, \tag{6.5}$$

where we are assuming, without loss of generality, that the history of the temperature is prescribed as zero for $t \leqq 0$. The functions a, b are usually assumed to be decaying exponentials with positive coefficients. The energy balance (6.1) applied to (6.4), (6.5) leads to the integrodifferential equation

$$b_0 u_t(t,x) - c_0 u_{xx}(t,x) = \int_0^t [a(t-s)u_{xx}(s,x) - b'(t-s)u(s,x)]\,ds + h(t,x) - b(0)u(t,x). \tag{6.6}$$

EXAMPLE 1. We consider a generalized semilinear version of (6.6). Let $R = (-\infty, +\infty)$ and consider the following pure initial-value problem

$$u_t - (k(t,x)u_x)_x = \int_0^t [a(t,s)(\sigma(\rho(s,x)u_x(s,x))_x + \gamma(t,s)u(s,x)]\,ds + h(t,x) + f(u(t,x)), \qquad t \geqq 0, \quad x \in R, \tag{6.7}$$

$$u(0,x) = u_0(x), \qquad x \in R. \tag{6.8}$$

We assume that all of the functions k, a, σ, ρ, γ, h and f are real-valued. In what follows, the letter C will be used to denote a positive constant depending only on T. We assume that k is a positive continuous function having a continuous first partial derivative k_x on $0 \leqq t < \infty$, $x \in R$ and satisfying

(i) $0 < k_0 \leqq k(t,x) < k_0'$,
(ii) $|k_x(t,x)| \leqq k_1$,
(iii) $|k(t,x) - k(\tau,x)| \leqq C|t-\tau|^\alpha$
(iv) $|k_x(t,x) - k_x(\tau,x)| \leqq C|t-\tau|^\alpha$

for all $0 \leqq t, \tau \leqq T$, $x \in R$, where k_0, k_0', k_1 are constants depending only on T.

We assume that ρ is continuous with a continuous first partial derivative ρ_x on $0 \leqq t < \infty$, $x \in R$ and

$$|\rho(t,x)| \leqq C, \qquad |\rho_x(t,x)| \leqq C, \qquad 0 \leqq t \leqq T, \quad x \in R. \tag{6.10}$$

We assume that the nonlinear functions σ, f satisfy (see [7, Example 2]), $f(0) = 0$ and

$$f, \sigma \in C^1(-\infty, +\infty), \quad |f'(s)| \leqq M, \quad |\sigma'(s)| \leqq M, \quad s \in R \tag{6.11}$$

and for each $r > 0$ there is a constant $c_r > 0$ such that

$$|\sigma'(t) - \sigma'(s)| \leqq c_r|t-s|, \qquad |t| \leqq r, \quad |s| \leqq r,$$
$$|f'(t) - f'(s)| \leqq c_r|t-s|, \qquad |t| \leqq r, \quad |s| \leqq r. \tag{6.12}$$

We assume that a and γ are continuous on $[0,\infty) \times [0,\infty)$ and that a satisfies hypothesis (A5) for each $T > 0$.

Let $c_0 > 0$ be a fixed constant. The change of variables $u \to e^{-c_0 t}u$ converts (6.7) into an equivalent problem

$$u_t - (k(t,x)u_x)_x + c_0 u = \int_0^t [a(t,s)(\sigma(\rho(s,x)u_x(s,x)))_x + \gamma(t,s)u(s,x)]\,ds + h(t,x) + f(t,u(t,x)), \qquad t \geqq 0, \quad x \in R, \tag{6.13}$$

with $a(t,s)$, $\rho(t,s)$ $\gamma(t,s)$, $h(t,x)$ in (6.7) replaced by $e^{-c_0 t}a(t,s)$, $e^{c_0 t}\rho(t,x)$, $e^{-c_0(t-s)}\gamma(t,s)$, $e^{-c_0 t}h(t,x)$, respectively, and $f(t,y) = e^{-c_0 t}f(e^{c_0 t}y)$. It is clear that the

basic assumptions on a, ρ, γ, h are unchanged. From now on we discuss the pure initial-value problem (6.13), (6.8).

Let $X = L^2(-\infty, +\infty)$, the complex Hilbert space with inner product defined by

$$(u, v) = \int_{-\infty}^{+\infty} u(x)\,\overline{v(x)}\,dx$$

and corresponding norm $|u| = \sqrt{(u, u)}$. Let $H^2(-\infty, +\infty)$ denote the usual (Hilbert) Sobolev space space with norm defined by

$$\|u\|^2 = |u|^2 + |Du|^2 + |D^2u|^2,$$

where D denotes differentiation with respect to x. For each $t \geqq 0$, we define a linear operator $A(t)$ in X by

$$A(t)u(x) = -\frac{\partial}{\partial x}(k(t, x)Du(x)) + c_0u(x) \quad \text{a.e. } x \in R,$$

where

$$u \in D(A) = D(A(t)) \equiv H^2(-\infty, +\infty).$$

Define operators g_0, f_1 on $[0, \infty) \times D(A)$ by

$$g_0(t, u)(x) = \frac{\partial}{\partial x}\sigma(\rho(t, x)\ \mathrm{Re}\ Du(x)) \quad \text{a.e. } x \in R,$$

$$f_1(t, u)(x) = e^{-c_0t}f(e^{c_0t}\ \mathrm{Re}\ u(x)) \quad \text{a.e. } x \in R.$$

We define g_1 on $[0, \infty) \times [0, \infty) \times D(A)$ by

$$g_1(t, s, u)(x) = \gamma(t, s)\ \mathrm{Re}\ u(x), \qquad x \in R. \tag{6.14}$$

Finally, we define $f_0(t)(x) = h(t, x)$ and assume that f_0 satisfies hypothesis (A4) for each $T > 0$. We now consider the Cauchy problem (1.1), (1.2) as an integrodifferential equation in $L^2(-\infty, \infty)$.

It is standard that $\{A(t): 0 \leqq t < \infty\}$ satisfies hypotheses (A1), (A2), (A3) for every $t > 0$. Consider the nonlinear operator g_0. Using (6.10), (6.11) we see that

$$\int_{-\infty}^{+\infty} |g_0(t, u)|^2\,dx \leqq C(|Du|^2 + |D^2u|^2), \qquad 0 \leqq t \leqq T.$$

So g_0 maps $[0, \infty) \times D(A)$ into X and has sublinear growth. By the Sobolev embedding theorem there exists a continuous injection $H^2(-\infty, +\infty) \to C_b^1(-\infty, +\infty)$, the Banach space of bounded continuous functions on $(-\infty, +\infty)$ having a bounded continuous first derivative. So if $B \subset D(A)$ is a bounded set then there is a positive number r such that

$$\|u\|_{C_b^1(-\infty, +\infty)} \leqq r \quad \text{for all } u \in B.$$

From (6.12) it then follows that

$$\int_{-\infty}^{+\infty} |g_0(t, u) - g_0(t, v)|^2\,dx \leqq C\|u - v\|^2, \qquad u, v \in B, \quad 0 \leqq t \leqq T,$$

where C depends only on T and B. Similar type estimates show that

$$g_0(t_n, u_n) \to g_0(t, u) \quad \text{strongly in } L^2(-\infty, +\infty)$$

as $t_n \to t$, $u_n \to u$. Thus we conclude that g_0 satisfies (A6)′ with $W = D(A)$. Using the same type of argument we can show that the operator f_1 will satisfy (A6)′ with $\mu = \frac{1}{2}$ and $X_{1/2} = D(A^{1/2}(0)) = H^1(-\infty, +\infty)$. It is also clear that g_1 satisfies (A6)′ and that both f_1 and g_1 have sublinear growth. Therefore by Corollary 2, for each $u_0 \in H^2(-\infty, +\infty)$, the Cauchy problem (1.1), (1.2) has a unique global strong solution $u(t)$ on $[0, \infty)$ which belongs to $C([0, \infty); H^2(-\infty, +\infty)) \cap C^1([0, \infty); L^2(-\infty, +\infty))$. If the initial function $u_0(x)$ is real-valued then by (1.1) the imaginary part Im $u(t, x)$ is zero and we obtain a strong solution of (6.13), (6.8).

Example 2. As an application of the results of § 5 we consider the following nonlinear initial-boundary value problem:

$$u_t - (k(t, x)u_x)_x = \int_0^t [a(t, s)(\sigma(s, x, u(s, x), u_x(s, x)))_x + \gamma(t, s)u(s, x)]\, ds + h(t, x) + f(u(t, x), u_x(t, x)), \qquad t > 0, \quad 0 < x < 1, \tag{6.15}$$

$$u(t, 0 = u(t, 1) = 0, \qquad t > 0, \tag{6.16}$$

$$u(0, x) = u_0(x), \qquad 0 < x < 1. \tag{6.17}$$

As in the first example, the functions k, a, σ, γ, h and f are real-valued. We assume $k(t, x)$ satisfies the usual hypotheses on $0 \leqq t < \infty$, $0 \leqq x \leqq 1$ (compare (6.9)) and $a(t, s)$ satisfies (A5) for each $T > 0$. Suppose that $\gamma(t, s)$ is continuous on $[0, \infty) \times [0, \infty)$. Assume that the nonlinear function $\sigma(t, x, u, v)$ is continuous from $[0, \infty) \times [0, 1] \times R \times R$ to R and has continuous first partial derivatives with respect to x, u and v. Also, assume there is a constant $c_1 > 0$ and positive continuous functions $c_i(t, x)$ such that

$$|\sigma_2(t, x, u, v)| \leqq s_1(|u| + |v|) + c_2(t, x),$$
$$|\sigma_i(t, x, u, v)| \leqq c_i(t, x), \qquad i = 3, 4$$

for all (t, x, u, v) in $[0, \infty) \times [0, 1] \times R \times R$, where σ_i denotes the first partial derivative with respect to the ith variable. We assume that the nonlinear function $f(u, v)$ is C^1 on $R \times R$ and satisfies

$$f(0, v) = 0 \quad \text{for all } v \in R, \tag{6.18}$$

$$\nabla f \text{ is bounded and locally Lipschitz continuous.} \tag{6.19}$$

We let $X = L^2(0, 1)$ and $H^2(0, 1)$ denote the usual (complex) Hilbert and Sobolev spaces with norm $|\cdot|$ and $\|\cdot\|$, respectively. For each $t \geqq 0$ we define $A(t)$ in X by

$$A(t) = -\frac{\partial}{\partial x}(k(t, x)Du(x)) \quad \text{a.e. } 0 \leqq x \leqq 1,$$

with domain

$$D(A) = \{u \in H^2(0, 1) : u(0) = u(1) = 0\}.$$

Then the family $\{A(t) : 0 \leqq t < \infty\}$ satisfies (A1), (A2), (A3) on each bounded interval $[0, T]$. Property (5.1) is a consequence of the Sobolev embedding theorem.

By Nirenberg [14], there are positive constants M and κ, $0 < \kappa < \frac{1}{2}$, such that $H^2(0, 1) \to C^{1,\kappa}[0, 1]$ is a continuous embedding, where $C^{1,\kappa}[0, 1]$ denotes the space of functions having uniformly Hölder continuous first derivative with exponent κ, and

furthermore

$$\|u\|_{C^{1,\kappa}[0,1]} \leqq M\|u\|^a |u|^{1-a}, \qquad u \in H^2(0,1),$$

where $a = (\kappa + \frac{3}{2})/2$. We choose $\kappa < \omega < \frac{1}{2}$, then for each $t \geqq 0$ the operator $A^{-\omega}(t)$ is a continuous (compact) mapping from X into $C^{1,\kappa}[0,1]$ (see [1, Proposition 4.1]). By a result of Kato [11] the space $X_\omega = X_\omega(t)$ does not depend on t. So if we let $E = C^{1,\omega}[0,1]$ then (5.2) is satisfied.

Now define $q(t, \varphi, \psi)$ on $[0,\infty] \times E \times X_1$ by

$$\begin{aligned} q(t,\varphi,\psi)(x) = {} & \sigma_2(t, x, \operatorname{Re}\varphi(x), \operatorname{Re} D\varphi(x)) \\ & + \sigma_3(t, x, \operatorname{Re}\varphi(x), \operatorname{Re} D\varphi(x)) \operatorname{Re} D\psi(x) \\ & + \sigma_4(t, x, \operatorname{Re}\varphi(x), \operatorname{Re} D\varphi(x)) \operatorname{Re} D^2\psi(x) \end{aligned}$$

for a.e. $0 \leqq x \leqq 1$. Then q is a continuous mapping from $[0,\infty) \times E \times X_1$ to X and

$$|q(t,\varphi,\psi_1) - q(t,\varphi,\psi_2)| \leqq b(t,\varphi)\|\psi_1 - \psi_2\|_1,$$

where

$$b(t,\varphi) = \sup_{0 \leqq x \leqq 1} \sum_{i=2}^{4} |\sigma_i(t, x, \operatorname{Re}\varphi(x), \operatorname{Re} D\varphi(x))|.$$

It follows that q satisfies (B1)′, (B2) (with $W = X_1$) and (5.4).

We define the operator f_1 on X_1 by

$$f_1(u)(x) = f(\operatorname{Re} u(x), \operatorname{Re} Du(x)), \qquad 0 \leqq x \leqq 1.$$

Using assumptions (6.18), (6.19) it follows that $f_1 : X_1 \to H_0^1(0,1) \equiv D(A^{1/2}(0))$ is well-defined, Lipschitz continuous on bounded subsets of X_1 and has sublinear growth. So if we define $g_0 : [0,\infty) \times X_1 \to X$ by

$$g_0(t,\psi)(x) = q(t,\psi(x),\psi(x)) = \frac{\partial}{\partial x}\sigma(t, x, \operatorname{Re}\psi(x), \operatorname{Re} D\psi(x)),$$

and g_1 by (6.14) then from Theorem 6, for each $u_0 \in D(A)$, there is a global strong solution $u(t)$ of (1.1), (1.2) on $[0,\infty)$ which belongs to $C([0,\infty); H^2(0,1)) \cap C^1([0,\infty); L^2(0,1))$. If the initial function $u_0(x)$ is real, we then obtain a global strong solution of (6.15), (6.16), (6.17).

7. Remarks on regularity. In this section we discuss the regularity of solutions of (6.13), (6.8) in the special case $f(t, u(t,x)) = c_1 u(t,x)$, where c_1 is a constant. We show that if the functions $u_0(x)$, $k(t,x)$, $\rho(t,x)$, $\sigma(s)$, $a(t,s)$, $\gamma(t,s)$ and $h(t,x)$ are sufficiently smooth, then (6.13), (6.8) has a classical solution.

We shall always assume that the hypotheses made in § 6 regarding (6.13) hold. In particular, we assume that (6.9), (6.10), (6.11) are true. We make the following additional hypotheses:

(H1) The partial derivatives $k_t(t,x)$, $k_{tx}(t,x)$, $k_{xx}(t,x)$ exist and are continuous on $[0,\infty) \times R$. Also, for each $T > 0$, there is a constant $C = C(T) > 0$ such that

(i) $\sup_R |k_t(t,x)| \leqq C$, $0 \leqq t \leqq T$;

(ii) $\sup_R |k_{tx}(t,x)| \leqq C$, $0 \leqq t \leqq T$;

(iii) $\sup_R |k_{xx}(t, x)| \leqq C,\ 0 \leqq t \leqq T;$

(iv) $\sup_R |k_t(t, x) - k_t(s, x)| \leqq C|t-s|^{\alpha},\ 0 \leqq s, t \leqq T;$

(v) $\sup_R |k_{tx}(t, x) - k_{tx}(s, x)| \leqq C|t-s|^{\alpha},\ P \leqq s, t \leqq T.$

(H2) The function $a(t, s)$ satisfies hypothesis (a_1) of § 4 for each $T > 0$.

(H3) The function σ belongs to $C^2(-\infty, +\infty)$.

(H4) There is a constant $\delta \in (0, 1]$ such that for each $T > 0$ there is $C = C(T) > 0$ such that

$$\sup_R |\rho(t, x) - \rho(s, x)| \leqq C|t-s|^{\delta},$$

$$\sup_R |\rho_x(t, x) - \rho_x(s, x)| \leqq C|t-s|^{\delta}$$

for all $0 \leqq t, s \leqq T$.

(H5) The function $f_0(t)$ defined by $f_0(t)(x) = h(t, x)$ satisfies hypothesis (f_0) *of* § 4 *for each* $T > 0$. Furthermore, $f_0 \in C([0, \infty); H^1(-\infty, +\infty))$.

(H6) The partial derivative $(\partial\gamma/\partial t)\ (t, s)$ is continuous on $[0, \infty) \times [0, \infty)$.

By taking the positive constant c_0 sufficiently large we may assume without loss of generality that $c_1 = 0$. By the results of § 6, given a real-valued initial-function $u_0 \in H^2(-\infty, +\infty)$, we have a unique global strong solution $u(t)$ of (6.13), (6.8) (with $f(t, u(t, x)) \equiv 0$) which belongs to $C([0, \infty); H^2(-\infty, +\infty)) \cap C^1([0, \infty); L^2(-\infty, +\infty))$. We make the following additional hypotheses.

(H7) $f_0(0) - A(0)u_0$ belongs to $D(A)$.

Under the hypotheses (H1)–(H7) the assumptions of Theorem 4 are satisfied. Hence $u(t)$ belongs to $C^1([0, \infty); H^2(-\infty, +\infty)) \cap C^2([0, \infty); L^2(-\infty, +\infty))$. Using the Sobolev embedding theorem we have the following result.

LEMMA 6. *After redefinition on a set of measure zero the functions* $u(t, x)$, $u_x(t, x)$, $u_t(t, x)$, $u_{tx}(t, x)$ *are all classical derivatives and are continuous on* $[0, \infty) \times R$.

Now consider the distributional derivative $u_{xx}(t, x)$ of $u(t, x)$. It is a measurable function of (t, x) and satisfies a local L^2 condition of the type

$$\int_{t_0}^{t_1} \int_{-\infty}^{+\infty} |u_{xx}(t, x)|^2\, dx\, dt < \infty, \qquad 0 \leqq t_0 < t_1 < \infty.$$

Furthermore, from equation (6.13)

$$\begin{aligned} & k(t, x)u_{xx}(t, x) + \int_0^t a(t, s)\sigma'(\rho(s, x)u_x(s, x))\rho(s, x)u_{xx}(s, x)\, ds \\ & = u_t(t, x) + c_0 u(t, x) - k_x(t, x)u_x(t, x) - h(t, x) \\ & \quad - \int_0^t [a(t, s)\sigma'(\rho(s, x)u_x(s, x))\rho_x(s, x)u_x(s, x) + \gamma(t, s)u(s, x)]\, ds. \end{aligned} \tag{7.1}$$

Let

$$K(t, s, x) = \frac{1}{k(t, x)} \{a(t, s)\sigma'(\rho(s, x)u_x(s, x))\rho(s, x) + \gamma(t, s)u(s, x)\}$$

and let $F(t, x)$ denote the right-hand side of (7.1) divided by $k(t, x)$. *Then* $u_{xx}(t, x)$ is a solution of the Volterra integral equation

$$u_{xx}(t, x) + \int_0^t K(t, s, x) u_{xx}(s, x)\, ds = F(t, x) \quad \text{a.e. } t \geqq 0, \quad x \in R.$$

Since K and F are continuous, it follows that $u_{xx}(t, x)$ can be redefined on a set of measure zero so as to be a continuous function on $[0, \infty) \times R$. It then follows that $u_{xx}(t, x)$ is the classical second partial derivative of $u(t, x)$ with respect to x. Thus $u(t, x)$ is a classical solution of (6.13), (6.8).

We remark that a slightly better regularity result is true, namely, $u(t, x) \in H^3_{\text{loc}}(-\infty, +\infty)$. This is proven by showing that for any bounded interval $[a, b]$ and any $\varepsilon_0 > 0$, $T > 0$ there is a constant $C = C(a, b, T, \varepsilon_0) > 0$ such that

$$\int_c^d |u^h_{xx}(t, x)|^2\, dx \leqq C^2$$

for all $0 \leqq t \leqq T$, $a \leqq c < d \leqq b$, $0 < |h| < \varepsilon_0$. Here the notation u^h_{xx} denotes the difference quotient

$$u^h_{xx}(t, x) = \frac{1}{h}[u_{xx}(t, x + h) - u_{xx}(t, x)].$$

The last inequality is obtained by estimating pointwise the function $t \to u^h_{xx}(t, x)$ as a solution of an appropriate Volterra integral equation.

Acknowledgment. The author wishes to thank the referee for valuable suggestions which improved and broadened the results in the final manuscript.

REFERENCES

[1] H. AMANN, *Periodic solutions of semilinear parabolic equations*, in Nonlinear Analysis, L. Cesari, ed., Academic Press, New York, 1978, pp. 1–29.

[2] V. BARBU, *Integro-differential equations in Hilbert spaces*, An. Sti. Univ. "Al. I. Cuza" Iasi Sect. Ia Math., 19 (1973), pp. 365–383.

[3] ———, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, the Netherlands, 1976.

[4] V. BARBU AND M. A. MALIK, *Semilinear integro-differential equations in Hilbert space*, J. Math. Anal. Appl., 67 (1979), pp. 452–475.

[5] A. BELLENI-MORANTE, *An integro-differential equation arising from the theory of heat conduction in rigid material with memory*, Boll. Un. Mat. Ital., 15-B (1978), pp. 470–482.

[6] B. D. COLEMAN AND V. J. MIZEL, *Thermodynamics and departure from Fourier's law of heat conduction*, Arch. Rational Mech. Anal., 13 (1963), pp. 245–261.

[7] M. G. CRANDALL, S.-O. LONDEN AND J. A. NOHEL, *An abstract nonlinear Volterra integrodifferential equation*, J. Math. Anal. Appl., 64 (1978), pp. 701–735.

[8] W. E. FITZGIBBON, *Semilinear integrodifferential equations in Banach space*, Preprint.

[9] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[10] A. FRIEDMAN AND M. SHINBROT, *Volterra integral equations in Banach space*, Trans, Amer. Math. Soc., 126 (1967), pp. 131–179.

[11] T. KATO, *Fractional powers of dissipative operators*, J. Math. Soc. Japan., 13 (1961), pp. 246–274.

[12] ———, *On linear differential equations in Banach spaces*, Comm. Pure Applied Math., IX, (1956), pp. 479–486.

[13] R. C. MACCAMY, *Stability theorems for a class of functional differential equations*, SIAM J. Appl. Math., 30, (1976), pp. 557–576.

[14] L. NIRENBERG, *On elliptic partial differential equations*, Ann. Scuola Norm. Sup Pisa Sci. Fis. Mat., 13 (1959), pp. 115–161.

[15] J. A. NOHEL, *Nonlinear Volterra equations for heat flow in materials with memory*, MRC Tech. Summary Rep. 2081, Mathematics Research Center, University of Wisconsin, Madison.

[16] J. W. NUNZIATO, *On heat conduction in materials with memory*, Quarterly Appl. Math., 29 (1971), pp. 187–204.

[17] C. RENNOLET, *Existence and boundedness of solutions of abstract nonlinear integrodifferential equations of nonconvolution type*. J. Math. Anal. Appl., 70 (1979), pp. 42–60.

[18] P. E. SOBOLEVSKY, *Equations of parabolic type in a Banach space*, Amer. Math. Soc. Translations (2), 49 (1966), pp. 1–62.

[19] H. TANABE, *On the equations of evolution in a Banach space*, Osaka J. Math., 12 (1960), pp. 363–376.

[20] I. I. VRABIE, *The nonlinear version of Pazy's local existence theorem*, Israel J. Math., 32 (1979), pp. 221–235.

[21] G. F. WEBB, *An abstract semilinear Volterra integrodifferential equation*, Proc. Amer. Math. Soc., 69, (1978), pp. 225–260.

[22] ———, *Abstract Volterra integrodifferential equations and a class of reaction-diffusion equations*, Lecture Notes in Mathematics, 737, Springer-Verlag, New York, (1979), pp. 295–303.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0007 $01.00/0

GENERAL SOLUTION OF THE PRICE-DIVIDEND INTEGRAL EQUATION*

N. A. DERZKO† AND S. P. SETHI‡

Abstract. This paper reports some new closed-form formulas of financial valuation for a deterministic firm with general financing policies and a time-dependent discount rate. A model of the firm is described which includes the price-dividend-balance integral equation whose solution yields the time path of share price, number of shares, and the value of the firm. The solution technique depends on deriving an equivalent system of differential equations. A broad class of firms for which the solution formulas are valid is characterized.

Introduction. In this paper we study the valuation of a firm described by a deterministic model in which the discount rate or the rate of return required by the stockholders of the firm is assumed to be time-dependent and exogenously given. Special versions of this model are available in Gordon [2] and Miller and Modigliani [3]. For our purposes the firm is defined by two functions on $[0, \infty)$; $D(t)$ gives the rate at which dividends are paid, and $E(t)$ gives the rate at which external equity capital is raised. The rate of dividend per share is given by $D(t)$ divided by the number of shares outstanding. External capital is raised by selling the firm's stock at the current market price. This process increases the number of outstanding shares when $E(t) > 0$. When $E(t) < 0$, the firm is buying back its own stock, thereby reducing the number of outstanding shares.

Therefore, the essential piece of information for the valuation problem is the price of one share which in our ideal world is assumed to be the discounted present value of future dividend payments to that share [2]. An alternate approach is indicated at the end of this paper.

Within this context, the governing equations of the model are developed and solved under fairly minimal assumptions on D and E and a variable discount rate $k(t)$.

1. Preliminary remarks. Traditionally a firm is defined in terms of the total rate of earnings $X(t)$, the dividend rate $D(t)$ and the rate of external equity funding $E(t)$. It is normally assumed that $D(t) \leqq X(t)$, and $0 \leqq E(t)$ for all t. A nonnegative function $r(t)$ giving the rate of return on the firm's capital is also introduced.

The governing differential equation for $X(t)$ is then

$$\begin{aligned} &X'(t) = r(t)[X(t) + E(t) - D(t)], \\ &X(0) = X_0 \quad \text{given.} \end{aligned} \tag{1.1}$$

This section concludes by showing that once X_0, D and E are given, it is possible to find $X(t)$ and $r(t)$, satisfying all the usual assumptions.

THEOREM 1.1. *Let $D(t)$ and $E(t)$ be nonnegative Borel measurable functions on $[0, \infty)$ and let X_0 be positive. Then there exist nonnegative functions $r(t)$ and $X(t)$ on $[0, \infty)$ such that*

1. *$r(t)$ is locally integrable and positive,*
2. *$X(t)$ is absolutely continuous, positive, and $X(0) = X_0$,*
3. *(1.1) is satisfied* a.e.

* Received by the editors November 4, 1980 and in final form March 23, 1981.
† Department of Mathematics, University of Toronto, Toronto, Canada M5S 1A1.
‡ Faculty of Management Studies, University of Toronto, Toronto, Canada, M5S 1V4.

Proof. We rearrange (1.1) as follows:

$$\dot{X} - rX = r(E - D). \tag{1.2}$$

(1.2) is a first order linear differential equation for X and has a formal solution

$$X(t) = \exp\left(\int_0^t r(\tau)\,d\tau\right)\left\{\int_0^t (E(s) - D(s))\,d\left[-\exp\left(-\int_0^s r(\tau)\,d\tau\right)\right] + X_0\right\}. \tag{1.3}$$

We let $\phi(s) = -\exp(-\int_0^s r(\tau)\,d\tau)$, and note that ϕ is:

(1.4)
1. negative with $\phi(0) = -1$,
2. increasing,
3. absolutely continuous.

Furthermore, for any function ϕ satisfying the above requirements there exists a corresponding r.

The term within { } on the right side of (1.3) can be written in terms of ϕ as

$$\int_0^t (E(s) - D(s))\,d\phi\,(s) + X_0. \tag{1.5}$$

It is clear that by controlling the rate of increase of ϕ on the sets where $E(\cdot) - D(\cdot)$ is negative or very large we can arrange that $E(\cdot) - D(\cdot)$ is integrable with respect to $d\phi\,(\cdot)$ and that (1.5) remains positive. This completes the proof.

We return now to D and E as starting points for a model.

2. Definition of the present model. We begin with the discount rate. Let $k(t)$ be an instantaneous discount rate such that the discounted value of M_2 at time t_2 to the present time t_1 is

$$M_1 = M_2 \exp\left(-\int_{t_1}^{t_2} k(s)\,ds\right).$$

It is convenient to think of the exponential multiplier as a "present value" operator pv (t_1, t_2). The present value operator has the useful property

$$\text{pv}\,(t_1, t_3) = \text{pv}\,(t_1, t_2)\,\text{pv}\,(t_2, t_3). \tag{2.1}$$

The minimal assumptions needed to write the basic equations of the model are:

(A1) E, D, k are real-valued locally integrable functions on $[0, \infty)$,

(A2) D, k are nonnegative on $[0, \infty)$.

A firm shall be the pair (D, E).

To write down the basic equations of the model we require a financial theory: the price of a single share is the discounted present value of future dividends which that share earns. Thus, if $P(t)$ denotes price and $N(t)$ denotes the number of shares outstanding at time t,

$$P(t) = \int_t^\infty \frac{\text{pv}\,(t, \tau)D(\tau)\,d\tau}{N(\tau)}, \tag{2.2}$$

$$N(t) = N_0 + \int_0^t \frac{E(s)}{P(s)}\,ds. \tag{2.3}$$

Of course, the value of the firm $V(t)$ at time t can be defined as $V(t) = N(t)P(t)$. Note

that N_0 is the number of shares at time $t=0$ and effectively establishes share denomination.

We seek a solution $P(t)$, $N(t)$ in the class of positive measurable functions on $[0, \infty)$ for which the integrands of (2.2) and (2.3) are Lebesgue integrable. The equations then imply that P and N must be absolutely continuous. It turns out that the calculations are greatly facilitated if we work with the present value variables

$$\begin{aligned} \mathscr{P}(t) &= \text{pv}\,(0, t)P(t), \\ \mathscr{D}(t) &= \text{pv}\,(0, t)D(t), \\ \mathscr{E}(t) &= \text{pv}\,(0, t)E(t). \end{aligned} \tag{2.4}$$

If we multiply (2.2) by pv $(0, t)$, we obtain

$$\text{pv}\,(0, t)P(t) = \int_t^\infty \frac{\text{pv}\,(0, t)\,\text{pv}\,(t, \tau)D(\tau)\,d\tau}{N(\tau)}.$$

Also (2.3) can be written

$$N(t) = N_0 + \int_0^t \frac{\text{pv}\,(0, s)E(s)}{\text{pv}\,(0, s)P(s)}\,ds.$$

Using the variables of (2.4), these equations become

$$\mathscr{P}(t) = \int_t^\infty \frac{\mathscr{D}(\tau)}{N(\tau)}\,d\tau, \tag{2.5}$$

$$N(t) = N_0 + \int_0^t \frac{\mathscr{E}(s)}{\mathscr{P}(s)}\,ds; \tag{2.6}$$

a solution pair $(\mathscr{P}, N)$ consists of a pair of positive and absolutely continuous functions on $[0, \infty)$, for which the integrals in (2.5) and (2.6) are Lebesgue and the equations are satisfied.

Differentiation of (2.5) and (2.6) yields

$$\mathscr{P}'(t) = -\frac{\mathscr{D}(t)}{N(t)}, \tag{2.7}$$

$$N'(t) = \frac{\mathscr{E}(t)}{\mathscr{P}(t)}, \tag{2.8}$$

t-a.e. Note that (2.7), which can also be stated as

$$P'(t) = k(t)P(t) - \frac{D(t)}{N(t)},$$

is the well-known arbitrage equation of Miller and Modigliani [3].

3. Solution of the (D, E) model. The natural class of functions in which to seek solutions to (2.7) and (2.8) is larger than the natural solution class pertaining to (2.4) and (2.5). To define it we need only require that there be a nonempty interval $[0, T]$ on which $\mathscr{P}$ and N are positive and absolutely continuous. We shall call these *blocked solutions.*

The initial value problem at $t=0$ for system (2.7), (2.8) possesses a solution on some nonempty interval $[0, T]$ for each set of positive initial values $\mathscr{P}(0)$, N_0. However, we shall see shortly that under fairly lenient assumptions on $\mathscr{D}$ and $\mathscr{E}$, for each N_0 there

is a unique $\mathscr{P}(0)$ which yields a positive solution on $[0, \infty)$. This initial price $\mathscr{P}(0)$ coincides with the accepted value derived using finance arguments [3], [2],

$$\Pi_0 = \frac{1}{N_0} \int_0^\infty (\mathscr{D}(s) - \mathscr{E}(s))\, ds. \tag{3.1}$$

We clarify the connection between (2.5), (2.6) and (2.7), (2.8) in the following.

THEOREM 3.1. *$\mathscr{P}(t), N(t), 0 \leqq t < \infty$ are positive solutions of* (2.5, 2.6) *if and only if they are positive solutions of* (2.7, 2.8) *and*

$$N(0) = N_0, \tag{3.2}$$

$$\lim_{t \to \infty} \mathscr{P}(t) = 0, \tag{3.3}$$

$$\frac{\mathscr{D}(\cdot)}{N(\cdot)} \in L_1(0, \infty), \tag{3.4}$$

$$\frac{\mathscr{E}(\cdot)}{\mathscr{P}(\cdot)} \in L_1^{\text{loc}}(0, \infty). \tag{3.5}$$

Proof. ($\Rightarrow$). Properties (3.2)–(3.5) are consequences of our assumption that the integrals in (2.5), (2.6) are Lebesgue. In addition it follows that $\mathscr{P}$ and N are absolutely continuous and can be differentiated to yield (2.7), (2.8).

($\Leftarrow$). Assumptions (3.2)–(3.5) enable us to integrate (2.7), (2.8) to obtain

$$\mathscr{P}(t) - \mathscr{P}(T) = \int_t^T \frac{\mathscr{D}(\tau)}{N(\tau)}\, d\tau$$

and

$$N(t) = N_0 + \int_0^t \frac{\mathscr{E}(\tau)}{\mathscr{P}(\tau)}\, d\tau,$$

which after taking $\lim_{T \to \infty}$ yields (2.5), (2.6). This completes the proof.

To solve the differential system we proceed as follows. From (2.7), (2.8) we obtain

$$\mathscr{E}(t) - \mathscr{D}(t) = N(t)\mathscr{P}'(t) + N'(t)\mathscr{P}(t) = \frac{d}{dt}(N(t)\mathscr{P}(t)). \tag{3.6}$$

Integration from 0 to t and application of the initial conditions yields

$$N(t)\mathscr{P}(t) = N_0\mathscr{P}(0) - \int_0^t (\mathscr{D}(s) - \mathscr{E}(s))\, ds. \tag{3.7}$$

We now use (3.7) together with (2.7) to obtain

$$\frac{\mathscr{P}'(t)}{\mathscr{P}(t)} = \frac{-\mathscr{D}(t)}{N_0\mathscr{P}(0) - \int_0^t (\mathscr{D}(\tau) - \mathscr{E}(\tau))\, d\tau}$$

which has the solution

$$\mathscr{P}(t) = \mathscr{P}(0) \exp \int_0^t \frac{-\mathscr{D}(\tau)\, d\tau}{N_0\mathscr{P}(0) - \int_0^\tau (\mathscr{D}(s) - \mathscr{E}(s))\, ds}. \tag{3.8}$$

Similarly, using (3.7) together with (2.8) yields

$$N(t) = N_0 \exp \int_0^t \frac{\mathscr{E}(\tau)\, d\tau}{N_0\mathscr{P}(0) - \int_0^\tau (\mathscr{D}(s) - \mathscr{E}(s))\, ds}. \tag{3.9}$$

Furthermore, it is clear from formulas (3.8), (3.9) that, if N_0 and $\mathscr{P}(0)$ are positive, the initial value problem has the unique solution given by the formulas on some interval $[0, T]$, $T>0$. It is also evident that positive solutions cannot exist beyond any point T for which

$$N_0\mathscr{P}(0)=\int_0^T (\mathscr{D}(s)-\mathscr{E}(s))\,ds.$$

Such points T will always exist if

$$\sup_T \int_0^T (\mathscr{D}(s)-\mathscr{E}(s))\,ds=\gamma_1=\infty.$$

We shall therefore assume that $\gamma_1<\infty$. This being the case, consider the choice of $\mathscr{P}(0)$. If $N_0\mathscr{P}(0)>\gamma_1$, it follows from (3.8) that to satisfy $\lim_{t\to\infty}\mathscr{P}(t)=0$, we need $\int_0^\infty \mathscr{D}(\tau)\,d\tau=\infty$, that is, an infinite present value of dividend payments. It follows also that $\sup_T \int_0^T \mathscr{E}(s)\,ds=\infty$. Thus, we have an explosive scenario in which stock is being sold just to pay dividends on shares outstanding, which is not satisfactory from the financial point of view. Considerations of this kind motivate the assumptions of the following section.

4. Further assumptions. Let us assume, in addition to (A1), (A2), that

$$\mathscr{D}(\cdot)-\mathscr{E}(\cdot)\in L_1(0,\infty),$$

and

(A3) $$0<\int_\tau^\infty (\mathscr{D}(s)-\mathscr{E}(s))\,ds \quad \forall \tau\geqq 0.$$

The effect of (A3) is to fix a value of $\mathscr{P}(0)$ in (4.1) for which

$$N_0\mathscr{P}(0)-\int_0^\tau (\mathscr{D}(s)-\mathscr{E}(s))\,ds>0 \quad \forall\tau,$$

and for which the expression on the left approaches 0 as $\tau\to\infty$. In the remainder of this section we shall let

(4.1) $$\mathscr{P}(0)=\frac{1}{N_0}\int_0^\infty (\mathscr{D}(s)-\mathscr{E}(s))\,ds.$$

Then, (3.8), (3.9) become

(4.2) $$\mathscr{P}(t)=\mathscr{P}(0)\exp\int_0^t \frac{-\mathscr{D}(\tau)\,d\tau}{\int_\tau^\infty (\mathscr{D}(s)-\mathscr{E}(s))\,ds},$$

(4.3) $$N(t)=N_0\exp\int_0^t \frac{\mathscr{E}(\tau)\,d\tau}{\int_\tau^\infty (\mathscr{D}(s)-\mathscr{E}(s))\,ds}.$$

We note that these closed-form solutions have been obtained in this general setting for the first time.

Our final assumption requires that $N(t)$ does not approach 0, that is,

(A4) $$\gamma=\inf_t \int_0^t \frac{\mathscr{E}(\tau)\,d\tau}{\int_\tau^\infty (\mathscr{D}(s)-\mathscr{E}(s))\,ds}>-\infty.$$

Under (A4), we are able to obtain the present value $V(t)$ pv $(0, t)$ of the firm from (3.7), which now becomes

$$V(t) \text{ pv } (0, t) = N(t)\mathcal{P}(t) = \int_t^\infty (\mathcal{D}(s) - \mathcal{E}(s))\, ds, \tag{4.4}$$

and also conclude that $\lim_{t\to\infty} \mathcal{P}(t) = 0$.

At this point we have completed the proof of the key theorem.

THEOREM 4.1. *If (D, E) is a firm satisfying* (A1)–(A4), *then for an initial number of shares N_0,* (4.2), (4.3) *is the unique positive solution existing on* $[0, \infty)$ *to the basic equations* (2.2), (2.3) *governing share price and number of shares at time t.*

We note that formula (4.4) is well known, if not mathematically rigorously derived, in the finance literature [2]. It states that the value of the firm at time t is the present value of the total dividends accruing to the stockholders of record at time t. The integral of the first term in the integrand represents the total present value of dividends issued by the firm in the interval $[t, \infty)$. A portion of this dividend is obviously going to stocks issued in the interval (t, ∞). However, this portion in an efficient market under certainty, i.e., where no arbitrage possibilities exist, must equal the integral of the second term in the integrand of (4.4).

Clearly, the residual represented by (4.4), which came to the stockholder of record t, can now be interpreted as the present value of the firm at time t.

We also note that the steps of the foregoing analysis are reversible in the sense that we could have started out with (4.4) as the formula for the value of the firm and derive the price-dividend balance equation (2.2) and obtain the share price formula (4.2) with $P(0)$ as in (4.1).

Finally, it should be mentioned that the (D, E) model of a firm is meaningful under weaker assumptions than (A1)–(A4). It is possible, for example, to define very general solution classes to (2.4)–(2.6) by allowing the blocked solutions to extend to $+\infty$ with value 0 and adopting the convention that $1/0 = \infty$ and $1/\infty = 0$ in the integral equations. The financially meaningful solution is then defined as the supremum of a solution class. Such an approach has the advantage of producing a financially acceptable solution for certain examples excluded by assumptions (A1)–(A4). Furthermore, it is also possible to extend the model for the case when (D, E) is an arbitrary stochastic process.

REFERENCES

[1] E. J. ELTON AND M. J. GRUBER, *Valuation and asset selection under alternative investment opportunities*, J. Finance, 31 (1976), pp. 525–549.

[2] M. J. GORDON, *The Investment, Financing and Valuation of the Corporation*, Richard D. Irwin Inc., Homewood, Il, 1962.

[3] M. H. MILLER AND F. MODIGLIANI, *Dividend policy, growth, and the valuation of shares*, J. Business, 34 (1961), pp. 411–433.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0008 $01.00/0

NONZERO SOLUTIONS OF NONLINEAR INTEGRAL EQUATIONS MODELING INFECTIOUS DISEASE*

LYNN R. WILLIAMS† AND RICHARD W. LEGGETT‡

Abstract. Sufficient conditions to insure the existence of periodic solutions to the nonlinear integral equation, $x(t)=\int_{t-\tau}^{t} f(s, x(s))\,ds$, are given in terms of simple product and product integral inequalities. The equation can be interpreted as a model for the spread of infectious diseases (e.g., gonorrhea or any of the rhinovirus viruses) if $x(t)$ is the proportion of infectives at time t and $f(t, x(t))$ is the proportion of new infectives per unit time.

1. Summary of results.

The nonlinear integral equation

$$x(t)=\int_{t-\tau}^{t} f(s, x(s))\,ds \tag{1.1}$$

can be interpreted as a model for the spread of a number of infectious diseases with periodic contact rate that varies due to certain seasonal factors. This model was formulated and discussed at some length by Cooke and Kaplan in [2]. Briefly, $x(t)$ represents the proportion of infectives (the number of individuals in the population who are infectious divided by the size of the population) at time t, $f(t, x(t))$ is the proportion of new infectives per unit time ($f(t, 0)=0$), and the positive constant τ is the length of time an individual remains infectious. Cooke and Kaplan consider functions which generalize $f(t, x)=a(t)x(1-x)$, where $a(t)$ is the effective contact rate. Equation (1.1) represents an *S-I-S* model, that is, it is assumed that the population is divided into susceptibles S and infectives I and that the disease is not lethal and confers no immunity. Furthermore, there is assumed to be no latent period between being exposed and becoming infectious.

Assuming that f is a continuous, bounded, nonnegative function which is ω-periodic in t for some $\omega>0$, Cooke and Kaplan show that (1.1) has a nontrivial ω-periodic solution provided

$$\alpha \equiv \inf_{t\in\mathbb{R}} a(t) > \frac{1}{\tau},$$

where $a(t)$ is the uniform limit of $f(t, x)/x$ as x decreases to zero. This result can be interpreted physically as implying that, for the type of disease modeled by (1.1), the infection can remain endemic to the population and the number of infectives can oscillate periodically provided the contact rate $a(t)$, the average of effective contacts with other individuals per infective per time period, exceeds $1/\tau$. Thus the result is a type of *threshold theorem*, since it asserts that the infection attains a "periodic steady state" in the population provided $\alpha\tau$ remains above the threshold level of 1.

Nussbaum [6] considered (1.1) in terms of the linear operator L_τ defined by

$$L_\tau x(t)=\int_{t-\tau}^{t} a(s)x(s)\,ds \tag{1.2}$$

* Received by the editors September 6, 1979 and in final revised form March 17, 1981.

† Department of Mathematics, Indiana University at South Bend, South Bend, Indiana 46615. The research of this author was partially supported by a Summer Faculty Fellowship from Indiana University at South Bend.

‡ Health and Safety Research Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830. The research of this author was supported by Union Carbide Corporation under contract W-7405-eng-26 with the U.S. Department of Energy.

and established the existence of a number τ_0 such that the spectral radius $r(L_\tau)$ satisfies

$$r(L_\tau)\begin{cases}<1 & \text{for } \tau<\tau,\\ =1 & \text{for } \tau=\tau_0,\\ >1 & \text{for } \tau>\tau_0.\end{cases}$$

The number τ_0 represents a threshold in the sense that if $\tau>\tau_0$, then (1.1) has a nontrivial periodic solution, but if $\tau\leqq\tau_0$, then, for most functions of interest, (1.1) has no nontrivial periodic solution. Smith [7] obtained similar results and established the existence of a number $\tau^0>\tau_0$ such that if $\tau_0<\tau<\tau^0$, then (1.1) has a unique positive periodic solution. Smith also showed that if $\tau_0<\tau_1<\tau_2<\tau^0$, then the solution at τ_1 is smaller than the solution at τ_2. Crude estimates are given in [6] and [7] which show that (1.1) has a positive periodic solution provided

$$\inf_{0\leqq t\leqq\omega}\int_{t-\tau}^{t} a(s)\,ds>1.$$

Nussbaum [6] also gives estimates on $r(L_\tau)$ in terms of approximating operators with finite dimensional range; however, these results may be difficult to apply and are numerically implementable only for special contact rates $a(s)$.

We establish the existence of positive ω-periodic solutions to (1.1) in terms of a simple product or product integral. The results are easily implemented and only require that one have basic information about upper and lower bounds of the contact rate on certain subintervals of $[0,\omega]$. We obtain periodic solutions without requiring that the average contact rate exceed a threshold level on each time interval of length equal to the duration of infection. Even though the average contact rate is small during some time intervals, the disease may remain endemic to the population provided the contact rate is sufficiently large during the remaining intervals. Although we do not specifically consider the operator L_τ, our results do have implications for $r(L_\tau)$ and may be viewed as giving a computable, sufficient condition to ensure that $r(L_\tau)>1$.

We assume throughout that τ and ω are positive constants and make the following assumptions on f and a:

H1. The function $f(t,x)$ is continuous from $(-\infty,\infty)\times[0,\infty)$ into $[0,\infty)$.

H2. For each $t\in\mathbb{R}$ and $x\geqq 0$, $f(t,x)=f(t+\omega,x)$ and $f(t,0)=0$.

H3. The function $a(t)$ is the uniform limit as x approaches zero of $f(t,x)/x$ and $a(t)$ is bounded away from zero.

H4. There exists $R>0$ such that $f(t,x)\leqq R/\tau$ for all $(t,x)\in[0,\omega]\times[0,R]$.

Assumptions H1, H2, and H3 are precisely as in [2] and [6]. Instead of H4, Cooke and Kaplan [2] require f to be bounded above and Nussbaum [6] requires $\lim_{x\to\infty}(f(t,x)/x)=0$.

In our first result, we use the function

$$P(x,y)=\begin{cases}e^{x\tau}-1-\dfrac{x}{x-y}e^{(x-y)\tau}+\dfrac{x}{x-y} & \text{if } x\neq y,\\ e^{x\tau}-1-x\tau & \text{if } x=y.\end{cases}$$

If M_1 and M_2 represent minimal contact rates on consecutive intervals of length τ, then $P(M_1,M_2)$ provides a measure of the carryover of infectives from one interval to the next.

THEOREM 1. *Assume that f and a satisfy* H1–H4. *Suppose $\omega=N\tau+\gamma$, for some integer $N\geqq 0$ and some $\gamma\geqq 0$, and let $b,c\in[0,\omega]$ satisfy $b+\gamma=c$. For each integer j, set $d_j=b-(j-1)\tau$, and let $M_j=\inf\{a(t)\colon d_j\leqq t\leqq d_{j-1}\}$. Then* (1.1) *has a positive ω-periodic*

solution provided

(i) $$\int_s^{s+\tau} a(t)\,dt > 1, \qquad b \leqq s \leqq c,$$

(ii) $$M_{N+1}\tau > 1,$$

(iii) $$(1-e^{-M_N\tau})\left(M_1\tau e^{M_2\tau} - \frac{M_1}{M_2}e^{M_2\tau} + \frac{M_1}{M_2}\right)\prod_{j=2}^{N-1} P(M_{j+1}, M_j) > 1.$$

With $N = 0$ in Theorem 1, conditions (i), (ii) and (iii) can be reduced to the requirement that

$$\int_s^{s+\tau} a(t)\,dt > 1, \qquad 0 \leqq s \leqq \omega$$

(cf. Nussbaum [6, Lemma 7]).

Example 1. In studying the incidence of chickenpox, mumps and measles in New York City and Baltimore for a thirty-five year period, London and Yorke [5] concluded that the contact rate for each disease is 1.7 to 2 times higher in the winter months (when school is in session) than in the summer months. Although these diseases are not *S-I-S* diseases, the large variance in contact rate may be exhibited by other diseases. Motivated by the work of London and Yorke, we let $\omega = 365$, $\tau = 15$, and assume that $a(t) \geqq 1.4/15$ on $[0; 260]$ and that $a(t) \geqq .7/15$ on $[260, 365]$. Choosing $b = 0$ and $c = 5$ gives

$$M_1 = 1.4/15,$$

$$M_j = .7/15, \qquad 2 \leqq j \leqq 8,$$

and

$$M_j = 1.4/15, \qquad 9 \leqq j \leqq 25.$$

Thus, $M_{25}\tau > 1$, and

$$(1-e^{-M_{24}\tau})\left(M_1\tau e^{M_2\tau} - \frac{M_1}{M_2}e^{M_2\tau} + \frac{M_1}{M_2}\right)\prod_{j=2}^{23} P(M_{j+1}, M_j) \cong 1.1214 > 1.$$

Also, for $b \leqq s \leqq c$, $\int_s^{s+\tau} a(t)\,dt \geqq 1.4 > 1$. The existence of a nonzero solution to (1.1) follows from Theorem 1 provided f and a satisfy H1–H4, even though $a(t) < 1/\tau$ on an interval of length 7τ.

Theorem 1 can be improved by replacing conditions (ii) and (iii) with an integral condition; however, the integral condition is typically not as easily verified as condition (iii).

THEOREM 2. *Assume that f and a satisfy* H1–H4, *and that N, γ, b, c, $d_1, \cdots, d_N$ and $M_1, \cdots, M_N$ are as in the statement of Theorem* 1. *Then* (1.1) *has a positive ω-periodic solution provided*

(i) $$\int_s^{s+\tau} a(t)\,dt > 1, \qquad b \leqq s \leqq c,$$

(ii) $$M_{N+1}(c+\tau-s) + g_N(s-\omega) > 1, \qquad c \leqq s \leqq c+\tau,$$

where $g_1(s) = M_1(s-d_2)$ and for $j \geqq 1$

$$g_{j+1}(s) = \int_{d_{j+1}}^{d_j} g_j(t)M_{j+1}\,e^{M_{j+1}(t-d_{j+1})}\,dt - \int_{s+\tau}^{d_j} g_j(t)M_{j+1}\,e^{M_{j+1}(t-s-\tau)}\,dt.$$

Although verification of condition (ii) of Theorem 2 is not as easily done as condition (iii) of Theorem 1, the determination of $g_N(s)$ is straightforward. At each stage the integrands involve only expressions of the form c, ct, ce^{Mt}, or cte^{Mt}, for constants c and M. The next example illustrates the dramatic improvement that can occur with this extra computational effort.

Example 2. Consider the case described by the conditions $\omega = 40$, $\tau = 10$, $a(t) \geqq .2$ on $[0, 10]$, $a(t) \geqq .1$ on $[20, 40]$ and $a(t) \geqq .075$ on $[10, 20]$. We take $N = 4$. Since $M_{N+1}\tau$ must be larger than 1, we must choose $b = c = 0$. Then $M_1 = .2$, $M_2 = M_3 = .1$ and $M_4 = .075$. Conditions (i) and (ii) of Theorem 1 (or Theorem 2) are satisfied, but the product in condition (iii) of Theorem 1 is

$$(1 - e^{-.75})(2e - 2e + 2)P(.1, .1)P(.075, .1) \cong .3437.$$

Thus, the existence of a positive solution to (1.1) does not follow from Theorem 1. However,

$$g_1(s) = .2(s + 10),$$
$$g_2(s) = .2(s + 20),$$
$$g_3(s) = .2(s + 30),$$

and

$$g_4(s - 40) = \tfrac{2}{3}e^{.75}(-1 + e^{-.075s}) + .2s.$$

Then $.2(10 - s) + g_4(s - 40) = 2 + \frac{2}{3}e^{.75}(-1 + e^{-.075s})$, and its minimum value on $[0, 10]$ occurs when $s = 10$. This minimum is $\frac{8}{3} - \frac{2}{3}e^{.75} \simeq 1.2553$. Thus, if f and a satisfy H1–H4, the existence of a nonzero solution to (1.1) follows from Theorem 2.

2. Proofs of theorems. The proofs employ the compression of the cone theorem of Krasnosel'skii [3] for completely continuous operators on ordered Banach spaces. Let E be a real Banach space. A closed, convex set $K \subset E$ is called a (*positive*) *cone* if the following conditions are satisfied:

(i) if $x \in K$, then $\lambda x \in K$ for $\lambda \geqq 0$;

(ii) if $x \in K$ and $-x \in K$, then $x = 0$.

A cone K in E induces a partial ordering $\leqq$ in E by

$$x \leqq y \quad \text{if and only if} \quad y - x \in K.$$

(We will write $x \not\leqq y$ if $y - x \notin K$.) A Banach space E with a partial ordering $\leqq$ induced by a cone K is called an *ordered Banach space*. By a *completely continuous map* we mean a continuous function which takes bounded sets into relatively compact sets. We consider completely continuous maps which take some subset K_c, $0 < c \leqq \infty$, of a cone K back into K, where $K_c = \{x \in K : \|x\| \leqq c\}$, $0 < c < \infty$, and $K_\infty = K$.

The following theorem is essentially due to Krasnosel'skii [3, p. 137] (see also [1] and [4]).

THEOREM 3 (compression of a cone). *Let $A: K \to K$ be a completely continuous operator. If r and R are real numbers, $0 < r < R$, such that*

(i) $Ax \not\leqq x$ *if* $x \in K_r$ *and* $x \neq 0$,

(ii) *for each* $\varepsilon > 0$, $(1 + \varepsilon)x \not\leqq Ax$ *if* $x \in K$ *and* $\|x\| = R$, *then A has at least one nonzero fixed point in K.*

Proof of Theorem 1. Let E be the sup-normed Banach space of continuous real-valued functions on $\mathbb{R}$ which are ω-periodic, and let K be the cone of nonnegative functions in E. Define the operator $A: K \to K$ by

$$Ax(t) = \int_{t-\tau}^{t} f(s, x(s))\, ds. \tag{2.1}$$

It is easy to see, using the Arzela–Ascoli theorem, that A is completely continuous. The proof of Theorem 1 involves showing that under the given assumptions, the operator A defined in (2.1) satisfies the conditions of Theorem 3. It follows from assumption H4 that A maps K_R into K_R, so that A satisfies condition (ii) of Theorem 3. It remains to be shown that A satisfies condition (i) of Theorem 3 for some $r > 0$.

Let k be a real number such that $0 < k < 1$,

$$k \int_{s}^{s+\tau} a(t)\, dt > 1, \qquad b \leqq s \leqq c,$$

kM_{N+1}, $\tau > 1$, and

$$(1 - e^{-kM_N\tau})\left(kM_1\tau\, e^{kM_2\tau} - \frac{M_1}{M_2}\, e^{kM_2\tau} + \frac{M_1}{M_2}\right) \prod_{j=2}^{N-1} P(kM_{j+1}, kM_j) > 1.$$

It follows from assumption H3 that there exists a real number r, $0 < r < R$, such that $f(t, s) \geqq ka(t)s$, $0 \leqq s \leqq r$. Suppose $x \in K_r$, $x \neq 0$ and $Ax \leqq x$. We will show that this assumption leads to a contradiction. It will then follow that A satisfies the conditions of Theorem 3 and has a nonzero fixed point in K.

Since $Ax \leqq x$,

$$\begin{aligned}
\int_{b}^{c+\tau} a(s)x(s)\, ds &\geqq \int_{b}^{c+\tau} a(s)Ax(s)\, ds \\
&= \int_{b}^{c+\tau} a(s)\left(\int_{s-\tau}^{s} f(t, x(t))\, dt\right) ds \\
&\geqq \int_{b}^{c+\tau} a(s)\left(\int_{s-\tau}^{s} ka(t)x(t)\, dt\right) ds.
\end{aligned}$$

By changing the order of integration, it follows that

$$\begin{aligned}
\int_{b}^{c+\tau} a(s)x(s)\, ds \geqq &\int_{b-\tau}^{b} a(t)x(t)\left(\int_{b}^{t+\tau} ka(s)\, ds\right) dt \\
&+ \int_{b}^{c} a(t)x(t)\left(\int_{t}^{t+\tau} ka(s)\, ds\right) dt \\
&+ \int_{c}^{c+\tau} a(t)x(t)\left(\int_{t}^{c+\tau} ka(s)\, ds\right) dt.
\end{aligned} \tag{2.2}$$

We now transform the integral over $[b-\tau, b]$ in (2.2) to an integral over $[c, c+\tau]$ from which the product in condition (iii) is derived.

Suppose g is any bounded, nonnegative, integrable function, l is a real number, and $M = \inf\{a(t): l-\tau \leqq t \leqq l\}$. Then

$$\begin{aligned}
\int_{l-\tau}^{l} a(t)x(t)g(t)\,dt &\geqq \int_{l-\tau}^{l} a(t)Ax(t)g(t)\,dt \\
&= \int_{l-\tau}^{l} a(t)g(t)\left(\int_{t-\tau}^{t} f(s, x(s))\,ds\right) dt \\
&\geqq \int_{l-\tau}^{l} a(t)g(t)\left(\int_{t-\tau}^{t} ka(s)x(s)\,ds\right) dt \\
&= \int_{l-2\tau}^{l-\tau} a(s)x(s)\left(\int_{l-\tau}^{s+\tau} ka(t)g(t)\,dt\right) ds \\
&\quad + \int_{l-\tau}^{l} a(s)x(s)\left(\int_{s}^{l} ka(t)g(t)\,dt\right) ds \\
&\geqq \int_{l-2\tau}^{l-\tau} a(s)x(s)\left(\int_{l-\tau}^{s+\tau} kMg(t)\,dt\right) ds \\
&\quad + \int_{l-\tau}^{l} a(s)x(s)\left(\int_{s}^{l} kMg(t)\,dt\right) ds.
\end{aligned}$$

From this result, it follows that

$$\begin{aligned}
&\int_{l-\tau}^{l} a(s)x(s)\left(\int_{s}^{l} kMg(t)\,dt\right) ds \\
&\quad \geqq \int_{l-2\tau}^{l-\tau} a(s)x(s)\left(\int_{l-\tau}^{s+\tau} kM\left(\int_{t_2}^{l} kMg(t_1)\,dt_1\right) dt_2\right) ds \\
&\quad\quad + \int_{l-\tau}^{l} a(s)x(s)\left(\int_{s}^{l} kM\left(\int_{t_2}^{l} kMg(t_1)\,dt_1\right) dt_2\right) ds.
\end{aligned}$$

With repeated application to the integral over $[l-\tau, l]$, it follows that

$$\begin{aligned}
&\int_{l-\tau}^{l} a(s)x(s)g(s)\,ds \\
&\quad \geqq \int_{l-2\tau}^{l-\tau} a(s)x(s)\left(\int_{l-\tau}^{s+\tau} kMg(t_1)\,dt_1 + \sum_{j=2}^{\infty} (kM)^j \int_{l-\tau}^{s+\tau}\int_{t_j}^{l}\cdots\int_{t_2}^{l} g(t_1)\,dt_1\cdots dt_j\right) ds.
\end{aligned}$$

For $j \geqq 2$,

$$\begin{aligned}
&\int_{l-\tau}^{s+\tau} \int_{t_j}^{l} \cdots \int_{t_2}^{l} g(t_1)\, dt_1 \cdots dt_j \\
&\quad = \int_{l-\tau}^{s+\tau} \int_{t_j}^{l} g(t_1) \int_{t_j}^{t_1} \int_{t_j}^{t_2} \cdots \int_{t_j}^{t_{j-2}} dt_{j-1} \cdots dt_2\, dt_1\, dt_j \\
&\quad = \int_{l-\tau}^{s+\tau} \int_{t_j}^{l} g(t_1) \frac{(t_1 - t_j)^{j-2}}{(j-2)!}\, dt_1\, dt_j \\
&\quad = \int_{l-\tau}^{s+\tau} g(t) \left(\int_{l-\tau}^{t} \frac{(t-u)^{j-2}}{(j-2)!}\, du \right) dt + \int_{s+\tau}^{l} g(t) \left(\int_{l-\tau}^{s+\tau} \frac{(t-u)^{j-2}}{(j-2)!}\, du \right) dt \\
&\quad = \int_{l-\tau}^{s+\tau} g(t) \frac{(t-l+\tau)^{j-1}}{(j-1)!}\, dt + \int_{s+\tau}^{l} g(t) \left(\frac{(t-l+\tau)^{j-1}}{(j-1)!} - \frac{(t-s-\tau)^{j-1}}{(j-1)!} \right) dt \\
&\quad = \int_{l-\tau}^{l} g(t) \frac{(t-l+\tau)^{j-1}}{(j-1)!}\, dt - \int_{s+\tau}^{l} g(t) \frac{(t-s-\tau)^{j-1}}{(j-1)!}\, dt.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\int_{l-\tau}^{l} a(s)x(s)g(s)\, ds \\
&\quad \geqq \int_{l-2\tau}^{l-\tau} a(s)x(s) \left(\int_{l-\tau}^{l} g(t) \sum_{j=1}^{\infty} (kM)^j \frac{(t-l+\tau)^{j-1}}{(j-1)!} \right. \\
&\qquad \left. - \int_{s+\tau}^{l} g(t) \sum_{j=1}^{\infty} (kM)^j \frac{(t-s-\tau)^{j-1}}{(j-1)!}\, dt \right) ds.
\end{aligned}$$

Hence,

$$\begin{aligned}
\int_{l-\tau}^{l} a(s)x(s)g(s)\, ds &\geqq \int_{l-2\tau}^{l-\tau} a(s)x(s) \left(\int_{l-\tau}^{l} g(t) kM e^{kM(t-l+\tau)}\, dt \right. \\
&\qquad \left. - \int_{s+\tau}^{l} g(t) kM e^{kM(t-s-\tau)}\, dt \right) ds,
\end{aligned} \tag{2.3}$$

and by replacing $\int_{s+\tau}^{l} g(t) kM e^{kM(t-s-\tau)}\, dt$ with the larger quantity $\int_{l-\tau}^{l} g(t) kM e^{kM(t-s-\tau)}\, dt$, it follows that

$$\begin{aligned}
&\int_{l-\tau}^{l} a(s)x(s)g(s)\, ds \\
&\quad \geqq \int_{l-2\tau}^{l-\tau} a(s)x(s) \left(\int_{l-\tau}^{l} g(t) kM e^{kM(t-l+\tau)} (1 - e^{kM(l-s-2\tau)})\, dt \right) ds.
\end{aligned} \tag{2.4}$$

Application of inequality (2.4) to the integral on $[b-\tau, b]$ in (2.2) yields

$$\begin{aligned}
&\int_{b-\tau}^{b} a(s)x(s)\left(\int_{b}^{s+\tau} ka(t)\,dt\right) ds \\
&\quad \geqq \int_{d_2}^{d_1} a(s)x(s)(kM_1(s+\tau-b))\,ds \\
&\quad \geqq \int_{d_3}^{d_2} a(s)x(s)\left(\int_{d_2}^{d_1} kM_1(t-d_2)kM_2\, e^{kM_2(t-d_2)}(1-e^{kM_2(d_3-s)})\,dt\right) ds \\
&\quad = \int_{d_3}^{d_2} a(s)x(s)(1-e^{kM_2(d_3-s)})\,ds\left(kM_1\tau\, e^{kM_2\tau} - \frac{M_1}{M_2}e^{kM_2\tau} + \frac{M_1}{M_2}\right).
\end{aligned}$$

Also, for any j,

$$\begin{aligned}
&\int_{d_{j+1}}^{d_j} a(s)x(s)(1-e^{kM_j(d_{j+1}-s)})\,ds \\
&\quad \geqq \int_{d_{j+2}}^{d_{j+1}} a(s)x(s)\left(\int_{d_{j+1}}^{d_j} (1-e^{kM_j(d_{j+1}-t)})kM_{j+1}\, e^{kM_{j+1}(t-d_{j+1})}\right. \\
&\qquad\qquad \left. \cdot\,(1-e^{kM_{j+1}(d_{j+2}-s)})\,dt\right) ds \\
&\quad \geqq \int_{d_{j+2}}^{d_{j+1}} a(s)x(s)(1-e^{kM_{j+1}(d_{j+2}-s)})\,ds \cdot P(kM_{j+1}, kM_j).
\end{aligned}$$

Using the ω-periodicity of a and x gives

$$\begin{aligned}
&\int_{d_{N+1}}^{d_N} a(s)x(s)(1-e^{kM_N(d_{N+1}-s)})\,ds \\
&\qquad = \int_{c-\omega}^{c+\tau-\omega} a(s)x(s)(1-e^{kM_N(c-\omega-s)})\,ds \\
&\qquad = \int_{c}^{c+\tau} a(s-\omega)x(s-\omega)(1-e^{kM_N(c-s)})\,ds \\
&\qquad = \int_{c}^{c+\tau} a(s)x(s)(1-e^{kM_N(c-s)})\,ds.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\int_{b-\tau}^{b} a(s)x(s)\left(\int_{b}^{s+\tau} ka(t)\,dt\right) ds \\
&\quad \geqq \int_{c}^{c+\tau} a(s)x(s)\left((1-e^{kM_N(c-s)})\left(kM_1\tau\, e^{kM_2\tau} - \frac{M_1}{M_2}e^{kM_2\tau} + \frac{M_1}{M_2}\right)\right. \\
&\qquad\qquad \left. \cdot \prod_{j=2}^{N-1} P(kM_{j+1}, kM_j)\right) ds.
\end{aligned}$$

Since $a(t) \geqq M_{N+1}$ on $[c, c+\tau]$, it follows from (2.2) that

$$\begin{aligned}
&\int_b^{c+\tau} a(s)x(s)\,ds \\
&\quad \geqq \int_b^c a(s)x(s)\left(\int_s^{s+\tau} ka(t)\,dt\right) ds \\
&\quad\quad + \int_c^{c+\tau} a(s)x(s)\Big(kM_{N+1}(c+\tau-s) + (1-e^{kM_N(c-s)}) \\
&\quad\quad\quad \cdot \left(kM_1\tau e^{kM_2\tau} - \frac{M_1}{M_2}e^{kM_2\tau} + \frac{M_1}{M_2}\right)\prod_{j=2}^{N-1} P(kM_{j+1}, kM_j)\Big)\,ds.
\end{aligned} \tag{2.5}$$

By assumption, $\int_s^{s+\tau} ka(t)\,dt > 1$ for $b \leqq s \leqq c$. Also, the function

$$kM_{N+1}(c+\tau-s) + (1-e^{kM_N(c-s)})\left(kM_1\tau e^{kM_2\tau} - \frac{M_1}{M_2}e^{kM_2\tau} + \frac{M_1}{M_2}\right)\prod_{j=2}^{N-1} P(kM_{j+1}, kM_j)$$

has only one critical value, and since that value is a maximum, the function attains its minimum on $[c, c+\tau]$ at one of the endpoints of the interval. When $s = c$, we obtain $kM_{N+1}\tau$ and when $s = c+\tau$, the value is

$$(1-e^{-kM_N\tau})\left(kM_1\tau e^{kM_2\tau} - \frac{M_1}{M_2}e^{kM_2\tau} + \frac{M_1}{M_2}\right)\prod_{j=2}^{N-1} P(kM_{j+1}, kM_j).$$

Both these values are greater than 1. This leads to a contradiction in (2.5) unless $\int_b^{c+\tau} a(s)x(s)\,ds = 0$. If $\int_b^{c+\tau} a(s)x(s)\,ds = 0$, it follows that $x \equiv 0$ on $[b, c+\tau]$. Then $0 = x(b) \geqq Ax(b) = \int_{b-\tau}^b f(s, x(s))\,ds \geqq \int_{b-\tau}^b ka(s)x(s)\,ds$, so that $x \equiv 0$ on $[b-\tau, b]$. A continuation of this argument leads to the conclusion that x is the zero function, which contradicts the original assumption. Hence, A satisfies the conditions of Theorem 3 and has a nonzero fixed point in K. □

Proof of Theorem 2. Using an argument analogous to that in the proof of Theorem 1 with inequality (2.3) substituted for (2.4), one obtains a nonzero solution to (1.1) provided

$$\int_s^{s+\tau} a(t)\,dt > 1, \qquad b \leqq s \leqq c,$$

and

$$\int_s^{c+\tau} a(t)\,dt + g_N(s-\omega) > 1, \qquad c \leqq s \leqq c+\tau, \tag{2.6}$$

where g_N is defined in the statement of Theorem 2. The desired conclusion then follows from the inequality

$$\int_s^{c+\tau} a(t)\,dt \geqq M_{N+1}(c+\tau-s), \qquad c \leqq s \leqq c+\tau.$$ □

In Theorems 1 and 2, $a(t)$ is required to exceed $1/\tau$ for some values of t. If $f(t, x) < a(t)x$ for $x > 0$ (as is the case for the function $f(t, x) = a(t)x(1-x)$), this condition is, in fact, necessary, for if $a(t) \leqq 1/\tau$ for each t, it is easy to see that (1.1) has only the zero solution.

Theorems 1 and 2 give conditions which insure that the spectral radius of L_τ is larger than one; however, better estimates of the spectral radius can be obtained by a slight change in the proofs. Suppose x is a nonzero periodic function, $\lambda > 0$, and $\lambda x(t) = L_\tau x(t) = \int_{t-\tau}^{t} a(s)x(s)\,ds$. By repeating the argument in Theorem 1 for the operator $(1/\lambda)L_\tau$ instead of the operator A, one obtains a contradiction if

(i) $$\int_s^{s+\tau} \frac{1}{\lambda} a(t)\,dt > 1, \qquad b \leqq s \leqq c,$$

(ii) $$\frac{1}{\lambda} M_{N+1}\tau > 1,$$

(iii) $$(1 - e^{-M_N\tau/\lambda})\left(\frac{M_1\tau}{\lambda} e^{M_2\tau/\lambda} - \frac{M_1}{M_2} e^{M_2\tau/\lambda} + \frac{M_1}{M_2}\right) \prod_{j=2}^{N-1} P\left(\frac{M_{j+1}}{\lambda}, \frac{M_j}{\lambda}\right) > 1.$$

Hence, if λ^* is any number for which (i), (ii) and (iii) above hold, then $r(L_\tau) > \lambda^*$.

Upper estimates for $r(L_\tau)$ can be obtained by replacing the minimum of $a(t)$ on each interval with the corresponding maximum and using the inequalities in Theorem 2. Specifically, suppose $U_j = \sup\{a(t): d_j \leqq t \leqq d_{j-1}\}$ and λ^* satisfies

(iv) $$\int_s^{s+\tau} \frac{1}{\lambda^*} a(t)\,dt < 1, \qquad b \leqq s \leqq c,$$

and

(v) $$\frac{1}{\lambda^*} U_{N+1}(c + \tau - s) + g_N(s - \omega) < 1, \qquad c \leqq s \leqq c + \tau,$$

where $g_1(s) = (1/\lambda^*)M_1(s - d_2)$, and for $j \geqq 1$,

$$g_{j+1}(s) = \int_{d_{j+1}}^{d_j} g_j(t) \frac{1}{\lambda^*} U_{j+1} \exp\left(\frac{1}{\lambda^*} U_{j+1}(t - d_{j+1})\right) dt$$
$$- \int_{s+\tau}^{d_j} g_j(t) \frac{1}{\lambda^*} U_{j+1} \exp\left(\frac{1}{\lambda^*} U_{j+1}(t - s - \tau)\right) dt.$$

Then $r(L_\tau) < \lambda^*$.

REFERENCES

[1] T. B. BENJAMIN, *A unified theory of conjugate flows*, Philos. Trans. Roy. Soc. London, 269 (1971), pp. 587–647.

[2] K. L. COOKE AND J. L. KAPLAN, *A periodicity threshold theorem for epidemics and population growth*, Math. Biosci., 31 (1976), pp. 87–104.

[3] M. A. KRASNOSEL'SKII, *Positive Solutions of Operator Equations*, Noordhoff, Groningen, 1964.

[4] R. W. LEGGETT AND L. R. WILLIAMS, *A fixed point theorem with application to an infectious disease model*, J. Math. Anal. and Appl., 76 (1980), pp. 91–97.

[5] W. P. LONDON AND J. A. YORKE, *Recurrent outbreaks of measles, chickenpox and mumps I. (Seasonal variation in contact rates)*, Amer. J. Epid., 98 (1973), pp. 453–468.

[6] R. NUSSBAUM, *A periodicity threshold theorem for some nonlinear integral equations*, this Journal, 9 (1978), pp. 356–376.

[7] H. L. SMITH, *On periodic solutions of a delay integral equation modelling epidemics*, J. Math. Biol., 4 (1977), pp. 69–80.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0009 $01.00/0

UNIQUE AND MULTIPLE SOLUTIONS OF A FAMILY OF DIFFERENTIAL EQUATIONS MODELING CHEMICAL REACTIONS*

LYNN R. WILLIAMS† AND RICHARD W. LEGGETT‡

Abstract. Uniqueness and multiplicity of solutions are studied for the boundary value problem

$$\beta x''(t) - x'(t) + pf(x(t)) = 0, \qquad 0 \leqq t \leqq 1, \quad \beta > 0, \quad p > 0,$$

$$\beta x'(0) - x(0) = 0, \quad x'(1) = 0,$$

which arises in chemical reactor theory. The "reaction rate" f is given by

$$f(x) = (q - x) \exp[-k/(1+x)], \qquad k > 0, \quad q > 0.$$

Uniqueness is shown for (1) sufficiently small p, (2) certain regions of p and sufficiently small β, (3) large p and sufficiently large β and (4) fixed β and sufficiently large p. Regions of points (β, p, q, k) are identified where there are at least three solutions. The combination of these results gives an improved picture of the behavior of the number of solutions as p and β vary.

1. Introduction.

The boundary value problem

$$\beta x''(t) - x'(t) + pf(x(t)) = 0, \qquad 0 \leqq t \leqq 1, \quad \beta > 0, \quad p > 0, \tag{1.1}$$

$$\beta x'(0) - x(0) = 0, \qquad x'(1) = 0 \tag{1.2}$$

arises in chemical reactor theory and describes steady-state reactor concentration and reaction temperature along a one-dimensional adiabatic dispersed-plug flow tubular reactor. The function f is the Arrhenius reaction rate given by

$$f(x) = (q - x) \exp[-k/(1+x)], \qquad k > 0, \quad q > 0. \tag{1.3}$$

It is known from experimental results that reactions modeled by (1.1)–(1.3) may exhibit either unique or multiple steady states, depending on the constants β, p, q and k. Cohen [4] has shown that the bvp (1.1)–(1.3) has a unique solution whenever $k \leqq 4 + 4/q$. The existence of at least three solutions to (1.1)–(1.3) in the case $k > 4 + 4/q$ has been suggested for some values of p and β by heuristic arguments [4], numerical methods [3], and results of Amann [1], and has been demonstrated rigorously by Leggett and Williams [5], [7] for an interval of values of p and numbers β greater than some β_p.

The purpose of this paper is to expand upon known results concerning the number of solutions of (1.1)–(1.3). In particular, we extend the uniqueness result of Cohen to the case $k > 4 + 4/q$ by showing uniqueness for (1) sufficiently small p, (2) certain regions of p and sufficiently small β, (3) large p and sufficiently large β, and (4) fixed β and sufficiently large p. Additionally, we describe a method (based on results in [5] for abstract ordered Banach spaces) for identifying regions of points (β, p, q, k) where (1.1)–(1.3) has at least three solutions. The combination of these results gives an improved picture of the behavior of the number of solutions of (1.1)–(1.3) as p and β vary. This is illustrated for the special case $q = 1.1$, $k = 10$.

* Received by the editors October 1, 1980.

† Department of Mathematics, Indiana University at South Bend, South Bend, Indiana 46615. The research of this author was partially supported by a Summer Faculty Fellowship from Indiana University at South Bend.

‡ Health and Safety Research Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830. The research of this author was supported by Union Carbide Corporation under contract W-7405-eng-26 with the U.S. Department of Energy.

We have attempted to make this article as complete and self-contained as was practical. In particular, a proof of Cohen's uniqueness result is presented, and a proof of a suitably modified version of the abstract result [5] needed for the multiple solutions results is included.

Other uniqueness conditions for the same general type of chemical reaction problem considered here are given in [6]. However, the heat generation function considered in [6] is slightly different from the function f considered here, and the methods of [6] are different from ours.

2. Uniqueness. It follows from the strong maximum principle (see also [5]) that every solution to (1.1)–(1.3) satisfies $0 \leqq x(t) \leqq q$, $0 \leqq t \leqq 1$. Furthermore the solutions to (1.1)–(1.3) are precisely the fixed points of the completely continuous operator A defined on $C([0, 1])$ by

$$Ax(t) = \int_0^1 G(t, s) p f(x(s))\, ds, \tag{2.1}$$

where

$$G(t, s) = \begin{cases} e^{(t-s)/\beta}, & 0 \leqq t \leqq s \leqq 1, \\ 1, & 0 \leqq s < t \leqq 1. \end{cases}$$

Since $G(t, s)$ is an increasing function of t for fixed s, it follows that any solution to (1.1)–(1.3) is strictly increasing. Further, Cohen [4] established the existence of minimal and maximal solutions to (1.1)–(1.3). The uniqueness result of Cohen follows readily from this result and the following lemma.

LEMMA 2.1. *If u and v are solutions to* (1.1)–(1.3), *a and b are real numbers such that $a \leqq u(t) \leqq v(t) \leqq b$, $0 \leqq t \leqq 1$, and $f(x)/x$ decreases on (a, b), then $u = v$.*

Proof. Assume $u \neq v$. Then $f(u(s)) \geqq (f(v(s))/v(s))u(s)$, $0 \leqq s \leqq 1$, with strict inequality holding for some s. Let $r = \min_{0 \leqq s \leqq 1} u(s)/v(s)$ and choose t_0 so that $rv(t_0) = u(t_0)$. Then

$$\begin{aligned} rv(t_0) = u(t_0) &= \int_0^1 G(t_0, s) p f(u(s))\, ds \\ &> \int_0^1 G(t_0, s) p f(v(s)) \frac{u(s)}{v(s)}\, ds \\ &\geqq r \int_0^1 G(t_0, s) p f(v(s))\, ds \\ &= rv(t_0). \end{aligned}$$

This contradiction proves Lemma 2.1. □

THEOREM 2.2. (Cohen). *If $k \leqq 4 + 4/q$, then* (1.1)–(1.3) *has exactly one solution for each value of p and β.*

Proof. If $k \leqq 4 + 4/q$, then $f(x)/x$ decreases on $(0, \infty)$. □

The cases of interest occur, therefore, when $k > 4 + 4/q$. Then $f(x)/x$ decreases in $(0, r_1]$, where

$$r_1 \equiv \frac{1}{2k + 2q} (kq - 2q - (kq(kq - 4q - 4))^{1/2}), \tag{2.2}$$

increases in $[r_1, r_2]$, where

$$r_2 \equiv \frac{1}{2k+2q}(kq-2q+(kq(kq-4q-4))^{1/2}), \tag{2.3}$$

and decreases in $[r_2, \infty)$. Also, f has a maximum at

$$r_0 \equiv \tfrac{1}{2}((4k+k^2+4kq)^{1/2}-2-k), \tag{2.4}$$

f increases in $[0, r_0]$, f decreases in $[r_0, \infty)$, and $r_1 < r_2 < r_0$.

Placing restrictions on p and β forces solutions to (1.1)–(1.3) to lie in intervals where $f(x)/x$ decreases. Thus, using Lemma 2.1, we extend the uniqueness result of Cohen to the case $k > 4+4/q$. The following lemma will be used to obtain some of these uniqueness results.

LEMMA 2.3. *Suppose p and β are fixed and x is a solution to* (1.1)–(1.3) *such that either $x(1) \leqq r_0$ or $pf(r_0) \leqq r_0$ (r_0 is given in* (2.4)). *Then $pf(x(1)) > x(1)$.*

Proof. Assume $pf(x(1)) \leqq x(1)$. If $x(1) \leqq r_0$, then $pf(x(t)) < pf(x(1)) \leqq x(1)$, $0 \leqq t < 1$. On the other hand, if $x(1) > r_0$ and $pf(r_0) \leqq r_0$, then $pf(x(t)) \leqq pf(r_0) \leqq r_0 < x(1)$, $0 \leqq t \leqq 1$. Thus, in either case, $x(1) = \int_0^1 pf(x(t))\,dt < \int_0^1 x(1)\,dt = x(1)$, a contradiction. □

From this result, the uniqueness of the solution for small p follows easily.

THEOREM 2.4. *If $k > 4+4/q$ and $p \leqq r_2/f(r_2)$, then there is only one solution to* (1.1)–(1.3) *for each $\beta > 0$ (r_2 is defined in* (2.3)).

Proof. If $p \leqq r_2/f(r_2)$, then $pf(r_0) < r_0$ since $f(x)/x$ decreases in $[r_2, \infty)$. Thus, if x is any solution to (1.1)–(1.3), $pf(x(1)) > x(1)$ by Lemma 2.3. The maximum of $f(y)/y$ on $[r_1, \infty)$ occurs when $y = r_2$ and $pf(r_2) \leqq r_2$. It follows that $x(1) < r_1$. Therefore, by Lemma 2.1, there is only one solution, since $f(x)/x$ decreases in $(0, r_1]$. □

The next theorem shows that for fixed p, $p > r_1/f(r_1)$, there is a unique solution for all large β.

THEOREM 2.5. *Suppose $k > 4+4/q$, β_0 is the solution to the equation*

$$\frac{r_1}{f(r_1)} = \frac{r_2}{f(r_2\, e^{1/\beta_0})},$$

and

$$p_0 = \frac{r_1}{\beta_0(1-e^{-1/\beta_0})f(r_1)}.$$

Then the boundary value problem (1.1)–(1.3) *has a unique solution provided*

$$\text{(i)} \qquad \frac{r_1}{f(r_1)} < p < p_0 \quad \text{and} \quad \beta(1-e^{-1/\beta}) \geqq \frac{r_1}{pf(r_1)}$$

or

$$\text{(ii)} \qquad p_0 \leqq p \quad \text{and} \quad \beta(1-e^{-1/\beta})f(r_2 e^{1/\beta}) \geqq \frac{r_2}{p}.$$

Note. The expression $\beta(1-e^{-1/\beta})$ is an increasing function of β on $(0, \infty)$ with range $(0, 1)$. Thus if condition (i) is satisfied for β_1, it is satisfied for $\beta \geqq \beta_1$. The same is true for condition (ii), although this fact is certainly not as obvious. Suppose condition (ii) holds for β_1 and p and suppose $\beta \geqq \beta_1$. Since $r_1/f(r_1) > r_2/f(r_2)$, it follows that $f(r_2) > f(r_2\, e^{1/\beta_0})$. Thus if $\beta_1 \leqq \beta_0$, then

$$\beta(1-e^{-1/\beta})\, f(r_2\, e^{1/\beta}) \geqq \beta_1(1-e^{-1/\beta_1})f(r_2\, e^{1/\beta_1}).$$

On the other hand, if $\beta_1 > \beta_0$, then $\beta > \beta_0$ and

$$\frac{r_2}{pf(r_2 e^{1/\beta})} < \frac{r_2}{pf(r_2 e^{1/\beta_0})} = \frac{r_1}{pf(r_1)} \leqq \frac{r_1}{p_0 f(r_1)} = \beta_0(1-e^{-1/\beta_0}) \leqq \beta(1-e^{-1/\beta}).$$

Thus, in either case, condition (ii) holds for β.

Proof of Theorem 2.5. We will show that if condition (i) or condition (ii) holds and x is a solution to (1.1)–(1.3), then $x(0) \geqq r_2$. The theorem will then follow from Lemma 2.1, since $f(x)/x$ decreases on $[r_2, \infty)$.

Suppose p and β satisfy condition (i). Then

$$\beta(1-e^{-1/\beta}) \geqq \frac{r_1}{pf(r_1)} \geqq \frac{r_1}{p_0 f(r_1)} = \beta_0(1-e^{-1/\beta_0}),$$

so that $\beta \geqq \beta_0$ and $f(r_2 e^{1/\beta}) \geqq f(r_2 e^{1/\beta_0})$. Therefore,

$$\beta(1-e^{-1/\beta}) f(r_2 e^{1/\beta}) \geqq \frac{r_1}{pf(r_1)} f(r_2 e^{1/\beta})$$

$$= \frac{r_2}{pf(r_2 e^{1/\beta_0})} f(r_2 e^{1/\beta}) \geqq \frac{r_2}{p}.$$

Now suppose p and β satisfy condition (ii). If $\beta \geqq \beta_0$, then

$$\beta(1-e^{-1/\beta}) \geqq \beta_0(1-e^{-1/\beta_0}) = \frac{r_1}{p_0 f(r_1)} \geqq \frac{r_1}{pf(r_1)},$$

and if $\beta < \beta_0$, then

$$\beta(1-e^{-1/\beta}) \geqq \frac{r_2}{pf(r_2 e^{1/\beta})} \geqq \frac{r_2}{pf(r_2 e^{1/\beta_0})} = \frac{r_1}{pf(r_1)}.$$

Hence, if p and β satisfy either condition (i) or condition (ii), then

$$\beta(1-e^{-1/\beta}) \geqq \max\left\{\frac{r_1}{pf(r_1)}, \frac{r_2}{pf(r_2 e^{1/\beta})}\right\}.$$

Assume x is a solution to (1.1)–(1.3) and $x(0) < r_2$. Then,

$$x(1) = \int_0^1 pf(x(s))\, ds \leqq e^{1/\beta} \int_0^1 e^{-s/\beta} pf(x(s))\, ds = e^{1/\beta} x(0) < e^{1/\beta} r_2$$

and

$$x(0) = \int_0^1 e^{-s/\beta} pf(x(s))\, ds > \min_{0 \leqq t \leqq 1} \{pf(x(t))\} \int_0^1 e^{-s/\beta}\, ds$$

$$= \min\{pf(x(0)), pf(x(1))\}\beta(1-e^{-1/\beta}).$$

Thus, if $f(x(0)) \leqq f(x(1))$, then

$$x(0) > pf(x(0))\beta(1-e^{-1/\beta}) = \frac{pf(x(0))}{x(0)} \beta(1-e^{-1/\beta}) x(0)$$

$$\geqq \frac{pf(r_1)}{r_1} \beta(1-e^{-1/\beta}) x(0) \geqq x(0),$$

a contradiction. However, if $f(x(1)) < f(x(0))$, then $x(1) > r_0$ and

$$x(0) > pf(x(1))\beta(1-e^{-1/\beta}) \geqq pf(e^{1/\beta}r_2)\beta(1-e^{-1/\beta}) \geqq r_2,$$

also a contradiction. Thus $x(0) \geqq r_2$ and the desired conclusion follows.

The next result establishes the existence of only one solution for fixed β provided p is sufficiently large.

THEOREM 2.6. *Suppose $k > 4 + 4/q$ and β_1 is the solution to the equation*

$$qe^{-k} = (q - r_2 e^{1/\beta_1}) \quad \exp\left(\frac{-k}{1+r_2 e^{1/\beta_1}}\right).$$

Then (1.1)–(1.3) *has a unique solution, provided*

(i) $\beta \geqq \beta_1$ *and* $p \geqq e^k$,

or

(ii) $\beta < \beta_1$ *and* $p \geqq \beta_1 e^k/\beta$.

Proof. Suppose x is a solution to (1.1)–(1.3) and $x(0) < r_2$. Then

$$x(1) = \int_0^1 pf(x(s))\, ds < e^{1/\beta}\int_0^1 e^{-s/\beta} pf(x(s))\, ds = e^{1/\beta}x(0) \leqq e^{1/\beta}r_2.$$

If condition (i) is satisfied, then $\beta \geqq \beta_1$ so that $f(x(1)) \geqq f(r_2 e^{1/\beta_1}) = f(0)$. Clearly, $f(x(0)) \geqq f(0)$ so that

$$x(1) = \int_0^1 pf(x(s))\, ds > p \min\{f(x(0)), f(x(1))\}$$

$$\geqq pf(0)) = pq e^{-k} \geqq q,$$

a contradiction. If condition (ii) is satisfied, let $t = \beta/\beta_1$. Then $0 < t < 1$ and

$$x(1) = \int_0^1 pf(x(s))\, ds > \int_0^t pf(x(s))\, ds$$

$$\geqq tp \min\{f(x(0)), f(x(t))\}.$$

But

$$x(t) = \int_0^t pf(x(s))\, ds + \int_t^1 e^{(t-s)/\beta} pf(x(s))\, ds$$

$$< \int_0^1 e^{(t-s)/\beta} pf(x(s))\, ds = e^{t/\beta}x(0) = e^{1/\beta_1}x(0) < e^{1/\beta_1}r_2.$$

Thus, $f(x(t)) > f(e^{1/\beta_1}r_2) = f(0)$. Also, $f(x(0)) \geqq f(0)$, so that $x(1) > tpf(0) = (\beta/\beta_1)pq e^{-k} \geqq q$, another contradiction. Therefore we must have $x(0) \geqq r_2$, and the desired conclusion follows from Lemma 2.1. □

The final uniqueness result gives uniqueness for small β when $p < r_1/f(r_1)$.

THEOREM 2.7. *If $k < 4 + 4/q$, $p < r_1/f(r_1)$, $\beta(1-e^{-1/\beta}) < r_1/pf(r_0)$ and*

$$\beta \leqq \frac{r_1 - pf(r_1)}{pf(r_0) + pf(r_1)[-1 + \ln(f(r_1)/f(r_0))]},$$

then (1.1)–(1.3) *has exactly one solution.*

Proof. Suppose x is a solution to (1.1)–(1.3). Then

$$x(0) = \int_0^1 e^{-s/\beta} f(x(s))\, ds < pf(r_0) \int_0^1 e^{-s/\beta}\, ds$$
$$= pf(r_0)\beta(1 - e^{-1/\beta}) \leqq r_1.$$

If we show $x(1) \leqq r_1$ as well, then the result will follow from Lemma 2.1.

Suppose there exists $t \in (0, 1]$ such that $x(t) = r_1$. Then

$$r_1 = \int_0^t pf(x(s))\, ds + \int_t^1 e^{(t-s)/\beta} pf(x(s))\, ds$$
$$< tpf(r_1) + pf(r_0) \int_t^1 e^{(t-s)/\beta}\, ds$$
$$= tpf(r_1) + pf(r_0)\beta(1 - e^{(t-1)/\beta}) \equiv g(t).$$

Now $g'(t) = pf(r_1) - pf(r_0)\, e^{(t-1)/\beta}$, so that the maximum value of g occurs when

$$t = 1 + \beta \ln \frac{f(r_1)}{f(r_0)}$$

and

$$g\left(1 + \beta \ln \frac{f(r_1)}{f(r_0)}\right) = \left(1 + \beta \ln \frac{f(r_1)}{f(r_0)}\right) pf(r_1) + pf(r_0)\beta\left(1 - \frac{f(r_1)}{f(r_0)}\right)$$
$$= pf(r_1) + \beta\left(pf(r_1) \ln \frac{f(r_1)}{f(r_0)} + pf(r_0) - pf(r_1)\right)$$
$$\leqq pf(r_1) + r_1 - pf(r_1) = r_1,$$

a contradiction. Thus, $x(t) < r_1$ for each t. □

3.1. Multiple solutions. Our basic multiple solutions result is a special case of a theorem concerning completely continuous operators on ordered Banach spaces [5, Thm. 3.4]. However, for the sake of completeness, a short sketch of the basic ideas in the proof will be given.

THEOREM 3.1. *Suppose $k < 4 + 4/q$ and $r_2/f(r_2) < p \leqq r_1/f(r_1)$. Choose $a \in (r_1, r_2]$ and let b be the number in (r_0, q) such that $f(a) = f(b)$. If there exists $\gamma \in (0, 1]$ such that β satisfies the inequality*

$$\beta(1 - e^{-\gamma/\beta}) \geqq \max\left\{\frac{a}{pf(a)} - (1 - \gamma)\frac{f(0)}{f(a)},\ \gamma - \frac{b - a}{pf(r_0)}\right\}, \tag{3.1}$$

then (1.1)–(1.3) *has at least three solutions.*

Proof. For $x \in C([0, 1])$, let $\alpha(x) = \min_{1-\gamma \leqq t \leqq 1} x(t)$, and consider the sets of continuous functions on $[0, 1]$ defined by

$$U_1 = \{x : 0 \leqq x(t) < r_1,\ 0 \leqq t \leqq 1\},$$
$$U_2 = \{x : 0 \leqq x(t) \leqq pf(r_0),\ 0 \leqq t \leqq 1,\ \alpha(x) > a\}$$

and

$$U = \{x : 0 \leqq x(t) \leqq pf(r_0),\ 0 \leqq t \leqq 1\}.$$

The operator $A_0x(t)$ given by

$$A_0x(t) = \int_0^1 G(t, s)pf_0(x(s))\, ds,$$

where

$$f_0(x) = \begin{cases} f(x), & x \leqq q, \\ 0\ , & x \geqq q \end{cases}$$

is a completely continuous map which leaves U and U_1 invariant and has the same set of fixed points as A defined in (2.1) (see [5, Example 2.2]). Further, A_0 has no fixed points on the boundary of U_1 in U, so that the fixed point index $i(A_0, \cdot, U)$ is defined for U_1 and U, and $i(A_0, U_1, U) = i(A_0, U, U) = 1$. (See [2] for properties of the fixed point index.)

If x is an increasing function such that $0 \leqq x(t) \leqq b$, $t \in [0, 1]$ and $\alpha(x) \geqq a$, then

$$\begin{aligned} \alpha(A_0x) &= A_0x(1-\gamma) \\ &= \int_0^{1-\gamma} pf_0(x(s))\, ds + \int_{1-\gamma}^1 e^{(1-\gamma-s)/\beta} pf_0(x(s))\, ds \\ &> (1-\gamma)pf(0) + pf(a)\int_{1-\gamma}^1 e^{(1-\gamma-s)/\beta}\, ds \\ &= (1-\gamma)pf(0) + pf(a)\beta(1-e^{-\gamma/\beta}) \geqq a. \end{aligned}$$

Also, if $x \in U$ and $A_0x(1) > b$, then

$$\begin{aligned} \alpha(A_0x) &= A_0x(1-\gamma) \\ &= \int_0^{1-\gamma} pf_0(x(s))\, ds + \int_{1-\gamma}^1 e^{(1-\gamma-s)/\beta} pf_0(x(s))\, ds \\ &= \int_0^1 pf_0(x(s))\, ds - \int_{1-\gamma}^1 (1-e^{(1-\gamma-s)/\beta}) pf_0(x(s))\, ds \\ &\geqq A_0x(1) - pf(r_0)\int_{1-\gamma}^1 (1-e^{(1-\gamma-s)/\beta})\, ds \\ &> b - pf(r_0)(\gamma - \beta(1-e^{-\gamma/\beta})) \geqq a. \end{aligned}$$

It follows from the above that A_0 has no fixed points in the boundary of U_2 in U. For if $A_0x = x$ and $x(1) \leqq b$, then, since x is increasing, we have $\alpha(A_0x) > a$. On the other hand, if $A_0x = x$ and $x(1) > b$ then $A_0x(1) > b$ and $\alpha(A_0x) > a$. Thus, the fixed point index $i(A_0, U_2, U)$ is also defined. Let x_0 be an increasing element of U_2 with $x_0(1) \leqq b$ and define $h\colon [0, 1] \times \bar{U}_2 \to U$ by

$$h(t, x) = (1-t)A_0x + tx_0.$$

Then h is continuous with compact range. Suppose $h(t, x) = x$ for some $x \in \partial U_2$. If $A_0x(1) > b$ then $\alpha(A_0x) > a$, so that $\alpha(x) = \alpha((1-t)A_0x + tx_0) > (1-t)a + ta = a$. Also, if $A_0x(1) \leqq b$, then

$$x(1) = (1-t)A_0x(1) + tx_0(1) \leqq b,$$

and since x is increasing, $\alpha(A_0x) > a$, and $\alpha(x) = \alpha((1-t)A_0x + tx_0) > a$. Thus, $i(h(t, \cdot),$

$U_2, U)$ is defined for each t and is independent of t. Therefore

$$i(A_0, U_2, U) = i(h(1, \cdot), U_2, U) = 1.$$

By the additivity properties of the fixed point index, it follows that $i(A, U\backslash(U_1 \cup U_2), U) = -1$. Therefore the index is nonzero on U_1, U_2 and $U\backslash(U_1 \cup U_2)$ and it follows that A_0 (and hence A) has a fixed point in each of these sets. □

3.2. Determination of the smallest β satisfying (3.1). The expression $\beta(1-e^{-\gamma/\beta})$ defines an increasing function of β with limit γ at $+\infty$. Hence, if (3.1) holds for $\beta = \beta_1$ and γ, then it holds for all $\beta \geqq \beta_1$. If $a/pf(a) \geqq 1$, then

$$\frac{a}{pf(a)} - (1-\gamma)\frac{f(0)}{f(a)} \geqq 1 - (1-\gamma) = \gamma,$$

so that (3.1) will not hold for any β. Let a_p be the number in $(r_1, r_2]$ such that $a_p/pf(a_p) = 1$, choose $a \in (a_p, r_2]$, and let $b \in [r_0, q]$ be such that $f(a) = f(b)$. We will determine the smallest value $\beta(a, p)$ of β satisfying (3.1) for some $\gamma \in (0, 1]$.

Let

$$P_0 = P_0(a) = \frac{a}{f(a)} + \frac{b-a}{f(r_0)},$$

$$P_1 = P_1(a) = \frac{-a \ln (f(0)/f(a))}{f(a) - f(0)},$$

and

$$\Gamma = \Gamma(a) = \left(\frac{a}{pf(a)} - \frac{f(0)}{f(a)} + \frac{b-a}{pf(r_0)}\right)\frac{f(a)}{f(a)-f(0)},$$

and note that $\Gamma > 0$. We shall divide our discussion into cases described in terms of P_0, P_1, and Γ. However, some preliminary discussion is required.

Inequality (3.1) will be satisfied if and only if both the inequalities

$$\beta(1-e^{-\gamma/\beta}) + (1-\gamma)\frac{f(0)}{f(a)} - \frac{a}{pf(a)} \geqq 0 \tag{3.1}'$$

and

$$\beta(1-e^{-\gamma/\beta}) - \gamma + \frac{b-a}{pf(r_0)} \geqq 0 \tag{3.1}''$$

are satisfied. For $\gamma \leqq \Gamma$, (3.1) reduces to (3.1)′ and for $\gamma \geqq \Gamma$, (3.1) reduces to (3.1)″, since $\gamma \leqq \Gamma$ if and only if

$$\frac{a}{pf(a)} - (1-\gamma)\frac{f(0)}{f(a)} \geqq \gamma - \frac{b-a}{pf(r_0)}.$$

Furthermore, it is easy to verify that $\Gamma \geqq 1$ if and only if $p \leqq P_0$.

Note that, for fixed β, the left side of (3.1)″ is a decreasing function of γ and is positive for γ sufficiently close to zero. In particular, smaller values of γ yield smaller values of β satisfying (3.1)″.

For fixed β, the left side of (3.1)′ has its maximum value when $\gamma = -\beta \ln (f(0)/f(a))$, increases in $(0, -\beta \ln (f(0)/f(a)))$ and decreases for $\gamma \geqq$

$-\beta \ln (f(0)/f(a))$. For $\gamma = -\beta \ln (f(0)/f(a))$, (3.1)′ reduces to

$$\beta\left(1-\frac{f(0)}{f(a)}+\frac{f(0)}{f(a)}\ln\frac{f(0)}{f(a)}\right)+\frac{pf(0)-a}{pf(a)}\geqq 0,$$

and the left side of this inequality is an increasing function of β which is positive if and only if

$$\beta \geqq \frac{a-pf(0)}{p\left(f(a)-f(0)+f(0)\ln\dfrac{f(0)}{f(a)}\right)} \equiv \beta_0(a,p).$$

Thus $\beta_0(a,p)$ is the smallest number for which (3.1)′ holds for some number γ; however, the required γ is $-\beta_0(a,p)\ln(f(0)/f(a))$, which may be greater than 1. In fact, $-\beta_0(a,p)\ln(f(0)/f(a))\leqq 1$ if and only if $p\geqq P_1$.

We are now prepared for our case-by-case discussion.

Case 1. $p\leqq P_0$ and $p\geqq P_1$. These inequalities imply that

$$-\beta_0(a,p)\ln\frac{f(0)}{f(a)}\leqq 1 \quad \text{and} \quad 1\leqq\Gamma.$$

Since $\gamma\leqq\Gamma$ for all $\gamma\in(0,1]$, (3.1) reduces to (3.1)′ and $\beta(a,p)=\beta_0(a,p)$.

Case 2. $p\leqq P_0$ and $p<P_1$. In this case $-\beta_0(a,p)\ln(f(0)/f(a))>1$ and $1\leqq\Gamma$. Again $\gamma\leqq\Gamma$ for each $\gamma\in(0,1]$, so that (3.1) reduces to (3.1)′. Since $\beta_0(a,p)$ does not satisfy (3.1)′ for any $\gamma\in(0,1]$, it follows that $\beta(a,p)>\beta_0(a,p)$ and that $-\beta(a,p)\ln(f(0)/f(a))>1$. Therefore if (3.1)′ holds for $\beta=\beta(a,p)$ and some $\gamma\in(0,1]$, it must hold for $\gamma=1$. Then $\beta(a,p)$ must be the solution to the equation

$$\beta(a,p)(1-e^{-1/\beta(a,p)})-\frac{a}{pf(a)}=0.$$

Case 3. $p\geqq P_0$ and $p\geqq P_1$. These inequalities imply that both Γ and $-\beta_0(a,p)\ln(f(0)/f(a))$ are less than 1. We will subdivide this case in terms of these two numbers.

Case 3a. $\Gamma\leqq-\beta_0(a,p)\ln(f(0)/f(a))$. For $\gamma\in(0,\Gamma]$, (3.1) reduces to (3.1)′ and (3.1)′ is not satisfied for $\beta<\beta_0(a,p)$. If $\beta\geqq\beta_0(a,p)$, $-\beta\ln(f(0)/f(a))\geqq\Gamma$, so that the smallest value of β satisfying (3.1)′ for some $\gamma\in(0,\Gamma]$ is the solution to

$$\beta(1-e^{-\Gamma/\beta})+(1-\Gamma)\frac{f(0)}{f(a)}-\frac{a}{pf(a)}=0.$$

If $\gamma\in[\Gamma,1]$, then (3.1) reduces to (3.1)″ and the smallest value of β satisfying (3.1)″ for some $\gamma\in[\Gamma,1]$ is the solution to

$$\beta(1-e^{-\Gamma/\beta})-\Gamma+\frac{b-a}{pf(r_0)}=0.$$

But, since

$$\beta(1-e^{-\Gamma/\beta})+(1-\Gamma)\frac{f(0)}{f(a)}-\frac{a}{pf(a)}=\beta(1-e^{-\Gamma/\beta})-\Gamma+\frac{b-a}{pf(r_0)},$$

$\beta(a,p)$ is the solution to

$$\beta(a,p)(1-e^{-\Gamma/\beta(a,p)})+(1-\Gamma)\frac{f(0)}{f(a)}-\frac{a}{pf(a)}=0.$$

Case 3b. $\Gamma > -\beta_0(a,p)\ln(f(0)/f(a))$. With $\gamma = -\beta_0(a,p)\ln(f(0)/f(a))$, (3.1) reduces to (3.1)′, and (3.1)′ is satisfied with $\beta = \beta_0(a,p)$. Since $\beta_0(a,p)$ is the smallest value of β satisfying (3.1)′ for any γ, it follows that $\beta(a,p) = \beta_0(a,p)$.

Case 4. $p \geqq P_0$ and $p < P_1$. These inequalities imply $\Gamma \leqq 1$ and $-\beta_0(a,p)\ln(f(0)/f(a)) > 1$. If $\gamma \in (0,\Gamma]$, then (3.1)′ is the pertinent inequality and best results are obtained when $\gamma = \Gamma$. If $\gamma \in [\Gamma, 1]$, then (3.1)″ is the pertinent inequality and again best results are obtained when $\gamma = \Gamma$. Thus $\beta(a,p)$ is the solution to

$$\beta(a,p)(1-e^{-\Gamma/\beta(a,p)}) + (1-\Gamma)\frac{f(0)}{f(a)} - \frac{a}{pf(a)} = 0.$$

4. An example. Consider the case $q = 1.1$ and $k = 10$. Then

$$r_0 \cong 0.78233, \qquad f(r_0) \cong 0.0011623,$$
$$r_1 \cong 0.15550, \qquad f(r_1) \cong 0.0001647,$$
$$r_2 \cong 0.63729, \qquad f(r_2) \cong 0.0010299,$$
$$\frac{r_1}{f(r_1)} \cong 944.199, \qquad \frac{r_2}{f(r_2)} \cong 618.797,$$

and $f(0) = 0.00004994$.

By Theorem 2.4 there is only one solution if $0 < p \leqq 618.797$ and $\beta > 0$. From Theorem 2.7, it follows that the solution is unique provided

$$p < 944.119 \quad \text{and} \quad \beta \leqq \frac{230.117}{p} - 0.243737.$$

Using Theorem 2.5 with $\beta_0 = 2.22$ and $p_0 = 1172.669$ establishes a unique solution if

$$944.119 < p \leqq 1172.669 \quad \text{and} \quad \beta(1-e^{-1/\beta}) \geqq \frac{944.119}{p},$$

or if

$$1172.669 \leqq p \quad \text{and} \quad \beta(1-e^{-1/\beta}) f(0.63729e^{1/\beta}) \geqq \frac{0.63729}{p}. \tag{4.1}$$

Finally, it follows from Theorem 2.6 with $\beta_1 = 1.85034$ that there is only one solution when

$$\beta \geqq 1.85034 \quad \text{and} \quad p \geqq e^{10}$$

or

$$\beta < 1.85034 \quad \text{and} \quad p \geqq \frac{40756.5}{\beta}. \tag{4.2}$$

The boundaries of the regions described by (4.1) and (4.2) intersect when $p = 22{,}082$ and $\beta = 1.8459$.

For multiple solutions, observe that the minimum value of $a/f(a) + (b-a)/f(r_0)$ for $a \in [r_1, r_2]$ is 846.99 and occurs when $a = r_2$. Also the minimum value of

$-(a \ln (f(0)/f(a)))/(f(a)-f(0))$ for $a \in [r_1, r_2]$ is 1616.9. Therefore if $p \leqq r_1/f(r_1)$, then $p \leqq P_1(a)$ for each $a \in (r_1, r_2]$. Now if $618.797 < p \leqq 846.99$, it follows from Case 2 that $\beta(a, p)$ is the solution to

$$\beta(a, p)(1-e^{-1/\beta(a,p)})-\frac{a}{pf(a)}=0,$$

and that $\beta_p \equiv \inf_{a_p < a \leqq r_2} \beta(a, p) = \beta(r_2, p)$. If $846.99 < p \leqq 944.119$, let a_1 be the number in $[a_p, r_2]$ such that

$$\frac{a_1}{f(a_1)}+\frac{b_1-a_1}{f(r_0)}=p.$$

For $a \leqq a_1$, it follows that $p \leqq P_0(a)$ and, from Case 2, that $\beta(a, p)$ is the solution to

$$\beta(a, p)(1-e^{-1/\beta(a,p)})-\frac{a}{pf(a)}=0.$$

On the other hand, if $a > a_1$ then $p > P_0(a)$ and it follows from Case 4 that $\beta(a, p)$ is the solution to

$$\beta(a, p)(1-e^{-\Gamma(a)/\beta(a,p)})+(1-\Gamma(a))\frac{f(0)}{f(a)}-\frac{a}{pf(a)}=0.$$

Clearly $\beta(a, p) \geqq \beta(a_1, p)$ for $a \leqq a_1$. Also with this choice of constants, $\beta(a, p) \geqq \beta(a_1, p)$ if $a \geqq a_1$. Thus, $\beta_p = \beta(a_1, \text{p})$. The following table gives the value of β_p for selected points $p \in (618.797, 944.119]$.

p	620	650	700	750	800	847	900	944
β_p	258	10.08	3.97	2.52	1.86	1.49	1.262	1.132

The combination of these results gives the following graphic description of the number of solutions when $q = 1.1$ and $k = 10$.

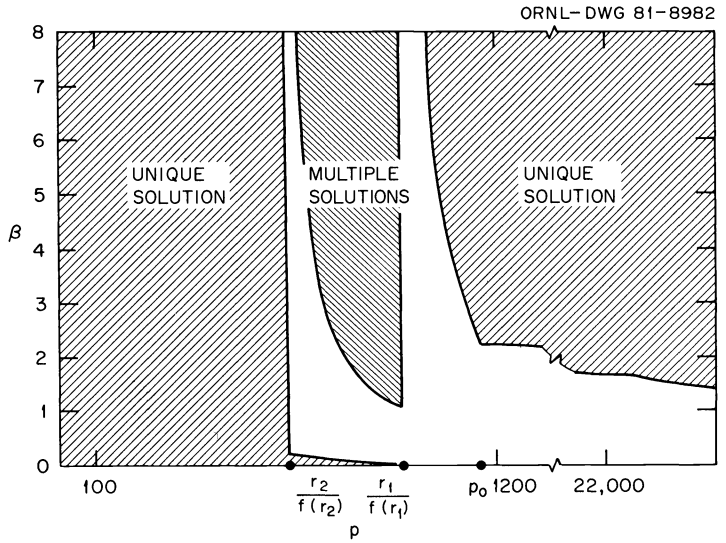

FIG. 1

REFERENCES

[1] H. AMANN, *On the number of solutions of nonlinear equations in ordered Banach spaces*, J. Funct. Anal., 11 (1972), pp. 346–384.

[2] ———, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[3] N. R. AMUNDSON AND D. LUSS, *Qualitative and quantitative observations on the tubular reactor*, Canad. J. Chem. Engr., 46 (1968), pp. 424–433.

[4] D. S. COHEN, *Multiple stable solutions of nonlinear boundary value problems arising in chemical reactor theory*, SIAM J. Appl. Math., 20 (1971), pp. 1–13.

[5] R. W. LEGGETT AND L. R. WILLIAMS, *Multiple positive fixed points of nonlinear operators on ordered Banach spaces*, Indiana Univ. Math. J., 28 (1979), pp. 673–688.

[6] A. VARMA AND R. ARIS, *Stirred pots and empty tubes*, Chemical Reactor Theory—A Review L. Lapidus and N. R. Amundson, eds., Prentice-Hall, Englewood Cliffs, NJ, 1977.

[7] L. R. WILLIAMS AND R. W. LEGGETT, *Multiple fixed point theorems for problems in chemical reactor theory*, J. Math. Anal. Appl. 69 (1979), pp. 180–193.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0010 $01.00/0

ON THE ESTIMATION OF FUNCTIONS OF SEVERAL VARIABLES FROM AGGREGATED DATA*

NIRA DYN† AND GRACE WAHBA‡

Abstract. This work was motivated by the problem of obtaining a smooth density function over a geographical region from data aggregated over irregular subregions. Minimization of a family of roughness criteria given "volume" data leads to smooth multivariate functions—Laplacian histosplines, having a certain order of the iterated Laplacian of constant value in each of the subregions and satisfying natural boundary conditions on the boundary of the region. For inexact data, e.g., in case of estimating an underlying density given counts of events by subregions, Laplacian smoothing histosplines are constructed, analogous to smoothing splines in the univariate case, and a method for choosing the smoothing parameter is presented.

For both cases of exact and inexact data, modified roughness criteria, independent of the region, are discussed, and results known for point-evaluation data are extended to the case of aggregated data.

1. Introduction. The work in this paper is motivated by the following problem: incidence rates of certain types of cancer are known to vary geographically; for example, persons living in areas with higher exposure to sunshine are more likely to get skin cancer than those in more northerly regions. Data on population density and disease occurrence is typically collected by bureaucratic subdivision. It is desired, from this aggregate data, to obtain an estimate $\hat{p}(x_1, x_2)$ of the probability $p(x_1, x_2)$ that a person living at (x_1, x_2) will contract the disease in a given year. Contour map representations of $\hat{p}$ can then be used to visually look for geographic patterns in p, and for apparent correlations with other geographically varying variables.

For concreteness, we consider data reported by state. Let Ω represent the contiguous 48 states of the U.S., and Ω_i the ith state. If $u(x_1, x_2)$ is the population density at point (x_1, x_2) (we pretend this is well defined), then the expected number of cases of our subject disease in state i is μ_i;

$$\mu_i = \int_{\Omega_i} p(x_1, x_2)u(x_1, x_2)\, dx_1\, dx_2.$$

The population $s_i = \int_{\Omega_i} u(x_1, x_2)\, dx_1\, dx_2$ of state i is assumed to be known exactly. The population of further subdivisions, e.g., countries, can also be assumed to be known exactly. In a particular year, the number Z_i of cases actually occurring in Ω_i is reported. If p is very small, then Z_i may be modelled as a Poisson random variable with mean μ_i. From this data it is desired to estimate $p(x_1, x_2)$, $(x_1, x_2) \in \Omega$. We will do this by first estimating $u(x_1, x_2)$ using only the population data $\{s_i\}$, and then estimating $g(x_1, x_2) \equiv p(x_1, x_2)u(x_1, x_2)$ using the disease count data $\{Z_i\}$. The estimate of p is then the quotient of these two estimates. For notational convenience we suppose that population data is aggregated at the same level (i.e., state) as the disease count data.

It is possible to obtain heuristically reasonable estimates of u and g by assuming that they are "smooth" in some sense, namely by minimizing certain measures of roughness. The roughness measures we will consider in most detail are defined by

$$J_1(u) = \int_{\Omega} (u_{x_1}^2 + u_{x_2}^2)\, dx_1\, dx_2 \tag{1.1}$$

* Received by the editors July 12, 1979. This research was sponsored by the U.S. Army under contracts DAAG29-75-C-0024 and DAAG29-77-G-0207.

† Mathematics Research Center, University of Wisconsin-Madison, visiting from the Department of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel.

‡ Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin 53706.

or

$$J_2(u) = \int_\Omega (u_{x_1x_1}^2 + 2u_{x_1x_2}^2 + u_{x_2x_2}^2)\, dx_1\, dx_2. \tag{1.2}$$

We will also briefly consider the more general measures

$$J_m(u) = \int_\Omega \sum_{i=0}^{m} \binom{m}{i} \left(\frac{\partial^m u}{\partial x_1^i\, \partial x_2^{m-i}}\right)^2 dx_1\, dx_2, \qquad m = 1, 2, 3, \cdots. \tag{1.3}$$

First we consider the problem of estimating u. With the roughness measures (1.1) and (1.2) our estimate $\hat{u}(x_1, x_2)$ of $u(x_1, x_2)$ will be the solution to one of the following problems.

Problems I-1/I-2. Find $u \in X$ (an appropriate space of functions on Ω) to minimize $J_1(u)/J_2(u)$ subject to the volume-matching constraints

$$\int_{\Omega_i} \int u(x, y)\, dx\, dy = s_i, \qquad i = 1, 2, \cdots, N, \tag{1.4}$$

where $\cup_{i=1}^N \Omega_i = \Omega$.

We obtain a characterization of the solution to a general problem of which Problems I-1 and I-2 are special cases.

Problem I-A. Let Ω be a smooth bounded subset of R^d, Euclidean d-space. Find $u \in H^m(\Omega)$ to minimize $J(u) = A(u, u)$, where

$$A(u, v) = \sum_{|\alpha|,|\beta|=m} \int_\Omega a_{\alpha\beta} D^\alpha u D^\beta v\, dx,$$

subject to

$$\int_\Omega \phi_i(x) u(x)\, dx = s_i, \qquad i = 1, 2, \cdots, N.$$

Here $H^m(\Omega)$ is the Sobolev space of functions with mixed partial derivatives up to order m in $L_2(\Omega)$, $x = (x_1, x_2, \cdots, x_d)$, $\alpha = (\alpha_1, \cdots, \alpha_d)$, $\beta = (\beta_1, \cdots, \beta_d)$, $|\alpha| = \sum_{i=1}^d \alpha_i$, $\sum_{i=1}^d \alpha_i = \sum_{i=1}^d \beta_i = m$, $D^\alpha u = (\partial^\alpha u / \partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d})$; $a_{\alpha\beta}$ are functions of x satisfying certain conditions specified in § 2, and the $\{\phi_i\}$ are linearly independent functions in $L_2(\Omega)$.

The characterization of the solution to Problem I-A is given in § 2. Certain further details are carried out in § 3 for the special cases of Problems I-1 and I-2. A simple example of Problem I-1, with concentric circles as subdomains, is worked out explicitly in § 4.

Numerical algorithms for computing the solutions to Problems I-1 and I-2 will appear in a separate paper.

The solutions to Problems I-1, I-2 and I-A are not required to be nonnegative, although it is known, of course, that $u(x_1, x_2)$ and $g(x_1, x_2)$ are nonnegative. In this paper, we sidestep the philosophical, theoretical and computational problems of imposing nonnegativity on the solution and hope to address this problem separately. The results of Lions and Stampacchia [12] will be relevant.

We know of very little literature specifically on the volume matching problem (although it is, of course, only a special case of the well-studied problem of estimating a function given the values of some linear functionals (see Golomb and Weinberger [9], Kimeldorf and Wahba [11]). Boneva, Kendall and Stefanov [2] discuss a special case in one dimension. Schoenberg and de Boor [16] discuss a volume matching

problem in two dimensions where the roughness measure has a tensor product structure and Ω is a rectangle with the Ω_i's a rectangular subdivision. Our interest in this problem was sparked by a paper of Tobler [18]. He proposed to solve the volume matching problem by minimizing $J_1(u) = \int_\Omega \int (u_{x_1}^2 + u_{x_2}^2)\, dx_1\, dx_2$ subject to volume matching conditions, positivity constraints, and certain boundary conditions, and suggested a numerical algorithm for doing this. Some of the results here are alluded to in our comments to his paper (Dyn, Wahba and Wong [8]).

Our results show that the solution to Problem I-A and the special cases I-1 and I-2 satisfies a certain elliptic boundary value problem with Neumann boundary conditions. Numerical implementation of these boundary value problems can be avoided if one is willing to modify the roughness criteria. Let X be a suitable space of functions on R^2 (to be defined), and define $\tilde{J}_m$ on X by

$$\tilde{J}_m(u) = \int_{R^2} \sum_{i=0}^{m} \binom{m}{i} \left(\frac{\partial^m u}{\partial x_1^i\, \partial x_2^{m-i}} \right)^2 dx_1\, dx_2.$$

Problem $\tilde{\text{I}}$-m. Find $u \in X$ to minimize $\tilde{J}_m(u)$ subject to

$$\int_{\Omega_i} u\, dx = s_i, \qquad i = 1, 2, \cdots, N.$$

If $\tilde{u}$ is the solution to this problem, we will have $\tilde{J}_m(\tilde{u}) \geqq J_m(\tilde{u}) \geqq J_m(\hat{u})$, with inequalities holding in general. This approach of using $\tilde{J}_m(u)$ as a roughness criterion has been extensively used for estimating surfaces given evaluation data by Duchon [6], [7], Meinguet [13], Paihua and Utreras [15] and Wahba [19]. Using these available results, we derive in § 7 an explicit expression for the solution of Problem $\tilde{\text{I}}$-m and a readily computable approximate solution. The results generalize easily to d dimensions.

We now proceed to the problem of estimating g. Since the data Z_i are only estimates of the μ_i we only want g to satisfy volume-matching conditions approximately. As in the case of smoothing splines (see [5] and references therein), we are led to another problem.

Problem II-m. Find $g \in X$ to minimize

$$\sum_{i=1}^{N} w_i \left(Z_i - \int_{\Omega_i} g(x, y)\, dx\, dy \right)^2 + \lambda J_m(g),$$

with $J_m(u)$ defined by (1.3). Here the $\{w_i\}$ should be equal to 1/variance Z_i. The parameter λ represents a trade-off between the roughness of g and the infidelity of g to the data. The variance of Z_i is μ_i, which is, of course, unknown. In practice, the w_i would have to be chosen iteratively. One could set $w_i = 1/Z_i$ initially, since Z_i is an estimate of μ_i. The resulting estimate of g is then used to get $\{w_i\}$ for a second estimate, etc.

In § 5 we characterize the solution to Problem II-m for J_m given by (1.3) and for given $w_1, \cdots, w_N$. In § 6 we indicate how λ may be chosen to approximately minimize the predictive mean square error. In § 7 we give an explicit representation for the solution to Problem II-m with J_m replaced by $\tilde{J}_m$ (Problem $\tilde{\text{II}}$-m). More significantly, we give explicit formulae for approximate solutions to Problem $\tilde{\text{II}}$-m which are suitable for numerical computation. In this context we also derive formulae for computing an optimal λ based on the results of § 6.

Hopefully, these results will provide the first step towards efficient methods for converting aggregate data to density maps.

2. Smooth surfaces on bounded domains matching integral data. Consider a bounded domain Ω of R^d with Γ its boundary, and a bilinear form

$$A(u, v) = \sum_{|\alpha|,|\beta|=m} \int_\Omega a_{\alpha\beta}(x) D^\alpha u D^\beta v, \qquad a_{\alpha\beta} \in L^\infty(\Omega) \tag{2.1}$$

where $x = (x_1, \cdots, x_d)$, $\alpha = (\alpha_1, \cdots, \alpha_d)$, $|\alpha| = \sum_{i=1}^d \alpha_i$, α_i a nonnegative integer, $D^\alpha = (\partial^{\alpha_1}/\partial x_1^{\alpha_1}) \cdots (\partial^{\alpha_d}/\partial x_d^{\alpha_d})$ (and similar notation for β). With this definition, $A(u, v)$ is continuous on $H^m(\Omega) \times H^m(\Omega)$, where $H^m(\Omega)$ is the Hilbert space

$$H^m(\Omega) = \{u | D^\alpha u \in L^2(\Omega), |\alpha| \leqq m\}, \qquad \|u\|^2_{H^m(\Omega)} = \sum_{|k| \leqq m} \|D^k u\|^2_{L^2(\Omega)}.$$

By assuming that

$$\sum_{|\alpha|,|\beta|=m} a_{\alpha\beta}(x) y_\alpha y_\beta > C_0 \sum_{|\alpha|=m} y_\alpha^2, \tag{2.2}$$

for all $y = (y_1, \cdots, y_k)$ and $k = \#\{\alpha | |\alpha| = m\}$, we have that $[A(u, u)]^{1/2}$ is a seminorm on $H^m(\Omega)$ with a null space Q (the space of all polynomials of total degree less than m, which is of dimension $M \equiv \binom{m+d-1}{d}$).

In this section we prove the existence and uniqueness of the solution to Problem I-A. For given $s_1, \cdots, s_N$, find $u \in H^m(\Omega)$ minimizing $A(u, u)$ among all functions in $H^m(\Omega)$ satisfying the integral data

$$\int_\Omega u \phi_i = s_i, \qquad i = 1, \cdots, N, \tag{2.3}$$

where $\phi_1, \cdots, \phi_N$ are N linearly independent functions in $L^2(\Omega)$.

In particular, we characterize the solution of Problem I-A as a solution of a certain boundary value problem.

We prove two lemmas.

LEMMA 2.1. *In the subspace H_0 of $H^m(\Omega)$ given by*

$$H_0 = \left\{ u | u \in H^m(\Omega), \int_\Omega D^\alpha u = 0, |\alpha| < m \right\}, \tag{2.4}$$

$\sqrt{A(u, u)}$ is an equivalent norm to $\|u\|_{H^m(\Omega)}$.

Proof. By (2.1) there exists $C_1 > 0$ such that

$$A(u, u) \leqq C_1 \|u\|^2_{H^m(\Omega)}, \qquad u \in H^m(\Omega). \tag{2.5}$$

Iterating the Poincaré inequality [14],

$$\int_\Omega u^2 \leqq C \left\{ \sum_{|\alpha|=1} \int_\Omega (D^\alpha u)^2 + \left[\int_\Omega u \right]^2 \right\}, \qquad u \in H^1(\Omega), \tag{2.6}$$

we obtain, for any $0 \leqq k < m$,

$$\sum_{|\alpha|=k} \int_\Omega (D^\alpha u)^2 \leqq C_2 \left\{ \sum_{|\alpha|=m} \int_\Omega (D^\alpha u)^2 + \sum_{k \leqq |\alpha| < m} \left[\int_\Omega D^\alpha u \right]^2 \right\}, \qquad u \in H^m(\Omega). \tag{2.7}$$

Thus, by (2.4) and (2.2),

$$\|u\|^2_{H^m(\Omega)} \leqq C_3 \sum_{|\alpha|=m} \int_\Omega (D^\alpha u)^2 \leqq \frac{C_3}{C_0} A(u, u), \qquad u \in H_0(\Omega). \tag{2.8}$$

Let $Q = \text{span}\{q_1, \cdots, q_M\}$. We assume that $N > M$ and that the N linear functionals in (2.3) are linearly independent over Q. Without loss of generality we can

assume that the matrix

$$\left[\int_\Omega q_i \phi_{N-M+j}\right]_{i,j=1}^{M} \tag{2.9}$$

is of rank M. Therefore, there exists a basis $\{\tilde{q}_1, \cdots, \tilde{q}_M\}$ of Q with the property

$$\int_\Omega \tilde{q}_i \phi_{N-M+j} = \delta_{ij}, \qquad i, j = 1, \cdots, M. \tag{2.10}$$

LEMMA 2.2. *In the subspace H_1 of $H^m(\Omega)$, given by*

$$H_1(\Omega) = \left\{ u \,|\, u \in H^m(\Omega), \int_\Omega u\phi_i = 0,\ i = N-M+1, \cdots, N \right\}, \tag{2.11}$$

$\sqrt{A(u,u)}$ *is an equivalent norm to* $\|u\|_{H^m(\Omega)}$.

Proof. For any $u \in H_1(\Omega)$ there exists $q \in Q$ such that $u_0 = u - q \in H_0(\Omega)$, and, therefore,

$$u = u_0 + \sum_{i=1}^{M} \tilde{q}_i \int_\Omega u_0 \phi_{N-M+i}.$$

Since, for any $\phi \in L_2(\Omega)$,

$$\left| \int_\Omega u\phi \right| \leqq \|\phi\|_{L^2(\Omega)} \|u\|_{H^m(\Omega)}, \qquad u \in H^m(\Omega), \tag{2.12}$$

we get, in view of Lemma 2.1,

$$\begin{aligned} \|u\|_{H^m(\Omega)} &\leqq \|u_0\|_{H^m(\Omega)} + \sum_{i=1}^{M} \|\tilde{q}_i\|_{H^m(\Omega)} \|\phi_{N-M+i}\|_{L^2(\Omega)} \|u_0\|_{H^m(\Omega)} \\ &\leqq C_4 \sqrt{A(u_0, u_0)} = C_4 \sqrt{A(u,u)}. \end{aligned}$$

This together with (2.5) completes the proof of the lemma. □

Let $u \in H^m(\Omega)$ satisfy (2.3). Then

$$\tilde{u} = u - \sum_{i=1}^{M} s_{N-M+i} \tilde{q}_i \in H_1, \tag{2.13}$$

$$\int_\Omega \tilde{u}\phi_j = s_j - \sum_{i=1}^{M} s_{N-M+i} \int_\Omega \tilde{q}_i \phi_j = \tilde{s}_j, \qquad j = 1, \cdots, N-M, \tag{2.14}$$

and $A(\tilde{u}, \tilde{u}) = A(u, u)$. Therefore, Problem I-A is equivalent to Problem (I-A)′.

Problem (I-A): Find $\tilde{u} \in H_1$ minimizing $A(u, u)$ among all functions of H_1 satisfying (2.14), or equivalently satisfying

$$\int_\Omega \tilde{u}\tilde{\phi}_j = \tilde{s}_j, \qquad j = 1, \cdots, N-M, \tag{2.15}$$

with

$$\tilde{\phi}_j = \phi_j - \sum_{i=1}^{M} \alpha_{ij} \phi_{N-M+i}, \qquad \{\alpha_{ij}\} \text{ arbitrary}. \tag{2.16}$$

In particular, it is possible by assumption (2.9) to choose $\{\alpha_{ij}\}$ such that

$$\int_\Omega q\tilde{\phi}_j = 0, \qquad j = 1, \cdots, N-M, \quad q \in Q. \tag{2.17}$$

By Lemma 2.2, the linear functionals

$$L_j(u) = \int_\Omega u\tilde{\phi}_j, \qquad j = 1, \cdots, N-M \tag{2.18}$$

are bounded in H_1 with respect to the norm $[A(u, u)]^{1/2}$. Invoking the Riesz representation theorem, we conclude the existence of $\xi_j \in H_1$, $j = 1, \cdots, N-M$, satisfying

$$A(u, \xi_j) = \int_\Omega u\tilde{\phi}_j, \qquad \text{all } u \in H_1, \tag{2.19}$$

and due to (2.17)

$$A(q, \xi_j) = \int_\Omega q\tilde{\phi}_j = 0, \qquad \text{all } q \in Q. \tag{2.20}$$

Since $\phi_1, \cdots, \phi_N$ are linearly independent so are $\xi_1, \cdots, \xi_{N-M}$, and the solution to Problem (I-A)′ is known to be the unique function in the span of $\{\xi_1, \cdots, \xi_{N-M}\}$ satisfying (2.16), (see [9]). The solution to Problem I-A is related to this solution according to (2.13). The following theorem summarizes the above findings.

THEOREM 2.1. *There exists a unique solution to Problem* I-A. *The solution is of the form*

$$\hat{u} = \sum_{i=1}^{N-M} c_i\xi_i + \sum_{i=1}^{M} s_{N-M+i}\tilde{q}_i, \tag{2.21}$$

where $\xi_1, \cdots, \xi_{N-M}$ *are the unique functions in* H_1 *determined by* (2.19), *and* $c_1, \cdots, c_{N-M}$ *are the solution of the nonsingular linear system*

$$\sum_{i=1}^{N-M} c_i A(\xi_i, \xi_j) = \tilde{s}_j = s_j - \sum_{l=1}^{M} s_{N-M+l}\int_\Omega \tilde{q}_l\phi_j, \qquad j = 1, \cdots, N-M. \tag{2.22}$$

An immediate consequence of Theorem 2.1, (2.19) and (2.20) is

COROLLARY 2.1. *The solution* $\hat{u}$ *of Problem* I-A *is uniquely determined by the variational characterization*

$$A(\hat{u}, v) = \int_\Omega \left(\sum_{i=1}^{N} \gamma_i\phi_i\right)v, \qquad v \in H^m(\Omega), \tag{2.23}$$

and the matching conditions

$$\int_\Omega u\phi_i = s_i, \qquad i = 1, \cdots, N. \tag{2.24}$$

In (2.23), $\gamma_1, \cdots, \gamma_N$ *are constants, which in particular satisfy*

$$\int_\Omega \left(\sum_{i=1}^{N} \gamma_i\phi_i\right)q = 0, \qquad q \in Q. \tag{2.25}$$

In case Ω is a smooth domain, the solution $\hat{u}$ of Problem I-A can be further characterized in terms of a boundary value problem. Since each ξ_i, $1 \leqq i \leqq N-M$, satisfies (2.19) and (2.20), namely

$$A(u, \xi_i) = \int_\Omega u\tilde{\phi}_i \quad \text{for all } u \in H^m(\Omega),$$

we conclude from Aubin [1, Corollary 2-2, pp. 219–220] that ξ_i is the unique solution

in H_1 to the boundary value problem

$$\Lambda\xi_i = \tilde{\phi}_i \quad \text{in } \Omega, \tag{2.26}$$

$$\delta_j\xi_i = 0 \quad \text{for } m \leqq j \leqq 2m-1 \text{ on } \Gamma. \tag{2.27}$$

In (2.26) Λ is the differential operator of order $2m$ given by

$$\Lambda u = \sum_{|\alpha|,|\beta|=m} (-1)^{|\beta|} D^{\beta}(a_{\alpha\beta}(x) D^{\alpha} u), \tag{2.28}$$

and in (2.27) $\delta = (\delta_{2m-1}, \cdots, \delta_m)$ is a differential operator of order $\geqq m$ mapping

$$H^m(\Omega, \Lambda) = \{u | u \in H^m(\Omega), \Lambda u \in L^2(\Omega)\} \tag{2.29}$$

into $\prod_{j=2m-1}^{m} H^{m-j-1/2}(\Gamma)$, such that the generalized Green formula holds

$$A(u, v) = \int_{\Omega} (\Lambda u) v + \sum_{j=0}^{m-1} \int_{\partial\Omega} (\delta_{2m-j-1} u)\left(\frac{\partial^j}{\partial n^j} v\right) \tag{2.30}$$

($\partial/\partial n$ is the operator of normal derivative to the boundary Γ).

The characterization (2.26), (2.27) of $\xi_i \in H_1$, together with Theorem 2.1, yields the following.

THEOREM 2.2. *The solution to Problem* I-A, *for a smooth domain* Ω *is uniquely determined as the solution to the boundary value problem*

$$\Lambda\hat{u} = \sum_{i=1}^{N} \gamma_i \phi_i \quad \text{in } \Omega, \tag{2.31}$$

$$\delta_j\hat{u} = 0, \qquad m \leqq j \leqq 2m-1 \quad \text{on } \Gamma, \tag{2.32}$$

which satisfies the matching conditions (2.24). *In* (2.31) $\gamma_1, \cdots, \gamma_N$ *are* N *constants satisfying* (2.25).

3. Laplacian histosplines—The volume-matching surfaces. In this section we specialize to the concrete problem of finding a smooth surface $u = u(x_1, x_2)$ having prescribed volumes over specified subdomains in R^2. We characterize the volume-matching surface as a function with the even order differential form $\Delta^m = [(\partial/\partial x_1)^2 + (\partial/\partial x_2)^2]^m$ of constant value in each of the subdomains. These surfaces are therefore strikingly analogous to even degree one-dimensional splines and are regarded as functions with a certain even order derivative of constant value in each subinterval. Following a suggestion of Professor Iso Schoenberg, we term these surfaces "Laplacian histosplines" in analogy to the univariate histosplines of Boneva, Kendall and Stefanov [2], which are the even degree univariate splines solving the "area matching" problem.

We consider in detail the following two problems. Let Ω be a smooth bounded domain in R^2 subdivided into N disjoint domains $\Omega_1, \cdots, \Omega_N$, $\Omega = \cup_{i=1}^{N} \Omega_i$.

Problem I-1. Find $u \in H^1(\Omega)$ minimizing

$$\int_{\Omega} (u_{x_1}^2 + u_{x_2}^2)\, dx_1\, dx_2 \tag{3.1}$$

among all functions in $H^1(\Omega)$ satisfying

$$\int_{\Omega_i} u = s_i, \qquad i = 1, \cdots, N. \tag{3.2}$$

Problem I-2. Find $u \in H^2(\Omega)$ minimizing

$$\int_\Omega (u_{x_1x_1}^2 + 2u_{x_1x_2}^2 + u_{x_2x_2}^2)\, dx_1\, dx_2 \tag{3.3}$$

among all functions in $H^2(\Omega)$ satisfying (3.2).

From a practical point of view these two problems are the most interesting, since computation of solutions of similar problems with higher order forms (2.1) becomes too complicated, with the increased complexity of the operators Λ and δ in Theorem 2.2.

Using Theorem 2.2 for the special setting of Problem I-1 together with the classical Green formula [3],

$$\int_\Omega u_{x_1}v_{x_1} + u_{x_2}v_{x_2} = \int_\Omega (-\Delta u)v + \int_\Gamma \frac{\partial u}{\partial n} v, \tag{3.4}$$

we obtain:

THEOREM 3.1. *The solution to Problem* I-1 *is uniquely determined by the conditions*:

$$\Delta\hat{u} = \sum_{i=1}^{N} \gamma_i \chi_{\Omega_i}, \qquad \chi_{\Omega_i} = \begin{cases} 1 & \text{in } \Omega_i, \\ 0 & \text{elsewhere,} \end{cases}$$

$$\sum_{i=1}^{N} \gamma_i \int_{\Omega_i} 1 = 0,$$

$$\frac{\partial \hat{u}}{\partial n} = 0 \quad \text{on } \Gamma,$$

$$\int_{\Omega_i} \hat{u} = s_i, \qquad i = 1, \cdots, N.$$

To get a similar result for Problem I-2, we first derive a more general Green formula for the bilinear form corresponding to the seminorm (3.3). By a repeated use of (3.4), we get

$$\int_\Omega u_{x_1}v_{x_1} + 2u_{x_1x_2}v_{x_1x_2} + u_{x_2x_2}v_{x_2x_2} = \int_\Omega (\Delta^2 u)v - \int_\Gamma \left(\frac{\partial}{\partial n}\Delta u\right)v + \int_\Gamma \nabla\frac{\partial u}{\partial n} \cdot \nabla v, \tag{3.5}$$

since on Γ, $\nabla u \cdot \nabla v = (\partial u/\partial n)(\partial v/\partial n) + (\partial u/\partial \tau)(\partial v/\partial \tau)$, where $\partial/\partial\tau$ is the tangential derivative to Γ, the last term in (3.5) becomes

$$\int_\Gamma \nabla\frac{\partial u}{\partial n} \cdot \nabla v = \int_\Gamma \frac{\partial^2 u}{\partial n^2}\frac{\partial v}{\partial n} + \int_\Gamma \frac{\partial^2 u}{\partial \tau \partial n}\frac{\partial v}{\partial \tau} = \int_\Gamma \frac{\partial^2 u}{\partial n^2} \cdot \frac{\partial v}{\partial n} - \int \frac{\partial^3 u}{\partial \tau^2 \partial n} v. \tag{3.6}$$

Comparing (3.5) and (3.6) with the generalized Green formula (2.30), we conclude that, for the seminorm (3.3), Λ and δ of Theorem 2.2 are

$$\Lambda = \Delta^2, \quad \delta = (\delta_3, \delta_2), \quad \delta_2 = \frac{\partial^2}{\partial n^2}, \quad \delta_3 = -\left(\Delta + \frac{\partial^2}{\partial \tau^2}\right)\frac{\partial}{\partial n}. \tag{3.7}$$

Thus, by Theorem 2.2:

THEOREM 3.2. *The solution* $\hat{u}$ *to Problem* I-2 *is uniquely determined by the*

conditions:

$$\Delta^2 \hat{u} = \sum_{i=1}^{N} \gamma_i \chi_{\Omega_i} \quad \text{in } \Omega,$$

$$\sum_{i=1}^{N} \gamma_i \int_{\Omega_i} q = 0, \qquad q = 1, x_1, x_2,$$

$$\frac{\partial^2 \hat{u}}{\partial n^2} = 0, \qquad \left(\frac{\partial^2}{\partial \tau^2} + \Delta\right)\frac{\partial \hat{u}}{\partial n} = 0 \quad \text{on } \Gamma$$

$$\int_{\Omega_i} \hat{u} = s_i, \qquad i = 1, \cdots, N.$$

Remark. It can be shown by Theorem 2.2 and repeated applications of the classical Green formula that, for the higher order roughness criterion

$$J_m(u) = \int_{\Omega} \sum_{i=1}^{m} \binom{m}{i} \left(\frac{\partial^m u}{\partial x_1^i \partial x_2^{m-i}}\right)^2 dx_1\, dx_2, \qquad m \geqq 3, \tag{3.8}$$

the solution to the volume-matching problem satisfies

$$(-1)^m \Delta^m u = \sum_{i=1}^{N} \gamma_i \chi_{\Omega_i} \quad \text{in } \Omega, \tag{3.9}$$

with the appropriate boundary conditions,

$$\delta_j u = 0 \quad \text{on } \Gamma, \qquad m \leqq j \leqq 2m-1. \tag{3.10}$$

4. A simple example of an explicit Laplacian histospline. Consider N subdomains in R^2,

$$\Omega_i = \{(x_1, x_2) | R_{i-1} < \sqrt{x_1^2 + x_2^2} < R_i\}, \qquad i = 1, \cdots, N, \tag{4.1}$$

with $R_0 \geqq 0$ and $\Omega = \cup_{i=1}^{N} \Omega_i$. In the following we derive the explicit form of the solution to the volume matching problem, I-1.

By the radial symmetry of the problem, $u = u(r)$ with $r = \sqrt{x_1^2 + x_2^2}$, and in view of Theorem 3.1, $-\Delta \hat{u} = \gamma_i$ in Ω_i, $i = 1, \cdots, N$. Since [3]

$$\Delta f(r) = \frac{1}{r}\frac{d}{dr}[rf'(r)],$$

$$\hat{u} = -\frac{\gamma_i}{4} r^2 + c_i \log r + b_i \quad \text{in } \Omega_i, \qquad i = 1, \cdots, N. \tag{4.2}$$

The coefficients γ_i, c_i, b_i, $i = 1, \cdots, N$ satisfy the following conditions implied by Theorem 3.1 and the continuity of $\hat{u}$ and $d\hat{u}/dr$:

$$\left.\frac{d\hat{u}}{dr}\right|_{r=R_N} = 0 = -\frac{\gamma_N}{2} R_N + \frac{c_N}{R_N} \qquad \text{(boundary condition)}, \tag{4.3}$$

$$\sum_{i=1}^{N} \gamma_i (R_i^2 - R_{i-1}^2) = 0, \qquad \left(\sum_{i=1}^{N} \gamma_i \int_{\Omega_i} 1 = 0\right), \tag{4.4}$$

$$c_i - c_{i+1} = (\gamma_i - \gamma_{i+1})\frac{R_i^2}{2}, \qquad i = 1, \cdots, N-1 \qquad \left(\text{continuity of } \frac{d\hat{u}}{dr}\right), \tag{4.5}$$

$$b_i - b_{i+1} = (\gamma_i - \gamma_{i+1})\frac{R_i^2}{4} - (c_i - c_{i+1}) \log R_i, \tag{4.6}$$

$$i = 1, \cdots, N-1 \qquad \text{(continuity of } \hat{u}\text{)},$$

$$\frac{\gamma_i}{16}(R_i^4 - R_{i-1}^4) + c_i\left\{\frac{R_i^2}{2}\left[\log R_i - \frac{1}{2}\right] - \frac{R_{i-1}^2}{2}\left[\log R_{i-1} - \frac{1}{2}\right]\right\} \tag{4.7}$$

$$+ b_i\frac{R_i^2 - R_{i-1}^2}{2} = \frac{s_i}{2\pi}, \qquad i = 1, \cdots, N \qquad \text{(volume matching)}.$$

The total number of linear equations (4.3)–(4.7) is $3n$, as is the total number of known coefficients. If $R_0 > 0$, there is an additional boundary condition to be satisfied;

$$\left.\frac{d\hat{u}}{dr}\right|_{r=R_0} = 0 = -\frac{\gamma_1 R_0}{2} + \frac{c_1}{R_0} \quad \text{if } R_0 > 0. \tag{4.8}$$

Claim. *If* $R_0 > 0$, (4.8) *is linearly dependent on* (4.3)–(4.5). *If* $R_0 = 0$, *then* (4.3)–(4.5) *imply* $c_1 = 0$.

Proof. Summing (4.5) for $i = 1, \cdots, N-1$ we get

$$c_1 - c_N = \sum_{i=1}^{N-1} (\gamma_i - \gamma_{i+1})\frac{R_i^2}{2} = \frac{1}{2}\sum_{i=1}^{N} \gamma_i(R_i^2 - R_{i-1}^2) + \frac{1}{2}\gamma_1 R_0^2 - \frac{1}{2}\gamma_N R_N^2,$$

which, in view of (4.4) and (4.3), can be written as

$$c_1 - c_N = \tfrac{1}{2}(\gamma_1 R_0^2 - \gamma_N R_N^2) = \tfrac{1}{2}\gamma_1 R_0^2 - c_N.$$

Therefore, $c_1 = \frac{1}{2}\gamma_1 R_0^2$, proving the claim. □

By integrating $r\hat{u}(r)$, one can transform this volume matching problem into an interpolation problem (similar equivalence exists between area-matching splines and interpolating splines in the one-dimensional case [16]). Thus, defining

$$U(r) = \int_{R_0}^{r} \rho u(\rho)\, d\rho, \qquad u(r) = \frac{1}{r}U'(r), \tag{4.9}$$

we have to construct an "interpolating spline" of the form

$$U(r) = A_i + B_i r^2 + c_i r^4 + D_i r^2 \log r, \qquad R_{i-1} \leqq r \leqq R_i, \quad i = 1, \cdots, N, \tag{4.10}$$

satisfying

$$U(r) \in C^2(R_0, R_N), \qquad U(R_i) = \frac{1}{2\pi}\sum_{j=1}^{i} s_j, \qquad i = 1, \cdots, N. \tag{4.11}$$

It is easy to check that the functions $1, r^2, r^4, r^2 \log r$ constitute an extended Chebyshev system on any interval of the form $(0, R_N)$. Thus $U(r)$, considered as a function of r, is a Chebyshev spline. (For the notion and construction of Chebyshev splines see, e.g., [10, Chapter 10].)

5. Laplacian histosplines for inexact data. In this section we consider the problem of finding a smooth function $\hat{g}$ given inexact volume data. Similar analysis can be done in the more general setting of § 2.

Problem II-m: For a given set of data $Z_1, \cdots, Z_N$, find $\hat{g} \in H^m(\Omega)$, minimizing

$$\sum_{i=1}^{N} w_i\left[\int_{\Omega_i} g - Z_i\right]^2 + \lambda J_m(g), \tag{5.1}$$

where $J_m(g)$ is defined in (3.8); Ω, $\Omega_1, \cdots, \Omega_N$ are as in § 2, and λ, $w_1, \cdots, w_N$ are fixed positive constants.

In the notation of § 2 any $g \in H^m(\Omega)$ can be represented as $g = g_1 + g_2 + g_3$ where $g_1 \in Q$, $g_2 \in \text{span}\{\xi_1, \cdots, \xi_{N-M}\}$ and g_3 satisfies

$$\int_{\Omega_i} g_3 = 0, \qquad i = 1, \cdots, N. \tag{5.2}$$

By (5.2), $g_3 \in H_1$ is orthogonal to $\xi_1, \cdots, \xi_{N-M}$ with respect to the inner-product in H_1 corresponding to the norm $\sqrt{J_m(\cdot)}$. Therefore g_3 does not affect the first term in (5.1), while $J_m(g_1 + g_2 + g_3) = J_m(g_2) + J_m(g_3)$ and necessarily the solution to Problem II-m is of the form

$$\hat{g} = \hat{g}_1 + \hat{g}_2 = \sum_{i=1}^{N-M} c_i \xi_i + \sum_{i=1}^{M} d_i \tilde{q}_i. \tag{5.3}$$

Since for the volume data, $\phi_1, \cdots, \phi_N$ in § 2 are of the form

$$\phi_i = \chi_{\Omega_i}, \qquad i = 1, \cdots, N,$$

then by (2.16), (2.17) and (2.26),

$$(-1)^m \Delta^m \xi_i = \begin{cases} 1 & \text{in } \Omega_i, \\ 0 & \text{in } \Omega_j, \quad j \neq i, \quad j = 1, \cdots, N-M, \\ \gamma_{ij} & \text{in } \Omega_j, \quad j = N-M+1, \cdots, N \end{cases} \tag{5.4}$$

with γ_{ij} satisfying

$$\sum_{j=N-M+1}^{N} \gamma_{ij} \int_{\Omega_j} \tilde{q}_l + \int_{\Omega_i} \tilde{q}_l = 0, \qquad l = 1, \cdots, M, \quad i = 1, \cdots, N-M. \tag{5.5}$$

In view of (5.4), (5.5) and (2.27), the solution $\hat{g}$ to Problem II-m, given by (5.3), satisfies the boundary value problem

$$(-1)^m \Delta^m g = \sum_{i=1}^{N} \gamma_i \chi_{\Omega_i} \quad \text{in } \Omega, \tag{5.6}$$

$$\delta_j \hat{g} = 0 \quad \text{on} \quad \Gamma, \qquad m \leqq j \leqq 2m-1, \tag{5.7}$$

with $\gamma_1, \cdots, \gamma_N$ being N constants restricted by

$$\sum_{i=1}^{N} \gamma_i \int_{\Omega_i} \tilde{q}_l = 0, \qquad l = 1, \cdots, M. \tag{5.8}$$

In (5.7), the boundary operators $\delta_m, \cdots, \delta_{2m-1}$ are as in the remark in § 3.

The following theorem relates the values of the constants $\gamma_1, \cdots, \gamma_N$ in (5.6) to the "smoothed data", namely to the values

$$\hat{Z}_i = \int_{\Omega_i} \hat{g}, \qquad i = 1, \cdots, N. \tag{5.9}$$

THEOREM 5.1. *The solution $\hat{g}$ of Problem* II-m *satisfies* (5.6) *with*

$$\gamma_i = \frac{w_i}{\lambda}(Z_i - \hat{Z}_i), \qquad i = 1, \cdots, N. \tag{5.10}$$

Proof. The coefficients in (5.3) satisfy the necessary conditions for minimizing (5.1), namely the vanishing of the partial derivatives of (5.1) with respect to

$c_1, \cdots, c_{N-M}$ and $d_1, \cdots, d_M$. In terms of the bilinear form $A_m(\cdot, \cdot)$ corresponding to $J_m(\cdot)$, these conditions become

$$\sum_{i=1}^{N-M} \left[w_i(\hat{Z}_i - Z_i) \int_{\Omega_i} \xi_j + \lambda A_m(\xi_i, \xi_i) c_i \right] = 0, \qquad j = 1, \cdots, N-M, \tag{5.11}$$

$$\sum_{i=N-M+1}^{N} w_i(\hat{Z}_i - Z_i) \int_{\Omega_i} \tilde{q}_j = 0, \qquad j = 1, \cdots, M. \tag{5.12}$$

In deriving (5.11), we recalled that

$$\int_{\Omega_i} \xi_j = 0, \qquad i = N-M+1, \cdots, N, \quad j = 1, \cdots, N-M. \tag{5.13}$$

Let K be the $(N-M)\times(N-M)$ matrix, with entries

$$K_{ij} = A_m(\xi_i, \xi_j) = \int_{\Omega_i} \xi_j = \int_{\Omega_j} \xi_i, \qquad i, j = 1, \cdots, N-M, \tag{5.14}$$

Let T be the $(N-M)\times M$ matrix with entries

$$T_{ij} = \int_{\Omega_i} \tilde{q}_j, \qquad i = 1, \cdots, N-M, \quad j = 1, \cdots, M,$$

and let

$$W = \operatorname{diag}\{w_1, \cdots, w_{N-M}\}, \qquad \underset{\sim}{W} = \operatorname{diag}\{w_{N-M+1}, \cdots, w_N\},$$
$$c = (c_1, \cdots, c_{N-M})', \quad z = (Z_1, \cdots, Z_{N-M})', \quad \underset{\sim}{z} = (Z_{N-M+1}, \cdots, Z_N)',$$
$$\hat{z} = (\hat{Z}_1, \cdots, \hat{Z}_{N-M})', \quad \hat{\underset{\sim}{z}} = (\hat{Z}_{N-M+1}, \cdots, \hat{Z}_N)'.$$

With this notation, (5.11) and (5.12) become

$$KW(z - \hat{z}) - \lambda K c = 0, \tag{5.15}$$

$$\underset{\sim}{z} - \hat{\underset{\sim}{z}} = -\underset{\sim}{W}^{-1} T' W(z - \hat{z}). \tag{5.16}$$

Since K, as defined in (5.14), is symmetric positive definite, (5.15) implies

$$c = \frac{1}{\lambda} W(z - \hat{z}), \tag{5.17}$$

while, by (5.3), (5.4) and (5.6),

$$c_i = (-1)^m \Delta^m \hat{g} = \gamma_i \quad \text{in } \Omega_i, \qquad i = 1, \cdots, N-M. \tag{5.18}$$

Therefore, (5.10) holds for $i = 1, \cdot\cdot, \mathrm{N}-\mathrm{M}$, and (5.8) becomes

$$(\gamma_{N-M+i}, \cdots, \gamma_N)' = -T'(\gamma_1, \cdots, \gamma_{N-M})' = -T'c = -\frac{1}{\lambda} T' W(z - \hat{z}). \tag{5.19}$$

Comparing (5.19) with (5.16) we conclude that (5.10) holds for $i = N-M+1, \cdots, N$ as well.

A direct consequence of Theorem 5.1, the representation (5.3) of g and (5.4), is:

COROLLARY 5.1. *The solution of Problem* II-m *is of the form*

$$\hat{g} = \frac{1}{\lambda} \sum_{i=1}^{N-M} w_i(Z_i - \hat{Z}_i)\xi_i + \sum_{i=1}^{M} \hat{Z}_{N-M+i} \tilde{q}_i, \tag{5.20}$$

and satisfies the integro differential equation

$$(-1)^m \Delta^m \hat{g} = \frac{1}{\lambda} \sum_{i=1}^{N} \chi_{\Omega_i} w_i \left[Z_i - \int_{\Omega_i} \hat{g} \right] \tag{5.21}$$

with boundary conditions

$$\delta_j \hat{g} = 0, \qquad m \leqq j \leqq 2m-1. \tag{5.22}$$

Equations (5.21), (5.22) indicate an alternative direct way for the computation of $\hat{g}$, avoiding the computation of the functions $\xi_1, \cdots, \xi_{N-M}$.

We conclude this section by deriving explicitly the relation between the vector of given data $Z = (Z_1, \cdots, Z_N)'$ and the vector of smoothed data $\hat{Z} = (\hat{Z}_1, \cdots, \hat{Z}_N)'$.

From (5.20) we get $\hat{z} = (1/\lambda) KW(z - \hat{z}) + T\hat{\xi}$ and, after substituting for $\hat{\xi}$ from (5.16),

$$\hat{z} = T\underset{\sim}{\xi} + T\underset{\sim}{W}^{-1} T' W(z - \hat{z}) + \frac{1}{\lambda} KW(z - \hat{z}). \tag{5.23}$$

With $B = (I + (1/\lambda)KW + T\underset{\sim}{W}^{-1} T' W)^{-1}$, (5.23) and (5.16) become

$$z - \hat{z} = B(z - T\underset{\sim}{\xi}), \qquad \underset{\sim}{\xi} - \hat{\xi} = -\underset{\sim}{W}^{-1} T' WB(z - T\underset{\sim}{\xi}). \tag{5.24}$$

Combining the last two expressions, we conclude that

$$\hat{Z} = A(\lambda) Z \tag{5.25}$$

with

$$I - A(\lambda) = \begin{pmatrix} B & -BT \\ -\underset{\sim}{W}^{-1} T' WB & \underset{\sim}{W}^{-1} T' WBT \end{pmatrix}. \tag{5.26}$$

6. The problem of choosing λ. We give a procedure for choosing λ in Problem II. In this section we suppose (inaccurately!) that the $\{w_i\}$ in the definition of Problem II are given positive constants. In the problem presented in the introduction we want $w_i = 1/\text{variance } Z_i = 1/\mu_i$. Since the μ_i are being estimated, the w_i can be chosen iteratively by one of several obvious ad hoc procedures. In what follows, the w_i are assumed fixed and given. It is likely that $w_i \equiv 1$ will give reasonable answers in most cases when the μ_i are all of the same order of magnitude.

A good criteria for choosing λ is the minimization of $R(\lambda)$, defined by

$$R(\lambda) = E \sum_{i=1}^{N} \theta_i \left(\mu_i - \int_{\Omega_i} \hat{g}_\lambda \right)^2, \tag{6.1}$$

where E is expected value, $\hat{g}_\lambda$ is the solution to Problem II and the θ_i are given positive weights. Since the μ_i are not known, we cannot minimize $R(\lambda)$. However, an unbiased estimate $\hat{R}(\lambda)$ of $R(\lambda)$ is available by generalizing an observation in Craven and Wahba [5]. Let $A(\lambda)$ be the $N \times N$ matrix satisfying

$$A(\lambda) Z = \begin{pmatrix} \int_{\Omega_1} \hat{g} \\ \int_{\Omega_2} \hat{g} \\ \vdots \\ \int_{\Omega_N} \hat{g} \end{pmatrix}.$$

Such a matrix is given explicitly in (5.25), (5.26).

Then (6.1) becomes

$$R(\lambda) = E\|D_\theta^{1/2}(\mu - A(\lambda)Z)\|^2,$$

where $D_\theta = \operatorname{diag}\{\theta_1, \cdots, \theta_N\}$ and $\mu = (\mu_1, \cdots, \mu_N)'$. Defining $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_N)'$ by $Z = \mu + \varepsilon$, we have

$$\begin{aligned} E\|D_\theta^{1/2}(\mu - A(\lambda)Z)\|^2 &= E\|D_\theta^{1/2}[(I - A(\lambda))\mu - A(\lambda)\varepsilon]\|^2 \\ &= \|D_\theta^{1/2}(I - A(\lambda)\mu\|^2 + \operatorname{Tr} D_\theta A(\lambda)\Sigma A'(\lambda), \end{aligned}$$

where $\Sigma = \operatorname{diag}\{\operatorname{var} Z_1, \operatorname{var} Z_2, \cdots, \operatorname{var} Z_N\} = \operatorname{diag}\{\mu_1, \mu_2, \cdots, \mu_N\}$.

Let $\hat{\Sigma} = \operatorname{diag}\{Z_1, \cdots, Z_N\}$. We claim that an unbiased estimate $\hat{R}(\lambda)$ of $R(\lambda)$ is given by

$$\begin{aligned} \hat{R}(\lambda) = \|D_\theta^{1/2}(I - A(\lambda))Z\|^2 &- \operatorname{Tr} D_\theta^{1/2}(I - A(\lambda))\hat{\Sigma}(I - A(\lambda)')D_\theta^{1/2} \\ &+ \operatorname{Tr} D_\theta^{1/2}A(\lambda)\hat{\Sigma}A(\lambda)'D_\theta^{1/2}. \end{aligned} \tag{6.2}$$

In fact, (6.2) simplifies to

$$\hat{R}(\lambda) = \|D_\theta^{1/2}(I - A(\lambda))Z\|^2 + \sum_{i=1}^{N} \theta_i Z_i - 2 \operatorname{Tr} D_\theta \hat{\Sigma}(I - A(\lambda)). \tag{6.3}$$

To confirm our claim, observe that

$$\begin{aligned} E\|D_\theta^{1/2}(I - A(\lambda))Z\|^2 &= \|D_\theta^{1/2}(I - A(\lambda))\mu\|^2 \\ &+ \operatorname{Tr} D_\theta^{1/2}(I - A(\lambda))\Sigma(I - A(\lambda)')D_\theta^{1/2}, \end{aligned} \tag{6.4}$$

and

$$E\hat{\Sigma} = \Sigma. \tag{6.5}$$

Substituting (6.4) into (6.2) and using (6.5), we obtain $E\hat{R}(\lambda) = R(\lambda)$. Thus, it is reasonable to choose λ by minimizing $\hat{R}(\lambda)$.

7. Laplacian histosplines for a modified smoothness criterion. Problems in coding a numerical algorithm for computing $\hat{u}$ and $\hat{g}$ related to solving the Neumann boundary value problem in an irregular domain can be avoided by modifying the smoothing criterion somewhat.

Whether or not this modified smoothing criterion gives results equally pleasing as the smoothing criterion previously used, and whether the computing time required is comparable or not remain to be seen. However, the coding of an algorithm for the modified criterion appears to be relatively straightforward and is similar to already existing codes for the case of point evaluation data [13], [15], [19].

The results below are modest generalizations of results given by Duchon [6], [7], and later discussed by Meinguet [13] and Wahba [19].

We let $d = 2$; however, the generalization to arbitrary d dimensions is immediate from the known results whenever $2m - d > 0$. Let X be a suitable[1] space of functions on R^2 for which

$$\tilde{J}_m(u) = \int\int_{R^2} \sum_{j=0}^{m} \binom{m}{j} \left(\frac{\partial^m u}{\partial x_1^{m-j}\, \partial x_2^j} \right)^2 \tag{7.1}$$

is well defined and finite.

[1] X is the Beppo–Levi space of all the Schwartz distributions for which all the partial derivatives in the distributional sense of total order m are square integrable in R^2 [13].

We modify Problems I-m and II-m to the following:

Problem $\tilde{\text{I}}$-m. Find $u \in X$ to minimize $\tilde{J}_m(u)$ subject to

$$\int_{\Omega_i} u(x_1, x_2)\, dx_1\, dx_2 = s_i, \qquad i = 1, 2, \cdots, N.$$

Problem $\tilde{\text{II}}$-m. Find $g \in X$ to minimize

$$\sum_{i=1}^{N} w_i \left[Z_i - \int_{\Omega_i} g(x_1, x_2)\, dx_1\, dx_2 \right]^2 + \lambda \tilde{J}_m(g).$$

Usually, we will only be interested in the restriction of u or g to Ω. If $\tilde{u}$ is the solution to Problem $\tilde{\text{I}}$-m, clearly $\tilde{J}_m(\tilde{u}) \geqq J_m(\tilde{u}) \geqq J_m(\hat{u})$, and equality will be obtained iff $\hat{u}$ can be extended to all of R^2 in such a way that the extension $\hat{\hat{u}}$ is in X and satisfies

$$\sum_{j=0}^{m} \binom{m}{j} \left(\frac{\partial^m \hat{\hat{u}}}{\partial x_1^j\, \partial x_2^{m-j}} \right)^2 = 0 \quad \text{for } (x_1, x_2) \notin \Omega.$$

Generally this is not possible, but is always possible in the case of one-dimensional histosplines. Moreover such an extension is also possible for domains with radial symmetry, as in the example of § 4, which is, essentially, a univariate problem in r. Indeed, by defining $\tilde{u}(r) = \hat{u}(r)$, $0 \leqq r \leqq R_n$, $\tilde{u}(r) = \hat{u}(R_n)$, $R_n \leqq r$, with $\hat{u}$, the solution in § 4, we get $J(\hat{u}) = \tilde{J}(\tilde{u})$, where both $\hat{u}$ and $\tilde{u}$ match the same volume data.

The solution to Problems $\tilde{\text{I}}$-m and $\tilde{\text{II}}$-m can be given explicitly; we do this later. However a representation of a computable approximate solution for $m \geqq 2$ can be obtained quickly from the known results, and we proceed to do this. Let $x = (x_1, x_2)$, and let $\{t_l\}_{l=1}^{n}$ be a fine regular mesh of points in Ω, $t_l = (x_1^l, x_2^l)$, such that

$$\int_{\Omega_i} u(x_1, x_2)\, dx_1\, dx_2 \cong a_i \sum_{t_l \in \Omega_i} u(t_l), \quad u \in H^m(\Omega),$$

where $a_i = |\Omega_i|/n_i$, $|\Omega_i|$ being the area of Ω_i and n_i the number of mesh points in Ω_i. We now consider

Problem $\tilde{\text{I}}$-m-$\{t_l\}$. Find $u \in X$ to minimize $\tilde{J}_m(u)$, subject to

$$a_i \sum_{t_l \in \Omega_i} u(t_l) = s_i, \qquad i = 1, 2, \cdots, N.$$

Problem $\tilde{\text{II}}$-m-$\{t_l\}$. Find $g \in X$ to minimize

$$\sum_{i=1}^{N} w_i \left[Z_i - a_i \sum_{t_l \in \Omega_i} g(t_l) \right]^2 + \lambda \tilde{J}_m(g).$$

THEOREM 7.1. *Suppose the $N \times M$ matrix T with*

$$T_{j\nu} = a_i \sum_{t_k \in \Omega_j} q_\nu(t_k) \tag{7.2}$$

is of rank M. Then the solutions to problems $\tilde{\text{I}}$-m-$\{t_l\}$ and $\tilde{\text{II}}$-m-$\{t_l\}$ are unique and have representations

$$\begin{aligned} u(x) &= \sum_{j=1}^{N} c_j \eta_j(x) + \sum_{\nu=1}^{M} d_\nu q_\nu(x), \\ g_\lambda(x) &= \sum_{j=1}^{N} c_j \eta_j(x) + \sum_{\nu=1}^{M} d_\nu q_\nu(x), \end{aligned} \tag{7.3}$$

where

$$\eta_i(x) = a_i \sum_{t_l \in \Omega_i} E_m(x - t_l),$$

$$E_m(x) = \theta_m |x|^{2m-2} \log |x|, \qquad \theta_m = \{2^{2m-1}\pi[(m-1)!]^2\}^{-1},$$

$$|x| = \sqrt{x_1^2 + x_2^2},$$

and $\{q_\nu(x)\}_1^M$ *span the space of polynomials of total degree less than m. The coefficients* $c = (c_1, \cdots, c_N)'$ and $d = (d_1, \cdots, d_M)'$ *satisfy the following equations*:

Problem $\tilde{\text{I}}$-m-$\{t_l\}$:

$$Kc + Td = s, \tag{7.4}$$

$$T'c = 0, \tag{7.5}$$

where K *is the* $N \times N$ *matrix with* ijth *entry*

$$K_{ij} = a_i a_j \sum_{\substack{t_k \in \Omega_i \\ t_l \in \Omega_j}} E_m(t_k, t_l), \qquad i, j = 1, \cdots, N, \quad s = (s_1, \cdots, s_N)'.$$

Problem $\tilde{\text{II}}$-m-$\{t_l\}$:

$$(K + \lambda W^{-1})c + Td = z, \tag{7.6}$$

$$T'c = 0, \tag{7.7}$$

where $W = \text{diag}\{w_1, \cdots, w_N\}$, *and* $Z = (Z_1, \cdots, Z_N)'$.

Proof. The special case $n_i = a_i = w_i = 1$, $i = 1, 2, \cdots, N$ is just the problem of interpolating or smoothing evaluation data, and in this case the result has been given explicitly in [6], [7], [13], [19]. The extension to the case of general n_i, a_i and w_i is straightforward from these results and is omitted.

Observe that the solution to Problem $\tilde{\text{I}}$-m-$\{t_l\}$ can be obtained by solving (7.6) and (7.7) for the solution of Problem $\tilde{\text{II}}$-m-$\{t_l\}$, with $\lambda = 0$ and Z replaced by s. We now put (7.6) and (7.7) in a form suitable for the computation of c, d and $\hat{R}(\lambda)$. Let R be any $N \times (N - M)$ matrix satisfying $R'T = 0$. Since $T'c = 0$, there exists a unique $N - M$ vector b, say, with $c = Rb$. Left multiplying (7.6) by R' and substituting $c = Rb$ gives

$$R'(K + \lambda W^{-1})Rb = R'Z. \tag{7.8}$$

We next assert that $R'KR$ is strictly positive definite. To prove this we use the following result [6]:

Suppose $t_1, \cdots, t_n$ do not fall on a straight line. Let $f = (f_1, \cdots, f_n)'$ be any nonzero vector satisfying

$$\sum_{i=1}^{n} f_i q_\nu(t_i) = 0, \qquad \nu = 1, 2, \cdots, M;$$

then $\sum_{i,j=1}^{n} (f_i f_j E_m(t_i - t_j)) > 0$. We need to show that if $r = (r_1, \cdots, r_N)'$ satisfies $T'r = 0$, then $r'Kr > 0$. Let F be the $n \times N$ matrix with jkth entry a_k if $t_j \in \Omega_k$ and 0 otherwise, let E be the $n \times n$ matrix with jkth entry $E_m(t_j - t_k)$, and let $\tilde{T}$ be the $n \times M$ matrix with $j\nu$th entry $q_\nu(t_j)$. Then $K = F'EF$ and $T = F'\tilde{T}$. Suppose $T'r = 0$. Then, if $f = F'r$, we have $\tilde{T}'f = T'r = 0$ and so $0 < f'Ef = r'F'EFr = r'Kr$.

In case $\lambda = 0$ or λ is a given positive constant, b is obtained from (7.8), $c = Rb$ and d is obtained from (7.6) as the solution of the system

$$(T'T)d = T(Z - (K + \lambda W^{-1})c). \tag{7.9}$$

We proceed to the case where we choose λ according to § 6. To compute $\hat{R}(\lambda)$ we first obtain an expression for $A(\lambda)$. The appropriate definition of $A(\lambda)$ is

$$A(\lambda)Z = \begin{pmatrix} a_1 & \sum_{t_l \in \Omega_1} \hat{g}(t_l) \\ a_2 & \sum_{t_l \in \Omega_2} \hat{g}(t_l) \\ & \vdots \\ a_N & \sum_{t_l \in \Omega_N} \hat{g}(t_l) \end{pmatrix}.$$

Using the fact that $a_i \sum_{t_l \in \Omega_i} (\eta_j(t_l)) = K_{ij}$, one obtains, from (7.3),

$$A(\lambda)Z = Kc + Td. \tag{7.10}$$

Combining (7.6) and (7.10), we get

$$(I - A(\lambda))Z = (K + \lambda W^{-1})c + Td - (Kc + Td) = \lambda W^{-1}c.$$

Since, by (7.8) and the definition of b, $c = Rb = R(R'(K + \lambda W^{-1})R)^{-1}R'Z$, we finally obtain

$$I - A(\lambda) = \lambda W^{-1}R[R'(K + \lambda W^{-1})R]^{-1}R'. \tag{7.11}$$

R can always be chosen so that $R'W^{-1}R = I_{N-M'}$ giving $I - A(\lambda) = \lambda W^{-1}R(B + \lambda I)^{-1}R'$, where $B = R'KR$ is a symmetric positive definite matrix. Now, let UD_BU' be the eigenvalue decomposition of B with $D_B = \operatorname{diag}\{b_1, \cdots, b_{N-M}\}$; then

$$I - A(\lambda) = \lambda W^{-1}RU(D_B + \lambda I)^{-1}U'R'. \tag{7.12}$$

Recalling the expression (6.3) for $\hat{R}(\lambda)$,

$$\hat{R}(\lambda) = \|D_\theta^{1/2}(I - A(\lambda))Z\|^2 + \sum_{i=1}^{N} \theta_i Z_i - 2 \operatorname{Tr}\{D_\theta \hat{\Sigma}(I - A(\lambda))\}, \tag{7.13}$$

and substituting (7.12), we obtain

$$\hat{R}(\lambda) = \lambda^2 \sum_{i,j=1}^{N-M} h_{ij} \frac{v_i}{b_i + \lambda} \frac{v_j}{b_j + \lambda} + \sum_{i=1}^{N} \theta_i Z_i - 2\lambda \sum_{i=1}^{N-M} \frac{l_{ii}}{b_i + \lambda}, \tag{7.14}$$

where

$$v = (v_1, \cdots, v_{N-M})' = U'R'Z,$$

$$H = \{h_{ij}\} = U'R'W^{-1}D_\theta W^{-1}RU = U'R' \operatorname{diag}\left\{\frac{\theta_1}{w_1^2}, \cdots, \frac{\theta_N}{w_N^2}\right\}RU,$$

$$L = \{l_{ij}\} = U'R'\hat{\Sigma}D_\theta W^{-1}RU = U'R' \operatorname{diag}\left\{\frac{Z_1\theta_1}{w_1}, \cdots, \frac{Z_N\theta_N}{w_N}\right\}RU.$$

In the special case $D_\theta = W$, the matrix H is I since $R'W^{-1}R = I$, and then (7.14) simplifies to

$$\hat{R}(\lambda) = \lambda^2 \sum_{i=1}^{N-M} \frac{v_i^2}{(b_i+\lambda)^2} + \sum_{i=1}^{N} w_i Z_i - 2\lambda \sum_{i=1}^{N-M} \frac{l_{ii}}{b_i+\lambda}. \tag{7.15}$$

With the expression for $\hat{R}(\lambda)$ in (7.14) (or (7.15)), repeated computations of $\hat{R}(\lambda)$ for different values of λ are straightforward, once the matrix H, the vector v and the diagonal of the matrix L are computed. Hence the value of λ minimizing $\hat{R}(\lambda)$ can be computed by standard minimization methods.

We remark here without proof that the arguments in [13] can be used here to prove that the solutions to Problems $\tilde{\text{I}}$-m and $\tilde{\text{II}}$-m have representations of the form

$$\sum_{j=1}^{N} c_j \psi_j(x) + \sum_{\nu=1}^{M} d_\nu q_\nu(x),$$

where

$$\psi_j(x) = \int_{\Omega_j} E_m(x, t)\, dt_1\, dt_2, \qquad t = (t_1, t_2) \tag{7.16}$$

and the $\{q_\nu\}$ are as before. The vectors c and d satisfy equations of the form (7.4) and (7.5) with K_{ij} and $T_{j\nu}$ given by

$$K_{ij} = \int_{\Omega_i}\int_{\Omega_j} E_m(x, t), \qquad T_{j\nu} = \int_{\Omega_j} q_\nu(x).$$

Since E_m is the fundamental solution of the iterated Laplacian (see [4] § V; [17] p. 47),

$$\Delta^m \psi_j(x) = 1,\ x \in \Omega_j, \qquad \Delta^m \psi_j(x) = 0,\ x \notin \Omega_j.$$

Therefore the solutions $\hat{u}$ and $\hat{g}$ to Problems $\tilde{\text{I}}$-m and II-m satisfy $\Delta^m \hat{u} = 0$, $\Delta^m \hat{g} = 0$ outside Ω and $\Delta^m \hat{u}$, $\Delta^m \hat{g}$ are constant on each Ω_i.

Note added in proof. Further results on the volume matching problem may be found in [20], [21]. Contour maps for some Wisconsin cancer mortality rates by county using the solution to problem $\tilde{\text{I}}$-2 may be found in [22].

REFERENCES

[1] J. P. AUBIN, *Approximation of Elliptic Boundary-Value Problems*, Wiley-Interscience, New York, 1972.

[2] L. BONEVA, D. KENDALL AND I. STEFANOV, *Spline transformations: Three new diagnostic aids for the statistical data analyst (with discussion)*. J. Roy. Statist. Soc. Ser. B, 33 (1971), pp. 1–70.

[3] R. COURANT, *Differential and Integral Calculus*, Vol. II, Interscience, New York, 1956.

[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. 1, Interscience, New York, 1953.

[5] P. CRAVEN AND G. WAHBA, *Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation*, Numer. Math., 31 (1979), pp. 377–403.

[6] J. DUCHON, *Interpolation des fonctions de deux variables suivante le principe de la flexion des plaques minces*, R.A.I.R.O. Analyse Numerique 10, 12 (1976), pp. 5–12.

[7] ———, *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*, in Constructive Theory of Functions of Several Variables, Proc. 1976 Oberwolfach Conference, W. Schempp and K. Zeller, eds., Springer-Verlag, Berlin, 1976, pp. 85–100.

[8] N. DYN, G. WAHBA AND W. H. WONG, *Comments to "Smooth pycnophylactic interpolation for geographical regions", by W. Tobler*, J. Amer. Statist. Assoc., 74 (1979), pp. 530–535.

[9] M. GOLOMB AND H. T. WEINBERGER, *Optimal approximation and error bounds*, On Numerical Approximation, R. E. Langer, ed., University of Wisconsin Press, Madison, 1959, pp. 117–190.
[10] S. KARLIN, *Total Positivity*, Vol. I. Stanford University Press, Stanford Ca., 1968.
[11] G. KIMELDORF AND G. WAHBA, *Some results on Tchebycheffian splines*, J. Math. Anal. Applic., 33 (1971), pp. 82–95.
[12] J. L. LIONS AND G. STAMPACCHIA, *Variational inequalities*, Comm. Pure Appl. Math., 20 (1967), pp. 493–519.
[13] J. MEINGUET, *Multivariate interpolation at arbitrary points made simple*, Zeit. Angew. Math. Phys., to appear.
[14] J. NEÇAS, *Les méthodes directes dans la théorie des équations aux derivées partielles*, Masson, Paris, 1967.
[15] L. PAIHUA MONTES AND F. UTRERAS DIAZ, *Un ensemble de programmes pour l'interpolation des fonctions, par des fonctions spline du type plaque mince*, R. R. no. 140, Mathématiques appliqués, Université Scientifique et Médicale de Grenoble, Oct. 1978.
[16] I. SCHOENBERG AND C. DE BOOR, *Splines and histograms*, TSR #1273, Mathematics Research Center, University of Wisconsin-Madison, October 1972.
[17] L. SCHWARTZ, *Théorie des distributions*, Hermann, Paris, 1966.
[18] W. TOBLER, *Smooth pycnophylactic interpolation for geographical regions*, J. Amer. Statist. Assoc., 74 (1979), pp. 519–529.
[19] G. WAHBA, *How to smooth curves and surfaces with splines and cross-validation*, TR #555, Statistics Department, University of Wisconsin-Madison, March 1979.
[20] N. DYN AND W. FERGUSON, *Numerical construction of smooth surface from aggregated data*, TSR #2129, Mathematics Research Center, University of Wisconsin-Madison, October 1980.
[21] W. H. WONG, *On constrained splines and their approximations*, TR # 127, Dept. of Statistics, University of Chicago, Chicago, IL, April 1981.
[22] G. WAHBA, *Numerical experiments with the thin plate histospline*, Commun. Statist., Series A, to appear.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0011 $01.00/0

SYMMETRIZATION WITH EQUAL DIRICHLET INTEGRALS*

MARIE-THÉRÈSE KOHLER-JOBIN†

Abstract. Using a symmetrization method first introduced in [7] and developed further in [8], [9], we give, in particular, isoperimetric bounds for the functionals

$$\inf_{v(x)} \left\{ \frac{[\int_D |\nabla v|^2 \, dx]^\alpha}{\int_D v^{2\alpha} \, dx} \right\}, \qquad \alpha > \tfrac{1}{2},$$

sharper than those given in Crooke [Applicable Anal., 3 (1974), pp. 345–378], [Colloq. Math., 38 (1978), pp. 263–267], Crooke and Sperb [SIAM J. Math. Anal., 9 (1978), pp. 671–681].

1. Introduction

1.1. Let D be a bounded domain of $\mathbb{R}^N$, with a piecewise analytic boundary ∂D. By $\mathcal{C}(D)$ we mean the space of all real-valued functions $v(x)$, vanishing on ∂D, continuous on $\bar{D}$ and piecewise continuously differentiable in D, for which the Dirichlet integral $\int_D |\nabla v|^2 \, dx$ is finite. ∇v is the gradient of $v(x)$, namely

$$\nabla v = \left(\frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2}, \cdots, \frac{\partial v}{\partial x_N} \right), \qquad x = (x_1, x_2, \cdots, x_N).$$

Let α be a positive number. We then define the decreasing domain functional

$$c^2(\alpha, D) = \inf_{v(x) \in \mathcal{C}(D)} \left\{ \frac{[\int_D |\nabla v|^2 \, dx]^\alpha}{\int_D v^{2\alpha} \, dx} \right\}. \tag{1}$$

We point out that in the case $\alpha = 1$, the solution of the variational problem above is given by the first eigenvalue of the Laplace operator, and in the case $\alpha = 1/2$ and D simply connected and plane, the variational problem (1) is actually the torsion problem [11].

In [3] Crooke, using the Schwarz symmetrization [1], [2], got the following result.

THEOREM 1. *Let D be a three-dimensional bounded domain and let $\tilde{D}$ denote the sphere of $\mathbb{R}^3$ having the same volume as D. Then*

$$c^2(2, D) \geqq c^2(2, \tilde{D}).$$

In a second paper [4], the same author computed $c^2(2, \tilde{D})$: If $\tilde{R}$ is the radius of $\tilde{D}$, then

$$c^2(2, \tilde{D}) = \frac{8\pi \cdot z_0^2 [y'(z_0)]^2}{\tilde{R}}, \tag{2}$$

where z_0 denotes the first positive zero of the Emden–Fowler initial value problem [6].

$$\begin{aligned} &y'' + \frac{2}{z} y' + y^3 = 0, \qquad y = y(z), \quad ' = \frac{d}{dz} \\ &y(0) = 1, \quad y'(0) = 0. \end{aligned} \tag{3}$$

As indicated in [4], the same method enables us to compute isoperimetric bounds for the variational problem (1). [5] deals with the case $N = 2$.

* Received by the editors December 2, 1980, and in revised form April 7, 1981. This work was supported by the Swiss National Foundation of Science.

† Department of Mathematics, Stanford University, Stanford, California 94305. Current address: Planchettes 14, CH-2900 Porrentruy, Switzerland.

1.2. The Euler–Lagrange equation corresponding to (1) is

$$\Delta U + \frac{c^2(\alpha, D)}{[\int_D |\nabla U|^2 \, dx]^{\alpha-1}} U^{2\alpha-1} = 0 \quad \text{in } D. \tag{4}$$

It follows from Pohožaev's results [10] that:

- If $N=2$, the variational problem (1) has a positive solution $U(x)$, which belongs to $C^2(D) \cap C^\circ(\bar{D})$ and satisfies (4) for every positive α.
- If $N \geqq 3$, the same statement holds, provided that $\alpha < 1+2/(N-2)$.
- Furthermore, if $N \geqq 3$ and $\alpha \geqq 1+2/(N-2)$, the eigenvalue problem (4) possesses no weak solution in a starlike domain.

From now on we have to assume that (4) has a solution. This means that we assume that α satisfies

$$\alpha < 1 + \frac{2}{N-2} \quad \text{if } N \geqq 3. \tag{Hα}$$

The solution of the problem (1) then satisfies (4) and thus belongs to $C^\infty(D) \cap C^\circ(\bar{D})$.

Using the same method as Crooke [3], [4] and Crooke and Sperb [5], we get the following statement.

THEOREM 2. *Let $\tilde{D}$ be the N-ball of $\mathbb{R}^N$, with radius $\tilde{R}$ having the same measure as D. Then, if α satisfies* (Hα), *we have*

$$c^2(\alpha, D) \geqq c^2(\alpha, \tilde{D}) = \left[\frac{\alpha \omega_N}{N+\alpha(2-N)}\right]^{\alpha-1} \cdot z_0^{2\alpha} \cdot [y'(z_0)]^{2\alpha-2} \cdot \tilde{R}^{-[N+\alpha(2-N)]}, \tag{5}$$

where z_0 is the first positive zero of the initial value problem

$$\begin{aligned} &y'' + \frac{N-1}{z} y' + y^{2\alpha-1} = 0, \qquad y = y(z), \quad ' = \frac{d}{dz} \\ &y(0) = 1, \qquad y'(0) = 0, \end{aligned} \tag{6}$$

and ω_N is the $(N-1)$-measure of the unit N-ball.

The proof is exactly the same as those given in [3], [4], [5]; and therefore we will not repeat it here.

In the case $N=3$, the problem (6) is the Emden–Fowler initial value problem. The values of z_0 and $y'(z_0)$ are listed in [6] for some α. This enables us to give the following numerical results:

(7)

α	$c^2(\alpha, D)\tilde{R}^{3-\alpha}$
$\frac{1}{2}$	1.892342
1	9.869582
$\frac{5}{4}$	19.91524
$\frac{3}{2}$	34.20395
$\frac{7}{4}$	64.29614
2	102.3737
$\frac{9}{4}$	148.8229
$\frac{5}{2}$	193.2528
$\frac{11}{4}$	212.6305

1.3. An auxiliary boundary value problem. We now consider the functionals

$$P[v] = -\int_D |\nabla v|^2\, dx + 2\int_D v(x)\, dx, \qquad v(x) \in \mathscr{C}(D), \tag{8}$$

$$P(D) = \max_{v(x)\in\mathscr{C}(D)} P[v]. \tag{9}$$

It is well known that the maximum above is given by the function $u(x)$ solving the boundary value problem

$$\Delta u = -1 \quad \text{in } D, \qquad u = 0 \quad \text{on } \partial D; \tag{10}$$

and that

$$P(D) = \int_D u(x)\, dx = \int_D |\nabla u|^2\, dx. \tag{11}$$

We point out that $P(D) = 1/c^4(\frac{1}{2}, D)$, and that both corresponding extremal problems (1) (for the choice $\alpha = \frac{1}{2}$) and (9) are equivalent. In the particular case where $N = 2$ and D is simply connected, $P(D)$ is actually the torsional rigidity of the plane domain D. Because of this, we call $P(D)$ the torsional rigidity of D, even if $N \geqq 3$ and/or D is not simply connected.

By the Schwarz symmetrization, we get $P(D) \leqq P(\tilde{D})$, where $\tilde{D}$ denotes the N-ball of $\mathbb{R}^N$, with radius $\tilde{R}$, having the same measure as D. This is the generalization of the Saint-Venant and Pólya theorem [11]. (For further information on this theorem or on the Schwarz symmetrization see [11], [2].)

Now let D^* be the N-ball of $\mathbb{R}^N$, with radius R^*, having the same torsional rigidity as D. Thus, by definition,

$$P(D^*) = P(D). \tag{12}$$

Since $P(D)$ is an increasing domain functional, we have $R^* \leqq \tilde{R}$. The goal of this paper is to show the statement below. The latter sharpens Theorem 1 and Theorem 2.

THEOREM 3. *Let D^* be the N-ball of $\mathbb{R}^N$ having the same torsional rigidity as D. If $\alpha > \frac{1}{2}$ and if further, α satisfies* (Hα), *we have*

$$c^2(\alpha, D) \geqq c^2(\alpha, D^*).$$

We note that by definition (12), $c^2(\frac{1}{2}, D) = c^2(\frac{1}{2}, D^*)$. The next two sections deal with the proof of the foregoing theorem.

2. Symmetrization with constant Dirichlet integrals

2.1. Let $t(x)$ be any positive function of class $C^\infty(D) \cap C^0(\bar{D})$, taking the value zero on the boundary ∂D. To it corresponds a lower bound $P(D; t(x))$ of $P(D)$ given by the maximum principle

$$P(D; t(x)) = \max_{\substack{v(x)\in\mathscr{C}(D) \\ v(x)=\varphi(t(x))}} P[v], \tag{13}$$

where $\varphi(\bar{t})$ denotes a real-valued function of the real variable $\bar{t}$. According to [7], [9], we call $P(D; t(x))$ the "modified torsional rigidity of D with respect to the function $t(x)$".

Throughout this paper we will use the following notation:

$$D_{\bar{t}} := \{x \in D \mid t(x) > \bar{t}\}, \tag{14}$$

$$\Gamma_{\bar{t}} := \{x \in D \mid t(x) = \bar{t}\}, \tag{15}$$

$$t_{\max} := \max_{x\in D} t(x), \tag{16}$$

$$a(\bar{t}) := \int_{D_{\bar{t}}} dx, \tag{17}$$

$$\gamma(\bar{t}) := \int_{\Gamma_{\bar{t}}} |\nabla t|\, ds. \tag{18}$$

Because of the regularity of $t(x)$, we have

$$-\frac{da}{d\bar{t}} = \int_{\Gamma_{\bar{t}}} \frac{ds}{|\nabla t|}. \tag{19}$$

Using $\bar{t}$ as an independent variable, the functional (8) can be written as

$$P[\varphi(t(x))] = -\int_0^{t_{\max}} \left|\frac{d\varphi}{d\bar{t}}\right|^2 \gamma(\bar{t})\, d\bar{t} + 2\int_0^{t_{\max}} \varphi(\bar{t})\left(-\frac{da}{d\bar{t}}\right) d\bar{t}.$$

We now get the following statement about the functional $P(D; t(x))$.

LEMMA 1. *The function* $\Phi(\bar{t})$, *defined by*

$$\Phi(\bar{t}) = \int_{\xi=0}^{\bar{t}} \frac{a(\xi)}{\gamma(\xi)}\, d\xi, \tag{20}$$

solves the maximum problem (13). *Furthermore,*

$$\begin{aligned} P(D, t(x)) &= \int_0^{t_{\max}} \frac{a^2(\bar{t})}{\gamma(\bar{t})}\, dt = \int_D \Phi(t(x))\, dx \\ &= \int_D |\nabla \Phi(t(x))|^2\, dx. \end{aligned} \tag{21}$$

Proof. Because the proof stands already in [7], [9] we only give an outline of it. An integration by parts leads to

$$P[\Phi(t(x))] - P[\varphi(t(x))] = \int_0^{t_{\max}} \left|\frac{d\Phi}{d\bar{t}} - \frac{d\varphi}{d\bar{t}}\right|^2 \gamma(\bar{t})\, d\bar{t},$$

which is never negative and takes the value zero if and only if $\Phi(\bar{t}) = \varphi(\bar{t})$.

Applying the foregoing lemma to the partial domain $D_{\bar{t}}$, (14), we get

$$p(\bar{t}) := P(D_{\bar{t}}; t(x) - \bar{t}) = \int_{\xi=\bar{t}}^{t_{\max}} \frac{a^2(\xi)}{\gamma(\xi)}\, d\xi. \tag{22}$$

According to (13) and (21), $p(\bar{t})$ is the "modified torsional rigidity of the domain $D_{\bar{t}}$ with respect to the function $t(x) - \bar{t}$". The case where $D_{\bar{t}}$ is not connected is not excluded in (22).

COROLLARY 1. *Let* ω_N *be the* $(N-1)$*-measure of the unit N-ball. Then, we have*

$$p(\bar{t}) \leqq \frac{[a(\bar{t})]^{(N+2)/N}}{(N+2) \cdot N^{(N-2)/N} \cdot \omega_N^{2/N}},$$

and equality holds if and only if D is a N-ball and the level sets $\Gamma_{\bar{t}}$ *are concentric spheres.*

Proof. Applying the Schwarz inequality to (18) and (19), we get

$$\gamma(\bar{t})\left(-\frac{da}{d\bar{t}}\right) \geqq \left[\int_{\Gamma_{\bar{t}}} ds\right]^2.$$

We now use the geometric isoperimetric inequality

$$\left[\int_{\Gamma_{\bar{t}}} ds\right]^2 \geqq N^{2(N-1)/N} \cdot \omega_N^{2/N} \cdot [a(\bar{t})]^{2(N-1)/N}.$$

The inequalities above lead to

$$\gamma(\bar{t}) \geqq N^{2(N-1)/N} \cdot \omega_N^{2/N} \cdot [a(\bar{t})]^{2(N-1)/N} \left(-\frac{da}{d\bar{t}}\right)^{-1}.$$

We now use the latter and (22) to finish the proof of the corollary.

We point out that in the case $N = 2$ and D simply connected, the corollary implies that for the particular choice $t(x) = u(x)$ (where $u(x)$ solves the boundary value problem (10)),

$$P(D) \leqq \frac{\left[\int_D dx\right]^2}{8\pi}.$$

This is exactly the statement of the Saint-Venant and Pólya theorem [11], [2].

By means of $t(x)$, (22) defines a function $p(t(x))$ in D; and, by (20) and (17),

$$\int_{\Gamma_{\bar{t}}} |\nabla p(t(x))|\, ds = a^2(\bar{t}). \tag{23}$$

2.2. Choice of the comparison domain. We compare D with the N-ball of $\mathbb{R}^N$, $D^*(t(x))$, defined by

$$P(D^*(t(x))) = P(D; t(x)). \tag{24}$$

On the left side of the definition (24) stands the torsional rigidity of the comparison domain, whereas on its right side stands the "modified torsional rigidity of the domain D with respect to the function $t(x)$." Clearly, to each function $t(x)$ corresponds such a ball, and since $P(D; t(x)) \leqq P(D)$, the radius of any of those balls is never greater than R^*, the radius of the N-ball D^* defined by $P(D^*) = P(D)$, (12).

Now let $v(x)$ be a nonnegative function of $\mathscr{C}(D)$ such that $v(x) = \varphi(t(x))$, $\varphi(\bar{t})$ being a piecewise continuously differentiable function of the positive variable $\bar{t}$. Our goal is to construct a function $v^*(x)$, defined in $D^*(t(x))$, belonging to $\mathscr{C}(D^*(t(x)))$ and such that

$$\int_{D^*(t(x))} |\nabla v^*|^2\, dx = \int_D |\nabla v|^2\, dx.$$

For this purpose, we define a one-to-one correspondence between the level sets $\Gamma_{\bar{t}}$ of $t(x)$, and the concentric spheres of $D^*(t(x))$. In the latter, we use the spherical coordinate $r = |x| = (\sum_{i=1}^N |x_i|^2)^{1/2}$. Let $p^*(\bar{r})$ be the torsional rigidity of the concentric N-ball with radius $\bar{r}$, (9). Then

$$\bar{t} \leftrightarrow \bar{r} \quad \text{if and only if} \quad p(\bar{t}) = p^*(\bar{r}). \tag{25}$$

The quantities $p(\bar{t})$ and $p^*(\bar{r})$ stand as "common measure" between the domains D and $D^*(t(x))$. Since $p(\bar{t})$ is decreasing in $\bar{t}$, $p(0) = P(D; t(x))$, $p(t_{\max}) = 0$, the function $v(x) = \varphi(t(x))$ can be regarded as a function of the variable $\bar{p}$. Without ambiguity, we

now write

$$\varphi(\bar{p}) = \varphi(t(x))\big|_{t(x)=\bar{t}} \quad \text{if } p(\bar{t}) = \bar{p}.$$

In the same way we write $a(\bar{p})$ for the measure of the partial domain $D_{\bar{t}}$ such that $p(\bar{t}) = \bar{p}$. From now on, we denote with an asterisk * all the quantities related to the comparison domain $D^*(t(x))$. For example $a^*(\bar{p}^*)$ is the measure of the N-ball of radius $\bar{r}$ such that $p^*(\bar{r}) = \bar{p}^*$.

In analogy to (22), we define

$$p^*(x)\big|_{|x|=\bar{r}} = p^*(\bar{r}). \tag{22*}$$

DEFINITION. The function $v^*(x) = \varphi^*(p^*(x))$, where $\varphi^*(\bar{p}^*)$ satisfies

$$a^*(\bar{p}^*)d\varphi^* = a(\bar{p})\, d\varphi \quad \text{for } \bar{p}^* = \bar{p}, \tag{26}$$

$$\varphi^*(P(D^*(t(x)))) = 0 \tag{27}$$

is the "symmetrization of $v(x)$ with equal Dirichlet integrals."

This symmetrization possesses the following properties:

LEMMA 2.

(A) $v^*(x) \geqq 0$ *in* $D^*(t(x))$, $\quad v^*(x) = 0$ *on* $\partial(D^*(t(x)))$.

(B) $\varphi^*(\bar{p}^*) \geqq \varphi(\bar{p})$ *for* $\bar{p}^* = \bar{p}$.

(C) $$\int_{D^*(t(x))} |\nabla v^*|^2\, dx = \int_D |\nabla v|^2\, dx.$$

Proof. The first statement is an immediate consequence of (26) and (27). Since the "modified torsional rigidity" is an increasing domain functional, Corollary 1 leads to the inequality $a^*(\bar{p}^*) \leqq a(\bar{p})$ for $p^* = \bar{p}$. The latter and (26), (27) prove the statement (B). The last one follows from (23),

$$\int_D |\nabla v|^2\, dx = \int_0^{P(D;t(x))} \left|\frac{d\varphi}{d\bar{p}}\right|^2 a^2(\bar{p})\, d\bar{p},$$

and from the same argument for $v^*(x)$,

$$\int_{D^*(t(x))} |\nabla v^*|^2\, dx = \int_0^{P(D^*(t(x)))} \left|\frac{d\varphi^*}{d\bar{p}^*}\right|^2 [a^*(\bar{p}^*)]^2\, d\bar{p}^*.$$

We point out that this symmetrization conserves not only the global Dirichlet integral, but also the Dirichlet integrals in corresponding partial domains (25).

We list another important property of "symmetrization with equal Dirichlet integrals."

LEMMA 3. *Let* $f(v)$ *be any positive, increasing function of the positive variable* v. *Let further* $F(v) = \int_{\xi=0}^{v} f(\xi)\, d\xi$. *Then*

$$\int_D F(v(x))\, dx \leqq \int_{D^*(t(x))} F(v^*(x))\, dx.$$

Proof. The left integral can be written as

$$\int_0^{P(D;t(x))} F(\varphi(\bar{p}))\, \frac{da}{d\bar{p}}\, d\bar{p}.$$

After an integration by parts, we get

$$\int_D F(v(x))\,dx = \int_0^{P(D;\,t(x))} f(\varphi(\bar{p}))\left(-\frac{d\varphi}{d\bar{p}}\right)a(\bar{p})\,d\bar{p}.$$

But, by Lemma 2(B), $\varphi(\bar{p}) \leqq \varphi^*(\bar{p}^*)$ for $\bar{p} = \bar{p}^*$, and since $f(\varphi)$ is increasing, $f(\varphi(\bar{p})) \leqq f(\varphi^*(\bar{p}^*))$ for $\bar{p} = \bar{p}^*$. Furthermore, by the definition of $v^*(x)$ (26), (27), we have

$$\int_D F(v(x))\,dx \leqq \int_0^{P(D;t(x))} f(\varphi^*(\bar{p}^*))\left(-\frac{d\varphi^*}{d\bar{p}^*}\right)a^*(\bar{p}^*)\,d\bar{p}^*.$$

But $P(D;\, t(x)) = P(D^*(t(x)))$, and thus, after an integration by parts,

$$\int_D F(v(x))\,dx \leqq \int_0^{P(D^*(t(x)))} F(\varphi^*(\bar{p}^*))\left(\frac{da^*}{d\bar{p}^*}\right)d\bar{p}^*.$$

The latter is actually $\int_{D^*(t(x))} F(v^*(x))\,dx$. This finishes the proof of Lemma 3.

It is also clear that by its very definition $v^*(x)$ belongs to $\mathscr{C}(D^*(t(x)))$.

This symmetrization method was first introduced in [7] and developed further in [8], [9].

3. Applications of Lemma 3

3.1. Proof of Theorem 3. Let $U(x)$ be the solution of the Euler–Lagrange equation (4). Then

$$c^2(\alpha, D) = \frac{[\int_D |\nabla U|^2\,dx]^\alpha}{\int_D U^{2\alpha}\,dx}.$$

Let $D^*(U(x))$ be the N-ball of $\mathbb{R}^N$ such that $P(D^*(U(x))) = P(D;\, U(x))$, (24). Let further $U^*(x)$ be the "symmetrization of $U(x)$ with equal Dirichlet integrals" (26), (27). Then according to Lemma 2

$$\int_D |\nabla U|^2\,dx = \int_{D^*(U(x))} |\nabla U^*|^2\,dx.$$

Furthermore, since $\alpha > \frac{1}{2}$, we can apply Lemma 3. We get

$$\int_D U^{2\alpha}\,dx \leqq \int_{D^*(U(x))} [U^*]^{2\alpha}\,dx.$$

From the three last relations, we have

$$c^2(\alpha, D) \geqq \frac{[\int_{D^*(U(x))} |\nabla U^*|^2\,dx]^\alpha}{\int_{D^*(U(x))} [U^*]^{2\alpha}\,dx}.$$

The latter leads to

$$c^2(\alpha, D) \geqq c^2(\alpha, D^*(U(x))) = \inf_{v(x)\in\mathscr{C}(D^*(U(x)))} \left\{\frac{[\int_{D^*(U(x))} |\nabla v|^2\,dx]^\alpha}{\int_{D^*(U(x))} v^{2\alpha}\,dx}\right\}.$$

Since $c^2(\alpha, D)$ is a decreasing domain functional and since $D^* \supseteq D^*(U(x))$, we conclude that

$$c^2(\alpha, D) \geqq c^2(\alpha, D^*) \quad \text{for } \alpha > \tfrac{1}{2}.$$

3.2. Generalization of Theorem 3. We consider the following extremal problem. Given are a bounded domain D of $\mathbb{R}^N$ with a piecewise analytic boundary ∂D, a positive number C, and a positive, increasing function $f(v)$ for $v \geqq 0$. We define $F(v) = \int_{\xi=0}^{v} f(\xi)\, d\xi$, and we look for

$$\sup_{\substack{v(x)\in\mathscr{C}(D) \\ v(x)\geqq 0 \\ \int_D |\nabla v|^2 dx = C}} \left\{ \int_D F(v)\, dx \right\} =: M(D, C). \tag{28}$$

(Notice that we do not exclude the case where $M(D, C) = \infty$.) If there exists a maximal sequence $u_n(x)$ such that each $u_n(x)$ belongs to $C^\infty(D) \cap C^\circ(\bar{D})$, then

$$M(D, C) \leqq M(D^*, C), \tag{29}$$

where D^* denotes, as before, the N-ball of $\mathbb{R}^N$ having the same torsional rigidity as D.

The proof is easy. To each $u_n(x)$ corresponds a N-ball $D^*(u_n(x))$ defined by $P(D^*(u_n(x))) = P(D; u_n(x))$. According to Lemma 3, if $u_n{}^*(x)$ is the "symmetrization of $u_n(x)$ with equal Dirichlet integrals", then

$$\int_D F(u_n)\, dx \leqq \int_{D^*(u_n(x))} F(u_n^*)\, dx.$$

Since $M(D, C)$ is an increasing domain functional (for fixed C), we conclude that

$$M(D, C) = \lim_{n\to\infty} \int_D F(u_n)\, dx \leqq M(D^*, C).$$

3.3. Remark. We can extend the results of the present paper to some more general problems. Indeed, let now D be a plane, simply connected, bounded domain, and let $\rho(x)$ be a positive function of class $C^\infty(D) \cap C^0(\bar{D})$ satisfying

$$\Delta \ln \rho + 2K\rho \geqq 0 \quad \text{in } D$$

for some constant K. Using an idea due to Bandle [1], [2], we get isoperimetric bounds for the functional

$$\inf_{v(x)\in\mathscr{C}(D)} \left\{ \frac{[\int_D |\nabla v|^2\, dx]^\alpha}{\int_D v^{2\alpha} \rho\, dx} \right\}, \qquad \alpha > \tfrac{1}{2},$$

corresponding to those of this paper. The method follows exactly as that described in [8], [9].

Acknowledgment. I am thankful to Professor J. Hersch for suggesting Lemma 3 to me, and to Professor L. E. Payne for introducing me to some of the problems discussed in the present paper.

REFERENCES

[1] C. BANDLE, *Konstruktion isoperimetrischer Ungleichungen der mathematischen Physik aus solcher der Geometrie*, Comm. Math. Helv. 46 (1971), pp. 182–213.

[2] ———, *Isoperimetric Inequalities and Applications*, Pitman, London 1980.

[3] P. S. CROOKE, *On two inequalities of Sobolev type*, Applicable Anal. 3 (1974), pp. 345–378.

[4] ———, *An isoperimetric bound for a Sobolev constant*, Colloq. Math. 38 (1978), pp. 263–267.

[5] P. S. CROOKE AND S. P. SPERB, *Isoperimetric inequalities in a class of nonlinear eigenvalue problems*, this Journal, 9 (1978), pp. 671–681.

[6] R. H. DAVIS, *Introduction to Nonlinear Differential and Integral Equations*, Dover, New York 1962.

[7] M.-Th. KOHLER-JOBIN, *Démonstration de l'inégalité isopérimétrique* $P \cdot \lambda^2 \geqq \pi j_0^2/4$ *conjecturée par Pólya et Szegö*, C. R. Acad. Sc. Paris 281 (1975), pp. 119–121.

[8] ———, *Une inégalité isopérimétrique entre la fréquence fondamentale d'une membrane inhomogène et l'énergie d'équilibre du problème de Poisson correspondant*, C. R. Acad. Sc. Paris, 283 (1976), pp. 65–69.

[9] ———, *Une méthode de comparaison isopérimétrique de fonctionnelles de domaines de la physique mathémathique* I, II, Z. Angew. Math. Phys., 29 (1978), pp. 757–766, 767–776.

[10] S. I. POHOŽAEV, *Eigenfunction of the equation* $\Delta u + \lambda f(u) = 0$, Dokl. Akad. Nauk. USSR 165 (1965), pp. 36–39, Soviet Math. Dokl. 6 (1965), pp. 1408–1411.

[11] G. PÓLYA AND G. SZEGÖ, *Isoperimetric Inequalities in Mathematical Physics*, Princeton University Press, Princeton, NJ, 1951.

SIAM J. MATH. ANAL.
Vol. 13, No. 1, January 1982

0036-1410/82/1301-0012 $01.00/0

AN EXPLICIT FORMULA FOR $f(\mathscr{A})$ AND THE GENERATING FUNCTIONS OF THE GENERALIZED LUCAS POLYNOMIALS*

MASSIMO BRUSCHI† AND PAOLO EMILIO RICCI†

Abstract. From $\mathscr{A}^n = \sum_{k=1}^{r} F_{k,n-1}(I_1, \cdots, I_r)\mathscr{A}^{r-k}$, where $\mathscr{A}$ is a $r \times r$ matrix and $I_1, \cdots, I_r$ are the invariants of $\mathscr{A}$ (elementary symmetric functions of the eigenvalues), we first derive a formula for $f(\mathscr{A})$. Then we obtain the generating functions for the $F_{k,n}$ and thence for the generalized Lucas polynomials $F_{1,n}$, $n \geqq -1$.

Introduction. Consider the expression (see [19], [4])

$$\text{(i)} \qquad \mathscr{A}^n = \sum_{k=1}^{r} F_{k,n-1}(I_1, \cdots, I_r)\mathscr{A}^{r-k},$$

where n is an integer, $\mathscr{A}$ an $r \times r$ matrix and $I_1, \cdots, I_r$ are the invariants of $\mathscr{A}$ (elementary symmetric functions of the eigenvalues). From (i) we first derive the following formula for $f(\mathscr{A})$ ($f(\lambda)$ a holomorphic function of the complex variable λ):

$$\text{(ii)} \qquad f(\mathscr{A}) = \sum_{h=0}^{r-1} \left[\frac{1}{2\pi i} \int_{+\gamma} \frac{f(\lambda) \sum_{j=0}^{r-h-1} (-1)^j I_j \lambda^{r-h-j-1}}{\sum_{j=0}^{r} (-1)^j I_j \lambda^{r-j}} \, d\lambda \right] \mathscr{A}^h, \qquad I_0 = 1.$$

Note that (ii) implies Sylvester's matrix interpolation formula and does not in general require the knowledge of the Jordan canonical form of $\mathscr{A}$. Recently formulae for $e^{\mathscr{A}}$ only have been given [1], [16], [12].

Furthermore we determine the generating functions of the $F_{k,n}$, $k = 1, \cdots, r$, and therefore of the generalized Lucas polynomials (in r variables)

$$\sum_{n=0}^{\infty} F_{k,n+r-k-1}(I_1, \cdots, I_r) z^n = \frac{\sum_{j=0}^{k-1} (-1)^j z^j I_j}{\sum_{j=0}^{r} (-1)^j z^j I_j}.$$

Various papers have been devoted to the study of the above polynomials (see [21]) and to the extension of the algebraic theory of the Lucas numerical functions (see, e.g., [13], [22]).

1. Formulae on the powers of a matrix. Consider the $r \times r$ ($r \geqq 2$) complex matrix $\mathscr{A}$ and let

$$(1.1) \qquad \Delta(\lambda) = |\lambda \mathscr{I} - \mathscr{A}| = \sum_{j=0}^{r} (-1)^j I_j \lambda^{r-j}, \qquad I_0 = 1$$

be its characteristic polynomial. It is known (see [4]) that for every positive integer n we have

$$(1.2) \qquad \mathscr{A}^n = \sum_{k=1}^{r} F_{k,n-1}(I_1, \cdots, I_r)\mathscr{A}^{r-k}.$$

Furthermore, if $\mathscr{A}$ is a nonsingular matrix (1.2) holds also for n a negative integer.

The functions $F_{k,n}(I_1, \cdots, I_r)$, $k = 1, \cdots, r$, $n \geqq -1$ are defined by the recurrence relations

$$(1.3) \qquad F_{k,n}(I_1, \cdots, I_r) = \sum_{j=1}^{r} (-1)^{j+1} I_j F_{k,n-j}, \qquad k = 1, \cdots, r, \quad n \geqq r-1$$

* Received by the editors October 19, 1979, and in final revised form March 2, 1981.
† Istituto di Matematica Applicata, Università degli Studi di Roma, Rome, Italy.

and the initial conditions

$$F_{r-k+1,h-2}(I_1,\cdots,I_r)=\delta_{k,h},\qquad k,h=1,\cdots,r. \tag{1.4}$$

If $I_r\neq 0$, the functions $F_{k,n}(I_1,\cdots,I_r)$ for $n<-1$ are defined instead by

$$F_{k,n}(I_1,\cdots,I_r)=F_{r-k+1,-n+r-3}\left(\frac{I_{r-1}}{I_r},\cdots,\frac{I_1}{I_r},\frac{1}{I_r}\right),\qquad k=1,\cdots,r,\quad n<-1, \tag{1.5}$$

and again satisfy (1.3).

The functions $F_{1,n}(I_1,\cdots,I_r)$, $n\geqq -1$, are called in the literature *generalized Lucas polynomials* (see [2], [21]). The above results have been extended (see [4]) to a matrix whose minimal polynomial is known. In this paper, however, we shall suppose, for simplicity, that the characteristic polynomial is known. It is easily seen that the results hold also if the minimal polynomial is known.

2. Representation of the resolvent matrix $(\lambda\mathscr{I}-\mathscr{A})^{-1}$. By the method used in § 1 we find a representation for the resolvent matrix of $\mathscr{A}$ as a polynomial in $\mathscr{A}$ of degree $r-1$ (see [5, pp. 93–95]).

Let $J_1(\lambda),\cdots,J_r(\lambda)$ be the invariants of the characteristic matrix of $\mathscr{A}$. $J_k(\lambda)$, $k=1,\cdots,r$ is then the kth elementary symmetric function of the eigenvalues of $\mathscr{A}$. It is easy to verify that

$$J_k=J_k(\lambda)=\sum_{j=0}^{k}(-1)^j\binom{r-j}{k-j}I_j\lambda^{k-j},\qquad k=1,\cdots,r. \tag{2.1}$$

Now we can show:

THEOREM 1. *Under the same notation and assumptions as in* § 1, *the resolvent matrix* $(\lambda\mathscr{I}-\mathscr{A})^{-1}$ *may be represented by the following formula*:

$$(\lambda\mathscr{I}-\mathscr{A})^{-1}=\frac{1}{\Delta(\lambda)}\sum_{h=0}^{r-1}\left[\sum_{j=0}^{r-h-1}(-1)^jI_j\lambda^{r-h-j-1}\right]\mathscr{A}^h. \tag{2.2}$$

Proof. Put $J_0=1$ and $J_r=\Delta(\lambda)$; by virtue of the formulae (1.2), (1.5), (1.3) and (1.4) we can write

$$\begin{aligned}
(\lambda\mathscr{I}-\mathscr{A})^{-1}&=\sum_{k=0}^{r-1}F_{r-k,-2}(J_1,\cdots,J_r)(\lambda\mathscr{I}-\mathscr{A})^k\\
&=\frac{1}{\Delta(\lambda)}\sum_{k=0}^{r-1}(-1)^kJ_{r-k-1}(\lambda\mathscr{I}-\mathscr{A})^k\\
&=\frac{1}{\Delta(\lambda)}\sum_{k=0}^{r-1}\sum_{h=0}^{k}(-1)^{h+k}\binom{k}{h}J_{r-k-1}\lambda^{k-h}\mathscr{A}^h\\
&=\frac{1}{\Delta(\lambda)}\sum_{h=0}^{r-1}\sum_{k=h}^{r-1}(-1)^{h+k}\binom{k}{h}J_{r-k-1}\lambda^{k-h}\mathscr{A}^h\\
&=\frac{1}{\Delta(\lambda)}\sum_{h=0}^{r-1}\sum_{k=h}^{r-1}\sum_{j=0}^{r-k-1}(-1)^{h+k+j}\binom{k}{h}\binom{r-j}{k+1}I_j\lambda^{r-h-j-1}\mathscr{A}^h\\
&=\frac{1}{\Delta(\lambda)}\sum_{h=0}^{r-1}\sum_{j=0}^{r-h-1}(-1)^j\left[\sum_{k=h}^{r-j-1}(-1)^{h+k}\binom{k}{h}\binom{r-j}{k+1}\right]I_j\lambda^{r-h-j-1}\mathscr{A}^h.
\end{aligned} \tag{2.3}$$

Hence (2.2) follows, since $\sum_{k=h}^{r-j-1}(-1)^{h+k}\binom{k}{h}\binom{r-j}{k+1}=1$ as a particular case of a known identity on Gauss' hypergeometric function (see, e.g., [17, Ex. 4, p. 69]).

3. A representation for the matrix function $f(\mathscr{A})$. Let $f(\lambda)$ be holomorphic in a domain C of the complex plane, let $D \subset C$ be a "Cauchy domain" (see [20, p. 288]) containing all the eigenvalues of the matrix $\mathscr{A}$ and such that $\bar{D} \subset C$. Let γ be the boundary of D. Then $f(\mathscr{A})$ is defined by the Dunford–Taylor formula (see [10, p. 44])

$$f(\mathscr{A}) = \frac{1}{2\pi i} \int_{+\gamma} f(\lambda)(\lambda \mathscr{I} - \mathscr{A})^{-1} \, d\lambda. \tag{3.1}$$

Formula (3.1) is analogous to Cauchy's integral formula in the theory of holomorphic functions and it is also ascribed to L. Fantappié (see [8]) or to F. Riesz (see [18]). An equivalent statement is the following:

THEOREM 2. *Under the above assumptions, we have*

$$f(\mathscr{A}) = \sum_{h=0}^{r-1} \left[\frac{1}{2\pi i} \int_{+\gamma} \frac{f(\lambda) \sum_{j=0}^{r-h-1} (-1)^j I_j \lambda^{r-h-j-1}}{\Delta(\lambda)} \, d\lambda \right] \mathscr{A}^h. \tag{3.2}$$

Note that the line integrals which appear in (3.2) may be easily evaluated by the theorem of residues if the eigenvalues of the matrix $\mathscr{A}$ are given. Moreover, if the Jordan canonical form of the matrix is known, it is possible to deduce Sylvester's interpolation formula from (3.2) (see [6], [9, vol. I, pp. 101–103]).

However, (3.2) does not require the knowledge of either this canonical form or the eigenvalues of $\mathscr{A}$. It is in fact sufficient to know a "Cauchy domain" D ($\bar{D} \subset C$) enclosing all the eigenvalues of the matrix (see, e.g., [7, p. 20]). For example, if $f(\mathscr{A}) = e^{\mathscr{A}}$ (a function important in the theory of linear differential equations with constant coefficients), (3.2) becomes

$$e^{\mathscr{A}} = \sum_{h=0}^{r-1} \left[\frac{1}{2\pi i} \int_{+\gamma} \frac{e^{\lambda} \sum_{j=0}^{r-h-1} (-1)^j I_j \lambda^{r-h-j-1}}{\Delta(\lambda)} \, d\lambda \right] \mathscr{A}^h,$$

where as curve γ we may choose any circle with center at the origin and radius greater than the spectral radius of $\mathscr{A}$.

4. The generating functions for the functions $F_{k,n}(I_1, \cdots, I_r)$, $k = 1, \cdots, r$. The classical method to derive the generating functions (see, e.g., [3, p. 358] or [15, pp. 31–32]) leads us to state:

THEOREM 3. *The functions*

$$\frac{\sum_{j=0}^{k-1} (-1)^j z^j I_j}{\sum_{j=0}^{r} (-1)^j z^j I_j}, \qquad k = 1, \cdots, r \tag{4.1}$$

are, respectively, the generating functions of the polynomials $F_{k,n}(I_1, \cdots, I_r)$, $n \geqq -1$. In fact, there exist the following expansions in power series of the complex variable z:

$$\sum_{n=0}^{\infty} F_{k,n+r-k-1}(I_1, \cdots, I_r) z^n = \frac{\sum_{j=0}^{k-1} (-1)^j z^j I_j}{\sum_{j=0}^{r} (-1)^j z^j I_j}, \qquad k = 1, \cdots, r. \tag{4.2}$$

If $k = 1$, (4.2) gives, in particular, the generating function of the generalized Lucas polynomials (in r variables):

$$\sum_{n=0}^{\infty} F_{1,n+r-2}(I_1, \cdots, I_r) z^n = \frac{1}{\sum_{j=0}^{r} (-1)^j z^j I_j}.$$

Finally, the preceding statement allows us to determine the generating functions of the functions $F_{k,n}(I_1, \cdots, I_r)$ for $n < -1$. In fact, by (1.5) it is easy to prove that:

THEOREM 4. *Under the same notation and assumptions as in the preceding section, there exists the following expansion in power series of the complex variable* z:

$$\sum_{n=-1}^{-\infty} F_{k,n+r-k-1}(I_1, \cdots, I_r) z^n = -\frac{\sum_{j=0}^{k-1} (-1)^j z^j I_j}{\sum_{j=0}^{r} (-1)^j z^j I_j}, \qquad k = 1, \cdots, r.$$

Acknowledgments. The authors are greatly indebted to the referees for many valuable suggestions to improve the paper.

REFERENCES

[1] R. BARAKAT, *The matrix operator* e^X *and the Lucas polynomials*, J. Math. and Phys., 43 (1964), pp. 332–335.

[2] R. BARAKAT AND E. BAUMANN, *Mth power of an* $N \times N$ *matrix and its connection with the generalized Lucas polynomials*, J. Math. Phys., 10 (1969), pp. 1474–1476.

[3] S. BARNARD AND J. M. CHILD, *Higher Algebra*, Macmillan, London, 1952.

[4] M. BRUSCHI AND P. E. RICCI, *Sulle potenze di una matrice quadrata della quale sia noto il polinomio minimo*, Pubbl. Ist. Mat. Appl. Fac. Ingegneria Univ. Stud. Roma, 204 Quad. 13, 1979, pp. 9–18.

[5] S. CHERUBINO, *Calcolo delle Matrici*, Cremonese, Roma, 1957.

[6] M. CIPOLLA, *Sulle matrici espressioni analitiche di un'altra*, Rend. Circ. Mat. Palermo, 56 (1932), pp. 144–154.

[7] R. CONTI, *Linear Equations and Control*, Academic Press, London-New York, 1976.

[8] L. FANTAPPIÉ, *Le calcul des matrices*, C.R. Acad. Sci., Paris, 186 (1928), pp. 619–621.

[9] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1960.

[10] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, Heidelberg, New York, 1966.

[11] O. D. KELLOGG, *Foundations of Potential Theory*, Dover, New York, 1953.

[12] R. B. KIRCHNER, *An explicit formula for* e^{At}, Amer. Math. Monthly, 74 (1967), pp. 1200–1204.

[13] D. H. LEHEMER, *An extended theory of Lucas functions*, Ann. Math., 31 (1930), pp. 419–448.

[14] E. LUCAS, *Théorie des Nombres*, Gauthier-Villars, Paris, 1891.

[15] P. MONTEL, *Leçons sur les récurrences et leurs applications*, Gauthier-Villars, Paris, 1957.

[16] E. J. PUTZER, *Avoiding the Jordan canonical form in the discussion of linear systems with constant coefficients*, Amer. Math. Monthly, 73 (1966), pp. 2–7.

[17] E. D. RAINVILLE, *Special Functions*, Macmillan, New York, 1967.

[18] F. RIESZ, *Les systèmes d'équations linéaires à une infinité d'inconnues*, Gauthier-Villars, Paris, 1913.

[19] B. SEGRE, *Sulle potenze di una matrice quadrata d'ordine* $n \geqq 2$, Abh. Math. Sem. Univ. Hamburg, 47 (1978), pp. 71–78.

[20] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley–Toppan, New York, Tokyo, 1958.

[21] H. C. WILLIAMS, *Some properties of the general Lucas polynomials*, Matrix Tensor Quart., 21 (1971), pp. 91–93.

[22] ———, *On a generalization of the Lucas functions*, Acta Arith., 20 (1972), pp. 33–51.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0001 $01.00/0

SYSTEMS OF DIFFERENTIAL EQUATIONS WHICH ARE COMPETITIVE OR COOPERATIVE. I: LIMIT SETS*

MORRIS W. HIRSCH†

Abstract. A vector field in n-space determines a competitive (or cooperative) system of differential equations provided all the off-diagonal terms of its Jacobian matrix are nonpositive (or nonnegative). The principal result is that limit sets of such systems cannot be more complicated than invariant sets of systems of one lower dimension. In fact orthogonal projection along any positive direction maps a limit set homeomorphically and equivariantly onto an invariant set of a Lipschitz vector field in a hyperplane. Limit sets are nowhere dense, unknotted and unlinked. In dimension 2 every trajectory is eventually monotone. In dimension 3 a compact limit set which does not contain an equilibrium is a closed orbit or a cylinder of closed orbits.

Introduction. One of the most interesting questions to ask about a dynamical system is: what is the long-run behavior of its trajectories? In many systems it is natural to expect, or at least hope, that almost all trajectories either converge to an equilibrium or asymptotically approach a closed orbit (= periodic trajectory). Unfortunately there are many systems that not only lack this convenient property, but cannot even be approximated by systems that have it. Such systems are often said to be "chaotic" or to possess "strange attractors".

To make matters worse, it is very hard to discover the long-run behavior of any but the simplest systems. Research on this problem has bifurcated into two quite different methodologies. A great deal of recent work has gone toward exploring the consequences of various assumptions about the large scale structure of the system, e.g., hyperbolicity of the nonwandering set, structural stability, ergodicity, and so forth. The basic examples come from geometry and physics; the mathematical techniques tend to be topological. For a recent overview of this work see Smale [15, Chapt. I].

This structural approach is very useful for the conceptual understanding of dynamical systems, but usually it is of little direct help to the researcher who wants to understand a particular system. Not only is it extremely difficult to decide whether a particular system has a given structural feature, but many systems do not satisfy any of the axiom systems commonly used in the structural approach. In consequence much research has gone into determining the long-run behavior of special systems (or classes of systems) that arise as models in biology, chemistry, economics and so forth. Algebraic techniques play a prominent role, but owing to the diversity of systems studied very few general principles have been developed.

In this and subsequent articles I hope to make a start at bridging the gap between these two approaches by using structural ideas to analyze a fairly broad class of systems, namely those which are competitive or cooperative (defined below). Such systems are sometimes associated with the concept of negative or positive feedback. They have been used to model a variety of biological, chemical and economic systems; see, e.g., [1], [5], [8], [9], [11], [12], [13], [14].

A general principle emerging from this analysis is that in such systems, especially cooperative ones, there is a strong tendency for bounded trajectories to converge to

* Received by the editors November 12, 1980, and in revised form May 12, 1981. This research was supported in part by the National Science Foundation under grant MCS 77-04242.

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

equilibria or to periodic trajectories. This will be made more precise in later articles, but Theorem C below can be viewed as an instance of this phenomenon.

An efficient way of investigating the long-term behavior of a trajectory $x(t)$ defined for all $t \geqq 0$ is to study its *ω-limit set* $\omega(x)$; the set of points which are limits of sequences $x(t_k)$ where $t_k \to +\infty$. Thus, to say that $\omega(x)$ consists of a single point p means that $x(t)$ converges to p; such a p is necessarily an equilibrium. On the other hand, if $\omega(x)$ is a closed orbit of period T then $x(t)$ will eventually oscillate with period approaching T.

In addition to ω-limit sets there are α-limit sets, defined similarly by letting $t_k \to -\infty$. These are less important in applications but are useful for technical reasons.

The basic theme of this paper is that there are strong geometrical and topological restrictions on the way limit sets are placed in Euclidean n-space. Section 1 contains the basic definition and states the main theorems. Basic technical results about limit sets are proved in § 2. The remaining sections contain the proofs of the main theorems.

1. The main results. Consider a C^1 system of differential equations in $\mathbb{R}^n$,

$$\frac{dx_i}{dt} = F_i(x_1, \cdots, x_n) = F_i(x), \qquad i = 1, \cdots, n. \tag{1.1}$$

The system is called

competitive if $\partial F_i/\partial x_j \leqq 0$ for $j \neq i$,

cooperative if $\partial F_i/\partial x_j \geqq 0$ for $j \neq i$.

A well-known type of competitive system is the model of *competing species*,

$$\frac{dx_i}{dt} = F_i(x) = x_i M_i(x), \tag{1.2}$$

where

$$\frac{\partial M_i}{\partial x_j} < 0 \quad \text{for } j \neq i \tag{1.3}$$

and x is restricted to the nonnegative orthant

$$\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x_i \geqq 0, i = 1, \cdots, n\}.$$

It is known that, for $n = 2$, *every bounded solution defined on* $[0, \infty)$ *or on* $(-\infty, 0]$ *converges.* (Compare [5], [6], [11], [13]. A stronger result is proved in Theorem 2.3 below.) In contrast to this, Leonard and May [9] give examples of 3 competing species having oscillatory solutions.

Smale [14] has proved the general result that any dynamical system in $\mathbb{R}^{n-1}$ can be embedded in a system of n competing species. Let $\Delta^{n-1} \subset \mathbb{R}^n$ be the simplex spanned by the unit vectors e_i, $i = 1, \cdots, n$, where the kth component of e_i is δ_{ik}.

THEOREM (Smale). *Let X be any C^1 vector field in Δ^{n-1}. Then there exists a C^1 vector field $F = (F_1, \cdots, F_n)$ in $\mathbb{R}^n$ satisfying* (1.2) *and* (1.3), *such that $F|\Delta^{n-1} = X$ and Δ^{n-1} is an attractor.*

This result means that for $n > 2$ there is no hope of proving an analogous convergence theorem. It seems to imply that the limiting behavior of competitive systems can be arbitrarily complicated. For example one can start with a strange attractor in Δ^3 and extend it to a structurally stable competitive system in $\mathbb{R}^4$. On the other hand the results below show that there are in fact important restrictions on the limit set structure of competitive systems.

Briefly put, our main result is that *compact limit sets in a competitive or cooperative system are unknotted and unlinked.* In a kind of converse to Smale's theorem, Theorem A shows that a limit set of such a system can be deformed isotopically and equivariantly into an invariant set of some Lipschitz system in one dimension lower; moreover, the deformation is very simple geometrically. (Theorem A also implies that Smale's choice of the simplex Δ^{n-1} is not entirely arbitrary; for example the conclusion is not true for any simplex containing 0 and a positive vector.)

Theorem B says that a finite family of disjoint compact limit sets can be isotoped into disjoint convex sets.

Theorem C concerns 3-dimensional systems; it says that a compact limit set which contains no equilibrium is either a closed orbit or a ribbon of closed orbits. Thus generically it is a closed orbit.

We now explain the main results in more detail.

By a *limit set* we mean either an α-limit set or an ω-limit set (full definitions are given in § 2).

In the rest of this section we assume that (1.1) is cooperative or competitive and is defined in $\mathbb{R}^n$ or $\mathbb{R}^n_+$. More general domains are described in § 2.

Let $L \subset \mathbb{R}^n_+$ be a set. Let $E^{n-1} \subset \mathbb{R}^n$ be a hyperplane orthogonal to a vector v. Define $\pi: \mathbb{R}^n \to E^{n-1}$ to be an orthogonal projection. We say L is *compressible along* v if $\pi|L$ is a homeomorphism with Lipschitz inverse, and π maps L equivariantly respecting the flow of some locally Lipschitz vector field Y in E^{n-1}. (*Equivariant* means π takes trajectories of (1.1) in L to trajectories of Y, respecting parameterization.)

A vector v is *positive* if $v_i > 0$, $i = 1, \cdots, n$.

THEOREM A. *Let L be a limit set (of a system (1.1) as above). Then L is compressible along any positive vector.*

This has the corollary that every trajectory is nowhere dense. Moreover, the Lipschitzian nature of $(\pi|L)^{-1}$ implies that the dimension, and even the Hausdorff dimension, of L is $\leqq n-1$.

The proof of Theorem A is given in § 3.

A collection $L_1, \cdots, L_r$ of disjoint subsets of $\mathbb{R}^n$ is *unlinked* if there is a diffeotopy of $\mathbb{R}^n$ carrying them into disjoint convex sets. In § 6 we prove:

THEOREM B. *Every finite collection of disjoint compact limit sets is unlinked.*

In dimension 3 Theorems A and B imply that closed orbits are unknotted and unlinked. Theorem A allows us to bring into play the Poincaré–Bendixson theorem in studying 3-dimensional competitive or cooperative systems. In § 4 we prove:

THEOREM C. *Suppose $n = 3$. Let L be a compact limit set which contains no equilibrium. Then:*

(a) *L is either a closed orbit or a cylinder of closed orbits.*

(b) *L is a closed orbit if the system is cooperative and L is an ω-limit set.*

(c) *L is a closed orbit if all closed orbits are hyperbolic.*

Theorem C has interesting implications about the *observed* long-term behavior of a bounded solution $x(t)$ of a 3-dimensional competitive or cooperative system.

First consider the case when the ω-limit set L contains an equilibrium p. Then $x(t)$ gets arbitrarily near p; moreover, it stays within any given neighborhood of p for arbitrarily long periods of time. An observer would be hard put not to conclude that $x(t)$ has stabilized at p.

Consider next the case when L does not contain any equilibrium. Then according to Theorem C $x(t)$ will either converge to a limit cycle, or it will oscillate with slowly varying period, the rate of variation tending to zero.

Section 5 also contains a proof that if a cooperative system has a certain generic behavior then the only compact ω-limit sets are closed orbits.

The proofs of Theorems A, B and C are based on a famous comparison principle of Kamke which implies that *the flow of a cooperative system preserves the vector ordering* (Kamke [7]; see also Coppel [2]). Recall that this ordering is defined by

$$x < y \quad \text{if } x_i < y_i \quad \text{for all } i.$$

We also write

$$x \leqq y \quad \text{if } x_i \leqq y_i \quad \text{for all } i.$$

This result can also be used to study competitive systems since these correspond to cooperative ones through *time-reversal*: changing the independent variable from t to $-t$.

Define vectors x, y to be *related* if $x < y$ or $y > x$, and to be *unrelated* otherwise, i.e., when there exist i, j with $x_i \leqq y_i$ and $x_j \geqq y_j$.

The form of Kamke's result we shall use is:

THEOREM D. *Let $x(t)$, $y(t)$ be solutions defined for $a \leqq t \leqq b$.*

(a) *Suppose the system is cooperative. If $x(a) < y(a)$ then $x(b) < y(b)$.*

(b) *Suppose the system is competitive. If $x(a)$ and $y(a)$ are unrelated then so are $x(b)$ and $y(b)$.*

This result is valid for systems defined in $\mathbb{R}^n$ or $\mathbb{R}^n_+$, and also for systems defined in sets Γ described in § 2.

2. Limit sets. In this section we consider a cooperative or competitive system

$$\frac{dx_i}{dt} = F_i(x_1, \cdots, x_n), \qquad i = 1, \cdots, n \tag{2.1}$$

defined by a C^1 vector field $F: \Gamma \to \mathbb{R}^n$. The precise assumptions on the domain $\Gamma \subset \mathbb{R}^n$ are given below following the statement of the main results of this section. They are satisfied if $\Gamma = \mathbb{R}^n_+$ or $\mathbb{R}^n$.

The first result is a useful criterion for a solution to converge. It can also be viewed as an existence criterion for certain kinds of equilibrium points.

A point $p \in \Gamma$ is an *equilibrium* if $F(p) = 0$.

THEOREM 2.1. *Assume* (2.1) *is cooperative. Let $x: [0, \infty) \to \Gamma$ be a solution whose image has compact closure in Γ. If $x(T)$ is related to $x(0)$ for some $T > 0$, then $x(t)$ converges to an equilibrium as $t \to \infty$.*

This implies that no two points of a closed orbit of a cooperative system can be related, and the same holds for competitive systems by time reversal. It follows easily that in dimension 3 a closed orbit cannot be knotted.

The following result expresses important limitations on the geometry of limit sets.

THEOREM 2.2. *Suppose* (2.1) *is cooperative or competitive. Then no two points of a limit set can be related. Moreover, if y is a limit point then the vector $F(y)$ is unrelated to the zero vector.*

The following result shows that 2-dimensional cooperative or competitive systems have trivial dynamics.

THEOREM 2.3. *Assume* (2.1) *is cooperative or competitive and $n = 2$. Let $y: [0, \tau) \to \Gamma$ be the solution through $y(0)$, $\tau = t_+(y)$. Then either $|y(t)| \to \infty$ as $t \to \tau$, or else $y(t)$ converges to some point of $\bar{\Gamma}$ as $t \to \tau$. In fact $[0, \tau)$ is the union of two intervals, in each of which both $y_1(t)$ and $y_2(t)$ are monotone.*

Before proving these results we describe the assumption on the domain Γ. We wish to cover the cases where $\Gamma = \mathbb{R}^n$ or $\mathbb{R}^n_+$. But there are interesting systems which are cooperative in some regions and competitive in others. Consider for example the following system in $\mathbb{R}^n_+$:

$$\frac{dx_i}{dt} = a_i(x_i)[b_i(x_i) + g(s)], \tag{2.2}$$

where $a_i \geqq 0$ and $s = x_1 + \cdots + x_n$. Evidently (2.2) is competitive where $g'(s) \leqq 0$ and cooperative where $g'(s) \geqq 0$. (Such systems are suggested by Grossberg's models of adaptive networks [5]. They also arise in models of economic competition.)

From now on Γ is a locally closed subset of $\mathbb{R}^n$ whose interior $\operatorname{Int}\Gamma$ is dense in Γ. This means that Γ is the intersection of an open set with the closure of an open set. The vector field $F: \Gamma \to \mathbb{R}^n$ is assumed to extend to a C^1 map on an open set in $\mathbb{R}^n$.

The final assumption is that Γ is *p-convex*: if $a, b \in \Gamma$ and $a \geqq b$ then Γ contains the line segment between a and b. Kamke's comparison principle (Theorem D above) is then valid. (But Theorem 2.3 is valid without p-convexity.)

The following statements on domains of solutions and limit sets are easily proved using standard theorems in differential equations and the assumptions about Γ.

Let $W \subset \mathbb{R}^n$ be an open set containing Γ and $G: W \to \mathbb{R}^n$ a C^1 vector field extending F. For any $y \in \Gamma$ there is a unique nonextendible solution $\xi(t)$, $\alpha < t < \beta$ to the initial value problem

$$\frac{d\xi}{dt} = G(\xi), \qquad \xi(0) = y.$$

Let $I(y) \subset \mathbb{R}$ be the connected component of 0 in the set

$$\{t: \alpha < t < \beta \text{ and } \xi(t) \in \Gamma\}.$$

The restriction of ξ to $I(y)$ is the solution (in Γ) to (2.1) passing through y at time $t = 0$. It is denoted by $y(t)$ or $\phi_t(y)$. Since the interior of Γ is dense this solution is independent of the choice of W and G.

Set

$$t_+(y) = \sup\{t: t \in I(y)\}, \qquad t_-(y) = \inf\{t: t \in I(y)\},$$

so that $-\infty \leqq t_-(y) \leqq 0 \leqq t_+(y) \leqq \infty$. We say the solution through y *terminates* if $t_+(y) \in I(y)$; otherwise it is *nonterminating*. In the nonterminating case if $t_+(y) < \infty$, then either $|y(t)| \to \infty$ or $y(t)$ approaches the boundary $\partial\Gamma$ of Γ as $t \to t_+(y)$. If $y(t)$ is nonterminating and the *forward orbit*

$$O_+(y) = \{y(t): 0 \leqq t < t_+(y)\}$$

has compact closure in Γ, then $t_+(y) = \infty$.

Suppose $y(t)$ is nonterminating. Its *ω-limit set* $\omega(y) = \omega(y(0))$ is defined to be the set of points $p \in \Gamma$ such that

$$p = \lim_{k \to \infty} y(t_k)$$

for some sequence t_k in $I(y)$ converging to $t_+(y)$. It is easy to prove that if p is an ω-limit point of y (i.e., $p \in \omega(y)$) then the solution through p is nonterminating. It is also noninitiating, i.e., $t_-(p) \notin I(p)$. Moreover, $\omega(y)$ is *invariant*: if $p \in \omega(y)$ then $\phi_t(p) \in \omega(y)$ for all $t \in I(p)$. It is well known that $O_+(y)$ has compact closure in Γ if and only if $\omega(y)$ is a nonempty compact connected set. In this case $t_+(y) = \infty$.

If $y(t)$ terminates then $\omega(y)$ is defined as the empty set.

The α-limit set $\alpha(y)$ is defined similarly, replacing $t_+(y)$ with $t_-(y)$; it has analogous properties.

A *closed orbit* γ of *period* $T \neq 0$ is the image of a solution $u: \mathbb{R} \to \Gamma$ such that $u(t+T) = u(t)$ for all t. Notice the set of periods of γ, together with 0, is a closed subgroup of $\mathbb{R}$.

PROPOSITION 2.4. *Assume* (2.1) *is cooperative and* $x: [0, \infty) \to \Gamma$ *is a solution. Let* $T > 0$ *be such that* $x(T) \geqq x(0)$ *or* $x(T) \leqq x(0)$. *Let* $p \in \Gamma$ *be a limit point of* $\{x(kT): k \in \mathbb{Z}_+\}$ *with* $T \in I(p)$. *Then* p *lies on a closed orbit* γ *of period* T *and* $\omega(x) = \gamma$.

Proof. We suppose $x(T) \geqq x(0)$, the other case being similar. Then

$$x(t+T) \geqq x(t) \quad \text{for all } t > 0.$$

In particular,

$$x((k+1)T) \geqq x(kT) \quad \text{for all } k \in \mathbb{Z}_+.$$

It follows easily that

$$p = \lim_{k \to \infty} x(kT) = \lim_{k \to \infty} x((k+1)T), \qquad k \in \mathbb{Z}_+. \tag{2.3}$$

Therefore $p(T) = p$, so p lies on a closed orbit γ of period T. From (2.3) and the continuity of solutions it follows that $\gamma = \omega(x)$. Q.E.D.

Proof of Theorem 2.1. Let $x: [0, \infty) \to \Gamma$ be as in Theorem 2.1. There is an open set $S \subset \mathbb{R}$ containing T such that $x(s) > x(0)$ [or $< x(0)$] for all $s \in S$. By Proposition 2.4, $\omega(x)$ is a closed orbit γ and γ has period s for all $s \in S$. It follows that γ consists of a single equilibrium p. Q.E.D.

The proof of Theorem 2.2 requires the following result, Proposition 2.5, which has some independent interest: it shows that solutions to (2.1) cannot oscillate with respect to the partial ordering $<$.

Let $y(t)$ be a curve in $\mathbb{R}^n$ defined on some interval $I \subset \mathbb{R}$. A subinterval $J = [a, b] \subset I$ is called an *up-interval* if $y(a) < y(b)$, and a *down-interval* if $y(a) > y(b)$.

The following result is due to L. Ito.

PROPOSITION 2.5. *Assume* (2.1) *is cooperative or competitive. Then a solution* $y: I \to \Gamma$ *cannot have an up-interval and a down-interval which are disjoint.*

Proof. We prove the cooperative case; the competitive case follows by time-reversal.

Suppose there are an up-interval K and a down-interval J with $J \cup K \subset I$, $J \cap K = \varnothing$. We assume $J < K$ (i.e., $u < v$ for all $u \in J$, $v \in K$), the other case being similar. Put

$$J = [a, r], \qquad K = [s, b], \qquad a < r < s < b.$$

Let $s' \in K$ be the smallest number such that $y(s') \leqq y(t)$ for all $t \in K$; we denote this by $y(s') = \inf y(K)$. Then $[s', b]$ is an up-interval disjoint from J, so we may replace K by $[s', b]$. Thus we may assume that $y(s) = \inf y(K)$.

We prove the proposition by showing that neither $r - a \leqq b - s$ nor $r - a > b - s$ can hold.

Assume $r - a \leqq b - s$. Then $s < s + r - a \leqq b$. Since $[s, s + r - a]$ is a translate to the right of $[a, r]$ it follows that $[s, s + r - a]$ is a down-interval. But then

$$y(s + r - a) < y(s) \quad \text{and} \quad s + r - a \in K,$$

contradicting $y(s) = \inf y(K)$.

Assume $r-a>b-s$. Then $a<a+b-r<s<b$. Since $[a+b-r, b]$ is a translate of $[a, r]$ to the right it follows that $[a+b-r, b]$ is a down-interval. Thus

$$y(a+b-r)>y(b)>y(s).$$

Define $c\in[a+b-r, s]$ to be the largest number such that $y(c)\geqq y(b)$, so that $c<s<b$. Then $y(c)>y(s)$, i.e., $[c, s]$ is a down-interval.

Suppose $s-c\leqq b-s$. Then translating $[c, s]$ by $s-c$ provides a down-interval $[s, 2s-c]\subset[s, b]$, contradicting $y(s)=\inf y(K)$.

Finally, suppose $s-c>b-s$. Then $c<c+b-s$; since $[c+b-s, b]$ is a right-translate of the down-interval $[c, s]$ we have

$$y(c+b-s)>y(b).$$

But this contradicts the definition of c. The proof of Proposition 2.5 is complete. Q.E.D.

In the cooperative case the following result also holds: *if $z(t)$ is a solution defined for all $t\geqq 0$ then it cannot have both an up-interval and a down-interval.* For one of the intervals could be translated to the right until it is disjoint from the other interval, in contradiction to Proposition 2.5. A similar (but less interesting) conclusion holds for a solution to a competitive system defined for all $t\leqq 0$.

Proof of Theorem 2.2. Suppose $p<q$. From the definition of limit set there must exist $t_1<t_2<t_3<t_4$ with $z(t_1)$, $z(t_4)$ so close to p, and $z(t_2)$, $z(t_3)$ so close to q, that $z(t_1)<z(t_2)$ and $z(t_4)<z(t_3)$. Therefore $[t_1, t_2]$ is an up-interval which is disjoint from the down-interval $[t_3, t_4)$, contradicting Proposition 2.5.

Suppose $F(p)>0$ and $p\in\omega(Z)$. Then $p(\varepsilon)>p$ for sufficiently small $\varepsilon>0$. Since $p(\varepsilon)\in\omega(z)$ this contradicts what has already been proved. The other cases are similar. Q.E.D.

Proof of Theorem 2.3. By time-reversal we may assume (2.1) is cooperative. We also assume $y(t)$ is not constant. The interval $I=\{t\geqq 0: y(t) \text{ is defined}\}$ is the union of the following sets:

$$A_1=\{t\in I: F_i(y(t))\geqq 0, i=1, 2\},$$

$$A_2=\{t\in I: F_2(y(t))>0>F_1(y(t))\},$$

$$A_3=\{t\in I: F_i(y(t))\leqq 0, i=1, 2\},$$

$$A_4=\{t\in I: F_2(y(t))<0<F_1(y(t))\}.$$

Notice that these sets are pairwise disjoint. And if $t\in A_1$ and $s>t$ and $s\in I$ then $s\in A_1$. For

$$D\phi_{s-t}(y(t))F(y(t))=F(y(s))$$

and, from Proposition 2.6 below, $D\phi_r(x)$ is a nonnegative matrix for all $x\in\Gamma$, $r\in I(x)$, $r>0$. Similarly if, $t\in A_3$ and $s>t$, $s\in I$ then $s\in A_3$. This proves that *either A_1 or A_3* is *empty.*

Let $k\in\{1, 2, 3, 4\}$ be such that $0\in A_k$. If $k=1$ or 3 then $I\subset A_k$, so $y_i(t)$ is monotone, $i=1, 2$. If $k=2$ or 4 and $I\not\subset A_k$ there must be a smallest $t_0\in I$ with $t_0\in A_j$, $j=1$ or 3. Then $I\subset A_k\cup A_j$. This proves the last sentence of Theorem 2.3. Q.E.D.

Theorem 2.3 is true even without p-convexity of Γ: the proof uses only the following result.

PROPOSITION 2.6. *Let $\{\phi_t\}$ denote the flow of a cooperative system defined in a set $\Gamma\subset\mathbb{R}^n$ which satisfies the assumptions above except that Γ is not assumed to be p-convex. Then $D\phi_t(x)$ is a nonnegative matrix for all $t\geqq 0$, $x\in\Gamma$.*

Proof. We apply Kamke's theorem (which holds for nonautonomous systems) to the variational equation along a fixed solution $x(t) = \phi_t(x)$:

$$\frac{dA}{dt} = DF(\phi_t(x))A. \tag{2.4}$$

Here $A(t)$ is an $n \times n$ matrix. It is easily seen that the right-hand side of (2.4), that is, the matrix function

$$G(t, A) = (DF(\phi_t(x))A,$$

satisfies

$$\frac{\partial G_{ij}}{\partial A_{rs}} \geqq 0 \quad \text{for } (i, j) \neq (r, s),$$

so that Kamke's comparison theorem applies to solutions to (2.4). Now the solution $B(t)$ to (2.4) with initial condition $B(0) = 0$ is the constant solution $B(t) = 0$, while the solution with initial condition $A(0) = I$ is $D\phi_t(x)$. Since $0 \leqq I$ we have $B(t) \leqq A(t)$, for all $t \geqq 0$, i.e., $D\phi_t(s)$ is a nonnegative matrix. Q.E.D.

Proposition 2.6 has an interesting consequence for competitive systems, due to S. Grossberg [5]. We say a solution $x(t)$ is *switched on* at time t_0 if $F_i(x(t_0)) > 0$ for some $i \in \{1, \ldots, n\}$. Grossberg calls the following result the *ignition principle.*

PROPOSITION 2.7. *Let Γ be as in Proposition 2.6 and let $x(t)$ be a solution to a competitive system in Γ. If $x(t)$ is switched on at t_0 then $x(t)$ is switched on all $t_1 > t_0$.*

Proof. Let the competitive system be

$$\frac{dx}{dt} = G(x).$$

Consider the corresponding cooperative system obtained by time-reversal,

$$\frac{dy}{dt} = F(y) \equiv -G(y).$$

Fix $t_1 > t_0$, and observe that the curve $y(t) = x(t_1 - t)$ is a solution to the cooperative system.

If $x(t)$ is not switched on at time t_1 then $G(x(t_1)) \leqq 0$, so $F(y(0)) \geqq 0$. Therefore $F(y(s)) \geqq 0$ for all $s > 0$ by Proposition 2.6, since

$$F(y(s)) = D\phi_s(y(0))F(y(0)),$$

and $D\phi_s(y(0)) \geqq 0$. Putting $s = t_1 - t_0$ shows that $F(y(t_1 - t_0)) \geqq 0$, so $G(x(t_0)) \leqq 0$, contradicting $x(t)$ being switched on at t_0. Q.E.D.

3. Extension and proof of Theorem A. In this section we consider a cooperative or competitive system

$$\frac{dx}{dt} = F(x). \tag{3.1}$$

THEOREM 3.1. *Let L be a limit set. Then L is compressible along any positive vector v.*

Proof. It suffices to consider only unit vectors v. Let $\pi_v : \mathbb{R}^n \to E^{n-1} = v^\perp$ be an orthogonal projection onto the hyperplane orthogonal to v. We first show that $\pi_v | L$ is injective.

Suppose $p, q \in L$ and $\pi_v(p) = \pi_v(q)$. Then $p - q = \lambda v$, $\lambda \in \mathbb{R}$. If $p \neq q$ then p, q are related since $\lambda \neq 0$ and $v > 0$. But this contradicts Theorem 2.2; therefore $\pi_v|L$ is injective.

For distinct nonzero vectors $a, b \in \mathbb{R}^n$ let $A(a, b)$ denote the positive acute angle between the lines $\{\lambda a: \lambda \in \mathbb{R}\}$ and $\{\lambda b: \lambda \in \mathbb{R}\}$. It is easy to see that if $\pi_v|L$ is *not* Lipschitz then there exist sequences p_j, q_j in L such that $p_j \neq q_j$ and

$$A(v, p_j - q_j) \to 0.$$

Suppose this holds, and let w_j denote the unit vector $(p_j - q_j)/|p_j - q_j|$. Then $A(v, w_j) \to 0$. Passing to a subsequence we may assume that $w_j \to \pm v$. Interchanging p_j and q_j where necessary we may assume $w_j \to v$. Choose k so large that $w_k > 0$ and set $w = w_k$. Then $\pi_w(p_k) = \pi_w(q_k)$, contradicting the injectivity of π_w proved above. This shows that $\pi_v|L$ is Lipschitz.

Set $\pi_v = \pi$ and define

$$H: \pi(L) \to E^{n-1}, \qquad H = \pi \circ F \circ (\pi|L)^{-1}.$$

Now F is C^1 so $F: L \to \mathbb{R}^n$ is locally Lipschitz. Therefore H is locally Lipschitz. By a result of McShane [14] H can be extended to a locally Lipschitz vector field on E^{n-1}. Notice that $H(\pi x) = 0$ if and only if $F(x) = 0$, by the last statement of Theorem 2.2.

To say that $\pi: L \to E^{n-1}$ is equivariant means that if $x(t)$ is an integral curve of F in L then $\pi(x(t))$ is an integral curve of H. This follows from the definition of H. Q.E.D.

We conclude this section with another compressibility theorem for cooperative systems.

THEOREM 3.2. *Suppose* (3.1) *is cooperative and* $K \subset \Gamma$ *is a compact set which is the closure of the image of a solution* $x: [0, \infty) \to K$. *If* $x(t)$ *does not converge to an equilibrium then* K *is compressible along any positive vector.*

Proof. Observe that $K = O_+(x) \cup \omega(x)$. Let $\pi_v: \mathbb{R}^n \to v^\perp$ be an orthogonal projection where $v > 0$. Using arguments similar to those above, one can show that if $\pi_v|K$ is not injective, or $(\pi_v|K)^{-1}$ is not Lipschitz, then $x(t_0)$ and $x(t_1)$ are related for some $t_0, t_1 \geqq 0$. But then by Theorem 2.1 $x(t)$ converges to an equilibrium. The rest of the proof is like that of Theorem 3.1. Q.E.D.

4. Proof of Theorem C. In this section we assume given a competitive or cooperative system

$$\frac{dx_i}{dt} = F_i(x_1, x_2, x_3), \qquad i = 1, 2, 3 \tag{4.1}$$

defined in a set $\Gamma \subset \mathbb{R}^3$ satisfying the conditions in § 2.

THEOREM 4.1. *Let* L *be a compact limit set which contains no equilibrium. Then*:

(a) L *is either a closed orbit or a cylinder of closed orbits.*

(b) L *is a closed orbit if the system is cooperative and* L *is an* ω*-limit set.*

(c) L *is a closed orbit if* L *contains a hyperbolic closed orbit.*

Proof. Let $\pi: \mathbb{R}^3 \to E^2$ be an orthogonal projection onto a plane perpendicular to a positive vector. By Theorem 3.1, π maps L homeomorphically and equivariantly onto an invariant set of some locally Lipschitz vector field Y in E^2. Clearly $\pi(L)$ is compact and connected, and it contains no equilibrium.

Let ψ_t denote the flow of Y.

The Poincaré–Bendixson theorem (see, e.g., Hartman [4]) implies that $\pi(L)$ is a union of closed orbits and trajectories that spiral down to closed orbits in both positive

and negative time. We shall prove that such spiralling cannot in fact occur, so that $\pi(L)$ is a union of closed orbits. From this part (a) of the theorem follows easily.

Let $z \in \pi(L)$. Suppose z is not on a closed orbit. Then, as $t \to \infty$, $\psi_t(z)$ spirals down to a closed orbit $\gamma \subset \pi(L)$.

Let A denote the component of $E^2 - \gamma$ which contains z; let B denote the other component. Let $T > 0$ be the period of γ.

It is well known that γ is an attractor for the flow restricted to $\bar{A}$. Thus there is a compact neighborhood N or γ in $\bar{A}$ such that

$$\psi_T(N - \gamma) \subset \operatorname{Int} N.$$

Define $W = \pi(L) \cap (B \cup N)$. Then W is a compact subset of $\pi(L)$ and

$$\psi_T(W) \subset \operatorname{Int} W.$$

Put $V = (\pi|L)^{-1}(W)$. Then V is a compact subset of L and

$$\phi_T(V) \subset \operatorname{Int}_L V.$$

This, however, is impossible for a compact limit set L by a result of Franke and Selgrade [3]. This contradiction shows that z, which is an arbitrary point of $\pi(L)$, must lie on a closed orbit. This completes the proof of (a).

Now assume the system is cooperative.

Suppose that L is not a single closed orbit, but is a cylinder of closed orbits. Then $\pi(L)$ contains a 2-disk D. Let $p \in L$ be such that $\pi(p)$ is the center of D.

Let $x(t)$, $t > 0$ be a solution of (1) whose ω-limit set is L. There exists $t_0 > 0$ such that $x(t_0) \in \operatorname{Int} D$. Let $q \in L$ be such that $\pi(x(t_0)) = \pi(q)$. It follows that $x(t_0)$ is related to q.

There exists $t_1 > t_0$ such that $x(t_1)$ is so near q that $x(t_0)$ is related to $x(t_1)$. It now follows from Proposition 2.1 that $x(t)$ converges to an equilibrium as $t \to \infty$. Thus L is an equilibrium; this contradiction completes the proof of (b).

Part (c) follows from (a) since a cylinder of closed orbits cannot contain a hyperbolic closed orbit. Q.E.D.

By exploiting Theorem 3.2 we can establish other criteria for L to be a closed orbit in the cooperative case. Suppose L is a compact ω-limit set of a cooperative system (4.1), say $L = \omega(x)$. Suppose L contains a nonequilibrium closed orbit γ. Then $x(t)$ does not converge, and so Theorem 3.2 implies that the closure of $\{x(t): t \geqq 0\}$ is compressible. Set $z = \pi x(0)$. Thus (in the notation above) $\psi_t(z)$ has the ω-limit set $\pi(L)$. Since $\pi(L)$ contains the closed orbit $\pi(\gamma)$, the Poincaré–Bendixson theorem implies $\pi(L) = \pi(\gamma)$. Since π is injective it follows that $L = \gamma$.

Now suppose L does not contain a closed orbit, L contains only a finite number of equilibria, and $x(t)$ does not converge. Then by using Theorem 3.2 and Poincaré–Bendixson one can show that L must contain a *cycle of equilibria* $p_1, \cdots, p_k$ $(k \geqq 1)$: this means that for each $i = 1, \cdots, k$ there is a solution $y_i(t)$ in L whose α-limit in p_i and whose ω-limit is p_{i+1}, with $p_{k+1} = p_1$. Thus we obtain:

THEOREM 4.2. *Assume that* (4.1) *is cooperative and contains no cycle of equilibria. Then every compact ω-limit set is a closed orbit* (*possibly an equilibrium*).

It is well known that existence of a cycle of equilibria is a highly unstable phenomenon. It cannot occur if all the equilibria are hyperbolic and their stable and unstable manifolds meet only transversely—a generic property of C^1 vector fields (see Smale [16], Abraham and Robbin [17]).

In applying approximation theorems to cooperative systems there arises the difficulty that the cooperative condition is not stable. However, the property of being

strongly cooperative—$\partial F_i/\partial x_j > 0$ for $i \neq j$—*is* stable. Given a cooperative field F every neighborhood of F in the compact-open C^1 topology contains a strongly cooperative field G of the form

$$G_i(x) = F_i(x) + \delta \sum_{j=1}^{n} x_j, \qquad \delta > 0.$$

The field G can then be approximated by fields having generic properties. In this way, using standard approximation methods of differentiable dynamical systems, one can prove the following result:

THEOREM 4.3. *Let F be a cooperative vector field in $\Gamma \subset \mathbb{R}^n$. Let $K \subset \Gamma$ be a compact set and ε a positive number. There exists a strongly cooperative vector field G on Γ with the following properties*:

(a) *For all $x \in K$,*

$$|F(x) - G(x)| + \|DF(x) - DG(x)\| < \varepsilon.$$

(b) *All equilibria and closed orbits of G are hyperbolic, and their stable and unstable manifolds meet only transversely.*

(c) *If $n = 3$ then every compact ω-limit set is a closed orbit (perhaps degenerate).*

5. A criterion for unlinking. In this section we prove a topological result, Proposition 5.2, needed for the proof of Theorem B.

An *isotopy* of $\mathbb{R}^n$ is a family of C^1 diffeomorphisms $h_t: \mathbb{R}^n \to \mathbb{R}$, $0 \leqq t \leqq 1$, such that h_0 is the identity and $h_t(x)$ is C^∞ in (t, x).

Let $\mathcal{A} = \{A_i\}$, $\mathcal{B} = \{B_i\}$ be two collections of subsets of $\mathbb{R}^n$ indexed by the same set S. We say $\mathcal{A}$ and $\mathcal{B}$ are *isotopic* if there is an isotopy h_t of $\mathbb{R}^n$ such that $h_1(A_i) = B_i$ for all $i \in S$. This is an equivalence relation.

The family $\mathcal{A}$ is *unlinked* if it is isotopic to a family $\mathcal{B}$ such that there exist disjoint convex sets $C_i \subset \mathbb{R}^n$ with $B_i \subset C_i$ for all $i \in S$.

Let $f_1, \cdots, f_r$ be continuous real-valued functions on $\mathbb{R}^{n-1}$. Let L_i be the graph of f_i; we consider L as a subset of $\mathbb{R}^{n-1} \times \mathbb{R} = \mathbb{R}^n$.

LEMMA 5.1. *If $f_1 < \cdots < f_r$ then $\{L_1, \cdots, L_r\}$ is unlinked.*

Proof. Given real numbers $u_1 < \cdots < u_r$ there is a single isotopy of $\mathbb{R}$ carrying each u_i into the open interval $(i, i+1)$. Moreover, the isotopy can be chosen to be C^∞ in the parameters $(u_1, \cdots, u_r)$. We assume such a family of isotopies has been chosen once and for all; for fixed $u_1 < \cdots < u_r$ we denote the isotopy by

$$y \to g(t, u_1, \cdots, u_r, y), \qquad y \in \mathbb{R},$$

where g is C^∞.

An isotopy of $\mathbb{R}^{n-1} \times \mathbb{R}$ is defined by

$$h_t: \mathbb{R}^{n-1} \times \mathbb{R} \to \mathbb{R}^{n-1} \times \mathbb{R},$$

$$h_t(x, y) = (x, g(t, f_1(x), \cdots, f_r(x), y)).$$

Evidently h_1 takes the graph of f_i into the convex set $C_i = \{(x, z) \in \mathbb{R}^{n-1} \times \mathbb{R}: i < z < i+1\}$. Since these sets are disjoint the L_i are unlinked. Q.E.D.

PROPOSITION 5.2. *For $i = 1, \cdots, r$ let $K_i \subset \mathbb{R}^{n-1}$ be a compact set and $g_i: K_i \to \mathbb{R}$ a continuous map. Let $L_i \subset \mathbb{R}^{n-1} \times \mathbb{R}$ denote the graph of g_i. Suppose that $g_i(x) < g_j(x)$ whenever $i < j$ and $x \in K_i \cap K_j$. Then $\{L_1, \cdots, L_r\}$ is unlinked.*

Proof. This follows from Lemma 5.1 provided the g_i extend to continuous maps $f_i: \mathbb{R}^{n-1} \to \mathbb{R}$ such that $f_i < f_j$ for $i < j$. Such extensions can be found as follows.

By Tietze's theorem the g_i extend to continuous maps $\hat{g}_i: \mathbb{R}^{n-1} \to \mathbb{R}$. Let

$$m_i < \min\{f_j(x): 2 \leqq j \leqq r, x \in K_j\}.$$

Let U be a neighborhood of K so small that $\hat{g}_i < m_i$ on U. Let $\rho: \mathbb{R}^{n-1} \to [0, 1]$ be a continuous function which is 1 on K_1 and 0 on $\mathbb{R}^{n-1} = U$.

Define

$$f_1: \mathbb{R}^{n-1} \to \mathbb{R},$$

$$f_1(x) = \rho(x)\hat{g}_1(x) + (1-\rho(x))m_1.$$

Then $f_1(x) < g_j(x)$ for $j > 1$, $x \in K_j$. A similar procedure extends g_2 to $f_2: \mathbb{R}^{n-1} \to \mathbb{R}$ with $f_2 > f_1$ and $f_2(x) < g_j(x)$ for $j > 2$, $x \in K_j$, etc. In this way the required f_j are successively defined. Q.E.D.

6. Extension and proof of Theorem B. Theorem B is a special case of a more general result, Theorem 6.1, proved below.

We suppose given a cooperative or competitive system

$$\frac{dx}{dt} = F(x) \tag{6.1}$$

defined in a set $\Gamma \subset \mathbb{R}^n_+$ as in § 2.

Let L be an invariant set. We call L a *pseudo-limit* set if it satisfies the following condition. Given two points of L and $\varepsilon > 0$, there is a trajectory (not necessarily in L) that comes within ε of each of the points. Evidently limit sets and orbit closures are examples of pseudo-limit sets.

A set L is *balanced* if p, q are unrelated for all p, q in L.

THEOREM 6.1. *Let $L_1, \cdots, L_s$ be disjoint compact pseudo-limit sets. Suppose that each L_i is balanced. Then $\{L_1, \cdots, L_s\}$ is unlinked.*

The proof depends on the following lemma. Define a relation $\prec$ on $\{L_1, \cdots, L_s\}$: $L_i \prec L_j$ if $p < q$ for some $p \in L_i$, $q \in L_j$.

LEMMA 6.2. *The relation $\prec$ is a partial ordering.*

Proof. Since each L_i is balanced it is impossible that $L_i \prec L_i$.

Suppose $L_i \prec L_j \prec L_m$. We want to prove $L_i \prec L_m$. There exist points

$$p \in L_i, \quad q, q' \in L_j, \quad r \in L_m$$

such that $p < q$, $q' < r$. Let U, U' be neighborhoods in Γ of q, q' respectively such that $p < U$, $U' < r$. Let $y(t)$ be a solution entering both U and U'. Suppose $y(t_0) \in U$, $y(t_1) \in U'$.

By time-reversal we assume $t_1 \geqq t_0$.

Suppose the system is cooperative. Let $x(t)$ be the solution such that $x(t_0) = p$. We have $x(t_0) < y(t_0)$. Since $t_0 \leqq t_1$, the order-preserving property (Theorem D of § 1) implies $x(t_1) < y(t_1)$. Now $y(t_1) \in U'$ so $y(t_1) < r$. Since $x(t_1) \in L_i$ it follows that $L_i \prec L_m$.

Suppose the system is competitive. Let $z(t)$ be the solution such that $z(t_1) = r$. By a similar argument one sees that

$$p = x(t_0) < y(t_0) < z(t_0)$$

and $z(t_0) \in L_m$. Thus in all cases $L_i \prec L_m$. This completes the proof of Lemma 6.2. Q.E.D.

Proof of Theorem 6.1. It follows from Lemma 6.1 that the L_k are partially ordered by $\prec$. We relabel the L_k so that if $L_i \prec L_j$ then $i < j$.

Let $E^{n-1} \subset \mathbb{R}^n$ be a hyperplane orthogonal to a positive unit vector v. Let $\pi: \mathbb{R}^n \to E^{n-1}$ be the orthogonal projection. Define continuous maps

$$g_k: \pi(L_k) \to \mathbb{R}, \qquad x \to \langle x, v \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product.

Identify $\mathbb{R}^n$ isometrically with $E^{n-1} \times \mathbb{R}$ in such a way that E^{n-1} is identified with $E^{n-1} \times 0$ in the natural way, and λv is identified with $(0, \lambda v)$ for all $\lambda \in \mathbb{R}$. Then L_k is identified with the graph of g_k.

The partial ordering of the L_k by $<$ implies that whenever $i < j$ then $g_j < g_i$ on $\pi(L_i) \cap \pi(L_j)$. It follows from Proposition 6.2 (with $K_i = \pi(L_i)$) that $\{L_1, \cdots, L_s\}$ is unlinked. Q.E.D.

REFERENCES

[1] J. G. CONLON, *A theorem in ordinary differential equations with an application to hyperbolic conservation laws*, Adv. Math., 35 (1980), pp. 1–18.

[2] W. A. COPPEL, *Stability and Asymptotic Behavior of Differential Equations*, D. C. Heath, Boston, 1965.

[3] J. E. FRANKE AND J. F. SELGRADE, *Abstract ω-limit sets, chain recurrent sets, and basic sets for flows*, Proc. Amer. Math. Soc., 60 (1976), pp. 309–316.

[4] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[5] S. GROSSBERG, *Competition, decision and consensus*, J. Math. Anal. Appl., 66 (1978), pp. 470–493.

[6] M. HIRSCH AND S. SMALE, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[7] E. KAMKE, *Zur Theorie der Systeme gewöhnlicher differential-gleichungen*, II, Acta Math., 58 (1932), pp. 57–85.

[8] A. LAJMANOVICH AND J. YORKE, *A deterministic model for gonorrhea in a nonhomogeneous population*, Math. Biosci., 28 (1976), pp. 221–236.

[9] W. J. LEONARD AND R. MAY, *Nonlinear aspects of competition between species*, SIAM J. Appl. Math., 29 (1975), pp. 243–275.

[10] E. J. MCSHANE, *Extension of range of functions*, Bull. Amer. Math. Soc., 40 (1934), pp. 837–842.

[11] A. RESCIGNO AND I. RICHARDSON, *The struggle for life*; I, *Two species*, Bull. Math. Biophys., 29 (1967), pp. 377–388.

[12] J. SELGRADE, *Mathematical analysis of a cellular control process with positive feedback*, SIAM J. Appl. Math., 36 (1979), pp. 219–229.

[13] ———, *Asymptotic behavior of solutions to single loop positive feedback systems*, J. Differential Equations, 38 (1980), pp. 80–103.

[14] S. SMALE, *On the differential equations of species in competition*, J. Math. Biol., 3 (1976), pp. 5–7.

[15] ———, *The Mathematics of Time*, Springer-Verlag, New York, 1980.

[16] ———, *Stable manifolds for differential equations and diffeomorphisms*, Annali della Scuola Normale Superiore di Pisa (Serie III), 17 (1963), pp. 97–116.

[17] R. ABRAHAM AND J. ROBBIN, *Transversal Mappings and Flows*, W. A. Benjamin, New York, 1967.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0002 $01.00/0

A CLOSURE PROBLEM FOR SIGNALS IN SEMIGROUP INVARIANT SYSTEMS*

JOHN J. BENEDETTO†

Abstract. Let V_f be the L^1 variety generated by $f \in L^1$. The closure statement, $V_f = L^1$, is characterized for a class of functions (damped signals) $f \in L^1$ in terms of a canonical semigroup Γ and a constructible measure ν; the characterization does not use Wiener's criterion $\{x: \hat{f}(x) = 0\} = \varnothing$, since this condition is difficult to verify in practice. Damped signals arise in certain variational and number theoretic problems. Applications include the verification of the closure statement in several cases as well as the construction of new nontrivial invariant subspaces of weighted L^1-spaces. The main technique involves the proper approximation of ν, which is supported in $\Gamma \subseteq [0, \infty)$; the main point is that for certain damped signals this approximation requires information from the distant past, that is, from a neighborhood of $-\infty$. The terminology, "semigroup invariant systems," in the title, is used to formulate a notion generalizing the "shift invariant systems" of digital signal processing.

Introduction. A bounded function S on the real line $\mathbb{R}$ is a *signal* if

$$S = \sum_{j=0}^{\infty} S(\lambda_j)\chi_{[\lambda_j, \lambda_{j+1})},$$

where $0 = \lambda_0 < \lambda_1 < \lambda_2 < \cdots$, $\lim \lambda_j = \infty$, $S(0) \neq 0$ and χ_A is the characteristic function of the set A. $\Lambda = \{\lambda_j: j = 0, 1, \cdots\}$ is the set of *jumps* of the signal S. $L^1(\mathbb{R})$ is the space of Lebesgue integrable functions on $\mathbb{R}$, $S_\alpha(t) = e^{-\alpha t}S(t)$ is the *damped signal* on $\mathbb{R}$ corresponding to the signal S and $\alpha > 0$, and V_α is the $L^1(\mathbb{R})$-closed translation invariant subspace generated by S_α. For a given signal S and $\alpha > 0$, we shall characterize the *closure statement*, $V_\alpha = L^1(\mathbb{R})$, in terms of the semigroup $\Gamma = \{\lambda + \gamma: \lambda, \gamma \in \Lambda\}$, a measure ν corresponding to S, and $\alpha > 0$. The closure statement for signals was originally discussed in Benedetto [5]; the main results of that paper are independent of the present work, although the characterization of closure that we give here is suggested by some of the results there.

The closure statement, $V_\alpha = L^1(\mathbb{R})$, is usually characterized by Wiener's Tauberian theorem: $V_\alpha = L^1(\mathbb{R})$ if and only if the Fourier transform, $\hat{S}_\alpha$, never vanishes. The reason that we have developed our characterization is because of the difficulty, in general, of determining whether or not $\hat{S}_\alpha$ vanishes. (Of course, Wiener's theorem is formulated for all of $L^1(\mathbb{R})$ and not only damped signals.)

In our paper [5] we stated two problems, one from analytic number theory and one from prediction theory, which deal with damped signals and depend on the validity of the closure theorem. We now mention two other applications. They concern the three ingredients, Γ, ν and α, of our closure theorem.

1. The usual deterministic sampling of continuous-time signals is periodic, and of course, an infinite periodic sequence of sampling times t beginning at $t = 0$ is a semigroup. It may happen that optimal (for a given project) sampling for a given continuous-time signal requires sampling over a semigroup more general than a periodic sequence. Our results provide a closure theory, which in turn is closely related to spectral theory, e.g., Benedetto [3, pp. 100–101], corresponding to semigroup sampling. Also, in the case of random sampling or deterministic periodic sampling in which noises may render some of the samples useless, our characterization of the

* Received by the editors August 27, 1980, and in final revised form April 24, 1981. This work was supported in part by a Faculty Research Grant at the University of Maryland.

† University of Maryland, College Park, Maryland 20740. Current address, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

closure statement shows the necessity of dealing with the semigroup generated by available sampling times in order to determine the range of integrable messages that can be formed from S in a damped system depending on α.

2. The results of § 5 and some preliminary facts we verified in [5, § 4] yield an explicit example of a nontrivial invariant subspace X for the weighted L^1-space,

$$L^1_{1/2} = \left\{ f: \int_{-\infty}^{\infty} |f(t)|\, e^{-t/2}\, dt < \infty \right\}.$$

The existence of such subspaces has been established by Domar using a device due to Vretblad. Our example is the subspace

$$X = \{f: f(t)\, e^{-t/2} \in V_{1/2}\},$$

in the case where $\Gamma = \{\log(n+1): n = 0, 1, \cdots\}$ and S alternatively takes the values 1 and 0. The characterization of the nontrivial invariant subspaces of weighted L^p-spaces, especially on half-lines, is difficult, and Domar, N. K. Nikolskiĭ and others have established a basic theory. Our semigroup technique provides a means of constructing these subspaces for certain weights.

Section 1 establishes notation. Our characterization of the closure statement is given in § 3 (Theorem 3.1) in a prediction theoretic setting; this characterization is more useful in checking for closure than the Tauberian theorem, where $Z\hat{S}_\alpha$ must be computed. Theorem 3.1 depends on an analytic approximation technique and the algebraic construction of a measure ν with the property that $S * \nu$ is the Heaviside function. The analytic technique involves the construction of a norm $\|\cdot\|_\alpha$, and Theorem 3.1 has the form

$$\inf_T \|T - \nu\|_\alpha = 0 \Leftrightarrow V_\alpha = L^1(\mathbb{R}),$$

where T belongs to a special class of distributions. The algebraic construction is made in § 2 (Theorem 2.1) and is the reason that Theorem 3.1 is applicable.

Theorem 3.1 has two parts. The first part provides a complete theory for a form of closure which is stronger than the conclusion $V_\alpha = L^1(\mathbb{R})$. This theory is given in § 4 (Theorems 4.1–4.3) and is applicable when ν can be approximated without recourse to the distant past. Such is the case when the differences between consecutive terms of Γ do not tend to 0 too quickly. Necessary conditions for this approximation in terms of first order distributional convergence are given in Theorem 4.4.

The second and more subtle part of Theorem 3.1 deals with the cases in which ν *can only* be approximated from the distant past. That such cases occur is verified in Proposition 5.1. The remainder of § 5 develops a theory (Theorems 5.1–5.2) to check on the closure theorem, $V_\alpha = L^1(\mathbb{R})$, when approximation from the distant past is required, and shows, by means of a weak compactness argument (Theorem 5.3), that this theory is best possible for the approximations we have constructed.

Section 6 contains some examples of closure for special semigroups as well as some constructive aspects of ν with which we did not deal in § 2.

1. Notation. If S is a signal with jumps at Λ and Γ is the semigroup generated by Λ, that is, $\Gamma = \{\lambda + \gamma: \lambda, \gamma \in \Lambda\}$, then we say that S *is a signal on the semigroup* $\Gamma = \{\gamma_j\}$. If S is a signal on Γ for which $S(\gamma_{2j}) = 1$ and $S(\gamma_{2j+1}) = 0$, $j = 0, 1, \cdots$, then S is a *0–1 signal.*

C^j_c, $j = 0, 1, \cdots, \infty$, denotes the space of j-times continuously differentiable functions on the real line $\mathbb{R}$ having compact support. For each j, C^j_c is topologized in the

usual way, e.g., Schwartz [20], and its dual is denoted by D_j. $D = D_\infty$ is the space of *distributions* and $D(X)$ is the space of distributions T having support, supp T, contained in X. D_0 is the space of *Radon measures* and it is clear that for all $j \geqq 0$

$$D_0(X) \subseteq D_j(X) \subseteq D_{j+1}(X) \subseteq D(X)$$

("X" will not be used in the case $X = \mathbb{R}$). $\mathcal{S}$ is the Schwartz space of infinitely differentiable functions f on $\mathbb{R}$: that is, f satisfies the condition $\lim_{|t|\to\infty} |t^m f^{(n)}(t)| = 0$, for each $m, n \geqq 0$. $\mathcal{S}$ is given the usual metric topology defined in terms of this polynomial condition for each derivative, e.g., Schwartz [20], and the dual space $\mathcal{S}'$ of $\mathcal{S}$ is the space of *tempered distributions*. We shall also deal with the space $D_+ = \{T \in D: \exists a = a_T$ for which supp $T \subseteq [a, \infty)\}$.

$M(X)$ is the space of *bounded Radon measures* supported by X (e.g., Benedetto [4]) and $M_f = \{\mu \in M:$ card supp $\mu < \infty\}$. It is clear that $L^1 \subseteq M \subseteq D_0$. The dual of $L^1(\mathbb{R})$ taken with the norm $\|f\|_{L^1(\mathbb{R})} = \int |f|$, is the space $L^\infty(\mathbb{R})$ of essentially bounded measurable functions. L^∞_{loc} (resp., L^1_{loc}) is the space of measurable functions each of whose elements is essentially bounded (resp., integrable) on every bounded set of $\mathbb{R}$. $C_b \subseteq L^\infty$ consists of the continuous elements of L^∞.

$\mathbb{R}^+ = \{t \in \mathbb{R}: t \geqq 0\}$ and H is the Heaviside function: $H(t) = 1$ if $t \in \mathbb{R}^+$ and $H(t) = 0$ if $t < 0$. The distributional derivative, H', of H is the Dirac δ "function", $\delta = \delta_0 \in M_f$. $\delta_\gamma \in M_f$ is the unit mass at $\gamma \in \mathbb{R}$. Finally χ_A denotes the characteristic function of the set $A \subseteq \mathbb{R}$.

2. Semigroups and a signal convolution equation.

PROPOSITION 2.1. *Let $\Gamma = \{\gamma_j: j = 0, 1, \cdots\} \subseteq [0, \infty)$ be a discrete additive semigroup and take integers $n \geqq m \geqq 0$. Then there is a unique integer $j \geqq 0$ for which*

$$[\gamma_j + \gamma_m, \gamma_{j+1} + \gamma_m) \supseteq [\gamma_n, \gamma_{n+1}). \tag{2.1}$$

Proof. Take the largest integer j such that $\gamma_j + \gamma_m \leqq \gamma_n$. Then $\gamma_{j+1} + \gamma_m > \gamma_n$. Since $\gamma_{j+1} + \gamma_m \in \Gamma$ and since $\gamma_{n+1} (> \gamma_n)$ is adjacent to γ_n, we have $\gamma_{j+1} + \gamma_m \geqq \gamma_{n+1}$, which is what we wanted to prove. □

Because of Proposition 2.1, we can make the following definition.

DEFINITION 2.1. Let $\Gamma = \{\gamma_j: j = 0, 1, \cdots\} \subseteq [0, \infty)$ be a discrete additive semigroup, and take integers $n \geqq m \geqq 0$. Then $j(m, n)$ is the uniquely determined integer j for which (2.1) holds.

We shall use the first part of the following example in Theorem 2.1.

Example 2.1. Let $\Gamma = \{\gamma_j: j = 0, 1, \cdots\} \subseteq [0, \infty)$ be a discrete additive semigroup with $0 = \gamma_0 < \gamma_1 < \gamma_2 < \cdots$, and let Λ be a set of generators of Γ.

a) There is M such that for all j, $\gamma_{j+1} - \gamma_j \leqq M$; in fact, we take $M = \gamma_1$. To see this, assume the opposite, viz., $\gamma_1 < \gamma_{j+1} - \gamma_j$ for some j. Take the largest integer n for which $n\gamma_1 \leqq \gamma_j$. By hypothesis, $(n+1)\gamma_1 < \gamma_{j+1}$, and this is a contradiction since γ_j and γ_{j+1} are adjacent elements of Γ and $\gamma_1 < (n+1)\gamma_1 < \gamma_{j+1}$.

b) In light of part a) we note that $\{\gamma_{j+1} - \gamma_j\}$ needn't be decreasing, even excluding the case that $\gamma_{j+1} - \gamma_j$ is eventually constant. For example, given γ_1, we choose $\gamma_2 \in (\gamma_1, 3\gamma_1/2)$ and $\gamma_3 = 2\gamma_1$. Then $\gamma_3 - \gamma_2 > \gamma_2 - \gamma_1$. On the other hand we always have

$$\gamma_2 < 2\gamma_1 \Rightarrow \gamma_3 \leqq 2\gamma_1,$$

since $2\gamma_1 \in \Gamma$.

c) Suppose $\Lambda = \{\gamma_1, \gamma_2\}$ is linearly independent over the rationals. Then $\lambda = \gamma_1/\gamma_2$ is irrational and so, by Hurwitz's theorem (e.g., Hardy and Wright [14, p. 164]), there are sequences $\{p_n\}, \{q_n\} \subseteq \mathbb{N}$ tending to infinity for which $|\lambda - (p_n/q_n)| < 1/(q_n^2\sqrt{5})$.

Using this we shall show that

$$\lim_{j\to\infty}(\gamma_{j+1}-\gamma_j)=0,$$

noting a similar result for any linearly independent set Λ. The limit will be verified if we show that for any $\varepsilon>0$ there is $K>0$ such that

$$\forall j', k'>K, \exists j, k\neq k' \text{ for which } |(j'-j)\lambda_1+(k'-k)\lambda_2|<\varepsilon.$$

For this ε we employ Hurwitz's theorem and choose N such that, for all $n>N$, $1/(\sqrt{5}\,q_n)<\varepsilon$. We complete the proof by writing p_n and q_n as differences k'_n-k_n and j'_n-j_n.

d) Let $G=\{\gamma-\lambda:\gamma,\lambda\in\Gamma\}$. G is a group and by the characterization of the discrete subgroups of $\mathbb{R}$ we see that if $\lim(\gamma_{j+1}-\gamma_j)=0$ then $\bar{G}=\mathbb{R}$.

THEOREM 2.1. *Let* $S=\sum_0 S(\lambda_j)\chi_{[\lambda_j,\lambda_{j+1})}\in L^\infty(\mathbb{R})$, $S(\lambda_0)\neq 0$, *be a signal with jumps at* Λ *and let* Γ *be the semigroup generated by* Λ. *Write* $S=\sum_{\gamma_j\in\Gamma}S(\gamma_j)\chi_{[\gamma_j,\gamma_{j+1})}$, *where we could have* $S(\gamma_j)=S(\gamma_{j+1})$.

a) *There is a discrete measure* (*possibly unbounded*)

$$\nu=\sum_{\gamma\in\Gamma}a_\gamma\delta_\gamma\in D_0[0,\infty) \tag{2.2}$$

such that

$$S*\nu=H \tag{2.3}$$

and

$$S^{-1}=\nu'=\sum_{\gamma\in\Gamma}a_\gamma\delta'_\gamma\in D_1[0,\infty) \tag{2.4}$$

(*that is,* $S*S^{-1}=\delta$). *Further,* ν *is the unique distributional solution of* (2.3).

b) *For integers* $n\geqq m\geqq 0$, *we write* $S(m,n)$ *for* $S(\gamma_{j(m,n)})$, *where* $j(m,n)$ *is defined in Definition* 2.1. *The coefficients* a_γ *in* (2.2) *are given by either of the formulas* (2.5) *or* (2.6):

$$\begin{aligned} a_0&=\frac{1}{S(0)},\\ a_{\gamma_n}&=a_0(1-a_0S(0,n)-a_{\gamma_1}S(1,n)-\cdots-a_{\gamma_{n-1}}S(n-1,n)) \end{aligned} \tag{2.5}$$

and, with $S(0)=1, \forall n>0$,

(2.6)

$$a_{\gamma_n}=1+\sum_{k=1}^{n}(-1)^{n-(n-k)}\sum_{0\leqq m_1<m_2<\cdots<m_k<n}S(m_1,m_2)S(m_2,m_3)\cdots S(m_{k-1},m_k)S(m_k,n).$$

The inner summand of (2.6) *has* $\binom{n}{k}$ *terms and the total sum has* 2^n *terms.*

c) *Let* $S(0)=1$ *and define*

$$\nu_n=\sum_0^n a_{\gamma_j}\delta_{\gamma_j}\quad\text{and}\quad H_n=S*\nu_n.$$

Then for all n

$$|a_{\gamma_n}|\leqq(1+\|S\|_\infty)^n,\qquad |H_n|\leqq(1+\|S\|_\infty)^{n+1}\quad \text{on } \mathbb{R},$$

$$H_n=1\quad \text{on } [0,\gamma_{n+1}),$$

and

$$H_n = \sum_{j=0}^{\infty} S(\gamma_j)\Big(\sum_{m=0}^{n} a_{\gamma_m}\chi_{[\gamma_j+\gamma_m,\gamma_{j+1}+\gamma_m)}\Big) \quad \text{on } \mathbb{R}.$$

Proof. i) Suppose (2.3) has been proved. By its definition in (2.2) we known that $\nu \in D[0, \infty)$ since Γ is discrete. Thus, by (2.3),

$$\delta = H' = S * \nu' = S * (\sum a_\gamma \delta'_\gamma), \tag{2.7}$$

and so (2.4) follows.

ii) Since S^{-1} exists, we automatically obtain the uniqueness in part a). In fact, if $S * T = H$, then $T = H * S^{-1} = \sum a_\gamma (H * \delta'_\gamma) = \sum a_\gamma \delta_\gamma = \nu$ by (2.4).

iii) The proofs of (2.3) and (2.5) are done together in parts iv)–vii) of the proof.

iv) Let $a_0 = 1/S(0)$. Then

$$a_0\delta_0 * S = \sum_0^{\infty} \left(\frac{S(\gamma_j)}{S(0)}\right)\chi_{[\gamma_j,\gamma_{j+1})}$$

and, in particular,

$$a_0\delta_0 * S = 1 \quad \text{on } [0, \gamma_1).$$

On $[\gamma_1, \gamma_2)$, we have

$$a_0\delta_0 * S = \frac{S(\gamma_1)}{S(0)},$$

and we want to choose a_{γ_1} such that

$$\forall t \in [\gamma_0, \gamma_2), (a_0\delta_0 + a_{\gamma_1}\delta_{\gamma_1}) * S(t) = 1.$$

To this end, we note that

$$S(t - \gamma_1) = S(0) \quad \text{on } [\gamma_1, \gamma_2),$$

since $t \in [\gamma_1, \gamma_2)$ implies $t - \gamma_1 \in [0, \gamma_2 - \gamma_1)$, which in turn is contained in $[0, \gamma_1)$ since $2\gamma_1 \geqq \gamma_2$. Thus we choose a_{γ_1} such that $(S(\gamma_1)/S(0)) + a_{\gamma_1}S(0) = 1$, i.e.,

$$a_{\gamma_1} = \frac{1}{S(0)} - S(\gamma_1)\frac{1}{S(0)^2}.$$

Consequently, for this a_{γ_1}, we have

$$(a_0\delta_0 + a_{\gamma_1}\delta_{\gamma_1}) * S = \sum_{j=0}^{\infty} S(\gamma_j)(a_{\gamma_0}\chi_{[\gamma_j,\gamma_{j+1})} + a_{\gamma_1}\chi_{[\gamma_1+\gamma_j,\gamma_1+\gamma_{j+1})})$$

and, in particular,

$$(a_0\delta_0 + \alpha_{\gamma_1}\delta_{\gamma_1}) * S = 1 \quad \text{on } [0, \gamma_2).$$

v) We shall now explicitly see how Proposition 2.1 and Definition 2.1 play a role in the proof. We want to choose a_{γ_2} such that

$$\forall t \in [\gamma_0, \gamma_3), \quad (a_0\delta_0 + a_{\gamma_1}\delta_{\gamma_1} + a_{\gamma_2}\delta_{\gamma_2}) * S(t) = 1. \tag{2.8}$$

$\delta_{\gamma_2} * S = 0$ in $[\gamma_0, \gamma_2)$ and

$$(a_0\delta_0 + a_{\gamma_1}\delta_{\gamma_1}) * S = 1 \quad \text{in } [0, \gamma_2).$$

Thus the first step in finding a_{γ_2} for which (2.8) is valid is to show that $\delta_0 * S$, $\delta_{\gamma_1} * S$ and $\delta_{\gamma_2} * S$ are constant on $[\gamma_2, \gamma_3)$.

From Example 2.1a we know that $\gamma_1 \geqq \gamma_{j+1} - \gamma_j$. Hence, if $t \in [\gamma_2, \gamma_3)$, then $t - \gamma_2 \in [0, \gamma_3 - \gamma_2) \subseteq [0, \gamma_1)$ and so $S(t - \gamma_2) = S(0)$.

Also, $\delta_{\gamma_1} * S = \sum S(\gamma_j)\chi_{[\gamma_j+\gamma_1, \gamma_{j+1}+\gamma_1)} = S(\gamma_{j(1,2)})$ on $[\gamma_2, \gamma_3)$, where $\gamma_{j(1,2)}$ is determined as in Proposition 2.1. $j(1, 2)$ could be 0 or 1 depending on Γ. Finally, $\delta_0 * S = S(\gamma_2)(= S(\gamma_{j(0,2)}))$ on $[\gamma_2, \gamma_3)$. We can now solve for a_{γ_2}, since a_{γ_1} and a_0 are known and since (2.8) leads to the equation $a_{\gamma_2}S(0) + a_{\gamma_1}S(\gamma_{j(1,2)}) + a_0 S(\gamma_2) = 1$. Thus,

$$a_{\gamma_2} = \frac{1}{S(0)}(1 - a_0 S(\gamma_{j(0,2)}) - a_{\gamma_1} S(\gamma_{j(1,2)})).$$

vi) We proceed in the same way for a_{γ_3}. The problem is to choose a_{γ_3} such that

$$\forall t \in [\gamma_0, \gamma_4), (a_0\delta_0 + a_{\gamma_1}\delta_{\gamma_1} + a_{\gamma_2}\delta_{\gamma_2} + a_{\gamma_3}\delta_{\gamma_3}) * S(t) = 1. \tag{2.9}$$

$\delta_{\gamma_3} * S = 0$ on $[0, \gamma_3)$ and $(a_0\delta_0 + a_{\gamma_1}\delta_{\gamma_1} + a_{\gamma_2}\delta_{\gamma_2}) * S = 1$ on $[0, \gamma_3)$ by the previous step. Because of Proposition 2.1 we see that on $[\gamma_3, \gamma_4)$ we have $\delta_{\gamma_3} * S = S(0)$,

$$\delta_{\gamma_2} * S = \sum S(\gamma_j)\chi_{[\gamma_j+\gamma_2, \gamma_{j+1}+\gamma_2)} = S(\gamma_{j(2,3)}),$$

$$\delta_{\gamma_1} * S = S(\gamma_{j(1,3)}) \quad \text{and} \quad \delta_0 * S = S(\gamma_3)(= S(\gamma_{j(0,3)}).$$

We can now solve for a_{γ_3}, since a_{γ_2}, a_{γ_1} and a_0 are known and since (2.9) leads to the equation

$$a_{\gamma_3}S(0) + a_{\gamma_2}S(\gamma_{j(2,3)}) + a_{\gamma_1}S(\gamma_{j(1,3)}) + a_0 S(\gamma_3) = 1;$$

thus,

$$a_{\gamma_3} = \frac{1}{S(0)}(1 - a_0 S(\gamma_{j(0,3)}) - a_{\gamma_1} S(\gamma_{j(1,3)}) - a_{\gamma_2} S(\gamma_{j(2,3)}).$$

vii) This method of construction is obviously valid for all n. Thus, as in (2.8) and (2.9), if we are given $a_0, a_{\gamma_1}, \cdots, a_{\gamma_{n-1}}$, then

$$\forall t \in [0, \gamma_{n+1}), \quad H_n(t) = \left(\sum_{j=0}^{n} a_{\gamma_j}\delta_{\gamma_j}\right) * S(t) = 1,$$

where a_{γ_n} is given by (2.5). Consequently, (2.3) and (2.5) are valid, as well as part of c). (2.3) and parts i) and ii) of the proof yield all of part a).

viii) The verification of (2.6) follows by systematically substituting the values a_{γ_j}, $0 \leqq j < n$, obtained by (2.5), into the formula for a_{γ_n} (also given by (2.5)). Thus, $a_{\gamma_1} = 1 - S(0, 1)$;

$$\begin{aligned} a_{\gamma_2} &= 1 - a_0 S(0, 2) - a_{\gamma_1} S(1, 2) \\ &= 1 - a_0 S(0, 2) - S(1, 2)(1 - a_0 S(0, 1)) \\ &= 1 - (S(0, 2) + S(1, 2)) + S(0, 1)S(1, 2), \end{aligned}$$

$$\begin{aligned} a_{\gamma_3} &= 1 - S(0, 3) - a_{\gamma_1} S(1, 3) - a_{\gamma_2} S(2, 3) \\ &= 1 - S(0, 3) - S(1, 3)(1 - S(0, 1)) - S(2, 3)(1 - (S(0, 2) + S(1, 2)) + S(0, 1)S(1, 2)) \\ &= 1 - (S(0, 3) + S(1, 3) + S(2, 3)) + (S(0, 1)S(1, 3) + S(0, 2)S(2, 3) + S(1, 2)S(2, 3)) \\ &\quad - S(0, 1)S(1, 2)S(2, 3), \end{aligned}$$

$$\begin{aligned} a_{\gamma_4} &= 1 - S(0, 4) - a_{\gamma_1} S(1, 4) - a_{\gamma_2} S(2, 4) - a_{\gamma_3} S(3, 4) \\ &= 1 - S(0, 4) - S(1, 4)(1 - S(0, 1)) - S(2, 4)(1 - S(0, 2) - S(1, 2) + S(0, 1)S(1, 2)) \end{aligned}$$

$$
\begin{aligned}
&-S(3,4)(1-S(0,3)-S(1,3)-S(2,3)+S(0,1)S(1,3)+S(0,2)S(2,3)\\
&\qquad +S(1,2)S(2,3)-S(0,1)S(1,2)S(2,3))\\
&=1-(S(0,4)+S(1,4)+S(2,4)+S(3,4))\\
&\quad +(S(0,1)S(1,4)+S(0,2)S(2,4)+S(1,2)S(2,4)\\
&\qquad +S(0,3)S(3,4)+S(1,3)S(3,4)+S(2,3)S(3,4))\\
&\quad -(S(0,1)S(1,2)S(2,4)+S(0,1)S(1,3)S(3,4)\\
&\qquad +S(0,2)S(2,3)S(3,4)+S(1,2)S(2,3)S(3,4))\\
&\quad +S(0,1)S(1,2)S(2,3)S(3,4).
\end{aligned}
$$

The general case follows in a similar way, and so (2.6) is verified.

Since we have already proved (2.5), the proof of part b) is complete.

ix) From (2.5) we have $a_{\gamma_1}=1-S(0,1)$, so that $|a_{\gamma_1}|\leqq 1+\|S\|_\infty$. From the definition of H_1,

$$
\begin{aligned}
|H_1|=|S+a_{\gamma_1}S(\cdot-\gamma_1)|&\leqq\|S\|_\infty(1+|a_{\gamma_1}|)\\
&\leqq\|S\|_\infty(1+1+\|S\|_\infty)\leqq 2\|S\|_\infty+\|S\|_\infty^2\leqq(1+\|S\|_\infty)^2.
\end{aligned}
$$

Next, $a_{\gamma_2}=1-S(0,2)-a_{\gamma_1}S(1,2)$, and so

$$
|a_{\gamma_2}|\leqq 1+\|S\|_\infty(1+|a_{\gamma_1}|)=1+\|S\|_\infty(1+1+\|S\|_\infty)=(1+\|S\|_\infty)^2;
$$

and, from the definition of H_2,

$$
\begin{aligned}
|H_2|=|S+a_{\gamma_1}S(\cdot-\gamma_1)+a_{\gamma_2}S(\cdot-\gamma_2)|&\leqq\|S\|_\infty(1+|a_{\gamma_1}|+|a_{\gamma_2}|)\\
&\leqq\|S\|_\infty(1+1+\|S\|_\infty+(1+\|S\|_\infty)^2)\\
&=3\|S\|_\infty+3\|S\|_\infty^2+\|S\|_\infty^3\leqq(1+\|S\|_\infty)^3.
\end{aligned}
$$

For a_{γ_3} we have

$$
\begin{aligned}
|a_{\gamma_3}|&\leqq 1+\|S\|_\infty+|a_{\gamma_1}|\|S\|_\infty+|a_{\gamma_2}|\|S\|_\infty\\
&\leqq 1+\|S\|_\infty+(1+\|S\|_\infty)\|S\|_\infty+(1+\|S\|_\infty^2)\|S\|_\infty\\
&=1+\|S\|_\infty(1+(1+\|S\|_\infty)+(1+\|S\|_\infty^2))=(1+\|S\|_\infty)^3,
\end{aligned}
$$

amd the general estimate for a_{γ_n} follows since

$$
1+r(1+(1+r)+(1+r)^2+\cdots+(1+r)^{n-1})=(1+r)^n.
$$

Similarly, for H_n we have

$$
|H_n|=|H_{n-1}+a_{\gamma_n}S(\cdot-\gamma_n)|\leqq(1+\|S\|_\infty)^n+|a_{\gamma_n}|\|S\|_\infty\leqq(1+\|S\|_\infty)^n(1+\|S\|_\infty).
$$

The final formula for H_n in part c) is clear from previous considerations. □

THEOREM 2.2. *Let S be a 0–1 signal and take ν as defined in* Theorem 2.1. *Then $a_0=1$ and*

$$
\forall n>0,\quad a_{\gamma_n}=-\sum_{i=1}^{j_n-1}(-1)^{m_i}a_{\gamma_{ni}},\tag{2.10}
$$

where the set $\{(m_i,n_i)\colon i=1,\cdots,j_n\}$ of ordered pairs is characterized by the property that

$$
\forall i=1,\cdots,j_n,\quad \gamma_{m_i}+\gamma_{n_i}=\gamma_n.\tag{2.11}
$$

In particular, $\gamma_{n_{j_n}} = \gamma_n$.

Proof.

$$\delta = S * S^{-1} = \sum_{k=0}^{\infty} \chi_{[\gamma_{2k},\gamma_{2k+1})} * \sum_{j=0}^{\infty} a_{\gamma_j}\delta'_{\gamma_j} = \sum\sum a_{\gamma_j}\delta'_{\gamma_j} * \chi_{[\gamma_{2k},\gamma_{2k+1})}$$

$$= \sum\sum a_{\gamma_j}(\delta_{\gamma_j} * \chi'_{[\gamma_{2k},\gamma_{2k+1})})$$

$$= \sum\sum a_{\gamma_j}\delta_{\gamma_j} * (\delta_{\gamma_{2k}} - \delta_{\gamma_{2k+1}})$$

$$= \sum\sum a_{\gamma_j}(\delta_{\gamma_{2k}+\gamma_j} - \delta_{\gamma_{2k+1}+\gamma_j}).$$

Consequently,

$$\delta = \sum_{j=0}^{\infty} a_{\gamma_j}(\delta_{\gamma_0+\gamma_j} - \delta_{\gamma_1+\gamma_j} + \delta_{\gamma_2+\gamma_j} - \delta_{\gamma_3+\gamma_j} + \delta_{\gamma_4+\gamma_j} - \delta_{\gamma_5+\gamma_j} + \cdots) \tag{2.12}$$

$$= a_{\gamma_0}\delta_{\gamma_0+\gamma_0} - a_{\gamma_0}\delta_{\gamma_1+\gamma_0} + a_{\gamma_0}\delta_{\gamma_2+\gamma_0} - a_{\gamma_0}\delta_{\gamma_3+\gamma_0} + a_{\gamma_0}\delta_{\gamma_4+\gamma_0} - a_{\gamma_0}\delta_{\gamma_5+\gamma_0} + \cdots$$
$$+ a_{\gamma_1}\delta_{\gamma_0+\gamma_1} - a_{\gamma_1}\delta_{\gamma_1+\gamma_1} + a_{\gamma_1}\delta_{\gamma_2+\gamma_1} - a_{\gamma_1}\delta_{\gamma_3+\gamma_1} + a_{\gamma_1+\gamma_1} - a_{\gamma_1}\delta_{\gamma_5+\gamma_1} + \cdots$$
$$+ a_{\gamma_2}\delta_{\gamma_0+\gamma_2} - a_{\gamma_2}\delta_{\gamma_1+\gamma_2} - a_{\gamma_2}2\delta_{\gamma_3+\gamma_2} + a_{\gamma_2}\delta_{\gamma_4+\gamma_2} - a_{\gamma_2}\delta_{\gamma_5+\gamma_2} + \cdots$$
$$+ \cdots.$$

We now combine like terms in (2.12), noting that the coefficient of δ_{γ_n}, $n \geqq 1$, is 0 since the left-hand side is δ. Thus, because of (2.12), we read off that $a_{\gamma_0} = 1$. Also, $a_{\gamma_1} = 1$ since $\delta_{\gamma_1} = \delta_{\gamma_m} + \delta_{\gamma_n}$ only for the cases $(m, n) = (0, 1), (1, 0)$; thus, from (2.12),

$$0 = \delta_{\gamma_1}(-a_{\gamma_0} + a_{\gamma_1}), \quad \text{i.e.,} \quad a_{\gamma_1} - a_{\gamma_0} = 0,$$

and so $a_{\gamma_1} = 1$. Generally, for each $n \geqq 1$, we define $\{(m_i, n_i)\}$ as in (2.11). From the array in (2.12), we have

$$\begin{aligned} n = m_1 > m_2 > \cdots > m_{j_n} = 0, \\ 0 = n_1 < n_2 < \cdots < n_{j_n} = n, \end{aligned} \tag{2.13}$$

so that we are scanning a triangle (in this array). Because of (2.13) we see that

$$\forall n \geqq 1, \quad 2 \leqq j_n \leqq n + 1.$$

Also, because of (2.12) we can now write the coefficient of δ_n, $n \geqq 1$, as

$$\sum_{i=1}^{j_n} (-1)^{m_i} a_{\gamma_{ni}} = 0.$$

Since $m_{j_n} = 0$ and $n_{j_n} = n$, we obtain our result. □

Example 2.2. a) If $\Gamma = \{0\} \cup \mathbb{N}$ *and* $S = \sum_0 \chi_{[2n,2n+1)}$ then

$$S * (\delta + \delta_1) = H,$$

and $a_0 = a_1 = 1$ and $a_n = 0$ for all $n > 1$.

b) In general, $\sum a_\gamma$ diverges and, in fact, if S is a 0–1 signal associated with an arbitrary semigroup Γ then $a_\gamma \in \mathbb{Z}$ for all $\gamma \in \Gamma$ and, hence, $\sum a_\gamma$ converges if and only if $a_\gamma = 0$ for all large γ.

Remark 2.1. The representation of a_γ in (2.6) is analogous to the structure of the polynomials constructed in [5, Thm. 3.3].

3. The closure statement for damped signals. Let S be a signal and let D_+ be the space of distributions on $\mathbb{R}$ supported on half-lines $[a, \infty)$, where a varies. If $\alpha > 0$, $T \in D_+$ and $T * S \in L^1_{\text{loc}}$, we define

$$\|T\|_\alpha = \int_{-\infty}^{\infty} e^{-\alpha t} |T * S(t)|\, dt. \tag{3.1}$$

We then set

$$Y_\alpha = \{T \in D_+ : T * S \in L^1_{\text{loc}}, \|T\|_\alpha < \infty, \text{ and } \exists \beta < \alpha \text{ such that } e^{-\beta t} T \in \mathcal{S}'\}. \tag{3.2}$$

Remark 3.1. a) If $T_1, T_2 \in D_+$, then $T_1 * T_2 \in D_+$ exists, and supp $T_1 * T_2 \subseteq$ supp $T_1 +$ supp T_2 (e.g., Schwartz [20, p. 172]).

b) Let $S = \chi_{[0,\gamma)}$, and let $T = \mu' \in D_+$, where μ is a measure having a discrete or singular part and where supp $\mu \subseteq [a, a + \gamma/2]$. Then

$$T * S \notin L^1_{\text{loc}};$$

in fact, $T * S = \mu * S' = \mu * \delta - \mu * \delta_\gamma = \mu - \tau_\gamma \mu$. We use the hypothesis on supp μ to ensure that $\mu - \tau_\gamma \mu$ does not cancel any of the "singularities" in μ.

PROPOSITION 3.1. *Given a signal S:*

a) *For all $\alpha > 0$, $(Y_\alpha, \| \|_\alpha)$ is a normed vector space.*

b) *If $\alpha_1 \leqq \alpha_2$, then $Y_{\alpha_1} \subseteq Y_{\alpha_2}$.*

c) *For all $\alpha > 0$, $M_f \subseteq Y_\alpha$.*

d) *For all $\mu \in D_0 \cap D_+$, $\mu * S \in L^\infty_{\text{loc}} \subseteq L^1_{\text{loc}}$.*

Proof. a) It is clear that $\| \cdot \|_\alpha$ is a seminorm on Y_α. If $T \in Y_\alpha$ and $\|T\|_\alpha = 0$, then $T * S = 0$ a.e. by properties of the Lebesgue integral. Since $T \in D_+$ and S is not the zero-function, we conclude that $T = 0$ by the Titchmarsh convolution theorem (e.g., Schwartz [20, p. 173]).

b) Take $\alpha_2 \geqq \alpha_1$ and $T \in Y_{\alpha_1}$. To prove $\|T\|_{\alpha_2} < \infty$ and $e^{-\beta t} T \in \mathcal{S}'$ for some $\beta < \alpha_2$.

$$\|T\|_{\alpha_2} = \int e^{-\alpha_2 t} |T * S(t)|\, dt = \int e^{-\alpha_1 t} e^{-(\alpha_2 - \alpha_1)t} |T * S(t)|\, dt \leqq e^{-c(\alpha_2 - \alpha_1)} \|T\|_{\alpha_1},$$

where $c = \min\{0, a\}$ and supp $T \subseteq [a, \infty)$. Thus, $\|T\|_{\alpha_2} < \infty$. Next, we write $\beta = \alpha_1 - \varepsilon$ and

$$e^{-(\alpha_2 - \varepsilon)t} T = e^{-(\alpha_1 - \varepsilon)t} e^{-(\alpha_2 - \alpha_1)t} T = e^{-\beta t} T_1.$$

By hypothesis we take β so that $e^{-\beta t} T \in \mathcal{S}'$, and since $\alpha_2 \geqq \alpha_1$, we also have $e^{-\beta t} T_1 \in \mathcal{S}'$.

c) If $\mu \in M_f$, then $\mu * S$ is a finite sum of translates of S and so the inclusion, $M_f \subseteq Y_\alpha$, is clear.

d) Part d) follows from the definition of the Stieltjes integral in the following way. By the Riesz representation theorem, $\mu = F'$ in the distributional sense, where $F \in BV_{\text{loc}}$,

$$F(x) = F(x+) \quad \text{and} \quad F = 0 \quad \text{on } (-\infty, a).$$

Thus,

$$\mu * S(x) = F(a) S(x) + \int_a^x S(x - t)\, d\mu(t).$$

$F(a) S(\,\cdot\,) \in L^\infty(\mathbb{R})$ and

$$\left| \int_a^x S(x - t)\, d\mu(t) \right| \leqq \|S\|_\infty \int_a^x d|\mu|(t) = \|S\|_\infty \|\mu \chi_{[a,x]}\|_1,$$

and so $\mu * S \in L^\infty_{\text{loc}}$.

Note that this result depends on the fact that if $\mu \in D_0$ and $K \subseteq \mathbb{R}$ is compact then $\mu\chi_K \in M(\mathbb{R})$. □

PROPOSITION 3.2. *Given a signal S, $\alpha > 0$, and an element $T \in Y_\alpha$ there is $\beta < \alpha$ such that the following are true for the bilateral Laplace transform L and $s = \sigma + i\tau$:*

a) *$L(T)(s)$ is an analytic function in the half-plane, $\sigma > \beta$.*

b) *$L(T * S)(s) = \int_{-\infty}^{\infty} e^{-st} T * S(t)\, dt$ for $\sigma > \beta$.*

c) *$L(T * S)(s) = L(T)(s)L(S)(s)$ for $\sigma > \beta$.*

Proof. By the definition of Y_α, we choose $\beta \in (0, \alpha)$ so that $e^{-\beta t}T \in \mathcal{S}'$. Then, using $\langle S, \varphi \rangle$ to indicate the functional value $S(\varphi)$, we see that

$$L(T)(s) = \langle T, e^{-st} \rangle = \langle T e^{-\beta t}, e^{-(s-\beta)t} \rangle$$

is well defined for $\sigma > \beta$. In fact, $T \in D_+$ implies that $e^{-(s-\beta)t} \in \mathcal{S}$ on supp $(e^{-\beta t}T)$.

The analyticity follows by the usual arguments (e.g., Benedetto [1]) and so a) follows. b) is immediate and c) follows since $L(S)(s)$ is analytic in the half-plane $\sigma > 0$ and because

$$L(T * S)(s) = \langle T_t, e^{-st} \rangle \langle S_\mu, e^{-s\mu} \rangle. \qquad \square$$

For the closure theorem, Theorem 3.1, we shall need the following two spaces:

$$X_\alpha = \{T \in Y_\alpha : \exists \{\mu_n\} \subseteq M_f \text{ such that } \|T - \mu_n\|_\alpha \to 0\},$$

$$X_{\alpha,0} = \{T \in Y_\alpha \cap D[0, \infty) : \exists \{\mu_n\} \subseteq M_f([0, \infty)) \text{ such that } \|T - \mu_n\|_\alpha \to 0\}.$$

Also, given a signal S and $\alpha > 0$, we define S_α as the product $S_\alpha(t) = e^{-\alpha t}S(t)$, and we let V_α (resp., V_α^+) be the L^1-closed (resp., $L^1(\mathbb{R}^+)$-closed) translation invariant (resp., right translation invariant) subspace generated by S_α.

Part a) of Theorem 3.1 depends on Wiener's Tauberian theorem, part b) depends on the Nyman–Beurling theory, and if the conditions of part b) are valid, then so also are those of part a).

THEOREM 3.1. (Closure theorem). *Given a signal S and $\alpha > 0$, the following hold:*

a) $\inf_{T \in X_\alpha} \|T - \nu\|_\alpha = 0$ *if and only if* $V_\alpha = L^1(\mathbb{R})$. *$X_\alpha$ can be replaced by Y_α or M_f.*

b) $\inf_{T \in X_{\alpha,0}} \|T - \nu\|_\alpha = 0$ *if and only if* $V_\gamma^+ = L^1(\mathbb{R}^+)$ *for all* $\gamma \geqq \alpha$.

c) *If $e^{-\beta t}\nu \in \mathcal{S}'$ for some $\beta < \alpha$, then $V_\alpha = L^1(\mathbb{R})$.*

d) $\|T - \nu\|_\alpha = 0$ *for some $T \in Y_\alpha$ if and only if $\nu \in Y_\alpha$. In this case $T = \nu$ since $\|\cdot\|_\alpha$ is a norm.*

Proof. c) If $e^{-Bt}\nu \in \mathcal{S}'$ for some $\beta < \alpha$, then $\nu \in Y_\alpha$ and so $\inf_{T \in Y_\alpha} \|T - \nu\|_\alpha = \|\nu - \nu\|_\alpha = 0$. The closure follows from part a).

d) If $\|T - \nu\|_\alpha = 0$ for some $T \in Y_\alpha$, then $S * (T - \nu) = 0$, a.e., and since S, $(T - \nu) \in D_+$, we conclude from the Titchmarsh convolution theorem that $T = \nu$. The converse is obvious.

b)i) Suppose $V_\alpha = L^1(\mathbb{R}^+)$; we don't have to consider any $\beta \geqq \alpha$ to verify this direction. Since $e_\alpha \in L^1(\mathbb{R}^+)$, where $e_\alpha(t) = e^{-\alpha t}\chi_{[0,\infty)}(t)$, we can approximate it in terms of V_α. Thus, given $\varepsilon > 0$, there is a finite set $F \subseteq [0, \infty)$ for which

$$\|e_\alpha(\cdot) - \sum_{t \in F} c_t e_\alpha(\cdot - t)S(\cdot - t)\|_{L^1(\mathbb{R}^+)} < \varepsilon. \tag{3.3}$$

Since $\sum c_t e_\alpha(x - t)S(x - t) = e_\alpha(x)S * \mu(x)$, where $\mu = \sum (c_t e_\alpha(-t))\delta_t$, (3.3) becomes

$$\|\nu - \mu\|_\alpha < \varepsilon,$$

and, of course, $\mu \in M_f[0, \infty) \subseteq X_{\alpha,0}$. Thus, we have verified the necessary conditions for closure.

ii) The proof of sufficiency is the difficult direction of part b) and involves the Nyman–Beurling theory (e.g., [17], [18]). By hypothesis, we choose $\{\mu_n\} \subseteq M_f[0, \infty)$ for which

$$\lim_{n\to\infty} \|\nu - \mu_n\|_\alpha = 0. \tag{3.4}$$

For any $s = \sigma + i\tau$, $\sigma \geqq 0$, we compute

$$|L(e_\alpha(S * \mu_n))(s) - L(e_\alpha)(s)| = \left| \int_0^\infty e^{-(s+\alpha)t}(1 - S * \mu_n(t))\, dt \right| \leqq \|\nu - \mu_n\|_\alpha,$$

since $\nu * S = H$. Consequently, because of (3.4),

$$\lim_n L(e_\alpha(S * \mu_n))(s) = \frac{1}{s+\alpha},$$

where the convergence is uniform in the half-plane $\sigma \geqq 0$. Therefore, for all s for which $\sigma \geqq 0$,

$$\exists n \quad \text{such that } |L(e_\alpha(S * \mu_n))(s)| > 0. \tag{3.5}$$

We shall use (3.5) to prove that

$$\forall s,\ \sigma \geqq 0, \quad |L(S_\alpha)(s)| > 0. \tag{3.6}$$

Suppose $L(S_\alpha)(s) = 0$ for some s, $\sigma \geqq 0$. Then, if $x \geqq 0$,

$$L(e_\alpha(\tau_x S))(s) = e^{-(s+\alpha)x} \int_{-x}^\infty e^{-(s+\alpha)t} S(t)\, dt = e^{-(s+\alpha)x} L(S_\alpha)(s)$$

and hence, $L(e_\alpha(S * \mu_n))(s) = 0$ for all n.

This contradicts (3.5), and so (3.6) is obtained. Nyman's theorem asserts that if $f \in L^1(\mathbb{R}^+)$ does not vanish identically in any neighborhood of 0, then the variety V^+ generated by f is all of $L^1(\mathbb{R}^+)$ if and only if $|L(f)(s)| > 0$ for each $s = \sigma + i\tau$, $\sigma \geqq 0$, e.g., Nyman [18], Koosis [17] and Benedetto [5, Remark 1.3].

Therefore, in our case, we have $V_\alpha^+ = L^1(\mathbb{R}^+)$ because of (3.6). The result is also true for all $\gamma \geqq \alpha$ since the analogous condition for γ corresponding to (3.6) is obviously true since we are dealing with exponentials.

a)i) We begin by verifying sufficiency. Suppose $V_\alpha \neq L^1(\mathbb{R})$. Then $\hat{S}_\alpha(\tau) = 0$ for some $\tau \in \mathbb{R}$ by the Tauberian theorem (e.g., Benedetto [3]). For this τ we choose any $\varepsilon > 0$ for which

$$0 < \varepsilon < \left| \int e^{-(\alpha+i\tau)t} H(t)\, dt \right| = \frac{1}{|\alpha + i\tau|}.$$

Then, by hypothesis, there is $T_\varepsilon \in X_\alpha$ (resp., Y_α or M_f) such that $\|\nu - T_\varepsilon\|_\alpha < \varepsilon$, and hence,

$$\begin{aligned}
\varepsilon &< \left| \int_{-\infty}^\infty e^{-(\alpha+i\tau)} S * \nu(t)\, dt \right| \\
&= \left| \int_{-\infty}^\infty e^{-(\alpha+i\tau)t} S * (\nu - T_\varepsilon)(t)\, dt + \int_{-\infty}^\infty e^{-(\alpha+i\tau)t} S * T_\varepsilon(t)\, dt \right| \\
&\leqq \|\nu - T_\varepsilon\|_\alpha + |L(S * T_\varepsilon)(\alpha + i\tau)| \\
&< \varepsilon + L(S)(\alpha + i\tau) L(T_\varepsilon)(\alpha + i\tau) = \varepsilon,
\end{aligned}$$

since $L(S)(\alpha + i\tau) = \hat{S}_\alpha(\tau)$. "$\varepsilon < \varepsilon$" is the desired contradiction, and so V_α must be all of $L^1(\mathbb{R})$.

In this calculation it was important to be able to use Proposition 3.2c.

ii) Suppose $V_\alpha = L^1(\mathbb{R})$. By this hypothesis and the fact that $e_\alpha \in L^1(\mathbb{R})$ we can find, for a given $\varepsilon > 0$, a finite set $F \subseteq \mathbb{R}$ for which

$$\|e_\alpha(\cdot) - \sum_{t\in F} c_t e_\alpha(\cdot - t)S(\cdot - t)\|_{L^1(\mathbb{R})} < \varepsilon.$$

As in (3.3), this expression can be written as $\|\nu - \mu\|_\alpha < \varepsilon$, and so we have verified the necessary conditions since $\mu \in M_f \subseteq X_\alpha \subseteq Y_\alpha$. □

PROPOSITION 3.3. *Given a signal S and $\alpha > 0$. Suppose there is a sequence $\{\mu_n\} \subseteq M_f[a, \infty)$, with a fixed, for which $\lim \|\nu - \mu_n\|_\alpha = 0$. Then, for all $\gamma \geqq \alpha$, $V_\gamma = L^1(\mathbb{R})$.*

Proof. Take $\gamma \geqq \alpha$. $\|\nu - \mu_n\|_\gamma = \|e_\alpha e_{(\beta-\alpha)}(S * (\nu - \mu_n))\|_{L^1(\mathbb{R})} \leqq e_{(\beta-\alpha)}(a)\|\nu - \mu_n\|_\alpha$, and this term tends to 0. The result follows from Theorem 3.1. □

4. Closure in terms of causal approximation to the ν-system. For a given signal S on a semigroup Γ, we found a measure ν in Theorem 2.1 for which $S * \nu = H$. For any signal $T = \sum T(\gamma_j)\chi_{[\gamma_j, \gamma_{j+1})}$ on Γ, we can associate the signal

$$T_\nu = \nu * T = \sum\sum a_{\gamma_j} T(\gamma_k)\chi_{[\gamma_k+\gamma_j, \gamma_{k+1}+\gamma_j)} = \sum T_\nu(\gamma_j)\chi_{[\gamma_j, \gamma_{j+1})}.$$

In particular, $\delta_\gamma * T$ is a signal on Γ for each $\gamma \in \Gamma$. Extending the usual terminology from signal processing, we say that ν is a *semigroup invariant system*, the *ν-system* associated with S.

The most obvious corollary of the closure theorem is the sufficiency of

$$\lim_{n\to\infty} \|\nu - \nu_n\|_\alpha = 0 \tag{4.1}$$

to ensure the closure, $V_\gamma^+ = L^1(\mathbb{R}^+)$ for all $\gamma \geqq \alpha$. The ν-system is *causal* since supp $\nu \subseteq \mathbb{R}^+$; cf. Oppenheim and Schafer [19, p. 16]. In this section we investigate the approximation of ν by means of the canonical sequence $\{\nu_n\}$. To this end, we take a signal S for which $S(0) = 1$ and a fixed number $\alpha > 0$, and we define

$$C = 1 + \|S\|_\infty \quad \text{and} \quad b_{n,\alpha} = \int_{\gamma_n}^{\gamma_{n+1}} e^{-\alpha t}\, dt. \tag{4.2}$$

THEOREM 4.1. *Given a signal S and $\alpha > 0$ the following hold:*

a) *If*

$$\lim_{m\to\infty} \sum_{j=m}^{\infty} \int_{\gamma_j}^{\gamma_{j+1}} e^{-\alpha t} \left| \sum_{k=m}^{j} a_{\gamma_k} S(t - \gamma_k) \right| dt = 0, \tag{4.3}$$

then (4.1) *is valid and so $V_\gamma^+ = L^1(\mathbb{R}^+)$ for each $\gamma \geqq \alpha$.*

b) *Condition* (4.3) *is satisfied if*

$$\lim_{m\to\infty} \sum_{j=m}^{\infty} b_{j,\alpha} \left(\sum_{k=m}^{j} |a_{\gamma_k}| \right) = 0. \tag{4.4}$$

c) *Condition* (4.4) *is satisfied if any of the following equivalent conditions is valid:*

$$\sum_{k=1}^{\infty} b_{k,\alpha} C^k < \infty, \tag{4.5}$$

$$\lim_{m\to\infty} \left(\sum_{j=m}^{\infty} b_{j,\alpha} \left(\sum_{k=m}^{j} C^k \right) \right) = 0, \tag{4.6}$$

$$\lim_{m\to\infty} \left(\sum_{j=m+1}^{\infty} b_{j,\alpha} \right) \left(\sum_{k=0}^{m} C^k \right) = 0. \tag{4.7}$$

Proof. a)

$$\|\nu-\nu_{m-1}\|_\alpha=\int_{\gamma_m}^{\infty}e^{-\alpha t}|S*(\nu-\nu_{m-1})(t)|\,dt=\int_{\gamma_m}^{\infty}e^{-\alpha t}\left|\sum_{k=m}^{\infty}a_{\gamma_k}S(t-\gamma_k)\right|\,dt.$$

If $t\in[\gamma_j,\gamma_{j+1})$ and $k>j\geqq m$, then $S(t-\gamma_k)=0$ since $t-\gamma_k<0$. Hence, $\|\nu-\nu_{m-1}\|_\alpha$ equals

$$\sum_{j=m}^{\infty}\int_{\gamma_j}^{\gamma_{j+1}}e^{-\alpha t}\left|\sum_{k=m}^{j}a_{\gamma_k}S(t-\gamma_k)\right|\,dt,$$

and part a) is complete.

b) Since $S\in L^\infty(\mathbb{R})$, it is clear that (4.4) implies (4.3).

c) Because of Theorem 2.1c, (4.4) follows from (4.6).

Since $C>0$, it is clear that (4.6) implies (4.5). The converse follows since

$$\sum_{k=m}^{j}C^k\leqq K\int_m^j C^x\,dx=K'(C^j-C^m)\leqq KC^j.$$

Thus, (4.5) is equivalent to (4.6). Similar observations provide the equivalence of (4.6) and (4.7). □

Another aspect of causal approximation to the ν-system deals with the condition,

$$\forall\varphi\in\mathcal{S},\quad\exists\lim_{n\to\infty}\langle e_\beta\nu_n,4\varphi\rangle. \tag{4.8}$$

If (4.8) is valid, then $e_\beta\nu\in\mathcal{S}'$ since $\mathcal{S}$ is barrelled (e.g., Horvath [15, p. 216]), and since $\langle e_\beta(\nu-\nu_n),\varphi\rangle\to 0$ on C_c^∞, a dense subspace of $\mathcal{S}$. Thus, we obtain the closure, $V_\alpha=L^1(\mathbb{R})$ for $\alpha>\beta$, because of Theorem 3.1c. In this regard we have:

THEOREM 4.2. *Given a signal S and $\alpha>0$, if there are constants $p\geqq 1$ and $\beta<\alpha$ such that*

$$\sup_n\sum_{0\leqq j<n}|a_{\gamma_j}|\,e^{-\beta\gamma_j}(\gamma_n-\gamma_j)^p<\infty, \tag{4.9}$$

then $e_\beta\nu\in\mathcal{S}'$, and hence $V_\alpha=L^1(\mathbb{R})$.

Proof. The basis of this verification is Schwartz's characterization of tempered distributions (e.g., Schwartz [20, pp. 239–240]): $T\in\mathcal{S}'$ if and only if there is $p\geqq 0$, $F\in C_b(\mathbb{R})$, and a positive polynomial P such that $T=(FP)^{(p)}$. Since ν only has a discrete part and $e_\beta\nu=\sum a_\gamma e^{-\beta\gamma}\delta_\gamma$, the continuity of F in the Schwartz criterion allows us to take $p\geqq 1$ in (4.9).

For each $p\geqq 1$, we define

$$F(t)=\begin{cases}0 & \text{if } t\leqq 0,\\ \dfrac{1}{p!}\displaystyle\sum_{\gamma<t}a_\gamma e^{-\beta\gamma}(t-\gamma)^p & \text{if } t>0\end{cases}$$

and $P(t)\equiv 1$. F is a finite sum for each t and, hence, F is a continuous function. It is also clear that $(FP)^{(p+1)}=\sum a_\gamma e^{-\beta\gamma}\delta_\gamma$.

The proof will be complete once we show that F is bounded. Take any $t>0$. Then $t\in(\gamma_{n-1},\gamma_n)$ for some $n\geqq 1$ and so $t-\gamma_j\leqq\gamma_n-\gamma_j$ for all $j\leqq n-1$. Therefore, the boundedness follows from (4.9). □

Remark 4.1. (4.9) can certainly be weakened as a sufficient condition that $e_\beta\nu$ be tempered. On the other hand, the requirement $e_\beta\nu\in\mathcal{S}'$ is a very strong condition for closure and cannot generally be expected. In this regard, note that if $e_\beta\nu\in\mathcal{S}'$,

then $e_\beta \nu' \in \mathcal{S}'$ since $(e_\beta \nu)' \in \mathcal{S}'$ and $e_\beta \nu' = (e_\beta \nu)' + \beta e_\beta \nu$. The converse is not obvious, but we do have:

THEOREM 4.3. *Given a signal S and $\alpha > 0$, if $e_\beta \nu' \in \mathcal{S}'$ for some $\beta < \alpha$, then $V_\alpha = L^1(\mathbb{R})$.*

Proof. Let $\chi_n = \chi_{[0,n]}$, then $\|e_\alpha(H - \chi_n)\|_{L^1(\mathbb{R})} \to 0$. Define $T_n = S^{-1} * \chi_n$. Then $\|T_n\|_\alpha < \infty$ and supp $T_n \subset [0, \infty)$. In fact, we have

$$T_n = \sum a_\gamma(\delta_\gamma - \delta_{\gamma+n}),$$

since $\delta'_\gamma * \chi_n = \delta_\gamma - \delta_{\gamma+n}$. Also, note that $e_\beta T_n = (e_\beta S^{-1}) * (e_\beta \chi_n)$. This follows because

$$\begin{aligned}\langle e_\beta T_n, \varphi\rangle &= \langle S^{-1} * \chi_n, e_\beta \varphi\rangle = \langle S_u^{-1}, \langle \chi_n(t), e_\beta(t+u)\varphi(t+u)\rangle\rangle \\ &= \langle e_\beta(u) S_u^{-1}, \langle e_\beta(t)\chi_n(t), \varphi(t+u)\rangle\rangle = \langle (e_\beta S^{-1}) * (e_\beta \chi_n), \varphi\rangle.\end{aligned}$$

Thus, since $e_\beta S^{-1} = e_\beta \nu' \in \mathcal{S}'$ and $e_\beta \chi_n$ has compact support, we see that $e_\beta T_n \in \mathcal{S}'$. The result follows by Theorem 3.1. □

We now investigate the relationship between α-convergence, $\|\mu_n - \nu\|_\alpha \to 0$, and weak $*$ convergence, $\mu_n \to T$ in $\sigma(D_1, C_c^1)$. The following result is included in this section since the boundedness condition (4.10) of part a) is related to the criteria in Theorem 4.1 and since the support condition of part b) is causal.

THEOREM 4.4. *Given a signal S and $\alpha > 0$, assume that $\lim \|\mu_n - \nu\|_\alpha = 0$ for some sequence $\{\mu_n\} \subseteq M_f$.*

a) *If*

$$\forall a < b, \quad \exists K_{a,b} \quad \text{such that} \quad \sup_{y<b}\left(e^{\alpha y} \sum_{a-y<\gamma<b-y} |a_\gamma|\right) \leqq K_{a,b} \tag{4.10}$$

then $\lim \mu_n = \nu$ *in* $\sigma(D_1, C_c^1)$.

b) *If*

$$\exists a \quad \text{such that} \quad \forall n, \ \text{supp}\ \mu_n \subseteq [a, \infty), \tag{4.11}$$

and

$$\exists \gamma > \alpha \quad \text{such that} \quad \sup_n \int e^{-\gamma t} |S * (\mu_n - \nu)(t)|^{\gamma/\alpha}\, dt < \infty, \tag{4.12}$$

then $\lim \mu_n = \nu$ *in* $\sigma(D_1, C_c^1)$.

Proof. a) Take $\varphi \in C_c^1$ with supp $\varphi \subseteq [a, b]$. We calculate

$$\begin{aligned}|\langle \mu_n - \nu, \varphi\rangle| &= |\langle ((\mu_n - \nu) * S) * S^{-1}, \varphi\rangle| \\ &= |\langle (\mu_n - \nu) * S(y), \langle S^{-1}(x), \varphi(x+y)\rangle\rangle| \\ &= |\langle e^{-\alpha y}(\mu_n - \nu) * S(y), e^{\alpha y}\langle S^{-1}(x), \varphi(x+y)\rangle\rangle| \\ &\leqq \|\mu_n - \nu\|_\alpha \sup_{y \in \mathbb{R}} |e^{\alpha y} \sum_{\gamma \in \Gamma} a_\gamma \varphi'(\gamma + y)|,\end{aligned} \tag{4.13}$$

recalling that $S^{-1} = \sum a_\gamma \delta'_\gamma$. If $y \geqq b$, then $\gamma + y \geqq b$ for all $\gamma \in \Gamma$. Hence, the supremum in (4.13) is really taken over the interval $y \in (-\infty, b)$. For any such y, $\varphi'(\gamma + y) = 0$, if $\gamma + y \geqq b$ or $\gamma + y \leqq a$, i.e., if $\gamma \geqq b - y$ or $\gamma \leqq a - y$.

Consequently, the supremum in (4.13) is dominated by the supremum in (4.10), and so the result is proved since $\|\mu_n - \nu\|_\alpha \to 0$.

b) Because of the convergence $\|\mu_n - \nu\|_\alpha \to 0$, there is a subsequence of $\{\mu_n\}$ (which we call $\{\mu_n\}$) such that

$$\lim_n g_n(t) = \lim_n e^{-\alpha t} S * (\mu_n - \nu)(t) = 0 \quad \text{a.e.} \tag{4.14}$$

By condition (4.12) there is $p = \gamma/\alpha > 1$ such that

$$\|g_n\|_{L^p(\mathbb{R})} = O(1), \qquad n \to \infty.$$

By a standard convergence theorem, e.g., Benedetto [4, Chapter 6], this norm boundedness coupled with (4.14) yields

$$\forall \varphi \in L^q(\mathbb{R}), \quad \int_{\mathbb{R}} g_n \varphi \to 0, \tag{4.15}$$

where $1/p + 1/q = 1$. Take $\varphi \in C_c^0$ and observe that $S * (\mu_n - \nu)$ is an element of $L^\infty(\mathbb{R})$ supported on a right half-line. We compute

$$\left| \int \varphi (S * (\mu_n - \nu)) \right| \leqq \left(\sup_{t \in \mathbb{R}} |e^{\alpha t} \varphi(t)| \right) \|\mu_n - \nu\|_\alpha,$$

and conclude that $\lim S * (\mu_n - \nu) = 0$ in $\sigma(D_0, C_c^0)$. As in (4.13), we obtain

$$\langle \mu_n - \nu, \varphi \rangle = \langle S * (\mu_n - \nu)(t), \sum_\gamma a_\gamma \varphi'(\gamma + t) \rangle \quad \text{for } \varphi \in C_c^1. \tag{4.16}$$

Note that, for each t, $\sum a_\gamma \varphi'(\gamma + t)$ is a finite sum and so $\sum a_\gamma \varphi'(\gamma + t)$ is a continuous function. Also, if supp $\varphi \subseteq [c, d]$, then $\sum a_\gamma \varphi'(\gamma + t) = 0$ for all $t \geqq d$. We now let $\psi \in C_c^0$ equal 1 on a neighborhood of $[b, d]$ where $b = \min\{0, a, c\}$ and so the right-hand side of (4.16) is

$$\langle S * (\mu_n - \nu)(t), \psi(t) \sum_\gamma a_\gamma \varphi'(\gamma + t) \rangle, \tag{4.17}$$

where $\psi(\cdot) \sum a_\gamma \varphi'(\gamma + \cdot) \in C_c^0$. We here use the hypothesis that supp $\mu_n \subseteq [a, \infty)$ independently of n. Combining (4.16), (4.17) and the fact that $S * (\mu_n - \nu) \to 0$ in $\sigma(D_0, C_c^0)$, we see that

$$\forall \varphi \in C_c^1, \quad \lim_n \langle \mu_n - \nu, \varphi \rangle = 0. \qquad \square$$

Example 4.1. Theorem 4.4 is a partial answer to the general question: when does $\|\mu_n - \nu\|_\alpha \to 0$ imply that $\mu_n \to \nu$ in a weak * topology or, more to the point, in what ways can this implication fail? In the case that $\|\mu_n - \nu\|_\alpha \to 0$ and $\mu_n \to T \in D$ in some $\sigma(D_j, C_c^j)$ topology we can conclude that $T = \nu$ when the following line of reasoning is valid. First, for $\varphi \in C_c^j$, we let $f_{\varphi,n}(y) = \langle \mu_n - T, \varphi(x + y) \rangle$ and so $S * (\mu_n - T) \to 0$ in a weak * topology if T and S are convolvable and $\lim \int S f_{\varphi,n} = 0$. Secondly, since $\|\mu_n - \nu\|_\alpha \to 0$, we have $S * (\mu_n - \nu) \to 0$ a.e. for a subsequence. Thus, in conjunction with Helly-type or equicontinuity criteria in the first case or Vitali weak compactness criteria in the second (of which (4.12) is a special example), we can conclude that $S * \nu = S * T$; and hence $\nu = T$ because S^{-1} exists. An essential ingredient of this approach is that $T * S$ exists. This occurs if the condition

$$\forall \varphi \in C_c^\infty, \quad T * \varphi|_{(-\infty,0)} \in L^1(-\infty, 0) \tag{4.18}$$

is satisfied. In fact, by basic convolution (of distributions) criteria, $T * S$ exists if and only if $(T * \varphi)(\check{S} * \psi) \in L^1(\mathbb{R})$ for all $\varphi, \psi \in C_c^\infty$ (e.g., Dierolf and Voigt [7]); and, for all $\psi \in C_c^\infty$, $\check{S} * \psi \in L^\infty(\mathbb{R})$ is supported in a left half-line. Note that (4.18) is satisfied if supp T is contained in a right half-line.

Remark 4.2. a) If Γ has the property that $\lim(\gamma_{j+1}-\gamma_j)=0$ and if S is a signal on Γ with the property that a_{γ_j} has order of magnitude $(1+\|S\|_\infty)^j$ infinitely often (this possibility is not precluded from Theorem 2.1) then (4.10) does not hold.

b) With regard to (4.10) and the criteria of Theorem 4.1 it is interesting to determine to what extent (4.1) characterizes the closure $V_\gamma^+ = L^1(\mathbb{R}^+)$, $\gamma \geqq \alpha$.

The situation dealing with H instead of ν is much simpler. We consider the limit,

$$\lim_{n\to\infty} S * \mu_n = H. \tag{4.19}$$

Proposition 4.1. *Given a signal* S, $\alpha>0$, *and* $\{\mu_n\}\subseteq M_f$, *assume* $\{S*\mu_n\}\subseteq L^\infty(\mathbb{R})$ *is uniformly bounded.*

a) *If* $\lim_{n\to\infty}\|\mu_n-\nu\|_\alpha=0$, *then* (4.19) *is valid pointwise a.e. and in the weak* $*$ *topology* $\sigma(L^\infty, L^1)$ *for some subsequence of* $\{\mu_n\}$.

b) *If* (4.19) *is valid pointwise a.e.* (*and hence in the weak* $*$ *topology* $\sigma(L^\infty, L^1)$ *by hypothesis*) *and* (4.11) *is also assumed, then* $\lim_{n\to\infty}\|\mu_n-\nu\|_\alpha=0$.

Proof. b) is clear by the dominated convergence theorem. For a) The hypothesis of α-convergence yields convergence in measure of $\{S*\mu_n\}$ to H, and, consequently, (4.19) is valid for a subsequence almost everywhere. Weak $*$ convergence follows since $\{S*\mu_n\}$ is uniformly bounded. □

5. Approximation to the ν-system from the distant past. In light of Theorem 3.1 and the fact that $\operatorname{supp}\nu=\Gamma\subseteq[0,\infty)$, the following is a natural question. Are there semigroups Γ, signals S, on Γ and $\alpha>0$ such that $V_\alpha=L^1(\mathbb{R})$ and every sequence $\{T_n\}\subseteq Y_\alpha$, for which $\lim\|T_n-\nu\|_\alpha=0$, has the property that "$\operatorname{supp} T_n\to-\infty$", i.e., for every $a\in\mathbb{R}$, there is $n=n_a$ such that $\{t\in\mathbb{R}: t\in\operatorname{supp}\mu_n\cap(-\infty,a)\}\neq\varnothing$? Proposition 5.1 answers this question in the affirmative. The rest of the section is devoted to a study of why and how recourse to the distant past is required to approximate the ν-system and, thus, ensure closure.

PROPOSITION 5.1. *Let* $\Gamma=\{\gamma_j=\log(j+1): j=0,1,\cdots\}$ *and let* S *be the* 0–1 *signal on* Γ. *Then*:

a) *There are* $\alpha_1<\alpha_2<\alpha_3$ *such that* $V_{\alpha_2}\subsetneqq L^1(\mathbb{R})$ *and* $V_{\alpha_1}=V_{\alpha_3}=L^1(\mathbb{R})$.

b) *There is* $\alpha\in(0,\frac{1}{2})$ *such that* $V_\alpha=L^1(\mathbb{R})$ *and* $\operatorname{supp} T_n\to-\infty$ (*in the sense defined above*) *whenever* $\{T_n\}\subseteq Y_\alpha$ *and* $\lim_{n\to\infty}\|T_n-\nu\|_\alpha=0$.

Proof. a) Let $\alpha_2=\frac{1}{2}$ and note that, by the analyticity of the Riemann zeta function, $\zeta(s)$, there are $\alpha_1\in(0,\frac{1}{2})$ and $\alpha_3\in(\frac{1}{2},1)$ for which $\zeta(\alpha_1+i\tau)$ and $\zeta(\alpha_3+i\tau)$ never vanish.

Because of the functional equation for $\zeta(s)$ and Theorems 4.3 and 4.4 of Benedetto [5, § 4], we obtain our result.

b) Using the analyticity of $\zeta(s)$ again, we choose $\alpha\in(0,\frac{1}{2})$, for which $\zeta(\alpha+i\tau)$ never vanishes. Consequently, by Theorem 4.3 of Benedetto [5, § 4], we have the closure $V_\alpha=L^1(\mathbb{R})$. By Theorem 3.1, we choose $\{T_n\}\subseteq Y_\alpha$, for which $\lim_{n\to\infty}\|T_n-\nu\|_\alpha=0$. Suppose there is $a\leqq 0$ such that, for all n, $\operatorname{supp} T_n\subseteq[a,\infty)$.

Then, taking any $\gamma\geqq\alpha$, we have

$$\|T_n-\nu\|_\gamma=\int_{-\infty}^{\infty} e^{-\alpha t}e^{-(\gamma-\alpha)t}|S*(T_n-\nu)(t)|\,dt\leqq e^{-(\gamma-\alpha)a}\|T_n-\nu\|_\alpha,$$

and so $\lim_{n\to\infty}\|T_n-\nu\|_\gamma=0$. In particular, $\lim_{n\to\infty}\|T_n-\nu\|_{1/2}=0$ and so $V_{1/2}=L^1(\mathbb{R})$, a contradiction since $\zeta((\frac{1}{2})+i\tau)=0$ for some τ, i.e., because of Theorem 4.4 of [5, § 4]. □

Remark 5.1. a) If Γ and S are as in Proposition 5.1, then, by Theorem 3.1a and the analyticity of $\zeta(s)$, $V_\alpha=L^1(\mathbb{R})$ for all but at most countably many $\alpha\in(0,1)$. Also,

Proposition 5.1b does not a priori preclude the possibility that $V_\alpha^+ = L^1(\mathbb{R}^+)$ for some $\alpha \in (\frac{1}{2}, 1)$, and thus, for all $\gamma \geqq \alpha$. On the other hand, we do have

$$\hat{S}_\alpha(\gamma) = \frac{\chi(s)((1-\alpha)+i\gamma)}{s} \quad \frac{e^{-s\log 2} - e^{-\log 2}}{e^{-((1-\alpha)+i\gamma)\log 2} - e^{-\log 2}} \hat{S}_{1-\alpha}(\gamma),$$

where $s = \alpha + i\gamma$ and $\chi(s) = \pi^{s-(1/2)}\Gamma(\frac{1}{2} - \frac{1}{2}s)/\Gamma(s/2)$, and so $\hat{S}_\alpha(\gamma) = 0$ if and only if $\hat{S}_{1-\alpha}(\gamma) = 0$, $\alpha \in (\frac{1}{2}, 1)$. Consequently, we expect that $V_\alpha^+ \subsetneqq L^1(\mathbb{R})$ for all $\alpha \in (0, 1)$.

b) In light of Proposition 5.1b, it is important to approximate ν in certain natural ways by means of sequences $\{T_n\} \subseteq Y_\alpha$ for which supp $T_n \to -\infty$. One approach, which is essentially causal and provides no insight into the phenomenon of Proposition 5.1, is the following. Let $T_n = \nu_n + \mu_n$, where $\mu_n \in M_f(-\infty, 0)$, supp $\mu_n \to -\infty$, supp $\mu_n \subseteq [a_n, \infty)$ and

$$\|\mu_n\|_\alpha \leqq \|S\|_\infty \|\mu_n\|_1 \int_{a_n}^{\infty} e^{-\alpha t}\, dt \to 0.$$

More effective approximations than that given in Remark 5.1b involve approximate identities. We modify the usual definition of approximate identity for our situation.

DEFINITION 5.1. A sequence $\{\rho_n\}$ of integrable functions on $\mathbb{R}$ is an *approximate identity* if $\|\rho_n\|_{L^1(\mathbb{R})} = O(1)$, $n \to \infty$, each ρ_n satisfies the conditions that $\int \rho_n = 1$, that supp ρ_n is compact and that for all $\varepsilon > 0$ there is $\rho_\varepsilon \in L^1(\mathbb{R})$ such that

$$\forall n, \quad |\rho_n| \leqq \rho_\varepsilon \quad \text{a.e. on } \mathbb{R}\backslash[-\varepsilon, \varepsilon]$$

and

$$\lim \rho_n = 0 \quad \text{a.e. on } \mathbb{R}\backslash[-\varepsilon, \varepsilon].$$

THEOREM 5.1. *Given a signal S on a semigroup Γ, let $\{\rho_n\}$ be an approximate identity. Then*:

a) $\{S * (\nu * \rho_n)\} \subseteq L^\infty(\mathbb{R})$ *is uniformly bounded and pointwise a.e. convergent to* $H(x)$. *In particular,* $\lim S * (\nu * \rho_n) = H$ *in the weak* $*$ *topology* $\sigma(L^\infty, L^1)$.

b) $\nu * \rho_n(x) = \sum_{\gamma \in R(x,n)} a_\gamma \rho_n(x - \gamma)$, *where*

$$R(x, n) = \{\gamma \in [x - r_n, x + r_n] \cap \Gamma\} \quad \text{if supp } \rho_n \subseteq [-r_n, r_n],$$

$$R(x, n) = \{\gamma \in [x, x + r_n] \cap \Gamma\} \text{ if supp } \rho_n \subseteq [-r_n, 0],$$

$$R(x, n) = \{\gamma \in [x - r_n, x] \cap \Gamma\} \quad \text{if supp } \rho_n \subseteq [0, r_n].$$

c) *Given $\alpha > 0$, if for each n there is $\beta_n \in (0, \alpha)$ such that*

$$\nu * \rho_n(t) = O(e^{\beta_n t}), \qquad t \to \infty, \tag{5.1}$$

then $\|\nu * \rho_n\|_\alpha < \infty$ *and* $e^{-\eta_n t}(\nu * \rho_n)(t) \in L^\infty \subseteq \mathcal{S}'$ *for some* $\eta_n \in (\beta_n, \alpha)$. *In particular,* $\nu * \rho_n \in Y_\alpha$.

d) *If for each $\varepsilon > 0$ there is $g_\varepsilon \in L^1(-\infty, -\varepsilon)$ such that*

$$\forall n, \quad \left| \int_{-\infty}^{t} \rho_n \right| \leqq e^{\alpha t} g_\varepsilon(t) \quad \text{a.e. on } (-\infty, -\varepsilon), \tag{5.2}$$

then

$$\lim_{n \to \infty} \|\nu * \rho_n - \nu\|_\alpha = 0. \tag{5.3}$$

Also, (5.2) *is satisfied if for each* $\varepsilon > 0$ *there are* $\gamma_\varepsilon > \alpha$ *and* $K_\varepsilon > 0$ *such that*

$$\forall n, \quad \left| \int_{-\infty}^{t} \rho_n \right| \leqq K_\varepsilon e^{\gamma_\varepsilon t} \quad \text{a.e. on } (-\infty, -\varepsilon).$$

Proof. a) Since each of the factors is convolvable with the other, we have that $S * (\nu * \rho_n) = H * \rho_n$. Also,

$$|S * (\nu * \rho_n)(t)| = \left| \int \rho_n(x) H(t-x)\, dx \right| \leqq \|\rho_n\|_{L^1(\mathbb{R})},$$

and so $\{S * (\nu * \rho_n)\}$ is uniformly bounded. Next, $|H(t) - S * (\nu * \rho_n)(t)| \leqq |\int \rho_n(x)(H(t) - H(t-x)|\, dx|$, and for a fixed t, $|t| > 0$, we have $H(t-x) = H(t)$ if $|x| < |t|$, and so

$$\forall n, \quad \int_{-|t|}^{|t|} \rho_n(x)(H(t) - H(t-x))\, dx = 0. \tag{5.4}$$

For the range $\mathbb{R} \backslash [-|t|, |t|]$, we see that

$$\int_{|x|>|t|} |\rho_n(x)|\, |H(t) - H(t-x)|\, dx \leqq 2 \int_{|x|>|t|} |\rho_n(x)|\, dx, \tag{5.5}$$

and the right-hand side goes to 0 for large n because of the dominated convergence theorem and the fact that $\{\rho_n\}$ is an approximate identity. Thus, (5.4) and (5.5) yield the pointwise convergence of $S * (\nu * \rho_n)$ to H, and part a) is complete.

Part b) is clear. For example, if x is fixed and supp $\rho_n \subseteq [-r_n, 0]$ then $x - \gamma \in [-r_n, 0]$ if and only if $-\gamma \in [-r_n - x, -x]$, which is the same as $\gamma \in [x, x + r_n]$.

c) Suppose that supp $\rho_n \subseteq [-r_n, \infty)$, $r_n \geqq 0$. Then we compute

$$\begin{aligned} \|\nu * \rho_n\|_\alpha &= \int_{-r_n}^{\infty} e^{-\alpha t} \left| \int_0^\infty S(u) \nu * \rho_n(t-u)\, du \right| dt \\ &\leqq K_n \int_{-r_n}^{\infty} e^{-\alpha t} \int_0^\infty e^{\beta_n (t-u)}\, du\, dt < \infty, \end{aligned} \tag{5.6}$$

noting that supp $\nu * \rho_n \subseteq [-r_n, \infty)$ so that the first inequality in (5.6) follows from (5.1).

Now, for any $\eta_n \in (\beta_n, \alpha)$, $e^{-\eta_n t} \nu * \rho_n(t)$ vanishes on $(-\infty, -r_n)$ and is bounded by $K e^{-(\eta_n - \beta_n)t}$ on $[-r_n, \infty)$. Thus, $e^{-\eta_n t} \nu * \rho_n(t) \in L^\infty(\mathbb{R}) \subseteq \mathcal{S}'(\mathbb{R})$, and part c) is complete.

d) The second part of d) is clear by setting $g_\varepsilon(t) = K_\varepsilon e^{(\gamma_\varepsilon - \alpha)t}$. To verify (5.3) given (5.2), we first compute

$$\|\nu - \nu * \rho_n\|_\alpha = \int_{-\infty}^{0} e^{-\alpha t} |H * \rho_n(t)|\, dt + \int_0^\infty e^{-\alpha t} |H(t) - H * \rho_n(t)|\, dt, \tag{5.7}$$

noting that the second integral in (5.7) tends to 0 from part a.

The first integral of (5.7) is

$$\int_{-\infty}^{0} e^{-\alpha t} \left| \int_0^\infty \rho_n(t-x)\, dx \right| dt = \int_{-\infty}^{0} e^{-\alpha t} \left| \int_{-\infty}^{t} \rho_n(u)\, du \right| dt.$$

Take any $\varepsilon > 0$. We first show that

$$\lim_{n \to \infty} \int_{-\infty}^{-\varepsilon} e^{-\alpha t} \left| \int_{-\infty}^{t} \rho_n(u)\, du \right| dt = 0. \tag{5.8}$$

In fact, if $t \in (-\infty, -\varepsilon)$, then $\lim_{n\to\infty} \int_{-\infty}^{t} \rho_n = 0$ since $\lim_{n\to\infty} \rho_n = 0$ a.e. and $|\rho_n| \leqq \rho_\varepsilon$.

Then using (5.2) and the dominated convergence theorem again we obtain (5.8).

We now observe that there is C such that

$$\forall n, \quad \int_{-\varepsilon}^{0} e^{-\alpha t}\left|\int_{-\infty}^{t} \rho_n(u)\,du\right| dt \leqq C\varepsilon. \tag{5.9}$$

This is clear since $\|\rho_n\|_{L^1(\mathbb{R})} = O(1)$, $n \to \infty$. Combining (5.8) and (5.9), we obtain that

$$\overline{\lim_{n\to\infty}} \int_{-\infty}^{0} e^{-\alpha t}\left|\int_{0}^{\infty} \rho_n(t-x)\,dx\right| dt \leqq C\varepsilon,$$

and so the result follows from (5.7) and the fact that ε is arbitrary. □

THEOREM 5.2. *Given a signal S and $\alpha > 0$, and let $\{\rho_n\}$ be an approximate identity.*

a) If

$$\forall n, \quad \exists \beta_n \in (0, \alpha) \quad \textit{such that } \nu * \rho_n(t) = O(e^{\beta_n t}), \quad t \to \infty, \tag{5.10}$$

and

$$\forall \varepsilon > 0, \quad \exists \gamma_\varepsilon > \alpha \textit{ and } \exists K_\varepsilon \quad \textit{such that } \forall n\ |\rho_n(t)| \leqq K_\varepsilon e^{\gamma_\varepsilon t} \textit{ a.e. on } (-\infty, -\varepsilon), \tag{5.11}$$

then $V_\alpha = L^1(\mathbb{R})$.

b) *If* (5.10) *and*

$$\sup_{t\in[-r_n,0]} \left|\int_{-r_n}^{t} \rho_n(u)\,du\right| = o(e^{-\alpha r_n}), \qquad n \to \infty, \tag{5.12}$$

where $\operatorname{supp} \rho_n \subseteq [-r_n, \infty)$ *and* $r_n \geqq 0$, *then* $V_\alpha = L^1(\mathbb{R})$.

Proof. a) (5.10) and Theorem 5.1c yield the fact that $\nu * \rho_n \in Y_\alpha$ for each n. Given $\varepsilon > 0$, (5.11) yields

$$\left|\int_{-\infty}^{t} \rho_n\right| \leqq K_\varepsilon \int_{-\infty}^{t} e^{\gamma_\varepsilon u}\,du = C_\varepsilon e^{\gamma_\varepsilon t}$$

for each n and each $t \in (-\infty, -\varepsilon)$, and this gives (5.2), as in the second part of Theorem 5.1d.

From (5.2) and Theorem 5.1d we obtain (5.3), and this gives the closure, $V_\alpha = L^1(\mathbb{R})$, because of Theorem 3.1 and the fact that $\nu * \rho_n \in Y_\alpha$ for each n.

b) Because of Theorem 5.1 and the role of (5.10) in part a) it is sufficient to prove that

$$\lim_{n\to\infty} \int_{-r_n}^{0} e^{-\alpha t}|H * \rho_n(t)|\,dt = 0. \tag{5.13}$$

The integral in (5.13) is bounded by

$$e^{\alpha r_n} \sup_{t\in[-r_n,0]} \left|\int_{0}^{\infty} \rho_n(t-x)\,dx\right| = e^{\alpha r_n} \sup_{t\in[-r_n,0]} \left|\int_{-r_n}^{t} \rho_n(u)\,du\right|,$$

and so the result follows from (5.12). □

(5.12) can be replaced by a weaker integral condition.

Remark 5.2. Let S be a signal on Γ. If the consecutive elements of Γ are well spaced, then verification of the closure $V_\alpha = L^1(\mathbb{R})$ can usually be made by means of the theory developed in § 4. On the other hand, if $\lim(\gamma_{j+1} - \gamma_j) = 0$, then there is good reason to expect the condition, $\operatorname{supp} T_n \to -\infty$, to occur in the situation that $T_n \in Y_\alpha$, $V_\alpha = L^1(\mathbb{R})$ and $\lim \|T_n - \nu\|_\alpha = 0$. In fact, for the case $T_n = \nu * \rho_n$, if $\operatorname{supp} \rho_n \to -\infty$, then $\nu * \rho_n$ involves long sums of consecutive coefficients, a_γ, e.g., Example 5.1,

and so there is greater opportunity for cancellation among the a_γ, thus allowing for (5.10) to be satisfied. Further, since a_{γ_n} could possibly be as large as $(1+\|S\|_\infty)^n$, e.g., Theorem 2.1, then, for nonnegative approximate identities ρ (with $\int \rho = 1$) (5.10) forces height restrictions on $\{|\rho(\gamma_n)|\}$, and, as a consequence, ρ must extend farther and farther to $-\infty$ in order that (5.11) also be satisfied.

Example 5.1. Let S be a signal on Γ and let $\alpha > 0$. For each $n \geqq 1$, let ρ_n be defined as

$$\rho_n(t) = \begin{cases} c_n & \text{if } x \in (-C_n, 0], \\ b_n & \text{if } x \in (-(C_n + B_n), -C_n], \\ 0 & \text{elsewhere,} \end{cases}$$

where b_n, c_n, B_n, $C_n > 0$. Assume $\lim b_n = \lim C_n = 0$, $\lim B_n = \lim c_n = \infty$ and $b_n B_n + c_n C_n = 1$. Thus, $\{\rho_n\}$ is a positive approximate identity. By Theorem 5.2 we obtain the closure $V_\alpha = L^1(\mathbb{R})$ if b_n, c_n, B_n, C_n can be chosen so that for all n there are $\beta_n < \alpha$ and K_n such that

$$\forall t \geqq 0, \quad \left| c_n \sum_{t \leqq \gamma \leqq t + C_n} a_\gamma + b_n \sum_{t + C_n < \gamma \leqq t + (B_n + C_n)} a_\gamma \right| \leqq K_n e^{\beta_n t},$$

and there are $\gamma > \alpha$ and K such that

$$\forall n, \quad |b_n| \leqq K e^{-\gamma(B_n + C_n)}.$$

Example 5.2. We now outline another procedure with which to minimize $\|T - \nu\|_\alpha$, $T \in Y_\alpha$, without using approximate identities.

a) Let S be a signal and define

$$S_{(n)} = \sum_{j=0}^{n} S(\gamma_j) \chi_{[\gamma_j, \gamma_{j+1})}.$$

Thus,

$$\hat{S}_{(n)}(\lambda) = \left(\frac{i}{\lambda}\right) \sum_{0}^{n} S(\gamma_j)(e^{-i\lambda\gamma_{j+1}} - e^{-i\lambda\gamma_j}),$$

and $\hat{S}_{(n)} \in A(\mathbb{R}) \cap L^2(\mathbb{R})$. Clearly, $\lim S_{(n)} = S$ in the weak $*$ topology $\sigma(L^\infty, L^1)$ and hence also in the topology of $\mathscr{S}'$. For the procedure below it is useful to deal with $\hat{H}$. If $h(t) = H(t) - H(-t)$, then $\hat{h}(\lambda) = pv(1/\lambda)$, and so $\hat{H}(\lambda) = \frac{1}{2}\delta + pv(1/\lambda)$ since $H = \frac{1}{2}(1+h)$, e.g., Horváth [16], cf., Schwartz [20, pp. 258–259].

b) We assume that $\hat{S}_{(n)}$ never vanishes and that $1/\hat{S}_{(n)} \ll P$ for some polynomial P; hence, $1/\hat{S}_{(n)} \in \mathscr{S}'$. Of course, in the case $n = 1$, $e^{-i\lambda\gamma_1} = 1$ for $\lambda = 2n\pi/\gamma_1$ and $n \in \mathbb{Z}$; therefore, $\hat{S}_{(1)}(\lambda) = 0$ for all such λ. In case $\hat{S}_{(n)}$ has zeros, and there can be at most a discrete zero set by analyticity, then $\hat{S}_{(n)}$ can be modified in the following formalism by inserting smaller and smaller "tubes" on which $\hat{S}_n$ must travel in neighborhoods of zeros.

c) Let $H_{(n)}$ have compact support and converge to H, and let $\{f_n\}$ have the properties that each supp f_n is contained in $[-R_n, R_n]$, $R_n \to \infty$, and $\{\hat{f}_n\}$ converges to δ, e.g., $f_n = \chi_{[-R_n, R_n]}$. The convergence of $\{H_{(n)}\}$ and $\{\hat{f}_n\}$ is specified according to the problem. We set

$$T_n = \left(\left(\frac{1}{\hat{S}_{(n)}} \right)^{\wedge} * H_{(n)} \right) f_n \tag{5.14}$$

and expect $\{T_n * S\}$ to converge to H in a reasonable way. In fact, T_n has compact support and $(S * T_n)^{\wedge} = \hat{S}((\hat{H}_{(n)}/\hat{S}_{(n)}) * \hat{f}_n$, which tends to $\hat{H}$. It is necessary that T_n be an element of $L^1(\mathbb{R})$ and that various continuity criteria be valid.

If $H_{(n)} \in C_c^\infty(\mathbb{R})$ then $(1/\hat{S}_{(n)})\hat{} * H_{(n)}$ is a slowly increasing function since $1/\hat{S}_{(n)} \in \mathcal{S}'$ (e.g., Schwartz [20, p. 239]) and so $T_n \in L^1 \cap L^\infty$.

d) Even with the assumptions in part b) it remains to verify the formal convergence, $T_n * S \to H$, in the context of closure problems for the $\|\cdot\|_\alpha$-topology. A very particular case of this is to ask under what circumstances $\lim T_n * S = H$ a.e. when $\lim T_n * S = H$ in the weak $*$ topology $\sigma(L^\infty, L^1)$, cf., Theorem 5.1a. Of course, in general, it need not happen that pointwise convergence is a consequence of $\sigma(L^\infty, L^1)$ convergence, e.g., Benedetto [4, Thm. 3.12].

If $\{S * T_n\}$ converges to H in the weak $*$ topology $\sigma(L^\infty, L^1)$, then $\sup_n \|S * T_n\|_\infty < \infty$. In the following result, we assume this boundedness. The hypothesis of convergence in measure is related to the discussion in part d of Example 5.2; Lebesgue proved that pointwise a.e. convergence on finite measure spaces implies convergence in measure. Finally, (5.16) and (5.17) reflect an assumption of weak compactness.

THEOREM 5.3. *Given a signal S and $\alpha > 0$, let $\{T_n\} \subseteq Y_\alpha$ be a sequence with the properties that $S * T_n \in L^\infty(\mathbb{R})$,*

$$\sup_n \|S * T_n\|_\infty = C < \infty, \tag{5.15}$$

and $\{S * T_n\}$ converges in measure to H on every compact subset of $\mathbb{R}$. Suppose

$$\lim_{r \to \infty} \int_{-\infty}^{-r} e^{-\alpha t} |S * T_n(t)|\, dt = 0 \quad \textit{uniformly in } n \tag{5.16}$$

and

$$\lim_{\theta \to 0} \int_{x+(-\theta,\theta)} e^{-\alpha t} |S * T_n(t)|\, dt = 0 \quad \textit{uniformly in } n \textit{ and } x \leqq 0. \tag{5.17}$$

Then $\lim_n \|\nu - T_n\|_\alpha = 0$ and (since $T_n \in Y_\alpha$) $V_\alpha = L^1(\mathbb{R})$.

Proof. For each n, we define $f_n(t)$ by the properties that $|f_n(t)| = e^{-\alpha t}$ and

$$f_n(t)(S * T_n(t) - H(t)) = e^{-\alpha t} |S * T_n(t) - H(t)|. \tag{5.18}$$

Take $\varepsilon > 0$. We'll find n_ε such that for all $n \geqq n_\varepsilon$

$$\left| \int f_n (S * T_n - H) \right| < \varepsilon, \tag{5.19}$$

and the result will follow by (5.18). We begin by choosing a compact set $K = K_\varepsilon \subseteq \mathbb{R}$ and $\theta = \theta_\varepsilon > 0$ such that

$$\forall n, \quad \int_{K^\sim} e^{-\alpha t} |T_n * S(t)|\, dt < \frac{\varepsilon}{4} \quad \text{and} \quad \int_{K^\sim \cap [0,\infty)} e^{-\alpha t}\, dt < \frac{\varepsilon}{4} \tag{5.20}$$

and, for all Borel sets satisfying the condition $|B| < \theta$,

$$\forall n, \quad \int_B e^{-\alpha t} |T_n * S(t)|\, dt < \frac{\varepsilon}{8} \quad \text{and} \quad \int_{B \cap [0,\infty)} e^{-\alpha t}\, dt < \frac{\varepsilon}{8}. \tag{5.21}$$

($K^\sim$ is the complement of K and $|B|$ is the Lebesgue measure of B.) (5.20) follows from (5.15) and (5.16), and (5.21) follows from (5.15) and (5.17). Using (5.20) we estimate, for each n,

$$\left| \int f_n (S * T_n - H) \right| \leqq \left| \int_K f_n (S * T_n - H) \right| + \frac{2\varepsilon}{4}. \tag{5.22}$$

Since $\{S * T_n\}$ converges in measure to H on K, there are $K_0 \subseteq K$ and n_0 such that

$$\forall n \geqq n_0, \quad \forall t \in K_0, \quad |S * T_n(t) - H(t)| < \varepsilon / \left(4 \int_K e^{-\alpha t}\, dt\right) \tag{5.23}$$

and $|K \cap \tilde{K_0}| < \theta$ (e.g., Benedetto [4, Exercise 3.19]). Consequently, using (5.21) we have

$$\begin{aligned}
&\left|\int_K f_n(T_n * S - H)\right| \\
&\quad \leqq \left|\int_{K_0}\right| + \left|\int_{K \cap \tilde{K_0}}\right| \\
&\quad \leqq \int_{K_0} e^{-\alpha t}|S * T_n(t) - H(t)|\, dt + \int_{K \cap \tilde{K_0}} e^{-\alpha t}|S * T_n(t)|\, dt + \int_{K \cap \tilde{K_0} \cap [0,\infty)} e^{-\alpha t}\, dt \\
&\quad < \int_{K_0} e^{-\alpha t}|S * T_n(t) - H(t)|\, dt + \frac{\varepsilon}{4}.
\end{aligned}$$

From the uniform convergence in (5.23) we see that

$$\int_{K_0} e^{-\alpha t}|S * T_n(t) - H(t)|\, dt < \frac{\varepsilon}{4}$$

for all $n \geqq n_0$. Thus, the right-hand side of (5.22) is less than $\varepsilon/2$ for all $n \geqq n_0$ and so (5.19) is obtained for $n_\varepsilon = n_0$. □

Remark 5.3. a) Given the hypothesis that $\{T_n\} \subseteq Y_\alpha$, the sufficient conditions in Theorem 5.3 are essentially necessary (e.g., Grothendieck [12, Cor. of Prop. 4 and Thm. 2]); cf. the comment after (5.24) in Remark 5.3b. The hypothesis $\{T_n\} \subseteq Y_\alpha$ of Theorem 5.3 is reflected by (5.10) in Theorem 5.2; and the weak compactness conditions, (5.16) and (5.17), of Theorem 5.3 are practically equivalent to (5.11) in Theorem 5.2. The verification of conditions (5.10) and (5.11) in Theorem 5.2 or their analogues in Theorem 5.3 depends on delicate estimates of the coefficients a_{γ_n}.

b) For this part of Remark 5.3 we discuss the weak compactness conditions (5.11) (or (5.16) and (5.17)). Generally speaking, an L^1-norm bounded set $\{F_n\} \subseteq L^1(\mathbb{R})$ is weakly precompact (that is, in the weak topology $\sigma(L^1, L^\infty)$) if and only if for all $\varepsilon > 0$ there is $K_\varepsilon \subseteq \mathbb{R}$, compact, such that

$$\forall n, \quad \int_{\tilde{K_\varepsilon}} |F_n| < \varepsilon \tag{5.16'}$$

and for all $\varepsilon > 0$ and for all $K \subseteq \mathbb{R}$, compact, there is $U \supseteq K$, open, such that

$$\forall n, \quad \int_{U \cap \tilde{K}} |F_n| < \varepsilon, \tag{5.24}$$

(e.g., Grothendieck [12, Theorem 2]). (5.16) and (5.16') are equivalent, whereas (5.24) is a weaker statement than (5.17); cf. (5.21) where (5.17) is used. The conditions for compactness are naturally more stringent. An L^1-norm bounded set $\{F_n\} \subseteq L^1(\mathbb{R})$ is precompact if and only if (5.16') and for all $\varepsilon > 0$ there is $\delta > 0$ such that

$$\forall |t| < \delta, \quad \forall n, \quad \|F_n(\cdot - t) - F_n(\cdot)\|_{L^1(\mathbb{R})} < \varepsilon \tag{5.25}$$

(e.g., Dunford and Schwartz [10, IV. 8.20 pp. 298ff.] and Grothendieck [13, pp. 287–289]). In this context we mention the result that every compact subset of a Banach space is contained in the normed closed convex hull of some null sequence.

6. Examples

Example 6.1. Let $\Gamma = c\mathbb{N} \cup \{0\} = \{\gamma_n = cn : n = 0, 1, \cdots\}$, $c > 0$.

a) Given $m < n$, we compute $j(m, n)$. $j(m, n)$ is the largest integer j for which $[\gamma_m + \gamma_j, \gamma_m + \gamma_{j+1}) \subseteq [\gamma_n, \gamma_{n+1})$, i.e., $[cm + cj, cm + cj + c) \supseteq [cn, cn + c)$, and so

$$\forall m < n, \quad \gamma_{j(m,n)} = c(n - m). \tag{6.1}$$

b) Given a signal S for which $S(0) = 1$ and a number $\alpha > (1/c) \log (1 + \|S\|_\infty)$, then $V_\gamma^+ = L^1(\mathbb{R}^+)$ for all $\gamma \geqq \alpha$. To verify this we use Theorem 4.1c. For the semigroup Γ, $\alpha b_{j,\alpha} = e^{-\alpha cj}(1 - e^{-\alpha c})$, and so condition (4.5) is equivalent to the convergence of the series $\sum \exp(-j(\alpha c - \log(1 + \|S\|_\infty)))$. Thus, we obtain the desired closure. Letting $T = \delta + \delta_1$, we see that $V_\alpha = L^1(\mathbb{R})$ for all $\alpha > 0$ by Theorem 3.1.

c) Given a signal S for which $S(0) = 1$ and a number $\alpha > (1/c) \log (1 + \|S\|_\infty)$, we can obtain the closure $V_\alpha = L^1(\mathbb{R})$ directly by means of Wiener's Tauberian Theorem and an elementary estimate. We compute

$$\begin{aligned} |\hat{S}_\alpha(\gamma)| &= |(1 - e^{-c(\alpha + i\gamma)})/(\alpha + i\gamma)| \left| \sum_0 S(cj)\, e^{-cj(\alpha + i\gamma)} \right| \\ &\geqq K \left| 1 + \sum_i S(cj)\, e^{-cj(\alpha + i\gamma)} \right| \geqq K \left(1 - \sum_i |S(cj)|\, e^{-\alpha cj} \right) \\ &\geqq K \left(1 - \|S\|_\infty \sum_1 (e^{-\alpha c})^j \right) = K \left(1 - \|S\|_\infty \frac{e^{-\alpha c}}{1 - e^{-\alpha c}} \right), \end{aligned}$$

where K depends on α, γ and c. Thus, $|\hat{S}_\alpha(\gamma)| > 0$ if $\alpha > (1/c)(1 + \|S\|_\infty)$; and so the desired closure follows the Tauberian theorem.

d) In light of the essential equivalence of parts b) and c), it is desirable to use Theorem 4.1a instead of Theorem 4.1c. Since Γ is a group, condition (4.3) of Theorem 4.1a becomes

$$\begin{aligned} &\lim_{m \to \infty} \sum_{j=m}^{\infty} e^{-\alpha cj} \left| \sum_{k=m}^{j} a_{ck} S(c(j-k)) \right| \\ &\quad = \lim_{m \to \infty} \sum_{j=m}^{\infty} e^{-\alpha cj} \left| \left(\sum_{k=0}^{j} a_{ck} \right) S(0) - \left(\sum_{k=0}^{m-1} a_{ck} \right) S(c(j-m)) \right. \\ &\qquad \left. + \sum_{n=m}^{j-1} \left(\sum_{k=0}^{n} a_{ck} \right) (S(c(j-n)) - S(c(j-(n+1)))) \right| = 0. \end{aligned} \tag{6.2}$$

Thus, estimates on $\sum_0^m a_{ck}$ and knowledge about successive values of S yield closure information beyond parts b) and c); cf. Benedetto [5, Example 5.1] for the best possible result for 0–1 signals.

e) Let $S = \chi_{[0,c)} + K\chi_{[c,\infty)}$. Hence, $\hat{S}_\alpha(\gamma) = 0$ if and only if $K = 1 - e^{(\alpha + i\gamma)c}$. In particular, for a given α and c, we have closure if K is chosen so that $|K - 1| \neq e^{\alpha c}$.

Example 6.2. Let $\Gamma = \{\gamma_j\}$, $\gamma_j = \log(j + 1)$.

a) Given $m < n$, we compute $j(m, n)$. In order that $\gamma_m + \gamma_j \leqq \gamma_n$, we have $\log((m + 1)(j + 1)) \leqq \log(n + 1)$, and hence, we set

$$j = j(m, n) = \left[\frac{n + 1}{m + 1} - 1 \right] = \left[\frac{n - m}{m + 1} \right]. \tag{6.3}$$

We need only check that $\gamma_{j+1} + \gamma_m \geqq \gamma_{n+1}$. Since j is the largest integer for which $\gamma_m + \gamma_j \leqq \gamma_n$, we must have $\gamma_m + \gamma_{j+1} > \gamma_n$, and since Γ is a semigroup, we have $\gamma_m + \gamma_{j+1} \geqq \gamma_{n+1}$.

b) Let S be a signal with jumps at Γ and take $\alpha > 0$. If

$$\lim_{m\to\infty} \sum_{j=m}^{\infty} \left(\sum_{k=m}^{j} |a_{\gamma_k}| \right) \Big/ j^{1+\alpha} = 0, \tag{6.4}$$

then $V_\gamma^+ = L^1(\mathbb{R}^+)$ for all $\gamma \geqq \alpha$. In fact,

$$\alpha \int_{\log j}^{\log(j+1)} e^{-\alpha t}\, dt \leqq \frac{1}{j^{1+\alpha}},$$

and so a direct application of Theorem 4.1b yields (6.4).

c) Let S be the 0–1 signal corresponding to Γ and fix a γ_n. We wish to find all ordered pairs (j, m), $j, m \geqq 0$, such that $\gamma_j + \gamma_m = \gamma_n$, i.e., $\log((j+1)(m+1)) = \log(n+1)$. Thus, for a fixed $n \geqq 1$, we wish to find all solutions (j, m), $j, m \geqq 0$, of the equation

$$jm + j + m = n. \tag{6.5}$$

Hence, when $m \in [0, n]$ we check to see if $(m+1) | (n-m)$. If "no", we throw out this m; if "yes", we keep the m and $((n-m)/(m+1), m)$ is a solution. Solutions of (6.5), for the first few n, are:

$n=1$,	none	$n=6$,	none
$n=2$,	none	$n=7$,	$(j, m) = (3, 1), (1, 3)$
$n=3$,	$(j, m) = (1, 1)$	$n=8$,	$(j, m) = (2, 2)$
$n=4$,	none	$n=9$,	$(j, m) = (4, 1), (1, 4)$.
$n=5$,	$(j, m) = (2, 1), (1, 2)$		

d) As in Theorem 2.2, j_n is the number of solution of (6.5). If $m+1 > n-m$ for $m \in [0, n]$, then $2m > n-1$ and so $m > (n-1)/2$. Thus $m+1 \nmid n-m$ for any $m > (n-1)/2$; hence, for all $i < j_n$, the corresponding $m = m_i$ is less than $n/2$.

Let us now estimate j_n for the case of the 0–1 signal on Γ. Fix a large integer K. Then we have, approximately, that

$$\frac{n-1}{2} \in \mathbb{N} \quad \text{for at most } \tfrac{1}{2} \text{ of } \{n : n \in [1, K]\},$$

$$\frac{n-2}{3} \in \mathbb{N} \quad \text{for at most } \tfrac{1}{3} \text{ of } \{n : n \in [1, K]\},$$

$$\frac{n-3}{4} \in \mathbb{N} \quad \text{for at most } \tfrac{1}{4} \text{ of } \{n : n \in [1, K]\}, \text{ etc.}$$

By adding all of these "solutions", we see that *there are at most* $K \log K$ *solutions of the "equation"*

$$\frac{n-m}{m+1} \in \mathbb{N}, \quad \text{where } 1 \leqq m < n \leqq K.$$

e) Let us now translate the $\mathcal{S}'$ condition (4.9) of Theorem 4.2 for the case of Γ and arbitrary S. For the sake of convenience we take $p = 1$ in (4.9). Larger p are not automatically ruled out, since it is not clear what happens for the two cases: $\gamma_n - \gamma_j =$

$\log(n+1)/(j+1)<1$ and $\gamma_n-\gamma_j=\log(n+1)/(j+1)\geqq 1$. We compute

$$\begin{aligned}
&\sum_{0\leqq j<n}|a_{\gamma_j}|\,e^{-\beta\gamma_j}(\gamma_n-\gamma_j)\\
&\quad=\sum_{j=0}^{n-1}\frac{|a_{\gamma_j}|}{(j+1)^{\beta}}\log\frac{n+1}{j+1}\\
&\quad=\left(\sum_{j=0}^{n-1}\frac{|a_{\gamma_j}|}{(j+1)^{\beta}}\right)\log\left(\frac{n+1}{n}\right)+\sum_{j=0}^{n-2}\left(\sum_{0\leqq m\leqq j}\frac{|a_{\gamma_m}|}{(m+1)^{\beta}}\right)\log\left(\frac{n+1}{j+1}-\log\frac{n+1}{j+2}\right)\\
&\quad\leqq\frac{1}{n}\sum_{j=0}^{n-1}\frac{|a_{\gamma_j}|}{(j+1)^{\beta}}+\sum_{j=0}^{n-2}\left(\sum_{0\leqq m\leqq j}\frac{|a_{\gamma_m}|}{(m+1)^{\beta}}\right)\log\left(1+\frac{1}{j+1}\right)\\
&\quad\leqq\frac{1}{n}\sum_{j=0}^{n-1}\frac{|a_{\gamma_j}|}{(j+1)^{\beta}}+\sum_{j=0}^{n-2}\frac{1}{j+1}\sum_{0\leqq m\leqq j}\frac{|a_{\gamma_m}|}{(m+1)^{\beta}}.
\end{aligned}$$

Thus, because of easy upper and lower bounds on the logarithm, (4.9) is valid if and only if

$$\sum_{j=0}^{\infty}\frac{1}{j+1}\left(\sum_{0\leqq m\leqq j}\frac{|a_{\gamma_m}|}{(m+1)^{\beta}}\right)<\infty. \tag{6.6}$$

Naturally, (6.6) should be compared with (6.4).

f) Let S be the 0–1 signal on Γ. The usefulness of our various closure criteria depends on the difficult number theoretic problem of determining if $[j/i]$ is even or odd. We illustrate this with two cases.

i) Let us apply Theorem 4.1 in a finer way than part b. Because of the integral estimate to be made in Theorem 4.1, we take t in the range, $[\gamma_j, \gamma_{j+1})$, of integration and ask in what way $S(t-\gamma_i)$, $i\leqq j$, can be evaluated. First note that

$$\left[\log\left[\frac{j+1}{i}\right],\log\left(\left[\frac{j+1}{i}\right]+1\right)\right)\supseteq\left[\log\frac{j+1}{i},\log\frac{j+2}{i}\right). \tag{6.7}$$

In fact, $(j+1)/i=[(j+1)/i]+m/i$, $0\leqq m<i$, and so $(j+2)/i=[(j+1)/i]+(m+1)/i\leqq[(j+1)/i]+1$. Now, $t\in[\gamma_j,\gamma_{j+1})$ implies $t-\log i\in[\log((j+1)/i),\log((j+2)/i)$, and, consequently, $S(t-\gamma_i)=S(\log[(j+1)/i])$ because of (6.7). Thus, Theorem 4.1 yields the closure, $V_{\gamma}^{+}=L^1(\mathbb{R}^+)$ for all $\gamma\geqq\alpha$, in case

$$\lim_{m\to\infty}\sum_{j=m}^{\infty}\left(\sum_{i=m}^{j}a_{\gamma_i}S(\log[(j+1)/i])\right)/j^{(1+\alpha)}=0. \tag{6.8}$$

Generally, it will be quite difficult to check if $[(j+1)/i]$ is even or odd (and, hence, $S(\log[(j+1)/i])$ is zero or one), but estimates such as (6.8) show the value of trying to estimate partial sums $\sum_{i=m}^{j}a_{\gamma_i}$.

ii) In order to apply Theorem 2.1b, we must compute $S(\gamma_{j(m,n)})$ where $j=j(m,n)$ is given by (6.3), and $S(\gamma_j)$ is 0 if j is odd and 1 if j is even. By the definition of $j(m,n)$, $m<n$, we see that

$$\forall m>\frac{n-1}{2},\quad S(\gamma_{j(m,n)})=1.$$

Thus, if n is odd we have from Theorem 2.1b that

$$\begin{aligned}
a_{\gamma_n}=1&-S(\gamma_{j(1,n)})-a_{\gamma_2}S(\gamma_{j(2,n)})-\cdots\\
&-a_{\gamma_{(n-1)/2}}S(\gamma_{j(n-1)/2,n)})-\sum_{m=(n+1)/2}^{n-1}a_{\gamma_m}.
\end{aligned}$$

The first few a_γ's given by the formula in Theorem 2.1b are

$$a_{\gamma_0}=1,\quad a_{\gamma_1}=1,\quad a_{\gamma_2}=-1,\quad a_{\gamma_3}=2,\quad a_{\gamma_4}=0,\quad a_{\gamma_5}=-1,\quad a_{\gamma_6}=-1.$$

g) We illustrate why the calculation in Example 6.1c does not work for $\Gamma=\{\log(j+1)\}$ but how it does provide an interval in which $\hat{S}_\alpha(\gamma)$ is nonzero. Fix $\alpha\in(0,1]$ and let S be the 0–1 signal on Γ. Then, if $s=\alpha+i\gamma$, we have

$$|s||\hat{S}_\alpha(\gamma)|\geqq 1-\sum_{j=1}^{\infty} S(\gamma_j)|1-e^{-s(\gamma_j-\gamma_{j+1})}|/(j+1)^\alpha. \tag{6.9}$$

By the mean value theorem we obtain

$$\left|1-\exp\left\{-s\log\frac{j+1}{j+2}\right\}\right|\leqq\frac{|s|}{j+1}\left|\frac{j+1}{j+2}\right|^{1-\alpha}\leqq\frac{|s|}{j}$$

and so

$$|\hat{S}_\alpha(\gamma)|\geqq\frac{1}{|s|}-\sum_{j=1}^{\infty}\frac{1}{(2j)^{1+\alpha}}\geqq\frac{1}{|s|}-\frac{1}{2^{1+\alpha}}\left(1+\frac{1}{\alpha}\right).$$

Consequently, $|\hat{S}_\alpha(\gamma)|>0$ if $|\gamma|<\alpha(2^{1+\alpha}/(1+\alpha)-1)^{1/2}$. Finer estimates yield longer "nonzero intervals."

h) Let S be the 0–1 signal on Γ. Then $\lim_{n\to\infty} S*\nu_n\neq H$ in the weak $*$ topology $\sigma(L^\infty, L^1)$. In fact, if there were weak convergence, then, since $\lim_{n\to\infty} S*\nu_n=H$ a.e., $\{S*\nu_n\}\subseteq L^\infty(\mathbb{R})$ is uniformly bounded and $\operatorname{supp} S*\nu_n\subseteq[0,\infty)$, we'd be able to conclude that $\lim_{n\to\infty}\|\nu_n-\nu\|_\alpha=0$ for each $\alpha>0$. This is a contradiction for the case $\alpha=\frac{1}{2}$.

i) We define a dyadic-log semigroup Γ_d. Let $d(1)=0$, $d(2)>0$, and $d(2^n)=n\,d(2)$ for $n\geqq 2$. If $2^n<j<2^{n+1}$, we define

$$d(j)=d(2^n)+\frac{j'}{2^n}d(2),$$

where $j=2^n+j'$. Next, we set $\gamma_j=d(j+1)$ for all $j\geqq 0$ and $\Gamma_d=\{\gamma_j\}$. For each $n\geqq 1$, there are 2^n+1 elements of Γ_d in the interval $[d(2^n), d(2^{n+1})]$; these elements are $d(2^n), d(2^n+1), d(2^n+2), \cdots, d(2^n+2^n)=d(2^{n+1})$. The interval $[d(2^n), d(2^{n+1})]$ has length $d(2)$ and for each $d(2^n+j)\in[d(2^n), d(2^{n+1}))$,

$$d(2^n+j+1)-d(2^n+j)=\frac{d(2)}{2^n}.$$

Note that $d(2^n)=\gamma_{2^n-1}$.

Γ_d is a semigroup. In fact, if $n\geqq m$, $0\leqq j<2^n$ and $0\leqq i<2^m$, then

$$d(2^n+j)+d(2^m+i)=d(2^n)+\frac{jd(2)}{2^n}+d(2^m)+\frac{id(2)}{2^m}=d(2^{n+m})+\frac{d(2)}{2^n}(j+i2^{n-m})$$

$$=d(2^{n+m})+\frac{d(2)}{2^{n+m}}(j2^m+i2^n)\in\Gamma_d,$$

where $0\leqq j2^m+i2^n<2\cdot 2^{m+n}$.

Example 6.3. Let S be a signal on a semigroup Γ, and assume $S(0)=1$. In this example we record some formulas for ν besides the basic one developed in Theorem 2.1. The major thrust of the example is to relate the measure ν with the polynomials $P_n(x_1,\cdots,x_n)$ that we constructed in [5, Theorem 3.3 and Example 3.2].

a) We begin by recalling the main property of $\{P_n\}$. First, we define

$$S_N(x) = \sum_{m=1}^{\infty} S\left(\frac{m}{N}\right) \chi_m(x),$$

where $\chi_m = \chi_{[(m-1)/N, m/N)}$. Then there is a sequence $\{P_n\}$ of polynomials (of nth degree and n variables) such that $\mu_N * S_N = H$, where

$$\mu_N = \sum_{n=0}^{\infty} \lambda_N^n \delta_{n/N}$$

and $\lambda_N^n = P_n(x_1, \cdots, x_n)$ for $x_j = S((j+1)/N)$.

b) Because of part a) and the definition of S^{-1} we can formulate ν in terms of $\Gamma_N = \{n/N: n = 0, 1, \cdots\}$ as follows:

$$\forall N, \quad \nu = \sum_{\substack{0 \leqq n, 1 \leqq m, \\ \gamma \in \Gamma}} a_\gamma S\left(\frac{m}{N}\right) \lambda_N^n (\delta_{\gamma+(m+n-1)/N} - \delta_{\gamma+(m+n)/N}). \tag{6.10}$$

Note that the right-hand side of (6.10) is independent of N. Thus, if $\gamma' < \lambda < \gamma''$ are consecutive elements of Γ and φ is a continuous function satisfying $\varphi(\gamma) = 1$ and supp $\varphi \in (\gamma', \gamma'')$, then

$$a_\lambda = \sum_{m,n,\gamma} a_\gamma S\left(\frac{m}{N}\right) \lambda_N^n \left(\varphi\left(\gamma + \frac{m+n-1}{N}\right) - \varphi\left(\gamma + \frac{m+n}{N}\right)\right), \tag{6.11}$$

independent of N. Since $S(0) = 1$, we obtain

$$\forall M \text{ and } \forall N > \frac{M+1}{\gamma_1}, \quad 1 = \sum_{n=0}^{M} \lambda_N^n. \tag{6.12}$$

c) Using a standard technique from integral equations, we compute

$$\nu = \sum_{n=0}^{\infty} (\delta - S)^n * H, \tag{6.13}$$

where $(\delta - S)^n$ denotes n-fold convolution.

REFERENCES

[1] J. BENEDETTO, *The Laplace transform of generalized functions*, Canad. J. Math., 18 (1965), pp. 357–374.

[2] ———, *The Wiener spectrum in spectral synthesis*, MIT Studies in Applied Math., 54 (1975), pp. 91–115.

[3] ———, *Spectral Synthesis*, Academic Press, New York, 1975.

[4] ———, *Real Variable and Integration*, B. G. Teubner, Stuttgart, 1976.

[5] ———, *The theory of constructive signal analysis*, MIT Studies in Applied Math., 65 (1981), pp. 37–80.

[6] R. B. BLACKMAN AND J. W. TUKEY, *The Measurement of Power Spectra*, Dover, New York, 1959.

[7] P. DIEROLF AND J. VOIGT, *Convolution and $\mathcal{S}'$-convolution of distributions*, preprint.

[8] J. DIEUDONNÉ, *Sur les espaces de Köthe*, J. d'Analyse 1 (1951), pp. 81–115.

[9] ———, *Sur la convergence des suites de mesures de Radon*, Anais da Acad. Bras. Ciencias, 23 (1951), pp. 21–38.

[10] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part* I, John Wiley, New York, 1967.

[11] H. DYM AND P. MCKEAN, *Gaussian Processes, Function Theory, and the Inverse Spectral Problem*, Academic Press, New York, 1976.

[12] A. GROTHENDIECK, *Sur les applications linéaires faiblement compactes d'espaces du type $C(K)$*, Canad. J. Math., 5 (1953), pp. 129–173.

[13] ———, *Espaces vectoriels topologiques*, 3.a Edição, Sociedade de Matematica de S. Paulo, São Paulo, 1964.

[14] G. H. HARDY AND E. M. WRIGHT, *An Introduction to the Theory of Numbers*, 4th ed., Oxford University Press, London, 1960.
[15] J. HORVÁTH, *Topological Vector Spaces and Distributions*, Addison-Wesley, Reading, MA, 1966.
[16] ———, *Finite parts of distributions*, ISNM, 20 (1972), pp. 142–158.
[17] P. KOOSIS, *Exposition of a theorem of Bertil Nyman*, Conference in Harmonic Anal., U. of Warwick, 1968.
[18] B. NYMAN, *On some groups and semigroups of translations*, Phd. Thesis, Univ. of Uppsala, 1950.
[19] A. OPPENHEIM AND R. SCHAFER, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.
[20] L. SCHWARTZ, *Théorie des distributions*, Hermann, Paris, 1966.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0003 $01.00/0

ON A GENERAL CONCEPT FOR SEPARATION OF VARIABLES*

J. HAINZL†

Abstract. A new definition of (generalized) separability for a linear operator L of arbitrary order is given which, when specialized to the Helmholtz operator, includes ordinary separability in nonorthogonal and "heat type" coordinates. On the basis of this definition, a set of commuting operators is explicitly constructed and various relationships between eigenvalue problems for these operators and separated equations corresponding to $Lu = 0$ are derived. Moreover, a transformation of Bäcklund type for separable operators is obtained.

1. Introduction. In a number of papers, W. Miller and others (see the references in [8]) obtained numerous interesting results concerning separation of variables in mathematical physics. A first attempt to deduce some of these results in a systematic way from an exact definition of variable separation was made in [2]. There, I succeeded only by using a rather narrow definition of separability which was a simplified version of Niessen's definition in [9]. For any operator separable according to this definition and satisfying certain additional assumptions, I was able in [2] to construct explicitly a set of mutually commuting operators with properties found earlier for many examples separately (see [8]).

In their recent paper [5], Kalnins and Miller notice that my definition used in [2] does not include separability of the Helmholtz operator in nonorthogonal and "heat type" coordinates. The present paper is a new attempt to overcome this problem. A new definition of (generalized) separability (G-separability) is proposed which extends Niessen's definition [9]. It covers, e.g., the nonorthogonal types of separable coordinates for the 4-dimensional complex Helmholtz equation which were listed by Kalnins and Miller in § 3 of their paper [4]. An interesting feature of the new definition is that in the corresponding "separated" equations there may occur more than one o.d.e. with respect to some fixed variable, and, in addition, p.d.e.'s are admitted, too. The present definition is still narrow enough to allow proving results which include the main statements in [2] as special cases. For instance, if L is a G-separable operator, it is now possible, without any essential further assumption, to construct a set of (partly) commuting operators such that there exist useful relationships between eigenvalue problems for these operators and the equation

$$Lu = 0. \tag{1.1}$$

The present paper may also be regarded as an attempt to settle some of the general questions concerning variable separation which were raised by Koornwinder in [6] and [7].

Let us say a little more about the contents. Section 2 contains the definition of G-separability of an operator L, which is justified by nontrivial examples and by Theorem 1 claiming that any solution of the "separated" equations also satisfies (1.1). In § 3 the operators S_{ij} and from these the operators S_i depending on separation constants are defined. Theorem 2 then states surprising commutation relations including the mutual commutability of all S_i. Central topics of § 4 are the interrelations between the simultaneous eigenvalue equations for the operators S_i and two kinds of separated equations (Theorems 3 and 4). A frequent special case is treated differently in Theorem 5. Section 5 investigates simultaneous eigenfunctions of S_{ij}. Here, various

* Received by the editors December 23, 1980, and in revised form June 11, 1981.
† Fachbereich Mathematik, Gesamthochschule/Universität Kassel, Kassel, West Germany.

results are obtained concerning the question when such eigenfunctions also satisfy (1.1). In § 6 we use Theorem 1 to derive results that can be interpreted as a Bäcklund transform for G-separable operators (Theorems 8 and 9). In the last section, § 7, some of the results achieved in the paper are illustrated by means of the examples given in § 2.

2. Definition and justification of generalized separability; special cases.

Assumptions. Let I_k, $k = 1, \cdots, N$ be open connected subsets of $\mathbb{R}$ or $\mathbb{C}$ and $\Omega := I_1 \times \cdots \times I_N$.

L_{ij} are assumed to be linear homogeneous (ordinary or partial) differential operators of arbitrary order with C^∞-coefficients defined on Ω, $i, j = 1, \cdots, n$. If we denote by a_{ij} the term of order zero in L_{ij}, and put

$$L_{ij} =: M_{ij} + a_{ij},$$

the following condition has to be satisfied.

$$\left.\begin{array}{l} M_{ij} \text{ commutes with } M_{kl} \text{ and } a_{kl}, \text{ i.e.,} \\ [M_{ij}, M_{kl}] := M_{ij}M_{kl} - M_{kl}M_{ij} = 0, \\ [M_{ij}, a_{kl}] = 0 \end{array}\right\} \quad \text{for } i, j, k, l = 1, \cdots, n, \quad k \neq i. \tag{2.1}$$

Then, obviously, $[L_{ij}, L_{kl}] = 0$ for $k \neq i$, and

$$\det (L_{ij})_{\substack{1 \leqq i \leqq n \\ 1 \leqq j \leqq n}}$$

is a well-defined linear homogeneous differential operator on Ω.

DEFINITION OF GENERALIZED SEPARABILITY (for short, *G-separability*). A linear homogeneous partial differential operator L of arbitrary order, expressed in the variables $x = (x_1, \cdots, x_N) \in \Omega$, is called *G-separable* if and only if for some $n \geqq N$ there exist operators L_{ij} $(i, j = 1, \cdots, n)$ satisfying the preceding assumptions and a C^∞-function f, defined and nowhere zero on Ω, such that

$$fL = \det (L_{ij}).$$

This definition of G-separability is actually unrelated to coordinates; it covers, however, the usual separability concept. We note the following special cases (i)–(iv):

(i) $n = N$; L_{ij} is an ordinary differential operator in the variable x_i, $i, j = 1, \cdots, n$. If $f \equiv 1$, this yields the definition of separability given by Niessen [9]. If, on the other hand, $M_{ij} = 0$ for $j = 2, \cdots, n$, we obtain the definition used in my paper [2].

(ii) L_{ij} is an ordinary differential operator in some variable x_k, where k depends on i, but is independent of j. In this case, condition (2.1) is not satisfied automatically and therefore has to be verified. As an example, we can take the complex 4-dimensional Helmholtz operator $L := \Delta_4 - \mu$, expressed in some nonorthogonal separable coordinate systems established by Kalnins and Miller in [4]. From that paper, the coordinate systems (3.1), (3.16) and (3.19) could be chosen. We give the details for the coordinate system (3.19). The connection between the cartesian coordinates $z_1, \cdots, z_4$ and $x_1, \cdots, x_4$ reads

$$\begin{aligned} z_1 + iz_2 &= x_1x_2 - \tfrac{1}{2}x_2^2 + 2x_4, & z_1 - iz_2 &= x_1, \\ z_3 + iz_4 &= x_1x_2 - \tfrac{1}{2}x_1^2 + 2x_3, & z_3 - iz_4 &= x_2, \end{aligned} \tag{2.2}$$

and L can be expressed in the following way as a determinant ($\partial_k := \partial/\partial x_k$):

$$\Delta_4 - \mu = 2\partial_1\partial_4 + 2\partial_2\partial_3 + x_1\partial_3^2 + x_2\partial_4^2 - \mu = \begin{vmatrix} 0 & -1 & 0 & x_1 & 2\partial_1 & 0 \\ -\mu & 1 & 2\partial_2 & 0 & 0 & x_2 \\ \partial_3 & 0 & -1 & 0 & 0 & 0 \\ \partial_3^2 & 0 & 0 & -1 & 0 & 0 \\ \partial_4 & 0 & 0 & 0 & -1 & 0 \\ \partial_4^2 & 0 & 0 & 0 & 0 & -1 \end{vmatrix}. \tag{2.3}$$

Thus, $n = 6$, $f \equiv 1$, $\Omega = \mathbb{C}^4$, and the operators L_{ij}, defined by the determinant, obviously satisfy (2.1).

(iii) For some indices i, L_{ij} is a *partial* differential operator with respect to a certain subset of variables (which is the same for all j), the remaining L_{ij} being of the same type as in case (ii). Condition (2.1) has to be verified.

Example. $L := \Delta_4 - \mu$ (complex). Choose the nonorthogonal separable coordinates given by Kalnins and Miller in [4, (3.34)]:

$$\begin{aligned} z_1 + iz_2 &= \sqrt{x_1x_2}\cosh x_4, \qquad z_3 + iz_4 = \sqrt{x_1x_2}\sinh x_4, \\ z_1 - iz_2 &= \left(\sqrt{\frac{x_1}{x_2}} + \sqrt{\frac{x_2}{x_1}}\right)\cosh x_4 - x_3\sqrt{x_1x_2}\sinh x_4, \\ z_3 - iz_4 &= -\left(\sqrt{\frac{x_1}{x_2}} + \sqrt{\frac{x_2}{x_1}}\right)\sinh x_4 + x_3\sqrt{x_1x_2}\cosh x_4, \end{aligned} \tag{2.4}$$

where $z_1, \cdots, z_4$ are the Cartesian coordinates. Then we find the following determinant expression:

$$(x_1 - x_2)(\Delta_4 - \mu) = \frac{4}{x_1}\partial_1(x_1^3\partial_1) - \frac{4}{x_2}\partial_2(x_2^3\partial_2) + \frac{x_1^2 - x_2^2}{(x_1x_2)^2}\partial_3^2 + \frac{2(x_1 - x_2)}{x_1x_2}\partial_3\partial_4 - \mu(x_1 - x_2)$$

$$= \begin{vmatrix} L_{11} & -1 & 0 & -x_1^{-2} & 0 & -2x_1^{-1} \\ L_{21} & 1 & 0 & x_2^{-2} & 0 & 2x_2^{-1} \\ \partial_3 & 0 & -1 & 0 & 0 & 0 \\ \partial_3^2 & 0 & 0 & -1 & 0 & 0 \\ \partial_4 & 0 & 0 & 0 & -1 & 0 \\ \partial_3\partial_4 & 0 & 0 & 0 & 0 & -1 \end{vmatrix} \tag{2.5}$$

where

$$L_{11} := 4x_1^2\partial_1^2 + 12x_1\partial_1 - \mu x_1, \qquad L_{21} := -4x_2^2\partial_2^2 - 12x_2\partial_2 + \mu x_2.$$

Condition (2.1) is clearly satisfied; $f(x) := x_1 - x_2$.

(iv) Any linear homogeneous partial differential operator with constant coefficients is G-separable. As an example showing the general procedure we take the operator

$$L := c_{11}\partial_1^2 + c_{12}\partial_1\partial_2 + c_{22}\partial_2^2 + c_1\partial_1 + c_0, \tag{2.6}$$

which can be written in the form

$$L = \begin{vmatrix} c_0 & c_1 & c_{12} & c_{11} & c_{22} \\ \partial_1 & -1 & 0 & 0 & 0 \\ \partial_1\partial_2 & 0 & -1 & 0 & 0 \\ \partial_1^2 & 0 & 0 & -1 & 0 \\ \partial_2^2 & 0 & 0 & 0 & -1 \end{vmatrix}. \tag{2.7}$$

The following result justifies our definition of G-separability by showing that it has the crucial property known for ordinary separability. (Henceforth, the functions u are always understood to be C^∞ on Ω.)

THEOREM 1. *Let L be G-separable with L_{ij} as above, and assume that for some $(\lambda_1, \cdots, \lambda_n) \in \mathbb{C}^n \backslash \{0\}$, the function u is a solution of the "separated" equations*

$$\sum_{k=1}^{n} \lambda_k L_{ik} u = 0, \qquad i = 1, \cdots, n. \tag{2.8}$$

Then u is a solution of $Lu = 0$, too.

Proof. Assume without restriction $\lambda_1 \neq 0$. In the following determinants involving *operators* L_{ij} and *functions* $L_{il}u$, it is understood that, when they are expressed as alternating sums of products, the factor $L_{il}u$ is written on the extreme right-hand side. With this agreement, we obtain by repeatedly using $[L_{ij}, L_{kl}] = 0$ for $k \neq i$:

$$\begin{aligned} fLu = \det(L_{ij})u &= \begin{vmatrix} L_{11}u & L_{12} & \cdots & L_{1n} \\ \vdots & \vdots & & \vdots \\ L_{n1}u & L_{n2} & \cdots & L_{nn} \end{vmatrix} \\ &= \begin{vmatrix} -\lambda_1^{-1} \sum_{k=2}^{n} \lambda_k L_{1k} u & L_{12} & \cdots & L_{1n} \\ \vdots & \vdots & & \vdots \\ -\lambda_1^{-1} \sum_{k=2}^{n} \lambda_k L_{nk} u & L_{n2} & \cdots & L_{nn} \end{vmatrix} \\ &= -\lambda_1^{-1} \sum_{k=2}^{n} \lambda_k \begin{vmatrix} L_{1k}u & L_{12} & \cdots & L_{1n} \\ \vdots & \vdots & & \vdots \\ L_{nk}u & L_{n2} & \cdots & L_{nn} \end{vmatrix} \\ &= -\lambda_1^{-1} \sum_{k=2}^{n} \lambda_k \underbrace{\begin{vmatrix} L_{1k} & L_{12} & \cdots & L_{1n} \\ \vdots & \vdots & & \vdots \\ L_{nk} & L_{n2} & \cdots & L_{nn} \end{vmatrix}}_{= 0 \text{ since two columns are identical}} u \\ &= 0. \end{aligned}$$

Hence, $Lu = 0$. □

3. Construction of commuting operators. From now on, we assume L to be G-separable, with f, M_{ij}, a_{ij}, for $i, j = 1, \cdots, n$ defined as in § 2. The rank of the matrix (a_{ij}) is assumed constant throughout Ω; more precisely, we postulate for some

m, $0 \leqq m \leqq n$,

$$\left.\begin{aligned} &\text{rank of } (a_{ij}(x))_{\substack{1\leqq i\leqq n\\ 1\leqq j\leqq n}} &&= n-m \\ &a(x) := \det (a_{ij}(x))_{\substack{m+1\leqq i\leqq n\\ m+1\leqq j\leqq n}} &&\neq 0 \end{aligned}\right\} \quad \text{for all } x \in \Omega. \tag{3.1}$$

Let A_{rs} denote the algebraic complement of a_{rs} in the matrix $(a_{ij})_{m+1\leqq i\leqq n, m+1\leqq j\leqq n}$. From (2.1), we then easily see that

$$[M_{pj}, A_{pi}] = 0, \quad [M_{kj}, A_{pi}] = 0, \quad [M_{kj}, a] = 0 \tag{3.2}$$
$$\text{for } i, p = m+1, \cdots, n, \quad k = 1, \cdots, m, \quad j = 1, \cdots, n.$$

Now define the operators S_{ij}, $i, j = 1, \cdots, n$ as follows:

$$S_{ij} := \begin{cases} -a^{-1} \sum\limits_{p=m+1}^{n} A_{pi} M_{pj} & \text{for } i = m+1, \cdots, n, \quad j = 1, \cdots, n, \\ -M_{ij} + a^{-1} \sum\limits_{p,q=m+1}^{n} a_{iq} A_{pq} M_{pj} & \text{for } i = 1, \cdots, m, \quad j = 1, \cdots, n. \end{cases} \tag{3.3}$$

A straightforward calculation then yields M_{ij} in terms of S_{kj}:

$$M_{ij} = \begin{cases} -S_{ij} - \sum\limits_{k=m+1}^{n} a_{ik} S_{kj} & \text{for } i = 1, \cdots, m, \quad j = 1, \cdots, n, \\ -\sum\limits_{k=m+1}^{n} a_{ik} S_{kj} & \text{for } i = m+1, \cdots, n, \quad j = 1, \cdots, n. \end{cases} \tag{3.4}$$

An important instrument in the calculations below is the following formula proved in [2, proof of Thm. 1] by means of Sylvester's determinant identity (see [1, p. 32, (28)]):

$$A_{ki} A_{pj} - A_{kj} A_{pi} = (-1)^{k+p+i+j} a A_{ij}^{kp}, \tag{3.5}$$

where $m+1 \leqq k < p \leqq n$, $m+1 \leqq i < j \leqq n$, and A_{ij}^{kp} is the determinant of the matrix arising from $(a_{rs})_{m+1\leqq r\leqq n, m+1\leqq s\leqq n}$ by cancelling the two rows with numbers k, p and the two columns with numbers i, j. From (2.1) we conclude

$$[A_{ij}^{kp}, M_{kl}] = 0, \qquad [A_{ij}^{kp}, M_{pl}] = 0 \quad \text{for } l = 1, \cdots, n. \tag{3.6}$$

With arbitrary but fixed complex numbers $\lambda_1, \cdots, \lambda_n$ we now define the operators S_i:

$$S_i := \sum_{k=1}^{n} \lambda_k S_{ik}, \qquad 1 \leqq i \leqq n. \tag{3.7}$$

Of course, we are interested only in the case where not all λ_i are zero.

THEOREM 2. (a) $[S_{ir}, S_{js}] + [S_{is}, S_{jr}] = 0$ *for* $i, j, r, s = 1, \cdots, n$.

(b) $[S_i, S_j] = 0$ *for* $i, j = 1, \cdots, n$.

Proof. As is easily seen, it suffices to prove (a) and (b) for $i < j$. We obtain

$$[S_i, S_j] = \sum_{r,s=1}^{n} \lambda_r \lambda_s [S_{ir}, S_{js}] = \sum_{1\leqq r<s\leqq n} \lambda_r \lambda_s ([S_{ir}, S_{js}] + [S_{is}, S_{jr}]) + \sum_{r=1}^{n} \lambda_r^2 [S_{ir}, S_{jr}];$$

hence (a) implies (b). To prove (a), we distinguish three possible cases, putting

$$X := [S_{ir}, S_{js}] + [S_{is}, S_{jr}].$$

Case 1. $m+1 \leqq i < j \leqq n$. Using (3.3), (3.2), (3.5), and (3.6), we calculate

$$X = a^{-1} \sum_{k,p=m+1}^{n} \begin{Bmatrix} A_{ki}M_{kr}a^{-1}A_{pj}M_{ps} - A_{pj}M_{ps}a^{-1}A_{ki}M_{kr} \\ + A_{ki}M_{ks}a^{-1}A_{pj}M_{pr} - A_{pj}M_{pr}a^{-1}A_{ki}M_{ks} \end{Bmatrix}$$

$$= a^{-1} \sum_{\substack{k,p=m+1 \\ k \neq p}}^{n} \begin{Bmatrix} M_{kr}A_{ki}A_{pj}a^{-1}M_{ps} - M_{pr}A_{ki}A_{pj}a^{-1}M_{ks} \\ + M_{ks}A_{ki}A_{pj}a^{-1}M_{pr} - M_{ps}A_{ki}A_{pj}a^{-1}M_{kr} \end{Bmatrix}$$

$$= a^{-1} \sum_{m+1 \leqq k < p \leqq n} \begin{Bmatrix} M_{kr}(A_{ki}A_{pj} - A_{kj}A_{pi})a^{-1}M_{ps} \\ - M_{pr}(A_{ki}A_{pj} - A_{kj}A_{pi})a^{-1}M_{ks} \\ + M_{ks}(A_{ki}A_{pj} - A_{kj}A_{pi})a^{-1}M_{pr} \\ - M_{ps}(A_{ki}A_{pj} - A_{kj}A_{pi})a^{-1}M_{kr} \end{Bmatrix}$$

$$= a^{-1} \sum_{m+1 \leqq k < p \leqq n} (-1)^{k+p+i+j} \underbrace{A_{ij}^{kp}(M_{kr}M_{ps} - M_{pr}M_{ks} + M_{ks}M_{pr} - M_{ps}M_{kr})}_{= 0 \text{ from (2.1)}}$$

$$= 0.$$

Case 2. $1 \leqq i \leqq m < j \leqq n$. From (3.4) we first obtain

$$X = -\left[M_{ir} + \sum_{k=m+1}^{n} a_{ik}S_{kr}, S_{js}\right] - \left[M_{is} + \sum_{k=m+1}^{n} a_{ik}S_{ks}, S_{jr}\right]$$

$$= -[M_{ir}, S_{js}] - [M_{is}, S_{jr}] - \sum_{k=m+1}^{n} [a_{ik}S_{kr}, S_{js}] - \sum_{k=m+1}^{n} [a_{ik}S_{ks}, S_{jr}].$$

Now we claim that the first two terms on the right-hand side vanish. More precisely,

(3.8) $$[M_{ip}, S_{jq}] = 0 \quad \text{for } 1 \leqq i \leqq m < j \leqq n, \quad p, q = 1, \cdots, n.$$

This follows from definition (3.3) together with (2.1), (3.2), noticing that $[M_{ip}, a] = 0$ when $p = 1, \cdots, n$ implies

$$aa^{-1}M_{ip} = M_{ip}aa^{-1} = aM_{ip}a^{-1}$$

and hence, $[M_{ip}, a^{-1}] = 0$. Again from (3.3) and (2.1) we deduce

(3.9) $$[a_{ik}, S_{jl}] = 0, \qquad k = m+1, \cdots, n, \quad l = 1, \cdots, n, \quad 1 \leqq i \leqq m < j \leqq n.$$

This implies

$$[a_{ik}S_{kr}, S_{js}] = a_{ik}[S_{kr}, S_{js}], \qquad [a_{ik}S_{ks}, S_{jr}] = a_{ik}[S_{ks}, S_{jr}].$$

Thus,

$$X = -\sum_{k=m+1}^{n} a_{ik}\underbrace{([S_{kr}, S_{js}] + [S_{ks}, S_{jr}])}_{= 0 \text{ from case 1}} = 0.$$

Case 3. $1 \leqq i \leqq j \leqq m$. Equations (3.4) yield

$$X = \underbrace{[M_{ir}, M_{js}]}_{= 0 \text{ from (2.1)}} + \underbrace{\left[M_{ir}, \sum_{k=m+1}^{n} a_{jk}S_{ks}\right]}_{= 0 \text{ from (2.1), (3.8)}} + \underbrace{\left[\sum_{k=m+1}^{n} a_{ik}S_{kr}, M_{js}\right]}_{= 0}$$

$$+ \left[\sum_{k=m+1}^{n} a_{ik}S_{kr}, \sum_{p=m+1}^{n} a_{jp}S_{ps}\right] + \underbrace{[M_{is}, M_{jr}]}_{= 0}$$

$$+\Big[\underbrace{M_{is}, \sum_{k=m+1}^{n} a_{jk}S_{kr}}_{=0}\Big]+\Big[\underbrace{\sum_{k=m+1}^{n} a_{ik}S_{ks}, M_{jr}}_{=0}\Big]+\Big[\sum_{k=m+1}^{n} a_{ik}S_{ks}, \sum_{p=m+1}^{n} a_{jp}S_{pr}\Big]$$

$$= \sum_{k,p=m+1}^{n} ([a_{ik}S_{kr}, a_{jp}S_{ps}]+[a_{ik}S_{ks}, a_{jp}S_{pr}]).$$

By means of (3.9), we obtain

$$a_{ik}S_{kr}a_{jp}S_{ps} - a_{jp}S_{ps}a_{ik}S_{kr} = a_{ik}a_{jp}S_{kr}S_{ps} - a_{jp}a_{ik}S_{ps}S_{kr} = a_{ik}a_{jp}[S_{kr}, S_{ps}],$$

and in the same way

$$[a_{ik}S_{ks}, a_{jp}S_{pr}] = a_{ik}a_{jp}[S_{ks}, S_{pr}].$$

Hence,

$$X = \sum_{k,p=m+1}^{n} a_{ik}a_{jp}(\underbrace{[S_{kr}, S_{ps}]+[S_{ks}, S_{pr}]}_{=0 \text{ from case 1}}) = 0.$$

This ends the proof of Theorem 2. □

4. Simultaneous eigenfunctions of $S_1, \cdots, S_n$. Remember that S_i are defined in (3.7) by means of n arbitrary constants $\lambda_1, \cdots, \lambda_n$ which are assumed fixed. We now prove the following equivalence:

THEOREM 3. *Let u be any function on Ω. Then the statements* (a) *and* (b) *are equivalent*:

(a) *u satisfies the n eigenvalue equations*

$$S_i u = \lambda_i u, \qquad 1 \leqq i \leqq n. \tag{4.1}$$

(b) *u solves the n "separated" equations*

$$\begin{aligned} \sum_{k=1}^{n} \lambda_k M_{ik}u &= -\lambda_i u - \sum_{l=m+1}^{n} \lambda_l a_{il}u, \qquad i = 1, \cdots, m, \\ \sum_{k=1}^{n} \lambda_k M_{ik}u &= -\sum_{l=m+1}^{n} \lambda_l a_{il}u, \qquad i = m+1, \cdots, n. \end{aligned} \tag{4.2}$$

Proof. (a) $\Rightarrow$ (b). Using (3.4) we calculate for $m+1 \leqq i \leqq n$

$$\begin{aligned} \sum_{k=1}^{n} \lambda_k M_{ik}u &= -\sum_{k=1}^{n} \lambda_k \sum_{l=m+1}^{n} a_{il}S_{lk}u \\ &= -\sum_{l=m+1}^{n} a_{il} \sum_{k=1}^{n} \lambda_k S_{lk}u = -\sum_{l=m+1}^{n} a_{il}S_l u = -\sum_{l=m+1}^{n} \lambda_l a_{il}u, \end{aligned}$$

and for $1 \leqq i \leqq m$

$$\sum_{k=1}^{n} \lambda_k M_{ik}u = -\sum_{k=1}^{n} \lambda_k \Big(S_{ik} + \sum_{l=m+1}^{n} a_{il}S_{lk}\Big)u = -S_i u - \sum_{l=m+1}^{n} a_{il}S_l u = -\lambda_i u - \sum_{l=m+1}^{n} \lambda_l a_{il}u.$$

(b) $\Rightarrow$ (a). From (3.7) and (3.3) we obtain for $m+1 \leqq i \leqq n$

$$\begin{aligned} S_i u &= -\sum_{k=1}^{n} \lambda_k a^{-1} \sum_{p=m+1}^{n} A_{pi} M_{pk} u \\ &= -a^{-1} \sum_{p=m+1}^{n} A_{pi} \sum_{k=1}^{n} \lambda_k M_{pk} u \\ &= a^{-1} \sum_{p=m+1}^{n} A_{pi} \sum_{l=m+1}^{n} \lambda_l a_{pl} u \\ &= a^{-1} \sum_{l=m+1}^{n} \lambda_l \underbrace{\sum_{p=m+1}^{n} A_{pi} a_{pl}}_{=a\delta_{il}} u \\ &= \lambda_i u. \end{aligned}$$

Using this result and (3.4), we finally obtain for $1 \leqq i \leqq m$

$$\begin{aligned} S_i u &= -\sum_{k=1}^{n} \lambda_k M_{ik} u - \sum_{k=1}^{n} \lambda_k \sum_{p=m+1}^{n} a_{ip} S_{pk} u \\ &= \lambda_i u + \sum_{l=m+1}^{n} \lambda_l a_{il} u - \sum_{p=m+1}^{n} a_{ip} \underbrace{\sum_{k=1}^{n} \lambda_k S_{pk} u}_{=S_p u = \lambda_p u} = \lambda_i u. \end{aligned}$$

□

Notice that (4.2) and (2.8) coincide in case $m = 0$. In any case, the following result can be derived:

THEOREM 4. *A solution u of* (4.2) (*or* (4.1)) *which vanishes nowhere on Ω also satisfies equations* (2.8)—*with the same* $(\lambda_1, \cdots, \lambda_n)$ *as in* (4.2)—*if and only if* $\lambda_1 = \lambda_2 = \cdots = \lambda_m = 0$.

Proof. Define the matrices $A_1, \cdots, A_4$ by decomposing $A := (a_{ij})_{1 \leqq i \leqq n, 1 \leqq j \leqq n}$ as follows:

$$A = \left(\begin{array}{c|c} A_1 & A_2 \\ \hline A_3 & A_4 \end{array} \right) \begin{matrix} \} m \\ \} n-m \end{matrix} \qquad \begin{matrix} \\ \\ \underbrace{}_{m} \end{matrix} \quad \begin{matrix} \\ \\ \underbrace{}_{n-m} \end{matrix} \tag{4.3}$$

By assumption (3.1), there exist A_4^{-1} and an $(n-m, m)$-matrix function C such that $A_1 = A_2 C$, $A_3 = A_4 C$. Thus, $C = A_4^{-1} A_3$ and

$$A_1 = A_2 A_4^{-1} A_3. \tag{4.4}$$

Now a simple calculation shows that a solution $u \neq 0$ of (4.2) also satisfies (2.8) if and only if

$$A_1 \tilde{\lambda} = \tilde{\lambda}, \qquad A_3 \tilde{\lambda} = 0 \quad \text{in } \Omega, \tag{4.5}$$

where $\tilde{\lambda}$ is the column vector $(\lambda_1, \cdots, \lambda_m)^t$. From (4.4) and (4.5) we finally conclude

$$\tilde{\lambda} = A_2 A_4^{-1} \underbrace{A_3 \tilde{\lambda}}_{=0} = 0. \qquad \square$$

Let us still consider the following special case which frequently occurs (for $m = 1$ see § 7):

(v) $\qquad M_{ij} = 0 \quad \textit{for } 1 \leqq i \leqq n, \quad m+1 \leqq j \leqq n.$

In this case we obtain $S_{ij}=0$ for $m+1\leqq j\leqq n$. Now, if for some $\lambda\in\mathbb{C}^n$, the nonzero function u satisfies (4.1) and (2.8), Theorem 4 yields $\lambda_1=\lambda_2=\cdots=\lambda_m=0$, hence

$$S_i=\sum_{k=1}^{n}\lambda_k S_{ik}=0 \quad \text{for } 1\leqq i\leqq n.$$

Using (4.1) again, we now obtain $\lambda_{m+1}=\cdots=\lambda_n=0$, and thus, $\lambda=0$.

Summarizing, we have seen that in case (v) (4.1) and (2.8) can be satisfied simultaneously only in a trivial way. Hence, in this case equations (4.1) cannot be used to find such nontrivial solutions of $Lu=0$ which also satisfy (2.8). However, the following substitute of Theorem 3 can be established (for $m=1$ see also [2, Thm. 8]):

THEOREM 5. *Suppose* (v) *is satisfied. Then*:

(a) *For any fixed* $(\lambda_{m+1},\cdots,\lambda_n)\in\mathbb{C}^{n-m}$ *and any function* u *on* Ω *the statements* (a_1) *and* (a_2) *are equivalent*:

(a_1) *u satisfies the equations*

$$\sum_{j=1}^{m} S_{ij}u=\begin{cases}0 & \text{for } i=1,\cdots,m,\\ \lambda_i u & \text{for } i=m+1,\cdots,n.\end{cases} \tag{4.6}$$

(a_2) *u satisfies the separated equations*

$$\sum_{j=1}^{m} M_{ij}u+\sum_{k=m+1}^{n}\lambda_k a_{ik}u=0, \qquad i=1,\cdots,n. \tag{4.7}$$

(b) *If* $m\neq 0$ *and u satisfies* (4.6) *or* (4.7) *for some* $(\lambda_{m+1},\cdots,\lambda_n)\in\mathbb{C}^{n-m}$, *then u is a solution of* $Lu=0$.

Proof. (a) The proof of this part is a straightforward calculation using (3.3), (3.4). Since it is quite similar to the proof of Theorem 3, we omit it.

(b) By using multilinearity of the determinant fL with respect to the first m columns, and owing to assumption (3.1), we obtain

$$\begin{aligned} fL&=\begin{vmatrix} M_{11}+a_{11} & \cdots & M_{1m}+a_{1m} & a_{1,m+1} & \cdots & a_{1n}\\ \vdots & & \vdots & \vdots & & \vdots\\ M_{n1}+a_{n1} & \cdots & M_{nm}+a_{nm} & a_{n,m+1} & \cdots & a_{nn}\end{vmatrix}\\ &=\begin{vmatrix} M_{11} & \cdots & M_{1m} & a_{1,m+1} & \cdots & a_{1n}\\ \vdots & & \vdots & \vdots & & \vdots\\ M_{n1} & \cdots & M_{nm} & a_{n,m+1} & \cdots & a_{nn}\end{vmatrix}.\end{aligned} \tag{4.8}$$

To this latter determinant, which is a G-separable operator again, we may now apply Theorem 1 with $(\lambda_1,\cdots,\lambda_n)=(1,\cdots,1,\lambda_{m+1},\cdots,\lambda_n)\neq 0$. The proof is completed by noticing that the role of equations (2.8) is taken over by equations (4.7). □

For $m=1$, another useful formula holds (see [2, (20), (22)]):

COROLLARY 1. *Assume* (V) *with* $m=1$. *Then*

$$S_{11}=-a^{-1}fL. \tag{4.9}$$

Proof. Expanding the determinant on the right-hand side of (4.8) ($m=1$) with respect to the first row, and using (3.3), (3.4), we obtain

$$fL=aM_{11}+a\sum_{k=2}^{n}a_{1k}S_{k1}=-aS_{11}. \qquad \square$$

In the next section, we investigate eigenfunctions u of the operators S_{ij} and ask when such functions also satisfy $Lu=0$.

5. Simultaneous eigenfunctions of the operators S_{ij}. Now assume that the function u on Ω is an eigenfunction of all operators S_{ij} with corresponding eigenvalues β_{ij}:

$$S_{ij}u = \beta_{ij}u, \qquad i,j = 1,\cdots,n. \tag{5.1}$$

Put

$$B := (\beta_{ij})_{\substack{1\leqq i\leqq n\\ 1\leqq j\leqq n}} \tag{5.2}$$

and let λ be the column vector with elements $\lambda_1,\cdots,\lambda_n$ occurring in (3.7). In case u does not vanish identically, we notice that u also satisfies (4.1) if and only if

$$B\lambda = \lambda. \tag{5.3}$$

We first prove the following lemma:

LEMMA 1. *If u satisfies* (5.1), *then*

$$fLu = gu. \tag{5.4}$$

Here, the function g is given by

$$g := (-1)^n \det D, \tag{5.5}$$

where $D = (d_{ij})$ is the following (n,n)-matrix:

$$d_{ij} := \begin{cases} \beta_{ij} + \sum\limits_{k=m+1}^{n} a_{ik}\beta_{kj} - a_{ij}, & 1\leqq i\leqq m,\\ \sum\limits_{k=m+1}^{n} a_{ik}\beta_{kj} - a_{ij}, & m+1\leqq i\leqq n, \end{cases} \qquad 1\leqq j\leqq n. \tag{5.6}$$

Proof. By (3.4) and (5.1) we obtain

$$L_{ij}u = (M_{ij} + a_{ij})u = \begin{cases} \Big(-\beta_{ij} - \sum\limits_{k=m+1}^{n} a_{ik}\beta_{kj} + a_{ij}\Big)u, & 1\leqq i\leqq m,\\ \Big(-\sum\limits_{k=m+1}^{n} a_{ik}\beta_{kj} + a_{ij}\Big)u, & m+1\leqq i\leqq n. \end{cases}$$

By (2.1), the factors of u on the right-hand side commute with L_{rs}, $r<i$. Hence, we successively calculate (π runs through all permutations of the figures $1,\cdots,n$):

$$\begin{aligned} fLu &= \sum_{\pi} (-1)^{\operatorname{sgn}\pi} L_{1\pi(1)}\cdots L_{n\pi(n)}u\\ &= \sum_{\pi} (-1)^{\operatorname{sgn}\pi}\Big(-\sum_{k=m+1}^{n} a_{nk}\beta_{k\pi(n)} + a_{n\pi(n)}\Big)L_{1\pi(1)}\cdots L_{n-1,\pi(n-1)}u\\ &= \qquad\qquad \cdots\cdots\\ &= \sum_{\pi} (-1)^{\operatorname{sgn}\pi} \prod_{i=1}^{m}\Big(-\beta_{i\pi(i)} - \sum_{k=m+1}^{n} a_{ik}\beta_{k\pi(i)} + a_{i\pi(i)}\Big) \prod_{i=m+1}^{n}\Big(-\sum_{k=m+1}^{n} a_{ik}\beta_{k\pi(i)} + a_{i\pi(i)}\Big)u\\ &= (-1)^n(\det D)u. \quad \square \end{aligned}$$

Next, we ask for conditions under which a function u satisfying (5.1) is also a solution of $Lu = 0$. By Lemma 1, this means that we have to find conditions for $\det D$ to vanish identically. First we prove some sort of necessary condition.

LEMMA 2. *With B as in* (5.2), *define the matrices* $B_1, \cdots, B_4$ by

$$B =: \left(\begin{array}{c|c} \underbrace{B_1}_{m} & \underbrace{B_2}_{n-m} \\ \hline B_3 & B_4 \end{array}\right) \begin{matrix} \} m \\ \} n-m \end{matrix}. \tag{5.7}$$

Then, in general, $\det D$ *does not vanish identically unless the following property* (P) *holds* (I = *unit matrix*):

$$\det \begin{pmatrix} B_1 & B_2 \\ Z & B_4 - I \end{pmatrix} = 0 \quad \textit{for all } (n-m, m)\textit{-matrices } Z. \tag{P}$$

Proof. It suffices to give an example. For this, we assume case (i) of § 2, and choose the matrix (a_{ij}) in the following way:

$$(a_{ij})_{\substack{1 \leqq i \leqq n \\ 1 \leqq j \leqq n}} := \begin{pmatrix} 0 & 0 \\ X & I \end{pmatrix} \quad \text{with } X := (x_i^j)_{\substack{m+1 \leqq i \leqq n \\ 1 \leqq j \leqq m}}.$$

(5.6) then yields

$$D = \begin{pmatrix} B_1 & B_2 \\ B_3 - X & B_4 - I \end{pmatrix}.$$

We claim that, by the special choice of X, all possible products p_α made up of elements from different rows and columns of $B_3 - X$ together with the function 1 are linearly independent. This can easily be verified by successively regarding those products consisting of a maximal number of factors. Thus, from the assumption $\det D \equiv 0$ we conclude that, in the usual expression of $\det D$ as an alternating sum, all the coefficients of such products p_α must vanish. This obviously yields property (P). □

Remark. From (P), we clearly obtain

$$\det \begin{pmatrix} B_1 & B_2 \\ Z & B_4 - I \end{pmatrix} = \det \begin{pmatrix} B_1 & B_2 \\ 0 & B_4 - I \end{pmatrix} = \det B_1 \cdot \det (B_4 - I) = 0.$$

(Here and in what follows, the determinant of a $(0, 0)$-matrix is defined to be 1.) Moreover, condition (P) implies the following weaker condition (Q):

$$\det \begin{pmatrix} B_1 & B_2 \\ Z & B_4 - I \end{pmatrix} \text{ is independent of the choice of the matrix } Z. \tag{Q}$$

This condition holds, for example, if $B_2 = 0$. For $m = 0$ and $m = n$, condition (Q) is, of course, empty.

LEMMA 3. *Let* $D = (d_{ij})$ *be defined by* (5.6), *and for the matrix* B, *decomposed as in* (5.7), *assume that condition* (Q) *is satisfied. Then we have*

$$\det D = a \det B_1 \det (B_4 - I). \tag{5.8}$$

Proof. For $m = 0$ or $m = n$, assertion (5.8) is seen directly; thus we may assume $0 < m < n$. Using the decomposition (4.3) of the matrix (a_{ij}) and (4.4), we obviously get the following form for D:

$$\begin{aligned} D &= \begin{pmatrix} B_1 + A_2 B_3 - A_1 & B_2 + A_2 B_4 - A_2 \\ A_4 B_3 - A_3 & A_4 B_4 - A_4 \end{pmatrix} \\ &= \begin{pmatrix} B_1 + A_2 B_3 - A_2 A_4^{-1} A_3 & B_2 + A_2 B_4 - A_2 \\ A_4 B_3 - A_3 & A_4 B_4 - A_4 \end{pmatrix}, \end{aligned}$$

and this latter matrix coincides with the product

$$\begin{pmatrix} I_m & A_2 \\ 0 & A_4 \end{pmatrix} \cdot \begin{pmatrix} B_1 & B_2 \\ B_3 - A_4^{-1}A_3 & B_4 - I_{n-m} \end{pmatrix}$$

where I_k is the (k, k)-unit matrix. Here, the first factor has determinant $\det A_4 = a$, and by assumption (Q), we obtain for the second factor

$$\det \begin{pmatrix} B_1 & B_2 \\ B_3 - A_4^{-1}A_3 & B_4 - I_{n-m} \end{pmatrix} = \det \begin{pmatrix} B_1 & B_2 \\ 0 & B_4 - I_{n-m} \end{pmatrix} = \det B_1 \cdot \det (B_4 - I_{n-m}).$$

This proves (5.8). □

Remark. Notice that the above proof still works if condition (Q) is replaced by the special assumption

$$B_3 - A_4^{-1}A_3 = 0.$$

Thus, in case (Q) is not satisfied, the assertion of Lemma 3 remains true as long as $A_3 = 0$ and $B_3 = 0$.

LEMMA 4. *Suppose* $0 < m < n$, *and assume that the* (n, n)*-matrix* $B = (\beta_{ij})_{1 \leqq i \leqq n, 1 \leqq j \leqq n}$ *decomposed as in* (5.7), *satisfies condition* (Q). *Then* $\det B_1 \cdot \det (B_4 - I_{n-m}) \neq 0$ *implies* $B_2 = 0$.

Proof. (a) We first treat the following special case:

$$B_1 = I_m, \qquad B_4 - I_{n-m} = I_{n-m}.$$

Choosing a matrix $Z = (z_{ij})$ of indeterminates z_{ij}, $1 \leqq i \leqq n - m$, $1 \leqq j \leqq m$, we calculate the determinant

$$d := \det \begin{pmatrix} I_m & B_2 \\ Z & I_{n-m} \end{pmatrix}$$

by Laplace's generalized expansion theorem applied to the first m columns of the matrix. A straightforward computation yields

$$d = 1 + \sum_{i=1}^{m} \sum_{j=1}^{n-m} (\pm 1) z_{ji} \beta_{i,m+j} + \text{second and higher order terms in } z_{ij}. \tag{5.9}$$

Since, by condition (Q), d is independent of z_{ij}, in particular the coefficients of all linear terms in (5.9) must vanish. This means $B_2 = 0$.

(b) In the general case, we have by assumption $\det B_1 \neq 0$ and $\det (B_4 - I_{n-m}) \neq 0$, hence B_1^{-1} and $(B_4 - I_{n-m})^{-1}$ exist.

Now, from

$$\begin{pmatrix} B_1 & B_2 \\ Z & B_4 - I_{n-m} \end{pmatrix} \begin{pmatrix} B_1^{-1} & 0 \\ 0 & (B_4 - I_{n-m})^{-1} \end{pmatrix} = \begin{pmatrix} I_m & B_2 (B_4 - I_{n-m})^{-1} \\ ZB_1^{-1} & I_{n-m} \end{pmatrix}$$

we obtain

$$\begin{aligned} \tilde{d} &:= \det \begin{pmatrix} I_m & B_2 (B_4 - I_{n-m})^{-1} \\ ZB_1^{-1} & I_{n-m} \end{pmatrix} \\ &= (\det B_1)^{-1} (\det (B_4 - I_{n-m}))^{-1} \det \begin{pmatrix} B_1 & B_2 \\ Z & B_4 - I_{n-m} \end{pmatrix}. \end{aligned}$$

Thus, owing to (Q), $\tilde{d}$ is independent of Z. Then, obviously, $\tilde{d}$ is also independent of $\tilde{Z} := ZB_1^{-1}$. Now, from part (a) of this proof, we have $B_2(B_4 - I_{n-m})^{-1} = 0$ and therefore $B_2 = 0$. □

Now, let us draw some conclusions from the above lemmas. As an immediate consequence of Lemmas 1 and 3 we obtain:

THEOREM 6. *Assume u to be a simultaneous eigenfunction of all operators S_{ij} with corresponding eigenvalue matrix B satisfying condition* (Q). *Then*

$$fLu = (-1)^n a \det B_1 \det (B_4 - I_{n-m}) u; \tag{5.10}$$

hence $u \not\equiv 0$ is a solution of $Lu = 0$ if and only if $\det B_1 = 0$ *or* $\det (B_4 - I_{n-m}) = 0$.

In the case $m = 0$, where condition (Q) is empty, the following corollary holds:

COROLLARY 2. *Suppose* $\det (a_{ij})_{1 \le i \le n, 1 \le j \le n} \neq 0$ *in* Ω, *and let $u \not\equiv 0$ be a simultaneous eigenfunction of all operators S_{ij} with corresponding eigenvalue matrix B. Then u is a solution of $Lu = 0$ if and only if u solves the "separated" equations* (4.2) *with some nonzero vector $\lambda = (\lambda_1, \cdots, \lambda_n) \in \mathbb{C}^n \backslash \{0\}$. In this case, λ is an arbitrary nonzero fixpoint of the matrix B.*

Proof. For any fixed $\lambda = (\lambda_1, \cdots, \lambda_n) \in \mathbb{C}^n \backslash \{0\}$, the following chain of equivalences holds:

u solves equations (4.2)

$$\overset{\text{(Theorem 3)}}{\Longleftrightarrow} u \text{ satisfies the eigenvalue equations (4.1)}$$

$$\overset{\text{(see (5.3))}}{\Longleftrightarrow} \lambda \text{ is a fixpoint of } B.$$

Now, B has a nonzero fixpoint if and only if $\det (B - I_n) = 0$, and this is true, by Theorem 6 ($m = 0$), if and only if $Lu = 0$. □

COROLLARY 3. *Suppose $0 < m < n$. Let $u \not\equiv 0$ satisfy the eigenvalue problems* (5.1) *with eigenvalue matrix B as in* (5.7). *Moreover, assume $B_2 = 0$ and* $\det B_1 \neq 0$. *If then u is a solution of $Lu = 0$, u also solves the equations* (4.2), *where $\lambda = (\lambda_1, \cdots, \lambda_n) \in \mathbb{C}^n \backslash \{0\}$ is an arbitrary nonzero fixpoint of B.*

Proof. Let u be a solution of $Lu = 0$. Owing to $B_2 = 0$, B satisfies condition (Q), and Theorem 6 then yields $\det (B_4 - I_{n-m}) = 0$. This implies $\det (B - I_n) = \det (B_1 - I_m) \cdot \det (B_4 - I_{n-m}) = 0$, and thus, B has a nonzero fixpoint $\lambda \in \mathbb{C}^n$. The equivalences established in the proof of Corollary 2 then end the proof. □

From Lemma 4 and Theorem 6 we further conclude:

THEOREM 7. *Suppose $0 < m < n$. Let $u \not\equiv 0$ be a simultaneous eigenfunction of all operators S_{ij} whose eigenvalue matrix B, decomposed as in* (5.7), *satisfies condition* (Q). *Then, if $B_2 \neq 0$, u is a solution of $Lu = 0$.*

COROLLARY 4. *Suppose $0 < m < n$. Let $u \not\equiv 0$ be a simultaneous eigenfunction of all S_{ij}. For the corresponding eigenvalue matrix B, decomposed as in* (5.7), *we assume that* (Q) *is satisfied and that B_2 has rank m. Then u is a solution of $Lu = 0$ which also solves the "separated" equations* (4.2) *with $(\lambda_1, \cdots, \lambda_n) \in \mathbb{C}^n \backslash \{0\}$ being an arbitrary nonzero fixpoint of B.*

Proof. $Lu = 0$ follows from Theorem 7, since the assumption on the rank implies $B_2 \neq 0$. Moreover, the assumption $\operatorname{rank} B_2 = m$ is equivalent to the existence of an $(n - m, m)$-matrix C such that $B_2 C = I_m$. Therefore we have

$$B - I_n = \begin{pmatrix} B_1 - I_m & B_2 \\ B_3 & B_4 - I_{n-m} \end{pmatrix} = \begin{pmatrix} B_1 & B_2 \\ B_3 + B_4 C - C & B_4 - I_{n-m} \end{pmatrix} \cdot \begin{pmatrix} I_m & 0 \\ -C & I_{n-m} \end{pmatrix}.$$

Hence, using property (Q) and $Lu = 0$, we obtain

$$\det (B - I_n) = \det B_1 \cdot \det (B_4 - I_{n-m}) = 0.$$

Thus, B possesses nonzero fixpoints, and by the equivalences in the proof of Corollary 2, the proof is finished. □

Remark. Notice that the condition rank $B_2 = m$, implying $2m \leqq n$, will sometimes contradict condition (Q). For instance, in case $2m = n$, (Q) implies $\det B_2 = 0$. But it is easy to find matrices B satisfying both conditions.

6. Transformations of Bäcklund type. For any function u vanishing nowhere on Ω, we now define $v := (v_1, \cdots, v_N)$ by

$$v_i := u^{-1}\partial_i u, \qquad 1 \leqq i \leqq N. \tag{6.1}$$

Moreover, let $\alpha = (\alpha_1, \cdots, \alpha_N)$, $\alpha_i \in \mathbb{N}_0$, be any multiindex with $|\alpha| := \sum_{i=1}^{N} \alpha_i \geqq 1$, and define the nonlinear differential operator N_α for n-tuples $w = (w_1, \cdots, w_N)$ of C^∞-functions w_i on Ω inductively as follows. Denote by δ_i the multiindex which is 1 in position i and zero elsewhere. Then:

$$N_{\delta_i}(w) := w_i, \qquad N_{\alpha+\delta_i}(w) := w_i N_\alpha(w) + \partial_i N_\alpha(w), \qquad 1 \leqq i \leqq N. \tag{6.2}$$

Obviously, N_α is an operator of order $|\alpha| - 1$, and from its definition and (6.1), we obtain

$$\partial^\alpha u = N_\alpha(v) \cdot u, \qquad (\partial^\alpha := \partial_1^{\alpha_1} \cdots \partial_N^{\alpha_N} \text{ for } \alpha = (\alpha_1, \cdots, \alpha_N)). \tag{6.3}$$

Finally, if

$$M_{ij} = \sum_\alpha c_{ij\alpha}(x)\partial^\alpha, \tag{6.4}$$

we put

$$Q_{ij}(w) := \sum_\alpha c_{ij\alpha}(x) N_\alpha(w). \tag{6.5}$$

Now, from Theorem 1 we easily derive:

THEOREM 8. *Assume that the function u, which is nowhere zero on Ω, is a solution of the first order system*

$$\partial_i u = v_i u, \qquad 1 \leqq i \leqq N, \tag{6.6}$$

where $v = (v_1, \cdots, v_N)$ satisfies the nonlinear differential system

$$\sum_{k=1}^{n} \lambda_k Q_{ik}(v) + \sum_{k=1}^{n} \lambda_k a_{ik} = 0, \qquad 1 \leqq i \leqq n \tag{6.7}$$

for some $(\lambda_1, \cdots, \lambda_n) \in \mathbb{C}^n \backslash \{0\}$. Then u also solves the equation $Lu = 0$.

Remark. Using Theorem 3 or Theorem 5 instead of Theorem 1, statements analogous to the one above could be established. For a special case, see [3, Thm. 2].

Systems (6.6), (6.7) may be regarded as Bäcklund-type transformation of the linear G-separable operator L. (See [10, pp. 81, 82].) To make this clearer, we now specify the assumptions on the operators L_{ij}. We assume

$$\{1, \cdots, N\} = E_1 \cup E_2 \cup \cdots \cup E_r \tag{6.8}$$

to be the finest possible partition of $\{1, \cdots, N\}$ into disjoint nonempty subsets E_ν with the following property:

(6.9) *The operators $L_{i1}, \cdots, L_{in}$ depend only on the variables with indices from one set E_{ν_i} $(1 \leqq i \leqq n)$.*

(In cases (i) and (ii) of § 2, E_ν are one-element subsets.) From (6.9) and (6.2)–(6.5), it follows that the ith equation in (6.7) explicitly depends only on the variables x_l, $l \in E_{\nu_i}$,

and contains only functions v_k, $k \in E_{\nu_l}$ and their derivatives with respect to such x_l. The functions v_k, however, may depend on the remaining variables, too. The following theorem states when this does not happen.

THEOREM 9. *Let u be a function nowhere vanishing on Ω and let $v = (v_1, \cdots, v_N)$ be defined by* (6.6). *Then the following statements* (a) *and* (b) *are equivalent*:

(a) *u is a product $u = u_1 u_2 \cdots u_r$ with C^∞-functions u_ν depending only on the variables x_l, $l \in E_\nu$.*

(b) *For all $\nu = 1, \cdots, r$ and all $k \in E_\nu$, the function v_k depends only on the variables x_l, $l \in E_\nu$.*

Proof. The implication (a) $\Rightarrow$ (b) is obvious. Let us prove (b) $\Rightarrow$ (a): For $\nu \neq \mu$ and $k \in E_\nu$, $j \in E_\mu$ we obtain (in the complex case) $\partial_j \partial_k (\ln u) = \partial_j (u^{-1} \partial_k u) = \partial_j v_k = 0$. By standard arguments, this yields

$$\ln u = \sum_{\nu=1}^{r} w_\nu,$$

where w_ν is a C^∞-function depending only on the variables x_l, $l \in E_\nu$. Thus, (a) follows by exponentiation. □

Remark. As is well known, solutions $v_1, \cdots, v_N$ of (6.7) with nontrivial $(\lambda_1, \cdots, \lambda_n)$ yield a solution $u \neq 0$ of (6.6), and hence of (1.1), if and only if $\partial_i v_j = \partial_j v_i$ holds for all $i, j = 1, \cdots, N$.

In the important special case where L is a second order operator separable according to case (i) or (ii) of § 2, the equations (6.7) are first order nonlinear ordinary differential equations, each for a single function v_l only.

In [10, pp. 81, 82], A. C. Scott treats the Bäcklund transform of the linear Klein–Gordon equation in polar coordinates (two space variables, $N = 3$). Here, one of the equations corresponding to (6.7) is a Riccati equation. Scott conjectures what was shown above, namely "that the procedure works for any separable, linear partial differential equation."

7. Examples.

(a) Consider the complex 4-dimensional Helmholtz operator $L := \Delta_4 - \mu$, expressed in the coordinates x defined by (2.4). First, we assume $\mu \neq 0$. Then, from (2.5), we obtain $m = 0$ and $a = \mu(x_2 - x_1)$. By (3.3) we calculate

$$S_{11} = \mu^{-1} \Delta_4 = \mu^{-1} L + 1,$$

$$S_{21} = \frac{x_1 x_2}{x_1 - x_2} ((x_1^{-3} - x_2^{-3}) \partial_3^2 + 2(x_1^{-2} - x_2^{-2}) \partial_3 \partial_4 - x_1^{-1} M_{11} - x_2^{-1} M_{21}),$$

where

$$M_{11} := 4x_1^2 \partial_1^2 + 12 x_1 \partial_1, \qquad M_{21} := -4x_2^2 \partial_2^2 - 12 x_2 \partial_2,$$

$$S_{31} = \partial_3, \quad S_{41} = \partial_3^2, \quad S_{51} = \partial_4, \quad S_{61} = \partial_3 \partial_4, \quad S_{ij} = 0 \quad \text{for } 1 \leqq i \leqq 6, \quad 2 \leqq j \leqq 6.$$

Since $m = 0$, (2.8) and (4.2) now coincide. Looking for solutions $u \neq 0$ of (2.8) with $\lambda = (\lambda_1, \cdots, \lambda_6) \neq 0$, we easily find that λ_1 must be nonzero; hence we assume without restriction $\lambda_1 = 1$. We calculate that (2.8) with $\lambda_1 = 1$ has a solution $u \neq 0$ if and only if $\lambda = (1, \lambda_2, \lambda_3, \lambda_3^2, \lambda_5, \lambda_3 \lambda_5)$ with $\lambda_2, \lambda_3, \lambda_5$ arbitrary complex. Moreover, we easily verify that any such solution u of (2.8) (and hence of (4.1)) is a simultaneous

eigenfunction of all S_{ij} with eigenvalue matrix

$$B = (\beta_{ij}) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \lambda_2 & \cdot & & \cdot \\ \lambda_3 & \cdot & & \cdot \\ \lambda_3^2 & \cdot & & \cdot \\ \lambda_5 & \cdot & & \cdot \\ \lambda_3\lambda_5 & 0 & \cdots & 0 \end{pmatrix}.$$

Such functions actually exist and, by Theorem 1 or Corollary 2, are solutions of $Lu = 0$.

Now assume $\mu = 0$. Then $m = 1$ and $a = \det(a_{ij})_{2\leqq i\leqq 6, 2\leqq j\leqq 6} = 1$. Since condition (v) of § 4 is satisfied, (4.9) yields

$$S_{11} = (x_2 - x_1)\Delta_4.$$

Moreover, we calculate

$$S_{21} = -M_{21} - x_2^{-2}\partial_3^2 - 2x_2^{-1}\partial_3\partial_4,$$

and the remaining S_{ij} turn out to be the same as in case $\mu \neq 0$. Here again, we may assume $\lambda_1 = 1$ in (2.8), which then obviously coincides with (4.7). Solutions $u \neq 0$ of (4.7) are seen to exist if and only if $(\lambda_2, \cdots, \lambda_6) = (\lambda_2, \lambda_3, \lambda_3^2, \lambda_5, \lambda_3\lambda_5)$ with $\lambda_2, \lambda_3, \lambda_5$ arbitrary complex. Any solution u of (4.7) is actually a solution of (5.1) with eigenvalue matrix

$$B = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \lambda_2 & \cdot & & \cdot \\ \lambda_3 & \cdot & & \cdot \\ \lambda_3^2 & \cdot & & \cdot \\ \lambda_5 & \cdot & & \cdot \\ \lambda_3\lambda_5 & 0 & \cdots & 0 \end{pmatrix}$$

and, by Theorem 5(b) or Theorem 6, satisfies $Lu = 0$.

(b) Take again $L := \Delta_4 - \mu$, but in coordinates x given by (2.2). Assume $\mu \neq 0$. From (2.3) we obtain $n = 6$, $m = 0$, $a = -\mu$ and

$$S_{11} = \mu^{-1}(x_1\partial_3^2 + x_2\partial_4^2), \quad S_{21} = x_1\partial_3^2, \quad S_{31} = \partial_3, \quad S_{41} = \partial_3^2,$$
$$S_{51} = \partial_4, \quad S_{61} = \partial_4^2, \quad S_{13} = 2\mu^{-1}\partial_2, \quad S_{15} = 2\mu^{-1}\partial_1, \quad S_{25} = 2\partial_1.$$

The remaining S_{ij} are zero.

Since $m = 0$, (2.8) and (4.2) coincide. Looking for solutions $u \neq 0$ of (2.8) with $\lambda = (\lambda_1, \cdots, \lambda_6) \neq 0$, we necessarily obtain $\lambda_1 \neq 0$, so we may choose $\lambda_1 = 1$. Then (2.8) has solutions $u \neq 0$ if and only if $\lambda = (1, \lambda_2, \lambda_3, \lambda_3^2, \lambda_5, \lambda_5^2)$ with $\lambda_2 \in \mathbb{C}$, λ_3 and $\lambda_5 \in \mathbb{C}\backslash\{0\}$. By Theorem 1, these solutions satisfy $Lu = 0$, but they don't satisfy (5.1), as is easily verified. For solutions $u \neq 0$ of (5.1) with eigenvalue matrix $B = (\beta_{ij})$, the only possible values are: β_{13}, β_{15} arbitrary complex, $\beta_{25} = \mu\beta_{15}$ and else $\beta_{ij} = 0$. Corresponding eigenfunctions are

$$u = u(x_1, x_2) = \text{const} \cdot \exp\frac{\mu}{2}(\beta_{15}x_1 + \beta_{13}x_2).$$

Since $\det(B - I) \neq 0$, we conclude from Theorem 6 that these functions u do not satisfy $Lu = 0$ and a fortiori, (2.8). Also, by Corollary 2, they fail to solve (4.1) with some nonzero λ.

Let us now look at the case $\mu = 0$. From (2.3) we obtain $m = 1$, $a = \det (a_{ij})_{2 \leqq i \leqq 6, 2 \leqq j \leqq 6} = 1$, and

$$S_{11} = -x_1 \partial_3^2 - x_2 \partial_4^2, \quad S_{21} = -x_2 \partial_4^2, \quad S_{13} = S_{23} = -2\partial_2, \quad S_{15} = -2\partial_1, \quad S_{25} = 0.$$

The remaining S_{ij} agree with those obtained in case $\mu \neq 0$.

Concerning solutions $u \neq 0$ of (2.8) with $\lambda \neq 0$, we now obtain the weaker condition:

$$\lambda = (1, \lambda_2, \lambda_3, \lambda_3^2, \lambda_5, \lambda_5^2), \qquad \lambda_2, \lambda_3, \lambda_5 \in \mathbb{C}.$$

By Theorem 1, these solutions satisfy $Lu = 0$ but, as a short calculation shows, they fail to satisfy (4.1) or (4.2).

As for solutions $u \neq 0$ of (5.1), the eigenvalue matrix $B = (\beta_{ij})$ is given by:

$$\beta_{13}, \beta_{15} \in \mathbb{C}, \quad \beta_{23} = \beta_{13}, \quad \text{all other } \beta_{ij} = 0.$$

Corresponding eigenfunctions of S_{ij} are now

$$u = u(x_1, x_2) = \text{const} \cdot \exp\left(-\tfrac{1}{2}(\beta_{15} x_1 + \beta_{13} x_2)\right), \tag{7.1}$$

and these are special solutions of (2.8) with $\lambda = (1, 0, 0, 0, 0, 0)$, the latter being arbitrary functions of x_1, x_2. That $Lu = 0$ is satisfied by (7.1) could also be seen from general results using Lemma 3 and the corresponding Remark: Notice that we have $A_3 = 0$ and $B_3 = 0$.

(c) Consider the operator (2.6), (2.7) as a representative of arbitrary linear homogeneous operators with constant coefficients. In case $c_0 \neq 0$ we have $m = 0$, $a = c_0$, $S_{11} = 1 - c_0^{-1} L$, and $S_{ij} = M_{ij}$ else. If $c_0 = 0$, we obtain $m = 1$, $a = 1$, $S_{11} = -L$, while the remaining S_{ij} coincide with M_{ij} again.

Let us investigate the systems (6.6), (6.7) for this example. With the notation

$$\alpha^2 := (1, 0), \quad \alpha^3 := (1, 1), \quad \alpha^4 := (2, 0), \quad \alpha^5 := (0, 2)$$

and using the definitions from (6.2)–(6.5), we obtain

$$Q_{i1}(w_1, w_2) = N_{\alpha^i}(w_1, w_2), \quad 2 \leqq i \leqq 5, \qquad Q_{ij} = 0 \text{ else.}$$

Case $c_0 \neq 0$. Without restriction, we may choose $\lambda_1 = 1$. Equations (6.7) then read

$$\begin{aligned} c_0 + \lambda_2 c_1 + \lambda_3 c_{12} + \lambda_4 c_{11} + \lambda_5 c_{22} &= 0, \\ N_{\alpha^i}(v_1, v_2) &= \lambda_i, \qquad 2 \leqq i \leqq 5. \end{aligned} \tag{7.2}$$

These equations yield conditions on $\lambda_2, \cdots, \lambda_5$ under which they are easily solved for v_1, v_2. Details are left to the reader.

Case $c_0 = 0$. Since condition (v) of § 4 is satisfied, we may use (4.7) instead of (2.8) (see Remark following Theorem 8). Then $u \neq 0$ is a solution of (4.7) if and only if v_1, v_2, defined by (6.6), satisfy the equations resulting from (7.2) by cancelling c_0.

REFERENCES

[1] F. R. GANTMACHER, *The Theory of Matrices*, vol. 1, Chelsea, New York, 1960.

[2] J. HAINZL, *Separation of variables and commuting operators*, Math. Meth. Appl. Sci., 1 (1979), pp. 468–479.

[3] ———, *Neue Aspekte der Separationsmethode*, Z. Angew. Math. Mech., 60 (1980), pp. T247–T249.

[4] E. G. KALNINS AND W. MILLER, JR., *Nonorthogonal separable coordinate systems for the flat 4-space Helmholtz equation*, J. Phys. A, 12 (1979), pp. 1129–1147.

[5] E. G. KALNINS AND W. MILLER, JR., *Killing tensors and nonorthogonal variable separation for Hamilton-Jacobi equations*, this Journal, 12 (1981), pp. 617–629.

[6] T. H. KOORNWINDER, *Book review: Symmetry and separation of variables*, by W. Miller, Jr., Bull. Amer. Math. Soc. (new series), 1 (1979), pp. 1014–1019.

[7] ———, *A precise definition of separation of variables*, Preprint, Stichting Mathematisch Centrum, Amsterdam, 1979.

[8] W. MILLER, JR., *Symmetry and Separation of Variables*, Addison-Wesley, Reading, MA, 1977.

[9] H.-D. NIESSEN, *Algebraische Untersuchungen über separierbare Operatoren*, Math. Z., 94 (1966), pp. 328–348.

[10] A. C. SCOTT, *The application of Bäcklund transforms to physical problems*, in Bäcklund Transformations, the Inverse Scattering Method, Solitons, and Their Applications, R. M. Miura, ed., Lecture Notes in Mathematics 515, Springer, Berlin–Heidelberg–New York, 1976, pp. 80–105.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0004 $01.00/0

A QUASILINEAR PARABOLIC EQUATION DESCRIBING THE ELONGATION OF THIN FILAMENTS OF POLYMERIC LIQUIDS*

M. RENARDY†

Abstract. We study the equation

$$\rho\ddot{u} = 3\eta\frac{\partial^2}{\partial x\,\partial t}\left(-\frac{1}{u_x}\right)+\frac{\partial}{\partial x}\int_{-\infty}^{t} a(t-s)\left(\frac{u_x(t)}{u_x^2(s)}-\frac{u_x(s)}{u_x^2(t)}\right)ds$$

where $u(x, t)$ is a real-valued function of $x\in[-1, 1]$ and $t\in\mathbb{R}$, with the boundary condition

$$3\eta\frac{\partial}{\partial t}\left(-\frac{1}{u_x}\right)+\int_{-\infty}^{t} a(t-s)\left(\frac{u_x(t)}{u_x^2(s)}-\frac{u_x(s)}{u_x^2(t)}\right)ds = f(t)$$

at $x=\pm1$. This equation is derived as a model for the elongation of thin filaments of polymeric liquids, u denoting the position of a fluid particle in space, a the memory kernel, and f the force acting on the ends of the filament. We study the evolution of u, assuming the initial condition $u(x, t=-\infty)=x$. It is shown that under appropriate conditions on a and f the boundary condition can be uniquely resolved with respect to u_x. The full problem is transformed in such a way that it is approachable by the Sobolevskii theory of quasilinear parabolic equations. This yields the existence of solutions to the initial value problem on sufficiently small time intervals. Moreover, we show that if $f(t)$ converges to zero exponentially as $t\to\pm\infty$ and is small in an appropriate norm, there exists a solution globally in time, which approaches a stationary limit as $t\to+\infty$.

1. Introduction. We study the following problem occurring in polymer processing: A thin filament of a viscoelastic liquid is subjected to a force f acting on its ends as shown in Fig. 1. We investigate the temporal evolution of the displacement. The

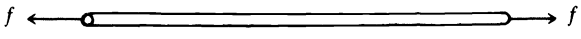

FIG. 1

equations that our analysis is based on involve the "rubberlike liquid" constitutive assumption for the stress–strain law [3] and certain approximations based on the thinness of the filament, which allow the reduction to a spatially one-dimensional problem. Using these assumptions, we shall derive the following equation

$$\rho\ddot{u} = 3\eta\frac{\partial^2}{\partial x\,\partial t}\left(-\frac{1}{u_x}\right)+\frac{\partial}{\partial x}\int_{-\infty}^{t} a(t-s)\left(\frac{u_x(t)}{u_x^2(s)}-\frac{u_x(s)}{u_x^2(t)}\right)ds \tag{1.1}$$

where $u(x, t)$ is a real-valued function of $x\in[-1, 1]$ and $t\in\mathbb{R}$. As usual, a subscript x denotes partial differentiation with respect to x and "dot" denotes partial differentiation with respect to t. The arguments (x, t) are suppressed unless needed for proper understanding. Equation (1.1) is supplemented by the nonlinear Neumann boundary condition

$$3\eta\frac{\partial}{\partial t}\left(-\frac{1}{u_x}\right)+\int_{-\infty}^{t} a(t-s)\left(\frac{u_x(t)}{u_x^2(s)}-\frac{u_x(s)}{u_x^2(t)}\right)ds = f(t) \tag{1.2}$$

at $x=\pm1$.

In these equations $u(x, t)$ denotes the position at time t of a fluid particle, which is at the position x in a certain reference state. This reference state will be identified

* Received by the editors April 20, 1981. This work was supported by the U.S. Army under contract DAAG29-80-C-0041 and by the Deutsche Forschungsgemeinschaft.

† Mathematics Research Center, University of Wisconsin–Madison, Madison, Wisconsin 53706.

with the state of the fluid at $t=-\infty$, i.e., we have $u(x, t=-\infty)=x$. The variable ρ denotes the density of the fluid, η the Newtonian part of the viscosity, and f the force acting on the ends of the filament. The memory kernel $a:[0,\infty)\to\mathbb{R}$ will be assumed to have the following properties, which we shall refer to as assumptions (a):

(i) a has the representation

$$a(t)=\int e^{-\lambda t}\,d\mu(\lambda) \tag{1.3}$$

where μ is a complex valued Borel measure on the complex plane $\mathbb{C}$ such that $1\in L^1(\mu)$ (i.e., 1 is integrable with respect to the total variation of μ), and supp μ is contained in $\{\lambda\in\mathbb{C}\,|\,-\varphi\leqq\arg\lambda\leqq\varphi, |\lambda|\geqq\varepsilon\}$ for some $\varphi<\pi/2$ and $\varepsilon>0$. Since a is real, we may and will assume that $d\mu(\bar{\lambda})=\overline{d\mu(\lambda)}$.

(ii) $a(t)\geqq 0$ for $t\in[0,\infty)$.

(iii) a is monotonically decreasing.

Note that (i) implies in particular that a is continuous and that $|a|$ can be estimated by a decaying exponential. The motivation for assumptions (a) will become apparent later in the paper. The sectorial condition for supp μ is needed to make the problem fit into the theory of parabolic equations; (ii) and (iii) will have important implications for the spectra of certain linear operators. In physical theories derived from "molecular network" or "bead-spring" models (see [4] and the references in [3, Ch. 6]) a turns out to be a finite sum of decaying exponentials. This is clearly a special case of assumptions (a), μ in this case being a finite sum of Dirac measures located on the real axis.

The boundary condition (1.2) agrees precisely with the equation describing the evolution of the length of the filament when inertial forces are neglected. This problem has been discussed previously by Lodge, McLeod and Nohel in [5] and by the author in [7]. Lodge, McLeod and Nohel consider the solution as known for $t<0$ and assume it is nondecreasing. They then assume $f=0$ for $t>0$ and study existence, asymptotic behavior and various monotonicity properties of solutions. In [7] the force f is a given continuous function $\mathbb{R}\to\mathbb{R}$. It is assumed that f converges to zero exponentially as $t\to-\infty$, and that either f converges to 0 exponentially as $t\to+\infty$ and is small in a suitable norm, or the size of f is arbitrary, but $f(t)$ vanishes identically for t greater than some finite t_0. (In the latter case we need the additional assumption that supp μ is on the real axis; in fact, in [7] we assumed that μ was a finite sum of Dirac measures on the real axis, but the same ideas can be applied to the more general situation as we demonstrate below.) In both cases it is proved that, given the initial condition $u_x(t=-\infty)=1$, a unique positive solution exists for all times t, and moreover $\lim_{t\to+\infty} u_x(t)$ exists and is strictly positive. Whereas the arguments in [5] rely mainly on monotonicity properties, the main tools in [7] are the implicit function theorem and the use of Lyapunov functions.

The present paper will be arranged as follows: In § 2 we explain the basic physical laws and the approximations leading to (1.1), (1.2). We start from the basic laws of continuum mechanics, using the "rubberlike liquid" constitutive relation. The equation of motion in the interior of the filament and the boundary conditions on the lateral surface are then solved formally by a power series expansion with respect to a "thinness parameter" in an analogous manner as was done in the theory of thin elastic rods [6]. The first order terms in this expansion lead to (1.1). The formal expansion does not in general fit given boundary conditions at the ends of the filament, and one is confronted with a "boundary layer" problem. Since we are only interested in a first

order approximation, we shall not deal with this situation here. Instead, we consider the balance of force, taking into account only terms not involving the small parameter. This leads to (1.2). Section 3 summarizes the results of [7] concerning (1.2) as explained above, taking into account the modifications required by the more general assumptions on a. As a result, we may subsequently consider u_x as being given on the boundary. In §§ 4 and 5 we finally deal with the full problem (1.1), (1.2). Using (1.3), this problem is transformed in such a way that it fits into the abstract theory of quasilinear parabolic equations introduced by Sobolevskii [2], [8]. An "initial condition" in the evolution problem so defined will not necessarily involve the whole history of u, but only certain of its moments; the choice of which depends on the support of μ. In § 4 we shall explain this transformation and as a consequence of the Sobolevskii theory obtain the existence and uniqueness of solutions to the initial value problem locally in time. Section 5 deals with the case where f converges to zero exponentially as $t \to \pm\infty$ and is small. We assume the filament is undeformed ($u = x$) at $t = -\infty$. It will be shown that a solution of the full problem exists globally in time, which approaches a stationary limit as $t \to +\infty$.

2. Derivation of the basic equations. We assign to each point in the fluid two different sets of coordinates. By $(\zeta^1, \zeta^2, \zeta^3)$ we denote "body coordinates", i.e., coordinates labelling a specific particle in the fluid. These coordinates can be identified with the position of the particle in space, when the fluid is in a certain "reference state". (It will later be convenient to take as a reference state the state of the fluid at time $t = -\infty$). On the other hand (y^1, y^2, y^3) will denote coordinates labelling a point in space. We are interested in finding trajectories of fluid particles, i.e., a functional dependence $y^i(\zeta^1, \zeta^2, \zeta^3, t)$. In our exposition of the equations describing this functional dependence we follow Lodge [3]. (For a summary of the relevant equations, see [3, pp. 206–207].)

To each point $(\zeta^1, \zeta^2, \zeta^3)$ there is assigned a body metric tensor γ and a body stress tensor π. γ is defined by the relation

$$\gamma_{ij} = \frac{\partial y^r}{\partial \zeta^i} \cdot \frac{\partial y^r}{\partial \zeta^j}.$$

π is related to γ by a constitutive law which expresses the specific properties of the material. We use the "rubberlike liquid" constitutive relation ([3, p. 143]):

$$\pi^{ij} + p\gamma^{ij} = -\eta \frac{\partial \gamma^{ij}}{\partial t} + \int_{-\infty}^{t} a(t-s)\gamma^{ij}(s)\, ds,$$

where γ^{ij} denote the components of γ^{-1}, i.e., $\gamma^{ij}\gamma_{jk} = \delta^i_k$; η is a positive material constant called the viscosity, and a is a given function, which will always be assumed to satisfy assumptions (a) stated in the introduction. The variable p is an unknown having the physical significance of a pressure. The introduction of this variable is necessary, since we assume the fluid is incompressible; i.e.,

$$\det \gamma = 1. \tag{2.1}$$

The evolution of y^k is determined by Newton's law, which for a Cartesian space coordinate system takes the form

$$\rho \ddot{y}^k = \frac{\partial y^k}{\partial \zeta^i} \left\{ \frac{\partial}{\partial \zeta^s} \pi^{si} + \Gamma^i_{rs} \pi^{rs} \right\}. \tag{2.2}$$

The Γ^i_{rs} denote the Riemann–Christoffel symbols associated with the metric tensor γ:

$$\Gamma^i_{rs}=\frac{1}{2}\gamma^{ij}\left\{\frac{\partial\gamma_{jr}}{\partial\zeta^s}+\frac{\partial\gamma_{js}}{\partial\zeta^r}-\frac{\partial\gamma_{rs}}{\partial\zeta^j}\right\}.$$

Equations (2.1) and (2.2) have to be supplemented by boundary conditions referring to either the displacement or the stresses on the boundary of the liquid. We shall here deal with stress conditions. Let Y^k denote the components of surface traction referred to space coordinates. Then the boundary conditions on a surface $\zeta^l=\text{const.}$ are given by

$$\pi^{il}\frac{\partial y^k}{\partial\zeta^i}(\gamma^{ll})^{-1/2}=Y^k. \tag{2.3}$$

In the problem of the elongated filament, the surface traction on the lateral surface is zero, whereas at the ends there is a longitudinal surface traction equal to f divided by the cross-sectional area of the filament. For convenience, we let ζ^1 and y^1, respectively, denote the coordinate in the direction of the filament and ζ^2, ζ^3 and y^2, y^3, respectively, the transversal coordinates. It is assumed that in the undeformed reference state (i.e., at $t=-\infty$) the filament is cylindrical and axi-symmetric, i.e., in this state we have $y^i=\zeta^i$, where ζ^1 (by appropriate normalization of length scale) ranges from -1 to 1, and $r=\sqrt{(\zeta^2)^2+(\zeta^3)^2}$ ranges from 0 to δ, the radius of the filament. Then, for small δ, equations (2.1) and (2.2) and the boundary conditions on the lateral surface can formally be solved by a series expansion in powers of r and δ. This expansion is analogous to that used by Nariboli [6] for the problem of longitudinal elastic waves in a thin rod.

We put $\hat\zeta^2=\zeta^2/\delta$, $\hat\zeta^3=\zeta^3/\delta$, $\hat r=r/\delta$, so that the lateral surface now corresponds to $\hat r=1$. We make the following ansatz:

$$y^1(\zeta^1,\delta\hat\zeta^2,\delta\hat\zeta^3)=\sum_{\nu=0}^{\infty}\delta^{2\nu}P_\nu(\zeta^1,\hat r^2),$$

$$y^2(\zeta^1,\delta\hat\zeta^2,\delta\hat\zeta^3)=\delta\hat\zeta^2\sum_{\nu=0}^{\infty}\delta^{2\nu}Q_\nu(\zeta^1,\hat r^2),$$

$$y^3(\zeta^1,\delta\hat\zeta^2,\delta\hat\zeta^3)=\delta\hat\zeta^3\sum_{\nu=0}^{\infty}\delta^{2\nu}Q_\nu(\zeta^1,\hat r^2),$$

$$p(\zeta^1,\delta\hat\zeta^2,\delta\hat\zeta^3)=\sum_{\nu=0}^{\infty}\delta^{2\nu}R_\nu(\zeta^1,\hat r^2),$$

where P_ν, Q_ν, R_ν are polynomials of νth degree in $\hat r^2$. This ansatz is inserted into (2.1), (2.2) and the lateral boundary conditions, which are supposed to be satisfied for all values of δ. Formally, this yields an infinite set of equations for the coefficients of P_ν, Q_ν and R_ν. We are only interested in deriving an equation for the first term $P_0(\zeta^1)$, and we shall in the following only carry out the series expansion as far as needed for this purpose.

When terms up to $O(\delta)$ are taken into account, we find for the metric tensor

$$\gamma=\begin{pmatrix}(\partial P_0/\partial\zeta^1)^2 & \delta\hat\zeta^2\cdot\varphi & \delta\hat\zeta^3\cdot\varphi\\ \delta\hat\zeta^2\cdot\varphi & Q_0^2 & 0\\ \delta\hat\zeta^3\cdot\varphi & 0 & Q_0^2\end{pmatrix},$$

where $\varphi=\partial P_0/\partial\zeta^1\cdot(\partial P_2/\partial\hat r)\hat r^{-1}+(\partial Q_0/\partial\zeta^1)Q_0$.

Putting $\delta = 0$, we find from (2.1)

$$\left(\frac{\partial P_0}{\partial \zeta^1}\right)^2 Q_0^4 = 1. \tag{2.4}$$

Next we consider the boundary conditions on the lateral surface. At a boundary point where $\hat{\zeta}^3 = 0$, these yield the equations $\pi^{21} = \pi^{22} = \pi^{23} = 0$. (Because of the radial symmetry it suffices to consider these boundary points; if the traction vanishes there, it does so everywhere else.) The quantity π^{23} vanishes identically as a result of the radial symmetry. For π^{22} we obtain the following terms of the order $O(1)$

$$\pi^{22} = -R_0 Q_0^{-2} - \eta \frac{\partial}{\partial t}(Q_0^{-2}) + \int_{-\infty}^{t} a(t-s) Q_0^{-2}(s)\, ds = 0. \tag{2.5}$$

Finally we have

$$\pi^{21} = -p\gamma^{21} - \eta \frac{\partial}{\partial t}(\gamma^{21}) + \int_{-\infty}^{t} a(t-s)\gamma^{21}(s)\, ds = 0. \tag{2.6}$$

All solutions we are going to consider shall satisfy $\lim_{t\to-\infty} p(t) = \int_{-\infty}^{0} a(-s)\, ds$ and $\lim_{t\to-\infty} \gamma^{21}(t) = 0$, the convergence being exponential. If (a) holds, it is then not difficult to prove that the only solution to (2.6) satisfying the specified conditions is $\gamma^{21} = 0$. In the first order in δ this yields $\varphi = 0$. The law of motion (2.2) now yields the following equation for $P_0 (\gamma^{11} = (\partial P_0/\partial \zeta^1)^{-2})$:

$$\begin{aligned}
\rho \ddot{P}_0 &= \frac{\partial P_0}{\partial \zeta^1}\left\{\frac{\partial}{\partial \zeta^1}\pi^{11} + \Gamma^1_{11}\pi^{11}\right\} \\
&= \frac{\partial P_0}{\partial \zeta^1}\left\{\frac{\partial}{\partial \zeta^1}\left(-R_0\left(\frac{\partial P_0}{\partial \zeta^1}\right)^{-2} - \eta\frac{\partial}{\partial t}\left(\frac{\partial P_0}{\partial \zeta^1}\right)^{-2} + \int_{-\infty}^{t} a(t-s)\left(\frac{\partial P_0}{\partial \zeta^1}\right)^{-2}(s)\, ds\right)\right. \\
&\qquad \left. + \left(\frac{\partial P_0}{\partial \zeta^1}\right)^{-1} \frac{\partial^2 P_0}{(\partial \zeta^1)^2}\left(-R_0\left(\frac{\partial P_0}{\partial \zeta^1}\right)^{-2} - \eta\frac{\partial}{\partial t}\left(\frac{\partial P_0}{\partial \zeta^1}\right)^{-2} + \int_{-\infty}^{t} a(t-s)\left(\frac{\partial P_0}{\partial \zeta^1}\right)^{-2}(s)\, ds\right)\right\}.
\end{aligned}$$

In order to simplify notation, we shall henceforth write u for P_0 and x for ζ^1. The last equation now yields (1.1) after a few manipulations, when (2.5) and (2.4) are used to express R_0 in terms of u.

Finally, we have to specify boundary conditions at the ends of the filament. As noted in [6], the asymptotic expansion which we used for the interior problem generally fails near the ends, and a "boundary layer" has to be taken into account. The boundary layer is discussed in a forthcoming paper by Reiss, which is summarized in [1], but not published yet. We are here only concerned with a first order approximation, and we shall ignore boundary layer effects. Instead, we take care of the force balance in the zeroth order with respect to δ. Namely, if one formally inserts our expansion into the boundary conditions at the ends, it is seen that all traction components transversal to the direction of the filament are $O(\delta)$. The longitudinal traction component gives the following terms of order $O(1)$:

$$\tau = \pi^{11}\frac{\partial P_0}{\partial \zeta^1}(\gamma^{11})^{-1/2} = \pi^{11}\left(\frac{\partial P_0}{\partial \zeta^1}\right)^2 = 3\eta\frac{\dot{u}_x}{u_x} + \int_{-\infty}^{t} a(t-s)\left(\frac{u_x^2(t)}{u_x^2(s)} - \frac{u_x(s)}{u_x(t)}\right) ds.$$

Since the cross-sectional area of the filament is in first approximation equal to u_x^{-1}, we shall require that $\tau = f \cdot u_x$. This yields (1.2).

3. The boundary problem. In this section we consider the problem of solving (1.2) for u_x, when f is given. The results we present slightly generalize those of [7], allowing for the more general class of kernels a satisfying assumptions (a). In order to simplify notation, we write y for u_x. Instead of (1.2) we study the slightly more general problem

$$3\eta\dot{y}+\int_{-\infty}^{t} a(t-s)\left(\frac{y^3(t)}{y^2(s)}-y(s)\right)ds=f(t)y^{\alpha} \tag{3.1}$$

where $0<\alpha<3$. Equation (1.2) corresponds to $\alpha=2$, and, as explained in [7], the case $\alpha=\frac{1}{2}$ is also physically interesting: namely, it describes the deformation of a sheet of the polymer, when inertia is neglected, as shown in Fig. 2.

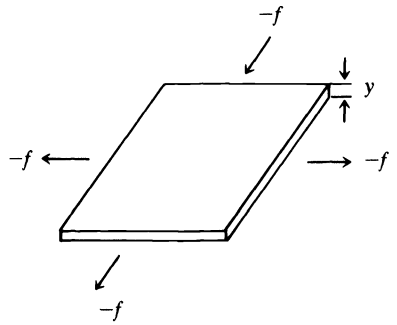

FIG. 2

We put $g(\lambda)=\int_{-\infty}^{t} e^{-\lambda(t-s)}y^{-2}(s)\,ds$, $h(\lambda)=\int_{-\infty}^{t} e^{-\lambda(t-s)}y(s)\,ds$,

$$\gamma(\lambda)=g(\lambda)y^2, \qquad \delta(\lambda)=h(\lambda)y^{-1}.$$

Then (3.1) is equivalent to either of the following systems:

$$\begin{aligned} 3\eta\dot{y}&=\int (h(\lambda)-g(\lambda)y^3)\,d\mu(\lambda)+fy^{\alpha},\\ \dot{g}(\lambda)&=-\lambda g(\lambda)+y^{-2},\\ \dot{h}(\lambda)&=-\lambda h(\lambda)+y; \end{aligned} \tag{3.2}$$

$$3\eta\dot{y}=y\cdot\int (\delta(\lambda)-\gamma(\lambda))\,d\mu(\lambda)+fy^{\alpha},$$

$$\dot{\gamma}(\lambda)=-\lambda\gamma(\lambda)+1-\frac{2}{3\eta}\gamma(\lambda)\int (\gamma(\lambda)-\delta(\lambda))\,d\mu(\lambda)+\frac{2}{3\eta}\gamma(\lambda)fy^{\alpha-1}, \tag{3.3}$$

$$\dot{\delta}(\lambda)=-\lambda\delta(\lambda)+1+\frac{1}{3\eta}\delta(\lambda)\int (\gamma(\lambda)-\delta(\lambda))\,d\mu(\lambda)-\frac{1}{3\eta}\delta(\lambda)fy^{\alpha-1}.$$

Both forms will be used in the following. Equations (3.2) or (3.3) will be regarded as evolution problems in the space $X=\mathbb{R}\times(L^s(\mu))^2$, $1\leqq s<\infty$. Here $L^s(\mu)$ denotes the space of all (equivalence classes of) functions $g:\mathbb{C}\to\mathbb{C}$ such that $g(\bar{z})=\overline{g(z)}$ and $|g|^s$ is integrable with respect to the total variation of μ. Clearly, the right side of (3.2) or (3.3) is the sum of an analytic generator and a smooth nonlinear term.

A trivial solution for $f=0$ is given by $y=1$, $g(\lambda)=h(\lambda)=\gamma(\lambda)=\delta(\lambda)=1/\lambda$, and we are interested in solutions converging to this trivial solution as $t\to-\infty$. As a first step we investigate the spectral properties of the linearization of (3.2) (of course (3.3) gives the same result) at this point, i.e., we study the inhomogeneous linear equation

$$
\begin{aligned}
&3\eta\beta y-\int (h(\lambda)-g(\lambda))\,d\mu(\lambda)+3y\int\frac{1}{\lambda}\,d\mu(\lambda)=3\eta\varphi,\\
&\beta g(\lambda)+\lambda g(\lambda)+2y=\psi_1(\lambda),\qquad (3.4)\\
&\beta h(\lambda)+\lambda h(\lambda)-y=\psi_2(\lambda).
\end{aligned}
$$

If $-\beta$ is not in the support of μ, the last two equations can be resolved with respect to $g(\lambda)$ and $h(\lambda)$. This inserted into the first equation of (3.4) yields

$$
3\eta\beta y-3y\int\frac{1}{\lambda+\beta}\,d\mu(\lambda)+3y\int\frac{1}{\lambda}\,d\mu(\lambda)=3\eta\varphi+\int\frac{\psi_2(\lambda)-\psi_1(\lambda)}{\lambda+\beta}\,d\mu(\lambda).
$$

Hence the resolvent exists at β, if and only if $-\beta$ is not in the support of μ and

$$
\rho(\beta):=3\eta\beta-3\int\left(\frac{1}{\lambda+\beta}-\frac{1}{\lambda}\right)d\mu(\lambda)\neq 0.
$$

Clearly, $\rho(\beta)$ vanishes for $\beta=0$. Namely, we have

$$
\rho(\beta)=3\beta\left(\eta+\int\frac{1}{\lambda(\lambda+\beta)}\,d\mu(\lambda)\right).
$$

Using the relationship between μ and the kernel a, we find

$$
\int\frac{1}{\lambda(\lambda+\beta)}\,d\mu(\lambda)=\int_0^\infty a(t)\frac{1-e^{-\beta t}}{\beta}\,dt.
$$

For $\beta\neq 0$, the real part of this expression is given by

$$
\int_0^\infty\frac{a(t)}{|\beta|^2}\{\operatorname{Re}\beta(1-e^{-t\operatorname{Re}\beta}\cos(t\operatorname{Im}\beta))+\operatorname{Im}\beta\, e^{-t\operatorname{Re}\beta}\sin(t\operatorname{Im}\beta)\}\,dt.
$$

If $\operatorname{Re}\beta\geqq 0$, condition (a) (ii) implies that the first contribution is positive, and condition (a) (iii) implies that the second contribution is positive, too. Hence $\operatorname{Re}\int 1/(\lambda(\lambda+\beta))\,d\mu(\lambda)\geqq 0$, whence certainly $\rho(\beta)\neq 0$.

For easier reference, let us put $Y=(y,g,h)\in\mathbb{R}\times(L^s(\mu))^2$ in (3.2) and write (3.2) in the form

$$
\dot{Y}=L(Y-Y_0)+N(Y-Y_0,f)\qquad (3.5)
$$

where L denotes the linearization of the right side at the trivial solution $Y_0=(1,1/\lambda,1/\lambda)$. Analogously, we put $Y'=(y,\gamma,\delta)$ and write (3.3) in the form

$$
\dot{Y}'=L'(Y'-Y_0)+N'(Y'-Y_0,f).\qquad (3.6)
$$

We have just proved

PROPOSITION 3.1. *The spectrum of L (or L') consists of the algebraically simple eigenvalue 0 (geometric simplicity is immediate, and algebraic simplicity follows from the fact that the resolvent has a first order pole) and a remainder contained in the left half-plane. Moreover, the restriction of L to the range of L generates an analytic semigroup of negative type.*

Before we can state our theorems, we must first define some spaces of functions.

DEFINITION 3.2. Let Z be a Banach space and σ a positive real number. Then

$$X_n^\sigma(Z) := \left\{ v \in C^n(\mathbb{R}, Z) \;\middle|\; \lim_{t\to\pm\infty} e^{\sigma|t|}\|v^{(k)}(t)\| = 0 \text{ for } k = 0, 1, \cdots, n, \right.$$
$$\left. v^{(k)} \text{ denoting the } k\text{th derivative} \right\},$$

$$Y_n^\sigma(Z) = \left\{ v \in C^n(\mathbb{R}, Z) \lim_{t\to\pm\infty} e^{\sigma|t|}\|v^{(k)}(t)\| = 0 \text{ for } k = 1, 2, \cdots, n, \right.$$
$$\left. \lim_{t\to-\infty} e^{-\sigma t}\|v(t)\| = 0, \lim_{t\to\infty} v(t) =: v(\infty) \text{ exists and } \lim_{t\to+\infty} e^{\sigma t}\|v(t) - v(\infty)\| = 0 \right\}.$$

A natural norm in X_n^σ is

$$\|v\| = \sum_{k=0}^{n} \sup_{t\in\mathbb{R}} e^{\sigma|t|}\|v^{(k)}(t)\|.$$

A natural norm in Y_n^σ is

$$\|v\| = \sum_{k=1}^{n} \sup_{t\in\mathbb{R}} e^{\sigma|t|}\|v^{(k)}(t)\| + \sup_{t\leqq 0} e^{-\sigma t}\|v(t)\| + \|v(\infty)\| + \sup_{t\geqq 0} e^{\sigma t}\|v(t) - v(\infty)\|.$$

THEOREM 3.3. *Let $\sigma > 0$ be small enough. Then the following holds: If $f \in X_n^\sigma(\mathbb{R})$ has sufficiently small norm,* (3.6) *has a unique solution Y' satisfying $\tilde{Y} = Y' - Y_0 \in Y_n^\sigma(\mathbb{R}) \times (X_n^\sigma(L^s(\mu)))^2$. Y' depends smoothly on f.*

Proof. We rewrite (3.6) in the form

$$G(\tilde{Y}, f) = \tilde{Y} - \left(\frac{d}{dt} - L'\right)^{-1} N'(\tilde{Y}, f) = 0. \tag{3.7}$$

It is a consequence of Proposition 3.1 that $(d/dt - L')^{-1}$ maps $X_n^\sigma(\mathbb{R}) \times X_n^\sigma(L^s(\mu))^2$ into $Z_n := Y_n^\sigma(\mathbb{R}) \times (X_n^\sigma(L^s(\mu)))^2$. Hence G is a smooth mapping from $Z_n \times X_n^\sigma(\mathbb{R})$ into Z_n and we have $D_{\tilde{Y}}G(0, 0) = \mathrm{id}$. By the implicit function theorem, (3.7) can therefore be resolved with respect to $\tilde{Y}$ in a sufficiently small neighborhood of $(0, 0)$.

$\tilde{Y}$ is clearly unique within that neighborhood. We want to show that it is in fact unique within the class of all functions converging to zero as $t \to -\infty$. To see this, let us first consider functions $\tilde{Y}$ satisfying $\lim_{t\to-\infty} e^{-\sigma' t}\|\tilde{Y}\| = 0$ for some σ' between 0 and σ. If $\tilde{Y}$ is such a function, then certainly $e^{-\sigma' t}\|\tilde{Y}(t)\|$ is smaller than ε on some interval $(-\infty, t_1]$. We can now apply an analogous implicit function argument as above, but rather than considering functions on all of $\mathbb{R}$, we consider only functions on $(-\infty, t_1]$. From this we see that $\tilde{Y}$ is unique in the class of all functions that approach zero exponentially as $t \to -\infty$. Finally, if we assume $\tilde{Y}$ converges to zero at all, it can be seen from the last two equations of (3.3) that $\gamma - 1/\lambda$ and $\delta - 1/\lambda$ converge to zero exponentially, because if only these two equations are considered, the zero eigenvalue in the linearization does not occur. From the first equation of (3.3) we find that $\dot{y}$ converges to zero exponentially, and hence the convergence of y to its limit has to be exponential, too. □

If further restrictions are made on μ, a global result can be proved that does not rely on the smallness of f.

THEOREM 3.4. *In addition to* (a), *assume* supp μ *is contained in the real axis and μ is positive real. Let $f: \mathbb{R} \to \mathbb{R}$ be continuous and such that $\lim_{t\to-\infty} e^{-\sigma t} f(t) = 0$ for some $\sigma > 0$ and $f(t) = 0$ for $t \geqq t_0$. For any such f, equation* (3.3) *has a unique solution*

satisfying $\lim_{t\to-\infty} Y'(t) = Y_0$. *This solution exists globally in time; moreover,* $\lim_{t\to+\infty} Y'(t) = (y(\infty), 1/\lambda, 1/\lambda)$ *exists and* $y(\infty) > 0$.

Proof. From the arguments in the proof of the last theorem we already know the existence and uniqueness of a solution on some interval $(-\infty, t_1]$. In order to prove that the solution exists globally in time, it is more convenient to look at (3.2) rather than at the equivalent equation (3.3). Solutions of (3.2) continue to exist as long as y stays away from zero or infinity. From the second and third equation of (3.2) one obtains positive lower bounds for $\int g(\lambda)\, d\mu(\lambda)$ and $\int h(\lambda)\, d\mu(\lambda)$ in every finite time interval, provided that y remains positive, and these bounds do not depend on any estimate for y. Hence, if y becomes too large, $y^3 \cdot \int g(\lambda)\, d\mu(\lambda)$ will dominate over fy^α and also over $\int h(\lambda)\, d\mu(\lambda)$ (the latter being less than some constant times $\max_{\tau\in(-\infty,t]} y(\tau)$). Analogously, if y becomes too small $\int h(\lambda)\, d\mu(\lambda)$ will be the dominant term. It is immediate from this that y cannot go to zero or infinity in finite time, and therefore the solution exists globally.

For $t > t_0$, we now have $f = 0$, and, putting $\alpha(\lambda) = \gamma(\lambda) - 1/\lambda$, $\beta(\lambda) = \delta(\lambda) - 1/\lambda$, we find from (3.3)

$$\int \left\{ \frac{3}{2}\eta \frac{\alpha\dot{\alpha}}{\alpha + \frac{1}{\lambda}} + 3\eta \frac{\beta\dot{\beta}}{\beta + \frac{1}{\lambda}} \right\} d\mu(\lambda) = -\int \left\{ \frac{3}{2}\eta\lambda \frac{\alpha^2}{\alpha + \frac{1}{\lambda}} + 3\eta\lambda \frac{\beta^2}{\beta + \frac{1}{\lambda}} \right\} d\mu(\lambda) - \left\{ \int (\alpha(\lambda) - \beta(\lambda))\, d\mu(\lambda) \right\}^2. \tag{3.8}$$

As we know that $\gamma(\lambda)$ and $\delta(\lambda)$ stay positive, the denominators $\alpha + 1/\lambda$ and $\beta + 1/\lambda$ are always positive, and the left side of (3.8) is therefore the derivative of a positive function that decreases along trajectories. (It is easy to prove that α and β are nice enough for all the integrals to make sense; namely, one sees from (3.2) that $\lambda g(\lambda)$ and $\lambda h(\lambda)$ and hence $\lambda\gamma(\lambda)$ and $\lambda\delta(\lambda)$ are bounded.) As a consequence, α and β converge to zero exponentially as $t \to \infty$ in the L^2-norm and a fortiori in the L^1-norm. From the first equation of (3.3) one sees then that $\dot{y}$ converges to zero exponentially, whence y must converge to a limit exponentially. Moreover, one easily concludes from the second and third equations of (3.3) that α and β in fact converge to zero in the L^∞-norm and not only in the L^2-norm. This concludes the proof. □

Remark. It is almost trivial to prove [7] that $y(\infty) > 1$ if $f > 0$ and $y(\infty) < 1$ if $f < 0$. Since the equation under study describes the evolution of the length of the filament, if inertia is neglected, this is a result that one would obviously expect. We have no analogue yet for the full problem (1.1), (1.2).

4. Local time existence. We now turn to the study of (1.1). According to what we have seen in the last section, we consider $u_x(t) = b(t) > 0$ as being given at $x = \pm 1$, where b is a smooth function of t. We want to reformulate (1.1) in such a way that it fits into the theory of quasilinear parabolic equations. For this purpose we make the following substitutions

$$p = u_x,$$

$$q = u_{xx},$$

$$r = \dot{u},$$

$$g_1(\lambda) = \int_{-\infty}^{t} e^{-\lambda(t-s)} (u_x(s) - u_x(t))\, ds,$$

$$g_2(\lambda)=\int_{-\infty}^{t} e^{-\lambda(t-s)}\left(\frac{u_{xx}(s)}{u_x^3(s)}-\frac{u_{xx}(t)u_x(s)}{u_x^4(t)}\right)ds,$$

$$g_3(\lambda)=\int_{-\infty}^{t} e^{-\lambda(t-s)}(u_x^{-2}(s)-u_x^{-2}(t))\,ds,$$

$$g_4(\lambda)=\int_{-\infty}^{t} e^{-\lambda(t-s)}\left(u_{xx}(s)-\frac{u_{xx}(t)u_x^2(t)}{u_x^2(s)}\right)ds.$$

Equation (1.1) now assumes the following form:

$$\dot p = r_x,$$

$$\dot q = r_{xx},$$

$$\rho\dot r = 3\eta p^{-2}r_{xx}-6\eta p^{-3}qr_x-2p\int g_2(\lambda)\,d\mu(\lambda)-\frac{1}{p^2}\int g_4(\lambda)\,d\mu(\lambda),$$

(4.1)
$$\dot g_1(\lambda)=-\lambda g_1(\lambda)-\frac{r_x}{\lambda},$$

$$\dot g_2(\lambda)=-\lambda g_2(\lambda)-\frac{r_{xx}}{p^4}\left(g_1(\lambda)+\frac{p}{\lambda}\right)+\frac{4r_xq}{p^5}\left(g_1(\lambda)+\frac{p}{\lambda}\right),$$

$$\dot g_3(\lambda)=-\lambda g_3(\lambda)+\frac{2r_x}{\lambda p^3},$$

$$\dot g_4(\lambda)=-\lambda g_4(\lambda)-r_{xx}p^2\left(g_3(\lambda)+\frac{1}{\lambda p^2}\right)-2r_xqp\left(g_3(\lambda)+\frac{1}{\lambda p^2}\right),$$

with boundary condition $p=b(t)$, $r_x=\dot b(t)$ at $x=\pm1$.

Since $\dot p = r_x$, the first boundary condition follows from the second, once it is satisfied initially, and we shall ignore it.

We will show that (4.1) can be treated by the Sobolevskii theory. For this we first introduce some notations. H^k will denote Sobolev spaces of functions on $[-1, 1]$, and $L^s(\mu, H^k)$ will denote the space of H^k-valued functions defined on $\mathbb{C}$, which are s-integrable with respect to the total variation of μ in the Bochner sense (for a precise definition, see, e.g., [10]). We put $X_s = H^2\times(H^1)^2\times(L^s(\mu, H^1))^4$. Moreover, in (4.1) we substitute $\hat r = r-\dot b(t)x$ and introduce the abbreviation $y=(p, q, \hat r, g_1, g_2, g_3, g_4)$. We rewrite (4.1) in the form

(4.2)
$$\dot y = A(y)y+f(y, t)$$

where $A(y)$ is defined as the following linear operator:

$$A(y)y' = \left(\hat r_x', \hat r_{xx}', \frac{1}{\rho}(3\eta p^{-2}\hat r_{xx}'-6\eta p^{-3}q\hat r_x'), -\lambda g_1'(\lambda)-\frac{1}{\lambda}\hat r_x',\right.$$

$$-\lambda g_2'(\lambda)-\left(g_1(\lambda)+\frac{p}{\lambda}\right)\frac{\hat r_{xx}'}{p^4}+\frac{4q}{p^5}\left(g_1(\lambda)+\frac{p}{\lambda}\right)\hat r_x',$$

$$\left.-\lambda g_3'(\lambda)+\frac{2}{\lambda p^3}\hat r_x', -\lambda g_4'(\lambda)-p^2\left(g_3(\lambda)+\frac{1}{\lambda p^2}\right)\hat r_{xx}'-2qp\left(g_3(\lambda)+\frac{1}{\lambda p^2}\right)\hat r_x'\right)$$

with the boundary conditions $\hat r_x'=0$ at $x=\pm1$.

We shall show that (4.2) satisfies all the requirements of the Sobolevskii theory when regarded as an evolution problem in $X_s (1 \leqq s < \infty)$. More precisely, we shall prove:

THEOREM 4.1. *Let* $1 \leqq s < \infty$ *be arbitrary. Let* $y_0 = (p_0, q_0, \hat{r}_0, g_{i,0}) \in X_s$ *be given such that* $\hat{r}_0 \in H^3$, $\hat{r}_{0,x} = 0$ *at* $x = \pm 1$, $\lambda g_{i,0} \in L^s(\mu, H^1)$ *and* $\min_{x \in [-1,1]} p_0(x) > 0$. *Then, for some* $T > 0$, *equation* (4.2) *has a unique solution* $y \in C^1([0, T], X_s)$ *such that* $y(0) = y_0$.

Proof. We shall deduce the result from [2, Thm. 16.2] ([8, Thm. 7] resp.). For this we have to verify the following conditions stated in [2] as (F1) and (F3)–(F5).

(F1) The operator $A_0 = A(y_0)$ is densely defined, closed and generates an analytic semigroup.

(F3) For v, w in a neighborhood of y_0 in X_s there exists an appropriate $\beta \in \mathbb{C}$ such that

$$\|(A(v) - A(w))(A(w) + \beta)^{-1}\| \leqq C\|v - w\|$$

with some constant c independent of v and w.

(F4) For v, w in a neighborhood of y_0 and $t, \tau \in [0, T]$ there is some constant c such that

$$\|f(v, t) - f(w, \tau)\| \leqq C(|t - \tau| + \|v - w\|).$$

(F5) $y_0 \in D(A_0)$.

(The conditions in [2] are more general, and we have only formulated the special case applying to our problem.)

(F4) and (F5) are trivial consequences of the smoothness of b and our assumptions on the initial data. (F3) is clear, if it is proved that the H^3-norm of the $\hat{r}$-component of $(A(w) + \beta)^{-1} y$ can be estimated by $\|y\|$. This will be immediate from the arguments leading to (F1) with y_0 replaced by w.

To prove (F1), consider the equation $(A_0 + \beta) y = y'$. In the $\hat{r}$-component this leads to

$$\frac{3\eta}{\rho} p_0^{-2} \hat{r}_{xx} - 6\eta p_0^{-3} q_0 \hat{r}_x + \beta \hat{r} = \hat{r}'$$

and the equations for the other components can be trivially resolved once $\hat{r}$ is known. It is now a simple consequence of Theorem 19.2 in [2] (which is from Agmon and Nirenberg [9]) that if β is in a sector not containing the positive real axis, and $|\beta|$ is large enough, we have an estimate of the form

$$\|\hat{r}\|_{H^3} + |\beta|^{1/2} \|\hat{r}\|_{H^2} + |\beta| \|\hat{r}\|_{H^1} \leqq C\|\hat{r}'\|_{H^1}.$$

This concludes the proof. □

5. Solutions for small forces. The goal of the present chapter is to establish an analogue of Theorem 3.3 for the equation (1.1), i.e., to prove existence of solutions globally in time for small forces f. $f = 0$ now corresponds to the boundary condition $u_x(t) = b(t) = 1 \Rightarrow \dot{b}(t) = 0$. In this case (4.1) has the trivial solution $p = 1$, $q = 0$, $r = 0$, $g_i(\lambda) = 0$. As a first step we shall study the linearization of (4.1) at this trivial solution with homogeneous boundary conditions $r_x = 0$. The linearized equation reads as

follows:

$$\dot{p} = r_x, \qquad \dot{q} = r_{xx},$$

$$\dot{r} = \frac{3\eta}{\rho} r_{xx} - \frac{2}{\rho}\int g_2(\lambda)\, d\mu(\lambda) - \frac{1}{\rho}\int g_4(\lambda)\, d\mu(\lambda),$$

$$\dot{g}_1(\lambda) = -\lambda g_1(\lambda) - \frac{r_x}{\lambda}, \qquad \dot{g}_2(\lambda) = -\lambda g_2(\lambda) - \frac{r_{xx}}{\lambda}, \tag{5.1}$$

$$\dot{g}_3(\lambda) = -\lambda g_3(\lambda) + \frac{2r_x}{\lambda}, \qquad \dot{g}_4(\lambda) = -\lambda g_4(\lambda) - \frac{r_{xx}}{\lambda}.$$

We abbreviate (5.1) in the form $\dot{y} = Ay$. We shall study the spectral properties of A as an operator in the space X_s of § 4 ($1 \leqq s < \infty$ is again arbitrary). Consider the resolvent equation $(A-\alpha)y = f = (f_1, f_2, f_3, f_4(\lambda), f_5(\lambda), f_6(\lambda), f_7(\lambda))$. If $-\alpha$ is not in the support of μ, this equation is immediately resolved with respect to p, q and $g_i(\lambda)$, yielding the following equation for r

$$\frac{3\eta}{\rho} r_{xx} + \frac{3\eta}{\rho} r_{xx} \cdot \int \frac{1}{\lambda(\lambda+\alpha)}\, d\mu(\lambda) - \alpha r = f_3 - \frac{2}{\rho}\int \frac{f_5(\lambda)}{\lambda+\alpha}\, d\mu(\lambda) - \frac{1}{\rho}\int \frac{f_7(\lambda)}{\lambda+\alpha}\, d\mu(\lambda).$$

As noted in § 2, $\int 1/(\lambda(\lambda+\alpha))\, d\mu(\lambda)$ has a positive real part for $\operatorname{Re}\alpha \geqq 0$. Moreover, this expression obviously goes to zero like $1/|\alpha|$ if $\alpha \to \infty$ in any sector $\{\alpha \in \mathbb{C} | -\pi + \varphi + \varepsilon \leqq \arg \alpha \leqq \pi - \varphi - \varepsilon\}$, φ being the angle of assumptions (a) and ε any positive number. From these properties it can easily be seen that the following holds:

PROPOSITION 5.1. *A is the generator of an analytic semigroup. Moreover, the spectrum of A consists of the semi-simple eigenvalue 0 and a remainder lying strictly in the left half-plane.*

Semi-simple here means that the resolvent has a simple pole at 0, or equivalently, that $R(A) \oplus N(A) = X_s$, $R(A)$ and $N(A)$ denoting the range and nullspace of A.

For technical reasons, the spaces X_n^σ, Y_n^σ of § 3 are not quite appropriate for the study of our present problem, and we shall use the following spaces, which are defined in a very similar manner.

DEFINITION 5.2. Let Z be a Banach space. Then $H^n(\mathbb{R}, Z)$ denotes the spaces of all functions $\mathbb{R} \to Z$ whose first n derivatives are square integrable in the sense of Bochner. Let moreover

$$\hat{X}_n^\sigma(Z) = \{v \in H^n(\mathbb{R}, Z) | e^{\sigma t} v,\ e^{-\sigma t} v \in H^n(\mathbb{R}, Z)\},$$

$$\hat{Y}_n^\sigma(Z) = \{v : \mathbb{R} \to Z | e^{-\sigma t} v \in H^n(\mathbb{R}, Z),\ \exists v_\infty \in Z \text{ such that } e^{+\sigma t}(v - v_\infty) \in H^n(\mathbb{R}, Z)\}.$$

Natural norms in $\hat{X}_n^\sigma$ and $\hat{Y}_n^\sigma$ are defined in an analogous way as for X_n^σ, Y_n^σ. The use of these definitions lies in the following lemma:

LEMMA 5.3. *Let the space X_2 and the operator A be as above, and let $\sigma > 0$ be small enough. Then the operator*

$$y(t) \to \left(A - \frac{d}{dt}\right)^{-1} y(t)$$

is bounded from $\hat{X}_n^\sigma(X_2)$ into $\hat{Y}_n^\sigma(N(A)) \oplus \hat{X}_n^\sigma(R(A) \cap D(A))$, where $N(A)$, $R(A)$ and $D(A)$ denote the nullspace, range and domain of A, respectively.

Proof. Note that since X_2 is a Hilbert space, the norm in $H^n(\mathbb{R}, X_2)$ can easily be expressed in terms of the Fourier transform, thus reducing the statement of the lemma to estimates on the resolvent of A. The latter follow from Proposition 5.1. (It is this argument that fails, if X_n^σ is chosen rather than $\hat{X}_n^\sigma$.)

With these preliminaries, it is now easy to establish an analogue of Theorem 3.2 for the nonlinear problem (4.1). Again we put $\hat{r} = r - \dot{b}(t)x$, and we put $y = (p-1, q, \hat{r}, g_1, g_2, g_3, g_4)$. Then (4.1) has the form

$$\dot{y} = Ay + \tilde{f}(y, \dot{b}) \tag{5.2}$$

where A is the operator studied above. $\tilde{f}$ is a smooth mapping from $\hat{Y}_n^\sigma(N(A)) \oplus \hat{X}_n^\sigma(D(A) \cap R(A)) \times \hat{X}_{n+1}^\sigma(\mathbb{R})$ into $\hat{X}_n^\sigma(X_2)$ for any $n \geqq 1$, and, according to Lemma 5.3, $(d/dt - A)^{-1}$ is (for σ small enough) a bounded linear mapping from $\hat{X}_n^\sigma(X_2)$ into $\hat{Y}_n^\sigma(N(A)) \oplus \hat{X}_n^\sigma(D(A) \cap R(A))$. The following result is now immediate from the implicit function theorem.

THEOREM 5.2. *Let $\sigma > 0$ be small enough. Then, in a neighborhood of $y = 0$, $\dot{b} = 0$ in $\hat{Y}_n^\sigma(N(A)) \oplus \hat{X}_n^\sigma(D(A) \cap R(A)) \times \hat{X}_{n+1}^\sigma(\mathbb{R})$, (5.2) has a unique resolution $y = y(\dot{b})$.*

REFERENCES

[1] S. ANTMAN, *The theory of rods*, in Handbuch der Physik VI a/2, Springer, Berlin, 1972, pp. 641–703.

[2] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart, Winston, New York, 1969.

[3] A. S. LODGE, *Body Tensor Fields in Continuum Mechanics*, Academic Press, New York, 1974.

[4] ———, *Constitutive equations from molecular network theories for polymer solutions*, Rheol. Acta, 7 (1968), pp. 379–392.

[5] A. S. LODGE, J. B. MCLEOD AND J. A. NOHEL, *A nonlinear singularly perturbed Volterra integrodifferential equation occurring in polymer rheology*, Proc. Roy. Soc. Edinburgh, 80A (1978), pp. 99–137.

[6] G. A. NARIBOLI, *Asymptotic theory of wave motion in rods*, Z. Angew. Math. Mech., 49 (1968), pp. 525–531.

[7] M. RENARDY, *Evolution of the shape of a polymer subjected to a force*, MRC Tech. Summary Rep. 2150, Mathematics Research Center, Univ. of Wisconsin, Madison, 1981.

[8] P. E. SOBOLEVSKII, *Equations of parabolic type in a Banach space*, Trudy Moskov Mat Obsch., 10 (1961), pp. 297–350, = Amer. Math. Soc. Transl., 49 (1966), pp. 1–62.

[9] S. AGMON AND L. NIRENBERG, *Properties of solutions of ordinary differential equations in Banach space*, Comm. Pure Appl. Math., 16 (1963), pp. 121–239.

[10] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, American Mathematical Society, Providence, RI, 1957.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0005 $01.00/0

THE UNIFORM ASYMPTOTIC EXPANSION OF A CLASS OF INTEGRALS RELATED TO CUMULATIVE DISTRIBUTION FUNCTIONS*

N. M. TEMME†

Abstract. An asymptotic expansion is given for a class of integrals for large values of a parameter, which corresponds with the degrees of freedom in a certain type of cumulative distribution functions. The expansion is uniform with respect to a variable related to the random variable of the distribution functions. Special cases include the chi-square distribution and the F-distribution.

1. Introduction.

We consider integrals of the type

$$F_a(\eta)=\left(\frac{a}{2\pi}\right)^{1/2}\int_{-\infty}^{\eta} e^{-a\zeta^2/2}f_a(\zeta)\,d\zeta \tag{1.1}$$

for large values of the positive parameter a. The independent variable η ranges over an unbounded domain

$$H_\delta=\{\eta=x+iy\,|\,x\in\mathbb{R},\,|y|<\delta\}, \tag{1.2}$$

where $\delta>0$; δ may depend on a but it must be bounded away from 0 when $a\to\infty$. The function $f_a(\zeta)$ is required to be an analytic function of the complex variable $\zeta\in H_\delta$ and we suppose that

$$\frac{d^k f_a(\zeta)}{d\zeta^k}=O(|\zeta|^{\lambda_k}e^{\omega\zeta^2}),\qquad x\to\pm\infty, \tag{1.3}$$

where $\zeta=x+iy\in H_\delta$ and where λ_k, ω are real numbers not depending on a. Hence, the integral (1.1) converges for $a>\omega$, $F_a(-\infty)=0$, $F_a(\infty)$ is finite and these limiting values do not depend on Im η, when $\eta\to\pm\infty$ in H_δ.

The function $f_a(\zeta)$ can play the role of a probability density function and $F_a(\eta)$ can be viewed as a cumulative distribution function. Then the variable η is related to the random variable of the underlying statistics and the parameter a corresponds with the degrees of freedom. As we will show in later sections, the gamma and beta distribution functions can be transformed into (1.1). When $f_a(\zeta)$ is a density function, it is supposed that it is positive for real ζ and that $F_a(+\infty)=1$. These conditions are not required here.

From further conditions for $f_a(\zeta)$, which will be given in the next section, it follows that (for large a) $F_a(\infty)$ has an asymptotic expansion with leading term 1. This expansion is written as

$$F_a(\infty)=\sum_{s=0}^{n-1}\frac{A_s}{a^s}+\frac{\bar{A}_n(a)}{a^n},\qquad n=0,1,2,\cdots, \tag{1.4}$$

where the coefficients A_s do not depend on a, $A_0=1$ and

$$\bar{A}_n(a)=O(1),\qquad a\to\infty,\quad n=0,1,2,\cdots.$$

* Received by the editors April 9, 1981.

† Stichting Mathematisch Centrum, Postbus 4079, 1009 AB Amsterdam, the Netherlands.

To describe some aspects of the expansion of $F_a(\eta)$ we suppose that η in (1.1) is real. For fixed values of η we have, by using well-known methods of asymptotics,

$$\begin{aligned} &\text{if } \eta<0 \quad F_a(\eta)=(2\pi a)^{-1/2}\, e^{-a\eta^2/2}\frac{f_a(\eta)}{-\eta}[1+O(a^{-1})],\\ &\text{if } \eta=0 \quad F_a(0)=\tfrac{1}{2}f_a(0)[1+O(a^{-1/2})],\\ &\text{if } \eta>0 \quad F_a(\eta)=1+O(a^{-1}). \end{aligned} \tag{1.5}$$

From these relations it follows that the asymptotic behavior of $F_a(\eta)$ is completely different in the three cases distinguished and, moreover, that the asymptotic forms do not pass into one another when η changes from negative values into positive ones. The above approximations are not uniform with respect to small values of $|\eta|$.

In asymptotics we describe these phenomena by saying that the end point of integration η may coalesce with the saddle point at $\zeta=0$. Several contributions in the literature deal with this aspect, for instance Erdélyi (1970), Olver (1974), Bleistein and Handelsman (1975) and Wong (1980).

The object of this paper is to give an asymptotic expansion which is uniform with respect to η. The uniform approximation is not only valid in a neighborhood of $\eta=0$ but in the whole domain H_δ. It is well known that such a uniform expansion cannot be described by elementary functions as arise in (1.5). In the present case we need the normal distribution function

$$P(x)=(2\pi)^{-1/2}\int_{-\infty}^{x} e^{-t^2/2}\,dt. \tag{1.6}$$

This is not surprising since we know from probability theory (the central limit theorem) that the normal distribution appears if in (1.1) a is large. It also follows by taking for $f_a(\zeta)$ a constant.

The form of the expansion for $F_a(\eta)$ is as follows. For $n=0, 1, 2, \cdots$, we write

$$\begin{aligned} &F_a(\eta)=F_a(\infty)P(\eta\sqrt{a})+R_a(\eta),\\ &R_a(\eta)=\frac{e^{-a\eta^2/2}}{\sqrt{2a\pi}}\left[\sum_{s=0}^{n-1}\frac{B_s(\eta)}{a^s}+\frac{\bar{B}_n(a,\eta)}{a^n}\right], \end{aligned} \tag{1.7}$$

where the functions $B_s(\eta)$ do not depend on a. In the next section we will give representations for $B_s(\eta)$ and $\bar{B}_n(a,\eta)$ from which information follows about the nature of the expansion in (1.7) and about the uniformity with respect to η. In (1.7), $F_a(\infty)$ can be replaced by (1.4).

The present paper extends the results of Temme (1979) on incomplete gamma functions to the more general class of integrals (1.1). In § 3 this special case will be considered again, together with other examples. We present our results for η in the strip H_δ. It is possible to introduce a more general simply connected unbounded open domain D, which should contain each real point η as an interior point. Depending on $f_a(\zeta)$, it is possible to introduce branch cuts, and D is not necessarily one-sheeted. This will not be done here because it makes the presentation less transparent. However, t is not difficult to modify our results for such a general case, for instance, by analytic continuation. Some modifications should be made in the assumptions: $F_a(\infty)$ may not be finite, since it may depend on the direction in which η approaches infinity.

2. Construction of the asymptotic expansion.

2.1. Conditions on $f_a(\zeta)$ and A_s. We suppose that, apart from the requirements on $f_a(\zeta)$ given in § 1, $f_a(\zeta)$ has an asymptotic expansion

$$f_a(\zeta)=\sum_{s=0}^{n-1}\frac{\phi_s(\zeta)}{a^s}+\frac{\bar{\phi}_n(a,\zeta)}{a^n},\qquad n=0,1,2,\cdots, \tag{2.1}$$

where $\phi_0(0)=1$, and $\phi_s(\zeta)$ do not depend on a and are analytic in H_δ. For the remainders $\bar{\phi}_n(a,\zeta)$, we assume that

$$\bar{\phi}_n(a,\zeta)=O(f_a(\zeta))\quad \text{as } a\to\infty, \tag{2.2}$$

uniform in H_δ (if $f_a(\zeta)$ happens to vanish in H_δ, this requirement should be modified in $\bar{\phi}_n(a,\zeta)=O(\max(1,|f_a(\zeta)|))$.

As will be shown, it is not possible to define the expansions (1.4) and (2.1) independently of one another. There is a relation between the coefficients A_s and the values of $\phi_s^{(k)}(0)$, the kth derivative of $\phi_s(\zeta)$ at $\zeta=0$. This follows from a well-known principle in asymptotics that says that the asymptotic expansion of the integral

$$\int_{-\infty}^{\infty}e^{-a\zeta^2/2}\phi(\zeta)\,d\zeta,\qquad a\to\infty$$

is obtained by expanding $\phi(\zeta)$ in power of ζ and integrating term by term.

In the underlying case each term in (2.1) has to be expanded. Writing

$$\phi_s(\zeta)=\sum_{t=0}^{\infty}\phi_{st}\zeta^t,\qquad |\zeta|<\delta, \tag{2.3}$$

we obtain for $s=0,1,2,3,\cdots$ and $a\to\infty$

$$\left(\frac{a}{2\pi}\right)^{1/2}\int_{-\infty}^{\infty}e^{-a\zeta^2/2}\phi_s(\zeta)\,d\zeta\sim\sum_{t=0}^{\infty}\phi_{s,2t}\left(\frac{2}{a}\right)^t\frac{\Gamma(t+\frac{1}{2})}{\Gamma(\frac{1}{2})}. \tag{2.4}$$

By rearranging the results for all terms in (2.1) and by collecting terms with equal powers of a^{-1}, we obtain (1.4) with

$$A_s=\sum_{t=0}^{s}2^t\frac{\Gamma(t+\frac{1}{2})}{\Gamma(\frac{1}{2})}\phi_{s-t,2t},\qquad s=0,1,2,\cdots. \tag{2.5}$$

2.2. Integration by parts procedure. Let us first suppose that $f_a(\zeta)$ in (1.1) does not depend on a. That is, we write $\phi_0(\zeta)$ instead of $f_a(\zeta)$, and we consider

$$F_a^{(0)}(\eta)=\left(\frac{a}{2\pi}\right)^{1/2}\int_{-\infty}^{\eta}e^{-a\zeta^2/2}\phi_0(\zeta)\,d\zeta, \tag{2.6}$$

with $\phi_0(0)=1$. Taking into account that the main contribution of the integral will come from a small neighborhood of the point $\zeta=0$, we write (cf. (1.6))

$$\begin{aligned}F_a^{(0)}(\eta)&=P(\eta\sqrt{a})+\left(\frac{a}{2\pi}\right)^{1/2}\int_{-\infty}^{\eta}e^{-a\zeta^2/2}[\phi_0(\zeta)-1]\,d\zeta\\&=P(\eta\sqrt{a})-(2\pi a)^{-1/2}\int_{-\infty}^{\eta}\frac{\phi_0(\zeta)-1}{\zeta}\,d\,e^{-a\zeta^2/2}\\&=P(\eta\sqrt{a})+(2\pi a)^{-1/2}\frac{1-\phi_0(\eta)}{\eta}e^{-a\eta^2/2}+(2\pi a)^{-1/2}\int_{-\infty}^{\eta}\phi_0^{(1)}(\zeta)\,e^{-a\zeta^2/2}\,d\zeta.\end{aligned}$$

The function $\phi_0^{(1)}(\zeta)$ is holomorphic in H_δ. It is given by

$$\phi_0^{(1)}(\zeta)=\frac{d}{d\zeta}\,\frac{\phi_0(\zeta)-1}{\zeta}.$$

Repeating this process, we obtain

$$F_a^{(0)}(\eta)\sim P(\eta\sqrt{a})\sum_{s=0}^{\infty}\frac{A_s^{(0)}}{a^s}+\frac{e^{-a\eta^2/2}}{\sqrt{2\pi a}}\sum_{s=0}^{\infty}\frac{B_s^{(0)}(\eta)}{a^s}, \tag{2.7}$$

where $A_s^{(0)}$ and $B_s^{(0)}(\eta)$ are special cases (i.e., with $t=0$) of a more general set of functions defined for $t=0, 1, 2, \cdots$.

$$\begin{aligned}
&A_s^{(t)}=\phi_t^{(s)}(0),\\
&B_s^{(t)}(\eta)=\frac{\phi_t^{(s)}(0)-\phi_t^{(s)}(\eta)}{\eta}, && s\geqq 0\\
&\phi_t^{(s)}(\eta)=\frac{d}{d\eta}\,\frac{\phi_t^{(s-1)}(\eta)-\phi_t^{(s-1)}(0)}{\eta}, && s\geqq 1,\\
&\phi_t^{(0)}(\eta)=\phi_t(\eta).
\end{aligned} \tag{2.8}$$

The same procedure can be used for each integral

$$F_a^{(t)}(\eta)=\left(\frac{a}{2\pi}\right)^{1/2}\int_{-\infty}^{\eta}e^{-a\zeta^2/2}\phi_t(\zeta)\,d\zeta, \qquad t=0, 1, 2, \cdots,$$

where $\phi_t(\zeta)$ appears in (2.1). The result is as in (2.7), with $F_a^{(0)}(\eta)$, $A_s^{(0)}$, $B_s^{(0)}(\eta)$ replaced by $F_a^{(t)}(\eta)$, $A_s^{(t)}$, $B_s^{(t)}(\eta)$, respectively.

For the complete asymptotic expansion of $F_a(\eta)$ we collect terms of equal powers of $1/a$ in the expansions of each $a^{-t}F_a^{(t)}(\eta)$, and the result is

$$F_a(\eta)\sim P(\eta\sqrt{a})\sum_{s=0}^{\infty}\frac{A_s}{a^s}+\frac{e^{-a\eta^2/2}}{\sqrt{2a\pi}}\sum_{s=0}^{\infty}\frac{B_s(\eta)}{a^s} \tag{2.9}$$

with

$$A_s=\sum_{t=0}^{s}A_{s-t}^{(t)}, \qquad B_s(\eta)=\sum_{t=0}^{s}B_{s-t}^{(t)}(\eta). \tag{2.10}$$

It is not yet clear how to interpret the formal expansion (2.9) as an asymptotic expansion. In the next subsection we will discuss a method that gives a more satisfactory relation for the coefficients A_s and $B_s(\eta)$ and from which a simple expression for the remainder in (2.9) follows (that is, for the function $\bar{B}_n(a, \eta)$ of (1.7)). The numbers A_s constructed here are the same as those defined by (1.4) and (2.5). This is proved first.

LEMMA 1. *Let A_s, $s=0, 1, \cdots$, be defined by* (2.8) *and* (2.10). *Then A_s satisfy* (2.5), *with ϕ_{st} given by* (2.3).

Proof. The function $\phi_t^{(s)}(\eta)$ defined in (2.8) are analytic in H_δ. Let us define $\phi_{tr}^{(s)}$ by writing

$$\phi_t^{(s)}(\eta)=\sum_{r=0}^{\infty}\phi_{tr}^{(s)}\eta^r, \qquad |\eta|<\delta, \quad s, t=0, 1, 2, \cdots$$

with $\phi_{tr}^{(0)}=\phi_{tr}$ of (2.3). By using the third line of (2.8), we obtain

$$\phi_{tr}^{(s)}=(r+1)\phi_{t,r+2}^{(s-1)}, \qquad s\geqq 1, \quad r\geqq 0, \quad t\geqq 0.$$

Applying this relation repeatedly, we can express $\phi_{tr}^{(s)}$ in terms of $\phi_{tr}^{(0)} = \phi_{tr}$. The result is

$$\phi_{tr}^{(s)} = (r+1)(r+3)\cdots(r+2s-1)\phi_{t,r+2s} = 2^s \frac{\Gamma(r/2+\frac{1}{2}+s)}{\Gamma(r/2+\frac{1}{2})}\phi_{t,r+2s}. \tag{2.11}$$

$A_s^{(t)}$ of (2.8) satisfies $A_s^{(t)} = \phi_{t0}^{(s)}$, giving for A_s of (2.10)

$$A_s = \sum_{t=0}^{s} A_{s-t}^{(t)} = \sum_{t=0}^{s} 2^{s-t} \frac{\Gamma(\frac{1}{2}+s-t)}{\Gamma(\frac{1}{2})}\phi_{t,2(s-t)}.$$

This proves the lemma. □

In the examples in § 3, the numbers A_s are usually obtained via (1.4). If $F_a(\infty) = 1$, then $A_s = 0$ $(s \geqq 1)$, and (2.5) gives an extra relation between ϕ_{tr}. Of course, it is possible to normalize $F_a(\eta)$ by dividing $f_a(\zeta)$ by $F_a(\infty)$. Then the scaled function $F_a(\eta)$ satisfies $F_a(\infty) = 1$.

LEMMA 2. *Let $B_s(\eta)$, $s = 0, 1, \cdots$, be defined by* (2.8) *and* (2.10); *let the numbers B_{st} be defined by*

$$B_s(\eta) = \sum_{t=0}^{\infty} B_{st}\eta^t, \qquad |\eta| < \delta. \tag{2.12}$$

Then

$$B_{st} = -\sum_{r=0}^{s} 2^r \frac{\Gamma(1+t/2+r)}{\Gamma(1+t/2)}\phi_{s-r,t+2r+1}. \tag{2.13}$$

Proof. This follows from (2.11) and the second line of (2.8) □

Remark. It may be rather difficult to compute the functions $B_s(\eta)$ from the definitions in (2.8) and (2.10), especially for values of η close to zero. Then the representations in (2.12) and (2.13) may be useful.

2.3. A simpler recursion for $B_s(\eta)$. Consider the function $R_a(\eta)$ defined in the first line of (1.7). From (1.1) and (1.6) it follows that

$$\frac{d}{d\eta} R_a(\eta) = (a/2\pi)^{1/2} e^{-a\eta^2/2}[f_a(\eta) - F_a(\infty)]. \tag{2.14}$$

The substitution of (1.4), (2.1) and the formal series

$$R_a(\eta) = \frac{e^{-a\eta^2/2}}{\sqrt{2a\pi}} \sum_{s=0}^{\infty} \frac{B_s(\eta)}{a^s}$$

into (2.14) shows that (2.14) is formally satisfied if

$$\begin{aligned} \eta B_0(\eta) &= 1 - \phi_0(\eta), \\ \eta B_s(\eta) &= A_s - \phi_s(\eta) + \frac{d}{d\eta} B_{s-1}(\eta), \qquad s \geqq 1. \end{aligned} \tag{2.15}$$

As outlined in foregoing subsections, the A_s are related with the numbers ϕ_{st} defined in (2.3); this relation is given in (2.5).

It is not clear, yet, that the functions $B_s(\eta)$ constructed in § 2.2 satisfy the recursion in (2.15). This will be proved in the theorem which follows.

THEOREM 1. *Let for $s = 0, 1, 2, \cdots, B_s(\eta)$ be defined in* (2.10) *with $B_r^{(t)}(\eta)$ defined in* (2.8). *Then $B_s(\eta)$ satisfy the recursion relation of* (2.15).

Proof. We use induction on s. For $s=0$, see the second line of (2.8) with $t=0$; it gives the first relation of (2.15). Suppose, next, that the theorem is correct for the function $B_r(t)$ of (2.10) with $0\leqq r\leqq s-1$, for some $s\geqq 1$. Then the right-hand side of the second line of (2.15) becomes for this s (see (2.10) and (2.8))

$$A_s-\phi_s^{(0)}(\eta)+\frac{d}{d\eta}\sum_{t=0}^{s-1}B_{s-1-t}^{(t)}(\eta)=A_s-\phi_s(\eta)+\sum_{t=0}^{s-1}\frac{d}{d\eta}\frac{\phi_t^{(s-1-t)}(0)-\phi_t^{(s-1-t)}(\eta)}{\eta}$$

$$=A_s-\phi_s(\eta)+\sum_{t=0}^{s-1}\phi_t^{(s-t)}(\eta)=A_s-\sum_{t=0}^{s}\phi_t^{(s-t)}(\eta)$$

$$=\sum_{t=0}^{s}[A_{s-t}^{(t)}-\phi_t^{(s-t)}(\eta)]=\sum_{t=0}^{s}[\phi_t^{(s-t)}(0)-\phi_t^{(s-t)}(\eta)]$$

$$=\eta\sum_{t=0}^{s}B_{s-t}^{(t)}(\eta)=\eta B_s(\eta),$$

which proves the theorem. □

2.4. Representations for the remainder.

THEOREM 2. *Consider for* (1.1) *representation* (1.7), *where the functions* $B_s(\eta)$ *are defined in* (2.15). *Then for the remainder* $\bar{B}_n(a,\eta)$ *of* (1.7) *we have for* $n=0,1,\cdots$

$$e^{-a\eta^2/2}\bar{B}_n(a,\eta)=-a\int_{-\infty}^{\eta}\zeta B_n(\zeta)\,e^{-a\zeta^2/2}\,d\zeta+\int_{-\infty}^{\eta}[\bar{\phi}_{n+1}(a,\zeta)-\bar{A}_{n+1}(a)]\,e^{-a\zeta^2/2}\,d\zeta, \tag{2.16}$$

where $\bar{A}_n(a)$ *and* $\bar{\phi}_n(a,\zeta)$ *are defined in* (1.4) *and* (2.1).

Proof. From (2.14) (or from (1.1), (1.6), (1.7)) we have

$$R_a(\eta)=\left(\frac{a}{2\pi}\right)^{1/2}\int_{-\infty}^{\eta}e^{-a\zeta^2/2}[f_a(\zeta)-F_a(\infty)]\,d\zeta. \tag{2.17}$$

Using

$$f_a(\zeta)=\phi_0(\zeta)+\frac{\bar{\phi}_1(a,\zeta)}{a},\quad F_a(\infty)=1+\frac{\bar{A}_1(a)}{a},\quad R_a(\eta)=\frac{e^{-a\eta^2/2}}{\sqrt{2a\pi}}\bar{B}_0(a,\eta),$$

we obtain

$$R_a(\eta)=\left(\frac{a}{2\pi}\right)^{1/2}\int_{-\infty}^{\eta}e^{-a\zeta^2/2}\left[\phi_0(\zeta)-1+\frac{\bar{\phi}_1(a,\zeta)}{a}-\frac{\bar{A}_1(a)}{a}\right]d\zeta$$

$$=(2\pi a)^{-1/2}\int_{-\infty}^{\eta}e^{-a\zeta^2/2}[-a\zeta B_0(\zeta)+\bar{\phi}_1(a,\zeta)-\bar{A}_1(a)]\,d\zeta,$$

which gives (2.16) with $n=0$. Considering (2.16) with $n\geqq 0$ we proceed by using the obvious relations

$$\bar{\phi}_{n+1}(a,\zeta)=\phi_{n+1}(\zeta)+\frac{\bar{\phi}_{n+2}(a,\zeta)}{a},$$

$$\bar{A}_{n+1}(a)=A_{n+1}+\frac{\bar{A}_{n+2}(a)}{a},$$

$$\bar{B}_n(a,\eta)=B_n(\eta)+\frac{\bar{B}_{n+1}(a,\eta)}{a}.$$

Then, by partial integration of the first integral in (2.16), this formula becomes

$$e^{-a\eta^2/2}\left[B_n(\eta)+\frac{\bar{B}_{n+1}(a,\eta)}{a}\right]$$
$$=e^{-a\zeta^2/2}B_n(\zeta)\Big|_{\zeta=-\infty}^{\zeta=\eta}+\int_{-\infty}^{\eta}\left[\phi_{n+1}(\zeta)-A_{n+1}-\frac{d}{d\zeta}B_n(\zeta)+\frac{\bar{\phi}_{n+2}(a,\zeta)}{a}-\frac{\bar{A}_{n+2}(a)}{a}\right]d\zeta.$$

The integrated term vanishes at $\zeta=-\infty$ (this follows from (1.3)), and by using (2.15) we obtain (2.16) with n replaced by $n+1$. □

THEOREM 3. *Under the same conditions as in Theorem 2 we have for* $n=0,1,2,\cdots$

$$e^{-a\eta^2/2}\bar{B}_n(a,\eta)=a\int_{\eta}^{\infty}\zeta B_n(\zeta)\,e^{-a\zeta^2/2}\,d\zeta+\int_{\eta}^{\infty}[\bar{A}_{n+1}(a)-\bar{\phi}_{n+1}(a,\zeta)]\,e^{-a\zeta^2/2}\,d\zeta. \tag{2.18}$$

Proof. From the definition of $R_a(\eta)$ in (1.7) it follows that

$$\lim_{x\to\pm\infty}R_a(\eta)=0,\qquad \eta=x+iy\in H_\delta.$$

Hence, we could have started with (2.17) with interval of integration $[\eta,\infty)$ and a different sign before the integral. The rest of the proof runs as in Theorem 2. □

It follows from both theorems that, for fixed real values of η, $\bar{B}_n(a,\eta)$ of (1.7) satisfy

$$\bar{B}_n(a,\eta)=O(1),\qquad a\to\infty. \tag{2.19}$$

For $\eta\leqq 0$ this follows from (2.16), for $\eta\geqq 0$ from (2.18). Hence, again for fixed η, $R_a(\eta)$ of (1.7) satisfy

$$R_a(\eta)=O(a^{-1/2}\,e^{-a\eta^2/2}),\qquad a\to\infty. \tag{2.20}$$

It is not difficult to show that (2.19) and (2.20) hold uniformly with respect to η in compact subsets of H_δ. In all relevant applications (see § 3) the number ω of (1.3) is 0. Also, bounds for $\bar{B}_n(a,\eta)$ can be constructed that are holding uniformly with respect to $\eta\in H_\delta$.

In our previous paper Temme (1979) on incomplete gamma functions such bounds for $\bar{B}_n(a,\eta)$ were computed for $n=0,1,\cdots,10$. In that paper we overlooked the remarkable point that both representations (2.16) and (2.18) can be used for constructing these bounds. Consequently, formula (2.17) of that paper should be revised for $\eta<0$ and (2.19) for $\eta>0$. Also, Remark 1 in the cited reference (p. 760) should be skipped. More details will be given in § 3.4.

In the next section some examples are worked out. In each example the functions are related with statistical distribution functions. Usually transformations are needed for a representation in the standard form (1.1).

3. Examples.

3.1. Incomplete normal distribution function.

We consider

$$F_a(\eta)=\left(\frac{a}{2\pi}\right)^{1/2}\int_{-\infty}^{\eta}e^{-a\zeta^2/2}\frac{d\zeta}{1+\zeta^2} \tag{3.1}$$

in the strip H_δ with $\delta=1$. This function finds wide application in probability theory, mathematical statistics and in problems involving the heat conduction equation. Jones

(1972) used it for describing the asymptotic expansion of a double integral. It constitutes a sort of generalization of the error integral. To see this note that

$$F_a(\infty) = (2a\pi)^{1/2}\, e^{a/2} Q(\sqrt{a}),$$

where $Q(x) = 1 - P(x) = \frac{1}{2}\operatorname{erfc}(x/\sqrt{2})$.

The function $f_a(\zeta)$ of (1.1) is here $1/(1+\zeta^2)$, and in this simple case the coefficients $B_s(\eta)$ can be constructed rather easily. We have $f_a(\zeta) = \phi_0(\zeta)$ and $\phi_s(\zeta) = 0$ $(s \geqq 1)$. From the known expansion of the error function we obtain

$$F_a(\infty) \sim \sum_{s=0}^{\infty} \frac{A_s}{a^s}, \quad A_0 = 1, \quad A_s = \frac{(-2)^s \Gamma(s+\frac{1}{2})}{\Gamma(\frac{1}{2})}, \quad s \geqq 1,$$

which can also be obtained from (2.5) with

$$\phi_{st} = \begin{cases} 0 & \text{if } s \geqq 1 \text{ or } t = 2r+1, \quad r \geqq 0, \\ (-1)^{t/2} & \text{else.} \end{cases}$$

Hence, from (2.12), (2.13) it follows that for $|\eta| < 1$

$$B_s(\eta) = \eta(-2)^s \sum_{t=0}^{\infty} \frac{\Gamma(t+s+\frac{3}{2})}{\Gamma(t+\frac{3}{2})} (-\eta^2)^t,$$

which can be expressed in terms of Gauss' hypergeometric function ${}_2F_1(a, b; c; z)$:

$$\begin{aligned} B_s(\eta) &= \eta(-2)^s \frac{\Gamma(s+\frac{3}{2})}{\Gamma(\frac{3}{2})} {}_2F_1(s+\tfrac{3}{2}, 1; \tfrac{3}{2}; -\eta^2) \\ &= \eta(-2)^s \frac{\Gamma(s+\frac{3}{2})}{\Gamma(\frac{3}{2})} (1+\eta^2)^{-1} {}_2F_1\left(1, -s; \tfrac{3}{2}; \frac{\eta^2}{\eta^2+1}\right). \end{aligned}$$

The second relation enables us to write

$$B_s(\eta) = \tfrac{1}{2}\eta(-2)^s \frac{\Gamma(s+\frac{3}{2})}{(1+\eta^2)\Gamma(\frac{3}{2})} \int_0^1 (1-t)^{-1/2} \left(1 - t\frac{\eta^2}{\eta^2+1}\right)^s dt,$$

from which we obtain

$$|B_s(\eta)| \leqq \frac{|\eta|}{1+\eta^2} 2^s \frac{\Gamma(s+\frac{3}{2})}{\Gamma(\frac{3}{2})}, \qquad \eta \in \mathbb{R}.$$

Hence, $B_s(\eta)$ are bounded on $\mathbb{R}$.

The representation for (3.1) becomes (see (1.7))

$$F_a(\eta) = F_a(\infty) P(\eta\sqrt{a}) + \frac{e^{-a\eta^2/2}}{\sqrt{2a\pi}} \left[\sum_{s=0}^{n-1} \frac{B_s(\eta)}{a^s} + \frac{\bar{B}_n(a, \eta)}{a^n}\right], \tag{3.2}$$

with $B_0(\eta) = \eta/(1+\eta^2)$, $B_1(\eta) = -\eta(3+\eta^2)/(1+\eta^2)^2$. In this example other $B_s(\eta)$ can be computed by using recursion relations of ${}_2F_1$-functions. It easily follows that

$$(1+\eta^2)B_{s+1}(\eta) = -2s(2s+1)B_{s-1}(\eta) - [(2s+1)\eta^2 + 4s + 3]B_s(\eta).$$

To estimate the remainder we remark that in (2.16) $\bar{\phi}_{n+1}(a, \zeta) = 0$ and $|\bar{A}_{n+1}(a)|$ can be replaced by A_{n+1}. This follows from a known result about the asymptotic expansion of the error function

$$F_a(\infty) = \left(\frac{a\pi}{2}\right)^{1/2} e^{a/2} \operatorname{erfc}\left(\frac{a}{2}\right)^{1/2} = \sum_{s=0}^{n-1} \frac{A_s}{a^s} + \frac{\bar{A}_n(a)}{a^n},$$

where $\bar{A}_n(a) = \theta_n A_n$, $n = 0, 1, \cdots$ with $0 < \theta_n < 1$.

It follows that (see (2.16))

$$|e^{-a\eta^2/2}\bar{B}_n(a,\eta)| \leqq \frac{2^n\Gamma(n+\frac{3}{2})}{\Gamma(\frac{3}{2})}\int_{-\infty}^{\eta}\frac{1+a\zeta^2}{1+\zeta^2}e^{-a\zeta^2/2}\,d\zeta.$$

The integral can be expressed in terms of $F_a(\eta)$ and $P(\eta\sqrt{a})$. This estimate is valid for $\eta\in\mathbb{R}$. However, for $\eta\geqq 0$ it is better to use the representation of (2.18) which gives the same expression with an interval of integration $[\eta,\infty)$.

It follows that (3.2) can be used throughout the strip H_1. It is not difficult to extend the results for a larger domain, or for complex values of a.

3.2. Sievert integral. This integral is defined as

$$I(\theta,a)=\int_{-\pi/2}^{\theta}e^{a(1-1/\cos\phi)}\,d\phi, \tag{3.3}$$

$a>0$, $-\pi/2<\theta<\pi/2$. The complete integral $I(\pi/2,a)$ is an integral of the modified Bessel function $K_0(x)$, that is,

$$I(\pi/2,a)=2e^a\int_a^{\infty}K_0(x)\,dx. \tag{3.4}$$

We need a transformation in order to bring (3.3) into the standard form (1.1). The appropriate change of variables is defined by

$$\begin{aligned} -\tfrac{1}{2}\zeta^2 &= 1-1/\cos\phi, \qquad \zeta=\frac{2\sin(\phi/2)}{\sqrt{\cos\phi}},\\ -\tfrac{1}{2}\eta^2 &= 1-1/\cos\theta, \qquad \eta=\frac{2\sin(\theta/2)}{\sqrt{\cos\theta}}, \end{aligned} \tag{3.5}$$

and the integral (3.2) becomes

$$I(\theta,a)=\int_{-\infty}^{\eta}e^{-a\zeta^2/2}\frac{d\phi}{d\zeta}\,d\zeta, \tag{3.6}$$

where $d\phi/d\zeta=[(1+\zeta^2/2)(1+\zeta^2/4)^{1/2}]^{-1}$, which is holomorphic in a strip H_δ in the ζ-plane with $\delta=\sqrt{2}$.

Remark. It is important to note that the mapping (3.5) of the ϕ-interval into the ζ-interval involves a square root $\zeta=2[\sin^2(\phi/2)/\cos\phi]^{1/2}$, where the sign of the square root has the sign of ϕ. In this way ζ becomes a holomorphic function of ϕ at $\phi=0$. The same for η and θ.

The standard form (1.1) is now achieved by writing

$$F_a(\eta)=\left(\frac{a}{2\pi}\right)^{1/2}I(\theta,a)=\left(\frac{a}{2\pi}\right)^{1/2}\int_{-\infty}^{\eta}e^{-a\zeta^2/2}f_a(\zeta)\,d\zeta, \qquad f_a(\zeta)=\frac{d\phi}{d\zeta}, \tag{3.7}$$

and $F_a(\infty)$ has the known expansion

$$\begin{aligned} F_a(\infty) &= 2\left(\frac{a}{2\pi}\right)^{1/2}e^a\int_a^{\infty}K_0(x)\,dx\sim\sum_{n=0}^{\infty}\frac{A_s}{a^s},\\ A_s &= (-1)^s\frac{\Gamma(s+\frac{1}{2})}{\Gamma(\frac{1}{2})}\sum_{t=0}^{s}2^{-t}\frac{\Gamma(t+\frac{1}{2})}{2t!\,\Gamma(\frac{1}{2})}. \end{aligned} \tag{3.8}$$

The asymptotic expansion as in (1.7) can now be given for $a \to \infty$, uniform with respect to $\eta \in H_\delta$ ($\delta = \sqrt{2}$). The relation between η and the original variable θ is given in (3.5). Complex values of η correspond to complex values of θ. In fact, (3.5) defines a conformal mapping between parts of the complex η and θ planes, and the asymptotic expansion is valid for complex θ in the image of H_δ under this mapping. As in the foregoing example, the coefficients $B_s(\eta)$ can be expressed in terms of hypergeometric functions.

The expansion reads

$$\left(\frac{a}{2\pi}\right)^{1/2} I(\theta, a) \sim F_a(\infty) P(2 \sin \tfrac{1}{2}\theta \sqrt{a/\cos\theta}) + \frac{e^{a(1-1/\cos\theta)}}{\sqrt{2\pi a}} \sum_{s=0}^{\infty} \frac{B_s(\eta)}{a^s}$$

(3.9) $$\text{as } a \to \infty, \quad \text{uniform in } \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right],$$

$$B_0(\eta) = \frac{\cos(\theta/2)\sqrt{\cos\theta} - \cos^2\theta}{\sin\theta}.$$

3.3. Incomplete beta function. This well-known function is given by

$$I_x(p, q) = \frac{1}{B(p, q)} \int_0^x t^{p-1}(1-t)^{q-1}\, dt. \tag{3.10}$$

We consider $0 \leqq x \leqq 1$, $p > 0$, $q > 0$; however, extension to complex values is possible. $B(p, q)$ is Euler's (complete) beta function

$$B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}\, dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

We consider the asymptotic expansion of $I_x(p, q)$ for large p (or q or both) uniformly with respect to $x \in [0, 1]$. Since

$$I_x(p, q) = 1 - I_{1-x}(q, p), \tag{3.11}$$

we can consider $p \geqq q$.

The incomplete beta function is a standard probability function. As special cases it has the (negative) binomial distribution, Student's distribution and the F-(variance-ratio) distribution.

We start with the symmetric case $p = q$, which is rather easy to handle compared with the general situation.

3.3.1. The symmetric case $p = q = a$. A trivial transformation gives

$$I_x(p, q) = \frac{4^{-a}}{B(a, a)} \int_0^x e^{a \ln[4t(1-t)]} \frac{dt}{t(1-t)}. \tag{3.12}$$

The next appropriate transformations are

$$-\tfrac{1}{2}\zeta^2 = \ln[4t(1-t)], \qquad 0 < t < 1, \quad \operatorname{sign}(\zeta) = \operatorname{sign}(t - \tfrac{1}{2}),$$
$$-\tfrac{1}{2}\eta^2 = \ln[4x(1-x)], \qquad 0 < x < 1, \quad \operatorname{sign}(\eta) = \operatorname{sign}(x - \tfrac{1}{2}).$$

Then we have the standard form (1.1), with

$$F_a(\eta) = I_x(a, a) = \left(\frac{a}{2\pi}\right)^{1/2} \int_{-\infty}^{\eta} e^{-a\zeta^2/2} f_a(\zeta)\, d\zeta,$$

$$f_a(\zeta) = \frac{4^{-a}}{B(a, a)} \left(\frac{\pi}{a}\right)^{1/2} [\tfrac{1}{2}\zeta^2/(1-e^{-\zeta^2/2})]^{1/2},$$

where the square root is positive for real values of its argument. Considered as a function of the complex variable ζ, $f_a(\zeta)$ is holomorphic in H_δ with $\delta = \sqrt{2\pi}$. By writing

$$\frac{4^{-a}(\pi/a)^{1/2}}{B(a, a)} = \frac{a^{-1/2}\Gamma(a+\frac{1}{2})}{\Gamma(a)} \sim \sum_{s=0}^{\infty} c_s a^{-s}, \qquad a \to \infty,$$

where c_s can be expressed in terms of Bernoulli polynomials (see Luke (1969, vol. I, p. 33)), it follows that

$$f_a(\zeta) \sim \phi_0(\zeta) \sum_{s=0}^{\infty} \frac{c_s}{a^s}, \qquad a \to \infty,$$

$$\phi_0(\zeta) = [\tfrac{1}{2}\zeta^2/(1-e^{-\zeta^2/2})]^{1/2} = \sum_{s=0}^{\infty} \frac{B_s^{(1/2)}(\frac{1}{2})}{s!} (\tfrac{1}{2}\zeta^2)^s.$$

The coefficients in this expansion are generalized Bernoulli polynomials; see again Luke (1969, vol. I, p. 18). In this case $F_a(\infty) = 1$; hence, $A_0 = 1$, $A_s = 0$ $(s \geqq 1)$. A further analysis shows that all $B_s(\eta)$ are bounded functions of η on $(-\infty, \infty)$.

3.3.2. The general case $p \geqq q$. Let us write in (3.10)

$$p = a \sin^2\theta, \quad q = a\cos^2\theta, \quad 0 \leqq \theta \leqq \frac{\pi}{2}.$$

Then

$$I_x(p, q) = \frac{1}{B(p, q)} \int_0^x \exp\left(a(\sin^2\theta \ln t + \cos^2\theta \ln(1-t))\right) \frac{dt}{t(1-t)}. \tag{3.13}$$

The maximum of the integrand occurs at $t = \sin^2\theta$, Hence, the transformation $t \to \zeta$ reads

$$-\tfrac{1}{2}\zeta^2 = \sin^2\theta \ln\frac{t}{\sin^2\theta} + \cos^2\theta \ln\frac{1-t}{\cos^2\theta}, \tag{3.14}$$

where the sign of ζ equals the sign of $t - \sin^2\theta$ (for real variables, for complex variables it is defined by analytic continuation of the real case). A similar transformation holds for $x \to \eta$ if t and ζ are replaced by x and η, respectively. From (3.14) we obtain

$$-\zeta\frac{d\zeta}{dt} = \frac{\sin^2\theta - t}{t(1-t)},$$

hence the representation of (3.13) in the standard form is

$$\begin{aligned} F_a(\eta) &= I_x(p, q) = \left(\frac{a}{2\pi}\right)^{1/2} \int_{-\infty}^{\eta} e^{-a\zeta^2/2} f_a(\zeta)\, d\zeta, \\ f_a(\zeta) &= \left(\frac{a}{2\pi}\right)^{-1/2} \frac{\exp[a(\sin^2\theta \ln\sin^2\theta + \cos^2\theta \ln\cos^2\theta)]}{B(p, q)\sin\theta\cos\theta} \frac{\zeta\cos\theta\sin\theta}{t - \sin^2\theta}. \end{aligned} \tag{3.15}$$

It follows, and this is well known, that the critical point $\eta = 0$ corresponds with $x = \sin^2\theta = p/(p+q)$. That is, when for large $a = p+q$ the parameter x crosses the value $p/(p+q)$, the function $I_x(p, q)$ changes suddenly from values close to zero into values close to 1.

It remains to show that $f_a(\zeta)$ of (3.15) has the properties as supposed in §§ 1 and 2. First we consider the ζ-part of it which is defined by

$$\phi_0(\zeta) = \frac{\zeta \cos\theta \sin\theta}{t - \sin^2\theta}, \tag{3.16}$$

where ζ and $t(\zeta)$ are related by (3.14); $t = \sin^2\theta$ corresponds with $\zeta = 0$. At this point $\phi_0(\zeta)$ is regular and $\phi_0(0) = 1$. Due to the many-valuedness of the logarithms in (3.14), we have other finite singularities in the ζ-plane. These singularities occur for $t = \sin^2\theta \exp(2\pi i n)$ or $(1-t) = \cos^2\theta \exp(2\pi i m)$, n, $m = \pm 1, \pm 2, \cdots$ in the many-sheeted t-plane. Corresponding ζ-values follow from (3.14)

$$-\tfrac{1}{2}\zeta_n^2 = 2\pi i n \sin^2\theta, \qquad -\tfrac{1}{2}\zeta_m^2 = 2\pi i m \cos^2\theta.$$

These ζ-values have imaginary parts

$$\pm \sin\theta\sqrt{2\pi n}, \qquad \pm \cos\theta\sqrt{2\pi m}.$$

If $p \geqq q$, we have $\sin\theta \geqq \cos\theta$. Hence, $\phi_0(\zeta)$ is holomorphic in H_δ with $\delta = \cos\theta\sqrt{2\pi}$, with $\cos\theta = \sqrt{q/(p+q)}$. It follows that $\delta \to 0$, if $q/(p+q) \to 0$ for large $a = p+q$. We can apply the methods of § 2 if δ is bounded away from zero. An important conclusion is that, when $p \to \infty$, q should also grow. Otherwise the strip H_δ will coincide with $\mathbb{R}$ since the singularities approach the origin $\zeta = 0$ (the saddle point).

We proceed by supposing that $p \to \infty$, $q \to \infty$, such that $q/(p+q) \geqq \varepsilon > 0$, where ε does not depend on p or q.

The remaining part of (3.15), that is, the part not including (3.16), can be expanded in terms of the large parameter $a = p+q$. By using the well-known Stirling approximations of the gamma function, we obtain

$$f_a(\zeta) \sim \phi_0(\zeta) \sum_{s=0}^{\infty} \frac{c_s(\theta)}{a^s}, \qquad a \to \infty,$$

with $c_0(\theta) = 1$, $c_1(\theta) = \frac{1}{12}[1 - 1/\sin^2\theta \cos^2\theta)]$. Hence, the function $f_a(\zeta)$ satisfies the requirements of §§ 1 and 2; $F_a(\infty) = 1$, and we can compute the functions $B_s(\eta)(s \geqq 1)$ from

$$B_0(\eta) = \frac{1 - \phi_0(\eta)}{\eta},$$

with $\phi_0(\eta)$ given in (3.16). The expansion

$$F_a(\eta) = I_x(p, q) \sim P(\eta\sqrt{a}) + \frac{e^{-a\eta^2/2}}{\sqrt{2\pi a}} \sum_{s=0}^{\infty} \frac{B_s(\eta)}{a^s}, \tag{3.17}$$

$a = p+q \to \infty$, is uniformly valid with respect to $x \in [0, 1]$ as long as $q/(p+q) \geqq \varepsilon > 0$ (for the case $p \geqq q$; if $q \geqq p$, we suppose that $p/(p+q) \geqq \varepsilon > 0$).

It is not surprising that we cannot obtain a uniform expansion in this way. The problem is to find an expansion for $a = p+q \to \infty$, which is uniform with respect to both x and θ, with $x \in [0, 1]$, $\theta \in [0, \pi/2]$. Thus there is one asymptotic variable (a) and two nonasymptotic variables (x and θ). It is expected that to cover the whole domain of interest in the $x \times \theta$ space, it will be necessary to resort to a transcendental function of two variables as approximant (see Olver (1975)).

It is possible to treat the case $p \to \infty$, q fixed, by using a method given by Erdélyi (1974) (see also Temme (1976)). Let us write (3.10) in the form

$$I_x(p, q) = \frac{1}{B(p, q)} \int_{-\ln x}^{\infty} e^{-pt} t^{q-1} \left[\frac{1-e^{-t}}{t}\right]^{q-1} dt,$$

which satisfies (by adapting the notation) the requirements for obtaining the asymptotic expansion for $p \to \infty$, uniformly valid in $x \in [0, 1]$; q is a fixed parameter. We restrict q to the interval $(0, 1)$, but there is no loss of generality in this assumption. Erdélyi's expansion is of the form

$$I_x(p, q) \sim \frac{1}{B(p, q)} \left[p^{-q} Q(q, -p \ln x) \sum_{s=0}^{\infty} \frac{A_s(q)}{p^s} + x^p \sum_{s=1}^{\infty} \frac{B_s(x, q)}{p^s}\right], \tag{3.18}$$

where $Q(a, z)$ is an incomplete gamma function (see next example). The construction of the coefficients $A_s(q)$ and $B_s(x, q)$ is outlined in Erdélyi's paper. Note that $Q(a, z)$ is a function of two variables. It is still an open problem how to modify (3.17) and (3.18) in order to obtain an expansion for $I_x(p, q)$ for $p \to \infty$, uniformly in $x \in [0, 1]$ and $q \geqq \varepsilon > 0$.

3.4. Incomplete gamma functions. This important example is considered earlier in Temme (1979). The present method gives the same asymptotic expansion. The incomplete gamma functions are

$$P(a, x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt, \qquad Q(a, x) = \frac{1}{\Gamma(a)} \int_x^{\infty} t^{a-1} e^{-t} dt.$$

The basic transformations for the integrals are applied on

$$P(a, x) = \frac{e^{-a} a^a}{\Gamma(a)} \int_0^{x/a} e^{-a(-\ln t + t - 1)} t^{-1} dt$$

by defining

$$\tfrac{1}{2}\zeta^2 = -\ln t + t - 1, \qquad \operatorname{sign} \zeta = \operatorname{sign} (t-1)$$

$$\tfrac{1}{2}\eta^2 = -\ln \lambda + \lambda - 1, \qquad \operatorname{sign} \eta = \operatorname{sign} (\lambda - 1), \qquad \lambda = \frac{x}{a}.$$

The result is

$$F_a(\eta) = P(a, x) = \left(\frac{a}{2\pi}\right)^{1/2} \int_{-\infty}^{\eta} e^{-a\zeta^2/2} f_a(\zeta)\, d\zeta,$$

$$G_a(\eta) = Q(a, x) = \left(\frac{a}{2\pi}\right)^{1/2} \int_{\eta}^{\infty} e^{-a\zeta^2/2} f_a(\zeta)\, d\zeta,$$

with

$$f_a(\zeta) = \frac{e^{-a} a^a (2\pi/a)^{1/2}}{\Gamma(a)} \frac{\zeta}{t-1} \sim \phi_0(\zeta) \sum_{s=0}^{\infty} \frac{\gamma_s}{a^s}, \qquad \phi_0(\zeta) = \frac{\zeta}{t-1},$$

and the numbers γ_s are the well-known coefficients appearing in the expansion of the reciprocal gamma function. For more details we refer to Temme (1979).

The expansions are (note that $F_a(\infty)=1$)

$$P(a,x)=P(\eta\sqrt{a})+\frac{e^{-a\eta^2/2}}{\sqrt{2\pi a}}\left[\sum_{s=0}^{n-1}\frac{B_s(\eta)}{a^s}+\frac{\bar{B}_n(a,\eta)}{a^n}\right],$$

$$Q(a,x)=Q(\eta\sqrt{a})-\frac{e^{-a\eta^2/2}}{\sqrt{2\pi a}}\left[\sum_{s=0}^{n-1}\frac{B_s(\eta)}{a^s}+\frac{\bar{B}_n(a,\eta)}{a^n}\right],$$

$$B_0(\eta)=\frac{1}{\eta}-\frac{1}{\lambda-1},\ B_1(\eta)=\frac{1}{(\lambda-1)^3}+\frac{1}{(\lambda-1)^2}+\frac{1}{12(\lambda-1)}-\frac{1}{\eta^3}.$$

$B_s(\eta)$ are the same as $-c_s(\eta)$ in our previous paper.

As mentioned in § 2.4, the bounds given there can be sharpened. By using Theorems 2 and 3, we obtain

$$|\bar{B}_n(a,\eta)|\leqq(\lambda+1)^{-\kappa}C_n\,e^{-a\eta^2/2}+|H_{n+1}(a)|e^a a^{-a}\Gamma(a)D(a,x),$$

where $\kappa=\frac{1}{2}$ for $n=0$, 1 for $n\geqq 1$,

$$D(a,x)=\min\,(P(a,x),\,Q(a,x)),\quad C_s=\sup_{\eta\in\mathbb{R}}[(1+\lambda)^{\kappa}|B_s(\eta)|],\quad s=0,1,\ldots$$

and $H_n(a)$ is the remainder in

$$\frac{1}{\Gamma(a)}=e^a a^{-a}\left(\frac{a}{2\pi}\right)^{1/2}\left[\sum_{s=0}^{n-1}\frac{\gamma_s}{a^s}+\frac{H_n(a)}{a^n}\right].$$

The numbers C_n and bounds for $H_n(a)$ are given in Temme (1979).

3.5. Pearson type IV probability function. We consider it in the form (see Fettis (1976))

$$I(\theta,\alpha,\beta)=\int_{-\pi/2}^{\theta}e^{2\beta\phi}\cos^{2\alpha}\phi\,d\phi,$$

for large positive vâlues of α; $-\pi/2<\theta<\pi/2$; β is a real parameter. It is known that

$$I(\tfrac{1}{2}\pi,\alpha,\beta)=\frac{\pi\Gamma(1+2\alpha)2^{-2\alpha}}{\Gamma(1+\alpha+i\beta)\Gamma(1+\alpha-i\beta)}.$$

The maximum of the integrand occurs at $\phi=\lambda:=\arctan\gamma$, with $\gamma=\beta/\alpha$. The appropriate transformations are

$$-\tfrac{1}{2}\zeta^2=\gamma(\phi-\lambda)+\ln\frac{\cos\phi}{\cos\lambda},\qquad -\tfrac{1}{2}\eta^2=\gamma(\theta-\lambda)+\ln\frac{\cos\theta}{\cos\lambda},$$

with sign $\zeta=$ sign $(\phi-\lambda)$, sign $\eta=$ sign $(\theta-\lambda)$. Hence, with $a=2\alpha$, we obtain

$$F_a(\eta)=\frac{I(\theta,\alpha,\beta)}{I(\pi/2,\alpha,\beta)}=\left(\frac{a}{2\pi}\right)^{1/2}\int_{-\infty}^{\eta}e^{-a\zeta^2/2}f_a(\zeta)\,d\zeta,$$

with

$$f_a(\zeta)=\frac{e^{a(\gamma\lambda+\ln\cos\lambda)}}{I(\pi/2,\alpha,\beta)}\left(\frac{2\pi}{a}\right)^{1/2}\frac{d\phi}{d\zeta}\sim\phi_0(\zeta)\sum_{s=0}^{\infty}\frac{c_s(\gamma)}{a^s},$$

$$\phi_0(\zeta)=\frac{1}{\cos\lambda}\,\frac{\zeta}{\tan\phi-\gamma},\qquad \phi_0(0)=1.$$

REFERENCES

N. BLEISTEIN AND R. A. HANDELSMAN (1975). *Asymptotic Expansion of Integrals*, Holt, Rinehart and Winston, New York.

A. ERDÉLYI (1970), *Uniform asymptotic expansion of integrals*, in Analytical Methods in Mathematical Physics, R. P. Gilbert, R. G. Newton, eds, Gordon and Breach, New York.

———, (1974), *Asymptotic evaluation of integrals involving a fractional derivative*, this Journal, 5, pp. 159–171.

H. E. FETTIS, (1976), *Fourier series expansions for Pearson type* IV *distributions and probabilities*, SIAM J. Appl. Math. 31, pp. 511–518.

D. S. JONES, (1972), *Asymptotic behaviour of an aperture integral*, Proc. Roy. Soc. Edinburgh (A), 71, pp. 9–29.

Y. L. LUKE, (1969), *The Special Functions and Their Approximations*, Academic Press, New York.

F. W. J. OLVER, (1974), *Asymptotics and Special Functions*, Academic Press, New York.

———, (1975), *Unsolved problems in the asymptotic estimation of special functions*, in Theory and Applications of Special Functions, R. Askey, ed., Academic Press, New York.

N. M. TEMME, (1976), *Remarks on a paper of A. Erdélyi*, this Journal, 7, pp. 767–770.

———, (1979), *The asymptotic expansions of the incomplete gamma functions*, this Journal, 10, pp. 757–766.

R. WONG. (1980), *Error bounds for asymptotic expansions of integrals*, SIAM Rev., 22, pp. 401–435.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0006 $01.00/0

ON SOLUTIONS OF ELLIPTIC EQUATIONS SATISFYING MIXED BOUNDARY CONDITIONS*

A. AZZAM† AND E. KREYSZIG†

Abstract. We consider the mixed boundary value problem for linear second order elliptic equations in a plane domain Ω whose boundary has corners, and obtain conditions sufficient for the solution to be in $C^{2+\alpha}(\bar{\Omega})$, where $0<\alpha<1$. This result means that under those conditions, solutions are as smooth as they would be in the absence of corners, so that, in this sense, the present result is best possible.

1. Introduction. We shall be concerned with the mixed boundary value problem for second order linear elliptic equations in a two-dimensional domain whose boundary has corners. More specifically, we shall study the effect of these corners on the Hölder smoothness of solutions. To motivate this investigation, we first give a general orientation about the development and present situation in this field, beginning with mixed problems in domains with a smooth boundary and then turning to the case of domains with corners on the boundary.

Early results on the regularity of solutions of boundary value problems concern *domains with a smooth boundary*, first for the Laplace and Poisson equations and corresponding Dirichlet and Neumann problems, and later for general second order elliptic equations as well as general boundary conditions. In particular, the *mixed problem* was first considered by Zaremba [54], and is often called *Zaremba's problem.* Further work on the mixed problem in domains with a smooth boundary up to about 1970 is reviewed by Miranda [32] (and a few additional references are given in [18] and [30]), so that it will suffice to mention some of the major contributions during that period and add an outline of some more recent basic results not yet included in any monograph. Of course, we shall be able to select only a small number of articles from the very extensive literature in the field.

Work by Signorini [41], almost contemporary with that of Zaremba, and similar results by Keldysh and Sedov [24] concern the mixed problem for harmonic functions in a half-plane. Slightly earlier than the latter two authors, Giraud [19] proposed a method of solving the mixed problem by first converting it to a Neumann problem on some Riemannian manifold. In 1949, Fichera [14] (cf. also [15]) proved a general existence theorem by transforming the problem into a system of Riesz–Fischer equations; this is known as Picone's method and is also of interest in numerical analysis. Direct methods of the calculus of variations were applied to the mixed problem by Stampacchia [45], whose results are particularly important since they also concern nonlinear equations. The existence of Hölder continuous solutions was proved by Miranda [31], using Schauder type estimates. The method of integral equations was first applied successfully to the mixed problem by Vekua [47]. See also Muskhelishvili [33], whose references reflect the development of that method until about 1955.

Beginning with a paper by Schechter [40], some of the work on (interior and boundary) regularity of solutions of the mixed problem is based on the Sobolev space approach and the use of coercive quadratic forms. For the general idea and setting (which also apply to other "non-Dirichlet problems"), we refer to Agmon [1]. An important contribution specifically devoted to the mixed problem is the thesis by Purmonen [37], which also contains numerous references. Purmonen's work concerns

* Received by the editors September 8, 1980, and revised form March 13, 1981.
† Department of Mathematics, University of Windsor, Windsor, Ontario, Canada N9B 3P4.

rather general mixed problems for linear elliptic equations in n variables, and his results include conditions sufficient for a priori regularity of strong solutions as well as for the existence and some regularity properties of weak solutions. Subsequently, Purmonen [38] studied the well-posedness of the two-dimensional mixed problem in Sobolev spaces. Another approach is the conversion of mixed problems to Wiener–Hopf type problems; this is known as Peetre's method and has recently been extended by Pryde [36].

It is clear that mixed problems in domains with smooth boundary are of great interest in physics, and there also exists an extensive literature on corresponding numerical methods. We cannot go into details, but want to mention that some of the references on domains with corners given below also include the case of domains with a smooth boundary; for further applications we refer to Sneddon [43] and a recent paper by Wendland, Stephan and Hsiao [52] on harmonic functions in two variables, in which two Fredholm equations resulting from the integral equation method are solved constructively, using finite element functions augmented by singular functions, an approach which would be difficult to extend to equations with *variable* coefficients, as is known (cf. Grisvard [20, p. 215] and Kawohl [23]).

We now turn to boundary value problems for *domains with corners on the boundary*. The interest in those problems and in regularity properties of corresponding solutions has several sources. The earliest impetus came from conformal mapping and boundary value problems for harmonic functions; see, for instance, Carleman [11], Kellogg [25] or Warschawski [51]. More recent results of importance, pertaining to the Laplace and Poisson equations in domains with corners, are those by Fufaev [17], Nikol'skiĭ [34] and Volkov [49]. Just as in the case of a smooth boundary, in addition to methods related to Hölder classes, as employed in the present paper, there are other approaches; we mention in particular Sobolev space methods as considered in the reviews by Grisvard [20], [21], then Kondratiev's extension [26] of the Sobolev–Slobodeckii space method by Eskin [13] and Višik [48] (cf. also [3]), furthermore a function theoretic approach by Lewy [29] and his school (see, for instance, Wigley [53]) and, finally, a recent method by Simon [42] based on geometric measure theory.

As a second source for the interest in regularity properties of solutions of boundary value problems in domains with corners we mention physical applications. In fact, it was recognized early that those investigations are important in connection with practical problems in heat conduction, fluid flow and elasticity theory; examples can be found in [12], [44] and other standard monographs. See also [43] and [52].

Thirdly, those problems play a role in numerical analysis, particularly in the study of the accuracy of finite element and finite difference approximations, acceleration of convergence, general convergence analysis, subtraction of singularities and other numerical techniques. Here, in error estimates and other tasks, one often experiences great difficulties caused by the presence of corners, since there may not exist an adequate theory covering such cases. Moreover, in this area there are various traps for the unwary; for a typical example, see [49, p. 157]. For the finite element method, a general characterization of the situation is given by Strang and Fix [46, Chap. 8]. More details are discussed by Fix, Gulati and Wakoff [16] as well as Babuska and Aziz [9, Chaps. 8, 9]; see also Babuska [8] and Babuska and Rosenzweig [10], who use the weighted Sobolev space approach. For the finite difference method, convergence in domains with nonsmooth boundary is studied in basic papers by Laasonen [27], [28]. For a combination of that method with the integral equation method and conformal mapping in the case of the two-dimensional Laplace equation, see Papamichael and Symm [35]. In accelerating convergence, rather natural ideas seem

to be the refinement of meshes near corners where convergence becomes poor and the choice of a net that confines slow convergence to small neighborhoods of corners, instead of "polluting" the whole domain; cf. Volkov [50] for finite differences, and the recent work of Schatz and Wahlbin [39] for finite elements in the case of the Poisson equation in the plane. Ref. [39] includes local estimates of convergence rates up to the boundary, estimates of the effect of systematic refinements and calculation procedures for stress intensity factors as well as the location of the maximum error.

Before we start on our actual problem, let us add a few words about the case of a smooth boundary as compared to that of a boundary with corners. The smoothness of solutions depends on that of the coefficients of the equation, of the boundary of the domain and of the boundary data. It is well known that if in a domain Ω with sufficiently smooth boundary, the regularity properties of the coefficients of the equation and of the boundary data improve, so do the regularity properties on $\bar{\Omega}$ of the solution of the first, second and third boundary value problems. This was first shown for special equations (Laplace and Poisson) and later for general elliptic equations; see Agmon, Douglis and Nirenberg [2]. However, the situation changes drastically in the case of corners at the boundary. Then the smoothness of solutions also depends on the interior angle at the corners. Roughly speaking, small angles are favorable with respect to smoothness of solutions. In addition, there also exist "exceptional angles" for which the smoothness is "exceptionally good", that is, is better than for values of the angle close to those exceptional ones. Our result will be typical in that respect, since it will illustrate this general pattern. We shall find conditions sufficient for the solutions to be as smooth as they would be in the absence of corners, the other conditions remaining the same; *hence our conclusion will be strongest possible.*

2. Problem and main result. We shall consider linear elliptic equations of the form

$$Lu = \sum_{i=1}^{2} \sum_{j=1}^{2} a_{ij}(x) u_{x_i x_j} + \sum_{i=1}^{2} a_i(x) u_{x_i} + a(x)u = f(x) \tag{2.1}$$

in a plane domain Ω whose boundary $\partial\Omega$ has corners. Here, $x = (x_1, x_2)$. We assume that Ω is simply connected and bounded and L is uniformly elliptic in Ω. The boundary conditions are of mixed type; we write them in the form

$$\chi_1(x)u(x) + \chi_2(x)u_n(x) = \chi_1(x)\phi(x) + \chi_2(x)\psi(x) \quad \text{on } \partial\Omega; \tag{2.2}$$

here, the subscript n denotes the outer normal derivative.

The following result in the "regular case" is well known. If $\partial\Omega$ is smooth of class $C^{2+\alpha}$, where $0 < \alpha < 1$, and if

(A) $a_{ij}, a_i, a, f \in C^{\alpha}(\bar{\Omega})$, L uniformly elliptic in Ω,

(B) $\chi_1, \phi \in C^{2+\alpha}(\partial\Omega)$, $\chi_2, \psi \in C^{1+\alpha}(\partial\Omega)$,

then

$$u \in C^{2+\alpha}(\bar{\Omega}). \tag{2.3}$$

See Agmon, Douglis and Nirenberg [2].

We now turn to the case when $\partial\Omega$ is not smooth, a case which we also considered in [5] and [7]. Then [2] implies that in a compact subregion Ω_1 of $\bar{\Omega}$ with positive distance from the corner points, u is smooth as before. More precisely we have the following. Without loss of generality we may assume that $\partial\Omega$ has a single corner, which is located at the origin $x = 0$, the interior angle being γ, $0 < \gamma < 2\pi$. Let Γ_1 and Γ_2 denote the two arcs of $\partial\Omega$ that form the corner at $x = 0$. Suppose that $\partial\Omega \backslash \{0\}$ is

smooth of class $C^{2+\alpha}$. Let u be a solution of (2.1) satisfying the boundary conditions (2.2). Further, assume that conditions (A) and

$$\text{(B*)} \qquad \begin{aligned} &\chi_1, \phi \in C^{2+\alpha}(\partial\Omega\backslash\{0\}), \\ &\chi_2, \psi \in C^{1+\alpha}(\partial\Omega\backslash\{0\}), \end{aligned}$$

as well as

$$\text{(C)} \qquad \begin{aligned} &\chi_2 = 0, \quad \phi = 0 \quad \text{on } \Gamma_1, \\ &\chi_1 = 0, \quad \psi = 0 \quad \text{on } \Gamma_2, \end{aligned}$$

hold true. Then, by [2],

$$u \in C^{2+\alpha}(\Omega_1) \cap C^0(\bar{\Omega}), \tag{2.4}$$

with Ω_1 as indicated before.

To characterize the smoothness of u near the corner point, we introduce

$$\omega = \arctan \frac{[a_{11}(0)a_{22}(0) - a_{12}^2(0)]^{1/2}}{a_{22}(0)\cot\gamma - a_{12}(0)}. \tag{2.5}$$

This is the angle obtained from γ in the transformation of the equation

$$\sum_{i=1}^{2}\sum_{j=1}^{2} a_{ij}(0)u_{x_i x_j} = 0$$

to normal form. In [7] we proved that, under assumptions (A), (B*), (C) and $\omega < \pi/2$, we have

$$u \in C^{\nu}(\bar{\Omega}), \qquad \nu = \min\left(\frac{\pi}{2\omega} - \varepsilon, 2\right), \tag{2.6}$$

with arbitrarily small $\varepsilon > 0$. Substantially improving that result, we shall now obtain sufficient conditions in order that even (2.3) be valid; those conditions will concern small angles as well as an exceptional angle $(\pi/4)$. Note well that (2.3) refers to the "regular case" of a smooth boundary. Accordingly, despite the presence of corners, *our result to be obtained is as strong as that in the case of the absence of corners*; in that sense, this result is best possible.

Our main result can be stated as follows.

THEOREM 1. *Let u be a bounded solution of* (2.1), (2.2) *in* Ω. *Suppose that* (A), (B*), (C) *hold true and* ω *in* (2.5) *satisfies the condition*

(D_1) $\omega < \pi/(4+2\alpha)$

or the condition

(D_2) $\omega = \pi/4$.

Then

$$u \in C^{2+\alpha}(\bar{\Omega}). \tag{2.7}$$

From the statement involving (2.4), we conclude that it suffices to prove Theorem 1 in

$$N = \{x \mid x \in \bar{\Omega}, |x| < r_0\}, \qquad r_0 > 0.$$

Furthermore, by [7] it is sufficient to consider the case of a circular sector and impose the additional condition

$$a_{ij}(0) = \delta_{ij}, \qquad i, j = 1, 2.$$

Indeed, the transition from this special setting to the general case is the same as in [7] (and is relatively simple), so that we need not reproduce it here.

At this point, we should notice that [7] concerns arbitrary n, whereas here we take $n=2$ because later (near the end of the paper) we have to use a result by Volkov which is known to hold for $n=2$ only. Actually, we need Volkov's result only in connection with condition (D_2), so that the assertion of Theorem 1 under condition (D_1) could be proved for any n by an argument similar to the present one.

3. The case of a sector. Let r, θ be polar coordinates defined by $x_1 = r\cos\theta$, $x_2 = r\sin\theta$ and consider the sector

$$\Omega_{2\sigma} = \{(r,\theta) \mid 0<r<2\sigma,\ 0<\theta<\omega\},$$

where $\sigma = \text{const} > 0$. Let

$$\Gamma_1\colon \theta = 0, \quad r<2\sigma, \qquad \Gamma_2\colon \theta = \omega, \quad r<2\sigma.$$

A theorem analogous to Theorem 1 but referring to the present setting can be stated as follows.

THEOREM 2. *Let u be a bounded solution of the mixed boundary value problem for the equation*

$$Lu = f \quad \text{in } \Omega_{2\sigma}, \tag{3.1}$$

with L as in (2.1) and $a_{ij}(0) = \delta_{ij}$ and assume that u satisfies the conditions

$$\text{(a) } u|_{\Gamma_1} = 0, \qquad \text{(b) } u_n|_{\Gamma_2} = 0. \tag{3.2}$$

Suppose that (A) *with Ω replaced by $\Omega_{2\sigma}$ and (D_1) or (D_2) hold. Then*

$$u \in C^{2+\alpha}(\bar{\Omega}_\sigma). \tag{3.3}$$

By what has been said, in order to obtain Theorem 1, it suffices to prove Theorem 2. The proof of the latter theorem will result from two lemmas.

In the first lemma, we obtain bounds for u and its first and second partial derivatives as well as a statement on the Hölder smoothness of u. Here, $D^k u$ denotes any kth partial derivative of u.

LEMMA 1. *Under the assumptions of Theorem 2 we have in $\bar{\Omega}_\sigma$*

$$\begin{aligned} &\text{a)} && |Du(x)| \leqq Mr^{\nu}, \\ &\text{b)} && |D^k u(x)| \leqq Mr^{\nu-k}, \qquad k = 1, 2, \\ &\text{c)} && u \in C^{\nu}(\bar{\Omega}_\sigma), \end{aligned} \tag{3.4}$$

where

$$\nu = \min\left(\frac{\pi}{2\omega} - \varepsilon,\ 2+\alpha\right).$$

Proof. a) We consider in $\bar{\Omega}_{2\sigma}$ the function

$$w(x) = Mr^{\nu}\cos\lambda(\omega-\theta),$$

with ν defined as in the lemma, $\lambda = (\pi - 2\delta)/2\omega$ and $\delta > 0$ so small that $\lambda > \nu$. Using the method developed in [7], one can show that w may serve as a barrier function for u, provided M is taken sufficiently large. In this way we obtain (3.4a).

b), c) From [5] it can be seen that in the case of the *Dirichlet problem*, the proof of the statements corresponding to our present (3.4b) and (3.4c) depends mainly on

the analog of our present (3.4a) and on a Schauder estimate of the form

$$\|u\|_{2+\alpha}^{\Omega^*} \leqq \kappa[\|u\|_0^{\Omega} + \|f\|_{\alpha}^{\Omega} + \|\phi\|_{2+\alpha}^{\Gamma}],$$

where $\bar{\Omega}^* \subset \bar{\Omega}$ and $\Gamma = \partial\Omega \cap \partial\Omega^*$ is of class $C^{2+\alpha}$. Such an estimate also holds for the *mixed* boundary value problem, the only difference being the absence of the last term. In this way, following the general idea in [5], we obtain (3.4b) and (3.4c). This completes the proof.

From (3.4c) it follows that Theorem 2 with condition (D_1) holds. Finally, we must prove Theorem 2 under condition (D_2). If (D_2) holds, then (3.4b) yields

$$|D^2 u(x)| \leqq M r^{-\varepsilon} \quad \text{in } \bar{\Omega}_\sigma$$

and (3.4c) gives

$$u \in C^{2-\varepsilon}(\bar{\Omega}_\sigma).$$

To prove Theorem 2 in the present case, we first investigate the nature of singular behavior of the second derivatives of u near the corner point.

LEMMA 2. *Let v be a solution of* (3.1) *in* $\Omega_{2\sigma}$ *satisfying* (3.2), *and suppose that the assumptions of Theorem* 2 *hold true. Suppose further that in* $\bar{\Omega}_\sigma$

$$|D^2 v(x)| \leqq M_1 r^{-\eta}, \qquad 0 \leqq \eta < 1.$$

Let $h \in C^{\tau}(\bar{\Omega}_\sigma)$, *where* $1 > \tau \geqq \eta$ *and* $h(0) = 0$. *Then*

$$hD^2 v \in C^{\mu}(\bar{\Omega}_\sigma), \qquad \mu = \min(\alpha, \tau - \eta). \tag{3.5}$$

Proof. In Ω_σ consider any two points $P_j : (r_j, \theta_j)$, $j = 1, 2$. By abuse of notation, we write $h(P_j)$ for $h(r_j, \theta_j)$ and so on. We must show that there exists a constant $H > 0$ such that

$$d(P_1, P_2)^{-\mu} |h(P_1) D^2 v(P_1) - h(P_2) D^2 v(P_2)| \leqq H. \tag{3.6}$$

Let $0 \leqq r_2 \leqq r_1 \leqq \sigma$, without restriction. If $r_2 \leqq \frac{1}{2} r_1$, then $d(P_1, P_2) \geqq \frac{1}{2} r_1$, and from

$$|h(P_j)| \leqq M_2 r_j^{\tau}, \qquad j = 1, 2,$$

we can obtain (3.6).

We consider the case $r_2 > \frac{1}{2} r_1$. Let

$$x = \xi y, \quad \xi = \frac{2r_1}{\sigma}, \quad y = (y_1, y_2).$$

This transformation maps

$$\Omega_0 = \{(r, \theta) | \tfrac{1}{2} r_1 \leqq r \leqq r_1, 0 < \theta < \tfrac{1}{4}\pi\}$$

onto

$$\Omega_1 = \{(\rho, \theta) | \tfrac{1}{4}\sigma \leqq \rho \leqq \tfrac{1}{2}\sigma, 0 < \theta < \tfrac{1}{4}\pi\},$$

where $\rho = r/\xi$. As in [6], it can be shown that in Ω_1 the function $V(y) = v(\xi y)$ satisfies

$$\|V\|_{2+\alpha}^{\Omega_1} \leqq M_3 r_1^{2-\eta}.$$

Now, for any $\mu \leqq \alpha$,

$$\xi^{2+\mu} H_{\mu}^{\Omega_0}(D^2 v) = H_{\mu}^{\Omega_1}(\tilde{D}^2 V) \leqq M_4 r_1^{2-\eta},$$

where $\tilde{D}^2 V$ denotes the partial derivative corresponding to $D^2 v$ and $H_{\mu}^{\Omega_0}$ is the Hölder coefficient. Hence,

$$H_{\mu}^{\Omega_0}(D^2 v) \leqq M_5 r_1^{-\mu-\eta}.$$

We now obtain (3.6) in the case $r_2 > \frac{1}{2}r_1$ as follows, writing $\delta = d(P_1, P_2)$:

$$\begin{aligned}
&|h(P_1)D^2v(P_1) - h(P_2)D^2v(P_2)|\delta^{-\mu} \\
&\quad \leqq |h(P_1)||D^2v(P_1) - D^2v(P_2)|\delta^{-\mu} \\
&\qquad + |D^2v(P_2)|\{|h(P_1) - h(P_2)|\delta^{-\tau}\}^{\mu/\tau}|h(P_1) - h(P_2)|^{1-\mu/\tau} \\
&\quad \leqq M_2 r_1^{\tau} M_5 r_1^{-\mu-\eta} + M_1 r_2^{-\eta} M_6 (2M_2 r_1^{\tau})^{1-\mu/\tau} \\
&\quad \leqq H.
\end{aligned}$$

This proves Lemma 2.

We can now prove Theorem 2 under assumption (D_2). We remember that by Lemma 1, under the assumptions of the theorem [with (D_1) or (D_2)] we have

$$u \in C^{2-\varepsilon}(\bar{\Omega}_\sigma)$$

and in $\bar{\Omega}_\sigma$

$$|D^2u(x)| \leqq Mr^{-\varepsilon},$$

as was stated above. Equation (3.1) can be written

$$\Delta u = f_1 = f - au - \sum_{i=1}^{2} a_i u_{x_i} - \sum_{i=1}^{2}\sum_{j=1}^{2} (a_{ij} - \delta_{ij}) u_{x_i x_j}. \tag{3.7}$$

Since f, a, $a_i \in C^{\alpha}(\bar{\Omega}_\sigma)$ and $u \in C^{2-\varepsilon}(\bar{\Omega}_\sigma)$, the first three expressions on the right-hand side of (3.7) are of class $C^{\alpha}(\bar{\Omega}_\sigma)$. Using Lemma 2 with

$$h = a_{ij} - \delta_{ij}, \quad \tau = \alpha, \quad \eta = \varepsilon,$$

we have

$$(a_{ij} - \delta_{ij}) u_{x_i x_j} \in C^{\alpha-\varepsilon}(\bar{\Omega}_\sigma).$$

Hence, $f_1 \in C^{\alpha-\varepsilon}(\bar{\Omega}_\sigma)$. From this and [49, p. 128], it follows that $u \in C^{2+\alpha-\varepsilon}(\bar{\Omega}_\sigma)$. Using this in the last term of (3.7) and applying again Lemma 2, with $\tau = \alpha$ and $\eta = 0$, we obtain (3.3). This completes the proof of Theorem 2.

From Theorem 2, our main result (Theorem 1) follows as indicated in § 2.

Acknowledgment. We want to thank the referee for helpful comments, in particular, for bringing basic literature to our attention and suggesting it for inclusion in the Introduction; this has led to a substantial improvement of the latter.

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, New York, 1965.

[2] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[3] A. ALVINO AND G. TROMBETTI, *The Dirichlet problem in a cone of R^n in L^p*, Ann. Mat. Pura Appl., (4), 120 (1979), pp. 269–291.

[4] A. K. AZIZ, ed., *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, 1972.

[5] A. AZZAM, *On Dirichlet's problem for elliptic equations in sectionally smooth n-dimensional domains*, this Journal, 11 (1980), pp. 248–253.

[6] ———, *The Dirichlet problem for linear elliptic equations in plane domains with corners*, Ann. Polon. Math., in press.

[7] A. AZZAM AND E. KREYSZIG, *Über das gemischte Randwertproblem für elliptische Gleichungen in n-dimensionalen Gebieten mit Kanten*, Ann. Acad. Sci. Fenn., in press.
[8] I. BABUSKA, *Finite element method for domains with corners*, Computing, 6 (1970), pp. 264–273.
[9] I. BABUSKA AND A. K. AZIZ, *Survey lectures on the mathematical foundations of the finite element method*, in [4], pp. 1–359.
[10] I. BABUSKA AND M. B. ROSENZWEIG, *A finite element scheme for domains with corners*, Numer. Math., 20 (1972), pp. 1–21.
[11] T. CARLEMAN, *Über das Neumann-Poincarésche Problem für ein Gebiet mit Ecken*, Dissertation, Uppsala, 1916.
[12] H. C. CARSLAW AND J. C. JAEGER, *Conduction of Heat in Solids*, 2nd ed., Clarendon Press, Oxford, 1959.
[13] G. I. ESKIN, *Boundary Problems for Elliptic Pseudo-Differential Operators*, Nauka, Moscow, 1973. (In Russian.)
[14] G. FICHERA, *Analisi esistenziale per le soluzioni dei problemi al contorno misti, relativi all'equazione e ai sistemi di equazioni del secondo ordine di tipo ellittico, autoaggiunti*, Ann. Scuola Norm. Super. Pisa, (3), 1 (1949), pp. 75–100.
[15] ———, *Sul problema della derivata obliqua e sul problema misto per l'equazione di Laplace*, Boll. Un. Mat. Ital., (3), 7 (1952), pp. 367–377.
[16] G. J. FIX, S. GULATI AND G. I. WAKOFF, *On the use of singular functions with finite element approximations*, J. Comp. Physics, 13 (1973), pp. 209–228.
[17] V. V. FUFAEV, *Conformal mappings of domains with corners and differential properties of solutions of the Poisson equation in domains with corners*, Soviet Math. Dokl., 4 (1963), pp. 1457–1459.
[18] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 1977.
[19] G. GIRAUD, *Problèmes mixtes et problèmes sur de variétés closes, relativement aux équations linéaires du type elliptique*, Ann. Soc. Polon. Math., 12 (1933), pp. 35–54.
[20] P. GRISVARD, *Behavior of the solutions of an elliptic boundary value problem in a polygonal or polyhedral domain*, in [22], pp. 207–274.
[21] ———, *Boundary value problems in non-smooth domains*, Lecture Notes 19, Department of Mathematics, University of Maryland, College Park, MD, 1980.
[22] B. HUBBARD, ed., *Numerical Solutions of Partial Differential Equations*, III, SYNSPADE 1975, Academic Press, New York, 1976.
[23] B. KAWOHL, *Über nichtlineare gemischte Randwertprobleme für elliptische Differentialgleichungen zweiter Ordnung auf Gebieten mit Ecken*, Dissertation, Technische Hochschule, Darmstadt, 1978.
[24] M. KELDYSH AND L. SEDOV, *Sur la solution effective de quelques problèmes limites pour les fonctions harmoniques*, Doklady Akad. Nauk SSSR, 16 (1937), pp. 7–10.
[25] O. D. KELLOGG, *On the derivatives of harmonic functions on the boundary*, Trans. Amer. Math. Soc., 33 (1931), pp. 486–509.
[26] V. A. KONDRATIEV, *Boundary problems for elliptic equations in domains with conical or angular points*, Trans. Moscow Math. Soc., (1967), pp. 227–313.
[27] P. LAASONEN, *On the degree of convergence of discrete approximations for the solutions of the Dirichlet problem*, Ann. Acad. Sci. Fenn. Ser. A. I., 246 (1957), 19 pp.
[28] ———, *On the solution of Poisson's difference equation*, J. Assoc. Comput. Mach., 5 (1958), pp. 370–382.
[29] H. LEWY, *Developments at the confluence of analytic boundary conditions*, Univ. of California Publ. Math., 1 (1950), pp. 247–280.
[30] J.-L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, 3 vols., Springer, Berlin, 1972–1973.
[31] C. MIRANDA, *Sul problema misto per le equazioni lineari ellittiche*, Ann. Mat. Pura Appl., 39 (1955), pp. 279–303.
[32] ———, *Partial Differential Equations of Elliptic Type*, 2nd ed., Springer, Berlin, 1970.
[33] N. I. MUSKHELISHVILI, *Singular Integral Equations*, reprinted, Wolters–Noordhoff, Groningen, 1972.
[34] S. M. NIKOL'SKIĬ, *Boundary properties of functions defined on a region with angular points*, Mat. Sbornik, 43 (1957), pp. 127–144. (In Russian.)
[35] N. PAPAMICHAEL AND G. T. SYMM, *Numerical techniques for two-dimensional Laplacian problems*, Comput. Meth. Appl. Mech. Engrg., 6 (1975), pp. 175–194.
[36] A. J. PRYDE, *Second order elliptic equations with mixed boundary conditions*, Bull. Amer. Math. Soc., 83 (1977), pp. 391–393.
[37] V. T. PURMONEN, *Über gemischte koerzitive elliptische lineare partielle Randwertaufgaben*, Ann. Acad. Sci. Fenn. Ser. A. I, Math. Dissertationes 5, 1975.

[38] ———, *Zur Korrektheit elliptischer fastlinearer gemischter Randwertaufgaben in* R^2, Ann. Acad. Sci. Fenn. Ser. A. I, Math., 3 (1977), pp. 277–300.

[39] A. H. SCHATZ AND L. B. WAHLBIN, *Maximum norm estimates in the finite element method on plane polygonal domains*, I, II, Math. Comp., 32 (1978), pp. 73–109, and 33 (1979), pp. 465–492.

[40] M. SCHECHTER, *Mixed boundary problems for general elliptic equations*, Comm. Pure Appl. Math., 13 (1960), pp. 183–201.

[41] A. SIGNORINI, *Sopra un problema al contorno nella teoria delle funzioni di variabile complessa*, Ann. Mat. Pura Appl., (3), 25 (1916), pp. 253–273.

[42] L. SIMON, *Regularity of capillary surfaces over domains with corners*, Res. Report, 10, University of Melbourne, Parkville, 1979.

[43] I. N. SNEDDON, *Mixed Boundary Value Problems in Potential Theory*, John Wiley, New York, 1966.

[44] I. S. SOKOLNIKOFF, *Mathematical Theory of Elasticity*, 2nd ed., McGraw-Hill, New York, 1956.

[45] G. STAMPACCHIA, *Problemi al contorno misti per equazioni del calcolo delle variazioni*, Ann. Mat. Pura Appl., (4) 40 (1955), pp. 193–209.

[46] G. STRANG AND G. J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[47] I. N. VEKUA, *New Methods for Solving Elliptic Equations*, John Wiley, New York, 1967.

[48] M. I. VIŠIK, *Sobolev–Slobodeckii spaces of variable order with weighted norms and their applications to mixed elliptic boundary problems*, Amer. Math. Soc. Transl., (2), 105 (1976), pp. 104–110.

[49] E. A. VOLKOV, *Differentiability properties of solutions of boundary value problems for the Laplace equation on a polygon*, Proc. Steklov Inst. Math., 77 (1965), pp. 127–159.

[50] ———, *The net-method for finite and infinite regions with a piecewise boundary*, Soviet Math. Dokl., 7 (1966), pp. 744–747.

[51] S. WARSCHAWSKI, *Ueber das Randverhalten der Ableitung der Abbildungsfunktion bei konformer Abbildung*, Math. Z., 35 (1932), pp. 321–456.

[52] W. L. WENDLAND, E. STEPHAN AND G. C. HSIAO, *On the integral equation method for the plane mixed boundary value problem of the Laplacian*, Math. Meth. Appl. Sci., 1 (1979), pp. 265–321.

[53] N. M. WIGLEY, *Mixed boundary value problems in plane domains with corners*, Math. Z., 115 (1970), pp. 33–52.

[54] S. ZAREMBA, *Sur un problème mixte relatif a l'équation de Laplace*, Bull. Int. Acad. Cracovie, Classe Math. Nat., Ser. A, (1910), pp. 313–344.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0007 $01.00/0

LAGUERRE SERIES AS BOUNDARY VALUES*

AHMED I. ZAYED†

Abstract. Let I denote the half-line $(0,\infty)$ and let $Q=\{(x,y)\,|\,x>0,\ y>0\}$. Define $S_I=\{\phi\,|\,\phi\in C^\infty(I),\ \sup_{x\in I}|x^p(d^q\phi(x)/dx^q)|<\infty\}$. A topology is defined on S_I by means of a countable family of seminorms. It is shown that $f\in S_I^*$ (the dual space of S_I) if and only if $f(x)=\sum_{n=0}^{\infty}a_nU_n(x)$ with $a_n=O(n^p)$ for some integer p, where $\{U_n(x)\}_{n=0}^{\infty}$ are Laguerre functions; $U_n(x)$ satisfies the Laguerre differential equation $xy''+y'+(n+\frac{1}{2}-x/4)y=0$. We use this result to construct a function $f(x,y)$ such that, for a continuous function $f\in S_I^*$, we have (i) $f(x,y)$ is harmonic in Q; (ii) $g(y)=\lim_{x\to 0+}f(x,y)$ is finite for $0<y<\infty$; (iii) $\lim_{y\to 0^+}f(x,y)=f(x)$ (in the sense of S_I^*). In addition, if we require that $\lim_{x\to 0^+}f(x)=c<\infty$, then $\sup_{0\leqq y<\infty}|g(y)|=M<\infty$ and $\lim_{y\to 0^+}f(x,y)=f(x)$, $0\leqq x<\infty$. In fact, the convergence is uniform on compact subsets of $(0,\infty)$. The construction of $f(x,y)$ uses Laguerre functions of the second kind, i.e., solutions of the Laguerre differential equation that vanish at $-\infty$.

Condition (i) is weakened if f is an arbitrary element of S_I^*.

1. Introduction. The connection between trigonometric Fourier series and boundary values of holomorphic and harmonic functions in the unit disk is very well known. There is a vast literature dealing with various relations and aspects of this subject. However, if one moves to the half-line $(0,\infty)$, no parallel theory based on series seems to exist. A natural orthonormal sequence of functions to use on the half-line is Laguerre functions.

The only known (to me) attempt to study Hermite and Laguerre expansions in analogy to ordinary Fourier series is the work of Muckenhoupt [3], [4], [5] and Walter [9]. Among other things, Muckenhoupt derived a representation for the Poisson integral $f(x,y)$ of a function $f(x)$ which belongs to a certain function space and for which the Laguerre series expansion exists. He also showed that $\lim_{y\to 0}f(x,y)=f(x)$ almost everywhere as well as in the mean, and that $f(x,y)$ satisfies the second order elliptic differential equation

$$f_{yy}+xf_{xx}+(1-x)f_x=0.$$

In [5] a notion of conjugacy was introduced by which one can define a conjugate Poisson integral $\tilde{f}(x,y)$ and hence a conjugate function $\tilde{f}(x)$. Moreover, $\tilde{f}(x,y)$ is shown to satisfy the differential equation

$$\tilde{f}_{yy}+x\tilde{f}_{xx}+(1-x)\tilde{f}_x-\left(\frac{1}{4x}+\frac{1}{2}\right)\tilde{f}=0.$$

This notion of conjugacy, while not as decisive as the classical one, still has some significant properties (see Vekua [7]).

One of the purposes of this paper is to extend some of Muckenhoupt's results to a wider class of functions. In fact, we shall be more interested in the case where the boundary values $f(x)$ are not given by a function but rather by a "generalized function". More specifically, we shall confine ourselves to the cases where $f(x)$ is in the dual space S_I^* of a certain testing-function space S_I that will be defined in the following section.

* Received by the editors January 23, 1980, and in final revised form March 10, 1981. This work formed part of the author's Ph.D. dissertation at the University of Wisconsin, Milwaukee.

† Department of Mathematics, California Polytechnic State University, San Luis Obispo, California 93407.

It should be pointed out that our approach has to differ considerably from Muckenhoupt's since pointwise convergence does not necessarily imply convergence in the sense of S_I^*. Clearly, the straightforward approach of continuing series of Laguerre functions to the complex plane fails because a series of aguerre functions which may converge uniformly on the positive real axis, together with all its derivatives, can diverge everywhere off that part of the axis. So it seems inevitable that one has to approach the problem in an indirect way. We do this by introducing series of Laguerre functions of the second kind $\{v_n(z)\}$ which are obtained from the analytic representation of Laguerre functions $\{u_n(x)\}$. We shall show that these $v_n(z)$'s have the important property of being analytic in the complex plane cut along the positive real axis, and their jumps across that part of the axis are the $u_n(x)$'s.

Naturally, one is led to considering the problem of characterizing the elements of the space S_I^* in terms of their Laguerre coefficients. Indeed, this is a special case of the general theory of expanding generalized functions in series of orthonormal functions. Among those who have contributed to the development of this theory are Korevaar [2], Zemanian [11] and Walter [8]. Although it is possible to find such a characterization directly for the space S_I^*, we shall use some of the already known results proved by Zemanian [11].

Zemanian has given a characterization for the elements of a certain testing function space $A(R, I)$, which depends on the differential operator R and the interval I, in terms of their expansions with respect to the eigenfunctions of the operator R. In a later work [12] he showed that for the Hermite differential operator R and $I = (-\infty, \infty)$ the space $A^*(R, (-\infty, \infty))$ is actually the space of tempered distributions. Analogous to Zemanian's result we shall show that for the Laguerre differential operator R and $I = (0, \infty)$ the space $A^*(R, (0, \infty))$ is indeed the space S_I^*.

2. Definitions and notation. In this section we list some properties and formulas pertaining to Laguerre polynomials and Laguerre functions that will be used later (see Szegö [6]). The spaces $A(R, I)$ and S_I are defined, and Zemanian's main theorem is cited for reference since it will be heavily used in the sequel.

Laguerre polynomials $L_n^\alpha(x)$ are defined by

$$L_n^\alpha(x) = e^x \frac{x^{-\alpha}}{n!} \frac{d^n}{dx^n}(e^{-x}x^{n+\alpha}), \qquad n = 0, 1, 2, \cdots, \tag{2.1}$$

with real $\alpha > -1$, and Laguerre functions $u_n^\alpha(x)$ are defined by

$$u_n^\alpha(x) = \left(\frac{n!}{\Gamma(n+\alpha+1)}\right)^{1/2} e^{-x/2}x^{\alpha/2}L_n^\alpha(x). \tag{2.2}$$

We will write $u_n(x)$ instead of $u_n^0(x)$. It is known that $L_n^\alpha(x)$ satisfies the differential equation

$$\left(x\frac{d^2}{dx^2} + (\alpha+1-x)\frac{d}{dx} + n\right)L_n^\alpha(x) = 0, \tag{2.3}$$

while $u_n^\alpha(x)$ satisfies

$$\left[x\frac{d^2}{dx^2} + \frac{d}{dx} + \left(n + \frac{\alpha+1}{2} - \frac{x}{4} - \frac{\alpha^2}{4x}\right)\right]u_n^\alpha(x) = 0. \tag{2.4}$$

Laguerre functions $\{u_n^\alpha(x)\}_{n=0}^\infty$ form a complete orthonormal system on the interval $(0, \infty)$; i.e., $\int_0^\infty u_n^\alpha(x)u_m^\alpha(x)\,dx = \delta_{nm}$. Equivalently,

$$\int_0^\infty e^{-x}x^\alpha L_n^\alpha(x)L_m^\alpha(x)\,dx = \frac{\Gamma(n+\alpha+1)}{n!}\delta_{nm}. \tag{2.5}$$

We also have the estimates

$$L_n^\alpha(x) = \pi^{-1/2}e^{x/2}x^{-\alpha/2-1/4}n^{\alpha/2-1/4}\cos\left[2(nx)^{1/2} - \frac{\alpha\pi}{2} - \frac{\pi}{4}\right] + O(n^{\alpha/2-3/4}) \tag{2.6}$$

for $x > 0$, and

$$L_n^\alpha(x) = O(n^a), \qquad a = \max\left(\frac{\alpha}{2} - \frac{1}{4}, \alpha\right), \qquad 0 \leqq x \leqq \omega \tag{2.7}$$

(see Szegö [6, formulas (8.22.1) and (7.6.11)]).

Definition of the space $A(R, I)$. Let I denote any open interval (a, b) on the real line where $a = -\infty$, $b = \infty$ are permitted, and let R denote a linear differential operator of the form

$$R = \theta_0 D^{n_1}\theta_1 D^{n_2}\cdots D^{n_\nu}\theta_\nu,$$

where $D = d/dx$, n_k, $k = 1, \cdots, \nu$ are positive integers and θ_k are C^∞-functions on I and never equal to zero anywhere there. We also require that the θ_k and n_k be such that

$$R = \bar{\theta}_\nu(-D)^{n_\nu}\cdots(-D)^{n_2}\bar{\theta}_1(-D)^{n_1}\bar{\theta}_0,$$

where $\bar{\theta}_k$ denotes the complex-conjugate function. Moreover, we assume that there exists a sequence $\{\lambda_n\}_{n=0}^\infty$ of real numbers called the eigenvalues of R, and a sequence $\{\psi_n\}_{n=0}^\infty$ of C^∞-functions in $L^2(I)$ called the eigenfunctions of R, such that $|\lambda_n| \to \infty$ as $n \to \infty$ and $R\psi_n = \lambda_n\psi_n$, $n = 0, 1, 2, \cdots$. The zero function (zero a.e.) is not allowed as an eigenfunction. We also assume that the $\{\psi_n\}_{n=0}^\infty$ form a complete orthonormal system on I.

Now we construct a testing-function space $A(R, I)$ which depends on the choices of I, R and $\{\psi_n\}_{n=0}^\infty$ as follows. A consists of all functions $\phi(x)$ that possess the following three properties:

(i) $\phi(x)$ is defined, complex-valued, and smooth on I; i.e., $\phi(x) \in C^\infty(I)$.

(ii) For all $k = 0, 1, 2, \cdots$ the quantity

$$\|\phi\|_k = \|R^k\phi\|_2 = \left[\int_a^b |R^k\phi_{(x)}|^2\,dx\right]^{1/2} < \infty.$$

(iii) For all n and k, $(R^k\phi, \psi_n) = (\phi, R^k\psi_n)$.

It is easy to see that $\{\|\cdot\|_k\}_{k=0}^\infty$ is a separating family of seminorms on A which is used to define a locally convex topology on A. In fact, Zemanian proved, among other things, that A equipped with this topology is a Fréchet space. Through his proof he showed that if $\{\phi_m\}_{m=1}^\infty$ converges in A, then $\{R^k\phi_m\}_{m=1}^\infty$ for each fixed k converges uniformly on every open interval Ω whose closure is compact in I. In addition to that, for each fixed k, $\{D^k\phi_m\}_{m=1}^\infty$ converges uniformly on Ω.

If R and I are understood from the context we will write A instead of $A(R, I)$ for short. Now we summarize Zemanian's main results in the following theorem.

THEOREM 2.1. *Let $\phi \in A$, $f \in A^*$ (the conjugate space of A), $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ be sequences of complex numbers. Then*

(i) $\phi = \sum_{n=0}^{\infty} (\phi, \psi_n)\psi_n$, *where the series converges in* A.

(ii) $\sum_{n=0}^{\infty} a_n\psi_n$ *converges in* A *if and only if* $\sum_{n=0}^{\infty} |\lambda_n|^{2k}|a_n|^2$ *converges for every nonnegative integer* k.

(iii) $f = \sum_{n=0}^{\infty} (f, \psi_n)\psi_n$, *where the series converges in* A^*.

(iv) $\sum_{n=0}^{\infty} b_n\psi_n$ *converges in* A^* *if and only if there exists a nonnegative integer* q *such that* $\sum_{\lambda_n \neq 0} |\lambda_n|^{-2q}|b_n|^2$ *converges. Furthermore, if* $f = \sum_{n=0}^{\infty} b_n\psi_n$, *then* $b_n = (f, \psi_n)$.

Definition of the space S_I. The space S_I where $I = (0, \infty)$ is defined as follows. S_I consists of all functions $\phi(x)$ with the following properties:

(i) $\phi(x)$ is defined and smooth on I; i.e., $\phi \in C^{\infty}(I)$.

(ii) $\sup_{x \in I} |x^k D^n \phi| < \infty$ for all $k, n = 0, 1, 2, \cdots$.

To define the topology in the space S_I, we introduce a countable system of norms

$$\|\phi\|_p = \sup_{\substack{x \in I \\ 0 \leqq k,q \leqq p}} |x^k D^q \phi|, \qquad p = 0, 1, 2, \cdots.$$

Evidently convergence of the sequence ϕ_ν to ϕ in this topology is equivalent to $D^k\phi_\nu \to D^k\phi$ uniformly in every compact subset of I, $k = 0, 1, 2, \cdots$ together with $|x^k D^q \phi_\nu| \leqq C_{kq}$ independent of ν. Unlike the space A, the space S_I is independent of the differential operator R. Zemanian [12] has shown that for the Hermite differential operator the space A is actually the space S of rapidly decreasing functions. Analogously, we will prove that for Laguerre differential operators the space A is, indeed, the space S_I. Now we are able to state and prove our main theorem.

3. Characterization of the space S_I. This section is devoted to the proof of the following theorem.

THEOREM 3.1. *For the Laguerre differential operator $R = e^{x/2} D\, e^{-x} x D\, e^{x/2}$, $I = (0, \infty)$, $\psi_n = u_n(x)$, we have $S_I = A(R, I)$.*

In general, S_I and A are different; e.g., if $\alpha \neq 0$, we have

$$R = x^{-\alpha/2}\, e^{x/2} D x^{\alpha+1}\, e^{-x} D x^{-\alpha/2}\, e^{x/2}$$

$$= DxD - \frac{x}{4} - \frac{\alpha^2}{4x} + \frac{\alpha+1}{2};$$

then we have $e^{-x} \in S_I$, but $e^{-x} \notin A$ since $\int_0^{\infty} (e^{-x}/x^2)\, dx$ diverges.

To prove the theorem we need the following lemmas.

LEMMA 3.1.

(i) *S_I is closed under R^k $(k = 0, 1, 2, \cdots)$; i.e. if $\phi \in S_I$, then $R^k\phi \in S_I$.*

(ii) $S_I \subset L^2(I)$.

(iii) $(R\phi, \psi) = (\phi, R\psi)$ *for all* $\phi, \psi \in S_I$.

(iv) $S_I \subseteq A$.

Proof.

(i) First observe that $R = e^{x/2} D\, e^{-x} x D\, e^{x/2} = xD^2 + D - x/4 + \frac{1}{2}$. By iteration, we can write $R^k = \sum_{i=0}^{2k} P_i(x) D^k$, where $P_i(x)$ is a polynomial in x. Since S_I is a linear space it suffices to show that S_I is closed under the operator $x^p D^q$ for $p, q = 0, 1, 2, \cdots$. But this follows immediately from the definition of S_I.

(ii) From the definition of S_I it follows that $\sup_{x \in I} |x\phi(x)| = M < \infty$ for $\phi \in S_I$.

Therefore, we have $|\phi(x)| \leqq M$ for all $x \in (1, \infty)$. Hence

$$\int_0^\infty |\phi(x)|^2\,dx = \int_0^1 |\phi(x)|^2\,dx + \int_1^\infty |\phi(x)|^2\,dx$$

$$\leqq \sup_{x\in(0,1]} |\phi(x)|^2 + \int_1^\infty \frac{M^2}{x^2}\,dx \leqq \|\phi\|_\infty^2 + M^2 < \infty.$$

(iii) We have

$$(R\phi, \psi) = \int_0^\infty \left(x\phi'' + \phi' - \frac{x}{4}\phi + \frac{1}{2}\phi\right)\psi\,dx.$$

Integration by parts together with $\phi(\infty) = 0 = \psi(\infty)$ yields $(R\phi, \psi) = (\phi, R\psi)$.

(iv) If $\phi \in S_I$, then $R^k\phi \in S_I$ from (i) and $\int_0^\infty |R^k\phi|^2\,dx < \infty$ for all k follows from (ii). This together with (iii) gives (iv). Q.E.D.

LEMMA 3.2. *If $\phi \in A$, then we have the following*:

(i) $x^p\phi \in A$, $p = 0, 1, 2, \cdots$.

(ii) $x\dfrac{d\phi}{dx} \in A$; *i.e., A is closed under the operator xD.*

(iii) $x^q\phi^{(p)} \in A$ *for all* $q \geqq p$.

(iv) ϕ *is bounded in* $(0, \infty)$.

Proof.

(i) From the relation

$$[(2n+1) - x]L_n(x) = (n+1)L_{n+1}(x) + nL_{n-1}(x), \tag{3.1}$$

after multiplying by $e^{-x/2}$ and writing $u_n(x) = e^{-x/2}L_n(x)$ we obtain

$$xu_n(x) = (2n+1)u_n(x) - (n+1)u_{n+1}(x) - nu_{n-1}(x). \tag{3.2}$$

Since $\phi \in A$ it follows from Theorem 2.1 parts (i) and (ii) that ϕ has the expansion $\phi(x) = \sum_{n=0}^\infty (\phi, u_n)u_n(x)$, where $\{(\phi, u_n)\}_{n=0}^\infty$ is a rapidly decreasing sequence; i.e., $\sum_{n=0}^\infty |(\phi, u_n)|n^k < \infty$ for all $k = 0, 1, 2, \cdots$. Hence by (3.2) we get

$$x\phi(x) = \sum_0^\infty (\phi, u_n)xu_n(x)$$

$$= \sum_0^\infty (\phi, u_n)[(2n+1)u_n(x) - (n+1)u_{n+1}(x) - nu_{n-1}(x)] \tag{3.3}$$

$$= \sum_{n=0}^\infty b_n u_n(x), \tag{3.4}$$

where $b_n = (2n+1)a_n - na_{n-1} - (n+1)a_{n+1}$, $a_n = (\phi, u_n)$, $n = 0, 1, 2$, $a_{-1} = 0$. Clearly, $\{b_n\}_{n=0}^\infty$ is also a rapidly decreasing sequence and consequently $x\phi \in A$ by Theorem 2.1 part (ii). Repeating the same argument yields $x^p\phi \in A$ for $p = 1, 2, \cdots$.

(ii) From $u_n(x) = e^{-x/2}L_n(x)$ and the relation

$$x\frac{dL_n(x)}{dx} = nL_n(x) - nL_{n-1}(x) \tag{3.5}$$

we obtain

$$\frac{du_n(x)}{dx} = e^{-x/2}\frac{dL_n(x)}{dx} - \frac{1}{2}u_n(x)$$

and hence

$$x\frac{du_n(x)}{dx}=nu_n(x)-nu_{n-1}(x)-\frac{1}{2}xu_n(x). \tag{3.6}$$

From (3.2) and (3.6) we get

$$x\frac{du_n(x)}{dx}=\frac{1}{2}(n+1)u_{n+1}-\frac{1}{2}u_n-\frac{n}{2}u_{n-1}. \tag{3.7}$$

Using this relation, we write

$$x\frac{d\phi}{dx}=\sum_{n=0}^{\infty}a_n(x)\frac{du_n(x)}{dx}=\sum_{n=0}^{\infty}b_nu_n(x), \tag{3.8}$$

where $b_n=\frac{1}{2}[na_{n-1}-a_n-(n+1)a_{n+1}]$. Again, the sequence $\{b_n\}$ is rapidly decreasing since $\{a_n\}$ is, and it follows as in part (i) that $x\,d\phi/dx\in A$. The validity of (3.8) follows from Theorem 2.1 and the remark preceding it.

(iii) First, we show that $x^p\phi^{(p)}\in A$ for all $p=1,2,\cdots$. We use an induction argument on p. The fact that it is true for $p=1$ follows from part (ii).

Assume $x^{p-1}\phi^{(p-1)}\in A$. From (ii) we obtain $x\,(d/dx)(x^{p-1}\phi^{(p-1)})\in A$. On the ther hand,

$$\begin{aligned}x\frac{d}{dx}(x^{p-1}\phi^{(p-1)})&=x(x^{p-1}\phi^{(p)}+(p-1)x^{p-2}\phi^{(p-1)})\\&=x^p\phi^{(p)}+(p-1)x^{p-1}\phi^{(p-1)}\end{aligned}$$

which gives

$$x^p\phi^{(p)}=x\frac{d}{dx}(x^{p-1}\phi^{(p-1)})-(p-1)x^{p-1}\phi^{(p-1)}\in A.$$

For $q\geqq p$, we have $x^q\phi^{(p)}=x^{q-p}x^p\phi^{(p)}\in A$ by part (i).

(iv) From the relation $L_n(x)=\int_1^x L_n'(t)\,dt+L_n(1)$, $x>1$ and (3.5) one gets

$$L_n(x)=n\int_1^x\frac{L_n(t)-L_{n-1}(t)}{t}\,dt+L_n(1).$$

Then

$$|u_n(x)|=|e^{-x/2}L_n(x)|\leqq n\,e^{-x/2}\int_1^x\left(\frac{|L_n(t)|+|L_{n-1}(t)|}{t}\right)dt+e^{-x/2}|L_n(1)|, \tag{3.9}$$

but by formula (2.7) we have $e^{-x/2}|L_n(1)|\leqq A$. Therefore,

$$\begin{aligned}|u_n(x)|&\leqq n\int_1^x e^{-t/2}\left(\frac{|L_n(t)|+|L_{n-1}(t)|}{t}\right)dt+A\\&\leqq n\int_1^\infty e^{-t/2}\left(\frac{|L_n(t)|+|L_{n-1}(t)|}{t}\right)dt+A.\end{aligned}$$

By the Schwarz inequality and the orthonormality of the system $\{e^{-t/2}L_n(t)\}_{n=0}^{\infty}$, we get

$$|u_n(x)|\leqq 2n\left(\int_1^\infty\frac{1}{t^2}\,dt\right)^{1/2}+A,\quad\text{i.e., }|u_n(x)|=O(n) \tag{3.10}$$

uniformly for $x>1$. Once more, (2.7) gives us

$$\sup_{0\leqq x\leqq 1}|u_n(x)|=O(1). \tag{3.11}$$

From Theorem 2.1 we have $\phi(x)=\sum_{n=0}^{\infty}(\phi, u_n)u_n(x)$, where the series converges uniformly on $[\varepsilon, w]$; $0<\varepsilon<w$. On the other hand,

$$|\phi(x)|\leqq\sum_{n=0}^{\infty}|(\phi, u_n)||u_n(x)|, \qquad 0\leqq x<\infty. \tag{3.12}$$

Then (3.10), (3.11), (3.12), and the fact that $\{|(\phi, u_n)|\}_{n=0}^{\infty}$ is a rapidly decreasing sequence gives us the boundedness of $\phi(x)$ in $[0, \infty)$.

LEMMA 3.3. $A\subset S_I$.

Proof. Let $\phi\in A$; we want to show that

$$\|\phi\|_k=\sup_{\substack{x\in I\\ 0\leqq p,q\leqq k}}|x^p\phi^{(q)}(x)|<\infty \quad \forall k=0, 1, 2, \cdots.$$

It follows from Lemma 2 parts (i), (ii), (iii) and (iv) that $|x^p\phi^{(q)}|\leqq C_{pq}$ for all p, q such that $0\leqq q\leqq p\leqq k$. We show that $|x^p\phi^{(q)}|$ is bounded for $0\leqq p<q\leqq k$ by examining this quantity in the two intervals $(0, 1]$, $(1, \infty)$. First, for $x\in(1, \infty)$, we have $|x^p\phi^{(q)}|\leqq|x^q\phi^{(q)}|<\infty$. As for $(0, 1]$, we consider the series expansion of ϕ,

$$\phi(x)=\sum_{n=0}^{\infty}(\phi, u_n)u_n(x).$$

Then, $|\phi^{(q)}(x)|\leqq\sum_{n=0}^{\infty}|(\phi, u_n)||u_n^{(q)}(x)|$.

Now we estimate $|u_n^{(q)}(x)|$ with the aid of $dL_n^{\alpha}(x)/dx=-L_{n-1}^{\alpha+1}(x)$ and $|L_n^{\alpha}(x)|=O(n^{\alpha})$, $\alpha\geqq 0$ uniformly in $[0, 1]$ (see (2.7)). Taking the qth derivative of $u_n(x)=e^{-x/2}L_n(x)$, we obtain by Leibniz's rule

$$u_n^{(q)}(x)=\sum_{j=0}^{q}\binom{q}{j}(e^{-x/2})^{(j)}(L_n(x))^{(q-j)}. \tag{3.13}$$

But

$$\frac{d^{q-j}L_n^{(x)}}{dx^{q-j}}=(-1)^{q-j}L_{n-(q-j)}^{q-j}(x), \qquad n\geqq q-j; \tag{3.14}$$

hence

$$\left|\frac{d^{q-j}}{dx^{q-j}}L_n(x)\right|=O(n^{q-j}) \quad \forall x\in[0, 1]. \tag{3.15}$$

Consequently, by (3.13), (3.14) and (3.15), we get $|u_n^{(q)}(x)|\leqq A(q)O(n^{q+1})$, and this gives

$$|\phi^{(q)}(x)|\leqq A(q)\sum_{n=0}^{\infty}|(\phi, u_n)|O(n^{q+1})<\infty$$

for all $x\in[0, 1]$, since $\{(\phi, u_n)\}_{n=0}^{\infty}$ is a rapidly decreasing sequence.

Finally, for $x\in[0, 1]$ we have $|x^p\phi^{(q)}(x)|\leqq|\phi^{(q)}(x)|<\infty$, $0\leqq p<q$, i.e., $|x^p\phi^{(q)}(x)|<\infty$ for all $p, q=0, 1, 2, \cdots$, and $x\in(0, \infty)$. Q.E.D.

Now we turn to the proof of Theorem 3.1.

Proof of Theorem 3.1. Recall that we want to show that $A=S_I$. By Lemma 3.1 we obtain $S_I\subseteq A$, and by Lemma 3.3 we obtain $A\subseteq S_I$. Therefore, we have $S_I=A$ as

sets. What remains to be shown is that the two topologies are equivalent. Assume that $\{\phi_\nu\}_{\nu=1}^{\infty} \subset S_I$ and $\phi_\nu \to 0$ in S_I as $\nu \to \infty$. We want to show that $\phi_\nu \to 0$ in A, i.e.,

$$\int_0^\infty |R^k \phi_\nu(x)|^2 \, dx \to 0 \quad \text{as } \nu \to \infty.$$

As indicated before R^k is of the form $R^k = \sum_{i=0}^{2k} P_i(x) D^i$, where $P_i(x)$ is a polynomial in x. Hence a typical term of $R^k \phi_\nu$ is of the form $x^p \phi_\nu^{(q)}(x)$. Therefore, it suffices to show that

$$\int_0^\infty |x^p \phi_\nu^{(q)}(x)|^2 \, dx \to 0 \quad \text{as } \nu \to \infty$$

with p, q fixed. Since $\phi_\nu \to 0$ in S_I, then $\phi_\nu^{(q)} \to 0$ uniformly on every compact subset of I and $|x^p \phi_\nu^{(q)}| \leqq C_{pq}$ independent of ν. For $x \geqq 1$, we have $|x^p \phi_\nu^{(q)}| \leqq |x^{p+1} \phi_\nu^{(q)}| \leqq C_{p+1,q}$, i.e., $|x^p \phi_\nu^{(q)}| \leqq C_{p+1,q}/x$.

From the Lebesgue dominated convergence theorem and the fact that $x^p \phi_\nu^{(q)} \to 0$ in I, it follows that

$$\int_a^\infty |x^p \phi_\nu^{(q)}(x)|^2 \, dx \to 0 \quad \text{as } \nu \to \infty \text{ and } a > 1.$$

As for $\int_0^a |x^p \phi_\nu^{(q)}(x)|^2 \, dx$, we have

$$\int_0^a |x^p \phi_\nu^{(q)}(x)|^2 \leqq a^{2p} \int_0^a |\phi_\nu^{(q)}(x)|^2 \, dx \to 0 \quad \text{as } \nu \to \infty,$$

since $\phi_\nu^{(q)} \to 0$ uniformly on compact subsets of I. Therefore, $\phi_\nu \to 0$ in A. Conversely, let $\phi_\nu \to 0$ in A. We need to show that $\phi_\nu^{(q)} \to 0$ uniformly on compact subsets of I for given p, q, $|x^p \phi_\nu^{(q)}(x)| \leqq C_{pq}$ independent of ν. That $\phi_\nu^{(q)} \to 0$ uniformly on compact subsets of I was proved by Zemanian [11]. We show only the second part in the following steps:

(i) If $p = 0 = q$, then for each ν, we have $\phi_\nu(x) = \sum_{n=0}^{\infty} a_{\nu n} u_n(x)$, where

$$a_{\nu n} = \int_0^\infty \phi_\nu(x) u_n(x) \, dx = \int_0^\infty \phi_\nu(x) \frac{R^k u_n^{(x)}}{(-n)^k} \, dx$$

$$= \frac{1}{(-n)^k} \int_0^\infty (R^k \phi_\nu(x)) u_n(x) \, dx, \qquad n \neq 0, \quad k = 1, 2, 3, \cdots.$$

Hence,

$$|a_{\nu n}|^2 \leqq \frac{1}{(n)^{2k}} \left\{ \left(\int_0^\infty |u_n(x)|^2 \, dx \right) \left(\int_0^\infty |R^k \phi_\nu(x)|^2 \, dx \right) \right\},$$

$$|a_{\nu n}|^2 \leqq \frac{1}{(n)^{2k}} \int_0^\infty |R^k \phi_\nu(x)|^2 \, dx \to 0 \quad \text{as} \quad \nu \to \infty.$$

Thus

$$|\phi_\nu(x)| \leqq \sum_{n=0}^{\infty} |a_{\nu n}| |u_n(x)| \leqq \sum_{n=0}^{\infty} \frac{1}{n^k} |u_n(x)| < \infty$$

is independent of ν.

(ii) *Let* $\phi_\nu \to 0$ in A and $\phi_\nu(x) = \sum_{n=0}^{\infty} a_{\nu,n} u_n(x)$. Then, as in the proof of Lemma 3.2, it can be easily shown that $x^p \phi_\nu^{(q)}(x) = \sum_{n=0}^{\infty} b_{\nu,n} u_n(x)$, where $b_{\nu,n}$ is a finite linear

combination of $\{a_{\nu,n-j}\}_{j=m_1}^{m_2}$, where m_1 and m_2 depend on p and q only. Therefore, since $\{a_{\nu,n}\}$ is a rapidly decreasing sequence in n and uniformly bounded in ν, then so is $\{b_{\nu,n}\}$ and this completes the proof.

4. The connection with the Hermite functions. The analogy between the results of the preceding section and Zemanian's results in the case of Hermite functions on the one hand, and the relationship between Hermite and Laguerre functions on the other, suggest connecting our results with Zemanian's. This will be our goal in this section.

We shall extend the generalized Laguerre functions from $(0, \infty)$ to the whole real line $(-\infty, \infty)$ and then use arguments similar to the ones given in § 3 to establish that connection. This approach was suggested by one of the referees.

One distinguishing feature between the results of § 3 and those of this section is the relative simplicity of the latter. This is due to the behavior of the test functions at the endpoints of the interval I in both cases. For example, if $\phi \in A(R, (0, \infty))$, where R is the Laguerre differential operator, then, as we saw in § 3, $\lim_{x\to\infty} \phi(x) = 0$ but $\lim_{x\to 0+} \phi(x)$ is not necessarily zero. On the contrary, if $\phi \in A(R, (-\infty, \infty))$, where R is the Hermite differential operator, then from Zemanian's results we evidently have $\lim_{x\to\infty} \phi(x) = \lim_{x\to-\infty} \phi(x) = 0$.

Now let us define

$$\mathscr{L}_n^\alpha(x) = \frac{L_n^\alpha(x)}{D_n^\alpha} e^{-1/2x}, \qquad \alpha > -1,$$

where

$$D_n^\alpha = \left[\frac{\Gamma(n+\alpha+1)}{n!}\right]^{1/2}.$$

Then we have

$$\int_0^\infty \mathscr{L}_n^\alpha(x)\mathscr{L}_m^\alpha(x)x^\alpha\,dx = \int_{-\infty}^\infty \mathscr{L}_n^\alpha(x^2)\mathscr{L}_m^\alpha(x^2)|x|^{2\alpha+1}\,dx$$

$$= \int_{-\infty}^\infty \mathscr{L}_n^\alpha(x^2)\mathscr{L}_m^\alpha(x^2)\,dm(x) = \delta_{nm},$$

where $dm(x) = |x|^{2\alpha+1}\,dx$. Let $R_\alpha = d^2/dx^2 + ((2\alpha+1)/x)\,d/dx - x^2$; then it is easy to see that $\{\mathscr{L}_n^\alpha(x^2)\}_{n=0}^\infty$ are eigenfunctions of R_α. In fact, we have

$$R_\alpha \mathscr{L}_n^\alpha(x^2) = -(4n+2\alpha+2)\mathscr{L}_n^\alpha(x^2).$$

Also, for an even function ϕ defined on $(-\infty, \infty)$, we have formally

$$\phi(x) = \sum_{n=0}^\infty \hat{\phi}(n)\mathscr{L}_n^\alpha(x^2)$$

if and only if

$$\hat{\phi}(n) = \int_{-\infty}^\infty \phi(x)\mathscr{L}_n^\alpha(x^2)|x|^{2\alpha+1}\,dx.$$

Throughout this section the space A^α will mean the space $A(R_\alpha, (-\infty, \infty))$ provided with the topology described in § 2. The analogues of the space S_I of § 3 will be the space S_{even} defined below.

First, let us introduce the space S as follows. S consists of all C^∞-functions in $(-\infty, \infty)$ that are rapidly decreasing; i.e. for $\phi(x) \in S$, we have $\sup_{x\in(-\infty,\infty)} |x^p \phi^{(q)}| < \infty$ for all nonnegative integers p and q.

The topology of S is the topology induced by the countable system of norms

$$\|\phi\|_k = \sup_{\substack{x\in(-\infty,\infty)\\ 0\leqq p,q\leqq k}} |x^p \phi^{(q)}|, \qquad k = 0, 1, 2, 3, \cdots.$$

As is known, in this topology convergence of the sequence $\{\phi_\nu\}$ to ϕ is equivalent to the following two conditions:

(i) For any compact subset K of the real line and any nonnegative integer p we have $\phi_\nu^{(p)} \to \phi^{(p)}$ uniformly on K.

(ii) For any nonnegative integers p and q, there exists a constant C_{pq} such that $|x^p \phi_\nu^{(q)}| \leqq C_{pq}$ independent of ν.

Now we define the space S_{even} as

$$S_{\text{even}} = \{\phi \mid \phi \in S \text{ and } \phi \text{ is even}\}.$$

LEMMA 4.1. $S_{\text{even}} \subseteq A^\alpha$.

Proof. First let us observe that $S_{\text{even}} \subseteq L^2(dm(x), (-\infty, \infty))$. Moreover, it is not hard to see that, if ϕ and ψ are in S_{even}, then so are $R_\alpha \phi$ and $R_\alpha \psi$ and hence

$$\|\phi\|_k = \int_{-\infty}^{\infty} |R_\alpha^k \phi|^2 \, dm < \infty.$$

Finally, straightforward calculations show that

$$\int_{-\infty}^{\infty} [(R_\alpha \phi)\psi - \phi(R_\alpha \psi)] \, dm = 0,$$

i.e., R_α is self-adjoint, and this establishes our claim.

Now let us define the space A^α_{even} as the space of all even functions in A^α. The space S_{odd} and the space A^α_{odd} are defined similarly.

LEMMA 4.2. *The assertion that $A^\alpha \subseteq S$ is equivalent to the following three conditions. If $\phi \in A^\alpha$, then*

(a) *ϕ is bounded on $(-\infty, \infty)$;*

(b) *$x^2 \phi \in A^\alpha$;*

(c) *$x^p \phi^{(p)} \in A^\alpha$ for all $p = 0, 1, 2, \cdots$.*

Proof. Recall that $A^\alpha \subseteq S$ is equivalent to $\sup_{x\in(-\infty,\infty)} |x^p \phi^{(q)}| < \infty$ for all nonnegative integers p, q and all $\phi \in A^\alpha$. In fact, it suffices to show that $\sup_{x\in J} |x^p \phi^{(q)}| < \infty$, where $J = (-\infty, -1) \cup (1, \infty)$ since

$$\sup_{x\in[-1,1]} |x^p \phi^{(q)}| \leqq \sup_{x\in[-1,1]} |\phi^{(q)}| < \infty.$$

If $p = q$, then from (c) and (a) it follows that $\sup_{x\in(-\infty,\infty)} |x^p \phi^{(q)}| < \infty$. If $p < q$, then we have $\sup_{x\in J} |x^p \phi^{(q)}| \leqq \sup_{x\in J} |x^q \phi^{(q)}| < \infty$ by (a) and (c). Finally, if $p > q$, we put $r = p - q$ and choose a positive integer k such that $r < 2k$. Thus, $\sup_{x\in J} |x^p \phi^{(q)}| = \sup_{x\in J} |x^r x^q \phi^{(q)}| \leqq \sup_{x\in J} |x^{2k} x^q \phi^{(q)}| < \infty$. The last inequality follows from (c), (a) and repeated application of (b).

LEMMA 4.3. $A^\alpha \subset S$.

Proof. We prove this assertion by proving the three conditions of the previous lemma.

(a) Consider the series expansion of $\phi(x)$ in terms of $\mathscr{L}_n^\alpha(x^2)$,

$$\phi(x)=\sum_{n=0}^{\infty} a_n\mathscr{L}_n^\alpha(x^2), \quad \text{where } \{a_n\} \text{ is rapidly decreasing.}$$

Hence

$$|\phi(x)|\leqq\sum_{n=0}^{\infty}|a_n||\mathscr{L}_n^\alpha(x^2)|<\infty.$$

The validity of the last inequality follows from the fact that $|\mathscr{L}_n^\alpha(x^2)|\leqq 1$ if $\alpha\geqq 0$ and $|\mathscr{L}_n^\alpha(x^2)|$ is polynomially bounded in n if $-1<\alpha<0$.

(b) From the recurrence relation

$$x^2D_n^\alpha\mathscr{L}_n^\alpha(x^2)=(2n+\alpha+1)D_n^\alpha\mathscr{L}_n^\alpha(x^2)-(n+1)D_{n+1}^\alpha\mathscr{L}_{n+1}^\alpha(x^2)-(n+\alpha)D_{n-1}^\alpha\mathscr{L}_{n-1}^\alpha(x^2)$$

we deduce that

$$x^2\phi(x)=\sum_{n=0}^{\infty}a_nx^2\mathscr{L}_n^\alpha(x^2)=\sum_{n=0}^{\infty}b_n\mathscr{L}_n^\alpha(x^2),$$

where

$$b_n=(2n+\alpha+1)a_n-n\frac{D_n^\alpha}{D_{n-1}^\alpha}a_{n-1}-(n+\alpha+1)\frac{D_n^\alpha}{D_{n+1}^\alpha}a_{n+1}, \qquad a_{-1}=0.$$

Clearly, $\{b_n\}$ is a rapidly decreasing sequence if and only if $\{a_n\}$ *is*. Thus $x^2\phi\in A^\alpha$.

(c) We show only that $x\phi^{(1)}\in A^\alpha$, since the case $x^p\phi^{(p)}$ for $p>1$ is proved by an induction argument as in the proof of Lemma 3.2. The relation

$$x\frac{d}{dx}\mathscr{L}_n^\alpha(x^2)=2n\mathscr{L}_n^\alpha(x^2)-2(n+\alpha)\frac{D_{n-1}^\alpha}{D_n^\alpha}\mathscr{L}_{n-1}^\alpha(x^2)-x^2\mathscr{L}_n^\alpha(x^2)$$

yields

$$x\phi^{(1)}=\sum_{n=0}^{\infty}a_nx\frac{d}{dx}\mathscr{L}_n^\alpha(x^2)=\sum_{n=0}^{\infty}b_n\mathscr{L}_n^\alpha(x^2)-x^2\phi,$$

where $b_n=2na_n-2(n+1+\alpha)(D_n^\alpha/D_{n+1}^\alpha)a_{n+1}$. Since the sequence $\{b_n\}$ is rapidly decreasing and $x^2\phi\in A^\alpha$ by part (b), it follows immediately that $x\phi^{(1)}\in A^\alpha$.

THEOREM 4.1. $A_{\text{even}}^\alpha=S_{\text{even}}$.

Proof. Lemma 4.1 gives the inclusion $S_{\text{even}}\subseteq A^\alpha$ which in turn gives $S_{\text{even}}\subseteq A_{\text{even}}^\alpha$. Similarly, Lemma 4.3 shows that $A^\alpha\subseteq S$, which implies that $A_{\text{even}}^\alpha\subseteq S_{\text{even}}$, hence $A_{\text{even}}^\alpha=S_{\text{even}}$ as sets. To prove the equivalence of the two topologies we use arguments similar to those used in the proof of Theorem 3.1.

For the fact that the convergence in S_{even} implies convergence in A_{even}^α, see the proof of Theorem 3.1. Conversely, let $\{\phi_\nu\}$ converge to zero in A_{even}^α. Then $\{\phi_\nu^{(p)}\}$ converges to zero uniformly on compact subsets of the real line for any nonnegative integer p (see Theorem 3.1).

To show that $\sup_{x\in(-\infty,\infty)}|x^p\phi_\nu^{(q)}|\leqq C_{pq}$ independent of ν, we write

$$\phi_\nu(x)=\sum_{n=0}^{\infty}a_{\nu,n}\mathscr{L}_n^\alpha(x^2);$$

hence

$$x^p\phi_\nu^{(q)}(x)=\sum_{n=0}^{\infty}a_{\nu,n}^{p,q}\mathscr{L}_n^\alpha(x^2),$$

where $a_{\nu,n}^{p,q}$ is a finite linear combination of $a_{\nu,n}$.

Since the operator R_α is self-adjoint, we can write

$$a_{\nu,n}^{p,q} = (x^p \phi_\nu^{(q)}, \mathscr{L}_n^\alpha(x^2)) = \frac{(-1)^k}{(4n+2\alpha+2)^k}(x^p \phi_\nu^{(q)}, R_\alpha^k \mathscr{L}_n^\alpha(x^2))$$

$$= \frac{(-1)^k}{(4n+2\alpha+2)^k}(R_\alpha^k x^p \phi_\nu^{(q)}, \mathscr{L}_n^\alpha(x^2))$$

$$= \frac{(-1)^k}{(4n+2a+2)^k}\int_{-\infty}^{\infty} \mathscr{L}_n^\alpha(x^2) R_\alpha^k(x^p \phi_\nu^{(q)})\, dm$$

for any positive integer k.

Thus by the Schwarz inequality one easily gets

$$|a_{\nu,n}^{p,q}|^2 \leqq \frac{1}{(4n+2\alpha+2)^{2k}}\int_{-\infty}^{\infty} |R_\alpha^k x^p \phi_\nu^{(q)}|^2\, dm \to 0 \quad \text{as } \nu \to \infty,$$

since $x^p \phi_\nu^{(q)}$ converges to zero in A^α. Therefore

$$\sup_{x\in(-\infty,\infty)} |x^p \phi_\nu^{(q)}| \leqq \sum_{n=0}^{\infty} |a_{\nu,n}^{p,q}||\mathscr{L}_n^\alpha(x^2)| \leqq \sum_{n=0}^{\infty} \frac{|\mathscr{L}_n^\alpha(x^2)|}{(4n+2\alpha+2)^k} < \infty.$$

The last series is independent of ν, and this finishes the proof. Q.E.D.

In essence, what Theorem 4.1 says is that any function $\phi \in S_{\text{even}}$ can be expanded in terms of $\{\mathscr{L}_n^\alpha(x^2)\}$ for any $\alpha > -1$, and that the coefficients of the expansion form a rapidly decreasing sequence of complex numbers. Now it is easy to see that if $\psi \in S_{\text{odd}}$ it can be expanded in terms of $\{x\mathscr{L}_n^\alpha(x^2)\}$ for $\alpha > -1$ with rapidly decreasing coefficients.

In particular, if $\alpha = -\frac{1}{2}$, we have for $\phi \in S_{\text{even}}$

$$\phi(x) = \sum a_n \mathscr{L}_n^{-1/2}(x^2),$$

while if $\alpha = \frac{1}{2}$, we have for $\psi \in S_{\text{odd}}$

$$\psi(x) = \sum b_n x \mathscr{L}_n^{1/2}(x^2),$$

where $\{a_n\}$ and $\{b_n\}$ are rapidly decreasing sequences.

Now Zemanian's result can easily be concluded from the fact that any function ϕ in S can be written as the sum of two functions, one even and the other odd, as well as the fact that

$$h_{2n}(x) = \frac{(-1)^n 2^n n!}{(2n!)^{1/2}\pi^{1/4}} D_n^{-1/2}\mathscr{L}_n^{-1/2}(x^2) = A_n \mathscr{L}_n^{-1/2}(x^2),$$

$$h_{2n+1}(x) = \frac{(-1)^n 2^{(2n+1)/2} n!}{\{(2n+1)!\}^{1/2}\pi^{1/4}} D_n^{1/2}\mathscr{L}_n^{1/2}(x^2) = B_n \mathscr{L}_n^{1/2}(x^2),$$

where $h_n(x)$ is the Hermite function and A_n and B_n are polynomially bounded in n.

5. Continuous functions of S_I^* as boundary values. In this section, the use of Fourier series of a function f defined on the unit circle to solve he Dirichlet problem for the unit disk will be emulated for a function f defined on $(0, \infty)$ to solve the following problem. Let Q be the upper right half-plane, i.e., $Q = \{(x, y) \mid x > 0, y > 0\}$, and let $f(x) \in S_I^*$. We wish to find a function $f(x, y)$ with the following properties:

(i) $f(x, y)$ is harmonic in Q.

(ii) $\lim_{x\to 0+} f(x, y)$ is finite for $0 < y < \infty$.

(iii) $\lim_{y+0+} f(x, y) = f(x)$ for $0 < x < \infty$.

Our solution may not be the most direct one, but it completes both our work and the work of Muckenhoupt on developing the analogy between the theory of Fourier series and that of Laguerre series. First we solve this problem when $f(x)$ is a continuous function; this can be easily extended to the case where $f(x)$ is any element of S_I^* with compact support. However, the case where $f(x)$ is an arbitrary element of S_I^* requires a new approach since there are two notions of convergence involved, pointwise convergence and convergence in the sense of S_I^*, neither of which in general implies the other. This case will be treated in the next section.

We have seen that each element of S_I^* has a Laguerre series expansion which converges to it in the sense of S_I^*, and the coefficients of the expansion form a slowly increasing sequence of real numbers. Unfortunately, we cannot in general continue this expansion to the complex plane since a series of Laguerre functions which may converge (even uniformly) on the positive real axis can possibly diverge off that part of the axis. To overcome this difficulty we introduce Laguerre functions of the second kind $\{V_n(z)\}$. These are solutions of the Laguerre differential equation $zV_n'' + V_n' + (n+\frac{1}{2}-z/4)V_n = 0$ which vanish at $-\infty$. The definition of the $V_n(z)$'s and some of their properties are given in the following lemma.

LEMMA 5.1. *Let*

$$V_n(z) = e^{z/2}\int_0^\infty \frac{e^{-t/2}u_n(t)}{t-z}\,dt, \quad n = 0, 1, 2, \cdots.$$

Then we have:

(i) $V_n(z)$ *is well defined and holomorphic in the complex plane cut along* $[0, \infty)$.

(ii) $V_n(z)$ *satisfies the same differential equation as* $u_n(z)$, *i.e.*, (2.4).

(iii) $\lim_{y\to 0}[V_n(x+iy) - V_n(x-iy)] = 2\pi i u_n(x)$; $x \in [0, \infty)$.

Proof. First, observe that the definition makes sense since by the Schwarz inequality we have

$$|V_n(z)|^2 \leqq \left(\int_0^\infty \frac{dt}{|t-z|^2}\right)\left(\int_0^\infty e^{-t}u_n^2(t)\,dt\right)e^x < \infty.$$

The proof of (i) and (ii) is straightforward and left to the reader.

(iii)

$$\begin{aligned} V_n(x+iy) - V_n(x-iy) &= e^{x/2}\left\{e^{iy/2}\int_0^\infty \frac{e^{-t/2}u_n(t)}{(t-z)}dt - e^{iy/2}\int_0^\infty \frac{e^{-t/2}u_n(t)}{(t-\bar{z})}\,dt\right\} \\ &= e^{x/2}\left\{\int_0^\infty \frac{e^{-t/2}u_n(t)}{(t-x)^2+y^2}\left\{\cos\frac{y}{2}(2iy) + 2i(t-x)\sin\frac{y}{2}\right\}dt\right\} \\ &= 2\pi i\, e^{x/2}\left\{\int_0^\infty \cos\frac{y}{2}\,e^{-t/2}u_n(t)P(t, z)\,dt\right. \\ &\qquad \left. + \frac{\sin(y/2)}{y}\int_0^\infty e^{-t/2}u_n(t)(t-x)P(t, z)\,dt\right\}, \end{aligned} \tag{5.1}$$

where $P(t, z)$ is the Poisson kernel for the upper half-plane. Since the limit of the first integral is $u_n(x)\exp(-x/2)$ and that of the second is zero, then taking the limit of (5.1) as $y \to 0$ yields the result.

DEFINITION. Since $V_n(z)$ satisfies the same differential equation as $u_n(z)$ and clearly $u_n(z)$ and $V_n(z)$ are linearly independent, we call $V_n(z)$ a Laguerre function of the second kind.

THEOREM 5.1. *Let $f \in S_I^*$ be a continuous function. Then*

(i) *f has Laguerre series expansion $f = \sum_{n=0}^{\infty} c_n u_n$, where the convergence is in the sense of S_I^*.*

(ii) *$\sum_{n=0}^{\infty} (ic_n/2\pi) V_n(z)$ converges pointwise in the plane cut along $[0, \infty)$.*

(iii) *$\sum_{n=0}^{\infty} (ic_n/2\pi) V_n(z)$ converges uniformly in compact subsets of the cut plane.*

(iv) *$\sum_{n=0}^{\infty} (ic_n/2\pi)[V_n(z) - V_n(\bar{z})]$ converges in Q to a real harmonic function $f(x, y)$.*

(v) *$\lim_{y \to 0} f(x, y) = f(x)$ pointwise; $x \in (0, \infty)$. Moreover, the convergence is uniform on compact subsets of $(0, \infty)$.*

Proof.

(i) The coefficients c_n given by $c_n = (f, u_n)$ are well defined, since $u_n \in S_I$ for all $n = 0, 1, 2, \cdots$. To show the convergence, we invoke Theorem 2.1 and write for any $\phi \in S_I$

$$(f, \phi) = \Big(f, \sum_{n=0}^{\infty} (\phi, u_n) u_n\Big) = \sum_{n=0}^{\infty} (f, u_n)(\phi, u_n) = \sum_{n=0}^{\infty} c_n (\phi, u_n) < \infty.$$

The last series converges since by Theorem 2.1 $\{c_n\}_{n=0}^{\infty}$ is a slowly increasing sequence, i.e., $c_n = O(n^k)$ for some k, and $\{(\phi, u_n)\}_{n=0}^{\infty}$ is rapidly decreasing.

(ii)

$$\sum_{n=0}^{\infty} \left(\frac{ic_n}{2\pi}\right) V_n(z) = \sum_{0}^{\infty} \left(\frac{ic_n}{2\pi}\right) e^{z/2} \int_0^{\infty} \frac{e^{-t/2} u_n(t)}{t - z}\, dt = \sum_{0}^{\infty} \left(\frac{ic_n}{2\pi}\right) (\phi_z(t), u_n(t)), \tag{5.2}$$

where $\phi_z(t) = \exp((z-t)/2)/(t-z) \in S_I$ for $z \notin [0, \infty)$. But the series on the right-hand side of (5.2) converges for the same reason as in (i).

(iii) Using the fact that $Ru_n(x) = -nu_n(x)$, hence $R^{k+2} u_n = (-n)^{k+2} u_n$, we write (except for a trivial modification when $n = 0$)

$$\sum_{n=1}^{\infty} \left(\frac{ic_n}{2\pi}\right) V_n(z) = \sum_{n=1}^{\infty} \left(\frac{ic_n}{2\pi}\right) (\phi_z, u_n) = \sum_{n=1}^{\infty} \left(\frac{ic_n}{2\pi}\right) \left(\phi_z, \frac{R^{k+2} u_n}{(-n)^{k+2}}\right)$$

$$= \sum_{n=1}^{\infty} \left(\frac{ic_n}{2\pi}\right) \frac{1}{(-n)^{k+2}} (R^{k+2} \phi_z, u_n).$$

The last equality holds since R is self-adjoint. Clearly, the series $\sum_{n=1}^{\infty} (ic_n/2\pi) 1/(-n)^{k+2}$ converges absolutely for sufficiently large k. We will show that $(R^{k+2} \phi_z, u_n)$ is uniformly bounded on compact subsets of the cut plane. From these two facts the uniform convergence follows.

To show that $(R^{k+2} \phi_z, u_n)$ is uniformly bounded, we recall that $R^{k+2} = \sum_{i=0}^{2k+4} P_i(t) D^i$, where $D^i = d^i/dt^i$ and $P_i(t)$ is a polynomial in t. Therefore, $R^{k+2} \phi_z(t) = P(t, z) \exp((z-t)/2)/(t-z)^{2k+4}$, where $P(t, z)$ is a polynomial in t and z.

Let K be any compact subset of the cut plane, $\delta = \operatorname{dist}(K, [0, \infty))$ and $M = \sup_{t \in [0,\infty), z \in K} (P(t, z)\, e^{z/2}\, e^{-t/4})$. Then we have

$$|(R^{k+2} \phi_z(t), u_n(t))| = \left| \int_0^{\infty} \frac{P(t, z)\, e^{z/2}\, e^{-t/2} u_n(t)}{(t - z)^{2k+4}}\, dt \right|$$

$$\leqq \frac{M}{\delta^{2k+4}} \int_0^{\infty} e^{-t/4} |u_n(t)|\, dt$$

$$\leqq \frac{M}{\delta^{2k+4}} \left\{ \left(\int_0^{\infty} e^{-t/2}\, dt \right)^{1/2} \left(\int_0^{\infty} u_n^2(t)\, dt \right)^{1/2} \right\} = \frac{\sqrt{2}\, M}{\delta^{2k+4}} < \infty.$$

(iv) This follows from part (iii) and part (i) of Lemma 5.1. We need only to show that $f(x, y)$ is real, but this follows from the relation

$$V(z)-V(\bar{z})=\int_0^\infty e^{-t/2}u_n(t)\left(\frac{e^{z/2}}{t-z}-\frac{e^{\bar{z}/2}}{t-\bar{z}}\right)dt$$

and the observation that the quantity in brackets is purely imaginary.

(v) This follows from part (iii) of Lemma 5.1 and the fact that the series $\sum_{n=0}^\infty c_n u_n(x)$ is Abel summable to $f(x)$. Q.E.D.

COROLLARY. *If, in addition to the hypothesis of Theorem* 5.1, $\lim_{x\to 0^+} f(x)=c<\infty$, *then* $f(x, y)$ *solves the following boundary value problem*:

(i) $f(x, y)$ *is harmonic in* Q.

(ii) $\lim_{y\to 0^+} f(x, y)=f(x)$ *for* $0\leqq x<\infty$, *where* $f(0)$ *means* $\lim_{x\to 0^+} f(x)$.

(iii) $\lim_{x\to 0^+} f(x, y)$ *is bounded for* $0\leqq y<\infty$.

Proof. We need only to prove (iii). From (5.1) it is easy to see that

$$g(y)=\lim_{x\to 0^+} f(x, y)=\sum_{n=0}^\infty \frac{ic_n}{2\pi}\int_0^\infty \frac{e^{-t/2}u_n(t)}{t^2+y^2}\left[2iy\cos\frac{y}{2}+2it\sin\frac{y}{2}\right]dt$$

is a continuous function of y which is bounded for $0<\varepsilon\leqq y<\infty$ and $\lim_{y\to 0^+} g(y)=f(0)$.

6. The space S_I^* as boundary values. In the previous section we considered the case where the boundary values were given by continuous functions in S_I^*. In fact, $f(x)$ does not have to be continuous for the proof to hold. Theorem 5.1 can be easily proved for functions that are continuous almost everywhere. With no difficulty one can also show that the results hold for any generalized function with compact support in $(0, \infty)$. However, rather than extending the results to a larger class of functions we shall consider the possibility of extending them to all of S_I^*.

Needless to say, imitating the theory of trigonometric series naturally leads one to look at the boundary behavior of the conjugate function $\tilde{f}(x, y)$ of $f(x, y)$.

As for extending the results to all of S_I^*, one difficulty immediately arises; that is, even uniform convergence on compact subsets of $(0, \infty)$ does not, in general, imply convergence in the sense of S_I^*.

As for the boundary behavior of the conjugate function, indeed, we have that the mapping $f\to\tilde{f}$ is continuous from L^P to $L^P(1<P<\infty)$ and it is of weak type (1, 1) (cf. [13, p. 239]). In addition, it is known that if $f(\theta)\in L'(T)$,where T is the unit circle, and if we associate with $f(\theta)$ its Fourier series $f(\theta)\to\sum_{n=-\infty}^\infty a_n e^{in\theta}$, its Poisson integral $f(r, \theta)=\sum_{n=-\infty}^\infty a_n r^{|n|}e^{in\theta}$ and its harmonic conjugate $\tilde{f}(r, \theta)$, then $\lim_{r\to 1}\tilde{f}(r, \theta)=\tilde{f}(\theta)$ is not in general in $L^1(T)$.

Analogously, if we take the conjugate function $\tilde{f}(x, y)$ of $f(x, y)$ of Theorem 4.1, one will easily find that $\lim_{y\to 0}\tilde{f}(x, y)$ is not necessarily in S_I^*.

To overcome this difficulty, it seems inevitable that one has to alter the definition of harmonic functions somewhat. More accurately, a solution will be found in the space of generalized harmonic functions as introduced by Vekua [7].

We begin by borrowing and modifying Muckenhoupt's conjugate functions [5] to introduce $f(x, y)$ and $\tilde{f}(x, y)$ as follows:

$$f(x, y)=\sum_{n=0}^\infty c_n e^{-\sqrt{n}y}u_n(x) \tag{6.1}$$

$$\tilde{f}(x, y)=-\sum_{n=1}^\infty c_n e^{-\sqrt{n}y}u_{n-1}^1(x) \tag{6.2}$$

$$x>0,\quad y>0,\quad c_n=O(n^P).$$

Some properties of $f(x, y)$ and $\tilde{f}(x, y)$ are given in the following lemma.

LEMMA 6.1. *Let $f(x, y)$ and $\tilde{f}(x, y)$ be given by* (6.1) *and* (6.2). *Then*:

(i) *$f(x, y)$ and $\tilde{f}(x, y)$ are well defined in Q, in fact, they are C^∞ functions there.*

(ii) *They satisfy the elliptic partial differential equations*

$$f_{yy} + xf_{xx} + f_x + \left(\frac{1}{2} - \frac{x}{4}\right) f = 0, \tag{6.3}$$

$$\tilde{f}_{yy} + x\tilde{f}_{xx} + \tilde{f}_x - \left(\frac{x}{4} + \frac{1}{4x}\right)\tilde{f} = 0. \tag{6.4}$$

(iii) *They are related by the analogues of the Cauchy–Riemann equations*:

$$\frac{\partial f}{\partial y} = e^{x/2} \frac{\partial}{\partial x}[e^{-x/2} x^{1/2} \tilde{f}] \tag{6.5}$$

and

$$\frac{\partial \tilde{f}}{\partial y} = -e^{-x/2} x^{1/2} \frac{\partial}{\partial x}[e^{x/2} f]. \tag{6.6}$$

Proof. We prove the first part of (i) by showing that the defining series converge uniformly on every compact subset of Q.

Let K be any compact subset of Q and $\delta = \operatorname{dist}\{K, [0, \infty)\}$. Since $|u_n(x)| \leqq 1$ for all $x \in [0, \infty)$ and $n = 0, 1, 2, \cdots$, then we have for any $(x, y) \in K$

$$|f(x, y)| \leqq \sum_{n=0}^{\infty} |c_n| \, e^{-\sqrt{n}\delta} < \infty.$$

A similar argument applies to $\tilde{f}(x, y)$. To prove the second part of (i) we use (3.15) to obtain $|d^p u_n(x)/dx^p| = O(n^{p+1})$, $p = 0, 1, 2, \cdots$, and since $|d^q e^{-\sqrt{n}y}/dy^q| = O(n^{q/2})$, $q = 0, 1, 2, \cdots$ it follows by the same argument as before that all the partial derivatives of $f(x, y)$ and $\tilde{f}(x, y)$ exist.

(ii) and (iii) With the aid of (2.4) and the relations

$$x \frac{du^1_{n-1}(x)}{dx} + \frac{1}{2}(1 - x) u^1_{n-1}(x) = \sqrt{n}\sqrt{x}\, u_n(x),$$

$$\frac{du_n(x)}{dx} + \frac{1}{2} u_n(x) = -\sqrt{n}\, x^{-1/2} u^1_{n-1}(x)$$

the reader should be able to finish the proof.

Now we are in a position to prove the following theorem.

THEOREM 6.1. *Let $f \in S_1^*$ be given by $f = \sum_{n=0}^{\infty} c_n u_n$. Then the function $f(x, y)$ given by* (6.1) *converges to f in the sense of S_I^* as $y \to 0^+$. The function $\tilde{f}(x, y)$ given by* (6.2) *converges in the sense of S_I^* to a generalized function $\tilde{f} \in S_I^*$ (the conjugate generalized function of f) as $y \to 0^+$, where $\tilde{f}$ is defined by the series $\sum_{n=1}^{\infty} c_n u^1_{n-1}(x)$.*

Proof. We want to show that $(f(\cdot, y), \phi) \to (f, \phi)$ as $y \to 0^+$. First, observe that for any fixed $y > 0$, $f(x, y)$ is in S_I^*, since for every $\phi \in S_I$ we have

$$(f(x, y), \phi) = \sum_{n=0}^{\infty} c_n \, e^{-\sqrt{n}y} (u_n, \phi) < \infty. \tag{6.7}$$

In fact, the series in (6.7) converges uniformly for $y \geqq 0$ since $|(u_n, \phi)| = O(n^{-q})$ for every positive integer q. Hence

$$|(f(x, y), \phi) - (f, \Phi)| = \left| \sum_{n=0}^{\infty} c_n (u_n, \phi)(1 - e^{-\sqrt{n}y}) \right|$$

$$\leqq \sum_{n=0}^{\infty} |c_n| |(u_n, \phi)| < \infty.$$

Therefore, we can take the limit as $y \to 0^+$ under the summation sign to get $|(f(x, y), \phi) - (f, \phi)| \to 0$ as $y \to 0^+$.

The proof for $\tilde{f}(x, y)$ and $\tilde{f}$ is exactly the same. Q.E.D.

Comments. The referee has kindly pointed out that the results of Lemma 5.1 and Theorem 5.1 rely on the fact that the second solution to the Laguerre differential equation is also a second solution to the three-term recurrence relation, and has indicated that the results can be extended to the case of the Jacobi polynomials.

Indeed, this is the case. In fact, the results have recently been extended by G. Walter and the author and G. Walter and P. Nevai [10] to a large class of orthogonal polynomials with relatively mild conditions on the weight function.

Acknowledgment. The author wishes to express his sincere gratitude to Professor Gilbert Walter for his helpful suggestions. This work is a part of a Ph.D. dissertation written under his supervision.

REFERENCES

[1] I. M. GEL'FAND AND G. E. SHILOV, *Generalized Functions*, vol. II, Academic Press, New York, 1964.

[2] J. KOREVAAR, *Distributions defined by fundamental sequences*, I, II, III, IV, V, Nederl Acad. Wetensch. Proc. Ser. A58, 1955.

[3] B. MUCKENHOUPT, *Poisson integral for Hermite and Laguerre expansions*, Trans. Amer. Math. Soc., 139 (1969), pp. 231–242.

[4] ———, *Hermite conjugate expansion*, Trans. Amer. Math. Soc., 139 (1969), pp. 243–260.

[5] ———, *Conjugate functions for Laguerre expansions*, Trans. Amer. Math. Soc., 147 (1970), pp. 403–418.

[6] G. SZEGÖ, *Orthogonal Polynomials*, AMS Colloquium Publications, 23, American Mathematical Society, Providence, RI, 1967.

[7] I. M. VEKUA, *Generalized Analytic Functions*, Fizmatgiz, Moscow, 1959; English transl., MR 21 #7288; 27 #321, Pergamon Press, London–Addison-Wesley, Reading, MA, 1962.

[8] G. WALTER, *Expansion of distributions*, Trans. Amer. Math. Soc., 116 (1965), pp. 492–510.

[9] ———, *Hermite series as boundary values*, Trans. Amer. Math. Soc., 147 (1970), pp. 403–418.

[10] G. WALTER AND P. NEVAI, *Series of orthogonal polynomials as boundary values*, this Journal, 12 (1981), pp. 502–513.

[11] A. H. ZEMANIAN, *Orthonormal series expansions of certain distributions and distributional transform calculus*, J. Math. Anal. Appl., 14 (1966), pp. 263–275.

[12] ———, Tech. Rept. No. 22, College of Engineering, State University at New York at Stony Brook, Nov. 1964.

[13] E. M. STEIN AND G. WEISS, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0008 $01.00/0

CHARACTERIZATION OF CONTINUOUS SELECTIONS OF THE METRIC PROJECTION FOR A CLASS OF WEAK CHEBYSHEV SPACES*

MANFRED SOMMER†

Abstract. In this paper we study the problem of existence of continuous selections of the metric projection for finite dimensional subspaces of $C[a, b]$. The problem of characterizing the spaces which admit continuous selections from among the finite dimensional subspaces of $C(X)$, X compact, has been posed by Lazar–Morris–Wulbert. In the case that $X = [a, b]$ partial solutions have been given by recent results of Nürnberger and Sommer. In this paper we complete the studies on continuous selections for finite dimensional subspaces of $C[a, b]$ by characterizing also those spaces which have not yet been considered in the above mentioned papers. Thus the problem of Lazar–Morris–Wulbert is solved in the case that $X = [a, b]$.

Introduction. Let X be a compact metric space, and let $C(X)$ be the space of all real-valued, continuous functions f on X under the uniform norm $\|f\| := \sup\{|f(x)| : x \in X\}$. If G is an n-dimensional subspace of $C(X)$, then for each $f \in C(X)$ we define $P_G(f) := \{g_0 \in G : \|f - g_0\| = \inf\{\|f - g\| : g \in G\}\}$ which is called the *set* of *best approximations* of f from G. This defines a set-valued mapping P_G from $C(X)$ into 2^G which is called the *metric projection* onto G. A continuous mapping s from $C(X)$ onto G is called *continuous selection* for P_G (or, more briefly, continuous selection) if $s(f) \in P_G(f)$ for each $f \in C(X)$.

Lazar–Morris–Wulbert [6] were the first to study the problem of existence of continuous selections. They have characterized those one-dimensional subspaces of $C(X)$ which admit a continuous selection and have posed the problem of characterizing the higher dimensional subspaces of $C(X)$. This question which is relevant to the convergence of algorithms for computing best approximations has also been posed in the book of Holmes [3].

Nürnberger–Sommer [8] have studied this problem with new methods, by using the theory of weak Chebyshev spaces. They have been able to establish the existence of continuous selections for all elements of a special class of weak Chebyshev subspaces of $C[a, b]$.

In further papers Nürnberger and Sommer have extended their studies to other classes of weak Chebyshev subspaces of $C[a, b]$ [9], [10], [11] and also to the class of those subspaces which fail to be weak Chebyshev [7]; e.g., they have shown in [9] that there exists a continuous selection for $P_{S_{m,k}}$, where $S_{m,k}$ is a space of spline functions of degree m with k fixed knots, if and only if $k \leqq m + 1$. Their results solve the problem of Lazar–Morris–Wulbert for all finite dimensional subspaces G of $C[a, b]$, except when G belongs to a special class of weak Chebyshev spaces which we will want to denote by $Z_n^2 \cup Z_n^3$ in the following (for definitions see § 1).

In this paper we complete the studies on existence of continuous slections for finite dimensional subspaces of $C[a, b]$ by characterizing also those elements of $Z_n^2 \cup Z_n^3$ which admit a continuous selection (Theorem 2.7 and Theorem 3.14). To do this we first define a subclass $\tilde{Z}_n^2$ of Z_n^2 and show the nonexistence of a continuous selection for all elements of $(Z_n^2 \setminus \tilde{Z}_n^2) \cup Z_n^3$ by using a fundamental lemma of Lazar–Morris–Wulbert. In contrast to this, for each $G \in \tilde{Z}_n^2$ we are able to construct a

* Received by the editors April 18, 1979, and in revised form March 18, 1981.

† Institut fur Angewandte Mathematik der Universität Erlangen-Nürnberg Martensstrasse 3, 8520 Erlangen, Germany.

continuous selection by modifying the methods of the construction established in [9] and in [11], for a class of spline spaces and for a class of generalized spline spaces, respectively. In both papers the construction essentially depends on the important property that for each $f \in C[a, b]$ all best approximations coincide on a subinterval of $[a, b]$. However, if $G \in \check{Z}_n^2$ this statement does not hold in general. Therefore, the techniques in [11] do not apply directly and the situation here is much more complicated than in that paper.

The construction of the selection is highly local and based on alternation properties and certain uniqueness conditions.

In addition to these results we give a complete characterization of those weak Chebyshev spaces which admit a continuous selection (Theorem 1.8). In this theorem we summarize all results established by Nürnberger and Sommer.

Thus the problem of Lazar–Morris–Wulbert is completely solved in the case that $X = [a, b]$.

1. The main result. In the following, let G be always an n-dimensional subspace of $C[a, b]$.

An important role is played in our studies by a very interesting class of finite dimensional subspaces of $C[a, b]$, the weak Chebyshev spaces.

DEFINITION 1.1. G is called a *Haar space* if each $g \in G$ has at most $n-1$ zeros on $[a, b]$. G is called *weak Chebyshev* if each $g \in G$ has at most $n-1$ sign changes; i.e., there do not exist points $a \leqq x_0 < x_1 < \cdots < x_n \leqq b$ such that $g(x_i)g(x_{i+1}) < 0$ for $i = 0, \cdots, n-1$.

We denote the *class of all n-dimensional weak Chebyshev subspaces* of $C[a, b]$ by W_n.

Recently, Nürnberger [7] has shown that all spaces G which fail to be weak Chebyshev do not admit any continuous selection:

THEOREM 1.2. *If G admits a continuous selection, then* $G \in W_n$.

Therefore, this result restricts our problem to the case that $G \in W_n$. This case requires a detailed knowledge of the properties of the elements of W_n. To simplify our considerations we divide W_n into certain subclasses. At first we define the following two subclasses of W_n:

$$Y_n := \{G \in W_n : \text{no nonzero } g \in G \text{ vanishes on an interval}\},$$

$$Z_n := \{G \in W_n : G \notin Y_n\}.$$

(Here and in the following we use the notation "interval" only for *nondegenerate* intervals.)

The class Y_n has been studied by Nürnberger–Sommer [8] and Sommer [10]. Their results characterize those elements of Y_n which admit continuous selections:

THEOREM 1.3. *Let* $G \in Y_n$. *Then the following conditions are equivalent*: (1.1) *There exists a continuous selection.* (1.2) *Each nonzero* $g \in G$ *has at most n distinct zeros.*

There still remains the case that $G \in Z_n$. This subclass of W_n seems to be the most interesting subclass of W_n, since the spline spaces and also the continuously composed Haar spaces belong to Z_n (see [12]). We now want to divide Z_n into three further subclasses. To do this we first need a condition on zero intervals given by Bartelt [1].

DEFINITION 1.4. We say that G satisfies *condition* (I) if there exists a $\delta > 0$ such that if $g \in G$ and $g \equiv 0$ on $[c, d] \subset [a, b]$ where $c < d$ and $c, d \in \overline{\{x : g(x) \neq 0\}} \cup \{a, b\}$, then $d - c \geqq \delta$.

If some G does not satisfy this condition, then there are elements of G having arbitrarily small zero intervals (see [1]).

Moreover, we need the following notation: A point $x_0 \in [a, b]$ is said to be *vanishing* (resp., *nonvanishing*) with respect to G if $g(x_0) = 0$ for all $g \in G$ (resp., if there is a $g \in G$ with $g(x_0) \neq 0$).

In the following, the term "with respect to G" will be omitted. Then we define:

$Z_n^1 := \{G \in Z_n : G$ satisfies condition (I) and all $x \in [a, b]$ are nonvanishing$\}$,

$Z_n^2 := \{G \in Z_n : G$ satisfies condition (I) and at least one $x \in [a, b]$ is vanishing$\}$,

$Z_n^3 := \{G \in Z_n : G$ does not satisfy condition (I)$\}$.

Then $Z_n = Z_n^1 \cup Z_n^2 \cup Z_n^3$.

In [11] we have studied the class Z_n^1 and have been able to characterize those elements of Z_n^1 which admit continuous selections. To give this result and to establish also a characterization of the elements of Z_n^2, we need the following property of the elements of $Z_n^1 \cup Z_n^2$ which we have shown in [12]:

LEMMA 1.5. *The following statements hold*:

(1.3) *If $G \in Z_n^1$, then there exists a minimal set of points (which we call knots) $a = x_0 < x_1 < \cdots < x_s = b$ such that the spaces $G|_{[x_i, x_{i+1}]}$ are Haar spaces with dimension $n_{i,i+1}$ for $i = 0, \cdots, s-1$.*

(1.4) *If $G \in Z_n^2$, then there exists a minimal set of points (which we call knots again) $a = x_0 < x_1 < \cdots < x_s = b$ such that the spaces $G|_{[x_i, x_{i+1}]}$ are weak Chebyshev with dimension $n_{i,i+1}$, $n_{i,i+1} \geqq 0$, for $i = 0, \cdots, s-1$. Furthermore, no nonzero $g \in G|_{[x_i, x_{i+1}]}$ vanishes on a subinterval of $[x_i, x_{i+1}]$.*

Statement (1.3) justifies the following natural convention:

DEFINITION 1.6. An element $G \in Z_n^1$ is called a *generalized spline space.*

The existence (resp., the nonexistence) of a continuous selection essentially depends on the behavior of certain error functions $f - g_0$, $f \in C[a, b]$, $g_0 \in P_G(f)$, in certain subintervals $[c, d]$ of $[a, b]$. One will see that if $G \in Z_n^1 \cup Z_n^2$, then these subintervals will be exactly the knot-intervals $[x_i, x_j]$ where the knots will be given by Lemma 1.5. Because of this observation and, since we want to completely characterize those weak Chebyshev spaces which admit continuous selections, we define a minimal set of knots for each $G \in Y_n \cup Z_n^3$, too.

We first need some definitions:

DEFINITION 1.7. We say that m zeros $x_1, x_2, \cdots, x_m$ (resp., m zero intervals $I_1 = [y_1, z_1]$, $I_2 = [y_2, z_2]$, $\cdots$, $I_m = [y_m, z_m]$) of a function $f \in C[a, b]$ are *separated* if there exist $m-1$ points $t_i \in (x_i, x_{i+1})$ (resp., $m-1$ points $t_i \in (z_i, y_{i+1})$) such that $f(t_i) \neq 0$ for $i = 1, \cdots, m-1$.

Furthermore, we denote by $Z(f)$ the *set of the zeros of f*, by bd A the *boundary of a set A* and by $|A|$ the *number of the elements of A*.

Then we set:

If $G \in Y_n$, then we define $x_0 := a$ and $x_1 := b$ and choose $\{x_0, x_1\}$ as a *minimal set* of *knots*. This is an immediate consequence of statement 1.4 of Lemma 1.5, since by definition of Y_n no nonzero $g \in G$ vanishes on a subinterval of $[a, b]$.

If $G \in Z_n^3$, then we distinguish two cases:

(i) There exists a $g \in G$ vanishing on two separated intervals. In this case we define $x_0 := a$, $x_1 := b$ and choose $\{x_0, x_1\}$ as a *minimal set* of *knots*.

(ii) There does not exist any $g \in G$ vanishing on two separated intervals. In Theorem 2.7 we will show the existence of a $g_0 \in G$ such that $g_0 \equiv 0$ on a subinterval $[c, d]$ of $[a, b]$ and

$$|\text{bd } Z(g_0)| \geqq \dim \{g \in G : g \equiv 0 \text{ on } [c, d]\} + 1.$$

In this case we define $x_0 := a$, $x_1 := c$, $x_2 := d$, $x_3 := b$ and choose the set of these points as a *minimal set* of *knots*.

It is easily verified that if $G \in Z_n^3$, then in general the above defined knots do *not* guarantee the same properties as the knots of G, in case $G \in Y_n \cup Z_n^1 \cup Z_n^2$. However, they are very helpful, summarizing all results on weak Chebyshev spaces and continuous selections in a characterization theorem.

We need some notation again: Let $G \in W_n$ and $\{x_i\}_{i=0}^s$ be a minimal set of knots according to Lemma 1.5 and to the above definition, respectively. Then for each $i, j \in \{0, \cdots, s\}$, $i < j$, we set:

$$\bar{G}_{ij} := \{g \in G: g \equiv 0 \text{ on } [x_i, x_j]\}, \qquad m_{ij} := \dim \bar{G}_{ij},$$
$$G^{ij} := G|_{[x_i, x_j]}, \qquad n_{ij} := \dim G^{ij}.$$

Now we are in position to completely characterize those weak Chebyshev spaces which admit continuous selections.

THEOREM 1.8. *Let $G \in W_n$ and $\{x_i\}_{i=0}^s$ be a minimal set of knots. Then the following statements* (1.5) *and* (1.6) *are equivalent*:

(1.5) *There exists a continuous selection.*

(i) *No $g \in G$ vanishes on two separated intervals*;

(1.6) (ii) *For all $i, j \in \{0, \cdots, s\}$, $i < j$, and each $g \in \bar{G}_{ij}$, $|\text{bd } Z(g)| \leqq m_{ij}$*;

(iii) *For each $i \in \{0, \cdots, s-1\}$ and each nonzero $g \in G^{i,i+1}$, $|Z(g)| \leqq n_{i,i+1}$.*

In the following sections we will prove this theorem in the case that $G \in Z_n^2 \cup Z_n^3$. If $G \in Y_n$, then the above theorem immediately follows from Theorem 1.3 by using the above defined set of knots for G. One easily sees that in this case (1.6i) and (1.6ii) are always satisfied. If $G \in Z_n^1$, then the above theorem immediately follows from Theorem 3.1 in [11]. In that paper we have completely characterized those generalized spline spaces which admit continuous selections. For such spaces condition (1.6iii) always holds because of Lemma 1.5.

Before studying the case that $G \in Z_n^2 \cup Z_n^3$ we first want to show by some examples that *none* of the conditions (1.6i), (1.6ii), (1.6iii) can be omitted.

Example 1. Let $g_1, g_2 \in C[0, 4]$ be defined by

$$g_1(x) := |x-2| \text{ and } g_2(x) := \begin{cases} x-1 & \text{if } x \in [0, 1], \\ 0 & \text{if } x \in [1, 3], \\ x-3 & \text{if } x \in [3, 4]. \end{cases}$$

Then $G = \text{span}\{g_1, g_2\} \in W_2$. Furthermore $G \in Z_2^2$, since $\tilde{x} = 2$ is vanishing. Lemma 1.5 implies the existence of a minimal set of knots $x_0 = 0$, $x_1 = 1$, $x_2 = 3$, $x_3 = 4$. Then it is easily verified that conditions (1.6i) and (1.6iii) always hold, but condition (1.6ii) fails, because $|\text{bd } Z(g_2)| = 2$ and $\bar{G}_{12} = \text{span}\{g_2\}$.

Example 2. Let $g_1, g_2, g_3 \in C[0, 4]$ be defined by

$$g_1(x) := \begin{cases} 0 & \text{if } x \in [0, 3], \\ x-3 & \text{if } x \in [3, 4], \end{cases} \qquad g_2(x) := \begin{cases} 1-x & \text{if } x \in [0, 1], \\ 0 & \text{if } x \in [1, 4], \end{cases} \qquad g_3(x) := |\cos(\pi/2)x|.$$

Then $G = \text{span}\{g_1, g_2, g_3\} \in W_3$. Furthermore $G \in Z_3^2$, since $\tilde{x} = 1$ and $\tilde{\tilde{x}} = 3$ are vanishing. Lemma 1.5 implies the existence of a minimal set of knots $x_0 = 0$, $x_1 = 1$, $x_2 = 3$, $x_3 = 4$. Then it is easily verified that the conditions (1.6i) and (1.6ii) always hold, but condition (1.6iii) fails, because $G^{12} = \text{span}\{g_3|_{[1,3]}\}$ and g_3 has two distinct zeros on $[1, 3]$.

Example 3. Let $g_1, g_2, g_3 \in C[0, 4]$ be defined by

$$g_1(x) := \begin{cases} 0 & \text{if } x \in [0, 3], \\ x-3 & \text{if } x \in [3, 4], \end{cases} \qquad g_2(x) := \begin{cases} 3-x & \text{if } x \in [0, 3], \\ 0 & \text{if } x \in [3, 4], \end{cases}$$

$$g_3(x) := \begin{cases} 0 & \text{if } x \in [0, 1], \\ -\cos(\pi/2)x & \text{if } x \in [1, 3], \\ 0 & \text{if } x \in [3, 4]. \end{cases}$$

Then $G = \operatorname{span}\{g_1, g_2, g_3\} \in W_3$. Furthermore $G \in Z_3^2$, since $\tilde{x} = 3$ is vanishing. As above, Lemma 1.5 yields a minimal of knots $x_0 = 0$, $x_1 = 1$, $x_2 = 3$, $x_3 = 4$. Then the conditions (1.6ii) and (1.6iii) always hold, but condition (1.6i) fails, because g_3 has two separated zero intervals.

At this point we want to remark that if $G \in Z_n^2$, then for all arguments which we will use in the following we always need a *minimal* set of knots. It does not suffice having a finite set of points such that G can be decomposed into weak Chebyshev spaces according to Lemma 1.5. This we want to make clearer by the following example.

Example. Let $g_1, g_2 \in C[0, 4]$ be defined by

$$g_1(x) := \begin{cases} 0 & \text{if } x \in [0, 2], \\ x-2 & \text{if } x \in [2, 4], \end{cases} \qquad g_2(x) := |\sin(\pi/2)x|.$$

Then $G = \operatorname{span}\{g_1, g_2\} \in W_2$. Furthermore $G \in Z_2^2$, since $\tilde{x} = 2$ is vanishing. Then Lemma 1.5 implies the existence of a *minimal* set of knots $x_0 = 0$, $x_1 = 2$, $x_2 = 4$ such that the spaces $G^{i,i+1}$ are weak Chebyshev and no nonzero $g \in G^{i,i+1}$ vanishes on a subinterval of $[x_i, x_{i+1}]$ for $i = 0, 1$. One easily verifies that condition (1.6iii) fails, because $G^{01} = \operatorname{span}\{g_2|_{[0,2]}\}$ and g_2 has two distinct zeros on $[0, 2]$. Therefore, by Theorem 1.8 G does not admit any continuous selection. However, choosing the points $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, $x_3 = 4$ one can see that these points also decompose the space G into the desired weak Chebyshev spaces. Furthermore, one easily verifies that for these points all conditions (1.6) hold. Then by Theorem 1.8 we would obtain a continuous selection for P_G. However, this conclusion does not hold, because the set $\{0, 1, 2, 4\}$ has four elements, while the preceding set of knots $\{0, 2, 4\}$ has only three elements. Therefore, the set $\{0, 1, 2, 4\}$ is *not* a minimal set of knots.

2. Nonexistence of continuous selections for each element of Z_n^3. In this section we will show the nonexistence of a continuous selection for P_G, in case $G \in Z_n^3$. This class is nonempty and it is not difficult to construct many elements (see Bartelt [1]).

To prove the nonexistence we need the following fundamental lemma established by Lazar–Morris–Wulbert:

LEMMA 2.1. *Let s be a continuous selection for P_G. If $f \in C[a, b]$, $\|f\| = 1$ and $O \in P_G(f)$, then there is a $g_0 \in P_G(f)$ such that:*

(i) *for every $x \in \operatorname{bd} Z(P_G(f)) \cap f^{-1}(1)$ and every $g \in P_G(f)$ there is a neighborhood U of x for which $g_0 \geqq g$ on U, and*

(ii) *for every $x \in \operatorname{bd} Z(P_G(f)) \cap f^{-1}(-1)$ and every $g \in P_G(f)$ there is a neighborhood V of x for which $g_0 \leqq g$ on V.*

Here we set: $Z(P_G(f)) := \{x \in [a, b]: g(x) = 0 \text{ for all } g \in P_G(f)\}$. As defined in § 1, bd denotes the set of the boundary points. Furthermore, we need the following lemmas:

LEMMA 2.2 (Stockenberg [14]). *Let $G \in W_n$. Then the following statements hold:*

(2.1) *If there is a $g \in G$ with n separated, nonvanishing zeros $x_1 < x_2 < \cdots < x_n$, then $g(x) = 0$ for all x with $x \leqq x_1$ and $x \geqq x_n$;*

(2.2) *No $g \in G$ has more than n separated, nonvanishing zeros.*

LEMMA 2.3 (Sommer [12]). *Let $G \in W_n$. Then for every subinterval $[c, d]$ of $[a, b]$ the space $\tilde{G} = G|_{[c,d]}$ is weak Chebyshev with dimension $m \leqq n$.*

Moreover, we need the following definition:

DEFINITION 2.4. A zero x_0 of an $f \in C[a, b]$ is said to be an *isolated zero* if there is a neighborhood U of x_0 such that $f(x) \neq 0$ for each $x \in U \backslash \{x_0\}$. A zero x_0 of f is said to be a *double zero* if x_0 is an isolated zero on (a, b) and f does not change sign at x_0. A zero x_0 of f is said to be a *simple zero* if x_0 is not a double zero of f or if $x_0 = a$ or $x_0 = b$. We denote the *set of all double zeros* of f by $Z_d(f)$ and by $|Z^*(f)|$ the *number of the zeros* of f counting simple zeros as one zero and double zeros as two zeros.

Furthermore, we set for two intervals $I_1 = [x_1, y_1]$, $I_2 = [x_2, y_2]$:

$$I_1 < I_2 \quad \text{if } y_1 < x_2,$$

$$[I_1, I_2] := [y_1, x_2] \quad \text{if } I_1 < I_2.$$

If $I_1, I_2, \cdots, I_m$ are finitely many subintervals of $[a, b]$ and G is a subspace of $C[a, b]$, then we define the space $G[I_1, I_2, \cdots, I_m]$ by

$$G[I_1, I_2, \cdots, I_m] := \{g \in G : g \equiv 0 \text{ on } \cup_{i=1}^{m} I_i\}.$$

We are now able to prove two lemmas on nonexistence. One will see that these lemmas hold not only for a subclass of Z_n^3, but also for a more general subclass of W_n.

LEMMA 2.5. *Let $G \in W_n$. Let at least one function $\tilde{g} \in G$ have two separated zero intervals. Then there does not exist any continuous selection.*

Proof. Lemma 2.2 shows that each $g \in G$ has at most n separated, nonvanishing zeros. Therefore, it is easily verified the existence of finitely many subintervals $I_1 < I_2 < \cdots < I_m$ of $[a, b]$ such that the following is satisfied: There exist two intervals $I_k, I_{k+1} \in \{I_1, I_2, \cdots, I_m\}$ such that $\dim G[I_1, \cdots, I_m]|_{[I_k, I_{k+1}]} \geqq 1$ and no $g \in G[I_1, \cdots, I_m]$, $g \not\equiv 0$ on $[I_k, I_{k+1}]$, vanishes on a subinterval of $[I_k, I_{k+1}] \backslash (Z(G) \cap [I_k, I_{k+1}])$ where $Z(G) := \{x \in [a, b] : g(x) = 0 \text{ for all } g \in G\}$ the set of all vanishing zeros with respect to G.

We set $\tilde{G} := G[I_1, \cdots, I_m]$. Since G is weak Chebyshev, there exists a $g_0 \in \tilde{G}$ such that g_0 has exactly $r \geqq 0$ sign changes in $[I_k, I_{k+1}]$ and no $g \in \tilde{G}$ has more than r sign changes in this interval. We set $\|g_0\| = 1$.

Without loss of generality we may assume that g_0 changes sign only at zero intervals. Since, by assumption, all zero intervals of g_0 which are subsets of $[I_k, I_{k+1}]$ must be subsets of $Z(G)$, there exist, therefore, r zero intervals $J_i = [y_i, z_i]$, $i = 1, \cdots, r$, $J_i \subset Z(G)$, of g_0 such that $[y_0, z_0] := I_k < J_1 < \cdots < J_r < I_{k+1} =: [z_{r+1}, y_{r+1}]$ and $g_0(z_{i-1} + \varepsilon) g_0(z_i + \varepsilon) < 0$, $i = 1, \cdots, r$ for $\varepsilon > 0$ sufficiently small. Moreover let $\varepsilon > 0$ such that $\{y_1, y_2, \cdots, y_r, z_{r+1}\} \cap [z_i, z_i + \varepsilon] = \varnothing$ for $i = 0, \cdots, r$. Without loss of generality we furthermore may assume that $g_0(x) > 0$ for each $x \in (z_0, z_0 + \varepsilon]$ and $g_0(x) \neq 0$ for each $x \in [z_{r+1} - \varepsilon, z_{r+1})$. This implies that $(-1)^{r+1} g_0(x) < 0$ for each $x \in [z_{r+1} - \varepsilon, z_{r+1})$.

We now construct a function $f \in C[a, b]$ as follows:

(a) $f(x) = (-1)^i$ for each $x \in [z_i, z_i + \varepsilon]$, $i = 0, \cdots, r$ and $f(z_{r+1}) = (-1)^{r+1}$.

(b) On the intervals $I_1, \cdots, I_m$ as follows: If $\dim G|_{I_1} = n_1 \geqq 1$, then it is easily verified that there are n_1 subintervals $[u_i, v_i]$, $i = 1, \cdots, n_1$ of I_1 such that we can interpolate by the elements of $G|_{I_1}$ at any n_1 tuple $(w_1, \cdots, w_{n_1})$ where $w_i \in [u_i, v_i]$ for $i = 1, \cdots, n_1$. We set: $f(u_i) = 1$ and $f(v_i) = -1$ for $i = 1, \cdots, n_{1_i}$. We define f analogously on the intervals $I_2, \cdots, I_m$.

(c) $\max\{-1 + g_0(x), -1\} \leqq f(x) \leqq \min\{1 + g_0(x), 1\}$ for each $x \in [a, b]$.

Then it follows that $\|f-0\| = \|f-g_0\| = 1$. Since by Lemma 2.3 each space $G|_{I_i}$ is weak Chebyshev with dimension n_i for $i=1,\cdots,m$ and $f-0$ has at least n_i+1 alternating extreme points on I_i, in case $n_i \geqq 1$, it is easily verified that 0 is a best approximation of f from $G|_{I_i}$ on I_i. Then $\|f\|=1$ implies that $0 \in P_G(f)$, too. Since we can interpolate at any n_1-tuple $(w_1,\cdots,w_{n_1})$ where $w_i \in [u_i, v_i]$, all best approximations of f from G must vanish identically on I_1 and, analogously, on I_i for $i=2,\cdots,m$. This implies that $P_G(f) \subset \tilde{G}$. Furthermore 0, $g_0 \in P_G(f)$ and z_0, $z_{r+1} \in \operatorname{bd} Z(P_G(f))$.

Now let $g \in P_G(f)$, $g \not\equiv 0$ on $[z_0, z_{r+1}]$. By definition of $\tilde{G}$ and, since g_0 does not vanish identically on $[z_i, z_i+\varepsilon]$ for $i=0,\cdots,r$, the function g does also not vanish identically on these intervals. Then it easily follows from the definitions of f and g_0 that g changes sign exactly r times on (z_0, z_{r+1}). Thus $(-1)^{r+1}g(x)<0$ for each $x \in [z_{r+1}-\varepsilon, z_{r+1})$.

We now apply Lemma 2.1: If there exists a continuous selection, then there would exist a $\bar{g} \in P_G(f)$ such that for z_0 and g_0 there is a neighborhood U of z_0 for which $\bar{g} \geqq g_0$ on U and for z_{r+1} and $0 \in P_G(f)$ there is a neighborhood V of z_{r+1} for which $(-1)^{r+1}\bar{g} \geqq 0$ on V. Since $\bar{g} \geqq g_0$ on U, the function $\bar{g}$ would be positive on $(z_0, z_0+\varepsilon] \cap U$. This would imply that $(-1)^{r+1}\bar{g}<0$ on $[z_{r+1}-\varepsilon, z_{r+1}) \cap V$ as has been shown above. But this is a contradiction to Lemma 2.1. This concludes the proof.

The next lemma studies the case that no $g \in G$ vanishes on two separated intervals. We first want to note that in § 1 for any two points x_i, $x_j \in [a, b]$ the space G_{ij} has been defined by $\bar{G}_{ij} = \{g \in G: g \equiv 0 \text{ on } [x_i, x_j]\}$.

LEMMA 2.6. *Let $G \in W_n$. Let no $g \in G$ vanish on two separated intervals. If there exist two points x_i, $x_j \in [a, b]$, $x_i < x_j$ and a function $\tilde{g} \in \bar{G}_{ij}$ such that $|\operatorname{bd} Z(\tilde{g})| \geqq m_{ij}+1$, then there does not exist any continuous selection.*

Proof. Assume that there are two points $x_i, x_j \in [a, b]$ and a $\tilde{g} \in \bar{G}_{ij}$ such that $|\operatorname{bd} Z(\tilde{g})| \geqq m_{ij}+1$ (for notations see § 1). Then we distinguish the following two cases. First, there is an integer q such that $|\operatorname{bd} Z(g)| \leqq q$ for each $g \in \bar{G}_{ij}$. Then we can proceed exactly as in the proof of Lemma 3.3 in [11] to prove the nonexistence. Second, for each integer p there exists a $g_p \in \bar{G}_{ij}$ such that $|\operatorname{bd} Z(g_p)| \geqq p$. Then for each $g \in \bar{G}_{ij}$ we denote the number of the sign changes of g in (a, x_i) by l_g and the number of the sign changes of g in (x_j, b) by r_g. Then it follows from $\bar{G}_{ij} \subset G$ and $G \in W_n$ that $l_g + r_g \leqq n-1$. We now choose a $g_0 \in \bar{G}_{ij}$, $\|g_0\| = 1$ such that $l_{g_0} + r_{g_0} \geqq l_g + r_g$ for each $g \in \bar{G}_{ij}$. Let $[x_k, x_l] \supset [x_i, x_j]$ be the maximal zero interval of g_0 (remember that g_0 has only one zero interval). Then we distinguish:

(i) There is an integer q such that $|\operatorname{bd} Z(g)| \leqq q$ for each $g \in \bar{G}_{kl}$. Then we may conclude as in the first case.

(ii) For each integer p there exists a $g_p \in \bar{G}_{kl}$ such that $|\operatorname{bd} Z(g_p)| \geqq p$.

This implies that g_p has at least $p-2$ separated zeros in $[a, x_k) \cup (x_l, b]$. Then Lemma 2.2 shows that at least $p-2-n$ of these zeros are vanishing. Hence for each integer p, $p \geqq n+2$, each $g \in G$ has at least $p-n-2$ vanishing zeros with respect to G in $[a, x_k) \cup (x_l, b]$. Since g_0 vanishes identically only on the interval $[x_k, x_l]$, by the preceding arguments this function must have at least one *double* vanishing zero $\tilde{x} \in (a, x_k) \cup (x_l, b)$. Without loss of generality let $\tilde{x} > x_l$.

We now only study the case that $a < x_k < x_l < b$, since the cases $a = x_k$ and $x_l = b$ follow analogously. Furthermore, we only study the case that $x_k \in \overline{\{x \in [a, b]: g_0(x) < 0\}}$ and $x_l \in \overline{\{x \in [a, b]: g_0(x) > 0\}}$. This implies that $g_0 < 0$ on $[x_k-\delta, x_k)$ and $g_0 > 0$ on $(x_l, x_l+\delta]$ for $\delta > 0$ sufficiently small. Now we set $r := l_{g_0}$, $s := r_{g_0}$ and $m := r+s+4$. Then the function g_0 has exactly r sign changes at the points $a = \tilde{z}_0 < \tilde{z}_1 < \cdots < \tilde{z}_r < \tilde{z}_{r+1} = x_k$ and exactly s sign changes at the points $x_l = \tilde{z}_{r+2} < \tilde{z}_{r+3} < \cdots < \tilde{z}_{m-2} < \tilde{z}_{m-1} = b$. Since $\tilde{x} \in Z_d(g_0)$ and $\tilde{x} > x_l$, there is an integer $p \in \{r+2, \cdots, m-2\}$ such that

$\tilde{x} \in (\tilde{z}_p, \tilde{z}_{p+1})$. We denote the points $\tilde{z}_0, \cdots, \tilde{z}_p, \tilde{x}, \tilde{z}_{p+1}, \cdots, \tilde{z}_{m-1}$ by $z_0, z_1, \cdots, z_m$. Then $\tilde{x} = z_{p+1}$.

We now choose $m-1$ points $t_i \in (z_i, z_{i+1})$, $i = 0, \cdots, m-1$, $i \neq r+1$, satisfying $g_0(t_i) \neq 0$ and we construct a function $f \in C[a, b]$ as follows:

(a) $f(x_k) = 1$ and $f(x) = 1$ for each $x \in [x_l, t_{r+2}]$;

(b) $f(x) = \operatorname{sgn} g_0(t_i)$ for each $x \in [z_i, t_i]$, $i = 0, \cdots, m-1$, $i \neq r+1, r+2, p+1$;

(c) $f(z_{p+1}) = -\operatorname{sgn} g_0(t_{p+1})$ and $f(x) = \operatorname{sgn} g_0(t_{p+1})$ for each $x \in [t_{p+1} - \varepsilon, t_{p+1} + \varepsilon]$ where we choose $\varepsilon > 0$ such that $z_{p+1} < t_{p+1} - \varepsilon$ and $t_{p+1} + \varepsilon < z_{p+2}$;

(d) On $[x_k, x_l]$: If $\dim G|_{[x_k, x_l]} = n_{kl} \geqq 1$, then it is readily verified that there are n_{kl} subintervals $[u_i, v_i]$, $i = 1, \cdots, n_{kl}$ of (x_k, x_l) such that we can interpolate by the elements of $G|_{[x_k, x_l]}$ at any n_{kl} tuple $(w_1, \cdots, w_{n_{kl}})$ where $w_i \in [u_i, v_i]$ for $i = 1, \cdots, n_{kl}$. We set: $f(u_i) = 1$ and $f(v_i) = -1$ for $i = 1, \cdots, n_{kl}$;

(e) $\max\{-1 + g_0(x), -1\} \leqq f(x) \leqq \min\{1 + g_0(x), 1\}$ for each $x \in [a, b]$.

Then it follows that $\|f - 0\| = \|f - g_0\| = 1$. Then using the same kind of arguments as in the proof of Lemma 2.5 we can show that $0, g_0 \in P_G(f)$ and $P_G(f) \subset \bar{G}_{kl}$. Furthermore $x_k, x_l \in \operatorname{bd} Z(P_G(f))$ and, since $\tilde{x} = z_{p+1}$ is vanishing, it follows that $\tilde{x} \in \operatorname{bd} Z(P_G(f))$, too.

Without loss of generality we may assume that $f(\tilde{x}) = 1$. Then $\tilde{x} \in \operatorname{bd} Z(P_G(f)) \cap f^{-1}(1)$. Now we apply Lemma 2.1: If there exists a continuous selection, then there would exist a $g_1 \in P_G(f)$ such that for x_l and g_0 there is a neighborhood U of x_l for which $g_1 \geqq g_0$ on U and for $\tilde{x} = z_{p+1}$ and $0 \in P_G(f)$ there is a neighborhood V of $\tilde{x}$ for which $g_1 \geqq 0$ on V.

Since $g_1 \geqq g_0 > 0$ on $(x_l, x_l + \delta)$ for $\delta > 0$ sufficiently small, the function g_1 does not vanish on a subinterval of $[x_l, b]$. Then $g_1 \geqq 0$ on V implies that $\tilde{x} \in Z_d(g_1)$. Furthermore, it immediately follows from the definition of f that $g_0 g_1 \geqq 0$ on $[z_i, t_i]$ for $i = 0, \cdots, m-1$, $i \neq r+1$, $p+1$ and $g_0 g_1 \geqq 0$ on $[t_{p+1} - \varepsilon, t_{p+1} + \varepsilon]$. Then it is easily verified that for all sufficiently large positive numbers c the function $g_0 + c g_1$ does not vanish on a subinterval of $[a, x_k] \cup [x_l, b]$ and has at least as many sign changes in (a, x_k) and on $(x_l, t_p) \cup (t_{p+1}, b)$, respectively, as g_0. In addition to this it follows from the definition of f that $g_1 \leqq 0$ on $[z_p, t_p] \cup [t_{p+1} - \varepsilon, t_{p+1} + \varepsilon]$. Hence for sufficiently large $c > 0$, $g_0 + c g_1$ has at least two sign changes in (t_p, t_{p+1}) and, therefore, it follows that $l_{g_0 + c g_1} + r_{g_0 + c g_1} \geqq 2 + 1_{g_0} + r_{g_0}$. However, since $g_0 + c g_1 \in \bar{G}_{kl}$, this contradicts our assumption that $l_{g_0} + r_{g_0} \geqq l_g + r_g$ for each $g \in \bar{G}_{kl}$. Hence there may not exist such a g_1. This concludes the proof.

Using the preceding lemmas we are now able to prove the nonexistence of continuous selections for P_G, in case $G \in Z_n^3$.

THEOREM 2.7. *Let $G \in Z_n^3$. Then there does not exist any continuous selection.*

Proof. We distinguish two possibilities. First, there exists a function $\tilde{g} \in G$ vanishing on two separated intervals. Then the statement follows from Lemma 2.5.

Second, there does not exist any $g \in G$ vanishing on two separated intervals. Then we will show that there exist two points $x_i, x_j \in [a, b]$, $x_i < x_j$ and a $\tilde{g} \in \bar{G}_{ij}$ such that $|\operatorname{bd} Z(\tilde{g})| \geqq m_{ij} + 1$.

We assume to the contrary that for any choice of $x_i, x_j \in [a, b]$, $x_i < x_j$ *and each* $g \in \bar{G}_{ij}$, $|\operatorname{bd} Z(g)| \leqq m_{ij}$. Then it is easily verified that the statements of [11, Lemma 2.1] hold. Using that lemma we show that for each integer i there is a $g_i \in G$ vanishing on an interval $I_i = [x_i, y_i]$ such that either $a < x_1 < x_2 < \cdots < b$ or $a < y_1 < y_2 < \cdots < b$. Let $h_1 \in G$ have the zero interval $\tilde{I}_1 = [\tilde{x}_1, \tilde{y}_1]$. Since $G \in Z_n^3$, there is a $h_2 \in G$ having a zero interval $\tilde{I}_2 = [\tilde{x}_2, \tilde{y}_2]$ such that $\tilde{y}_2 - \tilde{x}_2 < \frac{1}{2}(\tilde{y}_1 - \tilde{x}_1)$. Furthermore $h_2 \neq h_1$, since each $g \in G$ has at most one zero interval. Let, without loss of generality, $\tilde{y}_1 < \tilde{y}_2$. By a repeated application of the preceding argument, for each fixed integer m we find—by

rearranging—m functions $g_i \in G$ such that each g_i vanishes on $I_i = [x_i, y_i]$ and $y_1 < y_2 < \cdots < y_m < b$. Then using Lemma 2.1 in [11] we obtain m functions $\tilde{g}_i \in G$ satisfying $\tilde{g}_i = g_i$ on $[y_i, b]$ and $g_i \equiv 0$ on $[a, y_i]$. Then $y_1 < y_2 < \cdots < y_m$ implies that these functions must be linearly independent. This we can do for each integer m. Therefore, we get a contradiction of our assumption that G is finite dimensional. Thus we have shown that there must exist two points $x_i, x_j \in [a, b]$, $x_i < x_j$ and a $\tilde{g} \in G$ such that $|\text{bd } Z(\tilde{g})| \geqq m_{ij} + 1$. Then the statement of this theorem follows from Lemma 2.6.

Remark. If for each $G \in Z_n^3$ we choose such a set of knots as we have defined in § 1, then the proof of Theorem 1.8 immediately follows from Theorem 2.7.

3. Characterization of continuous selections for the elements of Z_n^2. In this section we will characterize those elements of Z_n^2 which admit continuous selections.

We first remember that by Lemma 1.5 for each $G \in Z_n^2$ there exists a minimal set of knots $\{x_i\}_{i=0}^s$ such that the spaces $G^{i,i+1}$ are weak Chebyshev and no nonzero $g \in G^{i,i+1}$ vanishes on a subinterval of $[x_i, x_{i+1}]$ for $i = 0, \cdots, s-1$.

Using such sets of knots we define the subclass $\tilde{Z}_n^2$ of Z_n^2 by

$$Z_n^2 := \{G \in Z_n^2 : \text{(i) No } g \in G \text{ vanishes on two separated intervals;}$$
$$\text{(ii) for all } i,j \in \{0, \cdots, s\},\ i < j, \text{ and each } g \in \bar{G}_{ij},\ |\text{bd } Z(g)| \leqq m_{ij};$$
$$\text{(iii) for each } i \in \{0, \cdots, s-1\} \text{ and each nonzero } g \in G^{i,j+1},\ |Z(g)| \leqq n_{i,i+1}\}.$$

Then we can show the nonexistence of a continuous selection for each element of $Z_n^2 \backslash \tilde{Z}_n^2$.

LEMMA 3.1. *Let $G \in Z_n^2 \backslash \tilde{Z}_n^2$. Then there does not exist any continuous selection.*

Proof. Let $G \in Z_n^2 \backslash \tilde{Z}_n^2$. If condition (i) or condition (ii) of the definition of $\tilde{Z}_n^2$ fails, then the statement of this lemma immediately follows from Lemma 2.5 and Lemma 2.6 respectively. Therefore, we have only yet to consider the case that there is a $k \in \{0, \cdots, s-1\}$ and a nonzero $g_0 \in G^{k,k+1}$ such that g_0 has at least $n_{k,k+1}+1$ distinct zeros on $[x_k, x_{k+1}]$. Furthermore, Lemma 1.5 shows that $G^{k,k+1}$ is weak Chebyshev with dimension $n_{k,k+1}$. Thus we have got the same situation as in [10] and we can proceed exactly as in that paper to show the nonexistence of a continuous selection.

This result implies that we have only yet to consider the case that $G \in \tilde{Z}_n^2$. In this case we are able to show the existence of a continuous selection. To construct such a selection we first need some lemmas.

LEMMA 3.2. *Let $G \in \tilde{Z}_n^2$. Then for all $i, j \in \{0, \cdots, s\}$, $i < j$, the following statements hold:*

(3.1) *G_{ij} is weak Chebyshev with dimension m_{ij}.*

(3.2) *For each $g_1 \in \bar{G}_{ij}$ there is a $\tilde{g}_1 \in G$ such that $\tilde{g}_1 = g_1$ on $[x_j, b]$ and $\tilde{g}_1 \equiv 0$ on $[a, x_j]$.*

(3.3) *For each $g_2 \in \bar{G}_{ij}$ there is a $\tilde{g}_2 \in G$ such that $\tilde{g}_2 = g_2$ on $[a, x_i]$ and $\tilde{g}_2 \equiv 0$ on $[x_i, b]$.*

This lemma can be proved analogously to [12, Lemma 4.3].

LEMMA 3.3 (Jones–Karlovtiz [4]). *The following statements are equivalent:*

(3.4) *$G \in W_n$.*

(3.5) *Given $a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b$ there exists a nonzero $g \in G$ such that $(-1)^i g(x) \geqq 0$, $t_{i-1} < x < t_i$, $i = 1, \cdots, n$.*

In the following, let $Z(G)$ be again the set of all vanishing zeros with respect to G. Then the definition of Z_n^2 shows that for each $G \in Z_n^2$ this set is nonempty.

In addition to this we show the following property:

LEMMA 3.4. *Let $G \in \tilde{Z}_n^2$. Then $Z(G) = [\tilde{x}, \tilde{\tilde{x}}]$ with $\tilde{x} \leqq \tilde{\tilde{x}}$.*

Proof. We distinguish two possibilities. First, $Z(G) = \{\tilde{x}\}$. Then the statement trivially holds. Second, there are at least two points $\tilde{x}, \tilde{\tilde{x}} \in Z(G)$. Assume that there is a $\tilde{g} \in G$ and a $y_0 \in (\tilde{x}, \tilde{\tilde{x}})$ such that $\tilde{g}(y_0) \neq 0$. We distinguish once more:

(i) Let $\tilde{g}$ vanish on an interval $[x_k, x_i]$ where $x_k, x_i \in \operatorname{bd} Z(\tilde{g}) \cup \{a, b\}$. Then $\tilde{g}(y_0) \neq 0$ implies that $x_k > y_0$ or $x_i < y_0$. Without loss of generality let $x_i < y_0$. Since $G \in \tilde{Z}_n^2$, Lemma 3.2 shows the existence of a $g_0 \in G$ such that $g_0 = \tilde{g}$ on $[x_i, b]$ and $g_0 \equiv 0$ on $[a, x_i]$. Furthermore, $\bar{G}_{0i}$ is weak Chebyshev with dimension $m_{0i} \geqq 1$. Let $\bar{x} < t_1 < t_2 < \cdots < t_{m_{0i}-1} < b$ be points where $\bar{x} = \max\{\tilde{\tilde{x}}, x_{s-1}\}$ if $\tilde{\tilde{x}} < b$ and $\bar{x} = x_{s-1}$ if $\tilde{\tilde{x}} = b$. Then, since no $g \in \bar{G}_{oi}$ vanishes on two separated intervals and all zero intervals of the elements of $\bar{G}_{oi}$ are knot intervals, by Lemma 3.3 there is a $\tilde{g}_0 \in \bar{G}_{oi}$ having exactly $m_{0i} - 1$ sign changes at the points t_p, $p = 1, \cdots, m_{0i} - 1$. This implies that $|\operatorname{bd} Z(\tilde{g}_0)| \geqq m_{0i}$. Since $\tilde{g}_0 \in \bar{G}_{0i}$, it follows that $\tilde{g}_0 \equiv 0$ on $[a, x_j]$ where $x_j \geqq x_i$. Assume that $x_j > x_i$. Then $g_0 \notin \bar{G}_{0j}$, because $x_i \in \operatorname{bd} Z(g_0)$. This implies that $m_{0j} < m_{0i}$. Thus we have shown the existence of a $\tilde{g}_0 \in \bar{G}_{0j}$ satisfying $|\operatorname{bd} Z(\tilde{g}_0)| \geqq m_{0i} \geqq m_{0j} + 1$. But this contradicts our assumption that $G \in \tilde{Z}_n^2$. This proves that $x_j = x_i$. Since $y_0 > x_i$, it follows that $\tilde{\tilde{x}} > x_i$, too. Then because of $\tilde{\tilde{x}} \in Z(G)$ the function $\tilde{g}_0$ has a further separated zero at $\tilde{\tilde{x}}$. This implies that $|\operatorname{bd} Z(\tilde{g}_0)| \geqq m_{0i} + 1$ and contradicts again our assumption that $G \in \tilde{Z}_n^2$. Thus we have shown that $\tilde{g} \equiv 0$ on $[\tilde{x}, \tilde{\tilde{x}}]$.

(ii) *Let* $\tilde{g}$ not vanish on an interval. We set: $x_p = \max\{x_l : x_l \leqq \tilde{x}\}$. In (i) we have shown that each $g \in G$ which vanishes on a subinterval of $[a, b]$ must in particular vanish on $[\tilde{x}, \tilde{\tilde{x}}]$. Then, since all zero intervals of the elements of G are knot intervals and the set of the knots is minimal, it immediately follows from the preceding argument that $\tilde{\tilde{x}} \leqq x_{p+1}$. This means that $[\tilde{x}, \tilde{\tilde{x}}] \subset [x_p, x_{p+1}]$. Since $\tilde{g}$ does not vanish on an interval, it follows that $\dim G^{p,p+1} = n_{p,p+1} \geqq 1$. We now choose $n_{p,p+1} - 1$ points $\tilde{x} < z_1 < \cdots < z_{n_{p,p+1}-1} < \tilde{\tilde{x}}$. Then Lemma 2.3 and Lemma 3.3 imply the existence of a nonzero $g_0 \in G^{p,p+1}$ such that g_0 has $n_{p,p+1} - 1$ zeros with sign changes at z_i for $i = 1, \cdots, n_{p,p+1} - 1$. Furthermore $g_0(\tilde{x}) = g_0(\tilde{\tilde{x}}) = 0$, because $\tilde{x}, \tilde{\tilde{x}} \in Z(G)$. But this contradicts our assumption that $G \in \tilde{Z}_n^2$. Hence there may only exist such functions $g \in G$ which vanish on intervals. Then it follows from (i) that each $g \in G$ must vanish on $[\tilde{x}, \tilde{\tilde{x}}]$. Thus $Z(G)$ has to be a subinterval of $[a, b]$. This concludes the proof.

The preceding lemma shows that only the cases that $Z(G) = \{\tilde{x}\}$ or $Z(G) = [\tilde{x}, \tilde{\tilde{x}}]$ are possible. To study these both cases we need for each $f \in C[a, b]$ certain best approximations $g \in P_G(f)$.

DEFINITION 3.5. If $f \in C[a, b]$, then $g \in P_G(f)$ is called an *alternation element* (AE) of f if there exist $n + 1$ points $a \leqq t_0 < t_1 < \cdots < t_n \leqq b$ such that $\varepsilon(-1)^i (f - g)(t_i) = \|f - g\|$, $i = 0, \cdots, n$, $\varepsilon = \pm 1$. Such points are called *alternating extreme points* of $f - g$.

Jones–Karlovitz [4] have completely characterized those spaces which admit for each $f \in C[a, b]$ at least one AE $g \in P_G(f)$. They have shown that this property is equivalent to weak Chebyshev:

THEOREM 3.6. *The following statements are equivalent*:

(3.6) $G \in W_n$.

(3.7) *For each* $f \in C[a, b]$ *there exists at least one* AE $g \in P_G(f)$.

In the following this theorem will play an important role in constructing continuous selections (see also [8], [9] and [11]). We first study the case that $Z(G) = [\tilde{x}, \tilde{\tilde{x}}]$ with $\tilde{x} < \tilde{\tilde{x}}$.

LEMMA 3.7. *Let* $G \in \tilde{Z}_n^2$ *and* $Z(G) = [\tilde{x}, \tilde{\tilde{x}}]$ *with* $\tilde{x} < \tilde{\tilde{x}}$. *Then there exists a continuous selection.*

Proof. Let $f \in C[a, b]$ and $g_0 \in P_G(f)$ arbitrary. Then by Lemma 3.4 there is an interval $[x_p, x_{p+1}]$ such that $[\tilde{x}, \tilde{\tilde{x}}] \subset [x_p, x_{p+1}]$. This implies that $g = g_0$ on $[x_p, x_{p+1}]$ for all $g \in P_G(f)$. By virtue of this we can proceed exactly as in the proof of Lemma 3.7 in [11] to construct a continuous selection by starting in $[x_p, x_{p+1}]$.

Hence there still remains the case that $Z(G) = \{\tilde{x}\}$. This is more difficult than the above studied case, since in general there is no interval on which all best approximations

coincide. But we are able to show the existence of a knot interval on which all AEs of f from G coincide.

In the following we may assume that $\tilde{x} < b$ and $\tilde{x} \in [x_p, x_{p+1})$. We first show an interpolation property.

LEMMA 3.8. *Let $G \in \tilde{Z}_n^2$. Let n points $a \leqq y_1 < \cdots < y_n \leqq b$ be given satisfying $\tilde{x} \notin \{y_1, \cdots, y_n\}$ and $y_{n-n_{is}} < x_i < y_{n_{0i}+1}$, $i = 1, \cdots, s-1$ (for $n - n_{is} = 0$ and $n_{0i} + 1 = n+1$ the first (resp., the second) inequality is omitted). Then for any n real numbers $\{z_i\}_{i=1}^n$ there exists exactly one $g_0 \in G$ with $g_0(y_i) = z_i$ for $= 1, \cdots, n$.*

Proof. Let any $g_0 \in G$ be given such that $g_0(y_i) = 0$ for $i = 1, \cdots, n$. Then the lemma is proved if we can show that $g_0 \equiv 0$. However, this proof follows directly from the proof of [11, Lemma 2.3] and, therefore, we get the desired statement. Using that proof we have only to note that, by the proof of Lemma 3.4, for each $g \in G$ vanishing on an interval I the point $\tilde{x}$ belongs to I. Furthermore, note that each point y_i is nonvanishing, because $\tilde{x} \notin \{y_1, \cdots, y_n\}$.

The next lemma gives a result on the numbers of zeros of a function having no zero intervals.

LEMMA 3.9. *Let $G \in \tilde{Z}_n^2$. Then for each $g_0 \in G$ having no zero intervals the following statements hold*:

(3.8) $|Z(g_0)| \leqq n$.

(3.9) $|Z^*(g_0)| \leqq n+1$ *if* $\tilde{x} \in Z_d(g_0)$ *and* $|Z^*(g_0)| \leqq n$ *if* $\tilde{x} \notin Z_d(g_0)$.

Proof. We first want to remember that $Z_d(g_0)$ is the set of the double zeros and $|Z^*(g_0)|$ the number of the zeros of g_0 counting multiplicities.

The first portion of this lemma follows directly from the proof of [11, Lemma 2.3]. To prove the second portion we assume that there is a $g_0 \in G$ having no zero intervals such that $\tilde{x} \in Z_d(g_0)$ and $|Z^*(g_0)| \geqq n+2$. By using [12, Theorem 3.1] it is easily verified that for each $i = 0, \cdots, s-1$, $i \neq p-1$, p, $G^{i,i+1}$ is even a Haar space with dimension $n_{i,i+1}$. Furthermore, since $G \in \tilde{Z}_n^2$, it follows that $|Z(g)| \leqq n_{i,i+1}$ for each nonzero $g \in G^{i,i+1}$, $i = p-1, p$.

Now we show by induction on the number of the knots that the preceding assumption does not hold. First, let $s = 2$. This means that there is only one knot $x_1 \in (a, b)$. Without loss of generality we may assume that $\tilde{x} < x_1$. The other case follows analogously. Then $|Z(g)| \leqq n_{12} - 1$ for each $g \in G^{12}$, since G^{12} is a Haar space, and $|Z(g)| \leqq n_{01}$ for each $g \in G^{01}$. This implies that $|Z(g)| \leqq n_{01} - 1$ for each $g \in G^{01}$ on $[a, x_1] \backslash \{\tilde{x}\}$. Then it follows from [5, Theorem 4.2, p. 23] that $|Z^*(g)| \leqq n_{12} - 1$ for each $g \in G^{12}$ and, using the arguments of that theorem, it is easily verified that $|Z^*(g)| \leqq n_{01} - 1$ for each $g \in G^{01}$ on $[a, x_1] \backslash \{\tilde{x}\}$. Now let $z_1, \cdots, z_n$ be zeros of g_0 on $[a, b] \backslash \{\tilde{x}\}$ where $z_i = z_{i+1}$, in case $z_i \in Z_d(g_0)$. Then the preceding arguments imply that $z_{n-n_{12}} < x_1 < z_{n_{01}+1}$. However, using the proof of [11, Lemma 2.5] and Lemma 3.8, it follows that $g_0 = 0$. But this contradicts our assumption on g_0. Let the statement be true for s knots.

Now given $s+1$ knots we choose again n zeros $z_1, \cdots, z_n$ of g_0 on $[a, b] \backslash \{\tilde{x}\}$ where $z_i = z_{i+1}$ if $z_i \in Z_d(g_0)$. Then each space G^{ij}, $j - i < s$, has at most s knots $x_i, \cdots, x_j$. Therefore, we may apply the assumption to these spaces and obtain the inequality $z_{n-n_{is}} < x_i < z_{n_{0i}+1}$, $i = 1, \cdots, s-1$ (here we have to note that at most $n_{ij} - 1$ of these zeros belong to $[x_i, x_j]$). As above it follows that $g_0 \equiv 0$. This contradicts our assumption that g_0 does not vanish on an interval. Thus we have shown that $|Z^*(g_0)| \leqq n+1$ if $\tilde{x} \in Z_d(g_0)$. The other statement of this lemma can be proved analogously.

In order to prove that all AEs of f coincide on an interval, we need the following lemma established by Nürnberger–Sommer [8]:

LEMMA 3.10. *Let $G \in W_n$ and $f \in C[a, b]$. If g_0, $\tilde{g}_0$ are two* AE*s of f, then at least one of the following statement holds*:

(3.10) *The function $g_0 - \tilde{g}_0$ has at least $n+1$ distinct zeros*;

(3.11) *The function $g_0 - \tilde{g}_0$ has at least $n+2$ zeros, counting multiplicities.*

LEMMA 3.11. *Let $G \in \tilde{Z}_n^2$ and $f \in C[a, b]$. If g_0, $\tilde{g}_0$ are two* AE*s of f, then $g_0 - \tilde{g}_0$ vanishes on an interval I which contains the point $\tilde{x}$.*

Proof. The preceding lemma shows that $|Z(g_0 - \tilde{g}_0)| \geqq n+1$ (resp., $|Z^*(g_0 - \tilde{g}_0)| \geqq n+2$). Then Lemma 3.9 implies that $g_0 - \tilde{g}_0$ must have a zero interval and the desired statement follows from Lemma 3.4.

Using this lemma we are able to show the following:

LEMMA 3.12. *Let $G \in \tilde{Z}_n^2$ and $f \in C[a, b]$. Then there is a subinterval I of $[a, b]$ such that $g_0 = \tilde{g}_0$ on I for all* AE*s g_0, $\tilde{g}_0$ of f.*

Proof. Remember that we always assume that $\tilde{x} \in [x_p, x_{p+1})$. Then it follows from Lemma 3.11 that we have yet to consider only the case that $\tilde{x} > a$. We distinguish two possibilities. First, $\tilde{x} \in (x_p, x_{p+1})$. By Lemma 3.11 any two AEs g_0, $\tilde{g}_0$ coincide on an interval I such that $\tilde{x} \in I$. Since $\tilde{x} \in (x_p, x_{p+1})$ and all zero intervals of the elements of G are knot intervals, it follows that $[x_p, x_{p+1}] \subset I$ and, therefore, all AEs of f coincide on $[x_p, x_{p+1}]$.

Second, $\tilde{x} = x_p$. If there is exactly one AE of f, then the statement is trivially satisfied and if there are exactly two AEs of f, then the statement follows directly from Lemma 3.11. Hence, there still remains the case that there are at least three AEs of f. Let $A(f) = \{g \in P_G(f) : g \text{ is an AE of } f\}$. Without loss of generality we may assume that $0 \in A(f)$. Furthermore, let $Z(A(f)) = \{x \in [a, b] : g(x) = 0 \text{ for each } g \in A(f)\}$. Then, since $\tilde{x} \in Z(G)$, it follows that $\tilde{x} \in Z(A(f))$. We now assume that $Z(A(f))$ does not contain any subinterval of $[a, b]$. Then using Lemma 3.11 there must exist three AEs 0, g_0, $\tilde{g}_0$ of f such that $g_0 \equiv 0$ on $[x_{p-1}, \tilde{x}]$ and $\tilde{g}_0 \equiv 0$ on $[\tilde{x}, x_{p+1}]$, however $g_0 \not\equiv 0$ on $[\tilde{x}, x_{p+1}]$ and $\tilde{g}_0 \not\equiv 0$ on $[x_{p-1}, \tilde{x}]$. Furthermore, Lemma 3.11 shows the existence of an interval $\tilde{I}$ on which $g_0 = \tilde{g}_0$. Without loss of generality we may assume that $\tilde{I} = [\tilde{x}, x_m]$. Then it follows that $g_0 = \tilde{g}_0 \equiv 0$ on $[\tilde{x}, x_{p+1}]$ and this contradicts our assumption that $Z(A(f))$ does not contain any subinterval of $[a, b]$. This concludes the proof.

Thus we have shown the interesting property that all AEs of f coincide on a subinterval of $[a, b]$. Unfortunately, for the set of all best approximations of f, an analogous result is not given in general, as we will show by the following example.

Example. Let $g_1, g_2 \in C[-1, 1]$ be defined by

$$g_1(x) := \begin{cases} 0 & \text{if } x \in [-1, 0], \\ x & \text{if } x \in [0, 1], \end{cases} \qquad g_2(x) := |x|.$$

Then $G = \operatorname{span}\{g_1, g_2\} \in W_2$. Furthermore $G \in \tilde{Z}_n^2$ and $x_1 = 0$ is the only knot of G. Then the function f, defined by $f(x) := 2|x| - 1$ has exactly one AE $g_0 \equiv 0$ for approximation from G and it is easily verified that $P_G(f) = \{a_1 g_1 + a_2 g_2 : 0 \leqq a_1 + a_2 \leqq 2, 0 \leqq a_2 \leqq 2\}$. But this implies that there does not exist any subinterval of $[-1, 1]$ on which all $g \in P_G(f)$ coincide.

Thus given any $G \in \tilde{Z}_n^2$ it is not true in general that all best approximations of f from G coincide on some interval. This shows an essential difference between the class $\tilde{Z}_n^2$ and the subclass $\tilde{V}_n$ of Z_n^1 which we have defined in [11], because for each $G \in \tilde{V}_n$ all best approximations of f from G always coincide on some interval. However, to construct a continuous selection for P_G, in case $G \in \tilde{Z}_n^2$, the statement of Lemma 3.12 is strongly enough.

LEMMA 3.13. *Let $G \in \tilde{Z}_n^2$. Then there exists a continuous selection.*

Proof. Let $f \in C[a, b]$ and $g_0 \in A(f)$ arbitrary where as above $A(f)$ is the set of all AEs of f. We furthermore may assume that $\tilde{x} < b$. Then $\tilde{x} \in [x_p, x_{p+1})$. The case $\tilde{x} = b$ follows analogously. We distinguish two possibilities. First, $\tilde{x} \in (x_p, x_{p+1})$. Then Lemma 3.11 shows that all $g \in A(f)$ coincide on $[x_p, x_{p+1}]$ and starting in $[x_p, x_{p+1}]$ with the function g_0 we construct a best approximation $s(f)$ of f step by step exactly as in the proof of Lemma 3.7 in [11]. To prove the continuity of this selection we can proceed again as in the proof of that lemma.

Second, $\tilde{x} = x_p$. This is more complicated than the first case. We construct again a best approximation $s(f)$ step by step analogously to the construction established in [11]:

(i) Local approximation: If $\dim \bar{G}_{0p} \geqq 1$, then we approximate $f - g_0$ in $[x_p, b]$ by $\bar{G}_{0p}$. Since by Lemma 3.2 $\bar{G}_{0p}$ is weak Chebyshev, Theorem 3.6 guarantees the existence of a local AE $g_1 \in P_{\bar{G}_{0p}}(f - g_0)$ (for definition of local AEs see [11]). Then $f - g_0 - g_1$ has at least $m_{0p} + 1$ local alternating extreme points in $[x_p, b]$. Furthermore:

$$\|f - g_0 - g_1\|_{[a,x_p]} = \|f - g_0\|_{[a,x_p]} \leqq \|f - g_0\|,$$

$$\|f - g_0 - g_1\|_{[x_p,b]} \leqq \|f - g_0 - 0\|_{[x_p,b]} \leqq \|f - g_0\|.$$

This implies that $g_0 + g_1 \in P_G(f)$. If $\bar{G}_{0p} = \{0\}$, then we define the function g_1 by $g_1 \equiv 0$.

(ii) Uniquenes of the local AEs on $[x_p, x_{p+1}]$: We will now show for approximation in $[x_p, b]$ that any two AEs $g_1, \bar{g}_1 \in P_{\bar{G}_{0p}}(f - g_0)$ are the same on $[x_p, x_{p+1}]$, i.e., $g_1 = \bar{g}_1$ on $[x_p, x_{p+1}]$. We assume to the contrary that $g_1 \neq \bar{g}_1$ on $[x_p, x_{p+1}]$. Since $g_1 = \bar{g}_1$ on $[a, x_p]$ and $G \in \tilde{Z}_n^2$, the function $g_1 - \bar{g}_1$ has no zero interval in $[x_p, b]$. Then Lemma 3.6 in [11] shows that $|Z^*(g_1 - \bar{g}_1)| \leqq m_{0p} + 1$ on $[x_p, b]$ and, since $G \in \tilde{Z}_n^2$, it follows that $|Z(g_1 - \bar{g}_1)| \leqq m_{0p}$ on $[x_p, b]$ (to use Lemma 3.6 in [11] we must note that $\tilde{x} = x_p \notin (x_p, b]$). But this contradicts the statements of Lemma 3.10. This implies that $g_1 = \bar{g}_1$ on $[x_p, x_{p+1}]$.

(iii) We show that if $\tilde{g}_0 \in A(f)$, $\tilde{g}_0 \neq g_0$, and $\tilde{g}_1 \in P_{\bar{G}_{0p}}(f - \tilde{g}_0)$ is a local AE for approximation in $[x_p, b]$, then $\tilde{g}_0 + \tilde{g}_1 = g_0 + g_1$ on $[x_p, x_{p+1}]$. Lemma 3.11 shows the existence of a subinterval I of $[a, b]$ such that $\tilde{x} \in I$ and $g_0 = \tilde{g}_0$ on I. We distinguish two cases:

If $[x_{p-1}, x_p] \not\subset I$, then $\tilde{x} \in I$ implies that $[x_p, x_{p+1}] \subset I$. First, we will show that $f - g_0$ and $f - \tilde{g}_0$ have at least $m_{0p} + 1$ alternating extreme points in $[x_p, b]$. We assume to the contrary that $f - g_0$ has at most m_{0p} such points in $[x_p, b]$. Since $g_0 \in A(f)$, the function $f - g_0$ has at least $n + 1$ alternating extreme points in $[a, b]$. Now it is easily verified that $n = m_{0p} + n_{0p}$ (remember that $n_{0p} = \dim G^{0p}$). Then it follows from our assumption that $f - g_0$ must have at least $n_{0p} + 1$ alternating extreme points in $[a, x_p - \varepsilon]$ for $\varepsilon > 0$ sufficiently small. Then, since $\tilde{x} = x_p$ and, therefore, each $x \in [a, x_p - \varepsilon]$ is nonvanishing, we can apply the proof of Lemma 2.5 in [11] to $f|_{[a,x_p-\varepsilon]}$ and to the space $\check{G} = G^{0p}|_{[a,x_p-\varepsilon]}$ and can show the existence of a subinterval $\check{I}$ of $[a, x_p - \varepsilon]$ on which all $g \in P_{\check{G}}(f|_{[a,x_p-\varepsilon]})$ coincide. Then it is easily verified that $g_0|_{[a,x_p-\varepsilon]} \in P_{\check{G}}(f|_{[a,x_p-\varepsilon]})$ and, furthermore, for each $g \in P_G(f)$ the function $g|_{[a,x_p-\varepsilon]}$ belongs to $P_{\check{G}}(f|_{[a,x_p-\varepsilon]})$. This implies that all $g \in P_G(f)$ coincide on $\check{I}$. Since $g_0, \tilde{g}_0 \in A(f) \subset P_G(f)$, it follows that $g_0 = \tilde{g}_0$ on $\check{I}$. However, this implies that $\check{I} \subset [a, x_{p-1}]$, because $g_0 \neq \tilde{g}_0$ on $[x_{p-1}, x_p]$. Then the function $g_0 - \tilde{g}_0$ has at least two separated zero intervals $\check{I}$ and $[x_p, x_{p+1}]$ which contradicts our assumption that $G \in \tilde{Z}_n^2$.

Thus we have shown that $f - g_0 = f - g_0 - 0$ and $f - \tilde{g}_0 = f - \tilde{g}_0 - 0$ have at least $m_{0p} + 1$ alternating extreme points in $[x_p, b]$. This implies that $0 \in \bar{G}_{0p}$ is an AE for approximation of $f - g_0$ and $f - \tilde{g}_0$ in $[x_p, b]$ by $\bar{G}_{0p}$. In (ii) we have shown that any two AEs $g_1, \bar{g}_1 \in P_{\bar{G}_{0p}}(f - g_0)$ are the same on $[x_p, x_{p+1}]$. Then, using this and the preceding

arguments, it immediately follows that $g_1 \equiv 0$ on $[x_p, x_{p+1}]$ and, analogously, $\tilde{g}_1 \equiv 0$ on $[x_p, x_{p+1}]$ where $\tilde{g}_1 \in P_{\bar{G}_{0p}}(f - \tilde{g}_0)$ is a local AE for approximation in $[x_p, b]$. This implies that $g_0 + g_1 = g_0 = \tilde{g}_0 = \tilde{g}_0 + \tilde{g}_1$ on $[x_p, x_{p+1}]$.

If $[x_{p-1}, x_p] \subset I$, then $g_0 - \tilde{g}_0 \in \bar{G}_{p-1,p}$. Then Lemma 3.2 implies that the function $\bar{g}_0$, defined by

$$\bar{g}_0 := \begin{cases} g_0 - \tilde{g}_0 & \text{on } [x_p, b], \\ 0 & \text{on } [a, x_p], \end{cases}$$

belongs to $\bar{G}_{0p}$. By definition, the functions $f - g_0 - g_1$ and $f - \tilde{g}_0 - \tilde{g}_1 = f - g_0 - (-g_0 + \tilde{g}_0 + \tilde{g}_1)$ have $m_{0p} + 1$ local alternating extreme points in $[x_p, b]$. Then the function $\bar{g}_1 := \tilde{g}_1 - \bar{g}_0 \in \bar{G}_{0p}$ is also a local AE of $f - g_0$ for approximation in $[x_p, b]$. Since according to (ii) all of these local AEs coincide on $[x_p, x_{p+1}]$, it follows that $\bar{g}_1 = \tilde{g}_1 - \bar{g}_0 = \tilde{g}_1 - g_0 + \tilde{g}_0$ on $[x_p, x_{p+1}]$. This proves that $\tilde{g}_0 + \tilde{g}_1 = g_0 + g_1$ on $[x_p, x_{p+1}]$.

(iv) This method will be continued in $[x_{p+1}, b]$ as follows: If $\dim \bar{G}_{0,p+1} \geqq 1$, then we approximate $f - g_0 - g_1$ in $[x_{p+1}, b]$ by $\bar{G}_{0,p+1}$. Then Theorem 3.6 guarantees the existence of a local AE $g_2 \in P_{\bar{G}_{0,p+1}}(f - g_0 - g_1)$. We can proceed exactly as in the proof of (ii) and (iii) above to show that all of these AEs coincide on $[x_{p+1}, x_{p+2}]$ and that $g_0 + g_1 + g_2 = \tilde{g}_0 + \tilde{g}_1 + \tilde{g}_2$ on $[x_p, x_{p+2}]$ for any choice of g_0, $\tilde{g}_0$, $g_0 + g_1$, $\tilde{g}_0 + \tilde{g}_1$. Furthermore, it follows that $g_0 + g_1 + g_2 \in P_G(f)$. If $\bar{G}_{0,p+1} = \{0\}$, then we define g_2 by $g_2 \equiv 0$.

(v) We continue this method up to the last interval $[x_{s-1}, b]$.

(vi) We use the same kind of arguments as in (i) to (v) for the interval $[a, x_p]$.

Thus we obtain a function

$$s(f) = g_{-p} + g_{-p+1} + \cdots + g_{-1} + g_0 + g_1 + \cdots + g_{s-p} \in P_G(f),$$

where for each $i \in \{1, \cdots, p\}$ $g_{-i} \in P_{\bar{G}_{p+1-i,s}}(f - g_0 - g_{-1} - \cdots - g_{-i+1})$ is a local AE for approximation in $[a, x_{p+1-i}]$ and for each $i \in \{1, \cdots, s-p\}$ $g_i \in P_{\bar{G}_{0,p+i-1}}(f - g_0 - g_1 - \cdots - g_{i-1})$ is a local AE for approximation in $[x_{p+i-1}, b]$. As we have shown above, this selection is independent of the choice of the AE g_0 and of the functions $g_{-1} + g_0, \cdots, g_{-p+1} + \cdots + g_0, g_0 + g_1, \cdots, g_0 + \cdots + g_{s-1-p}$. Now we can proceed as in the proof of Lemma 3.7 in [11] to prove the continuity of this selection. This concludes the proof.

Thus using Lemma 3.1 and Lemma 3.13 we get a complete characterization of those elements of Z_n^2 which admit a continuous selection.

THEOREM 3.14. *Let $G \in Z_n^2$. Then there exists a continuous selection if and only if $G \in \tilde{Z}_n^2$.*

If $G \in Z_n^2$, then, using the definition of $\tilde{Z}_n^2$, the proof of Theorem 1.8 follows directly from the preceding theorem. Thus the problem posed by Lazar–Morris–Wulbert is completely solved in the case that $X = [a, b]$.

Acknowledgment. I thank the referees for many helpful comments about the rewriting of this paper.

REFERENCES

[1] M. W. BARTELT, *Weak Chebyshev sets and splines*, J. Approx. Theory, 14 (1975), pp. 30–37.

[2] A. L. BROWN, *An extension to Mairhuber's Theorem: On metric projections and discontinuity of multivariate best uniform approximation*, preprint.

[3] R. B. HOLMES, *A Course on Optimization and Best Approximation*, Lecture Notes in Mathematics 257, Springer Verlag, Berlin–Heidelberg–New York, 1972.

[4] R. C. JONES AND L. A. KARLOVITZ, *Equioscillation under nonuniqueness in the approximation of continuous functions*, J. Approx. Theory, 3 (1970), pp. 138–145.

[5] S. KARLIN AND W. J. STUDDEN, *Tchebycheff Systems*, Interscience, New York 1966.
[6] A. J. LAZAR, P. D. MORRIS AND D. E. WULBERT, *Continuous selections for metric projections*, J. Funct. Anal., 3 (1969), pp. 193–216.
[7] G. NÜRNBERGER, *Nonexistence of continuous selections for the metric projection*, this Journal, 11 (1980), pp. 460–467.
[8] G. NÜRNBERGER AND M. SOMMER, *Weak Chebyshev subspaces and continuous selections for the metric projection*, Trans. Amer. Math. Soc., 238 (1978), pp. 129–138.
[9] ———, *Characterization of continuous selections of the metric projection for spline functions*, J. Approx. Theory, 22 (1978), pp. 320–330.
[10] M. SOMMER, *Nonexistence of continuous selections of the metric projection for a class of weak Chebyshev spaces*, Trans. Amer. Math. Soc., 260 (1980), pp. 403–409.
[11] ———, *Characterization of continuous selections for the metric projection for generalized splines*, this Journal, 11 (1980), pp. 23–40.
[12] ———, *Weak Chebyshev spaces and best L_1-approximation*, J. Approx. Theory, to appear.
[13] ———, *Continuous selections for metric projections*, in Quantitative Approximation, Bonn 1979, R. DeVore and K. Scherer, eds., Academic Press, New York, 1980, pp. 301–317.
[14] B. STOCKENBERG, *On the number of zeros of functions in a weak Tchebyshev space*, Math. Z., 156 (1977), pp. 49–57.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0009 $01.00/0

STRANGE EVALUATIONS OF HYPERGEOMETRIC SERIES*

IRA GESSEL† AND DENNIS STANTON‡

Abstract. Many evaluations of terminating hypergeometric series at arguments other than 1 are given. Some are equivalent to some unpublished work of Gosper, while others are new. In particular, two new evaluations of ${}_7F_6$'s with four parameters are stated. The main technique is a change of variables formula which is equivalent to the Lagrange inversion formula. A new proof of Whipple's transformation of a very well poised ${}_7F_6$ into a Saalschützian ${}_4F_3$ is a corollary.

1. Introduction. In 1836 Kümmer [14] compiled a list of quadratic transformations for ${}_2F_1$'s. Using Gauss's evaluation of ${}_2F_1(1)$, he was able to evaluate some ${}_2F_1$'s whose arguments were not 1. Goursat [12] used the same idea for third, fourth and sixth degree transformations of certain ${}_2F_1$'s. Recently, some evaluations for the generalized hypergeometric functions have been found which did not use this technique. G. Andrews found [1, eq. (1.12)]

$$(1.1)\qquad {}_3F_2\left[\begin{matrix}-n,\ n+3a,\ a\\ 3a/2,\ (3a+1)/2\end{matrix}\ \middle|\ \tfrac{3}{4}\right]=\begin{cases}0, & n\not\equiv 0 \pmod 3,\\ \dfrac{(3N)!\,(a+1)_N}{N!\,(3a+1)_{3N}}, & n=3N.\end{cases}$$

In a letter to R. Askey, R. Gosper gave a list of such mysterious-looking evaluations, for example,

$$(1.2)\qquad {}_5F_4\left[\begin{matrix}2a,\ 2b,\ 1-2b,\ 1+2a/3,\ -n\\ a-b+1,\ a+b+\frac{1}{2},\ 2a/3,\ 1+2a+2n\end{matrix}\ \middle|\ \tfrac{1}{4}\right]=\frac{(a+\frac{1}{2})_n(a+1)_n}{(a+b+\frac{1}{2})_n(a-b+1)_n},$$

$$(1.3)\qquad {}_5F_4\left[\begin{matrix}a,\ b,\ a+\frac{1}{2}-b,\ 1+2a/3,\ -n\\ 2a+1-2b,\ 2b,\ 2a/3,\ 1+a+n/2\end{matrix}\ \middle|\ 4\right]=\begin{cases}0, & n \text{ odd},\\ \dfrac{(2N)!\,(a+1)_N 2^{-2N}}{N!\,(a-b+1)_N(b+\frac{1}{2})_N}, & n=2N,\end{cases}$$

$$(1.4)\qquad {}_3F_2\left[\begin{matrix}\frac{1}{2}+3a,\ \frac{1}{2}-a,\ -n\\ \frac{1}{2},\ -3n\end{matrix}\ \middle|\ \tfrac{3}{4}\right]=\frac{(\frac{1}{2}-a)_n(\frac{1}{2}+a)_n}{(\frac{1}{3})_n(\frac{2}{3})_n},$$

$$(1.5)\qquad {}_3F_2\left[\begin{matrix}1+3a,\ 1-3a,\ -n\\ \frac{3}{2},\ -1-3n\end{matrix}\ \middle|\ \tfrac{3}{4}\right]=\frac{(1+a)_n(1-a)_n}{(\frac{2}{3})_n(\frac{4}{3})_n},$$

$$(1.6)\qquad {}_3F_2\left[\begin{matrix}2a,\ 1-a,\ -n\\ 2a+2,\ -a-\frac{1}{2}-3n/2\end{matrix}\ \middle|\ 1\right]=\frac{((n+3)/2)_n(n+1)(2a+1)}{(1+\frac{1}{2}(n+2a+1))_n(2a+n+1)}.$$

(Even though the argument in (1.6) is 1, it does not fit into the known ${}_3F_2(1)$ summation theorems.) Gosper also gave many ${}_2F_1$ evaluations with one free parameter.

* Received by the editors December 6, 1980, and in revised form May 5, 1981.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. The work of this author was supported in part by the Office of Naval Research under contract N00014-76-C-0366.

‡ School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

Although Gosper did not assume that the series terminate, we are concerned here only with the terminating cases. Thus in (1.4) and (1.5) the ${}_3F_2$'s terminate after $n+1$ terms. The ambiguity in these series can be removed by reversing them, but we do not do so.

Besides showing (1.1), (1.4) and (1.5), we extend (1.2), (1.3) and (1.6) to

$$(1.7)\qquad {}_7F_6\left[\begin{matrix}2a, 2b, 1-2b, 1+2a/3, a+d+n+\frac{1}{2}, a-d, -n\\ a-b+1, a+b+\frac{1}{2}, 2a/3, -2d-2n, 2d+1, 1+2a+2n\end{matrix}\;\middle|\;1\right] = \frac{(2a+1)_{2n}(b+d+\frac{1}{2})_n(d-b+1)_n}{(2d+1)_{2n}(a+b+\frac{1}{2})_n(a-b+1)_n},$$

$$(1.8)\qquad {}_7F_6\left[\begin{matrix}a, b, a+\frac{1}{2}-b, 1+2a/3, 1-2d, 2a+2d+n, -n\\ 2a+1-2b, 2b, 2a/3, a+d+\frac{1}{2}, 1-d-n/2, 1+a+n/2\end{matrix}\;\middle|\;1\right] = \begin{cases} 0, & n \text{ odd},\\ \dfrac{(b+d)_N(d-b+a+\frac{1}{2})_N(2N)!\,(a+1)_N}{(b+\frac{1}{2})_N(a+d+\frac{1}{2})_N(d)_N N!\,(a-b+1)_N}\,2^{-2N}, & n=2N,\end{cases}$$

$$(1.9)\qquad {}_3F_2\left[\begin{matrix}-sb+s+1, b-1, -n\\ b+1, s(-n-b)-n\end{matrix}\;\middle|\;1\right] = \frac{(1+s+sn)_n b(n+1)}{(1+s(b+n))_n(b+n)}.$$

The ${}_7F_6$'s in (1.7) and (1.8) have a curious property similar to the well-poised condition. Instead of just adding a numerator parameter to a denominator parameter, one of them is doubled first. In (1.7) the first three pairs give a constant $2a+2$, whereas the last three pairs give $2a+1$. The middle pair $(1+2a/3, 2a/3)$ can give either $2a+1$ or $2a+2$. These series are also Saalschützian (balanced).

We organize this paper in the following way. In § 2 we introduce our two methods of evaluation and discuss their relevance to some early work of Bailey. The first method—factorization of polynomials—gives (1.1) and many ${}_2F_1$ evaluations in § 3. The second method is a change of variables formula of Jacobi, and is explained in § 4. We give many examples of this method (including (1.2)–(1.9)) in § 5.

2. Methods of evaluation. G. Andrews and R. Gosper used unconventional techniques to find (1.1)–(1.6). Gosper (see [11]) used a splitting function to tear apart individual terms of the series and produce new series. He used the symbolic manipulation program MACSYMA to find the right splitting functions. Andrews [1] inverted the "Bailey transform" [1, Lemma 3], which was equivalent to finding the inverse of a certain lower triangular matrix. Our methods are very simple: equating coefficients in power series expansions. They are also variations on Bailey's first published paper [2]. In [2] he considered the following problem: Given integers λ and μ with $\lambda > 0$ and $\mu \neq 0$, for what values of $\alpha_1, \cdots, \alpha_r, \rho_1, \cdots, \rho_s, A$ and p can the coefficient of x^n in

$$(2.1)\qquad f(x) = (1-x)^p\,{}_rF_s\left[\begin{matrix}\alpha_1, \cdots, a_r\\ \rho_1, \cdots, \rho_s\end{matrix}\;\middle|\;\frac{Ax^\lambda}{(1-x)^\mu}\right]$$

be written as a quotient of gamma functions? (We use the notation $\langle f(x)|x^n\rangle$ for this coefficient.) Clearly $\langle f(x)|x^n\rangle$ can be expressed as a terminating hypergeometric series. Bailey relied upon the known summation theorems to find such examples. For example,

Saalschütz's theorem gives [2, (4.12)]

$$(2.2)\qquad \begin{aligned}(1-x)^{-a}\,{}_3F_2\left[\begin{matrix}a/2,(1+a)/2,1+a-b-c\\1+a-b,1+a-c\end{matrix}\,\middle|\,\frac{-4x}{(1-x)^2}\right]\\ ={}_3F_2\left[\begin{matrix}a,b,c\\1+a-b,1+a-c\end{matrix}\,\middle|\,x\right],\end{aligned}$$

Whipple's [15] quadratic transformation for a well poised ${}_3F_2$. Bailey listed the transformations he obtained in this way. Later, Bailey [3] multiplied these transformations by $(1-x)^B$ and equated coefficients of x^n. The results were transformations for series (mostly evaluated at 1) which implied the known summation theorems. However Bailey did not consider

(I) special factorizations of $f(x)$

or

(II) products of two transformations.

We use these two methods.

It is possible that Bailey considered (I) much too special to be of interest. However he knew that (II) could lead to interesting results. For example, by multiplying Euler's transformation by itself [4, p. 56], he derived the Saalschützian ${}_4F_3$ transformation. To use this method on (2.2), an expansion in terms of $-4x/(1-x)^2$ instead of x is needed. Otherwise double sums would occur in $\langle f_1(x)f_2(x)|x^n\rangle$. We accomplish this new expansion by using a change of variables formula equivalent to the Lagrange inversion formula.

Motivated by (2.2), Bailey did prove Whipple's transformation [16, (5.4)]. In fact he showed how to boost evaluations into transformations of series of higher order [16, § 4]. It is a special case of the Bailey transform. We would like to thank the referee for pointing this out to us.

3. Factorizations. We try to factor, or otherwise manipulate, the function $f(x)$ given in (2.1). In a sense this is trivial, so that we do not expect deep results. Rather than give the long list of transformations obtained in this way, we give only the most interesting examples.

For the bulk of this section we concentrate on $r=1$ and $s=0$ in (2.1). First, to show (1.1), we take $\lambda=1$ and $\mu=2$ so that

$$(3.1)\qquad \begin{aligned}f(x)&=(1-x)^{p+2a}((1-x)^2-Ax)^{-a},\\ \langle f(x)|x^n\rangle&=\frac{(-p)_n}{n!}\,{}_3F_2\left[\begin{matrix}-n,n-p,a\\-p/2,(1-p)/2\end{matrix}\,\middle|\,\frac{-A}{4}\right].\end{aligned}$$

For $A=-3$, $f(x)=(1-x)^{p+2a}(1+x+x^2)^{-a}=(1-x)^{p+3a}(1-x^3)^{-a}$, so

$$(3.2)\qquad \begin{aligned}\frac{(-p-3a)_n}{n!}\,{}_4F_3\left[\begin{matrix}-n/3,(1-n)/3,(2-n)/3,a\\(1+p+3a-n)/3,(2+p+3a-n)/3,(3+p+3a-n)/3\end{matrix}\,\middle|\,1\right]\\ =\frac{(-p)_n}{n!}\,{}_3F_2\left[\begin{matrix}-n,n-p,a\\-p/2,(1-p)/2\end{matrix}\,\middle|\,\tfrac{3}{4}\right].\end{aligned}$$

If $p=-3a$, (3.2) implies (1.1). We can also derive other evaluations from (3.2), for example, $p=1-3a$ gives

$$\frac{(3a-1)_n}{n!}\,{}_3F_2\left[\begin{matrix}-n, n+3a-1, a\\ (3a-1)/2, (3a)/2\end{matrix}\;\middle|\;\tfrac{3}{4}\right]=\begin{cases}0, & n=3k+2,\\ \dfrac{-(a)_k}{k!}, & n=3k+1,\\ \dfrac{(a)_k}{k!}, & n=3k.\end{cases}\tag{3.3}$$

We can also try other values of A. If $A=-4$, $f(x)=(1-x)^{p+2a}(1+x)^{-2a}$ and $\langle f(x)|x^n\rangle={}_2F_1(-1)$. From (3.1) we would get the transformation [4, p. 33] taking any ${}_2F_1(-1)$ into a ${}_3F_2(1)$. By writing $f(x)=(1-x^2)^{-2a}(1-x)^{p+4a}=(1-x^2)^{p+2a}(1+x)^{-p-4a}$, we could give two other ${}_3F_2(1)$ forms.

For $A=-2$, $f(x)=(1-x)^{p+2a}(1+x^2)^{-a}$, which gives a transformation of a ${}_3F_2(\frac{1}{2})$ into a ${}_3F_2(-1)$. No new evaluations come from this case or by writing $f(x)=(1-x)^{p+3a}(1+x)^a(1-x^4)^{-a}$.

For $A=-1$, $f(x)=(1-x)^{p+2a}(1-x+x^2)^{-a}=(1-x)^{p+2a}(1+x)^a(1+x^3)^{-a}$. This leads to transformations, if $p=-2a$ or $p=-a$, but no new interesting evaluations.

We give a few more examples.

If $r=1$, $s=0$, $\lambda=1$, $\mu=3$ and $A=-\frac{27}{4}$, then

$$f(x)=(1-x)^{p+3a}((1-x)^3+\tfrac{27}{4}x)^{-a}=(1-x)^{p+3a}(1+2x)^{-2a}(1-x/4)^{-a},$$

and

$$\langle f(x)|x^n\rangle=\frac{(-p)_n}{n!}\,{}_4F_3\left[\begin{matrix}-n, (n-p)/2, (n-p+1)/2, a\\ -p/3, (1-p)/3, (2-p)/3\end{matrix}\;\middle|\;1\right].\tag{3.4}$$

If $p=-3a$, we also have

$$\langle f(x)|x^n\rangle=\frac{(a)_n}{n!}\,4^{-n}\,{}_2F_1\left[\begin{matrix}-n, 2a\\ -a-n+1\end{matrix}\;\middle|\;-8\right],\tag{3.5}$$

so

$$\frac{(3a)_n}{n!}\,{}_3F_2\left[\begin{matrix}-n, (n+3a)/2, (n+3a+1)/2\\ a+\frac{1}{3}, a+\frac{2}{3}\end{matrix}\;\middle|\;1\right]=\frac{(a)_n}{n!}\,4^{-n}\,{}_2F_1\left[\begin{matrix}-n, 2a\\ -a-n+1\end{matrix}\;\middle|\;-8\right].\tag{3.6}$$

If $a=-n-\frac{1}{3}$, Vandermonde's theorem (${}_2F_1(1)$) implies

$${}_2F_1\left[\begin{matrix}-n, -2n-\frac{2}{3}\\ \frac{4}{3}\end{matrix}\;\middle|\;-8\right]=(-27)^n\frac{(\frac{5}{6})_n}{(\frac{3}{2})_n}.\tag{3.7}$$

Gosper gives a ${}_2F_1(\frac{8}{9})$ evaluation equivalent to (3.7). We can also use Vandermonde's theorem and a limit to evaluate the ${}_2F_1(-8)$ if $a=-n+\frac{1}{3}$, $-n+\frac{2}{3}$ or $-n+\frac{4}{3}$:

$${}_2F_1\left[\begin{matrix}-n, -2n+\frac{2}{3}\\ \frac{2}{3}\end{matrix}\;\middle|\;-8\right]=\frac{(-3n+1)_n4^n}{(\frac{1}{3})_n(\frac{2}{3})_n}[(\tfrac{1}{6})_n+\tfrac{1}{2}(\tfrac{1}{2})_n],\qquad n\geqq 1,\tag{3.8}$$

$${}_2F_1\left[\begin{matrix}-n, -2n+\frac{4}{3}\\ \frac{1}{3}\end{matrix}\;\middle|\;-8\right]=\frac{(-3n+2)_n4^n}{(-\frac{1}{3})_n(\frac{1}{3})_n}[(-\tfrac{1}{6})_n+\tfrac{1}{2}(-\tfrac{1}{2})_n],\qquad n\geqq 1,\tag{3.9}$$

$$(3.10)\quad {}_2F_1\left[\begin{matrix}-n, -2n+\frac{8}{3} \\ -\frac{1}{3}\end{matrix} \mid -8\right] = \frac{(-3n+4)_n 4^n}{(-\frac{1}{3})_n(-\frac{2}{3})_n}\left[(-\tfrac{5}{6})_n + (n/3-\tfrac{3}{4})(-\tfrac{1}{2})_{n-1}\right], \qquad n \geqq 2.$$

As a related example, put $\lambda = 2$, $\mu = 3$, $A = \frac{27}{4}$ and $p = -3a$ so that $f(x) = (1-4x)^{-a}(1+x/2)^{-2a}$. Instead of (3.6), we have for $\langle f(x)|x^n\rangle$

$$(3.11)\quad \frac{(3a)_n}{n!}\,{}_3F_2\left[\begin{matrix}-n/2, (1-n)/2, n+3a \\ a+\frac{1}{3}, a+\frac{2}{3}\end{matrix} \mid 1\right] = (-2)^{-n}\frac{(2a)_n}{n!}\,{}_2F_1\left[\begin{matrix}-n, a \\ 1-2a-n\end{matrix} \mid -8\right].$$

Putting $a = -n/2+\frac{1}{6}$, $-n/2+\frac{1}{3}$, $-n/2-\frac{1}{6}$, $-n/2-\frac{1}{3}$, $-n/2-\frac{2}{3}$ and applying Vandermonde's theorem in (3.11), we have

$$(3.12)\quad {}_2F_1\left[\begin{matrix}-n, -n/2+\frac{1}{6} \\ \frac{2}{3}\end{matrix} \mid -8\right] = 2\cdot 3^{(3n-1)/2}\cos(n\pi/2+\pi/6),$$

$$(3.13)\quad {}_2F_1\left[\begin{matrix}-n, -n/2+\frac{1}{3} \\ \frac{1}{3}\end{matrix} \mid -8\right] = \begin{cases} 3^{(3n-1)/2}(-1)^{(n+1)/2}, & n \text{ odd}, \\ 3^{3n/2-1}(-1)^{n/2}\left[1+2\dfrac{(\frac{1}{2})_{n/2}}{(\frac{1}{6})_{n/2}}\right], & n \geqq 2, \text{ even}, \end{cases}$$

$$(3.14)\quad {}_2F_1\left[\begin{matrix}-n, -n/2-\frac{1}{6} \\ \frac{4}{3}\end{matrix} \mid -8\right] = (-1)^{n/2}\frac{3^{3n/2}}{n+1}, \qquad n \text{ even},$$

$$(3.15)\quad {}_2F_1\left[\begin{matrix}-n, -n/2-\frac{1}{3} \\ \frac{5}{3}\end{matrix} \mid -8\right] = (-1)^{(n+1)/2}\frac{3^{(3n+1)/2}}{n+2}, \qquad n \text{ odd},$$

$$(3.16)\quad {}_2F_1\left[\begin{matrix}-n, -n/2-\frac{2}{3} \\ \frac{7}{3}\end{matrix} \mid -8\right] = (-1)^{(n+1)/2}\frac{5\cdot 3^{(3n+1)/2}}{(n+2)(n+4)}, \qquad n \text{ odd}.$$

Because the series in (3.12)–(3.16) are terminating, these ${}_2F_1(-8)$ evaluations can be stated as ${}_2F_1$ evaluations at $-\frac{1}{8}, \frac{8}{9}, \frac{9}{8}, \frac{1}{9}$ or 9. Special cases of (1.1) or (3.3) give ${}_2F_1(\frac{3}{4})$ evaluations. The arguments equivalent to $\frac{3}{4}$ are $\frac{4}{3}$, 4, $\frac{1}{4}$, $-\frac{1}{3}$ and -3. We give three transformations for these arguments which are similar in spirit to (3.6) and (3.11).

Put $\lambda = 1$, $\mu = -2$, $A = \frac{27}{4}$ and $p = 0$ so that

$$f(x) = (1-\tfrac{27}{4}x(1-x)^2)^{-a} = (1-3x)^{-2a}(1-\tfrac{3}{4}x)^{-a}.$$

Equating $\langle f(x)|x^n\rangle$ we have

$$(3.17)\qquad 9^n\,{}_3F_2\left[\begin{matrix}-2n/3, (1-2n)/3, (2-2n)/3 \\ -n+\frac{1}{2}, 1-a-n\end{matrix} \mid 1\right] = {}_2F_1\left[\begin{matrix}-n, 2a \\ 1-a-n\end{matrix} \mid 4\right].$$

For $\lambda = 3$, $\mu = 2$, $A = \frac{4}{27}$ and $p = -2a$,

$$(3.18)\qquad \begin{gathered} f(x) = ((1-x)^2 - \tfrac{4}{27}x^3)^{-a} = (1-x/3)^{-2a}(1-4x/3)^{-a}, \\ 3^n\,{}_3F_2\left[\begin{matrix}-n/3, (1-n)/3, (2-n)/3 \\ a+\frac{1}{2}, 1-2a-n\end{matrix} \mid 1\right] = {}_2F_1\left[\begin{matrix}-n, a \\ 1-2a-n\end{matrix} \mid 4\right]. \end{gathered}$$

Finally, if $\lambda = 3$, $\mu = 3$, $A = -1$ and $p = -3a$,

$$f(x) = ((1-x)^3 + x^3)^{-a} = (1-3x+3x^2)^{-a},$$

$$(3.19)\quad \frac{(3a)_n}{n!}\,{}_3F_2\left[\begin{matrix}-n/3, (1-n)/3, (2-n)/3 \\ a+\frac{1}{3}, a+\frac{2}{3}\end{matrix} \mid 1\right] = \frac{(a)_n}{n!}3^n\,{}_2F_1\left[\begin{matrix}-n/2, (1-n)/2 \\ 1-a-n\end{matrix} \mid \tfrac{4}{3}\right].$$

From (3.17), (3.18) and (3.19) many ${}_2F_1$ evaluations can be given.

We have not given the transformations which involve two series, neither of which is evaluated at 1. It is conceivable that the above evaluations could be used with these transformations to find even stranger evaluations. Also, we were unable to use Dixon's, Watson's, Whipple's or Saalschütz's theorem to evaluate the ${}_3F_2(1)$'s and find new evaluations.

Instead of putting $r=1$ and $s=0$ in (2.1) for $f(x)$, we could try using higher values of r and s with special values of $\alpha_1, \cdots, \alpha_r, \rho_1, \cdots, \rho_s$. Bailey had already considered this [3, p. 500] by expanding $(1-x)^a$ in powers of $x/(1-x)^{r+1}$. For example,

$$ {}_2F_1\left[\begin{matrix} a, a+\frac{1}{2} \\ 2a+1 \end{matrix} \,\middle|\, \frac{-4x}{(1-x)^2}\right] = (1-x)^{2a} $$

leads to

$$ (3.20)\qquad \frac{(-p)_n}{n!}\,{}_4F_3\left[\begin{matrix} a, a+\frac{1}{2}, -n, n-p \\ 2a+1, -p/2, (1-p)/2 \end{matrix} \,\middle|\, 1\right] = \frac{(-2a-p)_n}{n!} \qquad [3, (3.42)] $$

and

$$ (1-2x)\,{}_2F_1\left[\begin{matrix} a, a+\frac{1}{2} \\ 2a \end{matrix} \,\middle|\, 4x(1-x)\right] = (1-x)^{1-2a} $$

leads to

$$ (3.21)\qquad \frac{(1-p)_n(2+p)_n}{n!(1+p)_n}\,{}_4F_3\left[\begin{matrix} a, a+\frac{1}{2}, -n, 1+p \\ 2a, (2+p-n)/2, (3+p-n)/2 \end{matrix} \,\middle|\, 1\right] = \frac{(-1+2a-p)_n}{n!} \qquad [3, (3.43)]. $$

Andrews [1, (4.4)] proved (3.21) by the Bailey transform.

A related evaluation that Bailey did not explicitly state is

$$ (3.22)\qquad \frac{(-p)_n}{n!}\,{}_4F_3\left[\begin{matrix} a, -a, -n/2, (1-n)/2 \\ \frac{1}{2}, 1+p-n, -p \end{matrix} \,\middle|\, 1\right] = \frac{1}{2}\left[\frac{(-a-p)_n}{n!} + \frac{(a-p)_n}{n!}\right]. $$

It follows from

$$ (3.23)\qquad {}_2F_1\left[\begin{matrix} a, -a \\ \frac{1}{2} \end{matrix} \,\middle|\, \frac{x^2}{4(x-1)}\right] = \tfrac{1}{2}[(1-x)^a + (1-x)^{-a}] $$

or by putting $\beta = -\alpha$ in [3, (4.11)]. In § 5 we indicate a generalization of this procedure using (5.12).

4. Change of variables. If $f(x) = \sum_{j=-m}^{\infty} a_j x^j$ is a (formal) Laurent series, its residue is defined to be $\operatorname{Res} f(x) = a_{-1}$. Thus $\langle f(x) | x^n \rangle = \operatorname{Res} f(x)/x^{n+1}$. Note that $\operatorname{Res} f'(x)$ is always zero.

THEOREM 1. *Let* $G(x) = \sum_{j=-m}^{\infty} a_j x^i$ *and* $h(x) = \sum_{i=1}^{\infty} b_i x^i$, *where* $b_1 \neq 0$. *Then*

$$ \operatorname{Res} G(h(x))h'(x) = \operatorname{Res} G(x). $$

Proof. Since both sides are linear in G, it is sufficient to take $G(x) = x^m$. If $m \neq -1$, then

$$ \operatorname{Res} h^m(x)h'(x) = \frac{1}{m+1} \operatorname{Res} [h^{m+1}(x)]' = 0. $$

For $m=-1$, let $h(x)=b_1xf(x)$. Then

$$\operatorname{Res}\frac{h'(x)}{h(x)}=\operatorname{Res}\left[\frac{1}{x}+\frac{f'(x)}{f(x)}\right]=1+\operatorname{Res}\left[\log f(x)\right]=1.$$

(Since $f(0)=1$, $\log f(x)$ has a power series expansion.) □

This theorem and its proof are due to Jacobi [13]. In this important but little-known paper, Jacobi also gave the multivariable generalization, which is equivalent to Good's [10] multivariable Lagrange inversion formula.

The various forms of the Lagrange inversion formula are easily proved from the theorem. For example, we have the following proposition.

PROPOSITION 1. *Let f and g be power series with $f(0)=0$ and $g(0)\neq 0$ satisfying*

$$f(x)=xg(f(x)).$$

then for n, $k>0$,

$$(*)\qquad \langle f^k(x)|x^n\rangle=\frac{k}{n}\langle g^n(x)|x^{n-k}\rangle.$$

Proof. We have

$$\begin{aligned}\langle g^n(x)|x^{n-k}\rangle&=\operatorname{Res}g^n(x)/x^{n-k+1}=\operatorname{Res}g^n(f(x))f'(x)f^{n-k+1}(x)\\&=\operatorname{Res}f^{k-1}(x)f'(x)/x^n=\frac{1}{k}\left\langle\frac{d}{dx}f^k(x)|x^{n-1}\right\rangle\\&=\frac{n}{k}\langle f^k(x)|x^n\rangle.\end{aligned}$$

□

Bromwich [5, pp. 159–160], who referred to Jacobi [13], gave a similar proof of (*), but did not explicitly state the change of variables formula.

We give an application of the change of variables formula (Theorem 1) to identities.

THEOREM 2. *Suppose $A(x)$, $B(x)$, $C(x)$ and $D(x)$ are power series whose coefficients of x^k are A_k, B_k, C_k and D_k, respectively. Suppose*

$$(1-x)^{-\alpha}A(x/(1-x)^{\beta+1})=B(x),$$

$$(1+\beta x)(1-x)^{-\gamma}C(x/(1-x)^{\beta+1})=D(x).$$

If $n(\beta+1)=1-\alpha-\gamma$, then

$$\sum_{k=0}^{n}B_kD_{n-k}=\sum_{k=0}^{n}A_kC_{n-k}.$$

Proof. Clearly we have

$$B(x)D(x)/x^{n+1}=(1+\beta x)(1-x)^{-\alpha-\gamma}A(x/(1-x)^{\beta+1})C(x/(1-x)^{\beta+1})/x^{n+1}.$$

If $h(x)=x/(1-x)^{\beta+1}$, $(h'(x)=(1+\beta x)/(1-x)^{\beta+2})$ and $n(\beta+1)=1-\alpha-\gamma$, we obtain

$$B(x)D(x)/x^{n+1}=A(h(x))C(h(x))h'(x)/[h(x)]^{n+1}.$$

The residue of the left-hand side is $\sum_{k=0}^{n}B_kD_{n-k}$. Theorem 1 implies that the residue of the right-hand side is $\sum_{k=0}^{n}A_kC_{n-k}$. □

In § 5 we will be taking the transformations that Bailey had found for our $A(x)$ and $B(x)$ in Theorem 2. In this case $h(x)=x/(1-x)^{\mu/\lambda}$ and $\beta+1=\mu/\lambda$. However any transformation of this type gives a transformation between a $C(x)$ and a $D(x)$, as in Theorem 2. We give a general form of this companion transformation.

PROPOSITION 2. *Suppose* $A(x)=\sum_{k=0}^{\infty}A_kx^k$, $B(x)=\sum_{k=0}^{\infty}B_kx^k$ *and*

$$(1-x)^{-\alpha}A(x/(1-x)^{\beta+1})=B(x).$$

Then

$$(1+\beta x)(1-x)^{-\alpha-1}C(x/(1-x)^{\beta+1})=D(x),$$

where

$$C_k=A_k(\alpha+(\beta+1)k) \quad \text{and} \quad D_k=B_k(\alpha+(\beta+1)k).$$

Proof. Apply $\alpha I+(\beta+1)x\,d/dx$ to $(1-x)^{-\alpha}A(x/(1-x)^{\beta+1})=B(x)$. □

According to this proposition, the companion transformation of (2.2) is

$$(4.1)\qquad \begin{aligned}(1+x)(1-x)^{-a-1}{}_3F_2\left[\begin{matrix}(1+a)/2,\,(2+a)/2,\,1+a-b-c\\ 1+a-b,\,1+a-c\end{matrix}\;\middle|\;\frac{-4x}{(1-x)^2}\right]\\ ={}_4F_3\left[\begin{matrix}a,\,1+a/2,\,b,\,c\\ a/2,\,1+a-b,\,1+a-c\end{matrix}\;\middle|\;x\right].\end{aligned}$$

We can check that $(x/(1-x)^2)'=(1+x)/(1-x)^3$, so that the necessary $1+x$ factor is correct in (4.1). Bailey had given (4.1) previously by equating coefficients [4, p. 97].

5. Change of variables—examples. As our first example we prove Whipple's transformation [4, (4.3(4))], which takes a very well poised ${}_7F_6(1)$ into a Saalschützian (balanced) ${}_4F_3(1)$.

From the well poised ${}_3F_2$ quadratic transformation (2.2) we can put

$$(5.1)\qquad \begin{aligned}A_k&=(a/2)_k((1+a)/2)_k(1+a-b-c)_k(-4)^k/k!(1+a-b)_k(1+a-c)_k,\\ B_k&=(a)_k(b)_k(c)_k/k!(1+a-b)_k(1+a-c)_k,\end{aligned}$$

$\beta=1$ and $\alpha=a$ in Theorem 2. The companion transformation (4.1) (with d, e and f replacing a, b and c) allows

$$\begin{aligned}C_k&=((1+d)/2)_k((2+d)/2)_k(1+d-e-f)_k(-4)^k/k!(1+d-e)_k(1+d-f)_k,\\ D_k&=(d)_k(1+d/2)_k(e)_k(f)_k/k!(d/2)_k(1+d-e)_k(1+d-f)_k\end{aligned}$$

and $\gamma=d+1$ in Theorem 2. Thus, putting $2n=-a-d$, Theorem 2 implies

$$(5.2)\qquad \begin{aligned}&{}_7F_6\left[\begin{matrix}a,\,1+a/2,\,b,\,c,\,e+a+n,\,f+a+n,\,-n\\ a/2,\,1+a-b,\,1+a-c,\,1-e-n,\,1-f-n,\,a+n+1\end{matrix}\;\middle|\;1\right]\\ &\quad=\frac{(a+1)_n(a+e+f+n)_n}{(e)_n(f)_n}\,{}_4F_3\left[\begin{matrix}1+a-b-c,\,e+a+n,\,f+a+n,\,-n\\ 1+a-b,\,1+a-c,\,e+f+a+n\end{matrix}\;\middle|\;1\right].\end{aligned}$$

This is Whipple's theorem.

It is interesting to ask which theorem we would obtain by inverting only one of (2.2) or (4.1). If we set $c=1+a-b$, then (2.2) reduces to the binomial theorem and we are using only (4.1). Then (5.2) becomes the very well poised ${}_5F_4(1)$ evaluation. Similarly, for (2.2) we specialize (5.2) by $f=1+d-e=1-a-2n-e$. Again the very well poised ${}_5F_4(1)$ is the result. This is just multiplying (2.2) by $(1+x)(1-x)^{-d-1}/x^{n+1}$ and using the change of variables formula.

Next we complete the same steps on Bailey's [2, (4.05)] cubic transformation of a ${}_3F_2$;

$$(1-x)^{-3a}\,{}_3F_2\left[\begin{matrix}a,\ a+\frac{1}{3},\ a+\frac{2}{3}\\ (3a+b+1)/2,\ (3a-b+2)/2\end{matrix}\ \Big|\ \frac{-27x}{4(1-x)^3}\right] = {}_3F_2\left[\begin{matrix}3a,\ b,\ 1-b\\ (3a+b+1)/2,\ (3a-b+2)/2\end{matrix}\ \Big|\ \frac{x}{4}\right], \tag{5.3}$$

whose companion is

$$(1+2x)(1-x)^{-1-3c}\,{}_3F_2\left[\begin{matrix}c+\frac{1}{3},\ c+\frac{2}{3},\ c+1\\ (3c+d+1)/2,\ (3c-d+2)/2\end{matrix}\ \Big|\ \frac{-27x}{4(1-x)^3}\right] = {}_4F_3\left[\begin{matrix}3c,\ 1+c,\ d,\ 1-d\\ c,\ (3c+d+1)/2,\ (3c-d+2)/2\end{matrix}\ \Big|\ \frac{x}{4}\right]. \tag{5.4}$$

This time putting $c=-a-n$ gives

$${}_7F_6\left[\begin{matrix}3a,\ b,\ 1-b,\ 1+a,\ (1+n+3a-d)/2,\ (n+3a+d)/2,\ -n\\ (3a+b+1)/2,\ (3a-b+2)/2,\ 1+3a+2n,\ a,\ 1-n-d,\ d-n\end{matrix}\ \Big|\ 1\right] = \frac{(3a+1)_{2n}(-1)^n}{(d)_n(1-d)_n}\,{}_3F_2\left[\begin{matrix}-n,\ (1+n+3a-d)/2,\ (n+3a+d)/2\\ (3a+b+1)/2,\ (3a-b+2)/2\end{matrix}\ \Big|\ 1\right]. \tag{5.5}$$

The ${}_3F_2$ in (5.5) can be evaluated by Saalschütz's theorem. After relabeling the parameters, the ${}_7F_6(1)$ evaluation that we have is (1.7). Gosper's special case (1.2) occurs if we consider only one transformation, either (5.3) or (5.4).

Bailey's other cubic transformation [2, (4.06)]

$$(1-x)^{-3a}\,{}_3F_2\left[\begin{matrix}a,\ a+\frac{1}{3},\ a+\frac{2}{3}\\ b+\frac{1}{2},\ 3a-b+1\end{matrix}\ \Big|\ \frac{27x^2}{4(1-x)^3}\right] = {}_3F_2\left[\begin{matrix}3a,\ b,\ 3a-b+\frac{1}{2}\\ 2b,\ 6a-2b+1\end{matrix}\ \Big|\ 4x\right] \tag{5.6}$$

can be paired with

$$(1+x/2)(1-x)^{-3c-1}\,{}_3F_2\left[\begin{matrix}c+\frac{1}{3},\ c+\frac{2}{3},\ c+1\\ d+\frac{1}{2},\ 3c-d+1\end{matrix}\ \Big|\ \frac{27x^2}{4(1-x)^3}\right] = {}_4F_3\left[\begin{matrix}3c,\ 1+2c,\ d,\ 3c-d+\frac{1}{2}\\ 2c,\ 2d,\ 6c-2d+1\end{matrix}\ \Big|\ 4x\right] \tag{5.7}$$

to obtain (1.8). In the change of variables $h(x)=x(1-x)^{-3/2}$, so that the resulting sum is zero for odd powers of $x(1-x)^{-3/2}$. Again Gosper's (1.3) comes by inverting either transformation alone.

It is not necessary to pair up a transformation with its companion. We may use any two transformations to which the same change of variables can be applied. We could combine (4.1) with Gauss's quadratic transformation

$$(1-x)^{-2e}\,{}_2F_1\left[\begin{matrix}e,\ d\\ 2d\end{matrix}\ \Big|\ \frac{-4x}{(1-x)^2}\right] = {}_2F_1\left[\begin{matrix}e,\ e-d+\frac{1}{2}\\ d+\frac{1}{2}\end{matrix}\ \Big|\ x^2\right]. \tag{5.8}$$

If $a=-2e-2n$, we have

$$
\begin{aligned}
&{}_9F_8\left[\begin{matrix} e, e/2+1, e-d+\frac{1}{2}, -n/2, (1-n)/2, (b+n)/2+e, \\ e/2, d+\frac{1}{2}, e+1+n/2, e+\frac{1}{2}+n/2, (2-b-n)/2, \end{matrix}\right. \\
&\qquad\left.\begin{matrix} (b+n+1)/2+e, (c+n)/2+e, (c+n+1)/2+e \\ (1-b-n)/2, (2-c-n)/2, (1-c-n)/2 \end{matrix}\;\middle|\; 1\right] \\
&\qquad=\frac{(1+2e)_n(1-2e-2n-b-c)_n}{(b)_n(c)_n}\,{}_4F_3\left[\begin{matrix} d, -n, b+2e+n, c+2e+n \\ 2d, 1/2+e, b+c+2e+n \end{matrix}\;\middle|\; 1\right].
\end{aligned}
\tag{5.9}
$$

The ${}_9F_8$ in (5.9) is very well poised. Evidently there are many such transformations.

We can also find transformations for series not evaluated at 1. For example, we could pair up (5.4) with any transformation which takes a power series at $-27x/4(1-x)^3$ to a power series at Ax, $A\neq\frac{1}{4}$. One such transformation is

$$
{}_3F_2\left[\begin{matrix} a/3, (a+1)/3, (a+2)/3 \\ (a+1)/2, (a+2)/2 \end{matrix}\;\middle|\; \frac{-27x}{4(1-x)^3}\right]=(1-x)^a,
\tag{5.10}
$$

which is easily proved by Lagrange inversion. We now give another such transformation which has 3 parameters. A special case of it will be a nice generalization of (5.10).

We begin with the identity

$$
\sum_{k=0}^{n}(-1)^k\frac{(-a-(s-1)k)_{n-k}}{(n-k)!}\,\frac{(-a-sk)_k b}{k!(b+k)}=\frac{(sb-a)_n}{(b+1)_n},
\tag{5.11}
$$

which is just a partial fraction expansion. It is clear that (5.11) is equivalent to

$$
(1-x)^a\sum_k\frac{(-a-sk)_k b}{k!(b+k)}[-x(1-x)^{s-1}]^k={}_2F_1\left[\begin{matrix} 1, sb-a \\ b+1 \end{matrix}\;\middle|\; x\right].
\tag{5.12}
$$

For $s=-2$, we have a 2-parameter extension of (5.10)

$$
(1-x)^{-a}\,{}_4F_3\left[\begin{matrix} a/3, (a+1)/3, (a+2)/3, b \\ a/2, (a+1)/2, b+1 \end{matrix}\;\middle|\; \frac{-27x}{4(1-x)^3}\right]={}_2F_1\left[\begin{matrix} 1, a-2b \\ b+1 \end{matrix}\;\middle|\; x\right].
\tag{5.13}
$$

So (5.4) and (5.13) give ($a=-3c-3n$, $e=-b-n$),

$$
\begin{aligned}
&\frac{(3c+1)_n(e+1)_n}{n!(1+3c-2e)_n}\,{}_4F_3\left[\begin{matrix} -n, e, (1+3c+n)/2, (3c+n)/2 \\ e+1, (1+3c+d)/2, (1+3c-d)/2 \end{matrix}\;\middle|\; 1\right] \\
&\qquad={}_6F_5\left[\begin{matrix} 3c, d, 1-d, 1+c, e, -n \\ (3c-d+2)/2, (3c+d+1)/2, c, 1+3c-2e, -n \end{matrix}\;\middle|\; \tfrac{1}{4}\right].
\end{aligned}
\tag{5.14}
$$

We have written the $-n$ parameters in the ${}_6F_5$ only to indicate that the sum terminates after $n+1$ terms. If $e=-n$, (5.14) reduces to (1.2). The ${}_4F_3$ is again Saalschützian (balanced).

We can derive (1.9) from (5.12). The companion to (5.12) is

$$
\begin{aligned}
&(1-sx)(1-x)^{a-1}\sum_k\frac{(-a-sk)_k}{k!}\cdot\frac{b}{b+k}\cdot\frac{a+(s-1)k}{a}[-x(1-x)^{s-1}]^k \\
&\qquad={}_3F_2\left[\begin{matrix} 1, sb-a, 1+a/(s-1) \\ b+1, a/(s-1) \end{matrix}\;\middle|\; x\right].
\end{aligned}
\tag{5.15}
$$

Inverting (5.15) gives

$$(5.16)\qquad {}_4F_3\left[\begin{matrix}1, 1+a/(s-1), sb-a, -n\\ a/(s-1), b+1, 1-a-sn\end{matrix}\;\middle|\; 1\right]=\frac{b(a+sn)}{a(b+n)}.$$

Fields and Luke [7] had previously given (5.16). If $a=s-1$ in (5.15), the ${}_3F_2$ becomes a ${}_2F_1$ to which we can apply Euler's transformation

$$(5.17)\qquad \begin{aligned}(1-sx)(1-x)^{s-2}\sum_k \frac{(1-s-sk)_k}{k!}\frac{b}{b+k}\cdot(k+1)\cdot[-x(1-x)^{s-1}]^k\\ =(1-x)^{b-sb+s-2}\,{}_2F_1\left[\begin{matrix}b-sb+s, b-1\\ b+1\end{matrix}\;\middle|\; x\right].\end{aligned}$$

Inverting (5.17) gives (1.9), after s has been replaced by $s+1$.

We could obtain still more evaluations from (5.12) by multiplying by a power of $(1-x)$ and equating coefficients. The right-hand side yields a ${}_3F_2(1)$, which can be evaluated in special cases. We leave it to the interested reader to write out these identities.

We have not yet shown (1.4) and (1.5). These two evaluations follow from

$$(5.18)\qquad {}_2F_1\left[\begin{matrix}a, -a\\ \frac{1}{2}\end{matrix}\;\middle|\; \tfrac{27}{4}x(1-x)^2\right]={}_2F_1\left[\begin{matrix}3a, -3a\\ \frac{1}{2}\end{matrix}\;\middle|\; \frac{3x}{4}\right].$$

We could not find (5.18) stated as such in the literature. However, it is a version of the triple angle formula for cosines. If we multiply (5.18) by $x^{-n-2}(1-x)^{-3n-2}(1-3x)$ and change variables, the result is (1.5). Applying Euler's transformation to (5.18) yields

$$(5.19)\qquad (1-3x)\,{}_2F_1\left[\begin{matrix}\frac{1}{2}-a, \frac{1}{2}+a\\ \frac{1}{2}\end{matrix}\;\middle|\; \frac{27x(1-x)^2}{4}\right]={}_2F_1\left[\begin{matrix}\frac{1}{2}-3a, \frac{1}{2}+3a\\ \frac{1}{2}\end{matrix}\;\middle|\; \frac{3x}{4}\right].$$

Inverting (5.19) gives (1.4).

In an analogous way we can use a transformation for algebraic functions given by Bailey [2, (4.21)],

$$(5.20)\qquad (1-x)^{-2a}\,{}_2F_1\left[\begin{matrix}a, a+\frac{1}{2}\\ 2a+1\end{matrix}\;\middle|\; \frac{4x^3}{27(1-x)^2}\right]={}_2F_1\left[\begin{matrix}3a, 3a+\frac{1}{2}\\ 6a+1\end{matrix}\;\middle|\; \frac{4x}{3}\right]$$

to obtain

$$(5.21)\qquad {}_3F_2\left[\begin{matrix}3a+\frac{1}{2}, 3a+1, -n\\ 6a+1, -n/3+2a+1\end{matrix}\;\middle|\; \tfrac{4}{3}\right]=\begin{cases}0, & n\neq 3N,\\ \dfrac{(\frac{1}{3})_N(\frac{2}{3})_N}{(1+2a)_N(-2a)_N}, & n=3N.\end{cases}$$

The companion formula to (5.20) also gives (5.21), which is equivalent to (1.1).

Many ${}_2F_1$ evaluations with one parameter also can be obtained from the change of variables formula. We give a few examples which involve factorizations.

First take

$$1+27x/(1-x)^3=(1+2x)^2(1-x/4)/(1-x)^3,$$

so that

$$\begin{aligned}(1+2x)[1+27x/(1-x)^3]^a(1-x)^{3n-1}x^{-n-1}\\ =(1+2x)^{2a+1}(1-x/4)^a(1-x)^{3n-3a-1}x^{-n-1}.\end{aligned}$$

We find the residue of both sides, using the change of variables $h(x)=x/(1-x)^3$ on the left. The residue of the right-hand side is a single sum if either $a=-\frac{1}{2}$ or $a=n-\frac{1}{3}$. The results are

$$(5.22)\qquad {}_2F_1\left[\begin{matrix}-n, \frac{1}{2}\\ 2n+\frac{3}{2}\end{matrix}\;\middle|\;\frac{1}{4}\right]=\frac{(\frac{1}{2})_n}{(2n+\frac{3}{2})_n}\left(\frac{27}{4}\right)^n,$$

$$(5.23)\qquad {}_2F_1\left[\begin{matrix}-n, -\frac{1}{3}-2n\\ \frac{2}{3}\end{matrix}\;\middle|\;-8\right]=(-27)^n.$$

In fact, the $a=-\frac{1}{2}$ case explains why many special evaluations have a numerator parameter of $\frac{1}{2}$. (Another reason is the integral representation in terms of cosines and sines.) Also (5.23) is related to (3.12): the terminating parameter of (5.23) is the nonterminating parameter of (3.12).

A more exotic factorization is

$$1+27x^2(1-x)/64(1-9x/8)^2=(1-3x/4)^3/(1-9x/8)^2.$$

With the change of variables $h(x)=x(1-x)^{1/2}(1-9x/8)^{-1}$, the above process admits either $a=-2/3$ or $a=(n-1)/2$, yielding

$$(5.24)\qquad {}_2F_1\left[\begin{matrix}-n, n/2+1\\ \frac{4}{3}\end{matrix}\;\middle|\;\frac{8}{9}\right]=\begin{cases}0, & n \text{ odd},\\ \dfrac{(\frac{1}{2})_N}{(\frac{7}{6})_N}(-3)^{-N}, & n=2N \text{ even},\end{cases}$$

$$(5.25)\qquad {}_2F_1\left[\begin{matrix}-n, \frac{1}{2}\\ (n+3)/2\end{matrix}\;\middle|\;4\right]=\begin{cases}0, & n \text{ odd},\\ \dfrac{(\frac{1}{2})_N(\frac{3}{2})_N}{(\frac{5}{6})_N(\frac{7}{6})_N}, & n=2N \text{ even}.\end{cases}$$

This type of factorization can be found by looking at the arguments of the cubic transformations for a ${}_2F_1$ given in the Bateman project [6, p. 114]. The above example comes from (40).

We give a final example of change of variables. Gosper has given a number of ${}_4F_3(\frac{3}{128})$ evaluations with one parameter. In every case there is a numerator parameter which is one greater than a denominator parameter, for example $\frac{23}{15}-a$ and $\frac{8}{15}-a$. We give a ${}_4F_3(-27)$ evaluation which is similar in spirit.

We take the factorization

$$(5.26)\qquad 1+8x/3(1-x)^3=(1+3x^2)(1-x/3)/(1-x)^3.$$

If we raise (5.26) to the ath power, multiply by $(1+2x)(1-x)^{3n-1}x^{-n-1}$, change variables with $h(x)=x/(1-x)^3$ and put $a=n-\frac{1}{3}$, the result is

$$(5.27)\qquad {}_4F_3\left[\begin{matrix}\frac{1}{3}-n, -n/2, (1-n)/2, \frac{22}{21}-3n/7\\ \frac{5}{6}, \frac{4}{3}, \frac{1}{21}-3n/7\end{matrix}\;\middle|\;-27\right]=\frac{(-8)^n}{1-9n}.$$

The pair $(\frac{22}{21}-3n/7, \frac{1}{21}-3n/7)$ comes about because of the $1+2x$ term needed in the change of variables. Two coefficients are then added together to find the residue.

6. Open problems. Of the evaluations given on Gosper's list, only one evaluation with more than one parameter remains unproven. It is

$$
\begin{aligned}
&{}_7F_6\left[\begin{matrix} a+\frac{1}{2}, a, b, 1-b, c, (2a+1)/3-c, a/2+1 \\ \frac{1}{2}, (2a-b+3)/3, (2a+b+2)/3, 3c, 2a+1-3c, a/2 \end{matrix} \;\middle|\; 1\right] \\
&\quad = \frac{2\Gamma((2a-b+3)/3)\Gamma((2a+b+2)/3)\Gamma(c+\frac{1}{3})\Gamma(c+\frac{2}{3})}{\sqrt{3}\,\Gamma((2a+2)/3)\Gamma(2a/3+1)\Gamma(c-(b+2)/3)\Gamma(c+(b+1)/3)} \\
&\qquad \cdot \frac{\Gamma((2a+2)/3-c)\Gamma(2a/3-c+1)\sin(b+1)\pi/3}{\Gamma((2a-b)/3-c+1)\Gamma((2a+b+2)/3-c)}.
\end{aligned} \tag{6.1}
$$

An attractive terminating version of (6.1) is with $c=-n$,

$$
{}_7F_6[c=-n] = \frac{((2a+2)/3)_n(2a/3+1)_n((1+b)/3)_n((2-b)/3)_n}{((2a-b)/3+1)_n((2a+b+2)/3)_n(\frac{2}{3})_n(\frac{1}{3})_n}. \tag{6.2}
$$

Again we assume that the ${}_7F_6$ terminates after $n+1$ terms. This ${}_7F_6$ has a very strange property similar to the ${}_7F_6$'s in (1.7) and (1.8). Instead of doubling, tripling is needed! This prescription does not hold for $a+\frac{1}{2}$, a, $\frac{1}{2}$ parameters. It would be reasonable to hope that a ${}_5F_4$ version of (6.2) arises by inverting a quartic transformation of a ${}_3F_2$. A nice ${}_5F_4$ version is

$$
{}_5F_4\left[\begin{matrix} a+\frac{1}{2}, a, -n, (2a+1)/3+n, a/2+1 \\ \frac{1}{2}, -3n, 2a+1+3n, a/2 \end{matrix} \;\middle|\; 9\right] = \frac{((2a+2)/3)_n(2a/3+1)_n}{(\frac{1}{3})_n(\frac{2}{3})_n}. \tag{6.3}
$$

An iteration of (2.2) is

$$
(1-x)^{-a}\,{}_3F_2\left[\begin{matrix} a/4, a/4+\frac{1}{2}, a/4+\frac{1}{4} \\ \frac{1}{2}, 3a/4+\frac{3}{4} \end{matrix} \;\middle|\; \frac{-16x(1+x)^2}{(1-x)^4}\right] = {}_3F_2\left[\begin{matrix} a, a+\frac{1}{2}, a/4+\frac{1}{4} \\ \frac{1}{2}, 3a/4+\frac{3}{4} \end{matrix} \;\middle|\; -x\right]. \tag{6.4}
$$

Because of the a, $a+\frac{1}{2}$, $\frac{1}{2}$ terms and the tripling (note that the cubic transformation (5.6) involved doubling), we could hope that (6.4) is somehow related to (6.3). We were unable to show this.

If $a\to\infty$ in (6.2), we recover (1.4). Because n occurs in four places in (6.3), two Lagrange inversions (possibly involving (5.18)) might be indicated.

Gosper gives two ${}_2F_1$'s which we could not show:

$$
{}_2F_1\left[\begin{matrix} -n, -n+\frac{1}{4} \\ 2n+\frac{5}{4} \end{matrix} \;\middle|\; \frac{1}{9}\right] = \frac{(\frac{5}{4})_{2n}}{(\frac{2}{3})_n(\frac{13}{12})_n}\left(\frac{2^6}{3^5}\right)^n, \tag{6.5}
$$

$$
{}_2F_1\left[\begin{matrix} -n, -n+\frac{1}{4} \\ 2n+\frac{9}{4} \end{matrix} \;\middle|\; \frac{1}{9}\right] = \frac{(\frac{9}{4})_{2n}}{(\frac{4}{3})_n(\frac{17}{12})_n}\left(\frac{2^6}{3^5}\right)^n. \tag{6.6}
$$

With Lagrange inversion it is easy to translate (6.5) and (6.6) into expansions about certain ${}_2F_1$'s with no parameters, but we could not verify these.

Very little is known about q-analogues. The factorization method does not work, but something replaces it. Andrews [1, (4.7)] has a q-analogue of (1.1) which is equivalent to

$$
{}_3\phi_2\left[\begin{matrix} a, wa, w^2a \\ a^3x, q/x \end{matrix} \;\middle|\; q;q\right] = \frac{(a^6x^3;q^3)_\infty(x;q)_\infty}{(a^3x^3;q^3)_\infty(a^3x;q)_\infty}, \qquad w^3=1. \tag{6.7}
$$

We are considering (6.7) as a formal power series in x. It is the q-analogue of $(1-x)^{-3a}(1+3x/(1-x)^2]^{-a}=(1-x^3)^{-a}$.

Carlitz has found a q-analogue of (2.2), and it is not hard to write down a q-analogue of (4.1). Since the q-analogue of Whipple's theorem is known (Watson's theorem), a q-analogue of the change of variables could exist. This, in turn, may be related to the q-Lagrange inversion of Garsia [8] or Gessel [9].

Acknowledgment. We would like to thank R. Wm. Gosper, Jr., for access to his list of evaluations.

REFERENCES

[1] G. ANDREWS, *Connection coefficient problems and partitions*, AMS Proc. Symposia in Pure Mathematics 34, D. Ray-Chaudhuri, ed., American Mathematical Society, Providence, RI, 1979, pp. 1–24.

[2] W. BAILEY, *Products of generalized hypergeometric series*, Proc. London Math. Soc., 28 (1928), pp. 242–254.

[3] ———, *Transformations of generalized hypergeometric series*, Proc. London Math. Soc., 29 (1929), pp. 495–502.

[4] ———, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, 1935.

[5] T. J. I. BROMWICH, *An Introduction to the Theory of Infinite Series*, 2nd ed., Macmillan, London, 1942, pp. 159–160.

[6] A. ERDÉLYI, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[7] J. FIELDS AND Y. LUKE, *Problem* 68-20, *a hypergeometric series*, SIAM Rev., 12 (1970), pp. 586–588.

[8] A. GARSIA, *A q-analogue of the Lagrange inversion formula*, preprint.

[9] I. GESSEL, *A noncommutative generalization and q-analog of the Lagrange inversion formula*, Trans. Amer. Math. Soc., 257 (1980), pp. 455–482.

[10] I. GOOD, *Generalizations to several variables of Lagrange's expansion, with applications to stochastic processes*, Proc. Camb. Phil. Soc., 56 (1960), pp. 367–380.

[11] R. W. GOSPER, JR., *A calculus of series rearrangements*, Algorithms and Complexity, J. F. Traub, ed., Academic Press, New York, 1976, pp. 121–151.

[12] E. GOURSAT, *Sur l'équation différentielle linéaire qui admet pour intégral la série hypergéometrique*, Ann. Sci. École Norm. Sup., 10 (1881), pp. 3–142.

[13] C. JACOBI, *De resolution aequationum per series infinitas*, J. für die Reine und Angewandte Math., 6 (1830), 257–286.

[14] E. KÜMMER, *Über die hypergeometrische Reihe*

$$1+\frac{\alpha\cdot\beta}{1\cdot\gamma}x+\frac{\alpha(\alpha+1)\beta(\beta+1)}{1\cdot 2\cdot\gamma(\gamma+1)}x^2+\frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1\cdot 2\cdot 3\cdot\gamma(\gamma+1)(\gamma+2)}x^3+\cdots,$$

J. für die Reine und Angewandte Math., 15 (1836), pp. 39–83.

[15] F. WHIPPLE, *Some transformations of generalized hypergeometric series*, Proc. London Math. Soc., 26 (1927), pp. 257–272.

[16] W. BAILEY, *Some identities involving generalized hypergeometric series*, Proc. London Math. Soc., 29 (1929), pp. 503–516.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0010 $01.00/0

THE RECURSION FORMULAS FOR ORTHOGONAL POLYNOMIALS IN n VARIABLES*

M. A. KOWALSKI†

Abstract. In this paper we introduce a very promising matrix-vector notation for orthogonal polynomials in n variables. This enables us to prove some new properties of the related recursion formulas.

1. Introduction. Let $P_0, P_1, \cdots$ be a sequence of orthogonal polynomials in one real variable. We assume that P_k is of kth degree. It is well known that for $\{P_k\}$ the following three-term recurrence relation holds:

$$P_{k+1}(x) = (d_k x + e_k)P_k(x) + f_k P_{k-1}(x), \qquad k = 0, 1, \cdots, \tag{1.1}$$

where $P_{-1}(x) \equiv 0$.

The applications of this formula are very important both in theory and in computational practice. For instance, the availability of coefficients d_k, e_k, f_k, $k = 0, 1, \cdots$, allows us to simplify many numerical algorithms.

This paper deals with the corresponding relations for orthogonal polynomials in n variables (see [1], [2]). We propose to use a notation closely related to that of the scalar case. Its simplicity gives more insight into the subject and leads to some new results. Since the recursion derived here is not unique for $n > 1$, we point out a way of standardizing it. Some illustrative examples are also provided.

2. Preliminaries. Let Π_n^∞ be a vector space of all polynomials with real coefficients in n real variables and let Π_n^k be its subspace of polynomials whose total degree in n variables is not larger than k.

A real-valued bilinear functional $\langle \cdot, \cdot \rangle$ is said to be a quasi-inner product in Π_n^∞ if and only if

$$\langle t, u \rangle = \langle 1, tu \rangle \quad \text{for any } t, u \in \Pi_n^\infty. \tag{2.1}$$

For instance (2.1) is valid for any inner product expressible in the form

$$\langle f, g \rangle = \int_{R^n} f(x)g(x)w(x)\,dx,$$

where w is a weight function on R^n.

A sequence S of polynomials is called orthogonal (orthonormal) if and only if for any two elements t, u from S the alternative

$$\langle t, u \rangle = 0 \qquad \text{if } t \neq u,$$

$$\langle t, u \rangle \neq 0 \quad (=1) \quad \text{if } t = u$$

holds.

The sequence of monomials $x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}$, $i_1 + i_2 + \cdots + i_n = 0, \cdots, k$ is a basis of Π_n^k; therefore (see [3])

$$\dim \Pi_n^k = \sum_{i=0}^{k} r_n^i = \binom{n+k}{k},$$

where $r_n^k = \binom{n+k-1}{k}$ is the number of monomials in this basis whose degree is equal to k.

* Received by the editors May 27, 1980, and in revised form May 1, 1981.
† Institute of Informatics, University of Warsaw P.K.iN. 8p. 850, 00-901 Warsaw, Poland.

Let a basis in Π_n^∞ be denoted by $\{P_i^k\}_{k=0,i=1}^{\infty,r_n^k}$, each polynomial is of the degree indicated by its upper subscript. We define

$$\vec{P}_k(x)=[P_1^k(x), P_2^k(x), \cdots, P_{r_n^k}^k(x)]^T,$$

$$\overrightarrow{xP}_k(x)=[x_1\vec{P}_k(x)^T|x_2\vec{P}_k(x)^T|\cdots|x_n\vec{P}_k(x)^T]^T,$$

where $x=(x_1, x_2, \cdots, x_n)\in R^n$, $k=0, 1, \cdots$.

The following useful result holds (see [1]).

THEOREM 1. *If* $\{P_i^k\}_{k=0,i=1}^{\infty,r_n^k}$, $\{Q_i^k\}_{k=0,i=1}^{\infty,r_n^k}$ *are two orthonormal bases in* Π_n^∞, *then for each* k *there exists a* $r_n^k\times r_n^k$ *matrix* M_k *satisfying*

1) $$\vec{P}_k=M_k\vec{Q}_k,$$

2) $$M_kM_k^T=I \qquad (\textit{unit matrix}).$$

Conversely, if $\{Q_i^k\}_{k=0,i=1}^{\infty,r_n^k}$ *is an orthonormal basis in* Π_n^∞ *and* 1), 2) *hold, then* $\{P_i^k\}_{k=0,i=1}^{\infty,r_n^k}$ *is an orthonormal basis too.*

Proof. For specified k, i the polynomial P_i^k can be expressed in the form

$$P_i^k=\sum_{l=0}^{k}\sum_{j=1}^{r_n^l} m_{ij}^{(l)}Q_i^l.$$

Since P_i^k is orthogonal to every polynomial of degree lower than k, $m_{ij}^{(l)}$ must vanish for $l<k$. Hence $M_k=(m_{ij}^{(k)})$.

Our assumptions imply that

$$\sum_{t=1}^{r_n^k} m_{it}^{(k)}m_{jt}^{(k)}=\sum_{s,t=1}^{r_n^k} m_{it}^{(k)}m_{js}^{(k)}\delta_{ts}=\sum_{s,t=1}^{r_n^k} m_{it}^{(k)}m_{is}^{(k)}\langle Q_t^k, Q_s^k\rangle=\langle P_i^k, P_j^k\rangle=\delta_{ij}.$$

(Here, δ_{ij} is Kronecker's delta.)

Observe that orthonormality of polynomials P is equivalent to identity 2). This proves the theorem. □

Let the notation $A: i\times j$ mean that A is an $i\times j$ matrix.

3. The recursion formulas.

THEOREM 2. *If* $\{P_i^k\}_{k=0,i=1}^{\infty,r_n^k}$ *is an orthogonal basis in* Π_n^∞ *then there exist matrices* $A_k: nr_n^k\times r_n^{k+1}$, $B_k: nr_n^k\times r_n^k$, $C_k: nr_n^k\times r_n^{k-1}$, $D_k: r_n^{k+1}\times nr_n^k$, $E_k: r_n^{k+1}\times r_n^k$, $F_k: r_n^{k+1}\times r_n^{k-1}$ *such that*

(3.1) $$\overrightarrow{xP}_k=A_k\vec{P}_{k+1}+B_k\vec{P}_k+C_k\vec{P}_{k-1},$$

(3.2) $$\vec{P}_{k+1}=D_k\overrightarrow{xP}_k+E_k\vec{P}_k+F_k\vec{P}_{k-1},$$

$k=0, 1, \cdots$. ($\vec{P}_{-1}=0$, $C_0=F_0=0$.) *Furthermore, if relations* (3.1) *and* (3.2) *hold then for* $k=0, 1, \cdots$

(3.3) $$\operatorname{rank} A_k=r_n^{k+1},$$

(3.4) $$\operatorname{rank} D_k=r_n^{k+1},$$

(3.5) $$D_kA_k=I,$$

(3.6) $$E_k=-D_kB_k, \qquad F_k=-D_kC_k.$$

Proof. For specified i, j, k the polynomial $x_iP_j^k$, being an element of Π_n^{k+1}, can be written in the form

$$x_iP_j^k=\sum_{s=0}^{k+1}\sum_{t=1}^{r_n^s}\alpha_{st}P_t^s,$$

where $\alpha_{st} = \langle x_i P_j^k, P_t^s\rangle / \langle P_t^s, P_t^s\rangle$. By (2.1) we get

$$\alpha_{st} = \langle P_j^k, x_i P_t^s\rangle / \langle P_t^s, P_t^s\rangle.$$

Since $x_i P_t^s$ is a polynomial of degree $s+1$ and P_j^k is orthogonal to every polynomial of degree $<k$, α_{st} vanishes for $s \leqq k-1$. We see that the relation (3.1) is strictly an expansion of each $x_i P_j^k$ with respect to the basis $\{P_j^i\}_{i=0,j=1}^{\infty, r_n^k}$ described in the matrix form; so the matrices A_k, B_k, C_k are unique.

Let $k \geqq 0$ be an arbitrary integer. We now prove that rank $A_k = r_n^{k+1}$. Assume to the contrary, that any r_n^{k+1} rows of A_k are linearly dependent. From (3.1) it follows that any set of r_n^{k+1} polynomial coefficients of $\overrightarrow{xP}_k$ is linearly dependent with respect to Π_n^k. On the other hand, it is easy to verify that

$$\Pi_n^{k+1} = \operatorname{lin}(\Pi_n^k; x_i P_j^k, i = 1, \cdots, n, j = 1, \cdots, r_n^k), \tag{3.7}$$

where the right-hand side denotes the vector space spanned by the polynomials $x_i P_j^k$, $i = 1, \cdots, n$, $j = 1, \cdots, r_n^k$, and all elements of Π_n^k. Thus we obtain the contradiction.

By (3.3), columns of the matrix A_k are linearly independent. Hence we can choose as D_k any matrix D such that $DA_k = I$. Then we put $E_k = -D_k B_k$, $F_k = -D_k C_k$ and (3.2) holds.

Substituting $\overrightarrow{xP}_k$ in (3.2) for the right-hand side of (3.1) we get

$$(I - D_k A_k)\vec{P}_{k+1} - (E_k + D_k B_k)\vec{P}_k - (F_k + D_k C_k)\vec{P}_{k-1} = 0,$$

from which (3.5) and (3.6) follow.

By virtue of Silvester's theorem (see [3]) the identities (3.3) and (3.5) imply (3.4). This completes the proof. □

Since the recursion formula (3.2) yields an algorithm for calculating $\vec{P}_k(x)$ we give special attention to it. From our considerations it follows immediately that to determine all possible recursion formulas (3.2) it is enough to specify for each k the class of all matrices D_k such that $D_k A_k = I$. In the case $n = 1$, i.e., for orthogonal polynomials in one variable, the formulas (3.1) and (3.2) are equivalent and can be reduced to (1.1). We now discuss the more interesting case $n > 1$. Due to (3.3) and [4, Thm. (4.5)], A_k can be represented in the form

$$A_k = U_k \begin{bmatrix} \Sigma_k \\ 0 \end{bmatrix} V_k^T, \tag{3.8}$$

where $\Sigma_k = \operatorname{diag}(\delta_1^{(k)}, \cdots, \delta_{r_n^k}^{(k)})$, $\delta_1^{(k)} \geqq \delta_2^{(k)} \geqq \cdots \geqq \delta_{r_n^k}^{k} > 0$ are the singular values of A_k and U_k: $nr_n^k \times nr_n^k$, V_k: $r_n^{k+1} \times r_n^{k+1}$ are orthogonal matrices. This yields the following corollary.

COROLLARY. *Let X_k be any $r_n^{k+1} \times (nr_n^k - r_n^{k+1})$ matrix and V_k, Σ_k, U_k come from* (3.8). *Then the factorization*

$$D_k = V_k[\Sigma_k^{-1} | X_k] U_k^T$$

states all possible choices of matrix D_k. □

Hence we can standardize (3.2) specifying X_k for each k. In particular putting $X_k = 0$, $k = 0, 1, \cdots$, we obtain $D_k = A_k^+$—the generalized inverse of A_k.

Let us consider any two orthonormal bases of the space Π_n^∞, say $\{P_i^k\}_{k=0,i=1}^{\infty, r_n^k}$ and $\{Q_i^k\}_{k=0,i=1}^{\infty, r_n^k}$. We denote the matrix coefficients of formulas (3.1) introduced by these bases as A_k, B_k, C_k and A'_k, B'_k, C'_k respectively.

THEOREM 3. *The singular values of matrices A_k and A'_k, B_k and B'_k, C_k and C'_k respectively are identical.*

Proof. In view of Theorem 1 there exist orthogonal matrices M_k such that $\vec{P}_k = M_k\vec{Q}_k$, $k = 0, 1, \cdots$. Using (3.1) we now get

$$\overrightarrow{xM_k\vec{Q}_k} = A_kM_{k+1}\vec{Q}_{k+1} + B_kM_k\vec{Q}_k + C_kM_{k-1}\vec{Q}_{k-1}.$$

On the other hand we have

$$\overrightarrow{xM_k\vec{Q}_k} = N_k\overrightarrow{x\vec{Q}_k},$$

where N_k is an orthogonal $nr_n^k \times nr_n^k$ matrix of the following structure:

$$N_k = \begin{bmatrix} M_k & & 0 \\ & \boxed{M_k} & \\ 0 & & \frac{1}{2}\boxed{M_k} \end{bmatrix}, \qquad k = 0, 1, \cdots.$$

Hence $A'_k = N_k^TA_kM_{k+1}$, $B'_k = N_k^TB_kM_k$, $C'_k = N_k^TC_kM_{k-1}$ and the theorem follows. □

We now come back to a problem of standardization of formulas (3.2). We suppose the assumption of Theorem 2 holds. Let, for fixed k, the integers $i, j, \cdots, l$ denote the indices of linearly independent rows of matrix A_k. Recall that the number of such rows is r_n^{k+1}. If we additionally assume that D_k has nonvanishing columns with indices exactly equal to $i, j, \cdots, l$ then D_k is unique. It follows from (3.2) that, generally, in order to construct $\vec{P}_{k+1}$ from $\vec{P}_k$ and $\vec{P}_{k-1}$ it is necessary to compute $r_n^{k+1}((n+1)r_n^k + r_n^{k-1})$ coefficients of the matrices D_k, E_k, F_k. Nevertheless, taking D_k with the minimal number r_n^{k+1} of nonvanishing columns it is sufficient to determine only $r_n^{k+1}(r_n^{k+1} + r_n^k + r_n^{k-1})$ such coefficients.

PROPERTY 1. *If*

$$n(n-1) < \frac{1}{k+1}(n+k)(n+k-1) \tag{3.9}$$

holds and D_k has only r_n^{k+1} nonvanishing columns then $F_k \neq 0$.

Proof. Assume to the contrary that $\vec{P}_{k+1} = D_k\overrightarrow{x\vec{P}_k} + E_k\vec{P}_k$. Taking the quasi-inner product of P_l^{k-1} and all polynomials appearing in both vector sides of this identity we obtain $0 = D_k\vec{y}_l$, where

$$\vec{y}_l^T = [\vec{\mu}_{1l}^T|\vec{\mu}_{2l}^T|\cdots|\vec{\mu}_{nl}^T]$$

and

$$\vec{\mu}_{il}^T = [\langle P_1^k, x_iP_l^{k-1}\rangle, \cdots, \langle P_{r_n^k}^k, x_iP_l^{k-1}\rangle], \qquad i = 1, \cdots, n, \quad l = 1, 2, \cdots, r_n^{k-1}. \tag{3.10}$$

Let

$$M_i = [\vec{\mu}_{i1}|\vec{\mu}_{i2}|\cdots|\vec{\mu}_{ir_n^{k-1}}], \qquad i = 1, \cdots, n,$$
$$M^T = [M_1^T|M_2^T|\cdots|M_n^T].$$

We have $D_kM = 0$. Since all nonvanishing columns of D_k are linearly independent the matrix M has at least r_n^{k+1} vanishing rows.

We now prove that

$$\operatorname{rank} M_i = r_n^{k-1} \quad \text{for } i = 1, \cdots, n. \tag{3.11}$$

Indeed, $\sum_{j=1}^{r_n^{k-1}} \alpha_j \vec{\mu}_{ij} = 0$ implies

$$\left\langle P_l^k, x_i\left(\sum_{j=1}^{r_n^{k-1}} \alpha_j P_j^{k-1}\right)\right\rangle = 0 \quad \text{for } l = 1, \cdots, r_n^k,$$

which is possible if and only if $\alpha_j = 0$, $j = 1, \cdots, r_n^{k-1}$. Hence all columns of M_i are linearly independent, i.e., (3.11) holds.

From (3.11) it follows that M has at least nr_n^{k-1} nonvanishing rows. So we have proved that the $nr_n^k \times r_n^{k-1}$ matrix M consists of at least $r_n^{k+1} + nr_n^{k-1}$ rows. Hence $r_n^{k+1} + nr_n^{k-1} \leqq nr_n^k$, which contradicts (3.9). □

For the specific choice of r_n^{k+1} integers indicating the nonzero columns of D_k we can give the following stronger result.

PROPERTY 2. *If among r_n^{k+1} nonvanishing columns of matrix D_k there appear successively the $(i-1)r_n^k + 1$th to the ir_n^kth, where $i \leqq n$, then no column of F_k vanishes.*

Proof. Let us assume that there is a zero column of F_k, say the lth. Proceeding as in the proof of the previous property we get by (3.2) that $D_k \vec{y}_l = 0$. The linear independence of all nonvanishing columns of D_k and their distribution imply that $\vec{\mu}_{il} = 0$. Hence, according to (3.10), $x_i P_l^{k-1}$ is orthogonal to each P_j^k, $j = 1, \cdots, r_n^k$, which is not possible. □

Observe that distributions of nonvanishing columns of matrix D_k described in this property are admissible. This follows from a linear independence of the corresponding rows of A_k.

PROPERTY 3. *If C_k arises from* (3.1), *then*:

(1) *For every $i = 1, \cdots, r_n^k$ there is a $j \in \{1, \cdots, n\}$ such that $(C_k)_{i+(j-1)r_n^k} \neq 0$. ($(C_k)_l$ denotes the l*th *row of matrix C_k.)*

(2) *No column of C_k vanishes.*

Proof. (1) If $(C_k)_{i+(j-1)r_n^k} = 0$ for every $j = 1, \cdots, n$, then (3.1) yields

$$x_j P_i^k = (A_k)_{i+(j-1)r_n^k} \vec{P}_{k+1} + (B_k)_{i+(j-1)r_n^k} \vec{P}_k, \qquad j = 1, \cdots, n.$$

Hence $\langle P_i^k, x_j P_l^{k-1} \rangle = 0$ for $j = 1, \cdots, n$, $l = 1, \cdots, r_n^{k-1}$ which contradicts (3.7).

(2) The matrix F_k corresponding to an arbitrary D_k which appears in Property 2 consists of nonvanishing columns. On the other hand $F_k = -D_k C_k$. These facts can simultaneously hold if and only if no column of C_k vanishes.

4. Examples.

(I) Let $P_i^k(x_1, x_2) = T_{k-i+1}(x_1) T_{i-1}(x_2)$, $k = 0, 1, \cdots$, $i = 1, \cdots, k+1$, where the T_l are the Chebyshev polynomials of the first kind (see [5]). Polynomials $\{P_i^k\}$ are orthogonal over the region $[-1, 1] \times [-1, 1] \subset R^2$ with respect to the weight function $(1-x_1^2)^{-1/2}(1-x_2^2)^{-1/2}$. One can easily verify that

$$\vec{P}_{k+1} = D_k \overrightarrow{xP}_k + F_k \vec{P}_{k-1},$$

where

$$D_k = \begin{bmatrix} 2 & & & & & 0 & & & & & \\ & 2 & & & & & 0 & & & & 0 \\ & & 2u_1 & & & & & 2(1-u_1) & & & \\ & & & \ddots & & & & & \ddots & & \\ & & & & 2u_{k-2} & & & & & 2(1-u_{k-2}) & \\ 0 & & & & & 0 & & & & & 2 \\ & & & & & & 0 & & & & 2 \end{bmatrix},$$

$$F_k = \begin{bmatrix} -1 & & & & & & \\ 0 & -1 & & & & 0 & \\ u_1-1 & 0 & -u_1 & & & & \\ & \cdot & & \cdot & & & \\ & & \cdot & & \cdot & & \\ 0 & & & u_{k-2}-1 & 0 & -u_{k-2} & \\ & & & & -1 & 0 & \\ & & & & & -1 & \end{bmatrix}.$$

and $u_1, u_2, \cdots, u_{k-2}$ are arbitrary real numbers. ($k = 2, 3, \cdots$). In order to obtain the simplest structure of these matrices we put $u_1 = u_2 = \cdots = u_{k-2} = 0$ or 1.

(II) Let $P_i^k(x_1, x_2) = H_{k-i+1}(x_1)H_{i-1}(x_2)$, $k = 0, 1, \cdots, i = 1, \cdots, k+1$, where H_l are the Hermite polynomials (see [5]). Polynomials $\{P_i^k\}$ are orthogonal over R^2 with respect to the weight function $\exp(-x_1^2 - x_2^2)$. It is easy to verify that $\vec{P}_3 = D_2, \overrightarrow{xP}_2$ where

$$D_2 = \begin{bmatrix} 2 & 0 & 4 & 0 & -4 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & -4 & 0 & 4 & 0 & 2 \end{bmatrix}.$$

Hence for this specific choice of matrix D_2 we have $E_2 = F_2 = 0$.

(III) Let $P_i^k(x_1, x_2) = P_{k,i+1}^{\alpha,\beta,\gamma}(x_1, x_2)$, $k = 0, 1, \cdots, i = 1, \cdots, k+1$, where $P_{k,i+1}^{\alpha,\beta,\gamma}$ are the Koornwinder polynomials (see [6]). Polynomials $\{P_i^k\}$ are orthogonal over a region bounded by two lines $1 - x_1 + x_2 = 0$, $1 + x_1 + x_2 = 0$ and by the parabola $x_1^2 - 4x_2 = 0$ with respect to the weight function $(1 - x_1 + x_2)^\alpha (1 + x_1 + x_2)^\beta (x_1^2 - 4x_2)^\gamma$, ($\alpha, \beta, \gamma > 0$, $\alpha + \gamma + \frac{3}{2} > 0$, $\beta + \gamma + \frac{3}{2} > 0$). From formulas derived in [7] it follows that a standardization of (3.2) corresponding to the requirement that D_k has zero columns between the $(k+2)$nd and the $(2k+1)$th leads to a very simple and sparse structure of matrices D_k, E_k, F_k:

$$D_k = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & 0 & \\ & & \cdot & & & & \\ & & & \cdot & & & \\ & 0 & & & \cdot & & \\ & & & & & 1 & \\ & & & X & X & & 1 \end{bmatrix},$$

$$E_k = \begin{bmatrix} X & X & & & & & \\ X & X & X & & & 0 & \\ & X & & \cdot & & & \\ & & & & \cdot & & \\ & 0 & & & & \cdot & X \\ & & & & & X & X \\ & & & & X & X & X \end{bmatrix},$$

$$F_k = \begin{bmatrix} X & & & & \\ & & & 0 & \\ & X & & & \\ & & \cdot & & \\ & & & \cdot & \\ & 0 & & & X \\ & & & & 0 \\ & & & & X \end{bmatrix}.$$

However, if we assume that D_k has columns of zeros between the 2nd and the $(k+1)$th then the resulting matrices D_k, E_k, F_k are not sparse. Therefore the first standardization can be of greater practical interest.

Acknowledgments. I am greatly indebted to Prof. Stanisław Turski and Mrs. Jolanta Sokolnicka for help during the preparation of this paper.

REFERENCES

[1] D. JACKSON, *Formal properties of orthogonal polynomials in two variables*, Duke Math. J., 2 (1936), pp. 423–434.

[2] M. BERTRAN, *Note on orthogonal polynomials in ν variables*, this Journal, 6 (1975), pp. 250–257.

[3] A. MOSTOWSKI AND M. STARK, *Elements of Higher Algebra*, PWN, Warszawa, 1975. (In Polish.)

[4] G. H. FORSYTHE AND C. B. MOLER, *Computational Solution of Algebraic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[5] G. SZEGÖ, *Orthogonal Polynomials*, Colloquium Publications, American Mathematical Society, New York, 1939.

[6] T. H. KOORNWINDER, *Orthogonal polynomials in two-variables which are eigenfunctions of two algebraically independent partial differential operators, I, II*, Proc. Kon. Ned. Akad. Wetensch., A77 = Indag. Math., 36 (1974), pp. 48–66.

[7] I. G. SPRINKHUIZEN-KUYPER, *Orthogonal polynomials in two variables. A further analysis of polynomials orthogonal over a region bounded by two lines and a parabola*, this Journal, 7 (1976), pp. 501–518.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0011 $01.00/0

ORTHOGONALITY AND RECURSION FORMULAS FOR POLYNOMIALS IN n VARIABLES*

M. A. KOWALSKI†

Abstract. In this paper we obtain a necessary and sufficient condition for a sequence of polynomials in n variables to be orthogonal. A comparison criterion of orthogonality for these polynomials is also established. In conclusion an integral representation for a corresponding quasi-inner product is given.

1. Introduction. In 1935 J. Favard [1] gave, using a suitable recursion formula, a sufficient (and necessary) condition for a sequence of polynomials in one real variable to be orthogonal with respect to an integral inner product. In our paper a generalization of this result to the multivariate case is presented. We specify all recursion formulas for polynomials in n variables (in the sense of [2, Thm. 2]) which imply orthogonality. We introduce the comparison criterion of orthogonality for sequences of polynomials. Using this criterion we give an integral representation for a corresponding quasi-inner product.

2. Main results. Let $M_1, M_2, \cdots, M_n$ be any matrices with identical dimensions; we shall use the symbol bp (block permutation) to denote the operation

$$\operatorname{bp}([M_1|M_2|\cdots|M_n]) = \begin{bmatrix} M_1 \\ \hline M_2 \\ \hline \vdots \\ \hline M_n \end{bmatrix}.$$

Further we use the same notation and terminology as in [2].

The following theorem is a generalization of a well-known result for polynomials in one variable (see [3, Thms. 4.1, 4.4, pp. 18–22]).

THEOREM 1. *Let* $\{P_i^k\}_{k=0,i=1}^{\infty,r_n^k}$, $P_1^0 \neq 0$ *be an arbitrary sequence in the space* Π_n^∞. *Then the following statements are equivalent*:

1) *There exist a bilinear functional* $\langle \cdot, \cdot \rangle : \Pi_n^\infty \times \Pi_n^\infty \to R$ *and numbers* $M_{ki} \neq 0$ *such that*

$$\langle P_i^k, P_j^l \rangle = \langle 1, P_i^k P_j^l \rangle = M_{ki}\delta_{kl}\delta_{ij}.$$

2) *For each* $k = 0, 1, \cdots$ *there exist matrices* A_k, B_k, C_k *such that*

(a) $$\operatorname{rank} A_k = r_n^{k+1},$$

(b) $$\overrightarrow{xP}_k = A_k\vec{P}_{k+1} + B_k\vec{P}_k + C_k\vec{P}_{k-1}, \qquad \vec{P}_{-1} = 0,$$

and for an arbitrary sequence of matrices $D_0, D_1, \cdots$ *such that* $D_kA_k = I$ *the recursion*

(c) $$I_0 = [1], \qquad I_{j+1} = D_j \operatorname{bp}(I_jC_{j+1}^T), \qquad j = 0, 1, \cdots$$

produces diagonal and nonsingular matrices I_j.

Proof. 2) $\Rightarrow$ 1). First we prove that for $k = 0, 1, \cdots$ the polynomial coefficients of $\vec{P}_k$ are linearly independent with respect to Π_n^{k-1} ($\Pi_n^{-1} = \{0\}$). We use induction on k. The case $k = 0$ is an easy consequence of the assumption $P_1^0 \neq 0$. Suppose now that for some $k \geqq 0$ the polynomial coefficients of $\vec{P}_k$ are linearly independent with respect to Π_n^{k-1}. This yields

$$\Pi_n^{k+1} = \operatorname{lin}(\overrightarrow{xP}_k, \Pi_n^k),$$

* Received by the editors May 27, 1980, and in revised form May 1, 1981.
† Institute of Informatics, University of Warsaw P.K.iN. 8p. 850, 00-901 Warsaw, Poland.

where the right-hand side denotes the vector space spanned by the coefficients of $\overrightarrow{xP}_k$ and the elements of Π_n^k. On the other hand, assumptions (a), (b) imply that

$$\operatorname{lin}(A'_k\vec{P}_{k+1}, \Pi_n^k) = \operatorname{lin}(\overrightarrow{xP}_k, \Pi_n^k),$$

where A'_k is a nonsingular $r_n^k \times r_n^k$ matrix consisting of linearly independent rows of the matrix A_k. So we have

$$\operatorname{lin}(A'_k\vec{P}_{k+1}, \Pi_n^k) = \Pi_n^{k+1}.$$

This and the nonsingularity of the matrix A'_k proves that the polynomial coefficients of $\vec{P}_{k+1}$ are linearly independent with respect to Π_n^k.

This result means that $\{P_i^k\}_{k=0,i=1}^{\infty,r_n^k}$ forms a basis of the space Π_n^∞. Hence the linear functional

$$l(P_i^k) = \frac{\delta_{k0}}{P_1^0}, \qquad k = 0, 1, \cdots, \quad i = 1, \cdots, r_n^k, \tag{2.1}$$

is well defined on Π_n^∞. We show that

$$l(P_i^k P_j^l) = \delta_{kl}\delta_{ij}M_{ki}, \tag{2.2}$$

where $M_{ki} \neq 0$, $k, l = 0, 1, \cdots$, $i = 1, \cdots, r_n^k$, $j = 1, \cdots, r_n^l$. For convenience we introduce an operator $\mathscr{L}$ which maps any polynomial matrix $m = (m_{ij}(x))$ onto a real matrix $\mathscr{L}(m) = (l(m_{ij}))$. It is clear that $\mathscr{L}$ has the following property: For any polynomial matrices m_1, m_2 and any real matrices s_1, s_2 such that $s_1m_1 + m_2s_2$ is well defined, we have

$$\mathscr{L}(s_1m_1 + m_2s_2) = s_1\mathscr{L}(m_1) + \mathscr{L}(m_2)s_2.$$

The condition (2.2) takes the form

$$\mathscr{L}(\vec{P}_k\vec{P}_l^T) = \delta_{kl}\operatorname{diag} M_{ki}. \tag{2.3}$$

To prove (2.3) we first verify that $\mathscr{L}(\vec{P}_k\vec{P}_l^T) = 0$ for $k \neq l$.

Induction on k. In the case $k = 0$ this result immediately follows from (2.1). Suppose now that $\mathscr{L}(\vec{P}_i\vec{P}_j^T) = 0$ for $0 \leqq i \leqq k$ and $j > i$. Hence for $l \geqq k+1$ we have

$$\begin{aligned}
\mathscr{L}(\vec{P}_{k+1}\vec{P}_l^T) &= \mathscr{L}((D_k\overrightarrow{xP}_k + E_k\vec{P}_k + F_k\vec{P}_{k-1})\vec{P}_l^T) \\
&= D_k\mathscr{L}(\overrightarrow{xP}_k\vec{P}_l^T) \\
&= D_k \operatorname{bp}(\mathscr{L}(\vec{P}_k\overrightarrow{xP}_l^T)) \\
&= D_k \operatorname{bp}(\mathscr{L}(\vec{P}_k(A_l\vec{P}_{l+1} + B_l\vec{P}_l + C_l\vec{P}_{l-1})^T)) \\
&= D_k \operatorname{bp}(\mathscr{L}(\vec{P}_k\vec{P}_{l-1}^T)C_l^T),
\end{aligned}$$

which for $l > k+1$ concludes the induction step and for $l = k+1$ gives

$$\mathscr{L}(\vec{P}_{k+1}\vec{P}_{k+1}^T) = D_k \operatorname{bp}(\mathscr{L}(\vec{P}_k\vec{P}_k^T)C_{k+1}^T). \tag{2.4}$$

Thus, according to (c), we obtain $\mathscr{L}(\vec{P}_k\vec{P}_l^T) = \delta_{kl}I_k$ which proves (2.3) and (2.2). Now it is evident that the equality

$$\langle r, s\rangle = l(rs), \qquad r, s \in \Pi_n^\infty$$

defines the quasi-inner product fulfilling the requirements of 1).

1) $\Rightarrow$ 2). Relations (a), (b) and the existence of a matrix D_k such that $D_k A_k = I$, $k = 0, 1, \cdots$, has been proved in [2]. For the linear functional

$$l(r) = \frac{1}{M_{01}}\langle 1, r\rangle, \qquad r \in \Pi_n^\infty,$$

we have $\mathscr{L}(\vec{P}_k \vec{P}_l^T) = \delta_{kl} I_k$, where $I_k = \operatorname{diag} M_{ki}/M_{01}$ and $\mathscr{L}$ is defined as in the first part of the proof. Observe that the relation (2.4) remains true. This gives (c) and completes the proof. □

In order to obtain a similar result for the case when $\langle \cdot, \cdot \rangle$ is an inner product in Π_n^∞, it is necessary to change slightly an assumption made in 2), namely, that the matrices I_j should be positive definite and not merely non singular.

For convenience, by $\{Q_i^k\}_{k=0,i=1}^{\infty,r_n^k} = \{q_i\}_{i=1}^\infty$ we mean that

$$q_1 = Q_1^0, \quad q_2 = Q_1^1, \quad q_3 = Q_2^1, \quad \cdots, \quad q_{i+r_n^{k-1}} = Q_i^k, \quad \cdots.$$

Let μ be a measure on the δ-field of Borel sets in R^n. Then $L_2(R^n, \mu)$ will denote the space of all μ-measurable functions $f: R^n \to R$ such that the function f^2 is μ-integrable. Further two functions which are equal for μ for almost all $x \in R^n$ are considered to be the same. By the norm $\|f\|_2$ of an element f from $L_2(R^n, \mu)$ we mean the quantity

$$\|f\|_2 = \left(\int_{R^n} f^2 \, d\mu\right)^{1/2}.$$

It is well known that $L_2(R^n, \mu)$ is a Hilbert space with the inner product

$$\langle f, g\rangle = \int_{R^n} fg \, d\mu.$$

THEOREM 2 (Comparison criterion of orthogonality). *Suppose*:

1) $\{Q_i^k\}_{k=0,i=1}^{\infty,r_n^k} = \{q_i\}_{i=1}^\infty$ *is an orthogonal polynomial sequence in* $L_2(R^n, \mu)$.

2) $\{P_i^k\}_{k=0,i=1}^{\infty,r_n^k} = \{p_i\}_{i=1}^\infty$ *is a polynomial sequence satisfying the second condition of Theorem* 1.

Let

$$q_i = \sum_{j=1}^{\infty} C_{ij} p_j, \qquad i = 1, 2, \cdots.$$

If

$$\sum_{i=1}^{\infty} C_{i1}^2 < \infty, \tag{2.5}$$

then there exists a function $\rho \in L_2(R^n, \mu)$, $\rho \neq 0$ *and numbers* $m_i \neq 0$ *such that*

$$\int_{R^n} p_i p_j \rho \, d\mu = \delta_{ij} m_i, \qquad i, j = 1, 2, \cdots, \tag{2.6}$$

and conversely, the existence of $\rho \in L_2(R^n, \mu)$ *satisfying* (2.6) *implies* (2.5).

Proof. Sufficiency. By Theorem 1 there exist a linear functional l on Π_n^∞ and a corresponding quasi-inner product $\langle \cdot, \cdot \rangle$ such that

$$\langle p_i, p_j\rangle = l(p_i p_j) = \delta_{ij} m_i, \qquad i, j = 1, 2, \cdots, \tag{2.7}$$

where $m_i \neq 0$. Assume without loss of generality that $l(p_1) = 1$. We now prove that l is continuous with respect to the norm $\|\cdot\|_2$. Let w be an arbitrary polynomial. Define

$$d(w) = \min\{\dim \Pi_n^k | w \in \Pi_n^k\}.$$

We can express w in the form

$$w = \sum_{i=1}^{d(w)} a_i q_i. \tag{2.8}$$

Now let C be an $\infty \times \infty$ matrix such that

$$q_i = \sum_{j=1}^{\infty} C_{ij} p_j, \qquad i = 1, 2, \cdots. \tag{2.9}$$

The above sum is only formally infinite because

$$C_{ij} = 0 \quad \text{for } j > d(q_i), \quad i = 1, 2, \cdots.$$

From identities (2.8), (2.9) and this observation it follows that

$$w = \sum_{j=1}^{d(w)} \left(\sum_{i=1}^{d(w)} C_{ij} a_i \right) p_j. \tag{2.10}$$

By the orthonormality of polynomials q_j we obtain

$$\sum_{j=1}^{d(w)} a_j^2 = \|w\|_2^2. \tag{2.11}$$

Write now

$$\alpha_j = \sum_{i=1}^{d(w)} C_{ij} a_i, \qquad j = 1, 2, \cdots. \tag{2.12}$$

Using (2.10) and (2.12) for $j = 1$, we get

$$|l(w)| = |\alpha_1| = \left| \sum_{i=1}^{d(w)} C_{i1} a_i \right|.$$

Applying Schwarz's inequality and (2.11) we obtain

$$|l(w)| \leqq \left(\sum_{i=1}^{d(w)} C_{i1}^2 \right)^{1/2} \|w\|_2.$$

For $w = \sum_{j=1}^{d(w)} C_{j1} q_j$ this inequality becomes an equality. Hence we have proved that

$$\|l\| = \left(\sum_{i=1}^{\infty} C_{i1}^2 \right)^{1/2}.$$

Due to (2.5) this gives the continuity of the functional l.

We use now the Hahn–Banach theorem (see [4, pp. 62–63]) to obtain a continuous functional L on $L_2(R^n, \mu)$ as an extension of l. By virtue of the Riesz theorem (see [4, pp. 249–250]) there exists a unique function $\rho \in L_2(R^n, \mu)$ such that

$$L(f) = \int_{R^n} f\rho \, d\mu.$$

Hence

$$\int_{R^n} p_i p_j \rho \, d\mu = \delta_{ij} m_i.$$

This completes the first part of the proof.

Necessity. The linear functional L on $L_2(R^n, \mu)$ defined by the equation

$$L(f) = \int_{R^n} f\rho \, d\mu$$

has the norm $\|L\| = \|\rho\|_2$. On the other hand, proceeding as in the first part of the proof, we can verify that

$$\|l\|^2 = \sum_{i=1}^{\infty} C_{i1}^2,$$

where l is the restriction of L to Π_n^∞. So

$$\sum_{i=1}^{\infty} C_{i1}^2 = \|l\|^2 \leqq \|L\|^2 = \|\rho\|_2^2.$$

The proof is complete. □

Observe that the coefficients C_{ij} are completely determined by the recursion formulas for the polynomials P and Q.

THEOREM 3. *Let $\{P_i^k\}_{k=0,i=1}^{\infty,rk}$ be a basis in the space Π_n^∞ satisfying the identities*

$$\overrightarrow{xP}_k = A_k\vec{P}_{k+1} + B_k\vec{P}_k + C_k\vec{P}_{k-1}, \qquad k = 0, 1, \cdots, \quad C_0 = 0, \quad P_{-1} = 0 \tag{2.13}$$

for suitable matrices A_k, B_k, C_k. Then there exist a nondecreasing function ϕ with infinitely many points of increase and polynomials $\{Q_i^k\}_{k=0,i=1}^{\infty,rk}$ such that:

(i) $$\int_{R^n} Q_i^k(x_1, \cdots, x_n) Q_j^l(x_1, \cdots, x_n)\, d\phi(x_1) \cdots d\phi(x_n) = \delta_{kl}\delta_{ij},$$

(ii) for matrices G_{ij} such that $\vec{Q}_i = \sum_{j=0}^{i} G_{ij}\vec{P}_j$ we have

$$\|G_{ij}\|_\infty < 2^{-i+1}|G_{00}|, \qquad i = 1, 2, \cdots, \quad j = 0, 1, \cdots.$$

Proof. Let polynomials Q in n variables be defined by the recursion

$$\vec{Q}_{k+1} = d_k\overrightarrow{xQ}_k + f_k\vec{Q}_{k-1}, \qquad k = 0, 1, \cdots, \quad f_0 = 0, \quad \vec{Q}_{-1} = 0 \tag{2.14}$$

and let

$$\vec{Q}_i = \sum_{j=0}^{i} G_{ij}\vec{P}_j, \qquad i = 0, 1, \cdots. \tag{2.15}$$

(Since the polynomials P form a basis in Π_n^∞, the matrices G_{ij} exist and are uniquely determined.) Substituting (2.15) into (2.14) we get

$$\sum_{j=0}^{k+1} G_{k+1,j}\vec{P}_j = d_k \sum_{j=0}^{k} \overrightarrow{xG_{kj}P}_j + f_k \sum_{j=0}^{k-1} G_{k-1,j}\vec{P}_j.$$

Note that for any matrix M such that the product $M\vec{P}_j$ is well defined we have

$$\overrightarrow{xMP}_j = [M]\overrightarrow{xP}_j,$$

where $[M]$ is the n-block matrix of the form

$$[M] = \begin{bmatrix} M & & 0 \\ & \ddots & \\ 0 & & M \end{bmatrix}.$$

This and (2.13) lead to the equality

$$\sum_{j=0}^{k+1} G_{k+1,j}\vec{P}_j = d_k \sum_{j=0}^{k} [G_{kj}](A_j\vec{P}_{j+1} + B_j\vec{P}_j + C_j\vec{P}_{j-1}) + f_k \sum_{j=0}^{k} G_{k-1,j}\vec{P}_j, \qquad k = 0, 1, \cdots.$$

If G_{st}, for $s<0$, $t<0$ or $t>s$, are interpreted to be zero matrices, we can rewrite this relation in the form

$$\sum_{j=0}^{k+1} G_{k+1,j}\vec{P}_j = \sum_{j=0}^{k+1} \{d_k([G_{k,j-1}]A_{j-1} + [G_{kj}]B_j + [G_{k,j+1}]C_{j+1}) + f_k G_{k-1,j}\}\vec{P}_j,$$
$$k = 0, 1, \cdots.$$

So

$$G_{k+1,j} = d_k([G_{k,j-1}]A_{j-1} + [G_{kj}]B_j + [G_{k,j+1}]C_{j+1}) + f_k G_{k-1,j},$$
$$k = 0, 1, \cdots, \quad j = 0, 1, \cdots, k+1.$$

This yields

$$\|G_{k+1,j}\| \leqq \|d_k\| \cdot \|[G_{k,j-1}]\| \cdot \|A_{j-1}\| + \|[G_{kj}]\| \cdot \|B_j\| + \|[G_{k,j+1}]\| \cdot \|C_{j+1}\| + \|f_k\| \cdot \|G_{k-1,j}\|$$
$$\leqq \left(3\|d_k\| \max_{0\leqq l\leqq k} (\max(\|A_l\|, \|B_l\|, \|C_l\|)) + \|f_k\|\right) \max_{\substack{0\leqq l\leqq p \\ p=k,k-1}} \|G_{pl}\|, \qquad k = 0, 1, \cdots.$$

(Here and everywhere below, $\|\cdot\| = \|\cdot\|_\infty$. In the last inequality we use $\|M\| = \|[M]\|$ for any matrix M.) Let $\Delta_k = 3 \max_{0\leqq l\leqq k} (\max(\|A_l\|, \|B_l\|, \|C_l\|))$, $k = 0, 1, \cdots$. We have proved that

$$\max_{0\leqq l\leqq k+1} \|G_{k+1,l}\| \leqq (\|d_k\|\Delta_k + \|f_k\|) \max_{\substack{0\leqq l\leqq p \\ p=k,k-1}} \|G_{pl}\|. \tag{2.16}$$

We now prove that polynomials Q can be chosen to be orthonormal with respect to an integral inner product and to satisfy the inequality

$$\|d_k\|\Delta_k + \|f_k\| < \frac{1}{4}, \qquad k = 0, 1, \cdots.$$

Let $\{\alpha_k\}_{k=1}^{\infty}$, $\{\gamma_k\}_{k=1}^{\infty}$ be real sequences satisfying $(\alpha_{k-1}\gamma_k)/\alpha_k = 1$, $k = 2, 3, \cdots$, $\gamma_1 = 0$. Consider polynomials in one variable produced by the formula

$$q_{l+1}(y) = \alpha_{l+1} y q_l(y) - \gamma_{l+1} q_{l-1}(y), \qquad l = 0, 1, \cdots, \quad y \in R, \tag{2.17}$$
$$q_{-1}(y) = 0, \qquad q_0(y) = 1.$$

By Favard's theorem there exists a nondecreasing function ϕ with infinitely many points of increase such that

$$\int_R q_k(y) q_l(y)\, d\phi(y) = \delta_{kl}, \qquad k, l = 0, 1, \cdots.$$

(The integral is considered with respect to the Lebesgue–Stieltjes measure introduced by the function ϕ.) This yields that the sequence of polynomials

$$Q = \{Q_i^k\}_{k=0, i=1}^{\infty, r_k} = \{q_{i_1}(x_1) q_{i_2}(x_2) \cdots q_{i_n}(x_n)\}_{k=0, i_1+i_2+\cdots+i_n=k}^{\infty}$$

is orthonormal with respect to the inner product

$$\langle f, g\rangle = \int_{R^n} fg\, d\phi(x_1)\, d\phi(x_2) \cdots d\phi(x_n).$$

Let $Q_i^{k+1}(x_1, \cdots, x_n) = q_{j_1}(x_1) \cdots q_{j_n}(x_n)$ be a given polynomial of degree $k+1$ from the sequence Q. Since $j_1 + j_2 + \cdots + j_n = k+1$, there exists $s \in \{1, \cdots, n\}$ such that $j_s \geqq \lceil (k+1)/n \rceil$ ($\lceil y \rceil =$ the smallest integer $\geqq y$). Using (2.17) for $l = j_s - 1$ we get

$$Q_i^{k+1}(x_1, \cdots, x_n) = \alpha_{j_s} x_{j_s} Q_t^k(x_1, \cdots, x_n) - \gamma_{j_s} Q_u^{k-1}(x_1, \cdots, x_n),$$

where indices t, u are such that

$$Q_t^k(x_1, \cdots, x_n) = Q_i^{k+1}(x_1, \cdots, x_n) \frac{q_{j_s-1}(x_s)}{q_{j_s}(x_s)}$$

and

$$Q_u^{k-1}(x_1, \cdots, x_n) = Q_i^{k+1}(x_1, \cdots, x_n) \frac{q_{j_s-2}(x_s)}{q_{j_s}(x_s)}.$$

Thus, matrices d_k, f_k in the formula

$$\vec{Q}_{k+1} = d_k \overrightarrow{xQ}_k + f_k \vec{Q}_{k-1}, \qquad k = 0, 1, \cdots$$

can be chosen so that they have exactly one nonvanishing element in each row and these elements of d_k, f_k are contained in the sets $\{\alpha_{k+1}, \alpha_k, \cdots, \alpha_{\lceil (k+1)/n \rceil}\}$ and $\{\gamma_{k+1}, \gamma_k, \cdots, \gamma_{\lceil (k+1)/n \rceil}\}$, respectively. Let $S_p = \{i \in \mathcal{N} \mid \lceil (i+1)/n \rceil = p\}$. We now define positive numbers α_j, γ_j by the recursions

$$\alpha_1 \leqq \min_{i \in S_1} \frac{1}{8\Delta_i}, \qquad \alpha_j \leqq \min_{i \in S_j} \frac{\alpha_{j-1}}{8(1 + \Delta_i \alpha_{j-1})}, \qquad j = 2, 3, \cdots,$$

$$\gamma_1 = 0, \qquad \gamma_j = \frac{\alpha_j}{\alpha_{j-1}}, \qquad j = 2, 3, \cdots.$$

It is easy to verify that this definition guarantees us

$$\alpha_1 > 0, \quad \frac{\alpha_{i-1}\gamma_i}{\alpha_i} = 1, \quad 0 < \alpha_i < \frac{1}{8}\alpha_{i-1}, \quad \gamma_{i-1} < \frac{1}{8}, \qquad i = 2, 3, \cdots,$$

$$\Delta_k \alpha_{\lceil (k+1)/n \rceil} + \gamma_{\lceil (k+1)/n \rceil} < \frac{1}{8}, \qquad k = 0, 1, \cdots.$$

So for matrices d_k, f_k we have

$$\|d_k\| \leqq \alpha_{\lceil (k+1)/n \rceil}, \qquad \|f_k\| \leqq \frac{1}{8}, \qquad k = 0, 1, \cdots.$$

Thus

$$\|d_k\|\Delta_k + \|f_k\| \leqq \Delta_k \alpha_{\lceil (k+1)/n \rceil} + \frac{1}{8} \leqq \frac{1}{8} - \gamma_{\lceil (k+1)/n \rceil} + \frac{1}{8} < \frac{1}{4}.$$

Finally, according to (2.16), we get

$$\max_{0 \leqq l \leqq k+1} \|G_{k+1,l}\| < \frac{1}{4} \max_{\substack{0 \leqq l \leqq p \\ p = k, k-1}} \|G_{pl}\|, \qquad k = 0, 1, \cdots,$$

which gives

$$\max_{0 \leqq l \leqq k} \|G_{kl}\| < 2^{-k+1} |G_{00}|, \qquad k = 1, 2, \cdots.$$

This completes the proof of the theorem. □

Our next result states an integral representation for the quasi-inner product appearing in Theorem 1.

THEOREM 4. *Let* $\{P_i^k\}_{k=0,i=1}^{\infty,r_k}$ *be a polynomial sequence satisfying the second condition of Theorem* 1. *Then there exists a nondecreasing function* ϕ *with infinitely many points of increase and a function* $\rho: R^n \to R$ *such that*

$$\int_{R^n} \rho^2 \, d\phi(x_1) \cdots d\phi(x_n) < \infty, \tag{2.18}$$

$$\int_{R^n} P_i^k P_j^l \rho \, d\phi(x_1) \cdots d\phi(x_n) = \delta_{kl}\delta_{ij} M_{ki}, \tag{2.19}$$

where $M_{ki} \neq 0$, $k, l = 0, 1, \cdots$, $i = 1, \cdots, r_n^k$, $j = 1, \cdots, r_n^l$.

Proof. Consider a function ϕ and polynomials $\{Q_i^k\}_{k=0,i=1}^{\infty,r_k}$ obtained by Theorem 3. From Theorem 2 it follows that in order to establish (2.18), (2.19) it is enough to verify the condition

$$\sum_{i=0}^{\infty} \|G_{i0}\|_2^2 < \infty,$$

where $\|G_{i0}\|_2$ is the spectral norm of G_{i0}. By (ii) we get

$$\sum_{i=0}^{\infty} \|G_{i0}\|_2^2 \leqq \sum_{i=0}^{\infty} r_n^i \|G_{i0}\|^2 \leqq \sum_{i=0}^{\infty} \binom{n+i-1}{i} 2^{-2(i-1)} |G_{00}|^2$$

$$\leqq |G_{00}|^2 2^{n+1} \sum_{i=0}^{\infty} 2^{-i} = |G_{00}|^2 2^{n+2} < \infty,$$

as claimed. □

Compare this result with that of Shohat [5] for polynomials in one variable (see also [3, p. 75]).

The subject will be continued in our next paper.

REFERENCES

[1] J. FAVARD, *Sur les polynômes de Tchebicheff*, C.R. Acad. Sci. Paris, 200 (1935), pp. 2052–2053.

[2] M. A. KOWALSKI, *The recursion formulas for orthogonal polynomials in n variables*, this Journal, this issue, pp. 309–315.

[3] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Mathematics and Its Applications, vol. 13, Gordon and Breach, New York, 1978.

[4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators I*, Interscience Publishers, New York, 1958.

[5] J. A. SHOHAT, *Sur les polynômes orthogonaux généralisées*, C.R. Acad. Sci. Paris, 207 (1938), pp. 556–558.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0012 $01.00/0

PROJECTION FORMULAS AND A NEW PROOF OF THE ADDITION FORMULA FOR THE JACOBI POLYNOMIALS*

THOMAS P. LAINE†

Abstract. A new projection formula is used to give short proofs of the addition formulas for the Jacobi and Gegenbauer polynomials. A discrete analogue of the projection formula is also given.

1. Introduction. Since Koornwinder's original group-theoretic derivation of the addition formula for the Jacobi polynomials [6], several other proofs have been given, including two analytic ones [7], [8]. However, because of the importance of Koornwinder's addition theorem and because work still remains to be done in finding addition theorems for orthogonal polynomials, another analytic proof seems worthwhile. The relatively short proof given here uses essentially just Bateman's bilinear sum and a new "projection" formula for the Jacobi polynomials. As an intermediate step in the proof, we also use a corollary of the projection formula to give a new proof of the Gegenbauer addition theorem.

The Jacobi polynomials $R_n^{(\alpha,\beta)}(x)(\alpha, \beta > -1)$ are orthogonal on the interval $(-1, 1)$ with respect to the weight function $(1-x)^\alpha(1+x)^\beta$, and are normalized by $R_n^{(\alpha,\beta)}(1) = 1$. In terms of the hypergeometric function, they are given by

$$R_n^{(\alpha,\beta)}(x) = F(-n, n+\alpha+\beta+1; \alpha+1; \tfrac{1}{2}(1-x)). \tag{1.1}$$

The degenerate integrated Jacobi addition formula [7, (4.9)] is

$$\begin{aligned}(x-1)^{(k+l)/2}&(x+1)^{(k-l)/2}R_{n-k}^{(\alpha+k+l,\beta+k-l)}(x)\\ &= \frac{2^{k+1}(n-k)!\,\Gamma(\alpha+1+k+l)\Gamma(\beta+1+n-l)}{n!\,\Gamma(\tfrac{1}{2})\Gamma(\beta+\tfrac{1}{2})\Gamma(\alpha-\beta)\Gamma(\beta+1+n)}\\ &\quad\cdot\int_0^1\int_0^\pi\left[\frac{1}{2}(x+1)+\frac{1}{2}(x-1)r^2+i\sqrt{1-x^2}\,r\cos\phi\right]^n\\ &\quad\cdot R_l^{(\alpha-\beta-1,\beta+k-l)}(2r^2-1)r^{2\beta+k-l+1}(1-r^2)^{\alpha-\beta-1}\\ &\quad\cdot R_{k-l}^{(\beta-1/2,\beta-1/2)}(\cos\phi)(\sin\phi)^{2\beta}\,d\phi\,dr, \qquad 0\leqq l\leqq k\leqq n, \quad \alpha>\beta>-\tfrac{1}{2}.\end{aligned} \tag{1.2}$$

Formula (1.2), although simpler, is equivalent to the Jacobi addition theorem [7, (4.14)] by virtue of Bateman's bilinear sum,

$$R_n^{(\alpha,\beta)}(x)R_n^{(\alpha,\beta)}(y) = \sum_{k=0}^{n} b_{k,n}(x+y)^k R_k^{(\alpha,\beta)}\left(\frac{1+xy}{x+y}\right), \tag{1.3}$$

where the $b_{k,n}$ are defined when $y = 1$:

$$R_n^{(\alpha,\beta)}(x) = \sum_{k=0}^{n} b_{k,n}(x+1)^k. \tag{1.4}$$

For applying (1.3) on the left-hand side of (1.2) and (1.4) on the right-hand side gives (cf. [7]) an equivalent integrated version of the addition theorem. It therefore suffices to prove (1.2).

* Received by the editors July 12, 1979, and in revised form March 30, 1981.

† Department of Mathematics, University of Alabama, University, Alabama 35486. Current address 198 Seventh Avenue, Brooklyn, New York 11215.

In the classical Gegenbauer case $\alpha=\beta$, the degenerate integrated addition theorem is much simpler:

(1.5)
$$(y^2-1)^{1/2k}R_{n-k}^{(\beta+k,\beta+k)}(y)=\frac{2k(n-k)!\Gamma(\beta+1+k)}{n!\Gamma(\frac{1}{2})\Gamma(\beta+\frac{1}{2})}\int_0^{\pi}[y+i\sqrt{1-y^2}\cos\phi]^n \cdot R_k^{(\beta-1/2,\beta-1/2)}(\cos\phi)(\sin\phi)^{2\beta}\,d\phi.$$

The cases $k=l=0$ of (1.2) and $k=0$ of (1.5) are called the Jacobi and Gegenbauer Laplace representations, respectively. In [1], Askey showed that the Jacobi Laplace representation may be obtained by applying the projection formula

$$(1.6)\quad \frac{(1-x)^{\alpha}}{(1+x)^{n+\beta+1}}R_n^{(\alpha,\beta)}(x)=\frac{2^{\alpha-\beta}\Gamma(\alpha+1)}{\Gamma(\beta+1)\Gamma(\alpha-\beta)}\int_x^1\frac{(1-y)^{\beta}}{(1+y)^{n+\alpha+1}}R_n^{(\beta,\beta)}(y)(y-x)^{\alpha-\beta-1}\,dy,$$

$\alpha>\beta$, to the Gegenbauer Laplace representation and then making a change of variables. The integral (1.6) is in turn a consequence of Bateman's integral

$$(1.7)\quad F(a,b;c+\mu;x)=\frac{\Gamma(c+\mu)}{\Gamma(c)\Gamma(\mu)}\int_0^1 y^{c-1}(1-y)^{\mu-1}F(a,b;c;xy)\,dy,\qquad c,\mu>0,$$

the Pfaff transformation formula

$$(1.8)\qquad F(a,b;c;x)=(1-x)^{-a}F\left(a,c-b;c;\frac{x}{x-1}\right),$$

and (1.1). See [2].

This suggests that (1.2) could be obtained for $0\leqq l\leqq k\leqq n$ from (1.5) in much the same way, by means of a generalization of (1.6). This is in fact the case, and in § 2 we derive the required projection formula from Bateman's integral. In § 3, we apply a quadratic transformation to a special case of the projection formula to obtain (1.5), and then complete the proof of the Jacobi addition theorem by applying the general projection formula to (1.5) to obtain (1.2). Finally, in § 4 we find discrete analogues of the projection formula, which give projection formulas for the Hahn orthogonal polynomials.

2. Projection formulas. Let l be a nonnegative integer, and replace a, b, c, μ and y in (1.7) by $a+l$, $b+l$, $c+l$, $\mu+l$ and $1-y$, respectively, to obtain

$$(2.1)\quad F(a+l,b+l;c+\mu+2l;x)=\frac{-\Gamma(c+\mu+2l)}{\Gamma(c+l)\Gamma(\mu+l)}\cdot\int_0^1 y^{\mu+l-1}(1-y)^{c+l-1}F(a+l,b+l;c+l;x(1-y))\,dy.$$

Now, by termwise differentiation,

$$\frac{d^l}{dy^l}F(a,b;c;x(1-y))=\frac{(a)_l(b)_l}{(c)_l}x^l(-1)^lF(a+l,b+l;c+l;x(1-y)).$$

Hence, (2.1) may be rewritten as

$$(2.2)\quad x^lF(a+l,b+l;c+\mu+2l;x)=\frac{\Gamma(c+\mu+2l)(-1)^{l+1}}{\Gamma(c)\Gamma(\mu+l)(a)_l(b)_l}\int_0^1 y^{\mu+l-1}(1-y)^{c+l-1}\frac{d^l}{dy^l}F(a,b;c;x(1-y))\,dy.$$

Integrating by parts, we have

$$\int_0^1 y^{\mu+l-1}(1-y)^{c+l-1}\frac{d^l}{dy^l}F(a,b;c;x(1-y))\,dy$$

$$(2.3)\qquad =(-1)^l\int_0^1 \frac{d^l}{dy^l}[y^{\mu+l-1}(1-y)^{c+l-1}]F(a;b;c;x(1-y))\,dy,$$

since, with $\mu, c>0$, the integrated terms vanish. But, by [3, (17), p. 102],

$$(2.4)\qquad \frac{d^l}{dy^l}[y^{\mu+l-1}(1-y)^{c+l-1}]=(\mu)_l y^{\mu-1}(1-y)^{c-1}F(-l,c+\mu+l-1;\mu;y).$$

Hence, using (2.2)–(2.4) and replacing y by $1-y$ gives

$$x^l F(a+l,b+l;c+\mu+2l;x)$$

$$(2.5)\qquad =\frac{\Gamma(c+\mu+2l)}{\Gamma(c)\Gamma(\mu)(a)_l(b)_l}\int_0^1 y^{c-1}(1-y)^{\mu-1}F(-l,c+\mu+l-1;\mu;1-y)$$

$$\cdot F(a,b;c;xy)\,dy,\qquad c,\mu>0,\quad l=0,1,2,\cdots,$$

which is a generalization of (1.7).

Since the projection formula we need to go between (1.5) and (1.2) would generalize (1.6) instead, we continue as in the derivation of (1.6) from (1.7) in [2]. Apply (1.8) to the $F(a+l,b+l;c+\mu+2l;x)$ and $F(a,b;c;xt)$ in (2.5), replace $c-b$ by c, and let

$$z=\frac{x}{x-1},\qquad y=\frac{s(1-z)}{z(1-s)},$$

to obtain

$$\frac{z^{c+\mu+l-1}}{(1-z)^{c-a}}F(a+l,b+\mu+l;c+\mu+2l;z)$$

$$(2.6)\qquad =\frac{\Gamma(c+\mu+2l)(-1)^l}{\Gamma(c)\Gamma(\mu)(a)_l(c-b)_l}\int_0^z \frac{s^{c-1}}{(1-s)^{c+\mu-a}}(z-s)^{\mu-1}$$

$$\cdot F\left(-l,c+\mu+l-1;\mu;\frac{(z-s)}{(z-zs)}\right)F(a,b;c;s)\,ds.$$

Letting $a=k-n-l$, $b=n+k-l+2\beta+1$, $c=\beta+k-l+1$ and $\mu=\alpha-\beta$ and using (1.1) gives

$$\frac{(1-x)^{\alpha+k}}{(1+x)^{\beta+n+1}}R_{n-k}^{(\alpha+k+l,\beta+k-l)}(x)$$

$$=\frac{2^{\alpha-\beta+l}\Gamma(\alpha+1+k+l)}{\Gamma(\beta+1+k-l)\Gamma(\alpha-\beta)(k-n-l)_l(\beta+1+n-l)_l}$$

$$(2.7)\qquad \cdot\int_x^1 \frac{(1-y)^{\beta+k-l}}{(1+y)^{\alpha+n+1}}(y-x)^{\alpha-\beta-1}R_l^{(\alpha-\beta-1,\beta+k-l)}$$

$$\cdot\left(\frac{1-3y+3x-xy}{(1+y)(1-x)}\right)R_{n-k+l}^{(\beta+k-l,\beta+k-l)}(y)\,dy,$$

which is the required generalization of (1.6).

Formula (2.7) could also have been proved directly, by expanding the function

$$R_l^{(\alpha-\beta-1,\beta+k-l)}\left(\frac{1-3y+3x-xy}{(1+y)(1-x)}\right),$$

interchanging summation and integration, applying (1.6) to evaluate the resulting integrals and then using Saalschutz's theorem [3, (3), p. 188].

Formulas (2.5) and (2.6) appear to be new; in particular, they are not consequences of Erdélyi's generalizations of Bateman's integral [4].

3. Proof of the addition theorems. A corollary of (2.5) may be used to prove (1.5). Let $b=c$ and $\mu=c-b$ in (2.5), and use

$$F(a, b; b; xy)=(1-xy)^{-a},$$

to obtain

(3.1)

$$x^l F(a+l, b+l; c+2l; x)$$
$$=\frac{\Gamma(c+2l)}{\Gamma(b+l)\Gamma(c-b)(a)_l}\int_0^1 y^{b-1}(1-y)^{c-b-1}(1-xy)^{-a}\,F(-l, c+l-1; c-b; 1-y)\,dy,$$
$$c>b>0, \quad l=0, 1, 2, \cdots.$$

This is a generalization of Euler's integral [3, (10), p. 59], to which (3.1) reduces when $l=0$.

Now let $c=2b$, $x=4z(1+z)^{-2}$, $y=\frac{1}{2}(1-\cos\phi)$ in (3.1) to obtain

$$\frac{(4z)^l}{(1+z)^{2l+2a}}F(a+l, b+l; 2b+2l; 4z/(1+z)^2)$$
$$=\frac{\Gamma(2b+2l)}{\Gamma(b+l)\Gamma(b)(a)_l 2^{2b-1}}\int_0^\pi (\sin\phi)^{2b-1}(1+z^2+2z\cos\phi)^{-a}$$
$$\cdot F\left(-l, 2b+l-1; b; \frac{1}{2}(1+\cos\phi)\right)d\phi.$$

Using the quadratic transformation

$$F\left(a+l, b+l; 2b+2l; \frac{4z}{(1+z)^2}\right)=\frac{(1+z)^{2a+2l}}{(1+z^2)^{a+l}}F\left(a+l, 2b-a+l; b+\frac{1}{2}+l; \frac{z^2}{z^2-1}\right),$$

which follows from [3, (24), p. 64] and (1.8), this becomes

$$\frac{(4z)^l}{(1-z^2)^{a+l}}F\left(a+l, 2b-a+l; b+\frac{1}{2}+l; \frac{z^2}{z^2-1}\right)$$
$$=\frac{\Gamma(2b+2l)}{\Gamma(b+l)\Gamma(b)(a)_l 2^{2b-1}}\int_0^\pi (\sin\phi)^{2b-1}[1+z^2+2z\cos\phi]^{-a}$$
$$\cdot F\left(-l, 2b+l-1; b; \frac{1}{2}(1+\cos\phi)\right)d\phi.$$

Then letting $a=-n$, $b=\beta+\frac{1}{2}$, and $z^2=(y-1)(y+1)^{-1}$ gives

$$(y^2-1)^{l/2}R_{n-1}^{(\beta+l,\beta+l)}(y)$$
$$=\frac{\Gamma(2\beta+1+2l)}{\Gamma(\beta+\frac{1}{2}+l)\Gamma(\beta+\frac{1}{2})(-n)_l 2^{2\beta+l}}$$
$$\cdot\int_0^{\pi}[y+i\sqrt{1-y^2}\cos\phi]^n R_l^{(\beta-1/2,\beta-1/2)}(-\cos\phi)(\sin\phi)^{2\beta}\,d\phi,$$

which, apart from an application of Legendre's duplication formula [3, (11), p. 4] and the relation $R_l^{(\beta-1/2,\beta-1/2)}(-\cos\phi)=(-1)^l R_l^{(\beta-1/2,\beta-1/2)}(\cos\phi)$, is (1.5).

This may also be considered a proof of the Gegenbauer addition theorem itself since it is a consequence of (1.5) and Bateman's bilinear sum.

Formula (1.2) now follows easily. Replace k by $k-l$ in (1.5), and apply (1.9) to get

$$(x-1)^{(k+l)/2}(x+1)^{(k-l)/2}R_{n-k}^{(\alpha+k+l,\beta+k-l)}(x)$$
$$=\frac{2^{\alpha-\beta+k}(n-k)!\,\Gamma(\alpha+1+k+l)}{n!\,\Gamma(\frac{1}{2})\Gamma(\beta+\frac{1}{2})\Gamma(\alpha-\beta)(\beta+1+n-l)_l}$$
$$\cdot\int_x^1\frac{(1-y)^{\beta+(k-l)/2}}{(1-x)^{\alpha+(k-l)/2}}\frac{(1+x)^{n+\beta+1+(k-l)/2}}{(1+y)^{n+\alpha+1+(k-l)/2}}(y-x)^{\alpha-\beta-1}R_l^{(\alpha-\beta-1,\beta+k-l)}$$
$$\cdot\left(\frac{1-3y+3x-xy}{(1+y)(1-x)}\right)\int_0^{\pi}(y+i\sqrt{1-y^2}\cos\phi)^n R_{k-l}^{(\beta-1/2,\beta-1/2)}(\cos\phi)(\sin\phi)^n\,d\phi\,dy.$$

This becomes (1.2) after the same change of variables,

$$r^2=\frac{(1-y)(1+x)}{(1+y)(1-x)},$$

as made in [1].

4. Discrete analogues. Gasper [5] found the discrete analogue of Bateman's integral,

$$(4.1)\qquad {}_3F_2\left[\begin{matrix}-x,\,a,\,b\\ c+\mu,\,d\end{matrix};\right]=\sum_{y=0}^{x}\binom{x}{y}\frac{(c)_y(\mu)_{x-y}}{(c+\mu)_x}{}_3F_2\left[\begin{matrix}-y,\,a,\,b\\ c,\,d\end{matrix};\right],$$

where

$${}_3F_2\left[\begin{matrix}-n,\,a,\,b\\ c,\,d\end{matrix};\right]=\sum_{k=0}^{n}\frac{(-n)_k(a)_k(b)_k}{k!(c)_k(d)_k}.$$

Bateman's integral is a limit case of (4.1): replace x, y and d by Nx, Ny and $-N$, respectively, and let $N\to\infty$ to obtain (1.7) with xy replaced with y.

Since the Hahn polynomials $Q_n(x;\alpha,\beta,N)$ are given for $\alpha,\beta>-1$ by

$$(4.2)\qquad Q_n(x;\alpha,\beta,N)={}_3F_2\left[\begin{matrix}-x,\,n+\alpha+\beta+1,\,-n\\ \alpha+1,\,-N\end{matrix};\right],$$

(4.1) gives a projection formula for the Hahn polynomials. The $Q_n(x;\alpha,\beta,N)$ are othogonal with respect to a measure which has jumps at $x=0,1,\cdots,N$ of magnitudes

$$p(x;\alpha,\beta,N)=\binom{N}{x}\frac{(\alpha+1)_x(\beta+1)_{N-x}}{(\alpha+\beta+2)_N}.$$

Gasper also obtained the discrete analogue of (1.6) by applying

$$(4.3)\qquad {}_3F_2\left[\begin{matrix}-n, a, b\\ c, d\end{matrix};\right]=\frac{(d-b)_n}{(d)_n}{}_3F_2\left[\begin{matrix}-n, c-a, b\\ c, 1+b-d-n\end{matrix};\right]$$

to (4.1); (4.3) is a discrete analogue of (1.8) (let $b=Nx$, $d=N$, and let $N\to\infty$) and follows from the $b=c$ case of (4.1).

Formulas (2.5) and (2.6) similarly have discrete analogues. In fact, a formula which gives (2.5) in the same limit as used to obtain (1.7) from (4.1) can be written down almost immediately:

$$\begin{aligned}&\frac{(a)_l(b)_l(-x)_l}{(c+\mu)_{2l}(d)_l}{}_3F_2\left[\begin{matrix}-x+l, a+l, b+l\\ c+\mu+2l, d+l\end{matrix};\right]\\ &(4.4)\qquad =\sum_{y=0}^{x}\binom{x}{y}\frac{(c)_y(\mu)_{x-y}}{(c+\mu)_x}{}_3F_2\left[\begin{matrix}-l, c+\mu+l-1, \mu+x-y\\ \mu, c+\mu+x\end{matrix};\right]{}_3F_2\left[\begin{matrix}-y, a, b\\ c, d\end{matrix};\right],\\ &\qquad\qquad l=0, 1, 2, \cdots.\end{aligned}$$

Of course, the case $l=0$ of (4.4) is (4.1), and when (4.2) is used in (4.4) it becomes a projection formula for the Hahn polynomials. The discrete analogue of (2.6) follows from (4.4) by means of (4.3), just as (2.6) follows from (2.5) by means of (1.8); we omit the details.

Formula (4.4) is perhaps most easily proven by series manipulations. If S represents the sum on the right-hand side of (4.4), then

$$\begin{aligned}S&=\sum_{y=0}^{x}\sum_{j=0}^{l}\binom{x}{y}\frac{(c)_y(\mu)_{x-y}}{(c+\mu)_x}\frac{(-l)_j(c+\mu+l-1)_j(\mu+x-y)_j}{j!(\mu)_j(c+\mu+x)_j}{}_3F_2\left[\begin{matrix}-y, a, b\\ c, d\end{matrix};\right]\\ &=\sum_{j=0}^{l}\frac{(-l)_j(c+\mu+l-1)_j}{j!(c+\mu)_j}\sum_{y=0}^{x}\binom{x}{y}\frac{(c)_y(\mu+j)_{x-y}}{(c+\mu+j)_x}{}_3F_2\left[\begin{matrix}-y, a, b\\ c, d\end{matrix};\right]\\ &=\sum_{j=0}^{l}\frac{(-l)_j(c+\mu+l-1)_j}{j!(c+\mu)_j}{}_3F_2\left[\begin{matrix}-x, a, b\\ c+\mu+j, d\end{matrix};\right],\end{aligned}$$

where the last equality follows from (4.1) with μ replaced with $\mu+j$. Hence, using Vandermonde's theorem, we have

$$\begin{aligned}S&=\sum_{j=0}^{l}\sum_{k=0}^{x}\frac{(-l)_j(c+\mu+l-1)_j(-x)_k(a)_k(b)_k}{j!(c+\mu)_{j+k}k!(d)_k}\\ &=\sum_{k=0}^{x}\frac{(-x)_k(a)_k(b)_k}{k!(c+\mu)_k(d)_k}F(-l; \mu+c+l-1; c+\mu+k; 1)\\ &=\sum_{k=0}^{x}\frac{(-x)_k(a)_k(b)_k(k-l+1)_l}{k!(c+\mu)_k(d)_k(c+\mu+k)_l}.\end{aligned}$$

Finally, since $(k-l+1)_l=0$ if $k=0, 1, \cdots, l-1$, we can let $k=p+l$ in the last sum to get

$$S=\sum_{p=0}^{x-l}\frac{(-x)_{p+l}(a)_{p+l}(b)_{p+l}(p+1)_l}{(p+l)!(c+\mu)_{p+2l}(d)_{p+l}},$$

which, after some simplification, is (4.4).

Formula [2.10] of [5] may be generalized in the same way.

REFERENCES

[1] R. ASKEY, *Jacobi polynomials*, I, *New proofs of Koornwinder's Laplace type integral representation and Bateman's bilinear sum*, this Journal, 5 (1974), pp. 119–124.

[2] R. ASKEY AND J. FITCH, *Integral representations for Jacobi polynomials and some applications*, J. Math. Anal. Appl., 26 (1969), pp. 411–437.

[3] A. ERDÉLYI ET. AL., *Higher Transcendental Functions*, vol. I, McGraw-Hill, New York, 1953.

[4] A. ERDÉLYI, *Transformations of hypergeometric integrals by means of fractional integration by parts*, Quart. J. Math., 10 (1939), pp. 176–189.

[5] G. GASPER, *Projection formulas for orthogonal polynomials of a discrete variable*, J. Math. Anal. Appl., 42 (1973), pp. 438–451.

[6] T. KOORNWINDER, *The addition formula for Jacobi polynomials*, I, *Summary of results*, Indag. Math., 34 (1972), pp. 188–191.

[7] ———, *Jacobi polynomials*, III, *An analytic proof of the addition formula*, this Journal, 6 (1975), pp. 533–543.

[8] ———, *Yet another proof of the addition formula for Jacobi polynomials*, J. Math. Anal. Appl., 61 (1977), pp. 136–141.

SIAM J. MATH. ANAL.
Vol. 13, No. 2, March 1982

0036-1410/82/1302-0013 $01.00/0

A HURWITZ MATRIX IS TOTALLY POSITIVE*

J. H. B. KEMPERMAN†

Abstract. If the real polynomial $f(w)=\sum_0^n d_j w^{n-j}$ with $d_0>0$ has all its zeros in $\mathrm{Re}\,(w)\leqq 0$, then the infinite matrix H with elements $H_{i,j}=d_{2j-i}$ is totally positive. As a consequence, a real polynomial $\sum_j b_j w^j$ has at least $M=\max(\sigma_0,\sigma_1)$ zeros in each half plane $\mathrm{Re}\,(w)<0$ and $\mathrm{Re}\,(w)>0$, where σ_0 and σ_1 denote the number of changes of sign in $\{b_{2j}\}$ and $\{b_{2j-1}\}$, respectively, disregarding zero terms.

1. Main results.

In this paper all polynomials have real coefficients. Let

$$f(w)=d_0w^n+d_1w^{n-1}+\cdots+d_n \tag{1.1}$$

be a fixed polynomial of degree n. We may and will assume that $d_0>0$. If f has all its zeros in the open left half plane $\mathrm{Re}\,(w)<0$ then f is called a *Hurwitz polynomial.* This is because of a well-known criterion due to Hurwitz (see [1] and [3]) stating that $f(w)\neq 0$ throughout $\mathrm{Re}\,(w)\geqq 0$ if and only if

$$\Delta_1>0,\quad \Delta_2>0,\quad \cdots,\quad \Delta_n>0. \tag{1.2}$$

The Δ_p are defined by

$$\Delta_p=\det(d_{2j-i};\ i,j=1,\cdots,p). \tag{1.3}$$

Here and below, we take $d_j=0$ when $j<0$ or $j>n$. For instance, $\Delta_1=d_1$ and

$$\Delta_2=\begin{vmatrix} d_1 & d_3\\ d_0 & d_2\end{vmatrix},\qquad \Delta_3=\begin{vmatrix} d_1 & d_3 & d_5\\ d_0 & d_2 & d_4\\ 0 & d_1 & d_3\end{vmatrix}.$$

Also note that $\Delta_n=d_n\Delta_{n-1}$. Every Hurwitz polynomial f of degree n satisfies

$$d_j>0\quad \text{for } j=0,1,\cdots,n, \tag{1.4}$$

since f is a product of linear factors $ax+b$ and quadratic factors px^2+qx+r, each with positive coefficients. It is known (see [1, p. 196]) that, in the presence of (1.4), f is already a Hurwitz polynomial when $\Delta_{2j}>0$ for $2\leqq 2j\leqq n$ and also when $\Delta_{2j+1}>0$ for $3\leqq 2j+1\leqq n$.

THEOREM 1. *Let f be a Hurwitz polynomial. Then the associated (infinite) "Hurwitz matrix"*

$$H=(H_{i,j};\ i,j\in Z)\quad \textit{with } H_{i,j}=d_{2j-i} \tag{1.5}$$

is totally positive.

Here Z is the set of all integers. The total positivity of H is defined by the condition that

$$H\begin{pmatrix} i_1, i_2, \cdots, i_p\\ j_1, j_2, \cdots, j_p\end{pmatrix}\geqq 0, \tag{1.6}$$

for each choice of the integer $p>0$ and the integers i_r and j_s such that

$$i_1<i_2<\cdots<i_p\quad \text{and}\quad j_1<j_2<\cdots<j_p. \tag{1.7}$$

* Received by the editors May 28, 1980. This research was supported in part by the National Science Foundation.

† Department of Mathematics, University of Rochester, Rochester, New York 14627.

We further use the standard notation

$$H\begin{pmatrix} i_1, i_2, \cdots, i_p \\ j_1, j_2, \cdots, j_p \end{pmatrix} = \det\left(H_{i_r j_s}; \begin{matrix} r = 1, \cdots, p \\ s = 1, \cdots, p \end{matrix} \right)$$

for a subdeterminant of H. Because of (1.7), it has its rows in the same order as H, and similarly for the columns.

Remark 1. After this paper was submitted, we learned that Theorem 1 is actually due to B. A. Asner [6]; his proof is more involved, though it starts along similar lines.

Remark 2. The total positivity of $H = (d_{2j-i})$ carries over to the case where f has all its zeros in the *closed* left half plane $\mathrm{Re}\,(w) \leqq 0$, (while $d_0 > 0$): apply Theorem 1 to the Hurwitz polynomial $f_\varepsilon(z) = f(z+\varepsilon)$ and then let $\varepsilon \downarrow 0$.

Observe that condition (1.2) for a Hurwitz polynomial can be written as

$$\Delta_p = H\begin{pmatrix} 1, 2, \cdots, p \\ 1, 2, \cdots, p \end{pmatrix} > 0 \quad \text{for } p = 1, 2, \cdots, n.$$

If $p > n$ then $\Delta_p = 0$ due to a column of zeros.

The following result gives a more precise description of the class of nonzero subdeterminants of $H = (d_{2j-i})$.

THEOREM 2. *Let f be a Hurwitz polynomial and let H be the associated Hurwitz matrix $H = (d_{2j-i})$ as in* (1.5). *Let further i_r and j_r $(r = 1, \cdots, p)$ be given integers satisfying* (1.7). *Then in order that*

$$H\begin{pmatrix} i_1, i_2, \cdots, i_p \\ j_1, j_2, \cdots, j_p \end{pmatrix} > 0, \tag{1.8}$$

it is necessary and sufficient that all the diagonal elements $d_{2j_r - i_r}$ $(r = 1, \cdots, p)$ are positive, equivalently, that

$$0 \leqq 2j_r - i_r \leqq n \quad \text{for } r = 1, \cdots, p. \tag{1.9}$$

If (1.9) *fails to hold then the determinant of* (1.6) *is identically zero.*

Comments. Note that condition (1.9) of Theorem 2 depends only on the indices i_r, j_r and not on the values d_j themselves. The special case $p = 2$, $i_1 = j_1 = j$ and $i_2 = j_2 = j+1$ yields that each Hurwitz polynomial f satisfies

$$d_j/d_{j-1} > d_{j+2}/d_{j+1} \quad \text{when } 0 \leqq j \leqq n-1. \tag{1.10}$$

Thus, d_j/d_{j-1} is a decreasing function of j as long as j remains of the same parity (even or odd). It follows that there exist integers s and t with $0 \leqq s \leqq (n-1)/2$ and $-1 \leqq t \leqq (n-1)/2$ and such that

$$d_{2j} \geqq d_{2j-1} \quad \text{for } j \leqq s, \qquad d_{2j} \leqq d_{2j-1} \quad \text{for } j > s$$

and

$$d_{2j+1} \geqq d_{2j} \quad \text{for } j \leqq t, \qquad d_{2j+1} \leqq d_{2j} \quad \text{for } j > t.$$

Consequently, if f is a Hurwitz polynomial then the corresponding sequence of coefficients $\{d_j\}$ is increasing for $j \leqq \min(2s, 2t+1)$; it is decreasing for $j \geqq \max(2s+1, 2t+2)$; and in the intermediate range it has a DUDUDU $\cdots$ type of

behavior (D = down; U = up). For example,

$$(3w^2+w+3)^4$$
$$=81w^8+108w^7+378w^6+336w^5+595w^4+336w^3+378w^2+108w+81$$

is a Hurwitz polynomial with the behavior UUUDUDUDDD for the coefficients d_j between $d_{-1}=0$ and $d_{n+1}=d_9=0$. In a forthcoming paper [7], this up–down behavior of $\{d_j\}$ will be used to prove for any probability measure μ on $\{0, 1, 2\}$ with $\mu(\{1\})>0$ that, for n sufficiently large, the self-convolution $\mu^n=\mu*\mu*\cdots*\mu$ is unimodal over the *full* range $[0, 2n]$.

It is useful to associate to the polynomial (1.1) also the infinite matrix

$$P=(P_{i,j}=d_{j-1};\ i, j\in Z).$$

One has $H_{i,j}=d_{2j-i}=P_{i,2j}$. Thus, the inequality (1.6) refers to a special subdeterminant of the matrix P, namely, one with columns of even index only.

It is well known (see [4, Ch. 8]) that P itself is totally positive if and only if the polynomial (1.1) has all its zeros on the negative real axis $(-\infty, 0]$. A more precise result due to Schoenberg (see [4, p. 397]) states that the polynomial (1.1), with $d_0>0$ and $d_n>0$, has all its zeros in the sector

$$|\arg z|\geqq \pi r/(n+r-1)$$

about the negative real axis as soon as

$$P\begin{pmatrix} i_1, i_2, \cdots, i_p \\ j_1, j_2, \cdots, j_p \end{pmatrix}\geqq 0,$$

whenever $p\leqq r$ and (1.7) hold. Here, r is a fixed positive integer. This result is sharp in the sense that there exist examples where the polynomial f has a zero on the boundary $|\arg z|=\pi r/(n+r-1)$.

2. A simple lower bound on the number of zeros in a half plane. If $b_1, \cdots, b_m$ is a sequence of real numbers, then by $V[b_1, \cdots, b_m]$ we shall denote the number of changes of sign in the sequence which results from $b_1, b_2, \cdots, b_m$ by deleting all zero terms. By $V^+[b_1, \cdots, b_m]$ we denote the largest possible number of sign changes in any sequence obtained from $b_1, b_2, \cdots, b_m$ by replacing each zero term by a nonzero number (such as -1 or $+1$). For instance, $V^+[1, 0, 1, 0]=3$ and $V^+[1, 0, 1, 0, 0, -1]=5$.

Using Theorem 1, we shall prove the following result. Here

$$\varphi(w)=a_0w^n+a_1w^{n-1}+\cdots+a_n, \qquad a_0\neq 0 \tag{2.1}$$

is a polynomial of degree n with real coefficients. Let further

$$\begin{aligned} \sigma_0&=V[a_0, a_2, a_4, \cdots, a_{2[n/2]}], \\ \sigma_1&=V[a_1, a_3, a_5, \cdots, a_{2[(n-1)/2]+1}]. \end{aligned} \tag{2.2}$$

THEOREM 3. *The polynomial φ has at least*

$$M=\max(\sigma_0, \sigma_1) \tag{2.3}$$

zeros in the open left half plane $\operatorname{Re}(w)<0$ *(counting multiplicities) and also at least M zeros in the open right half plane* $\operatorname{Re}(w)>0$.

COROLLARY. *Let*

$$\sigma_0^+ = V^+[a_0, a_2, a_4, \cdots, a_{2[n/2]}],$$
$$\sigma_1^+ = V^+[a_1, a_3, a_5, \cdots, a_{2[(n-1)/2]+1}]. \tag{2.4}$$

Then the polynomial φ has at least

$$M^+ = \max(\sigma_0^+, \sigma_1^+) \tag{2.5}$$

zeros in the closed half plane $\operatorname{Re}(w) \leqq 0$ *and also at least* M^+ *zeros in* $\operatorname{Re}(w) \geq 0$. *More precisely, this assertion holds relative to at least one way of assigning each of the purely imaginary zeros to one of the two half planes (counting each zero according to its multiplicity).*

Proof of corollary. Perturbing the zero coefficients a_j in (2.1) by a small amount of suitable sign, one arrives at a polynomial $\tilde{\varphi}$ with $\tilde{\sigma}_0 = \sigma_0^+$ and $\tilde{\sigma}_1 = \sigma_1^+$. It follows from Theorem 3 that $\tilde{\varphi}$ has at least M^+ zeros in $\operatorname{Re}(w) < 0$ and also at least M^+ zeros in $\operatorname{Re}(w) > 0$. The assertion now follows by the continuity of the zeros as a function of the coefficients (see [5, p. 4)].

For a good understanding, let n_-, n_0 and n_+ denote the number of zeros (counting multiplicities) of $\varphi(w)$ in $\operatorname{Re}(w) < 0$, $\operatorname{Re}(w) = 0$ and $\operatorname{Re}(w) > 0$, respectively; thus $n_- + n_0 + n_+ = n$. Theorem 3 asserts that

$$\min(n_-, n_+) \geqq \max(\sigma_0, \sigma_1), \tag{2.6}$$

and the corollary yields that

$$\max_{0 \leqq \lambda \leqq 1} \min[n_- + \lambda n_0, n_+ + (1-\lambda) n_0] \geqq \max(\sigma_0^+, \sigma_1^+). \tag{2.7}$$

In particular,

$$\text{if } n_+ \geqq n_- + n_0 \quad \text{then } n_- + n_0 \geqq \max(\sigma_0^+, \sigma_1^+). \tag{2.8}$$

These assertions are most interesting in a situation where n_0 is somehow known to be small.

As an illustration of Theorem 3, if the sequence $\{a_{2j}\}$ of even coefficients has $\sigma_0 = 5$ changes of sign, then it follows that $\varphi(w)$ has at least 5 zeros in $\operatorname{Re}(w) < 0$ and also at least 5 zeros in $\operatorname{Re}(w) > 0$, *irrespective* of the behavior of the sequence $\{a_{2j+1}\}$ of odd coefficients.

This curious result is somewhat opposed to the Descartes rule of signs, which states that the number of real and positive zeros of $\varphi(w)$ (counting multiplicities) is equal to $\sigma - 2h$, where $2h$ denotes an even nonnegative integer, while

$$\sigma = V[a_0, a_1, a_2, \cdots, a_{n-1}, a_n] \tag{2.9}$$

denotes the *total number of changes of sign* in $\{a_j\}$ disregarding zeros.

Observe from (2.4) that

$$\sigma \leqq \sigma_0 \leqq [n/2], \qquad 0 \leqq \sigma_1 \leqq [(n-1)/2]. \tag{2.10}$$

Moreover, every pair of integers σ_0, σ_1 satisfying (2.10) can easily be realized by a suitable choice of the coefficients a_{2j} and a_{2j+1}. In that sense σ_0 and σ_1 do not "influence" each other. However, if the total number σ of changes of sign in $\{a_j\}$ is known, then there is some influence. For, if $0 < \sigma < n$ then it is impossible that both $\sigma_0 = 0$ and $\sigma_1 = 0$. The following theorem gives a precise description of all the possible combinations of σ_0, σ_1, and σ when the a_j are nonzero.

DEFINITION. Let S_n denote the collection of all triplets $(\sigma_0, \sigma_1, \sigma)$ of nonnegative integers which can be realized as in (2.2) and (2.9) by a suitable choice of the *nonzero* coefficients $a_0, a_1, \cdots, a_{n-1}, a_n$. One might as well restrict the a_j to $a_j \in \{-1, +1\}$, $j = 0, 1, \cdots, n$.

THEOREM 4. *Let $\sigma_0, \sigma_1, \sigma$ and n be nonnegative integers, $n > 0$. Then one has*

$$(\sigma_0, \sigma_1, \sigma) \in S_n \tag{2.11}$$

if and only if all of the following are true:

(i) $\sigma_0 \leqq n/2$ *and* $\sigma_1 \leqq (n-1)/2$;
(ii) $\sigma_0 \leqq \sigma$ *and* $\sigma_1 \leqq \sigma$;
(iii) $\sigma_0 + \sigma \leqq n$ *and* $\sigma_1 + \sigma \leqq n$;
(iv) *if n is even then $\sigma - \sigma_0$ is even*;
(v) *if $\sigma_0 = 0$ and $\sigma_1 = 0$ then either $\sigma = 0$ or $\sigma = n$.*

Observe that (ii) and (iii) yield that, for every triplet in S_n,

$$\max(\sigma_0, \sigma_1) \leqq \sigma \leqq n - \max(\sigma_0, \sigma_1).$$

Equivalently, (2.11) implies that

$$|\sigma - n/2| \leqq n/2 - \max(\sigma_0, \sigma_1). \tag{2.12}$$

It also follows from Theorem 4 that $S_k \subset S_n \cup \{(0, 0, k)\}$ if $1 \leqq k \leqq n$ and either n is odd or k is even.

It would also be interesting to determine the precise structure of the collection S_n^0 of all triplets $(\sigma_0, \sigma_1, \sigma)$ which can be realized by some choice of $a_0, a_1, \cdots, a_n$, this time allowing that $a_j = 0$. Clearly, $S_k \subset S_k^0 \subset S_n^0$ when $1 \leqq k \leqq n$; in particular, $(0, 0, k) \in S_n^0$ for all $0 \leqq k \leqq n$. Also of interest would be the set of possible triplets $(\sigma_0^+, \sigma_1^+, \sigma)$ as in (2.4), (2.9).

3. Proof of Theorems 1 and 2. We shall need the following result, to be proved below.

LEMMA 1. *Let f be a given Hurwitz polynomial of degree $n \geqq 1$ as in* (1.1). *Then there exist a unique Hurwitz polynomial*

$$f_1(w) = d_0' w^{n-1} + d_1' w^{n-2} + \cdots + d_{n-2}' w + d_{n-1}' \tag{3.1}$$

of degree $n-1$ and a unique constant c such that

$$d_{2j+1} = d_{2j}' \tag{3.2}$$

and

$$d_{2j} = d_{2j-1}' + c d_{2j}' \tag{3.3}$$

for any integer j. Here $d_j' = 0$ when $j < 0$ or $j \geqq n$. Note that $d_0' = d_1 > 0$ and

$$c = d_0/d_0' = d_0/d_1 > 0. \tag{3.4}$$

Proof of Theorems 1 *and* 2. The proof will be by induction with respect to n. Suppose first that $n = 0$, in which case $d_0 > 0$ and $d_j = 0$ for $j \neq 0$. Let $p \geqq 1$, i_r and j_r be as in (1.7) and put

$$\Delta = \det\left(H_{i_r, j_s} = d_{2j_s - i_r};\ \begin{matrix} r = 1, \cdots, p \\ s = 1, \cdots, p \end{matrix}\right). \tag{3.5}$$

One has $\Delta = 0$ unless each row contains at least one nonzero element, that is, unless to each row index i_r there corresponds a column index j_s with $2j_s - i_r = 0$. By (1.7) this

means that $i_r = 2j_r$, $r = 1, \cdots, p$, which is precisely condition (1.9) (with $n = 0$). And in this same case $\Delta = d_0^p > 0$.

All assertions are trivially true when $p = 1$. Let $n \geqq 1$ and $p \geqq 2$ be fixed and suppose the assertions of Theorems 1 and 2 hold when n is replaced by $n-1$. Further, let f be a Hurwitz polynomial of degree n as in (1.1).

We now apply Lemma 1. From the induction assumption, the matrix

$$H' = (H'_{i,j} = d'_{2j-i};\ i, j \in Z)$$

satisfies all the assertions of Theorems 1 and 2, provided n is replaced by $n-1$ and d_j by d'_j. Introducing

$$c(i) = \begin{cases} c & \text{if } i \text{ is even,} \\ 0 & \text{if } i \text{ is odd,} \end{cases} \tag{3.6}$$

$i \in Z$, we see from (3.2) and (3.3) that

$$H_{i,j} = d_{2j-i} = d'_{2j-i-1} + c(i) d'_{2j-i} = H'_{i+1,j} + c(i) H'_{i,j}.$$

Consequently, the determinant (3.5) equals

$$\Delta = \det\left(H'_{i_r+1,j_s} + c(i_r) H'_{i_r,j_s};\ \begin{matrix} r = 1, \cdots, p \\ s = 1, \cdots, p \end{matrix}\right). \tag{3.7}$$

We are still assuming (1.7); in particular, $i_{r-1} + 1 \leqq i_r$ for $r = 2, \cdots, p$. Thus the $i'_r \in \{i_r, i_r + 1\}$, $r = 1, \cdots, p$, always satisfy $i'_1 \leqq i'_2 \leqq \cdots \leqq i'_p$. Moreover, we know from the induction assumption that the matrix $H' = (H'_{i,j})$ is totally positive. Therefore, expanding the determinant in (3.7) in powers of c, one sees immediately that $\Delta \geqq 0$, showing that $H = (H_{i,j})$ is totally positive.

In order that $\Delta > 0$, it is at least necessary that (1.9) holds. For, if ever $i_r > 2j_r$, then also $i_r + 1 > 2j_r$. If ever $2j_r - i_r > n$, then $2j_r - i_r > 2j_r - (i_r + 1) > n - 1$. In each case one has $\Delta = 0$, by (3.7) and the induction assumption.

Conversely, suppose that (1.9) holds. From (3.6), (3.7) and the induction assumption, in showing that $\Delta > 0$ it suffices to prove that $i'_r \in \{i_r, i_r + 1\}$, $r = 1, \cdots, p$, can be found in such a way that

$$0 \leqq 2j_r - i'_r \leqq n - 1 \quad \text{for } r = 1, \cdots, p, \tag{3.8}$$

and that further $i'_1 < i'_2 < \cdots < i'_p$. It is also necessary that $i'_r = i_r + 1$ each time that i_r is odd.

In fact, let us choose $i'_r = i_r + 1$ unless $i_r = 2j_r$ in which case we choose $i'_r = i_r$, $r = 1, \cdots, p$. Then (3.8) is an immediate consequence of (1.9). Next consider the inequality $i'_{r-1} < i'_r$ with $1 < r \leqq p$. This inequality is obvious when $i'_r = i_r + 1$. If not then $i'_r = i_r = 2j_r$ and

$$i'_{r-1} \leqq i_{r-1} + 1 \leqq 2j_{r-1} + 1 \leqq 2(j_r - 1) + 1 = i'_r - 1.$$

This completes the proof of Theorems 1 and 2. □

Remark 3. If $n = 2m + 1$ is odd, then in the induction from n to $n-1$, one could also use the fact that $f(w) = (w + a) f_1(w)$ with $a > 0$ and f_1 as a Hurwitz polynomial of degree $n-1$ as in (3.1). Hence, $d_j = d'_j + a d'_{j-1}$ and $H_{i,j} = H'_{i,j} + a H'_{i+1,j}$, for all i and j.

Proof of Lemma 1. Take $c = d_0/d_1$; thus $c > 0$. Further define

$$d'_{2j} = d_{2j+1}, \qquad d'_{2j-1} = d_{2j} - c d_{2j+1}, \tag{3.9}$$

for each integer j. Then $d'_j = 0$ if $j < 0$ or $j \geqq n$; in particular, $d'_{-1} = d_0 - c d_1 = 0$.

It suffices to show that the resulting polynomial $f_1(w)$ in (3.1) is a Hurwitz polynomial of degree $n-1$. This fact is actually implicit in the literature and is the underlying idea behind the Routh algorithm (see [1, pp. 157, 204]).

In view of Hurwitz's criterion (1.2) applied to f_1 instead of f, it would be sufficient to show that $\Delta'_p>0$, $p=1,\cdots,n-1$, where

$$\Delta'_p=\det\left(H'_{i,j}=d'_{2j-i};\ \begin{matrix}i=1,\cdots,p\\ j=1,\cdots,p\end{matrix}\right).$$

Thus, it suffices to show that

$$\Delta_p=d_1\Delta'_{p-1},\qquad p=2,3,\cdots,n,$$

with Δ_p as in (1.3). This is easily seen by starting with the determinant Δ_p, subtracting c times the first row from the second row, c times the third row from the fourth, and so on. See [1, p. 169] for related results.

4. Proof of Theorems 3 and 4. In the proof of Theorem 3 we will need the following result due to Gantmacher and Krein [2]; see also Karlin [4, p. 223].

LEMMA 2. *Suppose $Q=(Q_{ij})$ is a totally positive $k\times m$ matrix. Let $c_1,\cdots,c_m$ be real numbers and define*

$$b_i=\sum_{j=1}^{m}Q_{i,j}c_j\quad \textit{for } i=1,\cdots,k. \tag{4.1}$$

Then

$$V[b_1,b_2,\cdots,b_k]\leqq V[c_1,c_2,\cdots,c_m]. \tag{4.2}$$

Proof of Theorem 3. Let n_- and n_+ denote the number of zeros of $\varphi(w)$ (counting multiplicities) with Re $(w)<0$ and Re $(w)>0$, respectively. It suffices to show that

$$n_+\geqq\sigma_0\quad\text{and}\quad n_+\geqq\sigma_1. \tag{4.3}$$

For, afterwards, applying (4.3) to $\tilde{\varphi}(w)=\varphi(-w)$ (which polynomial has precisely the same values σ_0 and σ_1), it also follows that $n_-\geqq\sigma_0$ and $n_-\geqq\sigma_1$. For brevity, let $q=n_+$.

One can factorize $\varphi(w)$ as

$$\varphi(w)=f(w)g(w), \tag{4.4}$$

where

$$f(w)=d_0w^{n-q}+d_1w^{n-q-1}+\cdots+d_{n-q}$$

has all its zeros in Re $(w)\leqq 0$, while

$$g(w)=e_0w^q+e_1w^{q-1}+\cdots+e_q$$

has all its zeros in Re $(w)>0$. Comparing (2.1) and (4.4), one has

$$a_j=\sum_{i=0}^{q}d_{j-i}e_i,\qquad j=0,1,\cdots,n, \tag{4.5}$$

provided $d_j=0$ when $j<0$ or $j>n-q$. In particular,

$$a_{2j}=\sum_{i=0}^{q}d_{2j-i}e_i\quad\text{for } j=0,1,\cdots,[n/2] \tag{4.6}$$

and

$$a_{2j-1} = \sum_{i=1}^{q+1} d_{2j-i} e_{i-1} \quad \text{for } j = 1, 2, \cdots, [(n+1)/2]. \tag{4.7}$$

Since f has all its zeros in the closed half plane $\mathrm{Re}\,(w) \leqq 0$, we have, from Theorem 1 and remarks following it, that the matrix

$$H = (H_{i,j} = d_{2j-i};\ i, j \in Z)$$

is totally positive. It follows from Lemma 2 and (4.6) that

$$\sigma_0 = V[a_0, a_2, \cdots, a_{2[n/2]}] \leqq V[e_0, e_1, \cdots, e_q] \leqq q.$$

Similarly, (4.7) yields that $\sigma_1 \leqq q$.

By the way, since $g(-w)$ is a Hurwitz polynomial of degree q it has all its coefficients positive so that $V[e_0, e_1, \cdots, e_q] = q$.

Proof of Theorem 4. Let Σ_n denote the collection of all 2^{n+1} sequences

$$\varepsilon = (\varepsilon_0, \varepsilon_1, \cdots, \varepsilon_n)$$

of length $n+1$ with $\varepsilon_i \in \{-1, +1\}$. To each ε in Σ_n we associate the numbers

$$\sigma_0 = V[\varepsilon_0, \varepsilon_2, \cdots, \varepsilon_{2[n/2]}],$$
$$\sigma_1 = V[\varepsilon_1, \varepsilon_3, \cdots, \varepsilon_{2[(n-1)/2]+1}],$$
$$\sigma = V[\varepsilon_0, \varepsilon_1, \cdots, \varepsilon_{n-1}, \varepsilon_n].$$

We will say that this triplet $(\sigma_0, \sigma_1, \sigma)$ corresponds to the sequence ε and also that ε is a realization of this triplet.

The set S_n is defined as the collection of all triplets which correspond to at least one sequence ε in Σ_n. One easily verifies Theorem 4 in the special case $n = 1$. Namely, S_1 consists of the two triplets $(0, 0, 0)$ and $(0, 0, 1)$. These are realized by $(++)$ and $(+-)$, respectively. Similarly, S_2 consists of the three triplets $(0, 0, 0)$, $(0, 0, 2)$ and $(1, 0, 1)$. These are realized by $(+++)$, $(+-+)$ and $(-++)$, respectively. The triplet $(1, 0, 1)$ also has a realization which ends with a change of sign, namely, $(++-)$.

The proof of Theorem 4 uses an induction on n. To enable the induction to proceed, we will prove a little more, namely, the additional validity of the following two assertions.

ASSERTION (A). *Suppose n is odd and let $(\sigma_0, \sigma_1, \sigma) \in S_n$ be realized by $\varepsilon = (\varepsilon_0, \varepsilon_1, \cdots, \varepsilon_{n-1}, \varepsilon_n)$ in Σ_n. Then ε must end with a stay $(\varepsilon_n = \varepsilon_{n-1})$ or a change of sign $(\varepsilon_n \neq \varepsilon_{n-1})$ depending on whether the difference $\sigma - \sigma_0$ is even or odd, respectively.*

ASSERTION (B). *Suppose n is even. Then $(\sigma_0, \sigma_1, \sigma) \in S_n$ possesses a realization $\varepsilon = (\varepsilon_0, \cdots, \varepsilon_{n-1}, \varepsilon_n)$ in Σ_n ending with a stay if and only if $\sigma_1 + \sigma < n$. Similarly, this same triplet possesses a realization in Σ_n ending with a change of sign if and only if $\sigma_1 < \sigma$.*

We now begin with the main body of the proof.

Necessity. Consider a sequence $\varepsilon = (\varepsilon_0, \varepsilon_1, \cdots, \varepsilon_n)$ in Σ_n with corresponding triplet $(\sigma_0, \sigma_1, \sigma)$.

The necessity of the properties (i) and (ii) of Theorem 4 is obvious. As to property (v), if $\sigma_0 = 0$ and $\sigma_1 = 0$ then $\varepsilon_{2j} = \varepsilon_0$ and $\varepsilon_{2j+1} = \varepsilon_1$ for all j, so that either $\sigma = 0$ or $\sigma = n$, depending on whether or not $\varepsilon_0 = \varepsilon_1$. In this special case Assertions (A) and (B) are easily verified.

Let us introduce the quantities

$$\xi_i = \begin{cases} 1 & \text{if } \varepsilon_{2i-2}\varepsilon_{2i} < 0, \\ 0 & \text{otherwise,} \end{cases} \qquad 1 \leqq i \leqq n/2,$$

$$\eta_j = \begin{cases} 1 & \text{if } \varepsilon_{2j-1}\varepsilon_{2j+1} < 0, \\ 0 & \text{otherwise,} \end{cases} \qquad 1 \leqq j \leqq (n-1)/2,$$

$$\zeta_k = \begin{cases} 1 & \text{if } \varepsilon_{k-1}\varepsilon_k < 0, \\ 0 & \text{otherwise,} \end{cases} \qquad 1 \leqq k \leqq n.$$

For instance, $\zeta_n = 0$ or 1 depending on whether the given sequence ε ends with a stay or a change of sign, respectively. Clearly,

$$\sigma_0 = \sum_{i=1}^{[n/2]} \xi_i, \quad \sigma_1 = \sum_{j=1}^{[(n-1)/2]} \eta_j, \quad \sigma = \sum_{k=1}^{n} \zeta_k.$$

Also observe that

$$\zeta_{2i-1} + \zeta_{2i} = \begin{cases} 0 & \text{implies } \xi_i = 0, \\ 1 & \text{implies } \xi_i = 1, \\ 2 & \text{implies } \xi_i = 0. \end{cases}$$

Thus $(\zeta_{2i-1} + \zeta_{2i} \pm \xi_i)$ can only take the values 0 or 2. If n is odd then

$$\sigma + \sigma_0 = \sum_{i=1}^{[n/2]} (\zeta_{2i-1} + \zeta_{2i} + \xi_i) + \zeta_n,$$

hence, $\sigma + \sigma_0 \leqq 2[n/2] + 1 = n$. Moreover, $\zeta_n = 0$ or 1 depending on whether $\sigma + \sigma_0$ is even or odd, respectively; equivalently, $\sigma - \sigma_0$ is even or odd, respectively. This proves Assertion (A).

If n is even then the above term ζ_n is missing so that $\sigma + \sigma_0 \leqq 2[n/2] = n$. In this same case, $\sigma + \sigma_0$ and thus $\sigma - \sigma_0$ is even which proves the necessity of property (iv) of Theorem 4.

If n is even then

$$\sigma + \sigma_1 = \sum_{j=1}^{[(n-1)/2]} (\zeta_{2j} + \zeta_{2j+1} + \eta_j) + \zeta_1 + \zeta_n,$$

where $(\zeta_{2j} + \zeta_{2j+1} \pm \eta_j)$ can take only the values 0 and 2. Thus, $\sigma + \sigma_1 \leqq 2[(n-1)] + 2 = n$. The inequality $\sigma + \sigma_1 \leqq n$ with n odd follows in a similar way, proving the necessity of (iii).

If n is even then $\sigma + \sigma_1 = n$ is only possible when $\zeta_1 = \zeta_n = 1$. Thus $\zeta_n = 0$ implies that $\sigma_1 + \sigma < n$, which is part of Assertion (B).

Similarly, if n is even then

$$\sigma - \sigma_1 = \sum_{j=1}^{[(n-1)/2]} (\zeta_{2j} + \zeta_{2j+1} - \eta_j) + \zeta_1 + \zeta_n \geqq 0.$$

Here, the equality sign is only possible when $\zeta_1 = \zeta_n = 0$. In particular, $\zeta_n = 1$ implies that $\sigma_1 < \sigma$. This establishes the necessity part of Assertion (B).

Sufficiency. It remains to prove that the stated conditions (i)–(v) are also sufficient for a triplet $(\sigma_0, \sigma_1, \sigma)$ to belong to S_n. We must also prove the existence part of Assertion (B). The proof is by induction with respect to n. Let $n \geqq 2$ be fixed and assume that Theorem 4 as well as Assertion (B) is true when n is replaced by $n-1$;

(Assertion (A) has been proved already). In particular, a triplet $(\sigma_0, \sigma_1, \sigma)$ *is realized by at least one* $\varepsilon = (\varepsilon_0, \cdots, \varepsilon_{n-1})$ in Σ_{n-1} as soon as it satisfies properties (i)–(v) with n replaced by $n-1$.

Consider a fixed triplet $(\sigma_0, \sigma_1, \sigma)$ of nonnegative integers satisfying (i)–(v) of Theorem 4. We must show that it can be realized by some $\varepsilon = (\varepsilon_0, \cdots, \varepsilon_{n-1}, \varepsilon_n)$ in Σ_n. In the case that n is even and $\sigma_1 + \sigma < n$ we must also show that this realization can be chosen such that $\zeta_n = 0$, that is, $\varepsilon_{n-1} = \varepsilon_n$ (this in view of Assertion (B)). Similarly, if n is even and $\sigma_1 < \sigma$ there must exist a realization with $\varepsilon_{n-1} \neq \varepsilon_n$.

We have already treated the case that $\sigma_0 = \sigma_1 = 0$; thus, assume that $\sigma_0 + \sigma_1 \geqq 1$. Let us first consider the case $\sigma_0 + \sigma_1 = 1$. To begin with let $\sigma_0 = 0$ and $\sigma_1 = 1$ so that $1 \leqq \sigma \leqq n-1$.

If $n = 2m$ is even then $\sigma - \sigma_0 = \sigma$ is even, by (iv), thus $\sigma = 2h$ with $1 \leqq h \leqq m-1$. Hence, the triplet on hand is $(0, 1, 2h)$ and it can be realized in Σ_n by choosing $\varepsilon_{2j} = 1$ for all j, further $\varepsilon_{2j-1} = -1$ for $j = 1, \cdots, h$ and $\varepsilon_{2j-1} = +1$ for $j = h+1, \cdots, m$. Observe that $\zeta_n = 0$ in this case. One can also attain that $\zeta_n = 1$, namely, by choosing $\varepsilon_{2(m-j)+1} = -1$ for $j = 1, \cdots, h$ and $\varepsilon_i = +1$, otherwise.

Next, consider the case $\sigma_0 = 0$, $\sigma_1 = 1$ with $n = 2m+1$ odd. One has $1 \leqq \sigma \leqq n-1 = 2m$. If $\sigma = 2h-1$ is odd $(1 \leqq h \leqq m)$ then a realization in Σ_n of $(\sigma_0, \sigma_1, \sigma) = (0, 1, 2h-1)$ is obtained by choosing $\varepsilon_{2(m-j)+1} = -1$, $j = 0, 1, \cdots, h-1$, and $\varepsilon_i = +1$, otherwise. If $\sigma = 2h$ is even $(1 \leqq h \leqq m)$ then choose $\varepsilon_{2j-1} = -1$, $j = 1, \cdots, h$, and $\varepsilon_i = +1$, otherwise.

Next, suppose that $\sigma_0 = 1$ and $\sigma_1 = 0$. The case with $n = 2m+1$ odd can be reduced to the previous case by means of the transformation

$$\varepsilon_i' = \varepsilon_{2m+1-i} \quad \text{for } i = 0, 1, \cdots, 2m+1. \tag{4.8}$$

For, the sequence ε in Σ_n is a realization of the triplet (x, y, z) if and only if the transformed sequence ε' in Σ_n is a realization of the triplet (y, x, z).

Finally, consider the case $\sigma_0 = 1$, $\sigma_1 = 0$ with $n = 2m$ even. Then $\sigma - \sigma_0 = \sigma - 1$ is even, by (iv), so that $\sigma = 2h-1$ with $1 \leqq h \leqq m$. A realization in Σ_n of $(1, 0, 2h-1)$ with $\zeta_n = 0$ is given by $\varepsilon_{2j} = -1$ if $j = 0, 1, \cdots, h-1$ and $\varepsilon_i = +1$, otherwise. A realization with $\zeta_n = 1$ is given by $\varepsilon_{2(m-j)} = -1$ if $j = 0, 1, \cdots, h-1$ and $\varepsilon_i = +1$, otherwise.

It remains to consider the case $\sigma_0 + \sigma_1 \geqq 2$. Suppose first that $n = 2m$ is *even* so that $\sigma - \sigma_0$ is even, by (iv). Since $\sigma_1 \leqq (n-1)/2$ one has

$$(n - \sigma_1 - \sigma) + (\sigma - \sigma_1) = n - 2\sigma_1 > 0.$$

Thus, *either* $\sigma_1 + \sigma < n$ *or* $\sigma_1 < \sigma$.

Consider the case $\sigma_1 + \sigma < n$. Suppose first that $\sigma_0 = 0$. One easily verifies that the triplet $(\sigma_0, \sigma_1, \sigma) = (0, \sigma_1, \sigma)$ satisfies (i)–(v) with n replaced by $n-1$ (in particular, $\sigma_1 \leqq [(n-1)/2] = (n-2)/2$), so that $(0, \sigma_1, \sigma) \in S_{n-1}$ from the induction assumption. But $\sigma = \sigma - \sigma_0$ is even while $n-1$ is odd. Hence, by Assertion (A), every $(\varepsilon_0, \cdots, \varepsilon_{n-2}, \varepsilon_{n-1})$ in Σ_{n-1} realizing $(0, \sigma_1, \sigma)$ satisfies $\varepsilon_{n-2} = \varepsilon_{n-1}$. Choosing $\varepsilon_n = \varepsilon_{n-1}$ one arrives at a realization of $(0, \sigma_1, \sigma)$ in Σ_n which ends with a stay, in agreement with Assertion (B).

Next, suppose $\sigma_0 \geqq 1$ while still $n = 2m$ and $\sigma_0 + \sigma_1 < n$. One easily verifies that $(\sigma_0 - 1, \sigma_1, \sigma) \in S_{n-1}$. Here $\sigma - (\sigma_0 - 1)$ is odd, hence, from (A), every realization $(\varepsilon_0, \cdots, \varepsilon_{n-2}, \varepsilon_{n-1})$ in Σ_{n-1} of the latter triplet has ε_{n-1} and ε_{n-2} of opposite sign. Choosing $\varepsilon_n = \varepsilon_{n-1}$, one obtains a realization in Σ_n of $(\sigma_0, \sigma_1, \sigma)$ which ends with a stay, again in agreement with (B).

Further consider the case $n = 2m$ with $\sigma_1 < \sigma$. Suppose first that $\sigma_0 = 0$. Then $(0, \sigma_1, \sigma - 1) \in S_{n-1}$ with $(\sigma - 1) - \sigma_0 = \sigma - 1$ odd. It has a realization $(\varepsilon_0, \cdots, \varepsilon_{n-2}, \varepsilon_{n-1})$ in Σ_{n-1} necessarily with ε_{n-1} and ε_{n-2} of opposite sign, by (A).

Choosing $\varepsilon_n = \varepsilon_{n-2}$, one obtains a realization in Σ_n of $(0, \sigma_1, \sigma)$ ending with a change of sign, in agreement with (B).

Next, let $n = 2m$, $\sigma_1 < \sigma$ and $\sigma_0 \geqq 1$. Then $(\sigma_0 - 1, \sigma_1, \sigma - 1) \in S_{n-1}$. Here $(\sigma - 1) - (\sigma_0 - 1) = \sigma - \sigma_0$ is even, hence, by (A), each realization $(\varepsilon_0, \cdots, \varepsilon_{n-2}, \varepsilon_{n-1})$ in Σ_{n-1} of the latter triplet has $\varepsilon_{n-2} = \varepsilon_{n-1}$. Choosing $\varepsilon_n = -\varepsilon_{n-1}$, one obtains a realization in Σ_n of $(\sigma_0, \sigma_1, \sigma)$ ending with a change of sign, in agreement with (B).

Let us now turn to the case where $n = 2m + 1$ is *odd*, still assuming that $\sigma_0 + \sigma_1 \geqq 2$. In view of the transformation (4.8), one may as well assume that $\sigma_1 \geqq \sigma_0$, thus $\sigma_1 \geqq 1$.

Suppose $\sigma - \sigma_0$ is even. Then one easily verifies that $(\sigma_0, \sigma_1 - 1, \sigma) \in S_{n-1}$, (in particular, $\sigma_0 + \sigma \leqq n - 1 = 2m$ since $\sigma_0 + \sigma$ is even). Here $n - 1$ is even while $\sigma_1 - 1 < \sigma$. Consequently, by (B), the triplet $(\sigma_0, \sigma_1 - 1, \sigma)$ is realized by at least one $(\varepsilon_0, \cdots, \varepsilon_{n-2}, \varepsilon_{n-1})$ in Σ_{n-1} having ε_{n-2} and ε_{n-1} of opposite sign. Choosing $\varepsilon_n = \varepsilon_{n-1}$, one obtains a realization in Σ_n of $(\sigma_0, \sigma_1, \sigma)$.

Finally, consider the case that $n = 2m + 1$, $\sigma_1 \geqq 1$ while $\sigma - \sigma_0$ is odd, thus $\sigma_0 < \sigma$. Then $(\sigma_0, \sigma_1 - 1, \sigma - 1) \in S_{n-1}$ with $(\sigma_1 - 1) + (\sigma - 1) < n - 1$ and $n - 1$ even. By (B), the latter triplet is realized by some $(\varepsilon_0, \cdots, \varepsilon_{n-2}, \varepsilon_{n-1})$ in Σ_{n-1} with $\varepsilon_{n-2} = \varepsilon_{n-1}$. Choosing $\varepsilon_n = -\varepsilon_{n-1}$, one obtains a realization in Σ_n of $(\sigma_0, \sigma_1, \sigma)$.

Remark 4. Implicit in the above proof is a well-defined construction for obtaining a realization in Σ_n of a given triplet $(\sigma_0, \sigma_1, \sigma)$ in S_n, that is, a triplet of nonnegative integers satisfying (i)–(v) of Theorem 4. For instance, the triplet (3, 2, 5) belongs to S_8 and has $\sigma_1 + \sigma < n = 8$. It is first reduced to $(2, 2, 5) \in S_7$, then to $(2, 1, 4) \in S_6$, then to $(1, 1, 4) \in S_5$, then to $(1, 0, 3) \in S_4$, leading to the realization $(-+-++--++)$ in Σ_8 of (3, 2, 5) such that $\varepsilon_7 = \varepsilon_8$.

REFERENCES

[1] F. R. GANTMACHER, *Matrizenrechung*, Teil II. VEB Deutscher Verlag der Wiss., Berlin, 1959.

[2] F. R. GANTMACHER AND M. G. KREIN, *Oscillatory Matrices and Kernels and Small Vibrations of Mechanical Systems*, 2nd ed., Nauka Moscow, 1950.

[3] A. HURWITZ, *Ueber die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Teilen besitzt*, Math. Ann., 46 (1895), pp. 273–284.

[4] S. KARLIN, *Total Positivity*, Vol. I. Stanford Univ. Press, Stanford, CA, 1968.

[5] M. MARDEN, *Geometry of Polynomials*, Mathematical Surveys 3, American Mathematical Society, Providence, RI, 1966.

[6] B. A. ASNER, JR., *On the total nonnegativity of the Hurwitz matrix*, SIAM J. Appl. Math., 18 (1970), pp. 407–414.

[7] P. L. BROCKETT AND J. H. B. KEMPERMAN, *On the unimodality of high convolutions*, Ann. Probab., to appear.

SIAM J. MATH. ANAL.
Vol. 13, No. 3, May 1982

0036-1410/82/1303-0001 $01.00/0

PERIODIC SOLUTIONS OF HAMILTONIAN SYSTEMS: A SURVEY*

PAUL H. RABINOWITZ†

Dedicated to Michael Golomb

Abstract. Recent contributions to the study of periodic solutions of Hamiltonian systems of ordinary differential equations are surveyed and their relationship to some earlier results is also considered.

Introduction. Hamiltonian systems of ordinary differential equations model the motion of a discrete mechanical system. During the past few years there has been a considerable amount of progress in the study of periodic solutions of such systems, with many new ideas and methods of solution being introduced. The purpose of this paper is to survey these recent developments and their connection with some earlier results. In particular, the main results that have been obtained will be stated and an indication will be given of their proofs. A few open questions will also be mentioned.

Let $p, q \in \mathbb{R}^n$ and $H : \mathbb{R}^{2n} \to \mathbb{R}$ be differentiable. An autonomous Hamiltonian system has the form

$$\dot{p} = -\frac{\partial H}{\partial q}(p, q), \qquad \dot{q} = \frac{\partial H}{\partial p}(p, q), \tag{0.1}$$

where $\cdot$ denotes d/dt. This system can be represented more concisely as

$$\dot{z} = \mathcal{J} H_z(z) \tag{HS}$$

where $z = (p, q)$ and $\mathcal{J} = \begin{pmatrix} 0 & -\mathcal{I} \\ \mathcal{I} & 0 \end{pmatrix}$, $\mathcal{I}$ being the identity matrix in $\mathbb{R}^n$. Also of interest is the forced analogue of (HS)

$$\dot{z} = \mathcal{J} H_z(t, z), \tag{FHS}$$

where H depends explicitly on t in a time periodic fashion.

There are many types of questions, both local and global, that have been studied for (HS) and (FHS). One set of questions has been motivated by the fact that H is an integral of the motion for (HS), i.e. if $z(t)$ satisfies (HS), $H(z(t))$ is independent of t. Thus one can seek solutions of (HS) having prescribed energy and ask what geometrical properties must an energy surface possess in order for there to exist periodic orbits of (HS) on it. Multiplicity questions are also natural: how many geometrically distinct periodic solutions can there be on a given energy level. Other questions of interest are the existence of solutions of (HS) having a prescribed period and of (FHS) having the given period of forcing. In the setting of (FHS) one can also study the existence of subharmonic solutions, i.e., solutions having a period which is an integer multiple of the period of forcing. Concerning local questions, perturbations of equilibrium or periodic solutions lead to problems of continuation and bifurcation.

The underlying theme in the recent treatment of these problems has been the use of the calculus of variations in finding solutions as critical points of a functional. There have been approaches to (HS) and (FHS) from three main directions: (i) differential geometry—obtaining solutions as geodesics in an appropriate metric; (ii) the direct methods of the calculus of variations—obtaining solutions by minimax arguments from

* Received by the editors September 25, 1980. This research was sponsored in part by the U.S. Office of Naval Research under Contract N00014-76-C-0300 and by the U.S. Army under contract DAAG29-80-C-0041.

† Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

indefinite functionals; (iii) convex analysis and optimization theory—obtaining solutions for convex H using tools such as the Legendre transformation to simplify the problem.

We will mainly concern ourselves with the existence of periodic solutions of (HS) and (FHS) in the large. However, to add perspective some local results also will be mentioned. This will be done in § 1. Global results will be described in § 2.

1. Local results. This section is concerned with some local results for (HS) and (FHS). The early work in this direction had an analytical flavor while the more recent research makes essential use of topological arguments.

We begin with a study of (HS). To normalize matters, let $H(0)=0$. We further assume $H_z(0)=0$, so $z\equiv 0$ is a solution of (HS). The question of interest then is the existence of time periodic solutions of (HS) which are small in amplitude. An old result of Lyapunov [1]—the Lyapunov center theorem—applies to this situation:

THEOREM 1.1. *Suppose H is twice continuously differentiable near 0, $H_z(0)=0$, and the spectrum of $JH_{zz}(0)$, $\sigma(JH_{zz}(0))=\{\pm i\zeta_1,\cdots,\pm i\zeta_m\}$, where ζ_j is real, $1\leqq j\leqq n$. If ζ_j/ζ_1 is not an integer for $j\neq 1$, then* (HS) *possesses a one-parameter family of periodic solutions $z_s(t)$ whose periods $T(s)\to 2\pi/\zeta_1$ as $s\to 0$.*

Actually, Lyapunov looked at a more general situation than (HS). After some simplifications, the proof of Theorem 1.1 can be reduced to the implicit function theorem. If further "nonresonance" assumptions are made on the numbers ζ_j, (HS) possesses n distinct one-parameter families of solutions near $z=0$. Thus, if $H(z)>0$ for small $z\neq 0$, these curves of solutions will pierce $H^{-1}(c)$ for small $c>0$ and $H^{-1}(c)$ contains n geometrically distinct periodic solutions of (HS). Many attempts were made to obtain similar results without having to impose nonresonance or irrationality assumptions on $\{\zeta_j\}$. See, e.g., Gordon [2] for such a partial result. No major successes were achieved, however, until 1973, when A. Weinstein [3] proved:

THEOREM 1.2. *Suppose H is twice continuously differentiable near 0, $H_z(0)=0$, and $H_{zz}(0)$ is positive definite. Then for all sufficiently small $c>0$,* (HS) *possesses n geometrically distinct periodic solutions on $H^{-1}(c)$.*

Other versions of the result permit the assumption on $H_{zz}(0)$ to be weakened somewhat [3], [4]. Weinstein's original proof of Theorem 1.2 relies on tools from the theory of Lagrangian manifolds. Moser [4] presents a simpler proof using a variant of the method of Lyapunov–Schmidt to reduce the problem to that of finding critical points of a C^1 function on S^{2n-1}, the function being invariant under a fixed-point-free S^1 action. A standard minimax theorem then provides n geometrically distinct critical points. Other results on perturbation of periodic solutions can be found in Bottkol [5] and Weinstein [6].

It is interesting to note that Theorem 1.2 can be interpreted as the $\dot{S}^1$ version, in its setting, of a bifurcation theorem involving functionals with a $\mathbb{Z}^2$ symmetry due to Böhme [7] and Marino [8]. They proved that if E is a real Hilbert space and $f\in C^2(E,\mathbb{R})$ with f even, $f'(u)=Lu+H(u)$, L being linear and $H(u)=o(\|u\|)$ at $u=0$, then if $\mu\in\sigma(L)$ is an isolated eigenvalue of multiplicity n, the equation $f'(u)=\lambda u$ has, for each sufficiently small $r>0$, at least n distinct pairs of solutions $(\lambda,\pm u)$ with $\|u\|=r$ near $(\mu,0)$. (Here $f'(u)$ denotes the Fréchet derivative of f. Using the duality between E and E', it can be interpreted as a mapping of E to E.) The work of Böhme and Marino motivated Fadell and the author to study the existence of solutions to $f'(u)=\lambda u$ as a function of λ for λ near μ [9]. Applying these ideas to (HS) where S^1 symmetries occur in a natural fashion when one is seeking periodic solutions, we considered solutions of (HS) near a bifurcation point as a function of the period [10] and showed:

THEOREM 1.3. *Suppose H is twice continuously differentiable near 0 and $H_z(0) = 0$. Let $\mathbb{R}^{2n} = E_1 \oplus E_2$, where E_1 and E_2 are invariant subspaces for the flow given by*

$$\dot{w} = J H_{zz}(0) w. \tag{1.4}$$

Suppose all solutions of (1.4) *with initial data in E_1 are T periodic, no solutions of* (1.4) *with initial data in $E_2 \backslash \{0\}$ are T periodic, and there are no equilibrium solutions of* (1.4) *in $E_1 \backslash \{0\}$. If the signature 2ν of the quadratic form $(H_{zz}(0)\zeta, \zeta)$, $\zeta \in E_1$ is nonsingular, then either* (i) *every neighborhood of $z = 0$ contains T periodic solutions of* (HS) *or* (ii) *there are a pair of integers $k, m \geqq 0$ such that $k + m \geqq |\nu|$ and left and right neighborhoods $\mathcal{I}_l, \mathcal{I}_r$ of T in $\mathbb{R}$ such that for all $\lambda \in \mathcal{I}_l$ (resp. $\mathcal{I}_r$), (HS) possesses at least k (resp. m) distinct nontrivial λ periodic solutions.*

The proof of Theorem 1.3 is related to that of Theorem 1.2 sketched above. To begin, one seeks solutions of (HS) in an infinite dimensional space of periodic functions. The method of Lyapunov–Schmidt reduces the problem to a finite-dimensional one, and a minimax argument relying on an S^1 symmetry inherent in the problem gives the solutions as critical points of a variational formulation of (HS). The minimax construction of the critical points here is more subtle than in [4] due to the fact that there is no analogue of the energy surface constraint of Theorem 1.2 here so one is working in a neighborhood of 0 rather than on a compact manifold. A special case of Theorem 1.3 was obtained independently by Chow and Mallet-Paret [11].

We conclude this section by stating a local theorem concerning the existence of subharmonic solutions of (FHS) due to Birkhoff and Lewis [12]–[14]. Some preliminary remarks are necessary. If H is smooth near $z = 0$, $H(t, 0) = 0$ and $H_z(t, 0) = 0$, then $H(t, z) = Q(t, z) + R(t, z)$ where Q is quadratic in z and $R(t, z) = o(|z|^2)$ at $z = 0$. If the Floquet exponents for the linear Hamiltonian system corresponding to Q are purely imaginary, Floquet theory and the Hamiltonian character of (FHS) permit canonical changes of dependent variables so that the transformed problem has a time-independent quadratic part of the form

$$\sum_{i=1}^{n} \lambda_i \frac{p_i^2 + q_i^2}{2} \tag{1.5}$$

(see Arnold [15]). Thus we can assume $Q = Q(z)$ and has the form (1.5). Let $\Lambda = (\lambda_1, \cdots, \lambda_n)$ and let k be a multiindex, $k = (k_1, \cdots, k_n) \in \mathbb{Z}^n$. Set $|k| = k_1 + \cdots + k_n$ and $\langle \Lambda, k \rangle = \sum_{i=1}^{n} \lambda_i k_i$. If H is C^4 with respect to z near 0 and $\langle \Lambda, k \rangle \notin \mathbb{Z}$ for all $|k| \leqq 4$, then there exists a canonical change of variables which transforms H into the Birkhoff normal form

$$Q(z) + \sum_{i,j=1}^{n} \alpha_{ij} \left(\frac{p_i^2 + q_i^2}{2} \right) \left(\frac{p_j^2 + q_j^2}{2} \right) + o(|z|^4) \tag{1.6}$$

at $z = 0$ (see [15] or Siegel–Moser [16]). Now we can state:

THEOREM 1.7. *Suppose H is C^4 near 0 and periodic in t with $H(t, z) = Q(z) + R(t, z)$, Q as in* (1.5) *and $R(t, z) = o(|z|^2)$ at $z = 0$. If $\langle \Lambda, k \rangle \notin \mathbb{Z}$ for all $|k| \leqq 4$ and $\det(\alpha_{ij}) \neq 0$ in* (1.6), *then* (FHS) *possesses a sequence of subharmonic solutions (Z_k) with arbitrarily large minimal periods.*

Above $\det(\alpha_{ij})$ denotes the determinant of the matrix (α_{ij}). There is also a variant of Theorem 1.7 for (HS) [17]. A proof of Theorem 1.7 but in the setting of maps can be found in Moser [18]. A less general version of Theorem 1.7 in the above setting is proved in Harris [17].

2. Global results. In this section the existence in the large of periodic solutions of (HS) and (FHS) will be studied. Our presentation will be roughly chronological. The first work we know of in a global setting is due to Seifert [19], who considered Hamiltonians consisting of the sum of a kinetic and a potential energy term. He essentially proved:

THEOREM 2.1. *Let* $H(p, q) = \sum_{i,j=1}^{n} a_{ij}(q)p_ip_j + V(q)$, *where* a_{ij}, $V \in C^2(\mathbb{R}^{2n}, \mathbb{R})$, *the matrix* $(a_{ij}(q))$ *is uniformly positive definite in* $\mathscr{D} = \{q \in \mathbb{R}^n \mid V(q) < 1\}$ *and* V *satisfies*

(V_1) $\mathscr{D}$ *is* C^2 *diffeomorphic to the unit ball in* $\mathbb{R}^n$.

(V_2) $\partial\mathscr{D}$ *is a manifold.*

Then there exist points $q^*, Q^* \in \partial\mathscr{D}$, $T > 0$, *and a solution* $(p(t), q(t))$ *of* (HS) *such that* $(p(0), q(0)) = (0, q^*)$, $(p(T), q(T)) = (0, Q^*)$, *and* $q(t) \in \mathscr{D}$ *for* $0 < t < T$.

Thus Theorem 2.1 gives us a solution whose motion begins and ends on the boundary of the potential well $\mathscr{D}$. Seifert actually assumed real analyticity for a_{ij} and V, but C^2 suffices for his arguments. Observing that H is even in p, a $2T$ periodic solution of (HS) on $H^{-1}(1)$ can be constructed by extending q as an even function and p as an odd function about 0 and T.

The existence of periodic solutions having a prescribed mean potential energy was studied by Berger [20] for a class of second order Hamiltonian systems which have a less general potential energy term than given above.

Theorem 2.1 was generalized by A. Weinstein [21] who permitted a wider class of kinetic energy terms:

THEOREM 2.2. *Suppose* $H(p, q) = K(p, q) + V(q)$, *where* $K \in C^2(\mathbb{R}^{2n}, \mathbb{R})$, $V \in C^2(\mathbb{R}^{2n}, \mathbb{R})$, V *satisfies* (V_1), (V_2) *and* K *satisfies*

(K_1) K *is even and strictly convex in* p *for each* $q \in \bar{\mathscr{D}}$.

(K_2) $K(0, q) = 0$ *and* $K(p, q) \to \infty$ *as* $|p| \to \infty$ *uniformly for* $q \in \bar{\mathscr{D}}$.

Then the conclusions of Theorem 2.1 *obtain.*

Seifert used ideas from differential geometry to prove Theorem 2.1. Roughly speaking he found the solution as a geodesic for a Riemannian metric (called the Jacobi metric) associated with the kinetic energy term in his Hamiltonian. Weinstein used a similar argument in his setting, the Riemannian metric being replaced by a Finsler metric associated with the more general K. Due to the fact that the metric degenerates on $\partial\mathscr{D}$, an approximation argument and a priori bounds which keep the approximate period away from 0 and ∞ are required in both cases.

Weinstein goes on in [21] to prove a result for general Hamiltonian systems:

THEOREM 2.3. *Suppose* $H \in C^2(\mathbb{R}^{2n}, \mathbb{R})$ *and* $H^{-1}(1)$ *is a manifold which bounds a compact convex region in* $\mathbb{R}^{2n}$. *Then* (HS) *possesses a periodic solution on* $H^{-1}(1)$.

The proof of Theorem 2.3 involves a clever application of Theorem 2.2. A new Hamiltonian on $\mathbb{R}^{4n}$ is constructed which satisfies the hypotheses of Theorem 2.2 (with n replaced by $2n$) and for which solutions of the type given in Theorem 2.2 correspond to periodic solutions of (HS) on $H^{-1}(1)$.

Simultaneous to Weinstein's work on Theorem 2.3, this author also was studying (HS), but from a totally different point of view, and obtained the following somewhat more general result [22]:

THEOREM 2.4. *Suppose* $H \in C^1(\mathbb{R}^{2n}, \mathbb{R})$ *and* $H^{-1}(1)$ *is a manifold which bounds a compact star-shaped region in* $\mathbb{R}^{2n}$, *i.e., there is a* $\xi \in \mathbb{R}^{2n}$ *such that, with* ξ *as origin,* $H^{-1}(1)$ *is radially diffeomorphic to* S^{2n-1}. *Then* (HS) *possesses a periodic solution on* $H^{-1}(1)$.

To describe the approach taken to (HS) in [22], observe first that the period of any periodic solution on $H^{-1}(1)$ is not known a priori. It is convenient to rescale the time variable and replace (HS) by

$$\dot{z} = \lambda J H_z(z), \tag{2.5}$$

where we now seek a 2π periodic function z and a nonzero scalar λ (essentially the unknown period) such that the pair satisfy (2.5). The idea now is to use the calculus of variations to find a solution of (2.5). Let $(\cdot, \cdot)_{\mathbb{R}^j}$ denote the inner product of two vectors in $\mathbb{R}^j$. Formally a critical point z of the action integral

$$A(z) = \int_0^{2\pi} (p, \dot{q})_{\mathbb{R}^n}\, dt, \tag{2.6}$$

subject to the constraint

$$\frac{1}{2\pi}\int_0^{2\pi} H(z)\, dt = 1, \tag{2.7}$$

has (2.5) as its Euler equation, λ appearing as a Lagrange multiplier due to (2.7). Moreover since z satisfies (2.5), $H(z(t)) \equiv c$, a constant, and by (2.7), $c = 1$.

It seems to be a difficult matter to make these heuristics precise in a direct fashion. Instead a finite dimensional approximation argument, looking for solutions in the class of trignometric polynomials, was used in [22]. Observe that $A(z)$ and the constraint are invariant if $z(t)$ is replaced by $z(t+\theta)$ for $\theta \in [0, 2\pi]$; i.e., the problem possesses an S^1 symmetry. Thus by employing an index theory for such S^1 actions [10] and minimax arguments, critical points can be obtained for an approximating finite dimensional problem. Appropriate bounds for approximate solutions and their associated Lagrange multipliers allow one to pass to a limit and solve (2.5).

An interesting geometrical result concerning the relation between the period and $H^{-1}(1)$ that comes up in the course of the proof of Theorem 2.4 is the following:

THEOREM 2.8. *Suppose $H \in C^1(\mathbb{R}^{2n}, \mathbb{R})$ is homogeneous of degree two and positive for $z \neq 0$. Let r and R denote respectively the radii of the smallest inscribed and largest circumscribed sphere for $H^{-1}([0, 1])$. Then* (HS) *has a T periodic solution on $H^{-1}(1)$ with $\frac{1}{2}r^2 \leqq T \leqq \frac{1}{2}R^2$.*

The bounds obtained in Theorem 2.8 play a role in the proof of Theorem 2.4. We depart from our chronological development for a moment to mention the following result which was motivated by Weinstein's Theorem 2.2:

THEOREM 2.9 [23]. *Suppose $H(p, q) = K(p, q) + V(q)$, where $K \in C^2(\mathbb{R}^{2n}, \mathbb{R})$, $V \in C^1(\mathbb{R}^n, \mathbb{R})$, V satisfies* (V_1)–(V_2), *and K satisfies* (K_2) *and*

$$(p, K_p(p, q))_{\mathbb{R}^n} > 0 \quad \text{for } p \neq 0. \tag{K_3}$$

Then (HS) *possesses a periodic solution on $H^{-1}(1)$.*

Thus Theorem 2.9 replaces the convexity hypothesis (K_1) by the "starshaped" assumption (K_3). The proof of Theorem 2.9 is based on that of Theorem 2.4, the bounds required here being somewhat more difficult to obtain. Theorems 2.1, 2.2, 2.4 and 2.9 all give sufficient geometrical conditions under which (HS) possesses periodic orbits on $H^{-1}(1)$. Just how general an energy surface one can take and still be guaranteed the existence of periodic orbits of (HS) on it remains an open question. See, e.g., [24] for some conjectures in this direction.

Our discussion up to this point has only dealt with global results for (HS) when the energy is prescribed. In [22] a study also was begun of the existence of periodic solutions of (HS) when the period is prescribed. The simplest such case treated in [22] is

THEOREM 2.10. *Suppose $H \in C^1(\mathbb{R}^{2n}, \mathbb{R})$ and satisfies*

(H_1) $H(z) \geqq 0$.

(H_2) $H(z) = o(|z|^2)$ *at $z = 0$.*

(H_3) *There are constants* $r>0$ *and* $\mu>2$ *such that, for all* $|z|>r$,

$$0<\mu H(z)\leqq(z, H_z(z))_{\mathbb{R}^{2n}}.$$

Then for any $T>0$, (HS) *has a nonconstant* T *periodic solution.*

A few remarks about this theorem are in order. First integrating the inequality in (H_3) shows there are constants $a_1, a_2>0$ such that

$$H(z)\geqq a_1|z|^{\mu}-a_2 \tag{2.11}$$

for all $z\in\mathbb{R}^{2n}$; i.e., H grows at a "superquadratic" rate as $|z|\to\infty$. Likewise (H_2) implies $H(z)\to 0$ as $|z|\to 0$ at a superquadratic rate.

The proof of Theorem 2.10 is in the same spirit as that of Theorem 2.4. Suppose for convenience we seek a 2π periodic solution of (HS). Then any critical point of

$$I(z)=A(z)-\int_0^{2\pi}H(z)\,dt \tag{2.12}$$

in the class of 2π periodic functions is a solution of (HS). To obtain a critical point of I, one proceeds as in Theorem 2.4 with three main differences: (i) no constraint is involved here; (ii) a minimax argument is given based on (H_1)–(H_3) which avoids the use of symmetries and the index theory of [10]; (iii) an additional difficulty is encountered here due to the presence of the trivial solution $z\equiv 0$. To overcome (iii), a comparison argument is employed which shows $I(z)>0$ for the solution constructed. Hence (H_1) and (2.12) imply z is nonconstant.

In research subsequent to [22], Benci and the author obtained a critical point theorem for indefinite functionals [25] which can be used to bypass the finite dimensional approximation arguments of [22] and get a critical point of (2.12) directly in the Sobolev space $(W^{1/2,2}(S^1))^{2n}$. See also Ekeland [26] who gave a direct proof of a special case of Theorem 2.10.

Several variants of Theorem 2.10 were also proved in [22] including one for (FHS). We will return to this result later when we discuss subharmonics for (FHS). Also discussed in [22] were some results for second order Hamiltonian systems:

$$\ddot{q}+V_q(t, q)=0, \tag{2.13}$$

where $q\in\mathbb{R}^n$, for forced or free V which satisfy hypotheses like (H_1)–(H_3).

Although in its setting Theorem 2.10 guarantees a nonconstant solution of period T of (HS) for all $T>0$, nothing is implied concerning the existence of a solution having minimal period T. We suspect that (H_1)–(H_3) are sufficient to give solutions of (HS) of minimal period T for any $T>0$. However if (H_1)–(H_2) are dropped, one cannot expect this to be the case. Indeed, suppose $n=1$ and consider $H(z)=g(|z|^2)$, where $g\in C^{\infty}(\mathbb{R}, \mathbb{R})$. Setting $\zeta=p+iq$, the corresponding Hamiltonian system can be written in complex form as

$$\dot{\zeta}=2ig'(|\zeta|^2)\zeta. \tag{2.14}$$

Thus $\zeta(t)=\zeta_0\exp[2ig'(|\zeta|^2)t]$, so if T is the minimal period of $\zeta(t)$, $T\leqq\pi[g'(|\zeta|^2)]^{-1}$. Consequently $g'\geqq 1$ implies $T\leqq 2\pi$.

On the other hand, if one is not interested in solutions having minimal periods, one can do much better than Theorem 2.10, namely:

THEOREM 2.15 [44]. *Suppose* $H\in C^1(\mathbb{R}^{2n}, \mathbb{R})$ *and satisfies* (H_3) *and* $H(z)\leqq a_3|z|^s+a_4$. *Then for all* $T, R>0$, (HS) *has a solution* z *of period* T *and satisfying* $\|z\|_{L^{\infty}}>R$.

Thus one can obtain arbitrarily large T periodic solutions (which as the above example shows may not have minimal period T). The proof of Theorem 2.15 is more complicated than that of Theorem 2.10, the structure given by (H_1)–(H_2) being replaced by the S^1 invariance of $I(z)$ as was employed in the proof of Theorem 2.4.

One final remark about the proof of Theorem 2.10: Although it appears to be rather different from Theorem 2.4, the two are closely related and in fact Theorem 2.10 can be used to give a short elementary proof of Theorem 2.4 [27].

At this point in the development of the theory, the convex analysts make their appearance. F. Clarke [28], [29] gave a new and simpler proof of Theorem 2.3. In addition he weakened the smoothness assumptions on H to merely convexity and continuity, so (HS) becomes an inclusion rather than an equation. However in our description here, we prefer to stay in the classical framework. Part of Clarke's idea is to employ a Legendre transformation to convert the problem to a simpler one. Unlike the usual situation in mechanics, he uses a Legendre transformation in all variables. Equation (HS) can be written as

$$-J\dot{z} = H_z(z) \tag{2.16}$$

with, in the setting of Theorem 2.3, H globally convex via a trick of [21] or [22]. Thus H_z is monotone. In essence, Clarke inverts H_z in (2.16), transforming it to

$$z = H_z^{-1}(-J\dot{z}). \tag{2.17}$$

This new equation in which $\dot{z}$ is taken to be the independent variable can be given a variational formulation for which a solution can be obtained as a minimum of the corresponding functional.

As another consequence of these ideas, Clarke and Ekeland [30] studied a situation complementary to that of Theorem 2.10 in which H is "subquadratic" at 0 and ∞; i.e.,

$$H(z)|z|^{-2} \to 0 \quad \text{as } |z| \to \infty \tag{H_4}$$

and

$$H(z)|z|^{-2} \to \infty \quad \text{as } |z| \to 0. \tag{H_5}$$

They proved

THEOREM 2.18. *Suppose $H \in C^1(\mathbb{R}^{2n}, \mathbb{R})$, H is convex with $H(0) = 0$, $H_z(0) = 0$, and satisfies (H_4)–(H_5). Then for all $T > 0$, (HS) has a solution having minimal period T.*

The minimality of T is a consequence of the characterization of the solution as a minimum of a variational problem. This theorem is the only result we know of in the context of general Hamiltonian systems which obtains information on minimal periods. See also Berger [20] or [31] for results on second order Hamiltonian systems.

In a further application of Legendre transformation ideas in conjunction with minimax arguments and the index theory of [10], Ekeland and Lasry proved a nice result which furnishes a partial globalization of Weinstein's bifurcation theorem (Theorem 1.2). Let B_ρ denote a Euclidean ball of radius ρ.

THEOREM 2.19. *Suppose $H \in C^1(\mathbb{R}^{2n}, \mathbb{R})$ and $H^{-1}(1)$ is a manifold which bounds a compact convex domain Ω. If there are positive numbers r and $R < \sqrt{2}r$ such that $B_r \subset \Omega \subset B_R$, then $H^{-1}(1)$ contains at least n geometrically distinct periodic solutions of* (HS).

Whether or not the restriction on the shape of Ω is essential remains an open question. Likewise, nothing is known about the number of periodic solutions of (HS)

in the settings of Theorems 2.1, 2.2 and 2.9. Also of interest are the number of solutions of (HS) having a given minimal period. No progress seems to have been made in this direction.

In addition to Theorem 2.18, there have been a considerable number of results obtained for subquadratic Hamiltonian systems, both autonomous and forced, satisfying variants of (H_4) and (H_5). We will not go into detail here but they include works by Benci [32]–[34], Benci and Rabinowitz [25], Brezis and Coron [35], and Coron [36]. Also Amann [37] and Amann and Zehnder [38]–[39] have studied problems which lie on the border between sub- and superquadratic, namely Hamiltonian systems which are quadratic near 0 and ∞ but have different signatures. Using a global Lyapunov–Schmidt reduction, minimax arguments and index theories, they obtain many existence and multiplicity results for (HS) and (FHS). Some earlier special cases were obtained by D. Clark [40].

Next we shall describe a contribution to subharmonic solutions of (FHS). As was mentioned earlier, a variant of Theorem 2.10 for (FHS) was given in [22]. It turns out that under the same hypotheses much more is true:

THEOREM 2.20 [41]. *Suppose* $H \in C^1(\mathbb{R}^{2n}, \mathbb{R})$ *and satisfies*

(H_6) *There is a constant* $T > 0$ *such that* $H(t+T, z) = H(t, z)$ *for all* $t \in \mathbb{R}$, $z \in \mathbb{R}^{2n}$.

(H_7) *There are constants* $\alpha, \beta > 0$ *such that for* $|z| > \beta$, $|H_z(t, z)| \leqq \alpha(z, H_z(t, z))_{\mathbb{R}^{2n}}$,

and (H_1)–(H_3) *with respect to* z. *Then, for each* $k \in \mathbb{N}$, (FHS) *possesses a* kT *periodic solution* $z_k(t)$. *Moreover infinitely many of the functions* z_k *are distinct.*

The proof of this result follows from the abstract critical point theorem of [25] combined with some bounds for the critical points and a simple indirect argument showing infinitely many must be distinct. Theorem 2.20 can be viewed as a global relative of the Birkhoff–Lewis theorem, where (H_3) plays the role of the condition on the quartic part of H in Theorem 1.7. In fact one can prove a local variation on Theorem 1.7 using (H_3) and Theorem 2.20 as a tool. Clarke and Ekeland [42] have also obtained a result on subharmonics for second order forced Hamiltonian systems (2.13) with convex subquadratic $V(q)$. See also [41] for another subquadratic case.

Our final result is a recent theorem of Gluck and Ziller [43] concerning the fixed energy case of (HS) which extends Theorem 2.2.

THEOREM 2.21. *Suppose* $H(p, q) = K(p, q) + V(q)$, *where* $K \in C^2(\mathbb{R}^{2n}, \mathbb{R})$, $V \in C^2(\mathbb{R}^n, \mathbb{R})$, K *satisfies* (K_1)–(K_2), *and* V *satisfies*

(V_3) $\mathscr{D} = \{q \in \mathbb{R}^n \mid V(q) \leqq 1\}$ *is compact and nonempty*

and (V_2). *Then* (HS) *possesses a periodic solution on* $H^{-1}(1)$.

The proof of Theorem 2.21 follows the geometrical approach of [19] and [21] together with some further topological ideas.

In conclusion, it should be mentioned that one of the main sources of inspiration for the development of Hamiltonian mechanics was the field of celestial mechanics. In this field, unlike the situations described above, one encounters Hamiltonians which possess singularities. We believe celestial mechanics is a very interesting and possibly fertile proving ground for the further development of the ideas and methods described in this survey.

REFERENCES

[1] A. LYAPUNOV, *Problème générale de la stabilite du mouvement*, Ann. Fac. Sci. Toulouse, 2 (1907), pp. 203–474.

[2] W. B. GORDON, *A theorem on the existence of periodic solutions to Hamiltonian systems with convex potential*, J. Differential Equations, 10 (1971), pp. 324–335.

[3] A. WEINSTEIN, *Normal modes for non-linear Hamiltonian systems*, Inv. Math., 20 (1973), pp. 47–57.
[4] J. MOSER, *Periodic orbits near an equilibrium and a theorem by Alan Weinstein*, Comm. Pure Appl. Math., 29 (1976), pp. 727–747.
[5] M. BOTTKOL, *Bifurcation of periodic orbits on manifolds and Hamiltonian systems*, PhD. thesis, New York University, 1977.
[6] A. WEINSTEIN, *Bifurcations and Hamilton's principle*, Math. Z., 159 (1978), pp. 235–248.
[7] R. BÖHME, *Die Lösung der Verzweigungsgleichungen für nichtlineare Eigenvectproblem*, Math. Z., 127 (1972), pp. 105–126.
[8] A. MARINO, *La biforcazione nel caso variazionale*, Conf. Sem. Mat. dell'Univ. Bari, 132, 1977.
[9] E. R. FADELL AND P. H. RABINOWITZ, *Bifurcation for odd potential operators and an alternative topological index*, J. Funct. Anal., 26 (1977), pp. 48–67.
[10] ———, *Generalized cohomological index theories for Lie group actions with an application to bifurcation questions for Hamiltonian systems*, Inv. Math., 45 (1978), pp. 139–174.
[11] S. N. CHOW AND J. MALLET-PARET, *Periodic solutions near an equilibrium of a nonpositive definite Hamiltonian system*, Michigan State University preprint.
[12] G. D. BIRKHOFF, *Une generalization à n-dimensions du dernier théorème de géometrié de Poincaré*, Compt. Rend. Acad. Sci., 192 (1931), pp. 196–198.
[13] G. D. BIRKHOFF AND D. C. LEWIS, *On the periodic motions near a given periodic motion of a dynamical system*. Ann. Mat. Pura Appl., 12 (1933), pp. 117–133.
[14] D. C. LEWIS, *Sulle oscillazioni periodiche d'une sistema dinamico*, Atti. Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 19 (1934), pp. 234–237.
[15] V. I. ARNOLD, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978.
[16] C. L. SIEGEL AND J. K. MOSER, *Lectures on Celestial Mechanics*, Springer-Verlag, New York, 1971.
[17] T. C. HARRIS, *Periodic solutions of arbitrary long period in Hamiltonian systems*, J. Differential Equations, 4 (1968), pp. 131–141.
[18] J. MOSER, *Proof of a generalized fixed point theorem due to G. D. Birkhoff*, Lecture Notes in Mathematics 597, Springer-Verlag, New York, 1977.
[19] H. SEIFERT, *Periodische Bewegungen mechanischen Systeme*, Math. Z., 51 (1948), pp. 197–216.
[20] M. S. BERGER, *On a family of periodic solutions of Hamiltonian systems*, J. Differential Equations, 10 (1971), pp. 17–26.
[21] A. WEINSTEIN, *Periodic orbits for convex Hamiltonian systems*, Ann. Math., 108 (1978), pp. 507–518.
[22] P. H. RABINOWITZ, *Periodic solutions of Hamiltonian systems*, Comm. Pure Appl. Math., 31 (1978), pp. 157–184.
[23] ———, *Periodic solutions of a Hamiltonian system on a prescribed energy surface*, J. Differential Equations, 33 (1979), pp. 336–352.
[24] A. WEINSTEIN, *On the hypotheses of Rabinowitz' periodic orbit theorems*, J. Differential Equations, 33 (1979), pp. 353–358.
[25] V. BENCI AND P. H. RABINOWITZ, *Critical point theorems for indefinite functionals*, Inv. Math., 52 (1979), pp. 336–352.
[26] I. EKELAND, *Periodic solutions of Hamiltonian equations and a theorem of P. Rabinowitz*, J. Differential Equations, 34 (1981), pp. 523–534.
[27] P. H. RABINOWITZ, *A variational method for finding periodic solutions of differential equations*, in Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, New York, 1978, pp. 225–251.
[28] F. CLARKE, *Periodic solutions to Hamiltonian inclusions*, J. Differential Equations, to appear.
[29] ———, *A classical variational principle for periodic Hamiltonian trajectories*, Proc. Amer. Math. Soc., 76 (1979), pp. 186–188.
[30] F. CLARKE AND I. EKELAND, *Hamiltonian trajectories having prescribed minimal period*, Comm. Pure Appl. Math., 33 (1980), 103–116.
[31] M. S. BERGER, *Nonlinearity and Functional Analysis*, Academic Press, New York, 1977.
[32] V. BENCI, *Some critical point theorems and applications*, Comm. Pure Appl. Math., 33 (1980), pp. 147–173.
[33] ———, *A geometrical index for the group S^1 and some applications to the research of periodic solutions of O.D.E.'s*, Comm. Pure Appl. Math., to appear.
[34] ———, *On the critical point theory for indefinite functionals in the presence of symmetry*, Trans. Amer. Math. Soc., to appear.
[35] H. BREZIS AND J. M. CORON, *Periodic solutions of nonlinear wave equations and Hamiltonian systems*, preprint.
[36] J. M. CORON, *Resolution de l'équation $Au+Bu=f$ où A est linéare autoadjoint et B deduit d'un potential convex*, Ann. Fac. Sci. Toulouse, to appear.

[37] H. AMANN, *Saddle points and multiple solutions of differential equations*, Math. Z., 169 (1979), pp. 127–166.

[38] H. AMANN AND E. ZEHNDER, *Nontrivial solutions for a class of nonresonance problems and applications to nonlinear differential equations*, Ann. Sc. Norm. Sup. Pisa, Ser. 4, 7 (1980), pp. 539–603.

[39] ———, *Periodic solutions of asymptotically linear Hamiltonian equations*, Manus. Math., 32 (1980), pp. 149–189.

[40] D. C. CLARK, *Periodic solutions of variational systems of ordinary differential equations*, J. Differential Equations, 28 (1978), pp. 354–368.

[41] P. H. RABINOWITZ, *Subharmonic solutions of Hamiltonian systems*, Comm. Pure Appl. Math., 33 (1980), pp. 609–633.

[42] F. CLARKE AND I. EKELAND, *Nonlinear oscillations and boundary value problems for Hamiltonian systems*, preprint.

[43] H. GLUCK AND W. ZILLER, *Existence of periodic motions of conservative systems*, preprint.

[44] P. H. RABINOWITZ, *On large norm periodic solutions of some ordinary and partial differential equations*, in Ergodic Theory and Dynamical Systems: Proc. Sp. Yr.-Maryland 79-80, A. Katok, ed., Birkhauser, Boston, 1981.

SIAM J. MATH. ANAL.
Vol. 13, No. 3, May 1982

0036-1410/82/1303-0002 $01.00/0

LONG-TIME BEHAVIOR OF A CLASS OF BIOLOGICAL MODELS*

H. F. WEINBERGER†

Dedicated to my friend and collaborator Michael Golomb

Abstract. It is shown that many of the asymptotic properties of the Fisher model for population genetics and population ecology can also be derived for a class of models in which time is discrete and space may or may not be discrete. This allows one to discuss the behavior of models in which the data consist of occasional counts on survey tracts, as well as that of computer models.

1. Introduction. The purpose of this work is to present some techniques by which results on long-term behavior previously obtained for continuous models in population genetics and population ecology can be shown to apply to a larger class of models in which the time and space variables are allowed to be discrete. Such an extension is useful for several reasons. In the first place it usually is impractical to measure more than finitely many (and often relatively few) aggregate quantities. For instance, one can measure the populations of various species or genotypes in a few census tracts from time to time, but one cannot measure (or possibly even define) the population densities at all points at all times.

Secondly, any deterministic model must rely on the assumption that the inevitable random events found in nature are averaged out. This is much more likely to be true of aggregates over relatively large space-time domains than of limits such as population densities in a continuous space.

Lastly, computations for continuous models are usually done by approximating them with discrete ones. On the other hand, continuous models often appear as approximations to discrete ones. (See, e.g., [2], [38], [39].) It is useful to know that all these models have the same qualitative large-time behavior.

We shall deal only with models which are deterministic, and in which the independent variable is a scalar. Both of these restrictions are serious ones. We shall briefly discuss the possibility of applying our results to stochastic models at the end of § 4.

The fact that the independent variable must be scalar-valued prevents the consideration of any very sophisticated interactions with the environment, other species, or other age classes. In particular, we consider species with nonoverlapping generations, and this naturally introduces a discrete time interval, the generation time.

The most frequently used model for studying the spread of a mutant or a new population in a homogeneous environment is the Fisher equation

$$\frac{\partial u}{\partial t} = D\Delta u + f(u). \tag{1.1}$$

This equation in one dimension was introduced by R. A. Fisher [15] as a model for the spread of an advantageous form (allele) of a single gene in a population of diploid individuals. It is assumed that there are only two allelic forms A and a of the gene under study. The function $u(t, x)$ represents the gene fraction, that is, the ratio of the number of A alleles to the number of A and a alleles at the time t in the population near the point x.

Fisher found that when $f(u) = u(1-u)$ there is a constant c^* with the property that, when $|c| \geqq c^*$, (1.1) has a traveling wave solution $u(t, x) = W(x - ct)$ with speed c, while

* Received by the editors November 12, 1980, and in revised form February 24, 1981.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. This research was partly supported by the National Science Foundation under grant MCS-7812182. Part of the work was done while the author was a visitor at the University of Queensland.

for $|c| < c^*$ there is no such wave. He conjectured that the minimal wave speed c^* is also the speed at which a newly arrived advantageous allele spreads into an established population.

Kolmogoroff, Petrowsky, and Piscounoff [31] proved this conjecture in the one-dimensional case where $u(0, x)$ is 0 for $x < 0$ and 1 for $x > 0$ provided $f(0) = f(1) = 0$ and f is concave.

The Fisher equation (1.1) is also used as a model for the spread of a population when u is interpreted as a population density.

In this work we shall consider discrete time models of the form

$$u_{n+1} = Q[u_n], \tag{1.2}$$

where $u_n(x)$ represents the gene fraction or population density at time n at the point x of the habitat and Q is an operator on a certain set of functions on the habitat. The habitat may be one-, two-, or three-dimensional. It may be discrete, in which case x ranges over a discrete set of niches (census tracts).

The model (1.2) is deterministic in the sense that u_{n+1} is uniquely determined by u_n. We continue to assume that u_n is scalar-valued and that the habitat is homogeneous. We shall confine our attention to propagation phenomena and shall not be concerned with clines [8], [14], [16], [17], [21], [38], [46].

We shall show that a few biologically reasonable hypotheses about Q permit us to derive results of the Fisher and Kolmogoroff–Petrowsky–Piscounoff types for this class of models. More specifically, we shall show that for each unit vector ξ there is a wave speed $c^*(\xi)$ with the property that, in an asymptotic sense, a new mutant or population which is initially confined to a bounded set spreads like the solution of a wave equation whose plane wave solutions with normals ξ have the speeds $c^*(\xi)$. Under some additional conditions we also show that (1.2) has a nonincreasing traveling wave solution $u_n(x) = W(x \cdot \xi - nc)$ when $c \geqq c^*(\xi)$ and not when $c < c^*(\xi)$.

Equation (1.1) implies (1.2) when $Q[v]$ is defined to be the solution at some generation time τ of the initial value problem for (1.1) with initial values $v(x)$. Therefore our results generalize the one-dimensional results of Fisher [15], Kolmogoroff, Petrowsky, and Piscounoff [31], Kanel' [25], [26], Kametaka [24], and Aronson and Weinberger [4], as well as the multidimensional results of Aronson and Weinberger [5] for the rotationally symmetric model (1.1).

Translational invariance (homogeneity) does not imply rotational invariance (isotropy). Because prevailing winds and other phenomena can produce preferred directions, the analogue of (1.1) without rotational symmetry is also of interest. In this case the operator $D\Delta$ in (1.1) is replaced by an elliptic operator of the form

$$\sum a_{ij} \frac{\partial^2}{\partial x_i \partial x_j} + \sum b_i \frac{\partial}{\partial x_i}.$$

This problem can again be put into the form (1.2) by defining $Q[v]$ to be the solution at time τ of the initial value problem with initial values $v(x)$. We can, in fact allow for seasonal variation by permitting the coefficients and f to be periodic functions of t of period τ.

Since the migrational behavior can be affected by the genotype [41] or by the population density [2], [3], [19], it is also of interest to consider the problem in which the coefficients a_{ij} and b_i in the above operator depend upon u. This problem can again be converted to the form (1.2) and treated by our methods.

Our results also generalize work by Diekmann [10], [11], Diekmann and Kaper [12], Thieme [48], [49] and the author [53], [54] on one-dimensional and rotationally symmetric discrete time models of the form (1.2) where Q is a nonlinear integral operator.

We have not been able to generalize the more delicate results on convergence in shape obtained by Kolmogoroff, Petrowsky, and Piscounoff [31], Kanel' [25], [26], Fife and McLeod [13], Bramson [7], and Uchiyama [50], [51], which show that for the one-dimensional Fisher equation the width of the transition layer between the high and low parts of a spreading pulse remains bounded and which approximate the space-time path of this layer.

The results presented here can also be applied to certain models for the spread of epidemics [1], [10], [11], [28], [29], [37], [47], [56].

Some examples of biological models are given in § 2. A precise statement of the hypotheses on the recursion operator Q is in § 3. Section 4 introduces our mathematical tools. In § 5 we show that the rather small set (3.1) of hypotheses leads to a wave speed $c^*(\xi)$ for disturbances which depend only on the single variable $x \cdot \xi$, where ξ is any unit vector.

We state our results in § 6. Essentially these are that, in an asymptotic sense, an initially confined disturbance is propagated along the ray cone which corresponds to the wave speed $c^*(\xi)$, and that there are plane wave solutions of the form $W(x \cdot \xi - nc)$ exactly when $c \geqq c^*(\xi)$. The remainder of the paper is devoted to proofs of these theorems, together with some examples and counterexamples.

Our results show that a rather large class which contains both continuous and discrete models is robust with respect to qualitative asymptotic behavior. This robustness means, of course, that a very crude model can give correct predictions about the asymptotic behavior of an ecological system. The other side of the coin of robustness, of course, is that a model which is found to predict such behavior correctly may be far from being a good model for predicting other phenomena.

2. Some models in population genetics and population growth. To obtain an example of a recursion of the form (1.2), we consider the so-called stepping stone model [27], [30], [35], [40], [45] in population genetics. We classify the individuals of a certain diploid population according to their genotypes with respect to a single gene locus, which occurs in two allelic (that is, variant) forms, which we label A and a. There are then three genotypes: the homozygotes AA and aa, and the heterozygote Aa.

The habitat is either naturally or arbitrarily divided into discrete regions or niches. The population which lives in any one niche during the nonmigratory part of a life cycle is called a deme. We suppose that the generations are nonoverlapping, and that members of different demes do not mix or interact except during a brief period of migration at the end of the life cycle. Migration moves the individuals to new niches, so that it changes the memberships of the various demes. Just after migration the individuals form gametes and die. The individuals are assumed to be monoecious (that is, hermaphroditic). The gametes mate at random to form the new generation of diploids.

The ratio of the number of alleles of type A to the total number of alleles of types A and a in a certain deme is called its *gene fraction*. Let $u_n(i)$ denote the gene fraction in the newly born individuals of the nth generation in the ith deme.

The Hardy–Weinberg law states that the ratios of the numbers of newly born individuals of genotypes AA, aA, and aa, respectively, are $u_n(i)^2 : 2u_n(i) \times (1-u_n(i)) : (1-u_n(i))^2$. These individuals undergo various hazards during the growth

cycle, and we assume that their abilities to survive these hazards depend only on their genotypes with respect to the gene under consideration. If the fitnesses (that is, survival rates to the migration stage) of the three genotypes have the ratios $1+s_i:1:1+\sigma_i$, then the ratios of the numbers of survivors at the time of migration are $(1+\sigma_i)u_n(i)^2:2u_n(i)(1-u_n(i)):(1+s_i)(1-u_n(i))^2$. We suppose that the total number of individuals in the ith deme who survive to migrate is a fixed carrying capacity p_i, which does not depend upon the genotypic makeup of the population. We further suppose that a fraction l_{ji} of the population of each genotype in the ith deme migrates to become part of deme j. (Of course the fraction l_{ii} remains in the ith deme.) Under our assumptions the gene fraction $u_{n+1}(j)$ born in the jth deme in the next generation is equal to the gene fraction in the jth deme after migration. It is given by the formula

$$u_{n+1}(j)=\sum_i m_{ji}g_i(u_n(i)), \tag{2.1}$$

where

$$g_i(u)=\frac{2(1+s_i)u^2+2u(1-u)}{2[(1+s_i)u^2+2u(1-u)+(1+\sigma_i)(1-u)^2]} \tag{2.2}$$

is the gene fraction of the ith deme just before the migration and

$$m_{ji}=\frac{l_{ji}p_i}{\sum_k l_{jk}p_k} \tag{2.3}$$

is the fraction of those individuals who are in the jth deme after the migration who came from the ith niche.

Thus the function $\{u_n(i): i=1, 2, \cdots\}$ satisfies a recursion of the form (1.2) with

$$Q[u](j)=\sum_i m_{ji}g_i(u(i)). \tag{2.4}$$

In the present work we will only deal with a homogeneous habitat. By this we mean that all the niches are identical, that they are obtained by applying the elements of a group of translations to any one niche, and that the migration function l_{ij} depends only on the translation which takes the ith niche into the jth niche.

Suppose, for instance, that the organisms are living in a plane $\mathscr{R}^2$. This plane is divided into the squares $\{(x, y)|(k-\frac{1}{2})h \leqq x < (k+\frac{1}{2})h,\ (l-\frac{1}{2})h \leqq y < (l+\frac{1}{2})h;\ k, l = 0, \pm 1, \pm 2, \cdots\}$ of side h, which are the niches. We label each niche by the coordinates of its center, which are integral multiples of h. Then we can think of the habitat as the set of points of $\mathscr{R}^2$ whose coordinates are integral multiples of h. We shall usually write these centers as vectors x. $\mathscr{H}$ can be obtained from any one niche by applying all translations by two-vectors whose components are multiples of h.

The fact that $\mathscr{H}$ is homogeneous means that the fitnesses s and σ and the adult population p are the same for all the niches, and the migration function l_{ij} depends only on the vector difference x_i-x_j between the centers of the niches. For the moment we also assume that s, σ, and p do not depend on u, so that they are constants. Since

$$\sum_k l_{jk}=\sum_k l(x_j-x_k)=\sum_i l(x_i)=\sum_i l_{io}=1$$

because all the individuals in niche x_o must go somewhere, (2.3) becomes

$$m_{ij}=l_{ij}\equiv m(x_i-x_j).$$

Thus the operator Q in the recursion (1.2) is

$$Q[u](x) = \sum_{y \in \mathcal{H}} m(x-y)g(u(y)), \tag{2.5}$$

where $\sum_{x \in \mathcal{H}} m(x) = 1$ and

$$g(u) = \frac{su^2 + u}{1 + su^2 + \sigma(1-u)^2}. \tag{2.6}$$

We see that

$$g(u) - u = \frac{u(1-u)[su - \sigma(1-u)]}{1 + su^2 + \sigma(1-u)^2}. \tag{2.7}$$

The definition of u shows that we must only consider functions $u(x)$ such that $0 \leqq u \leqq 1$. It is easily seen that g increases from 0 to 1 as u increases from 0 to 1. Therefore $Q[u]$ again has values on the interval [0, 1].

We see from (2.7) that there are three possible behaviors of the function g:

(i) If $s > 0 > \sigma$, the AA homozygotes are the most fit and the aa homozygotes are the least fit, and

$$g(u) > u \quad \text{for } 0 < u < 1. \tag{2.8}$$

This is called the *heterozygote intermediate* case. (Note that if $s < 0 < \sigma$, then $g(u) < u$. We can reduce this to the above case by replacing the variable u by $1 - u$, which amounts to interchanging the labels A and a.)

(ii) If s and σ are both negative, the heterozygote is more fit than either of the homozygotes, and we refer to this as the *heterozygote superior* case. (2.7) shows that this case is characterized by the fact that

$$g(u) \begin{cases} > u & \text{for } 0 < u < \pi_1, \\ < u & \text{for } \pi_1 < u < 1, \end{cases} \tag{2.9}$$

where

$$\pi_1 = \frac{\sigma}{s + \sigma}. \tag{2.10}$$

(iii) When s and σ are positive, we speak of the *heterozygote inferior* case. Here

$$g(u) \begin{cases} < u & \text{for } 0 < u < \pi_0, \\ > u & \text{for } \pi_0 < u < 1, \end{cases} \tag{2.11}$$

where π_0 is given by the same formula (2.10) as π_1.

There is, of course, no particular reason why the fitnesses $1 + s$ and $1 + \sigma$ should not depend upon u, that is, upon the composition of the competing population. As long as p is constant, we still obtain (2.5) and (2.6). The formula (2.7) shows that even when s and σ depend upon u, it is true that in the heterozygote intermediate case $s > 0 > \sigma$ the inequality (2.8) is satisfied.

However, the following example shows that the inequalities (2.8), (2.9), and (2.11) do not characterize the three cases when s and σ depend upon u.

Example 2.1. When $s = \varepsilon(1-u)(2 + \sin k\pi u)$, $\sigma = 2\varepsilon u$, where ε is any positive constant below $\frac{1}{3}$, the system is heterozygote inferior, but for $k = 1$ the corresponding g has the property (2.8), for $k = 2$ it has the property (2.9) with $\pi_1 = \frac{1}{2}$, and for larger k the

function $g-u$ has arbitrarily many sign changes. By multiplying $\sin k\pi u$ by $(-1)^k$ and making ε negative, one obtains a similar example for the heterozygote superior case.

The population size p before the migration may also depend on the genetic composition of the population and hence on u. Then (2.3) shows that in a homogeneous habitat m_{ji} is a function of $u(x_i)$ as well as of x_j-x_i. More generally, the migration pattern may depend upon genotype, as may the fertility.

If we assume no further selection after migration, one can hope to predict the numbers $l_A(x-y, u)$ of A gametes and $l_a(x-y, u)$ of a gametes produced in the niche at x by the deme born at y if its initial gene fraction is u. In this case we obtain the recursion (1.2) with the operator

$$Q[u](x)=\frac{\sum_{y\in\mathscr{H}} l_A(x-y, u(y))}{\sum_{y\in\mathscr{H}}[l_A(x-y, u(y))+l_a(x-y, u(y))]}. \tag{2.12}$$

If there is nonrandom mating, the new gene fraction is a function of the right-hand side.

The above model can be modified in various other ways. The niches may consist of parallelograms, hexagons, or other regions instead of squares. We can consider the limiting case of very small regions, so that $u_n(x)$ becomes a function of the continuous variable x and the formula (2.5) for Q is replaced by

$$Q[u](x)=\int_{\mathscr{R}^2} m(x-y)g(u(y))\,dy. \tag{2.13}$$

If $m(x)$ depends only on the Euclidean distance $|x|$, the habitat is isotropic (rotationally symmetric). This case was treated in previous works [11], [12], [48], [49], [53], [54]. A discrete habitat cannot be rotationally symmetric[1], and therefore we need to treat nonisotropic operators Q here. Because of such factors as prevailing winds and the position of the sun, the migration may well be nonisotropic even in a continuous homogeneous habitat.

A continuous analogue of the more general operator (2.12) can also be obtained by replacing the sums by integrals.

Fisher's equation (1.1) can be obtained from the recursion (1.2) with Q defined by (2.5) by taking a suitable limit in which both the size of the niches and the length of the time step approach zero.

On the other hand, if we define $u_n(x)=u(n\tau, x)$, where $u(t, x)$ is a solution of Fisher's equation (1.1), then u_n satisfies the recursion (1.2) where $Q[v](x)$ is defined to be $u(\tau, x)$ with $u(t, x)$ the solution of the initial value problem

$$\frac{\partial u}{\partial t}=D\Delta u+f(u), \qquad u(0, x)=v(x). \tag{2.14}$$

This remark remains true when D and f are periodic functions of t of period τ, so that seasonal variations in the rates of migration and selection can be taken into account. Nonisotropy in the migration can be put into the model by replacing the Laplace operator by a more general second order elliptic operator, whose coefficients are independent of x but may depend periodically on t. In fact, these coefficients and, in particular, the diffusion constant D in (2.14), may also depend on u.

[1] Rotational symmetry in the natural habitat can be destroyed by aggregating over a discrete set of niches. For example, if the population density in $\mathscr{R}^2$ is $(x^2+y^2)^4$, which is radially symmetric, the population in a square of side h centered at (ih, jh) is a polynomial in i^2+j^2 plus a multiple of i^2j^2, so that it is not invariant under rotations about the origin.

All the above models involve various assumptions which may or may not be satisfied. One can think of the recursion (1.2) itself as the model. It says that the gene fraction of each deme of one generation is uniquely determined by the set of all gene fractions of the preceding generation. If the niches are not too small relative to the maximum distance of migration, $Q[u](x)$ only depends on the values of u at a few nearby niches, and the operator Q can be determined experimentally by varying these values and watching the outcome.

The major assumption inherent in the class of models (1.2) is that the number of individuals born in each deme is so large that the numbers of individuals of the three genotypes who survive to adulthood do not depend upon the sizes of the initial populations but only on their gene fraction. It is this assumption which allows us to work with a scalar-valued rather than a vector-valued dependent variable.

One must, of course, also assume that all the external influences on growth, selection, and migration remain the same for all generations.

Recursions of the form (1.2) can serve as models for many other biological processes. We may for example consider the population growth of a single migrating species with synchronized nonoverlapping generations. Let the habitat again consist of the plane $\mathscr{R}^2$ divided into squares centered at (ih, jh), and let $u_n(x)$ represent the size of the population which is in the square centered at x at a certain part of the life cycle of the nth generation. If one assumes that as far as interaction with the species under consideration is concerned, the rest of the world does not change in time, one obtains a model of the form (1.2) where Q may be measured experimentally.

If $u_n(x)$ is the number of individuals in the square centered at x just after migration, these individuals produce $g(u_n)$ new individuals and die, where $g(u)$ is a function whose graph is called the *reproduction curve* [44]. A fraction $m(y-x)$ of these new individuals migrates to y. Thus one obtains the model (2.5). One can again obtain space limits such as (2.13) by letting the size of the squares approach zero and partial differential equations like the Fisher equation (1.1) by letting both the square size and the length of the life cycle approach zero.

In all the above models we have made the assumption of determinism. That is, we assume that u_{n+1} is completely determined by the function u_n, the assignment being given by the operator Q. In nature things are, of course, not that simple, and the map from u_n to u_{n+1} is more likely to be a Markov process. That is, given the function u_n, one has a probability distribution for the function u_{n+1}.

It may, however, happen that, if u_n is known, one can say with probability one that $u_{n+1}(x)$ lies in a certain interval $[v^-(x), v^+(x)]$ for each value of x. The function $v^-(x)$ and $v^+(x)$ are determined by u_n and hence are given by applying two operators Q^- and Q^+ to u_n. Thus, instead of the recursion (1.2), we are led to the "interval statement"

$$Q^-[u_n] \leqq u_{n+1} \leqq Q^+[u_n]. \tag{2.15}$$

We shall show in § 4 that if the operators Q^- and Q^+ have some of the properties mentioned in this section, the theorems of § 6 will give information about the behavior of u_n for large n.

3. Formulation of the problem. In this section we shall formulate a mathematical model which contains many of the models discussed in the preceding section as special cases.

Since we wish to treat spaces of one, two, and three dimensions, we shall work in a general Euclidean space $\mathscr{R}^N$ of N dimensions. We shall often identify a point of $\mathscr{R}^N$

with the vector from the origin to this point, so that we can speak of the vector sum of two points and of the point αx where α is a scalar.

DEFINITION. A *habitat* $\mathcal{H}$ is defined to be a set of points in $\mathcal{R}^N$ with the following property: *If x and y are in $\mathcal{H}$, then $x+y$ and $x-y$ are also in $\mathcal{H}$.* This property implies that the origin 0 is in $\mathcal{H}$. (In fact, $\mathcal{H}$ is a group under addition.)

In order to avoid trivialities we shall assume that N is actually the dimension of $\mathcal{H}$. That is, we suppose that *there is no unit vector which is orthogonal to all of $\mathcal{H}$.* We also suppose that $N \geqq 1$ so that $\mathcal{H}$ contains a point other than the origin. Since it contains all multiples of this point, $\mathcal{H}$ is unbounded.

In all our applications the dependent variable u has only nonnegative values. If u is a gene fraction, its values lie on the interval $[0, 1]$. A population size will normally vary between 0 and some large upper bound π_+, but this upper bound can conceivably be $+\infty$.

DEFINITION. B is the set of continuous functions defined on $\mathcal{H}$ with values on the interval $[0, \pi_+]$ or, if $\pi_+ = \infty$, $[0, \infty)$.

For any y in $\mathcal{H}$ the translation operator

$$T_y u(x) \equiv u(x-y)$$

is defined.

Our basic evolution law is the recursion

$$u_{n+1} = Q[u_n],$$

where u_n and u_{n+1} are elements of B and Q is a given operator on B. The properties of the model are the properties of Q.

Since we always deal with nonnegative quantities, we assume that $Q[0] \geqq 0$.

The homogeneity of the habitat is equivalent to the assumption that Q commutes with any translation. That is, $Q[T_y[u]] = T_y[Q[u]]$ for all u in B.

A constant function α is clearly translation invariant. That is, $T_y\alpha = \alpha$ for all y. Consequently, $T_y Q[\alpha] = Q[\alpha]$ for all y, which means that $Q[\alpha]$ is again a constant. This simply states that in a homogeneous habitat the effects of migration cancel when u is constant. Thus, the properties of the model in the absence of migration can be found by looking at what Q does to constant functions.

Because $\sum_{x \in \mathcal{H}} m(x) = 1$ in the model (2.5), we see that $Q[\alpha] = g(\alpha)$ there. We shall therefore require the function $Q[\alpha]$ to have properties like those of the function g. In the genetics model with constant p, g is given by (2.6) and as long as the fitnesses are independent of u we have the three possibilities (2.8), (2.9) and (2.11). We can cover all three by saying that there are constants π_0 and π_1 with $0 \leqq \pi_0 < \pi_1 \leqq 1$ such that

$$Q[\alpha] \begin{cases} < \alpha & \text{for } \alpha \in (0, \pi_0), \\ > \alpha & \text{for } \alpha \in (\pi_0, \pi_1), \\ < \alpha & \text{for } \alpha \in (\pi_1, 1). \end{cases}$$

Then $\pi_0 = 0$, $\pi_1 = 1$ give the heterozygote intermediate case, $\pi_0 = 0$, $\pi_1 \in (0, 1)$ the heterozygote superior case, and $\pi_0 \in (0, 1)$, $\pi_1 = 1$ the heterozygote inferior case. In the absence of migration, π_1 is a stable equilibrium point and π_0 is an unstable one.

Example 2.1 shows that when the fitnesses vary with u these correspondences may no longer apply, and the function $g(u) - u$ may have more sign changes. We shall formulate our hypotheses so that we can also treat these cases by only requiring that $Q[\alpha] > \alpha$ in some interval (π_0, π_1) with $Q[\pi_0] = \pi_0$ and $Q[\pi_1] = \pi_1$. There may be more than one such interval.

The above properties of $Q[\alpha]$, with 1 replaced by some upper limit π_+, also include any reproduction curves $g(\alpha)$. In this case π_0 is a critical population size below which the population tends to decrease to 0 and π_1 is a carrying capacity, to which the population size tends in the absence of migration. For some populations $g(u)-u$ may have more than two sign changes so that there are two or more stable equilibria. See [33], [34] for an example of such a population.

We shall also assume that increasing u everywhere in one generation increases u everywhere in the next one. More precisely, if $u(x) \geqq v(x)$ for all x, then $Q[u](x) \geqq Q[v](x)$ for all x. In the case of the model (2.5), this means that the function $g(u)$ is nondecreasing. This is certainly true of the function (2.6) when s and σ do not depend upon u and $1+s$ and $1+\sigma$ are positive.

The operator (2.12) has the above property if the number $l_A(x-y, u)$ of A gametes produced at x by the deme at y is an increasing function of the gene fraction u at y while the number l_a of a gametes is decreasing in u. This is a plausible property, although one can certainly think of migratory behavior in which it is violated.

In the model (2.5) for population growth we shall assume that the reproduction curve is nonincreasing. This excludes the case of a humped reproduction curve, which even in the absence of migration leads to oscillations [42], [52] and sometimes to chaos [23], [32], [36]. A more complicated population growth model can be treated as long as an increase in $u_n(i)$ results in increased migration to every deme.

In addition to these assumptions we shall require $Q[u]$ to behave continuously with respect to changes in u. We summarize these basic hypotheses:

(i) $Q[u] \in B$ for all $u \in B$.

(ii) $Q[T_y[u]] = T_y[Q[u]]$ for all $u \in B$, $y \in \mathcal{H}$.

(iii) There are constants $0 \leqq \pi_0 < \pi_1 \leqq \pi_+$ such that

$$Q[\alpha] > \alpha \quad \text{for } \alpha \in (\pi_0, \pi_1), \quad Q[\pi_0] = \pi_0, \quad Q[\pi_1] = \pi_1 \quad \text{if } \pi_1 < \infty. \tag{3.1}$$

(iv) $u \leqq v$ implies that $Q[u] \leqq Q[v]$.

(v) $u_n \to u$ as $n \to \infty$ uniformly on each bounded subset of $\mathcal{H}$ implies that $Q[u_n](x) \to Q[u](x)$ for each $x \in \mathcal{H}$.

The operator Q will be understood to satisfy these conditions throughout the remainder of this work.

As we shall see, these hypotheses suffice to define wave speeds and to determine the large-time behavior of solutions u_n which for $n=0$ are less than π_1 and vanish outside a bounded set. If we wish to consider cases where u_0 is at least π_1 at some points, we need some further hypotheses about the action of Q on such functions. We first assume that

$$Q[\alpha] < \alpha \quad \text{for } \alpha \in (\pi_1, \pi_+) \quad \text{if } \pi_1 < \pi_+. \tag{3.2}$$

The following hypothesis is a generalized form of a strong maximum principle.

(3.3) There are two convex subsets K_1 and K_2 of $\mathcal{R}^N$ at a positive distance from each other and an ε_1 which is positive if $\pi_1 < \pi_+$ and 0 if $\pi_1 = \pi_+$ with the property that, if $u \leqq \pi_1 + \varepsilon_1$ on $\mathcal{H}$ and $u < \pi_1$ on either K_1 or K_2, then $Q[u](0) < \pi_1$.

This hypothesis excludes, of course, the trivial case of no migration at all.

Finally we may wish to restrict the range of migration of any individual. In any biologically realistic model this range is finite, which implies that there is a constant b

such that

$$u(x)=0 \quad \text{for } |x|\leqq b \Rightarrow Q[u](0)=0. \tag{3.4}$$

This condition implies that $Q[0]=0$, so that it excludes mutation into the allele A in the genetic model and spontaneous generation in the population model.

In order to allow Q to be the solution operator of (1.1), which permits a small leakage to infinity in a finite time, we can make the following weaker form of this hypothesis.

(3.5) There is a nonincreasing function $h(s)$ defined for $s\geqq 0$ such that (i) if, for some unit vector ξ and any sufficiently small positive β, $u(x)\geqq\beta$ for $x\cdot\xi\leqq 0$, then $Q[u]\geqq\beta h(x\cdot\xi)$ for all $x\in\mathcal{H}$ with $x\cdot\xi\geqq 0$; (ii) to each integer m and each $\alpha\in(0,\pi_1)$ there correspond a positive integer k_m and a constant β_m such that if $u_{n+1}=Q[u_n]$ and if, for some unit vector ξ, $u_0(x)=0$ for $x\cdot\xi\geqq 0$, then $u_m(x)\leqq\alpha[h(x\cdot\xi/k_m)]^{k_m}$ for $x\cdot\xi\geqq\beta_m$. Moreover, $h(\infty)=0$.

If (3.4) is valid, this hypothesis clearly holds with $h(s)\equiv 0$ and $\beta_m=mb$. When Q is the solution at the period time τ of Fisher's equation (1.1) with $D(t)\not\equiv 0$ and $f(v,t)$ bounded above and below by constant multiples of v, one can let h be a multiple of

$$\operatorname{erfc}\left[\frac{1}{2}s\left\{\int_0^\tau D(t)\,dt\right\}^{-1/2}\right] \quad \text{and} \quad k_m=2m.$$

If we allow mutation to the allele A or spontaneous generation so that $Q[0]>0$, we do not need any restriction on migration, provided we assume the technical condition that there is a positive constant $\bar{\gamma}$ such that

$$\gamma<Q[\gamma]<\bar{\gamma} \quad \text{for } \gamma\in[0,\bar{\gamma}), \qquad Q[\bar{\gamma}]=\bar{\gamma}. \tag{3.6}$$

In order to establish the existence of traveling waves (Theorem 6.6) we shall need the following compactness property:

(3.7) Every sequence v_n of functions in B with $v_n\leqq\pi_1$ has a subsequence v_{n_i} such that the sequence $Q[v_{n_i}]$ converges uniformly on every bounded subset of $\mathcal{H}$.

We are given an initial function u_0 in B and we shall be interested in predicting the behavior for large n of the sequence of functions $u_n(x)$ which is determined by the recursion

$$u_{n+1}=Q[u_n].$$

We shall also be interested in traveling wave solutions of this recursion. By this we mean solutions of the form

$$u_n(x)=W(x\cdot\xi-nc),$$

where W is a nonconstant function of one variable, ξ is a fixed unit vector (the normal to the waves), and c is a parameter called the wave speed. The function $W(s)$ needs only to be defined at those values of s which are of the form $x\cdot\xi-nc$ with x in $\mathcal{H}$ and $n=0,1,2,\cdots$.

We note that the operator Q is easily extended to an operator on the set $\tilde{B}$ of continuous functions on all of $\mathcal{R}^N$ with values in $[0,\pi_+]$ by means of the definition

$$\tilde{Q}[u](x_0)=Q[T_{-x_0}[u]](0) \quad \forall x_0\in\mathcal{R}^N. \tag{3.8}$$

The function $T_{-x_0}[u]$ on the right is to be interpreted as the restriction of the function $u(x-x_0)$ to $\mathcal{H}$. Because of the translation invariance (3.1.ii), the restriction of $\tilde{Q}[u]$ to $\mathcal{H}$ coincides with the function obtained by applying Q to the restriction of u to $\mathcal{H}$, so that $\tilde{Q}$ is indeed an extension of Q. It is easily verified that if Q has any of the properties (3.1), (3.2), (3.3), (3.4), (3.5), or (3.6), the same is true of the extension $\tilde{Q}$ when $\mathcal{H}$ is replaced by $\mathcal{R}^N$.

For this reason there is no harm in thinking of $\mathcal{H}$ as $\mathcal{R}^N$ except in the last two theorems, Theorems 6.5 and 6.6, in which further assumptions about Q are made which do not necessarily extend to $\tilde{Q}$. In particular, the property (3.7) need not be preserved by this extension.

4. Two basic propositions. In this section we shall state and prove some propositions which will be used in our proofs. The first of these is a comparison principle.

PROPOSITION 4.1. (comparison principle) *Let R be an operator from B to B which is order preserving in the sense that*

$$v \geqq w \Rightarrow R[v] \geqq R[w]. \tag{4.1}$$

If the sequence v_n satisfies the inequalities

$$v_{n+1} \geqq R[v_n] \tag{4.2}$$

while the sequence w_n satisfies

$$w_{n+1} \leqq R[w_n], \tag{4.3}$$

and if $v_0 \geqq w_0$, then $v_n \geqq w_n$ for all n.

Proof. Suppose that $v_n \geqq w_n$. Then by (4.2), (4.1) and (4.3)

$$v_{n+1} \geqq R[v_n] \geqq R[w_n] \geqq w_{n+1}.$$

Thus the proposition is proved by induction.

PROPOSITION 4.2. *Let R have the order-preserving property* (4.1), *and suppose that $R[w_0] \geqq w_0$. If the sequence w_n is defined by the recursion*

$$w_{n+1} = R[w_n],$$

then $w_{n+1} \geqq w_n$ for all n.

Proof. Let $v_n = w_{n+1}$. Then $v_0 \geqq w_0$, and the statement follows from Proposition 4.1.

Proposition 4.1 can often be used to obtain a robustness result for a stochastic model. Suppose that the evolution $u_n \to u_{n+1}$ is a Markov process which is bounded by two deterministic processes in the sense that the inequalities (2.15) are satisfied with probability one, and suppose that Q^+ and Q^- are order preserving operators.

We see from Proposition 4.1 that if the sequences u_n^- and u_n^+ are defined by

$$u_{n+1}^- = Q^-[u_n^-], \qquad u_{n+1}^+ = Q^+[u_n^+],$$

and if $u_0^- \leqq u_0 \leqq u_0^+$, then

$$u_n^- \leqq u_n \leqq u_n^+$$

for all n.

If the operators Q^- and Q^+ satisfy the hypotheses (3.1) and possibly some of the other hypotheses of § 3, then the theorems of § 6 can be combined with these inequalities to give information about the behavior of the random sequence u_n for large n. In particular, one can sometimes determine the qualitative behavior of u_n without

knowing much about the probability distribution of the Markov process. Such a result can be interpreted as a statement of robustness.

5. The wave speed. In this section we shall define a wave speed $c^*(\xi)$ corresponding to any operator Q which satisfies the hypotheses (3.1). The name wave speed will be justified by the theorems which are stated in the next section. For the moment, $c^*(\xi)$ will be defined as a scalar-valued function of unit vectors ξ. We can think of $c^*(\xi)$ as the speed of plane waves whose normal is in the direction ξ.

In order to define c^* we begin by choosing a function $\varphi(s)$ of one real variable with the properties

(5.1) (i) φ is continuous and nonincreasing,

(ii) $\varphi(-\infty)\in(\pi_0, \pi_1)$,

(iii) $\varphi(s)=0$ for $s\geqq 0$.

For any real number c and any unit vector ξ we define the operator

$$R_{c,\xi}[a](s)\equiv\max\{\varphi(s), Q[a(x\cdot\xi+s+c)](0)\} \tag{5.2}$$

on continuous functions of one real variable $a(s)$ with $0\leqq a\leqq\pi_1$. Here the maximum just means the larger of two numbers for each s. The function $a(x\cdot\xi+s+c)$ is to be regarded as a function of x in $\mathcal{H}$ with ξ, s, and c fixed.

We shall suppose that Q has the properties (3.1). Then $R_{c,\xi}$ still has the property

$$R_{c,\xi}[\alpha]>\alpha \quad \text{for } \alpha\in(\pi_0, \pi_1) \tag{5.3}$$

and the order preserving property (3.1.iv).

We now define a sequence $a_n(c, \xi; s)$ by the recursion

$$a_{n+1}=R_{c,\xi}[a_n], \qquad a_0=\varphi. \tag{5.4}$$

We begin by establishing some simple properties of this sequence.

LEMMA 5.1. *The sequence $a_n(c, \xi; s)$ is nondecreasing in n, nonincreasing in s and c, and continuous in c, ξ, and s.*

Proof. Because $a_0=\varphi$, it is clear that $R_{c,\xi}[a_0]\geqq a_0$. Therefore Proposition 4.2 implies that a_n is nondecreasing in n.

We shall prove the other properties by induction. Suppose that a_n is nonincreasing in s and c. Then if $c'\leqq c$ and $s'\leqq s$,

$$a_{n+1}(c', \xi; s')=\max\{\varphi(s'), Q[a_n(c', \xi; x\cdot\xi+s'+c')](0)\}.$$

Since $s'\leqq s$ and $c'\leqq c$, $\varphi(s')\geqq\varphi(s)$ and $a_n(c', \xi; x\cdot\xi+s'+c')\geqq a_n(c, \xi; x\cdot\xi+s+c)$. The order preserving property of Q shows that

$$a_{n+1}(c', \xi; s')\geqq a_{n+1}(c, \xi; s).$$

Since $a_0(c, \xi; s)=\varphi(s)$, which is nonincreasing in c and s, all the a_n are nonincreasing in c and s by induction.

Finally, let (c_ν, ξ_ν, s_ν) be any sequence which converges to (c, ξ, s) as $\nu\to\infty$. Suppose that a_n is continuous in all its variables. Then $a_n(x\cdot\xi_\nu+s_\nu+c_\nu)$ converges to $a_n(x\cdot\xi+s+c)$ uniformly for x in any bounded subset of $\mathcal{H}$, and $\varphi(s_\nu)$ converges to $\varphi(s)$. We see from the property (3.1v) that $a_{n+1}(c_\nu, \xi_\nu; s_\nu)$ converges to $a_{n+1}(c, \xi; s)$; that is, a_{n+1} is continuous.

Because $a_0=\varphi(s)$, which is continuous, the continuity of all the a_n follows by induction, and the lemma is proved.

We remark that since Q may only be defined on continuous functions, the continuity of the a_n as functions of s is needed in order that the recursion (5.4) really define the sequence a_n.

We now consider the limits of a_n at $s = \pm\infty$.

LEMMA 5.2. *Define the sequence of constants α_n and γ_n by the recursions*

$$\begin{aligned} \alpha_{n+1} &= Q[\alpha_n], \qquad \gamma_{n+1} = Q[\gamma_n], \\ \alpha_0 &= \varphi(-\infty), \qquad \gamma_0 = 0. \end{aligned} \tag{5.5}$$

Then α_n increases to π_1 as $n \to \infty$, γ_n increases to the smallest nonnegative solution $\bar{\gamma}$ of the equation $\bar{\gamma} = Q[\bar{\gamma}]$, and for all n, c, and ξ

$$a_n(c, \xi; -\infty) = \alpha_n, \qquad a_n(c, \xi; +\infty) = \gamma_n. \tag{5.6}$$

Proof. By the property (3.1iii) $Q[\alpha_0] > \alpha_0$, while by (3.1i) $Q[\gamma_0] \geqq \gamma_0$. Therefore by Proposition 4.2 the sequences α_n and γ_n are nondecreasing. By Proposition 4.1, $\gamma_n \leqq \pi_0 < \alpha_n \leqq \pi_1$. Therefore α_n has a limit $\bar{\alpha}$, which may be $+\infty$ if $\pi_1 = +\infty$ and γ_n has a limit $\bar{\gamma}$.

If $\bar{\alpha} < \pi_1$, then $Q[\bar{\alpha}] > \bar{\alpha}$ by (3.1iii) and by (3.1v) $\alpha_{n+1} = Q[\alpha_n] > \bar{\alpha}$ for sufficiently large n. This contradicts the fact that α_n is nondecreasing, and we conclude that $\bar{\alpha} = \pi_1$.

To prove the limit properties of a_n we again use induction. The statement (5.6) is certainly true for $n = 0$. Suppose that $a_n(c, \xi; -\infty) = \alpha_n$. Then the sequence

$$u_k(x) \equiv a_n(c, \xi; x \cdot \xi - k + c)$$

converges to α_n as $k \to \infty$, uniformly on bounded subsets of $\mathscr{H}$. Therefore by (3.1v)

$$\lim_{k \to \infty} Q[a_n(c, \xi: x \cdot \xi - k + c)] = Q[\alpha_n] = \alpha_{n+1}.$$

Because $\alpha_{n+1} \geqq \alpha_0 \geqq \varphi$, we find that

$$\lim_{k \to \infty} a_{n+1}(c, \xi; -k) = \alpha_{n+1}.$$

Since a_{n+1} is nonincreasing in s, we conclude that

$$a_{n+1}(c, \xi; -\infty) = \alpha_{n+1}.$$

The same argument with k replaced by $-k$ shows that $a_{n+1}(c, \xi; +\infty) = \gamma_n$. Thus, (5.6) is established by induction, and the lemma is proved.

It follows from Lemma 5.1 that the sequence a_n increases to a limit function $a(c, \xi; s)$ as n goes to infinity and that $a(c, \xi; s)$ is nonincreasing in s and in c. Because $a_n(c, \xi; s) \leqq a_n(c, \xi; -\infty) = \alpha_n \leqq \pi_1$, we find that $a \leqq \pi_1$. Lemma 5.2 shows that

$$a(c, \xi; -\infty) = \pi_1.$$

(If $\pi_1 = +\infty$, a may have the value $+\infty$ on an interval.)

The value $a(c, \xi; +\infty)$ may or may not be π_1. Since a is nonincreasing in s, $a(c, \xi; +\infty) = \pi_1$ if and only if $a \equiv \pi_1$. We give a criterion to determine whether or not this is the case.

LEMMA 5.3. *The value $a(c, \xi; +\infty) = \pi_1$ if and only if there is an n such that*

$$a_n(c, \xi; 0) > \varphi(-\infty). \tag{5.7}$$

Proof. If $a(c, \xi; +\infty) = \pi_1$, then $a(c, \xi; s) = \pi_1$ for all s, and in particular for $s = 0$.

Because $\pi_1 > \varphi(-\infty)$ and $a_n(c, \xi; 0)$ converges to $a(c, \xi; 0)$, there must be an n such that (5.7) is valid.

Suppose, on the other hand that (5.7) holds for $n = n_0$. Because a_n and φ are nonincreasing, $\varphi = 0$ for $s \geqq 0$, and $\varphi \leqq \varphi(-\infty)$, it follows from the continuity of a_n and φ that there is a positive constant δ such that

$$a_{n_0}(c, \xi; s+\delta) \geqq \varphi(s) = a_0(c, \xi; s)$$

for all s.

Suppose that for some $k > 0$

$$a_{n_0+k}(c, \xi; s+\delta) \geqq a_k(c, \xi; s). \tag{5.8}$$

Then

$$\begin{aligned} a_{n_0+k+1}(c, \xi; s+\delta) &= \max\{a_{n_0}(c, \xi; s+\delta), Q[a_{n_0+k}(c, \xi; x\cdot\xi + s + \delta + c)](0)\} \\ &\geqq \max\{\varphi(s), Q[a_k(c, \xi; x\cdot\xi + s + c)](0)\} \\ &= a_{k+1}(c, \xi; s), \end{aligned}$$

because $a_{n_0+k+1} \geqq a_{n_0}$. Thus we have established the inequality (5.8) by induction. We let $k \to \infty$ to see that

$$a(c, \xi; s+\delta) \geqq a(c, \xi; s)$$

for all s. Since $\delta > 0$ and a is nonincreasing in s, it follows that a is independent of s. Since $a(c, \xi; -\infty) = \pi_1$, we conclude that $a(c, \xi; s) = \pi_1$ for all s, and the lemma is proved.

We now observe that since $a_1(0, \xi; -\infty) = \alpha_1 > \varphi(-\infty)$, there must be a nonnegative constant $t = t(\xi)$ such that

$$a_1(0, \xi; -t) > \varphi(-\infty).$$

Then since $\varphi(-t) \leqq \varphi(-\infty)$, $a_1(0, \xi; -t) = Q[\varphi(x\cdot\xi - t)](0)$, and

$$a_1(-t, \xi; 0) = \max\{\varphi(0), Q[\varphi(x\cdot\xi - t)](0)\} = \max\{0, a_1(0, \xi; -t)\} > \varphi(-\infty).$$

Thus, by Lemma 5.3, $a(-t, \xi; +\infty) = \pi_1$. Hence $a(c, \xi; +\infty) = \pi_1$ for $c \leqq -t$.

We now define

$$c^*(\xi) \equiv \sup\{c \,|\, a(c, \xi; +\infty) = \pi_1\}. \tag{5.9}$$

If $a(c, \xi, +\infty) = \pi_1$ for all c, we set $c^*(\xi) = +\infty$.

Since $a(c, \xi; s)$ is the limit of a nondecreasing family of continuous functions, it is a lower semicontinuous function of c, ξ, and s. Lemma 5.3 shows that $a(c, \xi; +\infty) = \pi_1$ if and only if $a(c, \xi; 0) > \varphi(-\infty)$, and the lower semicontinuity of a shows that the set of (c, ξ) where this is valid is open. This fact has two consequences:

PROPOSITION 5.1. *$a(c, \xi; +\infty) = \pi_1$ if and only if $c < c^*(\xi)$ and $c^*(\xi)$ is a lower semicontinuous function of ξ.*

The sequence a_n depends upon the choice of the function $\varphi(s)$. Consequently $c^*(\xi)$ appears to depend upon this choice. The following lemma shows that this is not the case.

LEMMA 5.4. *Let $\hat{\varphi}(s)$ be any continuous nonincreasing function of s with the properties*

$$\hat{\varphi}(-\infty) \in (\pi_0, \pi_1), \qquad \hat{\varphi}(s) = 0 \quad \text{for } s \geqq 0. \tag{5.10}$$

Define the sequence $\hat{a}_n(c, \xi; s)$ by the recursion

$$\hat{a}_{n+1}(s) = \max\{\hat{\varphi}(s), Q[\hat{a}_n(x\cdot\xi + s + c)](0)\}, \qquad \hat{a}_0(s) = \hat{\varphi}(s), \tag{5.11}$$

so that $\hat{a}_n$ *increases to a nonincreasing limit function* $\hat{a}(c, \xi; s)$. *Then* $\hat{a}(c, \xi; +\infty) = a(c, \xi; +\infty)$.

Proof. Define the sequence of constants $\hat{\alpha}_n$ by the recursion

$$\hat{\alpha}_{n+1} = Q[\hat{\alpha}_n], \qquad \hat{\alpha}_0 = \hat{\varphi}(-\infty).$$

Then $\hat{\alpha}_n$ increases to π_1, and because $\varphi(-\infty) < \pi_1$ there is an integer n_0 such that $\hat{\alpha}_{n_0} > \varphi(-\infty)$. By Lemma 5.2, $\hat{a}_{n_0}(c, \xi; -\infty) = \hat{\alpha}_{n_0}$. Therefore if $n_0 \geqq 1$ there is a $t = t(c, \xi)$ such that

$$\hat{a}_{n_0}(c, \xi; -t) > \varphi(-\infty). \tag{5.12}$$

Because $\hat{a}_{n_0}$ and $\varphi = a_0$ are nonincreasing in s and $\varphi = 0$ for $s \geqq 0$, it follows that

$$\hat{a}_{n_0}(c, \xi; s - t) > \varphi(s). \tag{5.13}$$

Since $\hat{a}_n$ is nondecreasing in n, $\hat{a}_{n+1}(c, \xi; s - t) \geqq \hat{a}_{n_0}(c, \xi; s - t) \geqq \hat{\varphi}(s)$ for $n \geqq n_0$, and hence

$$\begin{aligned} \hat{a}_{n_0+k+1}(c, \xi; s - t) &= \max \{\hat{a}_{n_0}(c, \xi; s - t), Q[\hat{a}_{n_0+k}(c, \xi; x \cdot \xi + s - t + c)](0)\} \\ &\geqq \max \{\varphi(s), Q[\hat{a}_{n_0+k}(c, \xi; x \cdot \xi + s - t + c)](0)\} \\ &= R_{c,\xi}[\hat{a}_{n_0+k}(c, \xi; s - t)](s) \end{aligned}$$

for $k \geqq 0$. We now see from (5.13) and Proposition 4.1 that $\hat{a}_{n_0+k}(c, \xi; s - t) \geqq a_k(c, \xi; s)$ for $k \geqq 0$. By letting $k \to \infty$ and then $s \to \infty$ we obtain the inequality

$$\hat{a}(c, \xi; +\infty) \geqq a(c, \xi; +\infty).$$

We now use the same proof with the roles of φ and $\hat{\varphi}$ reversed to obtain the opposite inequality. Therefore $\hat{a}(c, \xi; +\infty) = a(c, \xi; +\infty)$, which proves the lemma.

Lemma 5.4 shows that the definition (5.9) gives the same value of $c^*(\xi)$ when a is replaced by $\hat{a}$, so that c^* does not depend on the choice of the function φ.

It is important to know the behavior of a as $s \to +\infty$ when $c \geqq c^*(\xi)$.

PROPOSITION 5.2.

$$\lim_{s \to +\infty} a(c, \xi; s) \leqq \pi_0$$

uniformly in c *and* ξ *on every set of the form* $\{(c, \xi) | c \geqq c^*(\xi), c^*(\xi) \leqq C, |\xi| = 1\}$ *where* C *is a constant.*

Proof. We have shown in the proof of Lemma 5.4 that for any (c, ξ) there exist n_0 and t such that the inequality (5.12) is valid, and that this inequality implies that

$$\hat{a}_{n_0+k}(c, \xi; s - t) \geqq a_k(c, \xi; s). \tag{5.14}$$

Because $\hat{a}_{n_0}$ is continuous in c and ξ, the inequality (5.12) is valid in a neighborhood of (c, ξ). The set $\{(c, \xi) | c = c^*(\xi) \leqq C, |\xi| = 1\}$ is bounded, and it is closed because $c^*(\xi)$ is lower semicontinuous. We cover it with finitely many of the neighborhoods and take the largest of the n_0 and the smallest of the t to obtain the inequality (5.12) and hence (5.14) for all (c, ξ) on this set. By Lemma 5.3 $\hat{a}_{n_0+k}(c^*(\xi), \xi; 0) \leqq \hat{\varphi}(-\infty)$. We set $s = t$ in (5.14) to see that if $c = c^*(\xi) \leqq C$

$$a_k(c, \xi; t) \leqq \hat{\varphi}(-\infty).$$

Since a_k is nonincreasing in c, this inequality is valid whenever $c \geqq c^*(\xi)$ and $c^*(\xi) \leqq C$.

We let $k \to \infty$ to see that there is a t such that

$$a(c, \xi; s) \leqq \hat{\varphi}(-\infty) \quad \text{for } s \geqq t$$

when $c \geqq c^*(\xi)$ and $c^*(\xi) \leqq C$.

We now recall that $\hat{\varphi}(-\infty)$ is any number in the interval (π_0, π_1). In particular it may be chosen arbitrarily close to π_0. Since a is nonnegative and nondecreasing, the statement of the proposition follows.

Proposition 5.2 implies that for $c \geqq c^*(\xi)$, $a(c, \xi; +\infty) \leqq \pi_0$. The following lemma gives more precise information.

LEMMA 5.5. *If π_1 is finite, then for any c and ξ the limit $a(c, \xi; +\infty)$ satisfies the equation*

$$a(c, \xi; +\infty) = Q[a(c, \xi; +\infty)].$$

Proof. Suppose that

$$Q[a(c, \xi; \infty)] > a(c, \xi; \infty).$$

By continuity there is a number β such that

$$\beta < a(c, \xi; \infty) < Q[\beta].$$

Choose a continuous nonincreasing function $\psi(s)$ such that

$$\psi(s) = \begin{cases} \beta & \text{for } s \leqq -1, \\ 0 & \text{for } s \geqq 0. \end{cases}$$

For each positive integer k one can find n_k so that

$$a_{n_k}(c, \xi; k) \geqq \beta.$$

Then

$$a_{n_k}(c, \xi; s) \geqq \psi(s - k).$$

Hence

$$a(c, \xi; s) \geqq a_{n_k+1}(c, \xi; s) \geqq Q[\psi(x \cdot \xi + s + c - k)](0).$$

We let $k \to \infty$ and use (3.1.v) to see that $a(c, \xi; s) \geqq Q[\beta]$ for all s. This contradicts the condition $a(c, \xi; \infty) < Q[\beta]$. We conclude that $Q[a(c, \xi; \infty)] \leqq a(c, \xi; \infty)$.

Suppose now that

$$Q[a(c, \xi; \infty)] < a(c, \xi; \infty).$$

We choose $\hat{\beta}$ so that

$$Q[\hat{\beta}] < a(c, \xi; \infty) < \hat{\beta} < \pi_1.$$

Since π_1 is finite, we can choose a continuous nonincreasing function $\hat{\psi}(s)$ such that

$$\hat{\psi}(s) = \begin{cases} \pi_1 & \text{for } s \leqq -1, \\ \hat{\beta} & \text{for } s \geqq 0. \end{cases}$$

We can now choose s_1 so that $a(c, \xi; s_1) \leqq \hat{\beta}$. Then

$$a_n(c, \xi; s) \leqq a(c, \xi; s) \leqq \hat{\psi}(s - 1 - s_1)$$

for all n. Hence

$$a_{n+1}(c, \xi; s) \leqq \max\{\varphi(s), Q[\hat{\psi}(x \cdot \xi + s + c - 1 - s_1](0)\}.$$

We let $n \to \infty$ and then $s \to \infty$ to see that $a(c, \xi; \infty) \leqq Q[\hat{\beta}]$, which contradicts the condition $Q[\hat{\beta}] < a(c, \varphi; \infty)$.

We have shown that $Q[a(c, \xi; \infty)]$ can be neither larger nor smaller than $a(c, \xi; \infty)$, which proves the lemma.

Remark. When π_1 is finite and $\mathcal{H}$ is discrete, then $a(c, \xi; x \cdot \xi + s + c)$ is automatically a continuous function of x in $\mathcal{H}$ and the sequence $a_n(c, \xi; x \cdot \xi + s + c)$ converges to it uniformly on bounded sets. In this case the lemma follows immediately if we let $n \to \infty$ and then $s \to \infty$ in (5.4).

While each $a_n(c, \xi; s)$ is continuous in c, ξ, and s, Lemma 5.3 shows that the limit function $a(c, \xi; s)$ is discontinuous in c at $c = c^*(\xi)$. Moreover, a may also be discontinuous in s. Consider, for example, the operator

$$Q[u](s) = \tfrac{1}{3}u(s+1)[2 - u(s+1)] + \tfrac{2}{3}u(s-1)[2 - u(s-1)]$$

on $\mathcal{H} = \mathcal{R}^1$.

It is obvious from the definition that

$$a(c, 1; s) = 0 \quad \text{for } c \geqq 1, \quad s \geqq 0.$$

Moreover, because

$$Q[u](s+1) \geqq \tfrac{2}{3}u(s)[2 - u(s)] > u(s) \quad \text{when } u(s) < \tfrac{1}{2},$$

we find that if $\varphi(s) > 0$ for $s < 0$,

$$a(1, 1; s) \geqq \tfrac{1}{2} \quad \text{for } s < 0.$$

Thus the function $a(1, 1; s)$ has a jump at $s = 0$.

It is easily seen that $a(c, 1; s) \equiv 1$ for $c < 1$, so that $c^*(1) = 1$ in this example.

Of course, since $a(c, \xi; s)$ is lower semicontinuous and nonincreasing in c and in s, it is continuous from the right in c and in s.

The following three properties of the wave speed give an intuitive feeling for this function. Their proofs are obvious.

PROPOSITION 5.3. *If $c^*(\xi)$ is the wave speed corresponding to the operator Q, then the wave speed corresponding to the operator $T_y Q$ is $c^*(\xi) + y \cdot \xi$.*

PROPOSITION 5.4. *If Q has the property that for some $\delta \in [0, \pi_0]$*

$$u(x) \leqq \delta \quad \textit{for } |x| < b \quad \textit{implies } Q[u](0) \leqq \delta,$$

then

$$c^*(\xi) \leqq b \quad \textit{for all } \xi.$$

PROPOSITION 5.5. *Let Q_1 and Q_2 be two operators with the properties* (3.1), *where π_0 and π_1 are replaced by $\pi_0^{(1)}$, $\pi_1^{(1)}$ for Q_1 and by $\pi_0^{(2)}$, $\pi_1^{(2)}$ for Q_2. Let $c_i^*(\xi)$ be the wave speed corresponding to Q_i. If $\pi_0^{(2)} \leqq \pi_0^{(1)} < \pi_1^{(1)} \leqq \pi_1^{(2)}$ and*

$$Q_1[u] \leqq Q_2[u]$$

for all continuous functions u on $\mathcal{H}$ with values on $[0, \pi_1^{(1)}]$ then

$$c_1^*(\xi) \leqq c_2^*(\xi)$$

for all ξ.

6. Statement of the theorems. For the convenience of the reader we shall state our results in this section. The proofs will be presented in the subsequent sections.

For any set V of vectors in $\mathcal{R}^N$ we define

$$nV = \{v_1 + v_2 + \cdots + v_n | v_j \in V \text{ for } j = 1, \cdots, n\}. \tag{6.1}$$

It is easily seen that if V is convex, then

$$nV = \{nv | v \in V\}.$$

We denote the unit sphere $\{\xi \in \mathcal{R}^N \| \xi | = 1\}$ by S^{N-1}.

We shall show that, in an asymptotic sense, $c^*(\xi)$ is a propagation speed for arbitrary initial distrubances.

We define the convex set

$$\mathcal{S} = \{x \in \mathcal{R}^N | x \cdot \xi \leqq c^*(\xi) \forall \xi \in S^{N-1}\}. \tag{6.2}$$

If $c^*(\xi)$ were a true propagation speed, so that $u = 0$ for $x \cdot \xi \leqq 0$ implied that $Q[u] = 0$ for $x \cdot \xi \leqq c^*(\xi)$, but not that $Q[u] = 0$ for $x \cdot \xi \leqq c$ when $c < c^*(\xi)$, then it would follow that, if the initial disturbance u_0 is concentrated at the origin, the support of $u_1 = Q[u_0]$ is $\mathcal{S}$. More generally it would follow that if u_n is defined by the recursion $u_{n+1} = Q[u_n]$, the support of u_n is $n\mathcal{S} + \text{supp}\,(u_0)$. (Note that $n\mathcal{S} = \{x | x \cdot \xi \leqq nc^*(\xi) \forall \xi \in S^{N-1}\}$.)

Our first two theorems say that this property is approximately true when n is large. The first theorem says that if the support of u_0 is bounded, then at time n very little disturbance lies far beyond $n\mathcal{S}$, while the second theorem states that the disturbance at time n fills most of $n\mathcal{S}$.

In all the theorems the sequence u_n of functions in B is a solution of the recursion $u_{n+1} = Q[u_n]$, and Q satisfies the hypotheses (3.1).

THEOREM 6.1. *Suppose that the set $\mathcal{S}$ defined by* (6.2) *is bounded and not empty*[2]. *Let $\mathcal{S}'$ be any open set which contains $\mathcal{S}$. Suppose that $u_0 = 0$ outside a bounded set. If $u_0 < \pi_1$ or if $u_0 < \pi_+$ and Q satisfies the additional hypotheses* (3.2), (3.3) *and one of the hypotheses* (3.4), (3.5), *or* (3.6), *then*

$$\limsup_{n \to \infty} \max_{x \notin n\mathcal{S}'} u_n(x) \leqq \pi_0. \tag{6.3}$$

If $\mathcal{S}$ is empty and c^* is bounded, (6.3) *holds when the maximum is over the whole space $\mathcal{H}$.*

Note that if $\pi_0 = 0$, the statements say that the limit is 0.

THEOREM 6.2. *Suppose that the interior of $\mathcal{S}$ is not empty and let $\mathcal{S}''$ be any closed bounded subset of the interior of $\mathcal{S}$. For any $\sigma > \pi_0$ there exist a radius r_σ with the property that if $u_0(x) \geqq \sigma$ on a ball of radius r_σ and if $u_{n+1} = Q[u_n]$, then*

$$\liminf_{n \to \infty} \min_{x \in n\mathcal{S}''} u_n(x) \geqq \pi_1. \tag{6.4}$$

If $u_0 \leqq \alpha < \pi_+$ and if Q satisfies the additional hypothesis (3.2), *then*

$$\lim_{n \to \infty} \min_{x \in n\mathcal{S}''} u_n(x) = \lim_{n \to \infty} \max_{x \in \mathcal{H}} u_n(x) = \pi_1. \tag{6.5}$$

The next two theorems show how to estimate the wave speed in some cases.

[2] For the analogous results when $\mathcal{S}$ is unbounded or when $\mathcal{S}$ is empty but c^* is unbounded see [55].

THEOREM 6.3. *If $m(x, dx)$ is a bounded nonnegative measure on $\mathcal{H}$ with the property that for all continuous u with $0 \leqq u \leqq \pi_1$*

$$Q[u](x) \leqq \int u(x-y)m(y, dy), \tag{6.6}$$

then

$$c^*(\xi) \leqq \inf_{\mu>0} \frac{1}{\mu} \log \int e^{\mu x \cdot \xi} m(x, dx). \tag{6.7}$$

THEOREM 6.4. *Suppose that $l(x, dx)$ is a bounded nonnegative measure on $\mathcal{H}$ with the properties that*

$$\int l(x, dx) > 1$$

and that there is a positive ε such that for all continuous u with $0 \leqq u \leqq \varepsilon$

$$Q[u](x) \geqq \int u(x-y)l(y, dy).$$

Then $\mathcal{S}$ is not empty and

$$c^*(\xi) \geqq \inf_{\mu>0} \frac{1}{\mu} \log \int e^{\mu x \cdot \xi} l(x, dx), \tag{6.8}$$

where the right-hand side is $+\infty$ if the integral on the right diverges for all positive μ.

If the measure l is not concentrated on any hyperplane $x \cdot \xi = \text{constant}$, then $\mathcal{S}$ has interior points, and the radius r_σ in Theorem 6.2 can be chosen so that it does not depend upon σ.

We note the following immediate corollary of Theorems 6.3 and 6.4.

COROLLARY. *If the nonnegative bounded measure $m(x, dx)$ satisfies (6.6) when $0 \leqq u \leqq \pi_1$ and if for every positive δ there is a positive ε such that*

$$Q[u](x) \geqq (1-\delta) \int u(x-y)m(y, dy) \quad \text{when } 0 \leqq u \leqq \varepsilon, \tag{6.9}$$

then

$$c^*(\xi) = \inf_{\mu>0} \frac{1}{\mu} \int e^{\mu x \cdot \xi} m(x, dx) \tag{6.10}$$

for all unit vectors ξ.

The hypotheses of Theorem 6.4 and the corollary imply that $\pi_0 = 0$. This condition is also need for the last two theorems.

THEOREM 6.5. (hairtrigger effect) *Let Q and l satisfy the hypotheses of Theorem 6.4 and suppose that the support of l contains a set $\mathcal{W}$ of vectors in $\mathcal{H}$ with the property that any bounded subset of $\mathcal{H}$ is contained in a translate of the set $n\mathcal{W}$ for some integer n. Then if u_0 is not identically zero on $\mathcal{H}$ and $\mathcal{S}''$ is as in Theorem 2,*

$$\liminf_{n\to\infty} \min_{x \in n\mathcal{S}''} u_n(x) \geqq \pi_1. \tag{6.11}$$

This theorem states that r_σ in Theorem 6.2 may be chosen to be arbitrarily small, regardless of the value of σ.

The last theorem asserts the existence of a traveling wave of speed c with normal ξ whenever $c \geqq c^*(\xi)$. We again require π_0 to be 0 and we need the property (3.7).

THEOREM 6.6. *Suppose that* $\pi_0 = 0$ *and that, in addition to* (3.1), Q *has the compactness property* (3.7). *Then if* $c \geqq c^*(\xi)$, *there is a nonincreasing function* $W(s)$ *which is defined for all* s *of the form* $x \cdot \xi - nc$ *with* x *in* $\mathcal{H}$ *and* n *an integer such that the sequence* $u_n(x) = W(x \cdot \xi - nc)$ *satisfies the recursion* (1.2), $W(-\infty) = \pi_1$, *and* $W(+\infty) = 0$.

7. The rate of propagation. (Proof of Theorem 6.1) We shall use the construction of § 5 in order to prove Theorem 6.1. We begin with an easy lemma.

LEMMA 7.1. *Suppose* $0 \leqq u_0(x) \leqq \alpha < \pi_1$, *and that for some finite collection of unit vectors* $\xi_1, \cdots, \xi_K$ *there is a* ρ *such that for each* i

$$u_0(x) = 0 \quad \text{for } x \cdot \xi_i \geqq \rho - 1.$$

For a positive ε *define the set*

$$\mathcal{M} = \{x | x \cdot \xi_i \leqq c^*(\xi_i) + \varepsilon,\ i = 1, \cdots, K\} \tag{7.1}$$

Then, if u_n *is the solution of the recursion* $u_{n+1} = Q[u_n]$,

$$\limsup_{n \to \infty} \max_{x \notin n\mathcal{M}} u_n(x) \leqq \pi_0. \tag{7.2}$$

Proof. We choose a function $\varphi(s)$ with the properties (5.1) and the additional property

$$\varphi(s) = \alpha \quad \text{for } s \leqq -1.$$

Then

$$u_0(x) \leqq \varphi(x \cdot \xi_i - \rho) \quad \text{for } i = 1, \cdots, K. \tag{7.3}$$

We define the sequence $a_n(c^*(\xi_i), \xi_i; s)$ by the recursion (5.4). Then the sequence

$$v_n(x) = a_n(c^*(\xi_i), \xi_i; x \cdot \xi_i - nc^*(\xi_i) - \rho)$$

satisfies the recursion

$$v_{n+1}(x) = \max\{\varphi(x \cdot \xi_i - (n+1)c^*(\xi_i) - \rho), Q[v_n](x)\} \geqq Q[v_n](x),$$

while (7.3) shows that $v_0(x) \geqq u_0(x)$.

Proposition 4.1 implies that

$$\begin{aligned} u_n(x) \leqq v_n(x) &= a_n(c^*(\xi_i), \xi_i; x \cdot \xi_i - nc^*(\xi_i) - \rho) \\ &\leqq a(c^*(\xi_i), \xi_i; x \cdot \xi_i - nc^*(\xi_i) - \rho) \end{aligned} \tag{7.4}$$

for all n and for $i = 1, \cdots, K$.

If $x \notin n\mathcal{M}$, then $x \cdot \xi_i > n(c^*(\xi_i) + \varepsilon)$ for some i. Because a is nonincreasing in s, we find that

$$\max_{x \notin n\mathcal{M}} u_n \leqq \max_{1 \leqq i \leqq K} a(c^*(\xi_i), \xi_i; n\varepsilon - \rho). \tag{7.5}$$

The statement of the lemma now follows from Proposition 5.2.

In order to use Lemma 5.1 we need to show that there is a set $\mathcal{M}$ of the form (7.1) inside $\mathcal{S}'$.

LEMMA 7.2. *If $\mathcal{S}$ is bounded and nonempty, and if the open set $\mathcal{S}'$ contains $\mathcal{S}$, then there is a set $\mathcal{M}$ of the form* (7.1) *such that*

$$\mathcal{S} \subset \mathcal{M} \subset \mathcal{S}'.$$

Proof. We begin with the well-known fact [6, § 27] that a bounded convex set $\mathcal{S}$ can be approximated by an open convex polyhedron

$$P = \{x | x \cdot \eta_\alpha < b_\alpha, \alpha = 1, \cdots, L\}$$

in the sense that if $\mathcal{S}$ lies in the interior of $\mathcal{S}'$, there is a P such that $\mathcal{S} \subset P \subset \mathcal{S}'$. Because $\mathcal{S}$ is bounded, it lies in the interior of a closed ball B_R, and we may choose P so that its closure lies in the interior of B_R.

Because $\mathcal{S}$ lies in P, its intersection with each half-space $\{x | x \cdot \eta_\alpha \geqq b_\alpha\}$ is empty. Because of the definition (6.2) of $\mathcal{S}$ and because $\mathcal{S} \subset B_R$, we can write this fact in the form

$$\bigcap_{\substack{|\xi|=1 \\ \delta>0}} [B_R \cap \{x | x \cdot \eta_\alpha \geqq b_\alpha, x \cdot \xi \leqq c^*(\xi) + \delta\} = \varnothing].$$

Because of the presence of the set B_R, each of the sets in brackets is closed and bounded. Helly's theorem (the finite intersection property) (see, e.g., [6, p. 3]) states that there is a finite collection of the sets in brackets whose intersection is empty. That is, there are unit vectors $\xi_{1\alpha}, \cdots, \xi_{K_\alpha \alpha}$ and positive numbers $\delta_{1\alpha}, \cdots, \delta_{K_\alpha \alpha}$ such that

$$B_R \cap \{x | x \cdot \xi_{j\alpha} \leqq c^*(\xi_{j\alpha}) + \delta_{j\alpha}, j = 1, \cdots, K_\alpha\} \subset \{x | x \cdot \eta_\alpha < b_\alpha\}.$$

It follows that

$$B_R \cap \{x | x \cdot \xi_{j\alpha} \leqq c^*(\xi_{j\alpha}) + \delta_{j\alpha}, j = 1, \cdots, K_\alpha, \alpha = 1, , \cdots, L\} \subset P.$$

Because the closure of P lies in the interior of B_R and all the sets involved are convex, it follows that the set in braces is a subset of P. If we let $\xi_1, \cdots, \xi_K$ be an ordering of the vectors $\xi_{j\alpha}$ and $\varepsilon = \min \delta_{j\alpha}$ we find that the set $\mathcal{M}$ defined by (7.1) lies in P.

Because $P \subset \mathcal{S}'$, we have $\mathcal{M} \subset \mathcal{S}'$. It is obvious that $\mathcal{M} \supset \mathcal{S}$, so the lemma is proved.

If $u_0 < \pi_1$ and $u_0 = 0$ outside a bounded set, then because u_0 is continuous there is an $\alpha \in (\pi_0, \pi_1)$ such that $u_0 \leqq \alpha$. Thus the statement (6.3) of Theorem 6.1 is an immediate consequence of Lemmas 7.1 and 7.2 in this case.

Suppose now only that u_0 is less than π_+ and vanishes outside a bounded set, and that the hypotheses (3.2) and (3.3) are valid. Because u_0 is continuous there is a $\nu \in [\pi_1, \pi_+)$ such that $u_0 \leqq \nu$.

If we define the sequence of constants ν_n by the recursion $\nu_{n+1} = Q[\nu_n]$, $\nu_0 = \nu$, this sequence decreases to π_1 and there is an integer m_0 which depends only on ν such that $\nu_{m_0} \leqq \pi_1 + \varepsilon_1$, where ε_1 is the constant in (3.3). By Proposition 4.1,

$$u_n \leqq \pi_1 + \varepsilon_1 \quad \text{for } n \geqq m_0.$$

We recall the nondecreasing sequence γ_n defined in (5.5). We wish to show that $u_n \geqq \gamma_n$ and

$$\lim_{|x| \to \infty} u_n(x) = \gamma_n$$

for all n. Suppose this is true for some value of n. Then $u_{n+1} \geqq \gamma_{n+1}$.

Assume that there is a positive δ and a sequence x_k with $|x_k| \to \infty$ such that

$$u_{n+1}(x_k) \geqq \gamma_{n+1} + \delta.$$

By taking a subsequence if necessary, we may assume that $x_k/|x_k|$ converges to a unit vector η. Choose a nonincreasing continuous function $\psi(s)$ such that $\psi(\infty) = \gamma_n$ and

$$u_n(x) \leqq \psi(x \cdot \eta).$$

Then

$$u_{n+1}(y) \leqq Q[\psi(x \cdot \eta)](y) = Q[\psi(x \cdot \eta + y \cdot \eta)](0).$$

In particular,

$$u_{n+1}(x_k) \leqq Q[\psi(x \cdot \eta + x_k \cdot \eta)](0).$$

Since $x_k \cdot \eta \to \infty$, the sequence of functions $\psi(x \cdot \eta + x_k \cdot \eta)$ converges to γ_n uniformly on bounded sets. Thus by (3.1.v),

$$\limsup_{k \to \infty} u_{n+1}(x_k) \leqq \gamma_{n+1}$$

contrary to the assumption that $u_{n+1}(x_k) \geqq \gamma_n + \delta$. We conclude that $u_{n+1}(x)$ converges to γ_{n+1} as $|x| \to \infty$. Since $u_0 = \gamma_0 = 0$ outside a bounded set, we have shown by induction that

$$\lim_{|x| \to \infty} u_n(x) = \gamma_n \quad \text{for } n \geqq 0.$$

Since $\gamma_n \leqq \pi_0 < \pi_1$, we see that $u_{m_0}(x) < \pi_1$ outside a bounded set, whose diameter we call D.

Suppose that for some $m_1 > m_0$ there is a point $\bar{x}$ such that $u_{m_1}(\bar{x}) \geqq \pi_1$. By (3.3) and the translation invariance of Q, there must be vectors x_1 in K_1 and y_1 in K_2 such that $u_{m_1-1}(\bar{x} + x_1) \geqq \pi_1$ and $u_{m_1-1}(\bar{x} + y_1) \geqq \pi_1$. By the same reasoning, there must be x_2 in K_1 and y_2 in K_2 such that $u_{m_1-2}(\bar{x} + x_1 + x_2) \geqq \pi_1$ and $u_{m_1-2}(\bar{x} + y_1 + y_2) \geqq \pi_1$. Proceeding in this way, we find that $u_{m_0}(\bar{x} + x_1 + x_2 + \cdots + x_{m_1-m_0}) \geqq \pi_1$ and $u_{m_0}(\bar{x} + y_1 + \cdots + y_{m_1-m_0}) \geqq \pi_1$ where $x_i \in K_1$ and $y_i \in K_2$.

Because K_1 and K_2 are convex, $(m_1 - m_0)^{-1}(x_1 + \cdots + x_{m_1-m_0}) \in K_1$ and $(m_1 - m_0)^{-1}(y_1 + \cdots + y_{m_1-m_0}) \in K_2$. If d is the positive distance from K_1 to K_2, then the distance from $\bar{x} + x_1 + \cdots + x_{m_1-m_0}$ to $\bar{x} + y_1 + \cdots + y_{m_1-m_0}$ is a least $(m_1 - m_0)d$. Consequently, we must have $(m_1 - m_0)d \leqq D$. Thus we find that if we choose

$$m_1 = m_0 + \frac{D}{d} + 1,$$

then

$$u_{m_1} < \pi_1. \tag{7.6}$$

Because u_{m_1} is continuous and its limit superior at infinity is at most π_0, there is an $\alpha \in (\pi_0, \pi_1)$ such that $u_{m_1} \leqq \alpha$.

If the condition (3.4) is valid, (6.3) now follows from applying Lemmas 7.1 and 7.2 to the sequence u_{m_1+n}.

If (3.6) holds, the sequence γ_n is strictly increasing. We recall that $u_{m_1} \to \gamma_{m_1}$ as $|x| \to \infty$, that $a_{m_1+1}(c^*(\xi), \xi; +\infty) = \gamma_{m_1+1} > \gamma_{m_1}$, and that $u_{m_1}(x) \leqq \alpha < Q[\alpha] \leqq a(c^*(\xi), \xi; -\infty)$. It follows that there are k and ρ such that

$$u_{m_1}(x) \leqq a_{m_1+k}(c^*(\xi), \xi; x \cdot \xi - m_1 c^*(\xi) - \rho)$$

and hence by Proposition 4.2 that

$$u_n(x) \leqq a_{n+k}(c^*(\xi), \xi; x\cdot\xi - nc^*(\xi) - \rho)$$
$$\leqq a(c^*(\xi), \xi; x\cdot\xi - nc^*(\xi) - \rho)$$

for $n \geqq m_1$. Thus we obtain (7.4) and the rest of the proof proceeds as before.

Finally suppose that (3.5) is valid. We see from (3.5i) that if $\varphi(s) = \alpha$ for $s \leqq -1$ then for a sufficiently small β

$$a_1(c^*(\xi), \xi; s) \geqq \beta h(s + c^*(\xi) + 1) \quad \text{for } s \geqq -c^*(\xi) - 1.$$

Since a_1 is nonincreasing, we have for any $s_0 > \max\{0, -c^*(\xi)\}$

$$a_1(c^*(\xi), \xi; s) \geqq \beta h(s_0 + c^*(\xi)) \quad \text{for } s \leqq s_0 - 1.$$

If $h(s_0 + c^*(\xi)) \leqq 1$, then by (3.5.i)

$$a_2(c^*(\xi), \xi; s) \geqq \beta h(s_0 + c^*(\xi)) h(s + c^*(\xi) - s_0 + 1)$$
$$\geqq \beta h(s_0 + c^*(\xi))^2 \quad \text{for } s \leqq 2s_0 - 1.$$

By continuing in this manner and setting $s_0 = (s+1)/k$, we find that

$$a_k(c^*(\xi), \xi; s) \geqq \beta\left[h\left(\frac{s + 1 + kc^*(\xi)}{k}\right)\right]^k \tag{7.7}$$

for all unit vectors ξ as long as $h((s + 1 + kc^*(\xi))/k) \leqq 1$, and $s > \max\{0, -kc^*(\xi)\}$.

From (3.5.ii) and the fact that $u_0 = 0$ for $|x| \geqq r$ we see that

$$u_{m_1}(x + r\xi) \leqq \beta h\left(\frac{x\cdot\xi}{k_{m_1}}\right)^{k_{m_1}} \quad \text{for } x\cdot\xi \geqq \beta_{m_1}.$$

Thus (7.7) shows that there is a constant E such that for every ξ

$$u_{m_1}(x) \leqq a_{k_{m_1}}(c^*(\xi), \xi; x\cdot\xi - k_{m_1}c^* - r - 1) \quad \text{for } x\cdot\xi \geqq E. \tag{7.8}$$

Since $k_{m_1} \geqq 1$,

$$a_{k_{m_1}}(c^*(\xi), \xi; -\infty) \geqq Q[\alpha] > \alpha \geqq u_{m_1}.$$

Since a_k is nonincreasing in s, and continuous in c and ξ, we can find a ρ so that for any given set $\xi_1, \cdots, \xi_K$ of unit vectors

$$u_{m_1}(x) \leqq a_{k_{m_1}}(c^*(\xi_i), \xi_i; x\cdot\xi_i - \rho)$$

for $x\cdot\xi_i \leqq E$ and $i = 1, \cdots, K$. By (7.8) this inequality then holds for all x.

Proposition 4.1 now shows that

$$u_n(x) \leqq a_{k_{m_1}+n-m_1}(c^*(\xi_i), \xi_i; x\cdot\xi_i - (n - m_1)c^*(\xi_i) - \rho)$$
$$\leqq a(c^*(\xi_i), \xi_i; x\cdot\xi_i - nc^*(\xi_i) + m_1c^*(\xi_i) - \rho),$$

as in (7.4), and (6.3) follows from the remainder of the proof of Lemma 7.1 and from Lemma 7.2.

If c^* is bounded and $\mathcal{S}$ is empty, let $\xi_1, \cdots, \xi_{2N}$ be the unit vectors in the coordinate directions and their negatives. We have the representation

$$\mathcal{S} = \bigcap_{\substack{\xi\in S^{N-1} \\ \varepsilon>0}} \{x \mid x\cdot\xi_i \leqq c^*(\xi_i) + \varepsilon,\ i = 1, \cdots, 2N,\ x\cdot\xi \leqq c^*(\xi) + \varepsilon\}$$

of $\mathscr{S}$ as an intersection of closed bounded sets. Since $\mathscr{S}$ is empty, Helly's theorem (the finite intersection property) shows that there are finitely many unit vectors $\xi_{2N+1}, \cdots, \xi_K$ and a positive ε such that the intersection of the halfspaces $\{x | x \cdot \xi_i \leqq c^*(\xi_i)+\varepsilon\}$, $i=1, \cdots, K$ is empty. Then for each y in R^N there is an i such that $y \cdot \xi_i > c^*(\xi_i)+\varepsilon$. We take any x, set $y = n^{-1}x$ and insert the resulting inequality into (7.4). Since a is nonincreasing in s, we find that for every x in $\mathscr{H}$

$$u_n(x) \leqq \max_{1 \leqq i \leqq K} a(c^*(\xi_i), \xi_i; \varepsilon n - \rho),$$

and the last statement of Theorem 6.1 follows from Proposition 5.2.

Thus Theorem 6.1 is proved.

Remark. When the dimension N is one, the argument of Lemma 7.1 gives the slightly stronger version of (6.3)

$$\limsup_{\sigma \to \infty} \max_{d(x, n\mathscr{S}) \geqq \sigma} u_n(x) \leqq \pi_0.$$

The statement (6.3) of Theorem 6.1 is still valid when $\mathscr{S}$ is unbounded, provided $\mathscr{S}'$ satisfies slightly stronger conditions. The proof can be found elsewhere [55].

The following example shows that the set $\mathscr{S}$ may, indeed, be empty.

Example 7.1. Let $Q[v]$ be $u(1, x)$, where u is the solution of the initial value problem (2.14) and

$$f(u) = u(1-u)(u-\tfrac{3}{4}).$$

Phase plane analysis (see, e.g., [5]) shows that $c^*(\xi) < 0$ for all ξ, so that $\mathscr{S}$ is empty. The results of [5] also show that if $0 \leqq u \leqq 1$ and $u \not\equiv 1$, then v approaches zero as $t \to \infty$. In particular, $u_n(x) = v(n, x)$ approaches zero as $n \to \infty$.

We also present an example to show that extra conditions such as (3.3) and (3.4), (3.5) or (3.6) are needed to keep initial data which lie above π_1 from propagating at a speed higher than $c^*(\xi)$.

Example 7.2. Let $N = 1$ and let $\mathscr{H}$ be the set of integers. Define

$$Q[u](x) = q(u(x)) + [4q(\tfrac{1}{4}) - 1] \sum_{k=1}^{\infty} (4q(\tfrac{1}{4}))^{-k} p(u(x-k)),$$

where

$$q(u) = \begin{cases} 2u - 3u^2 + 2u^3, & 0 \leqq u \leqq \tfrac{1}{2}, \\ \tfrac{1}{2}(u + \tfrac{1}{2}), & \tfrac{1}{2} \leqq u \leqq 1 \end{cases}$$

and

$$p(u) = \begin{cases} 0, & 0 \leqq u \leqq \tfrac{1}{2}, \\ (u - \tfrac{1}{2})^2, & \tfrac{1}{2} \leqq u \leqq 1. \end{cases}$$

Here $\pi_0 = 0$, $\pi_1 = \tfrac{1}{2}$, $\pi_+ = 1$.

Since $p = 0$ for $u < \tfrac{1}{2} = \pi_1$, we see that $c^*(\pm 1) = 0$. However, if

$$u_0(x) = \begin{cases} \tfrac{3}{4}, & \text{for } x = 0, \\ 0 & \text{otherwise}, \end{cases}$$

we have

$$u_1(k) = [4q(\tfrac{1}{4}) - 1] p(\tfrac{3}{4}) [4q(\tfrac{1}{4})]^{-k} \quad \text{for } k > 0. \tag{7.9}$$

Because $q'' < 0$ for $u < \frac{1}{2}$, we see that $q(u) \geqq 4q(\frac{1}{4})u$ for $u \leqq \frac{1}{4}$. Consequently, if $u_n(k) \leqq \frac{1}{4}$ for $n < k$, then $u_{n+1}(k) \geqq 4q(\frac{1}{4})u_n(k)$. Starting with the values (7.9) at $n = 1$, we find that

$$u_k(k) \geqq [4q(\tfrac{1}{4}) - 1]p(\tfrac{3}{4})[4q(\tfrac{1}{4})]^{-1}$$

for all positive k. Thus an impulse of magnitude at least $\frac{3}{4}$ concentrated at 0 is propagated with a speed at least equal to 1, so that Theorem 6.1 is not valid.

It is not known whether this phenomenon can also occur when $u_0 \leqq \pi_1$.

8. Convergence to the equilibrium value. (Proof of Theorem 6.2). According to the hypotheses of Theorem 6.2 $\mathcal{S}$ has at least one interior point y_0. Then

$$y_0 \cdot \xi < c^*(\xi)$$

for all unit vectors ξ.

If $u_n(x)$ satisfies the recursion $u_{n+1} = Q[u_n]$ and if we define the sequence

$$v_n(x) = u_n(x + ny_0),$$

then v_n satisfies the recursion

$$v_{n+1} = T_{-y_0}Q[v_n].$$

By Proposition 5.3 $T_{-y_0}Q$ has the wave speed $c^*(\xi) - y_0 \cdot \xi$ in the direction ξ. Thus, the introduction of moving coordinates replaces Q by $T_{-y_0}Q$, $c^*(\xi)$ by $c^*(\xi) - y_0 \cdot \xi$, and the set $\mathcal{S}$ by its translate by $-y_0$. We then also translate $\mathcal{S}''$ by $-y_0$. It is easily seen that Theorem 6.2 for v_n in terms of $T_{-y_0}Q$ implies Theorem 6.2 for u_n in terms of Q.

Consequently we shall assume without loss of generality that $x = 0$ is an interior point of $\mathcal{S}$, so that

$$c^*(\xi) > 0$$

for all unit vectors ξ.

We now note that Theorem 6.2 is strengthened when $\mathcal{S}''$ is enlarged. We add to $\mathcal{S}''$ a closed neighborhood of the origin which also lies in the interior of $\mathcal{S}$. Because $\mathcal{S}$ is convex, we may replace this union by its convex hull and still obtain a closed bounded set in the interior of $\mathcal{S}$. It is well known that a closed bounded convex set may be approximated by a larger closed convex set whose boundary is analytic, has positive Gaussian curvature, and lies within an arbitrarily small distance of the original set (see [6, § 27]). Consequently we shall assume without loss of generality that *$\mathcal{S}''$ is a closed bounded subset of the interior of $\mathcal{S}$ whose boundary is analytic and has positive Gaussian curvature, and $x = 0$ is an interior point of $\mathcal{S}''$.*

We define the distance function

$$D(x) = \inf\left\{d > 0 \,\middle|\, \frac{1}{d}x \in \mathcal{S}''\right\},$$

$D(0) = 0$, so that $\mathcal{S}'' = \{x | D(x) \leqq 1\}$, and the boundary of $\mathcal{S}''$ has the polar coordinate representation $|x| = 1/D(x/|x|)$. We also define the vector field

$$\tau(x) = \frac{1}{|\operatorname{grad} D(x)|} \operatorname{grad} D(x).$$

Because the boundary of $\mathcal{S}''$ is analytic and has positive Gaussian curvature, $D(x)$ and $\tau(x)$ are analytic for $x \neq 0$. Moreover, $D(x)$ is positive homogeneous of degree one and $\tau(x)$ is positive homogeneous of degree zero. In fact, $\tau(x)$ is the unit outward

normal vector at that boundary point of $\mathscr{S}''$ which lies on the ray from the origin through x.

Because the boundary of $\mathscr{S}''$ has positive curvature, the implicit function theorem shows that the equation $D(x)\tau(x)=\xi$ can be inverted, and that the inverse function $x=\sigma(\xi)$ is analytic for $\xi\neq 0$. Since $|\tau(x)|=1$, $D(x)=|\xi|$. Hence, if $|\xi|=1$, $\sigma(\xi)$ lies on the boundary of $\mathscr{S}''$. It is the unique boundary point at which the unit outward normal is ξ.

The support function of $\mathscr{S}''$ is defined for $|\xi|=1$ by

$$S(\xi)=\max_{x\in\mathscr{S}''} x\cdot\xi.$$

It is easily seen that this maximum is attained at $x=\sigma(\xi)$. Thus, $S(\xi)=\xi\cdot\sigma(\xi)$, so that $S(\xi)$ is analytic.

Because of the definition of $S(\xi)$ and the homogeneity of $D(x)$,

$$\frac{x\cdot\xi}{D(x)S(\xi)}\leqq 1$$

for $x\neq 0$. Hence

$$D(x)=\max_{|\xi|=1}\frac{x\cdot\xi}{S(\xi)},$$

with equality when $\xi=\tau(x)$.

We suppose that $\mathscr{S}''$ contains a ball of radius ρ and lies in a ball of radius R, both centered at the origin. Then

$$\rho\leqq S(\xi)\leqq R \quad \text{for } |\xi|=1,$$

$$\frac{|x|}{R}\leqq D(x)\leqq\frac{|x|}{\rho} \quad \text{for all } x.$$

Because $\mathscr{S}''$ lies in the interior of $\mathscr{S}$, there is a positive constant ε such that the dilatation $(1+\varepsilon)\mathscr{S}''$ still lies in the interior of $\mathscr{S}$. It follows that

$$(1+\varepsilon)S(\xi)<c^*(\xi) \quad \text{for } |\xi|=1.$$

We shall begin the proof of Theorem 6.2 by constructing a comparison function which will force u_n to be uniformly positive on the set $n(1+\varepsilon)\mathscr{S}''$. This comparison function will be constructed out of the functions $a_n(c,\xi;s)$ which were defined in § 5.

We first choose some $\alpha\in(\pi_0,\pi_1)$ and a smooth nonincreasing function $\varphi(s)$ with the properties (5.1) and the additional property

$$\varphi(s)=\alpha \quad \text{for } s\leqq -1.$$

The family of nonincreasing functions $a_n((1+\varepsilon)S(\xi),\xi;s)$ is defined by the recursion (5.4) with $c=(1+\varepsilon)S(\xi)$. Since this c is less than $c^*(\xi)$, $a_n((1+\varepsilon)S(\xi),\xi;s)$ increases to the constant π_1 as $n\to\infty$. Because $a_n((1+\varepsilon)S(\xi),\xi;0)$ is continuous in ξ, Dini's theorem [9, p. 106] shows that the convergence is uniform on the bounded set $|\xi|=1$. Therefore there is an integer n_0 such that

$$a_n((1+\varepsilon)S(\xi),\xi;0)>\alpha \quad \text{for } n\geqq n_0, \quad |\xi|=1. \tag{8.1}$$

We wish to construct an N-dimensional comparison function of bounded support by patching together functions of one variable. For this purpose we shall first approximate $a_n((1+\varepsilon)S(\xi),\xi;s)$ by a function which is constant for small s and zero for large s. We begin by constructing a family of operators Q_k on B in the following manner.

Let $\zeta(s)$ be a smooth nonincreasing function of one variable with the properties

$$\zeta(s) = \begin{cases} 1 & \text{for } s \leqq \frac{1}{2}, \\ 0 & \text{for } s \geqq 1. \end{cases} \tag{8.2}$$

For any u in B define

$$Q_k[u](y) = Q\left[u(x+y)\zeta\left(\frac{|x|}{k}\right)\right](0), \tag{8.3}$$

where y is fixed and the argument of Q is considered as a function of x. We state the properties of this family of operators as a lemma.

LEMMA 8.1. *The family Q_k takes B into itself and has the following properties.*

(i) *Q_k has the properties* (3.1.i), (3.1.ii), (3.1.iv) *and* (3.1.v).

(ii) *For each u in B, the sequence $Q_k[u]$ is nondecreasing in k and converges to $Q[u]$ as $k \to \infty$.*

(iii) *$Q_k[u](x_0)$ depends only on the values of u in the ball of radius k centered at x_0. That is, if $u(x) = v(x)$ for $|x - x_0| \leqq k$, then*

$$Q_k[u](x_0) = Q_k[v](x_0).$$

Proof. Property (iii) is obvious from the definition. Property (i) follows immediately from the definition and the fact that Q satisfies (3.1). Property (ii) follows from the fact that for each fixed y the sequence $u(x+y)\zeta(|x|/k)$ increases to $u(x+y)$ uniformly on each bounded set as $k \to \infty$ together with (3.1.iv) and (3.1.v).

We now define the sequence $a_n^{(k)}(c, \xi; s)$ by means of the recursion

$$\begin{aligned} a_{n+1}^{(k)}(c, \xi; s) &= \max\{\varphi(s), Q_k[a_n^{(k)}(x \cdot \xi + s + c)](0)\}, \\ a_0^k(c, \xi; s) &= \varphi(s). \end{aligned} \tag{8.4}$$

Proposition 4.2 and the proof of Lemma 5.1 show that $a_n^{(k)}((1+\varepsilon)S(\xi), \xi; s)$ is continuous in ξ and s, nonincreasing in s, and nondecreasing in n. By Lemma 8.1, $a_n^{(k)}((1+\varepsilon)S(\xi), \xi; s)$ increases to $a_n((1+\varepsilon)S(\xi), \xi; s)$ as $k \to \infty$. Since all these functions are continuous, Dini's theorem states that the convergence is uniform on any bounded set of (ξ, s). In particular, we see from (8.1) that there is an integer k_0 such that

$$a_{n_0}^{(k_0)}((1+\varepsilon)S(\xi), \xi; 0) > \alpha \quad \text{for } |\xi| = 1. \tag{8.5}$$

We define the sequence $\alpha_n^{(k)}$ of constants by the recursion

$$\begin{aligned} \alpha_{n+1}^{(k)} &= Q_k[\alpha_n^{(k)}], \\ \alpha_0^{(k)} &= \alpha = \varphi(s) \quad \text{for } s \leqq -1. \end{aligned} \tag{8.6}$$

Then $\alpha_n^{(k)}$ is nondecreasing in k and converges to α_n, which is defined by the recursion $\alpha_{n+1} = Q[\alpha_n]$, $\alpha_0 = \alpha$, as $k \to \infty$.

By replacing the operator Q by $\min\{Q[u], \pi_1 - \tau(\pi_1 - u)\}$ with τ positive and small if necessary, we assume without loss of generality that $Q[\beta] < \pi_1$ for $\beta < \pi_1$. Then α_n is strictly increasing. In particular, $\alpha_{n_0+1} > \alpha_{n_0} \geqq \alpha_{n_0}^{(k)}$ for all k. We choose k_0 so large that in addition to (8.5) the inequality

$$\alpha_{n_0+1}^{(k_0)} > \alpha_{n_0} \geqq \alpha_{n_0}^{(k_0)} \tag{8.7}$$

is satisfied.

The advantage of $a_n^{(k_0)}$ over a_n is that it is constant for large s and for small s.

LEMMA 8.2. *$\alpha_n^{(k_0)}$ is increasing in n and for each n*

$$a_n^{(k_0)}(c, \xi; s) = \begin{cases} \alpha_n^{(k_0)} & \text{for } s \leqq -1 - n[k_0 + c], \\ 0 & \text{for } s \geqq n[k_0 - c]. \end{cases} \tag{8.8}$$

Proof. We note that, if $Q_k[\alpha] \leqq \alpha$, the proof of Proposition 4.2 shows that $\alpha_n^{(k)} \leqq \alpha$ for all n. On the other hand, by (3.1iii) $Q[\alpha] > \alpha$, so that $\alpha_n > \alpha$ for $n \geqq 1$. In particular $\alpha_{n_0} > \alpha$. Therefore we see from (8.7) that $Q_{k_0}[\alpha] > \alpha$ and then from Proposition 4.2 that $\alpha_n^{(k_0)}$ is increasing in n.

It is easily seen from the property (iii) of Lemma 8.1 and the fact that $\alpha_n^{(k_0)} \geqq \alpha \geqq \varphi$ that if (8.8) holds for n, it also holds for $n+1$. Since $a_0^{(k_0)}(c, \xi; s) = \varphi(s)$, about which we have assumed that (8.8) with $n = 0$ is valid, the lemma is proved.

Since $\alpha_n^{(k_0)} \leqq \alpha_n \leqq \pi_1$, and $\alpha_n^{(k_0)}$ is nondecreasing in n, the limit

$$\alpha^{(k_0)} = \lim_{n \to \infty} \alpha_n^{(k_0)} \tag{8.10}$$

exists. The proof of Lemma 5.3 shows that because of (8.5)

$$\lim_{n \to \infty} a_n^{(k_0)}((1+\varepsilon)S(\xi), \xi; s) \equiv \alpha^{(k_0)}. \tag{8.11}$$

The inequality (8.5) also implies that

$$a_n^{(k_0)}((1+\varepsilon)S(\xi), \xi; s) \geqq \varphi(s) \quad \text{for } n \geqq n_0.$$

Therefore the recursion (8.4) becomes

$$\begin{aligned} &a_{n+1}^{(k_0)}((1+\varepsilon)S(\xi), \xi; s) \\ &\quad = Q_{k_0}[a_n^{(k_0)}(x \cdot \xi + s + (1+\varepsilon)S(\xi))](0) \quad \text{for } n \geqq n_0, |\xi| = 1. \end{aligned} \tag{8.12}$$

We shall need the following lemma, in which we replace the variable s by $t = s/S(\xi)$.

LEMMA 8.3. *There is an integer $n_1 > n_0$ such that*

$$a_{n_1}^{(k_0)}((1+\varepsilon)S(\xi), \xi; S(\xi)t) \geqq a_{n_0}^{(k_0)}((1+\varepsilon)S(\xi'), \xi'; S(\xi')t) \quad \text{for } t \in \mathcal{R}, |\xi| = |\xi'| = 1. \tag{8.13}$$

Proof. We see from (8.7), (8.11), and Dini's theorem that there is an integer $n_1 > n_0$ such that

$$a_{n_1}^{(k_0)}\left((1+\varepsilon)S(\xi), \xi; n_0 k_0 \frac{R}{\rho}\right) > \alpha_{n_0}^{(k_0)} \quad \text{for } |\xi| = 1.$$

Because $a_{n_1}^{(k_0)}$ is nonincreasing in s, it follows that the same inequality holds for $s \leqq n_0 k_0 R/\rho$. Since $\rho \leqq S(\xi) \leqq R$ and $a_{n_0}^{(k_0)} \leqq \alpha_{n_0}^{(k_0)}$, the inequality (8.13) is then valid when $t \leqq n_0 k_0/\rho$. When $t \geqq n_0 k_0/\rho$, then $S(\xi')t \geqq n_0 k_0$ and because $c = (1+\varepsilon)S(\xi') > 0$ we see from (8.8) that the right-hand side of (8.13) is zero. Thus (8.13) is true for all t.

We recall that the ratio $x \cdot \xi/S(\xi)$ is homogeneous of degree zero in ξ and takes on its maximum value $D(x)$ when ξ is equal to the unit vector $\tau(x)$. Moreover, $S(\xi)$ is positive and analytic for $\xi \neq 0$. Then Taylor's theorem shows that $x \cdot \xi/S(\xi) = D(x) + O(|\xi - \tau(x)|^2)$ for ξ near $\tau(x)$. By looking at the bounded set $|\xi| = 1$, $|x| \leqq 1$ and using the homogeneity in x, we find that there is a constant m such that

$$\frac{x \cdot \xi}{S(\xi)} \geqq D(x) - m|x|\,|\xi - \tau(x)|^2 \quad \text{for } |\xi| = 1. \tag{8.14}$$

The function $\tau(x)$ is smooth for $x \neq 0$ and homogeneous of degree zero. Therefore if x and y are any two nonzero points in $\mathcal{R}^N$

$$|\tau(x)-\tau(y)| = \left|\tau\left(\frac{x}{|x|}\right)-\tau\left(\frac{y}{|y|}\right)\right| \leqq M\left|\frac{x}{|x|}-\frac{y}{|y|}\right|.$$

for some constant M. But

$$\left|\frac{x}{|x|}-\frac{y}{|y|}\right|^2 = 2\left(1-\frac{x\cdot y}{|x|\,|y|}\right) = \frac{1}{|x|\,|y|}[|x-y|^2-(|x|-|y|)^2] \leqq \frac{|x-y|^2}{|x|\,|y|},$$

so that

$$|\tau(x)-\tau(y)| \leqq M|x|^{-1/2}|y|^{-1/2}|x-y|. \tag{8.15}$$

We now choose a constant A which satisfies

$$A \geqq \rho^{-1}+n_1[k_0\rho^{-1}+1+\varepsilon]+(n_1-n_0)k_0\rho^{-1}+2mM^2\rho^{-1}\varepsilon^{-1}(n_1-n_0)^2k_0^2, \tag{8.16}$$

and define the comparison sequence

$$e_n(x) = a_{n_1}^{(k_0)}((1+\varepsilon)S(\tau(x)),\, \tau(x);\, S(\tau(x))[D(x)-A-(1+\tfrac{1}{2}\varepsilon)n]), \qquad n = 0, 1, 2, \cdots. \tag{8.17}$$

We recall that $\tau(x)$ is the unit outward normal to $\mathcal{S}''$ at the boundary point on the ray from the origin through x. Because $\rho \leqq S \leqq R$ and $c = (1+\varepsilon)S$, (8.8) shows that

$$e_n(x) = \begin{cases} \alpha_{n_1}^{(k_0)} & \text{for } D(x) \leqq A-\rho^{-1}-n_1[k_0\rho^{-1}+1+\varepsilon]+n(1+\frac{1}{2}\varepsilon), \\ 0 & \text{for } D(x) \geqq A+n_1[k_0\rho^{-1}-1-\varepsilon]+n(1+\frac{1}{2}\varepsilon). \end{cases} \tag{8.18}$$

Since $R^{-1}|x| \leqq D(x) \leqq \rho^{-1}|x|$, we see from (8.16) that $e_n = \alpha_{n_1}^{(k_0)}$ near $x = 0$, so that e_n is continuous, and e_n vanishes outside a bounded set. The most important property of e_n is given by the following lemma. As usual, we define $Q_{k_0}^r$ to be the rth iterate of the operator Q_{k_0}.

LEMMA 8.4. *If A satisfies the inequality* (8.16), *then the sequence $e_n(x)$ satisfies the inequality*

$$e_{n+n_1-n_0} \leqq Q_{k_0}^{n_1-n_0}[e_n] \quad \text{for } n = 0, 1, 2, \cdots. \tag{8.19}$$

Proof. Since e_n is obtained from e_0 by replacing A by $A+(1+\frac{1}{2}\varepsilon)n$, which is at least as large as A, it is sufficient to prove (8.19) for $n = 0$.

Because $D(x) \leqq |x|/\rho$, we see from (8.18) that if

$$|x_0| \leqq \rho\{A-\rho^{-1}-n_1[k_0\rho^{-1}+1+\varepsilon]\}-(n_1-n_0)k_0,$$

then $e_0(x) = \alpha_{n_1}^{(k_0)}$ for $|x-x_0| \leqq (n_1-n_0)k_0$. Because of the definition (8.3) of Q_k we find that for such x_0

$$Q_{k_0}^{n_1-n_0}[e_0](x_0) = \alpha_{2n_1-n_0}^{(k_0)} \geqq \alpha_{n_1}^{(k_0)} \geqq e_{n_1-n_0}(x_0),$$

so that (8.19) with $n = 0$ is true.

We now consider a point x_0 where

$$|x_0| > \rho\{A-\rho^{-1}-n_1[k_0\rho^{-1}+1+\varepsilon]\}-(n_1-n_0)k_0, \tag{8.20}$$

and let x be any point such that

$$|x-x_0| \leqq (n_1-n_0)k_0.$$

We then see from (8.15) and (8.16) that

$$m|x|\,|\tau(x)-\tau(x_0)|^2 \leqq \tfrac{1}{2}\varepsilon.$$

The latter, together with (8.14), shows that

$$D(x) \leqq \frac{x\cdot\tau(x_0)}{S(\tau(x_0))}+\frac{1}{2}\varepsilon \quad \text{for } |x-x_0| \leqq (n_1-n_0)k_0.$$

Since $a_{n_1}^{(k_0)}$ is nonincreasing in s, it follows that when x_0 satisfies (8.20),

$$e_0(x) \geqq a_{n_1}^{(k_0)}\left((1+\varepsilon)S(\tau(x),\tau(x);S(\tau(x)))\left[\frac{x\cdot\tau(x_0)}{S(\tau(x_0)}-A+\frac{1}{2}\varepsilon\right]\right)$$
$$\text{for } |x-x_0| \leqq (n_1-n_0)k_0.$$

We now use the inequality (8.13) to see that

$$e_0(x) \geqq a_{n_0}^{(k_0)}((1+\varepsilon)S(\tau(x_0)),\tau(x_0);x\cdot\tau(x_0)-(A-\tfrac{1}{2}\varepsilon)S(\tau(x_0))$$
$$\text{for } |x-x_0| \leqq (n_1-n_0)k_0.$$

Since $x_0\cdot\tau(x_0)=S(\tau(x_0))D(x_0)$, it follows from the order-preserving property of Q_{k_0} and the recursion (8.12) that

$$\begin{aligned} Q_{k_0}^{n_1-n_0}[e_0](x_0) &\geqq a_{n_1}^{(k_0)}((1+\varepsilon)S(\tau(x_0)),\tau(x_0);x_0\cdot\tau(x_0) \\ &\quad -(A-\tfrac{1}{2}\varepsilon)S(\tau(x_0))-(n_1-n_0)(1+\varepsilon)S(\tau(x_0)) \\ &\geqq a_{n_1}^{(k_0)}((1+\varepsilon)S(\tau(x_0)),\tau(x_0);S(\tau(x_0)[D(x_0)-A-(n_1-n_0)(1+\tfrac{1}{2}\varepsilon)] \\ &= e_{n_1-n_0}(x_0). \end{aligned}$$

We have established the inequality (8.19) for all points x_0, and, if we replace A by $A+(1+\frac{1}{2}\varepsilon)n$, for all n, so that the lemma is proved.

Next we show that if u_0 is uniformly positive on a sufficiently large ball, then the sequence e_n can be used as a comparison sequence.

LEMMA 8.5. *For any $\sigma\in(\pi_0,\pi_1)$ there are a radius r_σ and an integer l_σ such that, if $u_0(x)\geqq\sigma$ on the ball $\{x\,|\,|x|\leqq r_\sigma\}$ and if $u_{n+1}=Q[u_n]$, then*

$$u_l(x) \geqq e_0(x) \quad \text{for } l_\sigma \leqq l < l_\sigma + n_1 - n_0.$$

Proof. Define the sequence of constants μ_n by the recursion

$$\mu_{n+1}=Q[\mu_n], \qquad \mu_0=\sigma.$$

Then μ_n increases to π_1. Choose an integer l_σ such that

$$\mu_{l_\sigma} > \alpha_{n_1}^{(k_0)} \geqq e_0(x). \tag{8.21}$$

We use a nonnegative smooth function $\zeta(s)$ with the properties (8.2) to define the one parameter family of functions $v_n^{(r)}$ by the recursion

$$v_{n+1}^{(r)}=Q[v_n^{(r)}], \qquad v_0^{(r)}(x)=\sigma\zeta\left(\frac{|x|}{r}\right).$$

By (3.1.v), $v_l^{(r)}(x)$ increases to μ_l as $r\to\infty$, and by Dini's theorem the convergence is uniform on bounded sets. Because of the inequality (8.21) there is a value r_σ of r such that

$$v_l^{(r_\sigma)} \geqq e_0 \quad \text{for } l_\sigma \leqq l < l_\sigma + n_1 - n_0$$

on the bounded set where $e_0 > 0$. Since $v_l^{(r_\sigma)}$ is nonnegative, these inequalities then hold for all x.

Since $v_0^{(r_\sigma)} = 0$ for $|x| \geqq r_\sigma$ and $v_0^{(r_\sigma)} \leqq \sigma$, Proposition 4.1 shows that if $u_0 \geqq \sigma$ for $|x| \leqq r_\sigma$, then $u_l \geqq v_l^{(r_\sigma)} \geqq e_0$ for $l_\sigma \leqq l \leqq l_\sigma + n_1 - n_0$, which proves the lemma.

We now show that if $u_0 \geqq \sigma$ on a ball of radius r_σ, then for all sufficiently large n the function u_n is uniformly positive on a large set.

LEMMA 8.6. *Suppose that, for some $\sigma \in (0, \pi_0)$ and some $\bar{x}$ in $\mathscr{R}^N$, $u_o(x) \geqq \sigma$ for $|x - \bar{x}| \leqq r_\sigma$. Then for all sufficiently large n*

$$u_n(x) \geqq \alpha_{n_1}^{(k_0)} \quad \text{when } D(x) \leqq (1 + \tfrac{1}{4}\varepsilon)n. \tag{8.22}$$

Proof. Since $Q[u] \geqq Q_{k_0}[u]$ for all u, we see from Lemma 8.5, (8.19), and Proposition 4.1 applied to the operator $Q_{k_0}^{n_1 - n_0}$ that for $0 \leqq q < n_1 - n_0$ and $j \geqq 0$

$$u_{l_\sigma + q + j(n_1 - n_0)}(x) \geqq e_{j(n_1 - n_0)}(x - \bar{x}).$$

By (8.18)

$$u_n(x) \geqq \alpha_{n_1}^{(k_0)} \tag{8.23}$$

$$\text{for } D(x - \bar{x}) \leqq A - \rho^{-1} - n_1[k_0\rho^{-1} + 1 + \varepsilon] + (n - l_\sigma - n_1)(1 + \tfrac{1}{2}\varepsilon).$$

We note that

$$D(x - \bar{x}) = \max_{|\xi|=1} \frac{(x - \bar{x}) \cdot \xi}{S(\xi)} \leqq \max_{|\xi|=1} \frac{x \cdot \xi}{S(\xi)} + \max \frac{-\bar{x} \cdot \xi}{S(\xi)} = D(x) + D(-\bar{x}). \tag{8.24}$$

That is, the function D is subadditive. Hence if $D(x) \leqq (1 + \frac{1}{4}\varepsilon)n$ and n is sufficiently large, the condition on $D(x - \bar{x})$ in (8.23) is satisfied, which proves the lemma.

We have shown that, if $u_0 > \sigma$ on a ball of radius r_σ, then when n is large u_n is uniformly greater than π_0 on the set $(1 + \frac{1}{4}\varepsilon)n\mathscr{S}''$. We shall now show that this fact implies that u_n is arbitrarily near π_1 on the set $n\mathscr{S}''$.

LEMMA 8.7. *Suppose that $u_0(x) \geqq \sigma \in (\pi_0, \pi_1)$ on a ball $|x - \bar{x}| \in r_\sigma$. Then for any positive δ there is an n_δ such that if $m \geqq n_\delta$ and $D(x) \leqq m$, then $u_m(x) \geqq \pi_1 - \delta$.*

Proof. Because the sequence α_n defined by the recursion $\alpha_{n+1} = Q[\alpha_n]$, $\alpha_0 = \alpha$ converges to π_1, there is an integer n_2 such that $\alpha_{n_2} > \pi_1 - \delta$.

Let the sequence $w_n^{(r)}$ be defined by the recursion

$$w_{n+1}^{(r)} = Q[w_n^{(r)}], \qquad w_0^{(r)}(x) = \alpha\zeta(|x|/r)$$

where ζ is a smooth nonnegative function with the properties (8.2). Then $w_{n_2}^{(r)}(0)$ converges to α_{n_2} as $r \to \infty$. Hence there is an $\bar{r}$ such that

$$w_{n_2}^{(\bar{r})}(0) > \pi_1 - \delta.$$

Proposition 4.1 and the translation invariance of Q now show that if $u_n(x) \geqq \alpha$ on the ball $|x - x_1| \leqq \bar{r}$, then

$$u_{n+n_2}(x_1) > \pi_1 - \delta. \tag{8.25}$$

Because of the subadditivity (8.24), and the bound $D(x) \leqq |x|/\rho$, the inequalities $D(x_1) \leqq n + n_2$ and $|x - x_1| \leqq \bar{r}$ imply that

$$D(x - \bar{x}) \leqq D(x_1) + D(x - x_1) + D(-\bar{x})$$

$$\leqq n + n_2 + \rho^{-1}(\bar{r} + |\bar{x}|).$$

When n is sufficiently large, say $n \geqq n_\delta - n_2$, this inequality implies the condition in (8.23). Therefore $u_n(x) \geqq \alpha_{n_1}^{(k_0)} \geqq \alpha$ for $|x - x_1| \leqq \bar{r}$, and the inequality (8.25) follows.

Thus when $n+n_2 \geqq n_\delta$ and $D(x_1) \leqq n+n_2$, we have $u_{n+n_2}(x_1) \geqq \pi_1 - \delta$. This is the statement of the lemma with $m = n+n_2$.

Equation (6.4) is an immediate consequence of this lemma. Equation (6.5) is an immediate consequence of (6.4), (3.2), and Proposition 4.1. Thus, Theorem 6.2 is proved.

The following example of Fujita [18] shows that the radius r_σ in Theorem 6.2 must, in general, depend upon σ.

Example 8.1. Let $Q[u](x) = v(x, 1)$, where $v(x, t)$ is the solution of the N-dimensional Fisher equation

$$\frac{\partial v}{\partial t} = \Delta v + f(v)$$

with the initial values $v(x, 0) = u(x)$. If $f(0) = f(1) = 0$ and $0 < f(v) \leqq v^4$, the hypotheses (3.1), (3.3) and (3.5) are valid with $\pi_0 = 0$, $\pi_1 = \pi_+ = 1$. The function

$$w(x, t) = [6(t+1)]^{-1/3} e^{-|x|^2/4(t+1)}$$

satisfies the differential inequality $w_t \geqq \Delta w + f(w)$. It follows from the maximum principle for parabolic equations [43, Thm. 3.10] that if $u(x_0) = w(x, 0)$, then $u_n(x) \leqq w(x, n)$. Thus u_n does not approach π_1 and, in fact, goes to zero. The explanation is that for each σ the set where $u_0 \geqq \sigma$ does not contain a sufficiently large ball. We conclude that $r_\sigma > 2|\log 6^{1/3}\sigma|^{1/2}$, which approaches infinity as σ goes to 0.

The following variant of Theorem 6.2 gives a condition under which r_σ does not need to vary with σ. It will be used in the proof of Theorem 6.4.

LEMMA 8.8. *Suppose that Q satisfies the conditions of Theorem* 6.2 *and that it has the additional property that*

$$Q[\rho u] \geqq \rho Q[u] \tag{8.26}$$

for every constant ρ with $0 \leqq \rho \leqq 1$. Then there is a fixed radius r such that Theorem 6.2 *is valid when r_σ is replaced by r for any positive σ.*

Proof. Theorem 6.2 was proved by showing that when $u_0 \geqq \sigma$ on a ball of radius σ, then a translate of some u_0 lies above the first member e_0 of a sequence with the properties (8.18) and (8.19).

The additional property (8.26) shows that once one has constructed such a sequence, the sequence ρe_n with $\rho \in (0, 1)$ again satisfies (8.19) and (8.18) with the positive constant $\alpha_{n_1}^{(k_0)}$ replaced by the positive constant $\rho\alpha_{n_1}^{(k_0)}$.

Let $e_0 = 0$ outside the ball $|x| \leqq r$. Then if u_0 is uniformly positive in a ball $|x - \bar{x}| \leqq r$ of radius r, there is a positive ρ such that

$$u_0(x) \geqq \rho e_0(x - \bar{x}).$$

Proposition 5.1 then shows that

$$u_{m(n_1-n_0)}(x) \geqq \rho e_{m(n_1-n_2)}(x - \bar{x}) \quad \text{for } m \geqq 0.$$

Because of (8.19) it follows that there is an m_0 such that for $u_{m_0}(n_1 - n_0) \geqq \sigma$ on a ball of radius r_σ, where $\sigma = \rho\alpha_{n_1}^{(k_0)}$.

The proof of Lemma 8.6 still gives (8.22) for sufficiently large n, and the proof of Theorem 6.2 is finished as above.

9. Bounds for the wave speed. (Proof of Theorems 6.3 and 6.4). Theorems 6.3 and 6.4 will follow from Proposition 5.5 when we determine the wave speed of the

comparison operator

$$M[u](x) \equiv \min\left\{\beta, \int_{\mathscr{H}} u(x-y)m(y, dy)\right\}. \tag{9.1}$$

Here m is a nonnegative measure on $\mathscr{H}$ with the properties $\int_{\mathscr{H}} m(x, dx) < \infty$ and

$$\int_{\mathscr{H}} m(x, dx) > 1, \tag{9.2}$$

and β is any positive constant or $+\infty$. (In the latter case, M is linear.) M is defined on the set of continuous nonnegative functions. It has the properties (3.1) with $\pi_0 = 0$ and $\pi_1 = \beta$, so that its wave speed $\bar{c}^*(\xi)$ is defined. We shall obtain a formula for this wave speed.

LEMMA 9.1. *For each unit vector ξ the wave speed $\bar{c}^*(\xi)$ of M is given by the formula*

$$\bar{c}^*(\xi) = \inf_{\mu>0}\left\{\frac{1}{\mu}\log\int e^{\mu x\cdot\xi} m(x, dx)\right\}, \tag{9.3}$$

which is independent of β.

Moreover the set

$$\bar{\mathscr{S}} = \{x \mid x\cdot\xi \leqq \bar{c}^*(\xi) \forall \xi \in S^{N-1}\}$$

is not empty, and its interior is empty if and only if the measure m is concentrated on a hyperplane $x\cdot\xi =$ constant.

Proof. Consider a fixed unit vector ξ. For every $\mu > 0$. define

$$\Phi(\mu) = \frac{1}{\mu}\log\int e^{\mu x\cdot\xi} m(x, dx) \tag{9.4}$$

with $\Phi(\mu) = +\infty$ if the integral on the right diverges.

For any μ such that $\Phi(\mu)$ is finite, consider the function of one variable

$$w(s) = \min\{\beta, e^{-\mu s}\}.$$

Then $w(s) \leqq e^{-\mu s}$, and hence

$$\begin{aligned} M[w(x\cdot\xi + s + \Phi(\mu))](0) &\leqq \int e^{-\mu(-y\cdot\xi + s + \Phi(\mu))} m(y, dy) \\ &= e^{-\mu s} \end{aligned}$$

because of the definition (9.4) of $\Phi(\mu)$. The definition (9.1) of M also shows that $M[w(x\cdot\xi + s + \Phi(\mu))] \leqq \beta$. Consequently

$$M[w(x\cdot\xi + s + \Phi(\mu))](0) \leqq w(s).$$

If $\varphi(s)$ is any function with the properties (5.1) such that $\varphi(s) \leqq w(s)$, then also

$$\max\{\varphi(s), M[w(x\cdot\xi + s + \Phi(\mu))](0)\} \leqq w(s).$$

Therefore if the sequence $a_n(\Phi(\mu), \xi; s)$ is defined by the recursion (5.4), Proposition 4.1 shows that $a_n \leqq w$ for all n. Hence $a(\Phi(\mu), \xi; s) \leqq w(s)$. It follows that $a(\Phi(\mu), \xi; +\infty) = 0$. We conclude that

$$\bar{c}^*(\xi) \leqq \Phi(\mu) \tag{9.5}$$

for all positive μ, so that inf $\Phi(\mu)$ provides an upper bound for the wave speed $\bar{c}^*$.

We now wish to show that the infimum of $\Phi(\mu)$ is equal to $\bar{c}^*$. We shall assume for the moment that the measure m vanishes outside a ball $|x| \leqq b$. Then $\Phi(\mu)$ is

analytic in μ. We define the function

$$\psi(\mu) = (\mu\Phi)' = \frac{\int x \cdot \xi\, e^{\mu x \cdot \xi} m(x, dx)}{\int e^{\mu x \cdot \xi} m(x, dx)}. \tag{9.6}$$

Differentiation shows that

$$\psi'(\mu) = e^{-\mu\Phi(\mu)} \int [x \cdot \xi - \psi(\mu)]^2 e^{\mu x \cdot \xi} m(x, dx) \geqq 0, \tag{9.7}$$

so that ψ is nondecreasing. Moreover by (9.6)

$$\Phi'(\mu) = \frac{1}{\mu}[\psi - \Phi], \tag{9.8}$$

so that

$$(\mu^2\Phi')' = \mu\psi' \geqq 0. \tag{9.9}$$

We see from the definition (9.4) and the condition (9.2) that $\Phi(\mu) \to +\infty$ as μ decreases to zero, while $\psi(\mu)$ remains bounded as $\mu \to 0$ because the support of m is bounded. Then (9.8) shows that Φ is decreasing near $\mu = 0$, and (9.9) shows that Φ' changes its sign at most once.

It is easily seen that if

$$K = \sup \{x \cdot \xi | x \in \operatorname{supp} m\},$$

where supp m is the support of m, then

$$\lim_{\mu \to +\infty} \Phi(\mu) = \lim_{\mu \to +\infty} \psi(\mu) = K. \tag{9.10}$$

Moreover,

$$\Phi(\mu) = K + \frac{1}{\mu} \log \left\{ \int_{x \cdot \xi < K} e^{\mu(x \cdot \xi - K)} m(x, dx) + \int_{x \cdot \xi = K} m(x, dx) \right\},$$

and the first integral approaches zero as $\mu \to \infty$. Consequently, there are two possibilities:

(a) If

$$\int_{x \cdot \xi = K} m(x, dx) \geqq 1,$$

then $\Phi(\mu) \geqq K \geqq \psi(\mu)$ for all μ. Φ is nonincreasing by (9.8), and

$$\inf_{\mu > 0} \Phi(\mu) = \Phi(\infty) = K = \psi(\infty).$$

(b) If

$$\int_{x \cdot \xi = K} m(x, dx) < 1,$$

then $\Phi(\mu) < K$ for all sufficiently large μ. Hence there is a $\mu^* > 0$ such that $\Phi' < 0$ for $\mu < \mu^*$ and $\Phi' \geqq 0$ for $\mu > \mu^*$. Then

$$\inf_{\mu > 0} \Phi(\mu) = \Phi(\mu^*) = \psi(\mu^*), \tag{9.11}$$

and by (9.8)

$$\Phi(\mu) > \psi(\mu) \quad \text{for } \mu < \mu^*. \tag{9.12}$$

For the case (a) we define $\mu^* = +\infty$, so that (9.11) and (9.12) are also valid for case (a).

We now choose any $\mu \in (0, \mu^*)$, and introduce a positive parameter γ which is so small that the integral

$$\int e^{\mu y \cdot \xi} \cos \gamma y \cdot \xi m(y, dy)$$

is positive. We define the constant

$$z = \frac{1}{\gamma} \tan^{-1} \frac{\int e^{\mu y \cdot \xi} \sin \gamma y \cdot \xi m(y, dy)}{\int e^{\mu y \cdot \xi} \cos \gamma y \cdot \xi m(y, dy)}. \tag{9.13}$$

Clearly,

$$\lim_{\gamma \to 0} z = \psi(\mu).$$

We suppose that γ is so small that

$$\gamma(b + |z|) < \pi, \tag{9.14}$$

where $m(x, dx) = 0$ for $|x| \geqq b$.

For any ε such that $0 < \varepsilon \leqq \beta$ we define the function

$$v(s) = \begin{cases} \varepsilon e^{-\mu s} \sin \gamma s & \text{for } 0 \leqq s \leqq \dfrac{\pi}{\gamma}, \\ 0 & \text{elsewhere.} \end{cases}$$

Then

$$M[v(x \cdot \xi + s + z)](0) = \min \left\{ \beta, \int v(-y \cdot \xi + s + z)\, m(y, dy) \right\}.$$

When $0 \leqq s \leqq \pi/\gamma$ and $|y| \leqq b$, (9.14) implies that $-\pi/\gamma \leqq -y \cdot \xi + s + z \leqq 2\pi/\gamma$ so that

$$v(-y \cdot \xi + s + z) \geqq \varepsilon e^{-\mu(-y \cdot \xi + s + z)} \sin \gamma(-y \cdot \xi + s + z).$$

We then see from the definition (9.13) of z that for $0 < s < \pi/\gamma$

$$M[v(x \cdot \xi + s + z)] \geqq \min \left\{ \beta, \varepsilon \int e^{-\mu(-y \cdot \xi + s + z)} \sin \gamma(-y \cdot \xi + s + z)\, m(y, dy) \right\}$$

$$= \min \left\{ \beta, e^{-\mu z} \sec \gamma z \int e^{\mu y \cdot \xi} \cos \gamma y \cdot \xi m(y, dy) v(s) \right\}.$$

As γ goes to zero, the coefficient of $v(s)$ approaches $e^{\mu[\Phi(\mu) - \psi(\mu)]}$, which is greater than 1 because $\mu < \mu^*$. Since $v \leqq \varepsilon \leqq \beta$, we find that when γ is sufficiently small,

$$M[v(x \cdot \xi + s + z)] \geqq v(s) \tag{9.15}$$

for $0 \leqq s \leqq \pi/\gamma$. Since v vanishes outside this interval, this inequality is valid for all s.

The function v attains its maximum at a unique point $\bar{s} = \gamma^{-1} \tan^{-1}(\gamma/\mu)$. We define the function

$$\varphi(s) = \begin{cases} v(\bar{s}) & \text{for } s \leqq -\dfrac{\pi}{\gamma} + \bar{s}, \\ v\left(s + \dfrac{\pi}{\gamma}\right) & \text{for } s \geqq -\dfrac{\pi}{\gamma} + \bar{s}, \end{cases}$$

which clearly has the properties (5.1).
Moreover,

$$\varphi(s) = \max\left\{ v(s-t) \middle| t \leqq -\frac{\pi}{\gamma} \right\}.$$

Therefore (9.15) shows that

$$M[\varphi(x \cdot \xi + s + z)](0) \geqq \max\left\{ v(s-t) \middle| t \leqq -\frac{\pi}{\gamma} \right\} = \varphi(s).$$

For any $c < z$ we define the sequence $a_n(c, \xi; s)$ by the recursion (5.4) with Q replaced by M. Then

$$\begin{aligned} a_1(c, \xi; s) &= \max\{\varphi(s), M[\varphi(x \cdot \xi + s + c)]\} \geqq \max\{\varphi(s), \varphi(s + c - z)\} \\ &= \varphi(s + c - z) = a_0(c, \xi; s + c - z). \end{aligned}$$

Proposition 4.1 now shows that $a_{n+1}(c, \xi; s) \geqq a_n(c, \xi; s + c - z)$, so that $a(c, \xi; s) \geqq a(c, \xi; s + c - z)$. Since a is nonincreasing and $c - z < 0$, this implies that a is a constant. Therefore the inequality $c < z$ implies that $c < \bar{c}^*(\xi)$.

We conclude that when γ is sufficiently small, $z \leqq \bar{c}^*(\xi)$. Taking the limit of z as γ decreases to 0, we find that when $\mu < \mu^*$

$$\bar{c}^*(\xi) \geqq \psi(\mu).$$

We recall that as μ increases to μ^*, $\psi(\mu)$ increases to the infimum of $\Phi(\mu)$.

Thus we have shown that when the measure m has bounded support,

$$\bar{c}^*(\xi) \geqq \inf_{\mu > 0} \Phi(\mu). \tag{9.16}$$

If m does not have bounded support we choose a nonincreasing function ζ with the properties (8.2) and approximate m by the nonnegative measure

$$m_k(x, dx) = \zeta\left(\frac{|x|}{k}\right) m(x, dx),$$

which has bounded support. Clearly the sequence of measures m_k increases to m and for all sufficiently large k the measure m_k has the property (9.2). We define the operator M_k and the function $\Phi_k(\mu)$ by replacing m by m_k in the definitions (9.1) and (9.4) and let $\bar{c}_k^*(\xi)$ be the wave speed of M_k. Obviously, $M \geqq M_k$. Therefore by Proposition 5.5 $\bar{c}^*(\xi) \geqq \bar{c}_k^*(\xi)$, so that by (9.16)

$$\bar{c}^*(\xi) \geqq \inf_{\mu > 0} \Phi_k(\mu) \tag{9.17}$$

for all k. The right-hand side is nondecreasing in k, and its limit again provides a lower bound for $\bar{c}^*(\xi)$.

Suppose that ρ is any number which is above this limit. That is, suppose that

$$\rho > \inf_{\mu>0} \Phi_k(\mu)$$

for all k. Then the sets

$$S_k = \{\mu > 0 | \Phi_k(\mu) \leqq \rho\}$$

are nonempty, and S_k contains S_{k+1}. When k is so large that m_k satisfies (9.2), $\Phi_k(0) = +\infty$, so that S_k is closed.

If $x \cdot \xi$ is not bounded above on the support of m, we see from (9.10) that for sufficiently large k the set S_k is bounded above. We then see from Helly's theorem (the finite intersection property) that the intersection S_∞ of all the S_k is not empty. Since Φ_k converges to Φ, we see that $\Phi(\mu) \leqq \rho$ for $\mu \in S_\infty$; so that $\inf \Phi \leqq \rho$.

We conclude that if $x \cdot \xi$ is not bounded above on the support of m,

$$\inf_{\mu>0} \Phi(\mu) = \lim_{k\to\infty} \inf_{\mu>0} \Phi_k(\mu). \tag{9.18}$$

If, on the other hand, $x \cdot \xi$ is bounded above on the support of m, then for all sufficiently large k, $\sup\{x \cdot \xi | x \in \operatorname{supp} m_k\} = K$ which is independent of k. There are then two possibilities:

If

$$\int_{x\cdot\xi=K} m(x, dx) < 1,$$

then by the argument we have used earlier $\inf \Phi_k \leqq \inf \Phi < \Phi(\infty) = K = \Phi_k(\infty)$ for large k. Therefore we may carry out the above argument when

$$\inf \Phi_k < \rho < K \quad \text{for all } k \tag{9.19}$$

to reach the conclusion (9.18).

When

$$\int_{x\cdot\xi=K} m(x, dx) \geqq 1,$$

we easily see that $\Phi(\mu) \geqq K$ and $\Phi(\infty) = K$. If there were a ρ which satisfies (9.19), the same argument would lead to the conclusion that $\inf \Phi(\mu) \leqq \rho$, which is incompatible with (9.19) since $\inf \Phi = K$. Therefore (9.19) leads to a contradiction, and we conclude that

$$\lim_{k\to\infty} \inf_{\mu>0} \Phi_k(\mu) = K = \inf_{\mu>0} \Phi.$$

We have thus proved (9.18) in all cases.

The formula (9.3) now follows from (9.5), (9.17) and (9.18). Note that our proof is valid even when $\bar{c}^*(\xi) = +\infty$.

To prove that $\bar{\mathcal{S}}$ is not empty we again begin by considering the case where m has bounded support. We define the "drift velocity"

$$V = \frac{\int x m(x, dx)}{\int m(x, dx)}. \tag{9.20}$$

We note that for any unit vector ξ, $\psi(0) = V \cdot \xi$.

We recall that there is a $\mu^* \in (0, \infty]$ such that $\bar{c}^*(\xi) = \psi(\mu^*)$. Since (9.7) shows that ψ is nondecreasing, we conclude that

$$V \cdot \xi \leqq \bar{c}^*(\xi) \tag{9.21}$$

for all ξ. This shows that V lies in $\bar{\mathscr{S}}$ so that $\bar{\mathscr{S}}$ is not empty.

If the support of m is not concentrated on a plane normal to ξ, then (9.7) shows that ψ is strictly increasing so that

$$V \cdot \xi < c^*(\xi). \tag{9.22}$$

If the support of m is not contained in any plane, this inequality is valid for all ξ so that V is an interior point of $\bar{\mathscr{S}}$.

If the support of m is unbounded, we again define the measure $m_k = \zeta(|x|/k)m$ of bounded support and choose k so large that $\int m_k > 1$. If the support of m does not lie in any plane, we choose k so large that the same is true of m_k. We let

$$V_k = \frac{\int x m_k(x, dx)}{\int m_k(x, dx)}.$$

Then by the above arguments

$$V_k \cdot \xi \leqq \bar{c}_k^*(\xi) \leqq \bar{c}^*(\xi)$$

for all ξ, so that $\bar{\mathscr{S}}$ is not empty. If the support of m_k is not concentrated on any plane, then

$$V_k \cdot \xi < \bar{c}_k^*(\xi) \leqq \bar{c}^*(\xi)$$

for all ξ, so that V_k is an interior point of $\bar{\mathscr{S}}$.

The formula (9.3) shows that if the support of m lies in the plane $x \cdot \xi_0 = A$, then $\bar{c}^*(\xi_0) = A$ and $\bar{c}^*(-\xi_0) = -A$. Therefore $\bar{\mathscr{S}}$ again lies in the plane $x \cdot \xi_0 = A$, so that $\bar{\mathscr{S}}$ cannot have interior points.

We have thus proved all the statements of the lemma.

Theorem 6.3 is now an immediate consequence of Lemma 9.1 and Proposition 5.5 with $Q_1 = Q$ and $Q_2 = M$ where $\beta = \pi_1$.

The first two parts of Theorem 6.4 follow from Lemma 9.1 and Proposition 5.5 with $Q_2 = Q$ and $Q_1 = M$ where m is replaced by l and β by ε in the definition (9.1) of M. We must, of course, use the observation that increasing the function c^* enlarges the set $\mathscr{S}$.

To prove the last statement of Theorem 6.4 we first note that the definition (9.1) implies that $M[\rho u] \geqq \rho M[u]$ when ρ is any constant such that $0 \leqq \rho \leqq 1$. Thus M satisfies the conditions of Lemma 8.8. Because $Q \geqq M$, Lemma 8.8 implies that if u_0 is uniformly positive on a ball of radius r, then $u_n \geqq \frac{1}{2}\varepsilon$ on a set $n\bar{\mathscr{S}}''$ when n is sufficiently large. In particular, there is an m_0 such that $u_{m_0} \geqq \frac{1}{2}\varepsilon$ on a ball of radius $r_{\varepsilon/2}$, so that the proof of Lemma 8.6 still implies (8.22). The remainder of the proof of Theorem 6.2 then yields the last part of Theorem 6.4.

If, as in the corollary, the same measure m gives both the upper bound (6.6) and the lower bound (6.9) for arbitrarily small δ, then the linear operator M given by (9.1) with $\beta = \infty$ is the Fréchet derivative of Q at $u = 0$ in the space $C[\mathscr{K}]$ of continuous functions with the maximum norm.

For example if $Q[v]$ is again the value $u(\tau, x)$ of the solution u of (2.14) with $u(0, x) = v(x)$, its Fréchet derivative $M[v]$ is the solution at $t = \tau$ of the corresponding problem in which $f(u)$ is replaced by $f'(0)u$. The maximum principle for parabolic

equations [43, Thm. 3.6.10] shows that if $f(u) \leqq f'(0)u$ for $u \geqq 0$, then (6.7) and (6.9) are valid with $m(x, dx) = C_N \tau^{-N/2} \exp[f'(0)\tau - |x|^2/4\tau]dx$ where C_N is a known constant. The formula (6.10) gives $c^* = 2\sqrt{f'(0)\tau}$, which becomes the usual wave speed for the Fisher equation [15], [31] when τ is set equal to 1.

In the case of various integral operators the formula (6.10) has been found by Aronson [1], Diekmann [10], [11], Diekmann and Kaper [12], Thieme [48], [49] and the author [53], [54]. In a different context it was discovered by Hammersley [22].

In fact, as far as I know, the only explicit formula for an asymptotic speed which does not come from the corollary is one given by Hadeler and Rothe [20] for the Fisher equation in which $f(u)$ is a cubic polynomial.

The following example shows that the Fréchet derivative of Q at $u = 0$ need not satisfy the inequality (6.9) for arbitrarily small δ, even if it satisfies (6.6).

Example 9.1. Let $N = 1$, $\mathscr{H} = \{0, \pm 1, \cdots\}$. Define

$$Q[u](k) = \sinh 1 \sum_{j=-\infty}^{\infty} e^{-|j|}(1 - e^{-|j|u(k-j)}),$$

which satisfies the hypotheses (3.1) with $\pi_0 = 0$, $\pi_1 = \pi_+ = 1$. The Fréchet derivative at $u = 0$ is easily found to be

$$M[u](k) = \sinh 1 \sum_{j=-\infty}^{\infty} |j| e^{-|j|} u(k-j)$$

and the bound (6.6) is a consequence of the fact that the second derivative of the function $1 - e^{-t}$ is negative.

Let $\varepsilon > 0$, $\delta \in (0, 1)$, and $u(j) = \varepsilon \delta_{j0}$. Then

$$Q[u](k) - (1-\delta)M[u](k) = \sinh 1 \, e^{-|k|}(1 - (1-\delta)\varepsilon|k| - e^{-\varepsilon|k|}),$$

which is negative for all sufficiently large values of $|k|$. Thus the inequality (6.9) is not true for any positive value of ε.

10. The hairtrigger effect. (Proof of Theorem 6.5). The proof of Theorem 6.5 is based on the simple observation that when u and l are nonnegative

$$\text{supp}\left\{\int u(x-y)l(y, dy)\right\} = \text{supp}\,(u) + \text{supp}\,(l). \tag{10.1}$$

(We define the sum $V + W$ of two sets of vectors by

$$V + W = \{v + w | v \in V, w \in W\}.)$$

We see from (10.1) and Proposition 4.1 that, if $u_{n+1} = Q[u_n]$ and Q satisfies the inequality $Q[u] \geqq \int u(x-y)l(y, dy)$ for $0 \leqq u \leqq \varepsilon$, then

$$\text{supp}\,(u_n) \supset \text{supp}\,(u_0) + n \text{ supp}\,(l).$$

Suppose that m_0 is so large that the support of e_0, the function defined by (8.17), lies in the interior of a translate of m_0 supp (l). Then if $u_0 \not\equiv 0$ there are a $\rho \in (0, 1]$ and an $\bar{x}$ in R^N such that $u_{m_0}(x - \bar{x}) \geqq \rho e_0(x)$. Since $M[\rho u] \geqq \rho M[u]$ for $\rho \in [0, 1]$, the proof of Lemma 8.8 gives Theorem 6.5.

The following example shows that the extra condition in Theorem 6.5 is needed to produce the hairtrigger.

Example 10.1. If $N = 1$, $\mathscr{H} = \{0, \pm 1, \pm 2, \cdots\}$, and

$$Q[u](x) = \tfrac{1}{2}u(x-1)[2 - u(x-1)] + \tfrac{1}{2}u(x+1)[2 - u(x+1)],$$

Q satisfies the hypotheses of Theorem 6.4 with $\pi_0 = 0$, $\pi_1 = 1$, $\varepsilon = \frac{1}{2}$ and

$$l(x, dx) = \tfrac{3}{4}[\delta(x-1) + \delta(x+1)]$$

where δ is the usual Dirac measure. The set n supp (l) consists of those points whose coordinates are integers of the same parity as n. Hence the hypotheses of Theorem 6.5 are not satisfied. In fact, if $u_0(x) = 1$ for $x = 0$ and 0 elsewhere, $u_n(x) = 0$ when $n + x$ is odd. However, if u_0 is positive at two points of opposite parities, and in particular, at two adjacent points, the above argument leads to the conclusion of Theorem 6.4.

Example 10.2. If $\mathscr{H} = \{0, \pm 1, \pm 2, \cdots\}$,

$$Q[u] = \tfrac{1}{3}u(x-1) + \tfrac{1}{3}u(x)[2 - u(x)] + \tfrac{1}{3}u(x+1)$$

and

$$l = \tfrac{1}{3}[\delta(x-1) + \tfrac{3}{2}\delta(x) + \delta(x+1)],$$

the hypotheses of Theorem 6.5 are satisfied with $\varepsilon = \frac{1}{2}$, $\pi_0 = 0$ and $\pi_1 = 1$. However if $\mathscr{H}$ is the whole real line the hypotheses of Theorem 6.4 are satisfied but those of Theorem 6.5 are not. In fact, if $u_0(x)$ is positive for $-\frac{1}{4} < x < \frac{1}{4}$ and zero elsewhere, $u_n(x)$ is zero when x lies at distance at least $\frac{1}{4}$ from any integer, so that the statement of Theorem 6.5 is not valid for this Q.

Thus while the hypotheses of Theorems 6.1 to 6.4 are preserved when Q is extended from $\mathscr{H}$ to $\mathscr{R}$, the same need not be true of the extra hypothesis in Theorem 6.5.

11. The existence of traveling waves. (Proof of Theorem 6.6). We choose a function φ with the properties 5.1. For each positive integer k we define the sequence $a_n(c, \xi, k; s)$ by the recursion

$$\begin{aligned} &a_{n+1}(c, \xi, k; s) = \max\{k^{-1}\varphi(s), Q[a_n(c, \xi, k; x \cdot \xi + s + c](0)\}, \\ &a_0(c, \xi, k; s) = k^{-1}\varphi(s). \end{aligned} \tag{11.1}$$

As in § 5, we find that $a_n(c, \xi, k; s)$ is nonincreasing in c, k, and s and nondecreasing in n. As $n \to \infty$ it converges to a limit function $a(c, \xi, k; s)$, which is nonincreasing in c, k, and s. By Proposition 5.2

$$\begin{aligned} &\lim_{s \to -\infty} a(c, \xi, k; s) = \pi_1, \\ &\lim_{s \to +\infty} a(c, \xi, k; s) = 0 \quad \text{for } c \geqq c^*(\xi). \end{aligned} \tag{11.2}$$

We now use the additional property (3.7): Every sequence $Q[v_n]$ with $v_n \leqq \pi_1$ has a subsequence $Q[v_{n_i}]$ which converges uniformly on each bounded subset of $\mathscr{H}$.

Then for any given real number t there is a sequence n_i such that the sequence $Q[a_{n_i}(c, \xi, k; x \cdot \xi + t + c)](y)$ converges uniformly for y on bounded subsets of $\mathscr{H}$. Since the sequence a_n is nonincreasing in n and Q is order preserving, it follows that the whole sequence $Q[a_n(c, \xi, k; x \cdot \xi + t + c)](y)$ converges uniformly on bounded sets. Because of the translation invariance of Q, (11.1) then shows that for each t the sequence $a_n(c, \xi, k; y \cdot \xi + t)$ converges to $a(c, \xi, k; y \cdot \xi + t)$, uniformly for y in any bounded subset of $\mathscr{H}$. In particular, $a(c, \xi, k; y \cdot \xi + t)$ is a continuous function of y on $\mathscr{H}$.

Furthermore, because of the continuity property (3.1v) of Q we may now take limits in (11.1) to see that

$$a(c, \xi, k; s) = \max\{k^{-1}\varphi(s), Q[a(c, \xi, k; x \cdot \xi + s + c)](0)\}. \tag{11.3}$$

We choose $y_0 \in \mathcal{H}$ such that $y_0 \cdot \xi > 0$. For every integer l and any $c \geqq c^*(\xi)$ we define the sequence

$$K_k(l) = \tfrac{1}{2}[a(c, \xi, k; ly_0 \cdot \xi) + a(c, \xi, k; (l+1)y_0 \cdot \xi)]. \tag{11.4}$$

Then $K_k(l)$ is nonincreasing in l, $K_k(-\infty) = \pi_1$, and $K_k(+\infty) = 0$. Since a decreases from π_1 to 0 as s goes from $-\infty$ to $+\infty$,

$$K_k(l) - K_k(l-1) = \tfrac{1}{2}[a(c, \xi, k; (l+1)y_0 \cdot \xi) - a(c, \xi, k; (l-1)y_0 \cdot \xi)] \leqq \tfrac{1}{2}\pi_1.$$

Consequently there is an integer, which we call l_k, such that

$$\tfrac{1}{4}\pi_1 \leqq K_k(l_k) \leqq \tfrac{3}{4}\pi_1. \tag{11.5}$$

We now consider the sequence $a(c, \xi, k; x \cdot \xi + l_k y_0 \cdot \xi)$, $k = 1, 2, \cdots$.

Because of the equation (11.3) and the hypothesis (3.7) there is a subsequence k_i of the integers such that $a(c, \xi, k_i; x \cdot \xi + l_{k_i} y_0 \cdot \xi)$ converges uniformly for x on bounded subsets of $\mathcal{H}$ to a function $W(x \cdot \xi)$ defined on $\mathcal{H}$. This subsequence has a subsequence k_i' such that $a(c, \xi, k_i'; x \cdot \xi + l_{k_i'} y_0 \cdot \xi + c)$ also converges uniformly on bounded subsets of $\mathcal{H}$ to a function $W(x \cdot \xi + c)$. (Of course, if c is of the form $z \cdot \xi$ for some z in $\mathcal{H}$, $W(x \cdot \xi + c)$ is already defined and the convergence is implied by the convergence of the previous sequence, but c need not be of this form.) We take further subsequences to get uniform convergence on bounded subsets of $\mathcal{H}$ of $a(c, \xi, k_i^{(m)}; x \cdot \xi + l_{k_i^{(m)}} y_0 \cdot \xi + mc)$ to a function $W(x \cdot \xi + mc)$ for any (positive or negative) integer m. By taking a diagonal sequence $\bar{k}_i$, we obtain simultaneous convergence for all m. Because the convergence is uniform on bounded subsets of $\mathcal{H}$, we may take limits in the equation (11.3) with $k = \bar{k}_i$ and $s = y \cdot \xi + l_{\bar{k}_i} y_0 \cdot \xi - (n+1)c$ to find that

$$W(y \cdot \xi - (n+1)c) = Q[W(x \cdot \xi - nc)](y). \tag{11.6}$$

Thus $u_n(x) = W(x \cdot \xi - nc)$ is a traveling wave solution of the recursion $u_{n+1} = Q[u_n]$.

The definition (11.4) shows that the sequence $K_{\bar{k}_i}(l_{\bar{k}_i})$ converges to $\frac{1}{2}[W(0) + W(y_0 \cdot \xi)] \geqq W(y_0 \cdot \xi)$. By (11.5) this quantity must be at most $\frac{3}{4}\pi_1$. Theorem 6.2 proves that $W(-\infty) = \pi_1$, and therefore W is not constant. Because the functions a are nonincreasing in s, the same is true of $W(s)$. Therefore $W(s)$ has a limit as $s \to \infty$. By letting $n \to -\infty$ in (11.6) we see that $W(\infty) = Q[W(\infty)]$. Therefore $W(\infty)$ is a fixed point of Q whose value is below π_1. Hence it must be 0. This proves Theorem 6.6.

Remarks. 1. If $\mathcal{H}$ is a discrete set and π_1 is finite, the extra condition (3.7) is automatically satisfied. If $\mathcal{H}$ has limit points and π_1 is finite, the condition is normally verified by showing that the set of functions $\{Q[u] | u \in B\}$ is equicontinuous and applying Ascoli's theorem.

2. The set of points of the form $x \cdot \xi + mc$ where $W(s)$ is defined may depend upon ξ and c. As an example, let $\mathcal{H}$ be the set of points with integral coordinates in the plane $\mathcal{R}^2$. If ξ_1, ξ_2, and c are all integral multiples of a fixed real number ρ, then the set of points of the form $x \cdot \xi + mc$ lies in the discrete set of integral multiples of ρ, while if the ratio of two of these three numbers is irrational, the set of points of the form $x \cdot \xi + mc$ is dense on the real line.

If there is a y in $\mathcal{H}$ such that $y \cdot \xi = rc$ for some integer r, then increasing n by r in the function $W(x \cdot \xi - nc)$ translates the function by y. If there is no y in $\mathcal{H}$ such that $y \cdot \xi$ is an integral multiple of c, then the arguments of $W(x \cdot \xi - nc)$ for two different values of n are all different. However, the evolution of the nonincreasing function $W(x \cdot \xi - nc)$ with increasing n will still have the general appearance of a traveling wave.

3. The uniform convergence on bounded subsets of $\mathcal{H}$ does not imply that the limit functions a and W are continuous functions of s.

For example, let $\mathcal{H}$ be the set of points with integral coordinates in $\mathcal{R}^2$ and

$$Q[u] = u(x_1+1, x_2-2)[2-u(x_1+1, x_2-2)].$$

It is easy to see that $c^*(\xi) = -\xi_1 + 2\xi_2$, that if $\varphi(s) > 0$ for $s < 0$

$$a(c^*(\xi), k; s) = \begin{cases} 1 & \text{for } s < 0, \\ 0 & \text{for } s \geqq 0, \end{cases}$$

and that for $c = c^*(\xi)$

$$W(s) = \begin{cases} 1 & \text{for } s < 0, \\ 0 & \text{for } s \geqq 0. \end{cases}$$

Note that if ξ_2/ξ_1 is irrational, there is a sequence x_ν such that $x_\nu \cdot \xi$ decreases to zero. However, $|x_\nu|$ is unbounded so that the sequence x_ν does not lie in any bounded set and does not have a convergent subsequence.

4. It is known [13], [25], [26] that in the heterozygote inferior case of Fisher's equation (1.1) there is a monotone traveling wave when $c = c^*$, but not, in general, for $c > c^*$. Thus the statement of Theorem 6.6 can be false if the condition $\pi_0 = 0$ is dropped. We do not know whether a monotone traveling wave always exists for $c = c^*(\xi)$ when this condition is dropped.

REFERENCES

[1] D. G. ARONSON, *The asymptotic speed of propagation of a simple epidemic*, Nonlinear Diffusion, W. E. Fitzgibbon and H. F. Walker, eds., Research Notes in Mathematics 14, Pitman, London, 1977, pp. 1–23.

[2] ———, *Density dependent interaction diffusion systems*, Dynamics and Modeling of Reactive Systems, W. E. Stewart, W. H. Ray, and C. C. Conley, eds., Academic Press, New York, 1980, pp. 161–176.

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

[4] D. G. ARONSON AND H. F. WEINBERGER, *Nonlinear diffusion in population genetics, combustion, and nerve propagation*, Partial Differential Equations and Related Topics, J. Goldstein, ed., Lecture Notes in Mathematics 446, Springer, New York, 1975, pp. 5–49.

[5] ———, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. in Math., 30 (1978), pp. 33–76.

[6] T. BONNESEN AND W. FENCHEL. *Theorie der konvexen Körper*, Ergeb. d. Math. u. ihrer Grenzgeb., 3, Chelsea, New York, 1948.

[7] M. BRAMSON, *Maximal displacement of branching Brownian motion*, Comm. Pure Appl. Math., 31 (1978), pp. 531–581.

[8] C. CONLEY, *An application of Wazewski's method to a nonlinear boundary value problem which arises in population genetics*, J. Math. Biol., 2 (1975), pp. 241–249.

[9] R. COURANT, *Differential and Integral Calculus*, vol. I., Interscience, New York, 1936.

[10] O. DIEKMANN, *Thresholds and travelling waves for the geographical spread of infection*, J. Math. Biol., 6 (1978), pp. 109–130.

[11] ———, *Run for your life, A note on the asymptotic speed of propagation of an epidemic*, J. Differential Equations, 33 (1979), pp. 58–73.

[12] O. DIEKMANN AND H. G. KAPER, *On the bounded solutions of a nonlinear convolution equation*, J. Nonlin. Analysis, 2 (1978), pp. 721–737.

[13] P. C. FIFE AND J. B. MCLEOD, *The approach of solutions of nonlinear diffusion equations to travelling wave solutions*, A. M. S. Bull., 81 (1975), pp. 1076–1078; Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.

[14] P. C. FIFE AND L. A. PELETIER, *Nonlinear diffusion in population genetics*, Arch. Rational Mech. Anal., 64 (1977), pp. 93–109.

[15] R. A. FISHER, *The advance of advantageous genes*, Ann. Eugenics, 7 (1937), pp. 355–369.
[16] ———, *Gene frequencies in a cline determined by selection and diffusion*, Biometrics, 6 (1950), pp. 353–361.
[17] W. FLEMING, *A selection migration model in population genetics*, J. Math. Biol., 2 (1975), pp. 219–233.
[18] H. FUJITA, *On the blowing up of solutions of the Cauchy problem for* $u_t = \Delta u + u^{1+\alpha}$, J. Fac. Sci. Univ. Tokyo (I), 13 (1966), pp. 109–124.
[19] M. E. GURTIN AND R. C. MACCAMY, *On the diffusion of biological populations*, Math. Biosci., 33 (1977), pp. 35–49.
[20] K. P. HADELER AND F. ROTHE, *Travelling fronts in nonlinear diffusion equations*, J. Math. Biol., 2 (1975), pp. 251–263.
[21] J. B. S. HALDANE, *The theory of a cline*, J. Genetics, 48 (1948), pp. 277–284.
[22] J. M. HAMMERSLEY, *Postulates for subadditive processes*, Ann. Probab., 2 (1974), pp. 652–680.
[23] F. HOPPENSTEAD AND J. M. HYMAN, *Periodic solutions of a logistic difference equation*, SIAM J. Appl. Math, 32 (1977), pp. 73–81.
[24] Y. KAMETAKA, *On the nonlinear diffusion equations of Kolmogorov-Petrovsky-Piskunov type*, Osaka J. Math., 13 (1976), pp. 11–66.
[25] JA. I. KANEL', *Stabilization of solutions of the Cauchy problem for equations encountered in combustion theory*, Mat. Sbornik (N.S.) 101, 59 (1962) supplement, pp. 245–288.
[26] ———, *On the stability of solutions of the equations of combustion theory for finite initial functions*, Mat. Sbornik (N.S.) 107, 65 (1964), pp. 398–413.
[27] S. KARLIN, *Population subdivision and selection migration interaction*, in Population Genetics and Ecology, S. Karlin and E. Nevo, eds., Academic Press, New York, 1976, pp. 617–657.
[28] D. G. KENDALL, *Mathematical models of the spread of infection*, in Mathematics and Computer Science in Biology and Medicine, H.M.S.O., London, 1965, pp. 213–225.
[29] W. D. KERMACK AND A. G. MCKENDRICK, *A contribution to the mathematical theory of epidemics*, Proc. Royal Soc. A, 115 (1927), pp. 700–721.
[30] M. KIMURA, *Stepping-stone model of population*, Ann. Report National Inst. of Genetics of Japan, 3 (1953), pp. 62–63.
[31] A. KOLMOGOROFF, I. PETROVSKY AND N. PISCOUNOFF, *Étude de l'équations de la diffusion avec croissance de la quantité de matière et son application a un problème biologique*, Bull. Univ. Moscow, Ser. Internat., Sec. A, 1 # 6 (1937), pp. 1–25.
[32] T. Y. LI AND J. A. YORKE, *Period three implies chaos*, Amer. Math. Monthly, 82 (1975), pp. 985–992.
[33] D. LUDWIG, D. G. ARONSON AND H. F. WEINBERGER, *Spatial patterning of the spruce budworm*, J. Math. Biol., 8 (1979), pp. 217–258.
[34] D. LUDWIG, D. D. JONES, AND C. S. HOLLING, *Qualitative analysis of insect outbreak systems: the spruce budworm and the forest*, J. Anim. Ecol., 47 (1978), pp. 315–332.
[35] G. MALÉCOT, *Quelques schémas probabilistes sur la variabilité des populations naturelles*, Ann. Univ. Lyon Sci. A, 13 (1950), pp. 37–60.
[36] R. M. MAY AND G. F. OSTER, *Bifurcations and dynamic complexity in simple ecological models*, Amer. Naturalist, 110 (1976), pp. 573–599.
[37] D. MOLLISON, *Possible velocities for a simple epidemic*, Adv. Appl. Prob., 4 (1972), pp. 233–257.
[38] T. NAGYLAKI, *Conditions for the existence of clines*, Genetics, 80 (1975), pp. 595–615.
[39] ———, *A diffusion model for geographically structured populations*, J. Math. Biol., 6 (1978), pp. 375–382.
[40] ———, *The geographical structure of populations*, Studies in Mathematics, vol. 16: Studies in Mathematical Biology, Part II, S. A. Levin, ed., Math. Assoc. of America, Washington, 1978, pp. 588–623.
[41] T. NAGYLAKI AND M. MOODY, *Diffusion model for genotype dependent migration*, Proc. Nat. Acad. of Sci. U.S.A., 77(1980), pp. 4842–4846.
[42] A. J. NICHOLSON, *An outline of the dynamics of animal populations*, Austr. J. Zool., 2 (1954), pp. 9–65.
[43] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
[44] W. E. RICKER, *Stock and recruitment*, J. Fish. Res. Bd. Can., 11 (1954), pp. 559–623.
[45] S. SAWYER, *Results for the stepping stone model for migration in population genetics*, Ann. Probab., 4 (1976), pp. 699–728.
[46] M. SLATKIN, *Gene flow and selection in a cline*, Genetics, 75 (1973), pp. 733–756.
[47] H. R. THIEME, *A model for the spatial spread of an epidemic*, J. Math. Biol., 4 (1977), pp. 337–351.
[48] ———, *Asymptotic estimates of the solutions of nonlinear integral equations and asymptotic speeds for the spread of populations*, J. Reine Angew. Math., 306 (1979), pp. 94–121.

[49] ———, *Density-dependent regulation of spatially distributed populations and their asymptotic speeds of spread*, J. Math. Biol., 8 (1979), pp. 173–178.

[50] K. UCHIYAMA, *The behavior of solutions of the equation of Kolomogorov-Petrovsky-Piskunov*, Proc. Japan. Acad. Ser. A, 53 (1977), pp. 225–228.

[51] ———, *The behavior of solutions of some nonlinear diffusion equations for large time*, J. Math. Kyoto U., 18 (1978), pp. 453–508.

[52] S. UTIDA, *Studies on experimental population of the azuki bean weevil*, Mem. Coll. of Agricult., Kyoto Imp. Univ., 48 (1941), pp. 1–30.

[53] H. F. WEINBERGER, *Asymptotic behavior of a model in population genetics*, in Nonlinear Partial Differential Equations and Applications, J. Chadam, ed., Lecture Notes in Mathematics 648, Springer, New York, 1978, pp. 47–98.

[54] ———, *Asymptotic behavior of a class of discrete time models in population genetics*, in Applied Nonlinear Analysis, V. Lakshmikantham, ed., Academic Press, New York, 1979, pp. 407–422.

[55] ———, *Genetic wave propagation, convex sets, and semi-infinite programming*, in Constructive Approaches to Mathematical Models, C. V. Coffman and G. J. Fix, eds., Academic Press, New York, 1979, pp. 293–317.

[56] ———, *Some deterministic models for the spread of genetic and other alterations*, Biological Growth and Spread—Mathematical Theories and Applications, W. Jäger, H. Rost, and P. Tautu, eds., Lecture Notes in Biomathematics 38, Springer, Berlin-Heidelberg-New York, 1980, pp. 320–349.

SIAM J. MATH. ANAL.
Vol. 13, No. 3, May 1982

0036-1410/82/1303-0003 $01.00/0

GLOBAL SMOOTH SOLUTIONS TO THE INITIAL-BOUNDARY VALUE PROBLEM FOR THE EQUATIONS OF ONE-DIMENSIONAL NONLINEAR THERMOVISCOELASTICITY*

C. M. DAFERMOS†

Dedicated to Michael Golomb

Abstract. The system of balance laws of mass, momentum and energy for a class of one-dimensional thermoviscoelastic materials is considered, and the existence of globally defined smooth thermomechanical processes is established.

1. Introduction. This article is concerned with the problem of existence of globally defined smooth thermomechanical processes in one-dimensional nonlinear thermoviscoelasticity.

The referential (Lagrangian) form of the conservation laws of mass, momentum and energy for one-dimensional materials with reference density $\rho_0=1$ is

$$\begin{aligned} &u_t-v_x=0,\\ &v_t-\sigma_x=0,\\ &\left[e+\frac{1}{2}v^2\right]_t-[\sigma v]_x+q_x=0, \end{aligned} \tag{1.1}$$

while the second law of thermodynamics is expressed by the Clausius–Duhem inequality

$$\eta_t+\left(\frac{q}{\theta}\right)_x\geq 0. \tag{1.2}$$

Here u,v,e,σ,η,θ and q denote deformation gradient, velocity, internal energy, stress, specific entropy, temperature and heat flux, in that order. Note that u,e and θ may only take positive values.

For one-dimensional, homogeneous, thermoviscoelastic materials, internal energy, stress, entropy and heat flux are given by the constitutive relations

$$e=\hat{e}(u,\theta),\quad \sigma=\hat{\sigma}(u,\theta,v_x),\quad \eta=\hat{\eta}(u,\theta),\quad q=\hat{q}(u,\theta,\theta_x), \tag{1.3}$$

which, in order to comply with (1.2), must satisfy

$$\begin{aligned} &\hat{\sigma}(u,\theta,0)=\hat{\psi}_u(u,\theta),\qquad \hat{\eta}(u,\theta)=-\hat{\psi}_\theta(u,\theta),\\ &[\hat{\sigma}(u,\theta,w)-\hat{\sigma}(u,\theta,0)]w\geq 0,\qquad \hat{q}(u,\theta,g)g\leq 0, \end{aligned} \tag{1.4}$$

where $\psi=e-\theta\eta$ is the Helmholtz free energy.

We consider here a body with reference configuration the interval $[0,1]$ whose endpoints are stress-free and thermally insulated, that is,

$$\begin{aligned} &\sigma(0,t)=\sigma(1,t)=0,\qquad t\geq 0,\\ &q(0,t)=q(1,t)=0,\qquad t\geq 0, \end{aligned} \tag{1.5}$$

*Received by the editors September 25, 1981. This research was supported by the Science Research Council of Great Britain, the National Science Foundation under grants MCS-79-05774-02, CME 80-23824, and the U. S. Army under contract ARO-DAAG-29-79-C-0161.

†Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

and we prescribe the initial values of deformation gradient, velocity and temperature:

$$u(x,0)=u_0(x), \quad v(x,0)=v_0(x), \quad \theta(x,0)=\theta_0(x), \qquad 0\le x\le 1. \tag{1.6}$$

The question is whether the combined dissipative effects of viscosity and thermal diffusion may counterbalance the destabilizing influence of nonlinearity and thus induce the existence of globally defined smooth solutions to the initial-boundary value problem (1.1), (1.5), (1.6).

When the material is an ideal, linearly viscous gas with constant specific heats, i.e.,

$$e=c\theta, \quad \sigma=-R\frac{\theta}{u}+\mu\frac{v_x}{u}, \quad q=-\kappa\frac{\theta_x}{u}, \tag{1.7}$$

Kazhikhov [5] establishes the existence of globally defined smooth solutions to (1.1), (1.5), (1.6) by analysis which depends crucially upon the specific form of the constitutive relations (1.7). On the other hand, Dafermos and Hsiao [4] consider (1.1), (1.5), (1.6) for a fairly general class of solidlike linearly viscous materials,

$$e=\hat{e}(u,\theta), \quad \sigma=-\hat{p}(u,\theta)+\mu v_x, \quad q=\hat{q}(u,\theta,\theta_x), \tag{1.8}$$

in which shear viscosity μu is inversely proportional to density and, consequently, its dissipative effect is so weak at high density that existence of global solutions is to be expected only when the initial energy is not too large. Furthermore, the approach in [4] requires growth conditions on $\hat{e},\hat{p}$, and $\hat{q}$ which are not met, for example, when the heat flux satisfies Fourier's law with thermal conductivity that varies with temperature.

To overcome the above limitations we consider here the problem of existence of solutions to (1.1), (1.5), (1.6) for linearly viscous materials

$$e=\hat{e}(u,\theta), \quad \sigma=-\hat{p}(u,\theta)+\hat{\mu}(u)v_x, \quad q=-\hat{\kappa}(u,\theta)\theta_x \tag{1.9}$$

where the viscosity $\hat{\mu}(u)u$ is uniformly positive, that is,

$$\hat{\mu}(u)u\ge\mu_0>0, \qquad 0<u<\infty. \tag{1.10}$$

Unfortunately, our techniques cannot handle the situation where viscosity varies with temperature.

We assume that $\hat{e}(u,\theta)$, $\hat{p}(u,\theta)$, $\hat{\mu}(u)$ and $\hat{\kappa}(u,\theta)$ are twice continuously differentiable on $0<u<\infty$, $0\le\theta<\infty$ and are interrelated by

$$\hat{e}_u(u,\theta)=-\hat{p}(u,\theta)+\theta\hat{p}_\theta(u,\theta), \tag{1.11}$$

so as to comply with (1.4). Furthermore, we require that the elastic part of the stress be compressive at high density and tensile at low density, at any temperature; i.e., there are $0<\tilde{u}\le\tilde{U}<\infty$ such that

$$\begin{aligned} &\hat{p}(u,\theta)\ge 0, \qquad 0<u<\tilde{u}, \quad 0\le\theta<\infty,\\ &\hat{p}(u,\theta)\le 0, \qquad \tilde{U}<u<\infty, \quad 0\le\theta<\infty. \end{aligned} \tag{1.12}$$

Employing an idea of Andrews [2], we show that, as a consequence of (1.10) and (1.12), the deformation gradient is a priori confined in a bounded interval $0<\bar{u}<u(x,t)<\bar{U}$ and hence no restrictions are necessary on the behavior of $\hat{e}(u,\theta)$, $\hat{p}(u,\theta)$, $\hat{\kappa}(u,\theta)$ at $u=0+$ and $u=\infty$. As regards growth with respect to temperature, we assume that for any $0<\bar{u}<\bar{U}<\infty$ there are positive constants ν,κ_0 and N, possibly depending upon $\bar{u}$ and/or $\bar{U}$, such that, for any $\bar{u}<u<\bar{U}$, $0\le\theta<\infty$,

$$\hat{e}(u,0)\ge 0, \qquad \nu\le\hat{e}_\theta(u,\theta)\le N(1+\theta^{1/3}), \tag{1.13}$$

$$|\hat{p}_u(u,\theta)|\le N(1+\theta^{1+1/3}), \qquad |\hat{p}_\theta(u,\theta)|\le N(1+\theta^{1/3}), \tag{1.14}$$

$$(1.15)\qquad \kappa_0 \le \hat{\kappa}(u,\theta) \le N, \quad |\hat{\kappa}_u(u,\theta)| \le N, \quad |\hat{\kappa}_\theta(u,\theta)| \le N, \quad |\hat{\kappa}_{uu}(u,\theta)| \le N.$$

The above growth rates are by no means maximal. At the expense of routine complications in the analysis, $\frac{1}{3}$ in (1.13), (1.14) may be replaced by any $r < \frac{7}{18}$. The slower the growth rate of $\hat{e}(u,\theta)$, the faster the growth rate of $\hat{p}(u,\theta)$ that can be handled by our techniques. Moreover, if specific heat and/or heat conductivity grow with temperature, say $\hat{e}_\theta(u,\theta) \ge \nu(1+\theta^s)$ and/or $\hat{\kappa}(u,\theta) \ge \kappa_0(1+\theta^s)$, then higher growth rates of $\hat{e}(u,\theta)$ and $\hat{p}(u,\theta)$ may be tolerated.

As in [4], the solution to (1.1), (1.5), (1.6) will be sought in the space of Hölder continuous functions. $C^\alpha[0,1]$ denotes the Banach space of functions on $[0,1]$ which are uniformly Hölder continuous with exponent α, while $C^{\alpha,\alpha/2}(Q_T)$ stands for the Banach space of functions on $Q_T = [0,1]\times[0,T]$ which are uniformly Hölder continuous with exponents α in x and $\alpha/2$ in t. The existence theorem reads:

THEOREM 1.1. *Consider the initial-boundary value problem* (1.1), (1.5), (1.6) *under conditions* (1.9)–(1.15). *Assume that* $u_0(x)$, $u_0'(x)$, $v_0(x)$, $v_0'(x)$, $v_0''(x)$, $\theta_0(x)$, $\theta_0'(x)$, $\theta_0''(x)$ *are all in* $C^\alpha[0,1]$ *and* $u_0(x) > 0$, $\theta_0(x) > 0$, $0 \le x \le 1$. *Furthermore, let the initial data be compatible with the boundary conditions* (1.5) *at* (0,0) *and* (1,0). *Then there exists a unique solution* $\{u(x,t), v(x,t), \theta(x,t)\}$ *on* $[0,1]\times[0,\infty)$ *such that for every* $T > 0$ *the functions* $u, u_x, u_t, u_{xt}, v, v_x, v_t, v_{xx}, \theta, \theta_x, \theta_t, \theta_{xx}$ *are all in* $C^{\alpha,\alpha/2}(Q_T)$ *and* $u_{tt}, v_{xt}, \theta_{xt}$ *are in* $L^2(Q_T)$. *Moreover,* $\theta(x,t) > 0$, $\bar{u} < u(x,t) < \bar{U}$, *for* $0 \le x \le 1$, $0 \le t < \infty$, *where* $\bar{u}$ *and* $\bar{U}$ *are positive constants depending on the initial data.*

The theorem can be proved by the procedure devised in [4]; namely, solutions to (1.1), (1.5), (1.6) are visualized as fixed points of a map P on the Banach space $\mathfrak{B}$ of functions $\{U(x,t), V(x,t), \Theta(x,t)\}$ with $U, V, V_x, \Theta, \Theta_x$ in $C^{1/3,1/6}(Q_T)$ and existence is established by means of the Leray–Schauder fixed point theorem. The map P carries $\{U(x,t), V(x,t), \Theta(x,t)\}$ into the solution of a complicated linear "parabolic" system obtained by linearizing (1.1) about $\{U(x,t), V(x,t), \Theta(x,t)\}$. By virtue of the smoothing action of linear parabolic systems, P is completely continuous and its range is contained in the set of functions $\{u(x,t), v(x,t), \theta(x,t)\}$ with $u, u_x, u_t, u_{xt}, v, v_x, v_t, v_{xx}, \theta, \theta_x, \theta_t, \theta_{xx}$ in $C^{\alpha,\alpha/2}(Q_T)$. The construction of P and the precise statements and proofs of its aforementioned properties are presented in [4] and need not be repeated here. What remains to be done in order to complete the list of requirements for the application of the Leray–Schauder fixed point theorem is to show that any possible fixed point of P, i.e., any solution $\{u(x,t), v(x,t), \theta(x,t)\}$ of (1.1), (1.5), (1.6) satisfies the admissibility conditions $\theta(x,t) > 0$, $0 < \bar{u} < u(x,t) < \bar{U}$ and is contained in an a priori bounded set of $\mathfrak{B}$. The object of this paper is to establish these a priori bounds, under the current assumptions (1.9)–(1.15).

The conservation laws of momentum and energy may be visualized as diffusion equations for velocity and temperature. The difficulty lies in that these equations contain coupling terms of superlinear growth, albeit lower order, which may induce finite time blow-up of solutions. One may control the effect of such terms by means of interpolation inequalities, provided that a basic set of a priori bounds is already available. The bounds on total momentum and total energy were employed for that purpose in [4]. In order to improve the results of [4] we establish here an additional estimate (cf. (2.17)) which is motivated by the second law of thermodynamics and embodies the dissipative effects of viscosity and thermal diffusion.

The same techniques work and an identical existence theorem obtains when only one end of the body is stress-free while the other is fixed, say $\sigma(0,t) = 0$, $v(1,t) = 0$. On the other hand, when both ends are fixed, i.e., $v(0,t) = v(1,t) = 0$, a priori bounds on the deformation gradient are only known [6] in the ideal gas case (1.7) so further investigation is required.

Since we are not assuming that $\hat{p}(u,\theta)$ is monotone in u, it would be of interest to determine the asymptotic behavior of solutions, as $t\to\infty$. This has already been done by Andrews and Ball [3] in the framework of (isothermal) viscoelasticity.

2. Derivation of a priori estimates. Superimposing, if necessary, a trivial rigid motion, we normalize the initial data so that

$$\int_0^1 v_0(x)\,dx=0. \tag{2.1}$$

Throughout this section $\{u(x,t),\ v(x,t),\ \theta(x,t)\}$ will denote a fixed solution of (1.1), (1.5), (1.6) on $[0,1]\times[0,\infty)$ in the function class indicated in Theorem 1.1.

Integrating (1.1) over $[0,1]\times[0,t]$ and using the boundary conditions (1.5) we obtain the conservation laws of total momentum and energy:

$$\int_0^1 v(x,t)\,dx=\int_0^1 v_0(x)\,dx=0, \qquad 0\le t<\infty, \tag{2.2}$$

$$\int_0^1\left[e(x,t)+\frac{1}{2}v^2(x,t)\right]dx=\int_0^1\left[e(x,0)+\frac{1}{2}v_0^2(x)\right]dx\overset{\text{def}}{=}E_0, \qquad 0\le t<\infty. \tag{2.3}$$

Substituting σ from (1.9) we may write $(1.1)_2$ in the form

$$v_t+\hat{p}(u,\theta)_x=\left[\hat{\mu}(u)v_x\right]_x, \tag{2.4}$$

while combining $(1.1)_3$ with $(1.1)_2$ and using (1.9), (1.11) and $(1.1)_1$ we obtain

$$\hat{e}_\theta(u,\theta)\theta_t+\theta\hat{p}_\theta(u,\theta)v_x-\hat{\mu}(u)v_x^2=\left[\hat{\kappa}(u,\theta)\theta_x\right]_x. \tag{2.5}$$

Applying the maximum principle on (2.5), recalling that $\theta_0(x)>0$, $0\le x\le 1$, one deduces

PROPOSITION 2.1.

$$\theta(x,t)>0, \qquad 0\le x\le 1, \quad 0\le t<\infty. \tag{2.6}$$

Our next objective is to derive bounds on the deformation gradient. Using $(1.1)_1$ we rewrite (2.4) as

$$v_t+\hat{p}(u,\theta)_x=\hat{M}(u)_{xt}, \tag{2.7}$$

where

$$\hat{M}(u)\overset{\text{def}}{=}\int_1^u\hat{\mu}(w)\,dw. \tag{2.8}$$

By virtue of (1.10), $\hat{M}(u)$ is a strictly increasing function which maps $(0,\infty)$ onto $(-\infty,\infty)$.

PROPOSITION 2.2. *We have*

$$\underline{u}<u(x,t)<\overline{U}, \qquad 0\le x\le 1, \quad 0\le t<\infty, \tag{2.9}$$

where

$$\begin{aligned}\underline{u}&\overset{\text{def}}{=}\hat{M}^{-1}\left(\hat{M}\left(\min\left\{\tilde{u},\ \min_{[0,1]}u_0(\cdot)\right\}\right)-\sqrt{2E_0}\right),\\ \overline{U}&\overset{\text{def}}{=}\hat{M}^{-1}\left(\hat{M}\left(\max\left\{\tilde{U},\ \max_{[0,1]}u_0(\cdot)\right\}\right)+\sqrt{2E_0}\right).\end{aligned} \tag{2.10}$$

Proof. The following argument is based on an idea of Andrews [2]. We integrate (2.7) over $[0,y]\times[s,\tau]$, $0\le y\le 1$, $0\le s<\tau$, and use the boundary condition $(1.5)_1$ to get

$$\hat{M}(u(y,\tau))=\hat{M}(u(y,s))+\int_s^\tau p(y,t)\,dt+\int_0^y v(x,\tau)\,dx-\int_0^y v(x,s)\,dx. \tag{2.11}$$

By (2.3), (2.2) and (1.13),

$$\left[\int_0^y v(x,t)\,dx\right]^2\le\frac{1}{4}\int_0^1 v^2(x,t)\,dx<\frac{1}{2}E_0,\qquad 0\le t<\infty. \tag{2.12}$$

In particular, (2.10) and (2.12) imply $\bar{u}<u_0(x)<\overline{U}$, $0\le x\le 1$. Thus, if $u(x,t)>\bar{u}$ is violated on $[0,1]\times[0,\infty)$, there are $\tau>0$ and $y\in[0,1]$ such that $u(x,t)>\bar{u}$, for $0\le x\le 1$, $0\le t<\tau$, but $u(y,\tau)=\bar{u}$.

Now either $u(y,t)<\tilde{u}$, $0\le t<\tau$, or $u(y,t)<\tilde{u}$ for $0\le s<t\le\tau$ but $u(y,s)=\tilde{u}$. In the former case we apply (2.11) with $s=0$ and use (1.12) and (2.12) to infer

$$\hat{M}(u(y,\tau))>\hat{M}(u_0(y))-\sqrt{2E_0}, \tag{2.13}$$

while in the latter case (2.11) together with (1.12) and (2.12) yield

$$\hat{M}(u(y,\tau))>\hat{M}(\tilde{u})-\sqrt{2E_0}. \tag{2.14}$$

In either case, by $(2.10)_1$, $\hat{M}(u(y,\tau))>\hat{M}(\bar{u})$ which is a contradiction to $u(y,\tau)=\bar{u}$. This shows $u(x,t)>\bar{u}$, $0\le x\le 1$, $0\le t<\infty$.

The proof of $u(x,t)<\overline{U}$, $0\le x\le 1$, $0\le t<\infty$, is quite similar. □

We now fix $T>0$ and restrict our solution to the rectangle $Q_T=[0,1]\times[0,T]$. In the sequel, Λ will denote a generic constant which may depend at most on T,μ_0,ν,κ_0, N and upper bounds of the $C^\alpha[0,1]$ norm of $u_0,u_0',v_0,v_0',v_0'',\theta_0,\theta_0',\theta_0''$. Our principal objective is to show that $\{u(x,t),v(x,t),\theta(x,t)\}$ is a priori bounded in the Banach space $\mathfrak{B}$ referred to in the Introduction, that is:

PROPOSITION 2.3. *We have*

$$\begin{aligned}&\|u\|_{C^{1/3,1/6}(Q_T)}\le\Lambda,\\ &\|v\|_{C^{1/3,1/6}(Q_T)}\le\Lambda,\qquad \|v_x\|_{C^{1/3,1/6}(Q_T)}\le\Lambda,\\ &\|\theta\|_{C^{1/3,1/6}(Q_T)}\le\Lambda,\qquad \|\theta_x\|_{C^{1/3,1/6}(Q_T)}\le\Lambda.\end{aligned} \tag{2.15}$$

The proof of Proposition 2.3 will be partitioned into several steps.

The first observation is that, in view of (2.3) and (1.13),

$$\max_{[0,T]}\int_0^1\theta(x,t)\,dx\le\Lambda. \tag{2.16}$$

We now proceed to get estimates which are motivated by the second law of thermodynamics and embody the dissipative character of viscosity and thermal diffusion.

LEMMA 2.1. *We have*

$$\int_0^T\int_0^1\theta^{-4/3}\theta_x^2\,dx\,dt\le\Lambda, \tag{2.17}$$

$$\int_0^T\int_0^1\theta^{8/3}\,dx\,dt\le\Lambda, \tag{2.18}$$

$$\int_0^T\max_{[0,1]}\theta^{5/3}(\cdot,t)\,dt\le\Lambda. \tag{2.19}$$

Proof. We define

$$\hat{H}(u,\theta)=\int_0^\theta \xi^{-1/3}\hat{e}_\theta(u,\xi)\,d\xi \tag{2.20}$$

and note that, on account of (1.13),

$$|\hat{H}(u,\theta)|\le 2N(1+\theta). \tag{2.21}$$

After a short calculation, using (1.11), we deduce

$$\hat{H}_\theta(u,\theta)=\theta^{-1/3}\hat{e}_\theta(u,\theta), \tag{2.22}$$

$$\hat{H}_u(u,\theta)=\theta^{2/3}\hat{p}_\theta(u,\theta)-\hat{G}(u,\theta), \tag{2.23}$$

where

$$\hat{G}(u,\theta)\overset{\text{def}}{=}\frac{2}{3}\int_0^\theta \xi^{-1/3}\hat{p}_\theta(u,\xi)\,d\xi. \tag{2.24}$$

Thus, setting $H(x,t)=\hat{H}(u(x,t),\,\theta(x,t))$, multiplying (2.5) by $\theta^{-1/3}$ and using (2.22), (2.23), we get

$$H_t+\hat{G}(u,\theta)v_x-\theta^{-1/3}\hat{\mu}(u)v_x^2-\theta^{-1/3}\big[\hat{\kappa}(u,\theta)\theta_x\big]_x=0. \tag{2.25}$$

Integrating (2.25) over $[0,1]\times[0,T]$, integrating by parts with respect to x and recalling the boundary condition $(1.5)_2$ we obtain

$$\begin{aligned}\int_0^T\int_0^1 \hat{\mu}(u)\theta^{-1/3}v_x^2\,dxdt+\frac{1}{3}\int_0^T\int_0^1 \hat{\kappa}(u,\theta)\theta^{-4/3}\theta_x^2\,dxdt\\ =\int_0^1 H(x,T)\,dx-\int_0^1 H(x,0)\,dx+\int_0^T\int_0^1 \hat{G}(u,\theta)v_x\,dxdt.\end{aligned} \tag{2.26}$$

By virtue of (2.24) and (1.14), $|\hat{G}(u,\theta)|\le 2N(1+\theta)$ so, using (1.10), (2.9), (1.15), (2.21), (2.16), and the Cauchy–Schwarz inequality, we deduce

$$\begin{aligned}\mu_0\bar{U}^{-1}\int_0^T\int_0^1\theta^{-1/3}v_x^2\,dxdt+\frac{1}{3}\kappa_0\int_0^T\int_0^1\theta^{-4/3}\theta_x^2\,dxdt\\ \le\Lambda+4N^2\bar{U}\mu_0^{-1}\int_0^T\int_0^1\theta^{7/3}\,dxdt+\frac{1}{2}\mu_0\bar{U}^{-1}\int_0^T\int_0^1\theta^{-1/3}v_x^2\,dxdt.\end{aligned} \tag{2.27}$$

On the other hand, by (2.16) and the Cauchy–Schwarz inequality,

$$\begin{aligned}\int_0^T\int_0^1\theta^{8/3}\,dxdt&\le\Lambda\int_0^T\max_{[0,1]}\theta^{5/3}(\cdot,t)\,dt\\ &\le\Lambda+\Lambda\int_0^T\int_0^1\theta^{2/3}|\theta_x|\,dxdt\\ &\le\Lambda+\frac{1}{2}\int_0^T\int_0^1\theta^{8/3}\,dxdt+\Lambda\int_0^T\int_0^1\theta^{-4/3}\theta_x^2\,dxdt,\end{aligned} \tag{2.28}$$

whence

$$4N^2\bar{U}\mu_0^{-1}\int_0^T\int_0^1\theta^{7/3}\,dxdt\le\Lambda+\frac{1}{6}\kappa_0\int_0^T\int_0^1\theta^{-4/3}\theta_x^2\,dxdt. \tag{2.29}$$

Combining (2.27) with (2.29) we arrive at (2.17), and then (2.28) yields (2.18) and (2.19). □

Remark. Identity (2.26) should be compared with

$$\int_0^T\int_0^1\hat{\mu}(u)\theta^{-1}v_x^2\,dxdt+\int_0^T\int_0^1\hat{\kappa}(u,\theta)\theta^{-2}\theta_x^2\,dxdt=\int_0^1\eta(x,T)\,dx-\int_0^1\eta(x,0)\,dx. \tag{2.30}$$

To get (2.30) multiply (2.5) by θ^{-1} and then integrate over $[0,1]\times[0,T]$, noting that $\theta^{-1}\hat{e}_\theta=\eta_\theta$, $\hat{p}_\theta=\hat{\eta}_u$. In contrast to (2.26), (2.30) yields estimates independent of T and consequently it is expected to play a key role in any investigation of the asymptotic behavior of solutions. On the other hand, (2.30) is inadequate for our present purposes because the term $\theta^{-2}\theta_x^2$ is homogeneous and cannot be used to reduce the degree of superlinear terms.

As a corollary of Lemma 2.1 we have

LEMMA 2.2.

$$\int_0^T\int_0^1 v_x^2\,dxdt\le\Lambda. \tag{2.31}$$

Proof. We multiply (2.4) by v, integrate over $[0,1]\times[0,T]$, integrate by parts with respect to x and recall the boundary condition $(1.5)_1$ to get

$$\begin{aligned}&\frac{1}{2}\int_0^1 v^2(x,T)\,dx+\int_0^T\int_0^1\hat{\mu}(u)v_x^2\,dxdt\\ &\quad=\frac{1}{2}\int_0^1 v_0^2(x)\,dx+\int_0^T\int_0^1\hat{p}(u,\theta)v_x\,dxdt\\ &\quad\le\Lambda+\frac{1}{2}\mu_0\bar{U}^{-1}\int_0^T\int_0^1 v_x^2\,dxdt+\frac{1}{2}\mu_0^{-1}\bar{U}\int_0^T\int_0^1\hat{p}^2(u,\theta)\,dxdt\end{aligned} \tag{2.32}$$

from which (2.31) follows with the help of (1.10), (2.9), (1.14) and (2.18). □

In the following lemmas we employ the bounds obtained thus far in order to estimate by interpolation the square integral of various derivatives of the solution in terms of low powers of

$$Y\overset{\text{def}}{=}\max_{[0,T]}\int_0^1\theta_x^2(x,t)\,dx, \tag{2.33}$$

$$Z\overset{\text{def}}{=}\max_{[0,T]}\int_0^1 v_{xx}^2(x,t)\,dx. \tag{2.34}$$

To prepare the ground note that by (2.16) and Schwarz's inequality,

$$\theta^{3/2}(y,t)\le\Lambda+\frac{3}{2}\int_0^1\theta^{1/2}(x,t)|\theta_x(x,t)|\,dx\le\Lambda+\Lambda Y^{1/2},\qquad 0\le y\le 1, \tag{2.35}$$

whence

$$\max_{Q_T}\theta\le\Lambda+\Lambda Y^{1/3}. \tag{2.36}$$

Similarly, combining

$$v_x^2(y,t)\le\int_0^1 v_x^2(x,t)\,dx+2\int_0^1|v_x(x,t)||v_{xx}(x,t)|\,dx,\qquad 0\le y\le 1, \tag{2.37}$$

with the standard interpolation estimate (e.g., [1, §3])

$$\int_0^1 v_x^2(x,t)\,dx\le 108\int_0^1 v^2(x,t)\,dx+432\left\{\int_0^1 v^2(x,t)\,dx\right\}^{1/2}\left\{\int_0^1 v_{xx}^2(x,t)\,dx\right\}^{1/2} \tag{2.38}$$

and using (2.3), one obtains

$$\max_{Q_T}|v_x|\le\Lambda+\Lambda Z^{3/8}. \tag{2.39}$$

LEMMA 2.3.

$$\max_{[0,T]} \int_0^1 u_x^2(x,t)\,dx \le \Lambda + \Lambda Y^{1/9}. \tag{2.40}$$

Proof. We multiply (2.7) by $\hat{M}(u)_x - v$ and integrate over $[0,1]\times[0,t]$, $0<t\le T$, thus obtaining

$$\begin{aligned} &\frac{1}{2}\int_0^1 \left[\hat{M}(u(x,t))_x - v(x,t)\right]^2 dx - \frac{1}{2}\int_0^1 \left[\hat{M}(u_0(x))_x - v_0(x)\right]^2 dx \\ &\qquad = \int_0^t \int_0^1 \left[\hat{p}_u u_x + \hat{p}_\theta \theta_x\right]\left[\hat{M}(u)_x - v\right] dx d\tau. \end{aligned} \tag{2.41}$$

By virtue of (1.14), (2.8), and (2.9),

$$\begin{aligned} &\left|\int_0^t \int_0^1 \hat{p}_u u_x \left[\hat{M}(u)_x - v\right] dx d\tau\right| \\ &\qquad \le \Lambda \int_0^t \int_0^1 (1+\theta^{4/3})\left\{\left[\hat{M}(u)_x - v\right]^2 + v^2\right\} dx d\tau \\ &\qquad \le \Lambda \int_0^t \left\{1 + \max_{[0,1]} \theta^{5/3}(\cdot,\tau)\right\} \int_0^1 \left[\hat{M}(u(x,\tau))_x - v(x,\tau)\right]^2 dx d\tau \\ &\qquad\quad + \Lambda \int_0^t \left\{1 + \max_{[0,1]} \theta(\cdot,\tau)\right\} \int_0^1 v^2(x,\tau)\,dx d\tau. \end{aligned} \tag{2.42}$$

Similarly, using (1.14) and applying Schwarz's inequality,

$$\begin{aligned} &\left|\int_0^t \int_0^1 \hat{p}_\theta \theta_x \left[\hat{M}(u)_x - v\right] dx d\tau\right| \\ &\qquad \le N \int_0^t \left\{1 + \max_{[0,1]} \theta^{5/3}(\cdot,\tau)\right\} \int_0^1 \left[\hat{M}(u(x,\tau))_x - v(x,\tau)\right]^2 dx d\tau \\ &\qquad\quad + N\left\{1 + \max_{Q_T} \theta^{1/3}\right\} \int_0^t \int_0^1 \theta^{-4/3} \theta_x^2\, dx d\tau. \end{aligned} \tag{2.43}$$

Combining (2.41) with (2.42), (2.43), applying Gronwall's inequality and taking account of (2.3), (2.17), (2.36) and (2.19), we arrive at (2.40). □

LEMMA 2.4.

$$Y \le \Lambda + \Lambda Z^{3/4}, \tag{2.44}$$

$$\int_0^T \int_0^1 \theta_t^2\, dx dt \le \Lambda + \Lambda Z^{3/4}. \tag{2.45}$$

Proof. Define

$$\hat{Q}(u,\theta) \overset{\text{def}}{=} \int_0^\theta \hat{\kappa}(u,\xi)\,d\xi, \tag{2.46}$$

set $Q(x,t) = \hat{Q}(u(x,t),\theta(x,t))$, multiply (2.5) by Q_t and integrate over $[0,1]\times[0,t]$, $0<t\le T$. After an integration by parts with respect to x we obtain

$$\int_0^t \int_0^1 \left\{\hat{e}_\theta \theta_t + \theta \hat{p}_\theta v_x - \hat{\mu} v_x^2\right\} Q_t\, dx d\tau + \int_0^t \int_0^1 \hat{\kappa} \theta_x Q_{xt}\, dx d\tau = 0. \tag{2.47}$$

Note that

$$\begin{aligned} Q_t &= \hat{Q}_u v_x + \hat{\kappa}\theta_t, \\ Q_{xt} &= \left[\hat{\kappa}\theta_x\right]_t + \hat{Q}_u v_{xx} + \hat{Q}_{uu} v_x u_x + \hat{\kappa}_u u_x \theta_t. \end{aligned} \tag{2.48}$$

We estimate each term in (2.47) using (1.13), (1.14), (1.15), (2.17), (2.18), (2.19), (2.39), (2.9), (2.31) and noting that, by account of (2.46) and (1.15), $|\hat{Q}_u| \le N\theta$, $|\hat{Q}_{uu}| \le N\theta$.

$$\int_0^t \int_0^1 \hat{e}_\theta \hat{\kappa} \theta_t^2 \, dx d\tau \ge \nu \kappa_0 \int_0^t \int_0^1 \theta_t^2 \, dx d\tau, \tag{2.49}$$

$$\begin{aligned} \left| \int_0^t \int_0^1 \hat{e}_\theta \theta_t \hat{Q}_u v_x \, dx d\tau \right| &\le \frac{1}{8} \nu \kappa_0 \int_0^t \int_0^1 \theta_t^2 \, dx d\tau + \Lambda \left\{ \max_{Q_T} v_x^2 \right\} \int_0^t \int_0^1 (1 + \theta^{8/3}) \, dx d\tau \\ &\le \frac{1}{8} \nu \kappa_0 \int_0^t \int_0^1 \theta_t^2 \, dx d\tau + \Lambda + \Lambda Z^{3/4}, \end{aligned} \tag{2.50}$$

$$\begin{aligned} &\left| \int_0^t \int_0^1 \{ \theta \hat{p}_\theta v_x - \hat{\mu} v_x^2 \} \hat{Q}_u v_x \, dx d\tau \right| \\ &\quad \le \Lambda \left\{ \max_{Q_T} v_x^2 \right\} \int_0^t \int_0^1 \{ 1 + \theta^{7/3} + v_x^2 \} \, dx d\tau \le \Lambda Z^{3/4}, \end{aligned} \tag{2.51}$$

$$\begin{aligned} &\left| \int_0^t \int_0^1 \{ \theta \hat{p}_\theta v_x - \hat{\mu} v_x^2 \} \hat{\kappa} \theta_t \, dx d\tau \right| \\ &\quad \le \frac{1}{8} \nu \kappa_0 \int_0^t \int_0^1 \theta_t^2 \, dx d\tau + \left\{ \Lambda \max_{Q_T} v_x^2 \right\} \int_0^t \int_0^1 \{ 1 + \theta^{8/3} + v_x^2 \} \, dx d\tau \\ &\quad \le \frac{1}{8} \nu \kappa_0 \int_0^t \int_0^1 \theta_t^2 \, dx d\tau + \Lambda Z^{3/4}, \end{aligned} \tag{2.52}$$

$$\begin{aligned} &\int_0^t \int_0^1 \hat{\kappa} \theta_x [\hat{\kappa} \theta_x]_t \, dx d\tau \\ &\quad = \frac{1}{2} \int_0^1 \hat{\kappa}(u(x,t), \theta(x,t)) \theta_x^2(x,t) \, dx - \frac{1}{2} \int_0^1 \hat{\kappa}(u_0(x), \theta_0(x)) \theta_{0x}^2(x) \, dx \\ &\quad \ge \frac{1}{2} \kappa_0 \int_0^1 \theta_x^2(x,t) \, dx - \Lambda. \end{aligned} \tag{2.53}$$

$$\begin{aligned} \left| \int_0^t \int_0^1 \hat{\kappa} \theta_x \hat{Q}_u v_{xx} \, dx d\tau \right| &\le \left\{ \max_{Q_T} \theta^{5/6} \right\} \left\{ \int_0^t \int_0^1 \theta^{-4/3} \theta_x^2 \, dx d\tau \right\}^{1/2} \\ &\quad \cdot \left\{ \int_0^t \max_{[0,1]} \theta^{5/3}(\cdot, \tau) \int_0^1 v_{xx}^2(x,\tau) \, dx d\tau \right\}^{1/2} \\ &\le \Lambda + \frac{1}{16} \kappa_0 Y + \Lambda Z^{3/4}, \end{aligned} \tag{2.54}$$

$$\begin{aligned} &\left| \int_0^t \int_0^1 \hat{\kappa} \theta_x \hat{Q}_{uu} v_x u_x \, dx d\tau \right| \\ &\quad \le N^2 \left\{ \max_{Q_T} |v_x| \right\} \left\{ \max_{Q_T} \theta^{5/6} \right\} \left\{ \int_0^t \int_0^1 \theta^{-4/3} \theta_x^2 \, dx d\tau \right\}^{1/2} \\ &\qquad \cdot \left\{ \int_0^t \max_{[0,1]} \theta^{5/3}(\cdot, \tau) \int_0^1 u_x^2(x,\tau) \, dx d\tau \right\}^{1/2} \\ &\quad \le \Lambda + \frac{1}{16} \kappa_0 Y + \Lambda Z^{3/4}, \end{aligned} \tag{2.55}$$

$$\left|\int_0^t\int_0^1 \hat\kappa\theta_x\hat\kappa_u u_x\theta_t\,dxd\tau\right| \le \frac{1}{8}\nu\kappa_0\int_0^t\int_0^1\theta_t^2\,dxd\tau+\Lambda\int_0^t\int_0^1[\hat\kappa\theta_x]^2u_x^2\,dxd\tau. \tag{2.56}$$

To estimate the last term in (2.56),

$$\begin{aligned}\Lambda\int_0^t\int_0^1[\hat\kappa\theta_x]^2u_x^2\,dxd\tau &\le \Lambda\int_0^t\max_{[0,1]}[\hat\kappa\theta_x]^2\int_0^1u_x^2(x,\tau)\,dxd\tau\\ &\le[\Lambda+\Lambda Y^{1/9}]\int_0^t\int_0^1|\hat\kappa\theta_x|\,|(\hat\kappa\theta_x)_x|\,dxd\tau\\ &\le[\Lambda+\Lambda Y^{1/9}]\Big\{\max_{Q_T}\theta^{2/3}\Big\}\Big\{\int_0^t\int_0^1\theta^{-4/3}\theta_x^2\,dxd\tau\Big\}^{1/2}\\ &\quad\cdot\Big\{\int_0^t\int_0^1[(\hat\kappa\theta_x)_x]^2dxd\tau\Big\}^{1/2}\\ &\le\Lambda+\frac{1}{16}\kappa_0Y+\Lambda\Big\{\int_0^t\int_0^1[(\hat\kappa\theta_x)_x]^2dxd\tau\Big\}^{3/4}.\end{aligned} \tag{2.57}$$

Finally, to estimate the last term in (2.57) we appeal to (2.5) noting that

$$\int_0^t\int_0^1\hat e_\theta^2\theta_t^2\,dxd\tau\le[\Lambda+\Lambda Y^{2/9}]\int_0^t\int_0^1\theta_t^2\,dxd\tau, \tag{2.58}$$

$$\int_0^t\int_0^1\{\theta\hat p_\theta v_x-\hat\mu v_x^2\}^2dxd\tau\le\Lambda+\Lambda Z^{3/4}, \tag{2.59}$$

and deduce easily

$$\Lambda\Big\{\int_0^t\int_0^1[(\hat\kappa\theta_x)_x]^2dxd\tau\Big\}^{3/4}\le\Lambda+\frac{1}{16}\kappa_0Y+\frac{1}{8}\nu\kappa_0\int_0^t\int_0^1\theta_t^2\,dxd\tau+\Lambda Z^{3/4}. \tag{2.60}$$

Combining (2.47)–(2.60) we obtain

$$\frac{1}{2}\nu\kappa_0\int_0^t\int_0^1\theta_t^2\,dxd\tau+\frac{1}{2}\kappa_0\int_0^1\theta_x^2(x,t)\,dx\le\Lambda+\frac{1}{4}\kappa_0Y+\Lambda Z^{3/4} \tag{2.61}$$

from which (2.44) and (2.45) follow. □

LEMMA 2.5.

$$\max_{[0,T]}\int_0^1v_t^2(x,t)\,dx\le\Lambda+\Lambda Z^{11/12}, \tag{2.62}$$

$$\int_0^T\int_0^1v_{xt}^2\,dxdt\le\Lambda+\Lambda Z^{11/12}. \tag{2.63}$$

Proof. Differentiate formally (2.4) with respect to t, multiply by v_t and integrate over $[0,1]\times[0,t]$, $0<t\le T$. Integrating by parts with respect to x we get

$$\begin{aligned}&\frac{1}{2}\int_0^1v_t^2(x,t)\,dx-\frac{1}{2}\int_0^1v_t^2(x,0)\,dx+\int_0^t\int_0^1\hat\mu(u)v_{xt}^2\,dxd\tau\\ &\quad+\int_0^t\int_0^1\hat\mu'(u)v_x^2v_{xt}\,dxd\tau-\int_0^t\int_0^1\{\hat p_uv_x+\hat p_\theta\theta_t\}v_{xt}\,dxd\tau=0.\end{aligned} \tag{2.64}$$

Using (1.10), (2.9), (2.31), (2.39), (1.14), (2.18), (2.36), (2.44), and (2.45), we estimate every term of (2.64) as follows:

$$\int_0^t\int_0^1\hat\mu(u)v_{xt}^2\,dxd\tau\ge\mu_0\bar U^{-1}\int_0^t\int_0^1v_{xt}^2\,dxd\tau, \tag{2.65}$$

$$\left|\int_0^t\int_0^1 \hat{\mu}'(u)v_x^2 v_{xt}\,dxd\tau\right| \le \frac{1}{4}\mu_0\bar{U}^{-1}\int_0^t\int_0^1 v_{xt}^2\,dxd\tau + \Lambda\left\{\max_{Q_T} v_x^2\right\}\int_0^t\int_0^1 v_x^2\,dxd\tau$$
$$\le \frac{1}{4}\mu_0\bar{U}^{-1}\int_0^t\int_0^1 v_{xt}^2\,dxd\tau + \Lambda + \Lambda Z^{3/4}, \tag{2.66}$$

$$\left|\int_0^t\int_0^1 \hat{p}_u v_x v_{xt}\,dxd\tau\right| \le \frac{1}{4}\mu_0\bar{U}^{-1}\int_0^t\int_0^1 v_{xt}^2\,dxd\tau + \Lambda\left\{\max_{Q_T} v_x^2\right\}\int_0^t\int_0^1 (1+\theta^{8/3})\,dxd\tau$$
$$\le \frac{1}{4}\mu_0\bar{U}^{-1}\int_0^t\int_0^1 v_{xt}^2\,dxd\tau + \Lambda + \Lambda Z^{3/4}, \tag{2.67}$$

$$\left|\int_0^t\int_0^1 \hat{p}_\theta\theta_t v_{xt}\,dxd\tau\right| \le \frac{1}{4}\mu_0\bar{U}^{-1}\int_0^t\int_0^1 v_{xt}^2\,dxd\tau + \Lambda\left\{1+\max_{Q_T}\theta^{2/3}\right\}\int_0^t\int_0^1 \theta_t^2\,dxd\tau$$
$$\le \frac{1}{4}\mu_0\bar{U}^{-1}\int_0^t\int_0^1 v_{xt}^2\,dxd\tau + \Lambda + \Lambda Z^{11/12}. \tag{2.68}$$

Combining (2.64)–(2.68) we arrive at (2.62) and (2.63). □

LEMMA 2.6.

$$\max_{[0,T]}\int_0^1 v_t^2(x,t)\,dx \le \Lambda, \tag{2.69}$$

$$\int_0^T\int_0^1 v_{xt}^2\,dxdt \le \Lambda, \tag{2.70}$$

$$\max_{[0,T]}\int_0^1 v_{xx}^2(x,t)\,dx \le \Lambda. \tag{2.71}$$

Proof. In view of (2.62), (2.63), and (2.34), it suffices to show $Z \le \Lambda$. To this end we employ (2.4), whose terms can be estimated, with the help of (1.10), (2.9), (2.39), (1.14), (2.44), (2.36), as follows:

$$\int_0^1 \hat{\mu}^{-2}\hat{\mu}'^2 u_x^2 v_x^2\,dx \le \Lambda\left\{\max_{Q_T} v_x^2\right\}\int_0^1 u_x^2\,dx \le \Lambda + \Lambda Z^{5/6}, \tag{2.72}$$

$$\int_0^1 \hat{\mu}^{-2}\hat{p}_u^2 u_x^2\,dx \le 2N^2\left\{1+\max_{Q_T}\theta^{8/3}\right\}\int_0^1 u_x^2\,dx \le \Lambda + \Lambda Z^{3/4}, \tag{2.73}$$

$$\int_0^1 \hat{\mu}^{-2}\hat{p}_\theta^2\theta_x^2\,dx \le 2N^2\left\{1+\max_{Q_T}\theta^{2/3}\right\}\int_0^1 \theta_x^2\,dx \le \Lambda + \Lambda Z^{11/12}. \tag{2.74}$$

By virtue of (2.72), (2.73) and (2.74), (2.4) yields

$$Z \le \Lambda + \Lambda Z^{11/12}, \tag{2.75}$$

whence $Z \le \Lambda$. □

LEMMA 2.7.

$$\max_{[0,T]}\int_0^1 \theta_t^2(x,t)\,dx \le \Lambda, \tag{2.76}$$

$$\int_0^T\int_0^1 \theta_{xt}^2\,dxdt \le \Lambda, \tag{2.77}$$

$$\max_{[0,T]}\int_0^1 \theta_{xx}^2(x,t)\,dx \le \Lambda. \tag{2.78}$$

Proof. We differentiate formally (2.5) with respect to t, multiply by $\hat{e}_\theta \theta_t$, integrate over $[0,1]\times[0,t]$, $0<t\leq T$, and integrate by parts. After a lengthy sequence of routine estimations, which are recorded in [4] and thus need not be reproduced here, one obtains (2.76), (2.77), and (2.78). □

Proof of Proposition 2.3. By (2.76), (2.77) and Schwarz's inequality, $\theta(x,t)$ is uniformly Hölder continuous in t with exponent $\frac{1}{2}$ while (2.78) implies that $\theta_x(x,t)$ is uniformly Hölder continuous in x with exponent $\frac{1}{2}$. It then follows from a standard interpolation property (e.g. [7, II, Lemma 3.1]) that $\theta_x(x,t)$ is also uniformly Hölder continuous in t with exponent $\frac{1}{6}$, hence $\|\theta_x\|_{C^{1/3,1/6}(Q_T)}\leq\Lambda$. This immediately yields $\|\theta\|_{C^{1/3,1/6}(Q_T)}\leq\Lambda$. Similarly, using (2.69), (2.70), (2.71), we deduce that $\|v_x\|_{C^{1/3,1/6}(Q_T)}\leq\Lambda$ and thereby $\|v\|_{C^{1/3,1/6}(Q_T)}\leq\Lambda$, $\|u\|_{C^{1/3,1/6}(Q_T)}\leq\Lambda$. □

Acknowledgment. This paper was written while I was visiting Heriot–Watt University. I wish to thank John Ball, Jack Carr and Robin Knops for their hospitality and for providing a stimulating research environment.

REFERENCES

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

[2] G. ANDREWS, *On the existence of solutions to the equation* $u_{tt}=u_{xxt}+\sigma(u_x)_x$, J. Differential Equations, 35 (1980), pp. 200–231.

[3] G. ANDREWS AND J. M. BALL, *Asymptotic behavior and changes of phase in one-dimensional nonlinear viscoelasticity*, J. Differential Equations, to appear.

[4] C. M. DAFERMOS AND L. HSIAO, *Global smooth thermomechanical processes in one-dimensional nonlinear thermoviscoelasticity*, J. Nonlinear Analysis, to appear.

[5] A. V. KAZHIKHOV, *Sur la solubilité globale des problèmes monodimensionnels aux valeurs initiales-limitées pour les équations d'un gaz visqueux et calorifère*, C. R. Acad. Sci. Paris, Ser. A, 284 (1977), pp. 317–320.

[6] A. V. KAZHIKHOV AND V. V. SHELUKHIN, *Unique global solution with respect to time of initial-boundary value problems for one-dimensional equations of a viscous gas*, PMM, 41 (1977), 282–291; English translation, Appl. Math. Mech., 41 (1977), 273–282.

[7] O. A. LADYŽENSKAJA AND V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, (Translated from the Russian by S. Smith), American Mathematical Society, Providence, RI, 1968.

SIAM J. MATH. ANAL.
Vol. 13, No. 3, May 1982

0036-1410/82/1303-0004 $01.00/0

ASYMPTOTIC PROPERTIES OF BEST $L_p[0, 1]$ APPROXIMATION BY SPLINES*

D. D. PENCE† AND P. W. SMITH‡

Dedicated to Michael Golomb on the occasion of his seventieth birthday. We were fortunate to benefit from his scholarship and teaching.

Abstract. Precise asymptotic error expressions are derived for best spline approximation to smooth functions in $L_p[0, 1]$ when the knots are chosen by a knot quantile function. These results lead to error expressions for the free knot problem.

1. Introduction. The purpose of this paper is to extend, refine and clarify the results in Barrow and Smith [1]. Given a sequence of knots $\mathbf{t}^N := \{0 = t_0 < t_1 < \cdots < t_N = 1\}$, let $S_N^k(\mathbf{t}^N)$ denote the space of polynomial splines of order k with simple knots $\mathbf{t}^N$, i.e., the space of piecewise polynomials of degree less than k in C^{k-2} with breakpoints $\mathbf{t}^N$. Barrow and Smith [1], [2] dealt with the approximation by splines whose knots were determined by a diffeomorphism t, called a knot quantile function. Let t map $[0, 1]$ onto itself with $t' \geqq 0$, and let $S_N^k(t) := S_N^k(\{t(i/N)\}, i = 0, 1, \cdots, N)$. One of the major results of [1] was the determination of the precise behavior of the $L_2[0, 1]$ distance of some $f \in C^k[0, 1]$ from $S_N^k(t)$ as $N \to \infty$.

It has become apparent that one can significantly reduce the hypotheses on both t and f found in [1] while still preserving the conclusions there. In addition, one can obtain analogous results for $L_p[0, 1]$, $1 \leqq p < \infty$ by modifying the arguments of [1] and using a result of Zhensykbaev [9]. Specifically, we will prove under certain general hypotheses on f and t (see Theorem 4.4) that

$$\lim_{N\to\infty} N^k \|f - P_N^{k,p}(t)f\|_{p,[0,1]} = C_{k,p}(J_{k,p}(f, t))^{1/p}, \tag{1.1}$$

where $P_N^{k,p}(t)$ denotes the $L_p[0, 1]$ metric projection onto $S_N^k(t)$, $C_{k,p}$ is a known constant, and

$$J_{k,p}(f, t) := \int_0^1 |f^{(k)}(t(x))|^p (t'(x))^{kp+1}\, dx.$$

One can find a proof of (1.1) for $p = 2$, $f \in C^k[0, 1]$, and $0 < \delta < t' < M$ in [1]. By making slight modifications of that proof, it is possible to cover the case $p = 2$ where $f^{(k)}$ is piecewise continuous on $(0, 1)$ and monotone near 0 and 1, and where the set of zeros of t' has content zero. However this method does not extend easily to $p \neq 2$, and thus we use quite different methods in the next sections.

2. Lower bounds. Let $B_k(\cdot)$ denote the kth Bernoulli polynomial on $[0, 1]$. See [1], [2], [6] and [7] for various properties of these polynomials. In particular, we note that

$$B_k(x) = (-1)^{\lfloor (k-1)/2 \rfloor} k!(2\pi)^{-k} 2 \sum_{m=1}^{\infty} m^{-k} \cos\left(2\pi m x - \frac{1-(-1)^k}{4}\pi\right).$$

* Received by the editors June 5, 1981, and in final form July 3, 1981.

† Department of Mathematics, University of Vermont, Burlington, Vermont 05405. The research of this author was partially supported by University of Vermont Institutional Grant UVM-PS5.

‡ Department of Mathematical Sciences, Old Dominion University, Norfolk, Virginia 23508. The research of this author was partially supported by the U.S. Army Research Office under grants DAHC04-75-0816 and DAAG 29-78-G-0097.

Thus it is possible to recognize that Zhensykbaev uses a multiple of this Bernoulli polynomial in the proof of the theorem in [9]. There the following is demonstrated.

THEOREM 2.1. (Zhensykbaev). *Let* $\mathring{P}_N^{k,p}(\mathbf{Z}^N)$ *denote the* $L_p[0,1]$ *metric projection onto the subspace of* $S_N^k(\mathbf{Z}^N)$ *consisting of all one-periodic functions. Then*

$$\inf_{\mathbf{z}^N} N^k \|B_k(\cdot) - \mathring{P}_N^{k,p}(\mathbf{Z}^N) B_k(\cdot)\|_{p,[0,1]} = C_{k,p}^*,$$

where $C_{k,p}^* = \min_{\gamma \in R} \|B_k(\cdot) - \gamma\|_{p,[0,1]}$.

Following the argument in [1, p. 301], we now use the above theorem. Since $B_k(\cdot)$ is a monic polynomial, we can conclude for any polynomial g of degree k that

$$\begin{aligned} \inf_{\mathbf{t}^N} N^k \|g - P_N^{k,p}(\mathbf{t}^N) g\|_{p,[0,1]} &\geqq \inf_{\mathbf{z}^{N+2k}} N^k \|g - \mathring{P}_{N+2k}^{k,p}(\mathbf{Z}^{N+2k}) g\|_{p,[0,1]} \\ &\geqq \left(\frac{N}{N+2k}\right)^k C_{k,p} |g^{(k)}(0)|, \end{aligned} \tag{2.1}$$

where $C_{k,p} := C_{k,p}^*/k!$

Now consider $f \in C^k[0,1]$ and a subinterval $J \subseteq [0,1]$. Let $\bar{f}$ denote the kth degree Taylor expansion of f about some point $a \in J$, and let R denote the remainder function. Let $\tau: [0,1] \to J$ denote the linear change of variable, i.e., τ' is constantly equal to $|J|$, the length of the interval J. Suppose that only the $L-1$ points $t_{L_0+1}, \cdots, t_{L_0+L-1}$ from $\{\mathbf{t}^N\}$ lie in the interior of J and that $\tau(z_i^*) = t_{L_0+i}$, $i = 1, \cdots, L-1$. Let $\mathbf{z}^{*L} = \{0 = z_0^* < z_1^* < \cdots < z_{L-1}^* < z_L^* = 1\}$. Then

$$\begin{aligned} \|f - P_N^{k,p}(\mathbf{t}^N) f\|_{p,J} &= |J|^{1/p} \|f \circ \tau - (P_N^{k,p}(\mathbf{t}^N) f) \circ \tau\|_{p,[0,1]} \\ &\geqq |J|^{1/p} \{\|\bar{f} \circ \tau - (P_N^{k,p}(\mathbf{t}^N) f) \circ \tau\|_{p,[0,1]} - \|(f - \bar{f}) \circ \tau\|_{p,[0,1]}\} \\ &\geqq |J|^{1/p} \{\|\bar{f} \circ \tau - (P_L^{k,p}(\mathbf{z}^{*L}))(\bar{f} \circ \tau)\|_{p,[0,1]} - \|R \circ \tau\|_{p,[0,1]}\}. \end{aligned} \tag{2.2}$$

The first inequality follows from the triangle inequality while the second follows by replacing the spline $(P_N^{k,p}(\mathbf{t}^N) f) \circ \tau$ with the best approximation to $\bar{f} \circ \tau$ on $[0,1]$. Further

$$|J|^{1/p} \|R \circ \tau\|_{p,[0,1]} = \|R\|_{p,J} \leqq \frac{2|J|^{(kp+1)/p}}{k!(kp+1)^{1/p}} \max_{\xi \in J} |f^{(k)}(\xi) - f^{(k)}(a)|. \tag{2.3}$$

Let $\omega(f^{(k)}, a, |J|) = \max_{\xi \in J} |f^{(k)}(\xi) - f^{(k)}(a)|$. Now multiplying (2.2) by L^k, taking the infimum of the right-hand side over all $\mathbf{z}^L$, and using (2.1) and (2.3), we can establish that

$$L^k \|f - P_N^{k,p}(t) f\|_{p,J} \geqq \left[\left(\frac{L}{L+2k}\right)^k C_{k,p} |f^{(k)}(a)| - \frac{2L^k}{k!(kp+1)^{1/p}} \omega(f^{(k)}, a, |J|)\right] |J|^{k+1/p}. \tag{2.4}$$

We can now prove the following sequence of theorems giving the desired lower bounds.

THEOREM 2.2. *Let* $f \in C^k[0,1]$ *and the bijection* $t \in C^1[0,1]$. *Then for* $1 \leqq p < \infty$

$$\varliminf_{N \to \infty} N^k \|f - P_N^{k,p}(t) f\|_{p,[0,1]} \geqq C_{k,p} (J_{k,p}(f,t))^{1/p}.$$

Proof. Let L be an arbitrary positive integer. For each $N > L$, let

$$M_N = \left[\frac{N}{L}\right] \quad \left(\text{the greatest integer} \leqq \frac{N}{L}\right),$$

$$I_j^N = \left[(j-1)\frac{L}{N}, j\frac{L}{N}\right], \qquad j = 1, \cdots, M_N,$$

$$J_j^N = t(I_j^N), \qquad j = 1, \cdots, M_N.$$

In the estimates that follow $K_1, K_2, \cdots$ will denote positive constants independent of N and ξ_j^N will denote a point chosen from J_j^N, $j = 1, \cdots, M_N$. Using (2.4) we obtain

$$\begin{aligned} & N^{kp}\|f - P_N^{k,p}(t)f\|_{p,[0,1]}^p \\ & \quad \geqq \left(\frac{N}{L}\right)^{kp} \sum_{j=1}^{M_N} L^{kp}\|f - P_N^{k,p}(t)f\|_{p,J_j^N}^p \\ & \quad \geqq \left(\frac{N}{L}\right)^{kp} \sum_{j=1}^{M_N} \left| \left(\frac{L}{L+2k}\right)^k C_{k,p}|f^{(k)}(\xi_j^N)| - K_1 L^k \omega(f^{(k)}, \xi_j^N, J_j^N) \right|^p |J_j^N|^{kp+1} \\ & \quad \geqq \left(\frac{N}{L}\right)^{kp} \sum_{j=1}^{M_N} \left[\left(\frac{L}{L+2k}\right)^{kp} C_{k,p}^p |f^{(k)}(\xi_j^N)|^p - K_2 L^k \omega(f^{(k)}, \xi_j^N, J_j^N) \right] |J_j^N|^{kp+1}. \end{aligned} \tag{2.5}$$

There exist points $\eta_j^N \in I_j^N$ such that

$$|J_j^N| = t'(\eta_j^N)\left(\frac{L}{N}\right). \tag{2.6}$$

We can set $\xi_j^N = t(\eta_j^N)$ since these points were chosen arbitrarily from J_j^N, $j = 1, \cdots, M_N$. Combining (2.5) and (2.6) gives

$$\begin{aligned} N^{kp}\|f - P_N^{k,p}(t)f\|_{p,[0,1]}^p \geqq & \sum_{j=1}^{M_N} \left(\frac{L}{L+2k}\right)^{pk} C_{k,p}^p |f^{(k)}(t(\eta_j^N))|^p (t'(\eta_j^N))^{kp+1}\left(\frac{L}{N}\right) \\ & - \sum_{j=1}^{M_N} K_2 L^k \omega(f^{(k)}, \xi_j^N, J_j^N)(t'(\eta_j^N))^{kp+1}\left(\frac{L}{N}\right). \end{aligned}$$

The first summation lacks only a term for the subinterval $[M_N(L/N), 1]$ in order to be a Riemann sum for the interval $[0, 1]$. Thus as $N \to \infty$, it tends to

$$\left(\frac{L}{L+2k}\right)^{pk} C_{k,p}^p \int_0^1 |f^{(k)}(t(x))|^p (t'(x))^{kp+1}\, dx.$$

The second summation is bounded above by $K_2 L^{k+1}\omega(f^{(k)}, \{\max t'\}(L/N))\{\max t'\}^{kp+1}(M_N/N)$. This goes to zero as $N \to \infty$ because $f^{(k)}$ is continuous and its modulus of continuity, $\omega(f^{(k)}, h)$, tends to zero as $h \to \infty$.

Therefore,

$$\lim_{N\to\infty} N^{kp}\|f - P_N^{k,p}(t)f\|_{p,[0,1]}^p \geqq \left(\frac{L}{L+2k}\right)^{pk} C_{k,p}^p J_{k,p}(f, t). \tag{2.7}$$

But (2.7) must be true for all integers $L \geqq 1$. Taking the limit as $L \to \infty$ we obtain the pth power of the desired result.

THEOREM 2.3. *Let $f \in L_p[0,1] \cap C^k[t(a), t(b)]$ where $[a, b] \subset [0,1]$ and the bijection $t \in C^1[a, b]$. Then for $1 \leqq p < \infty$*

$$\lim_{N\to\infty} N^{kp} \|f - P_N^{k,p}(t) f\|^p_{p,[t(a),t(b)]} \geqq C^p_{k,p} \int_a^b |f^{(k)}(t(x))|^p (t'(x))^{kp+1}\, dx.$$

Proof. Actually Theorem 2.2 is a special case, and the only modification of the proof needed for this more general result is that all of the summations now should only be for $j = A_N, \cdots, B_N$ where $a \leqq (A_N - 1)L/N \leqq B_N L/N \leqq b$.

Before stating the next theorem we recall the definition of *content zero*. A subset G of the real numbers has *content zero* if for every $\varepsilon > 0$ there is a set D_ε which contains G and has Lebesgue measure less than ε, and D_ε is the finite union of intervals.

THEOREM 2.4. *Suppose the set of points where either $f^{(k)}(\cdot)$ or $t'(\cdot)$ fails to be continuous has content zero. Then for $1 \leqq p < \infty$*

$$\lim_{N\to\infty} N^k \|f - P_N^{k,p}(t) f\|_{p,[0,1]} \geqq C_{k,p} (J_{k,p}(f, t))^{1/p}.$$

Proof. For every $\varepsilon > 0$, there is a finite union of open subintervals D covering all of the points where either $f^{(k)}$ or t' fails to be continuous with $|D| < \varepsilon$. Theorem 2.3 applies on the finite subintervals in D^c, the complement of D in $[0, 1]$.

$$\begin{aligned} \lim_{N\to\infty} N^{kp} \|f - P_N^{k,p}(t) f\|^p_{p,[0,1]} &\geqq \lim_{N\to\infty} N^{kp} \|f - P_N^{k,p}(t) f\|^p_{p,D^c} \\ &\geqq C^p_{k,p} \int_{D^c} |f^{(k)}(t(x))|^p (t'(x))^{kp+1}\, dx. \end{aligned} \tag{2.8}$$

Letting $\varepsilon \to 0$ in (2.8) gives the pth power of the desired result. Note that the lower bound may be ∞ since $f^{(k)}$ and/or t' may be unbounded.

3. The operators Q_N. We have obtained appropriate lower bounds for the error in the previous section. This section develops the necessary tools and techniques which will be combined in § 4 to produce the proper upper bound.

Let $\{N_{l,k,t,N}\}_{l=1-k}^{N-1}$ denote the normalized B-spline basis for $S_N^k(t)$, using the knot set $t_i = 0$ for $i \leqq 0$, $t_i = t(i/N)$ for $i = 1, \cdots, N-1$ and $t_i = 1$ for $i \geqq N$. Let $h_i = t_{i+1} - t_i$ for all i. For simplicity we just write $N_l = N_{l,k,t,N}$ when the other parameters remain fixed. The subscript N will be added when we wish to emphasize that N is changing. We shall make use of several results of de Boor. A very nice survey of these and many other properties of B-splines can be found in [4].

LEMMA 3.1. *Let D_k be the smallest number with the property that for every knot set $\mathbf{t}$, every i, there exists a function $H_i \in L_\infty[0,1]$ such that*

$$\operatorname{supp} H_i \subseteq [t_i, t_{i+k}], \quad \|H_i\|_{\infty,[0,1]} < \frac{D_k}{(t_{i+k} - t_i)}, \quad \int_0^1 H_i N_j = \delta_{ij} \quad \text{for all } j.$$

Then $(\pi/2)^k/2 \leqq D_k \leqq 2k 9^{k-1}$.

LEMMA 3.2. *Let E be the diagonal matrix $[\cdots, (t_{i+k} - t_i)/k, \cdots]$. Then*

$$D_k^{-1} \|E^{1/p} \boldsymbol{\alpha}\|_p \leqq \|\textstyle\sum \alpha_l N_l\|_{p,[0,1]} \leqq \|E^{1/p} \boldsymbol{\alpha}\|_p$$

for all $\boldsymbol{\alpha} \in R^{N+k}$, $1 \leqq p \leqq \infty$, where $\|\cdot\|_p$ denotes the sequence norm on $l_p(R^{N+k})$.

Further, a class of local approximation operators $\mathcal{Q}_N$ which are linear projections from $L_p[0,1]$ onto $S_N^k(t)$ can be defined using the functions H_i from Lemma 3.1:

$$\mathcal{Q}_N f := \textstyle\sum \mu_l(f) N_l := \sum \left(\int_0^1 H_l f \right) N_l.$$

The norm of each operator $\not{p}_N$ is bounded by D_k for each $1 \leqq p \leqq \infty$, independently of both N and the knot function t, because

$$|\mu_l(f)| := \left|\int_0^1 H_l f\right| \leqq \left\{\int_{t_i}^{t_{i+k}} |f|^p\right\}^{1/p} \|H_l\|_{\infty,[0,1]} \leqq \|f\|_{p,[t_i,t_{i+k}]} \frac{D_k}{(t_{i+k}-t_i)^{1/p}}.$$

These operators $\not{p}_N$ are examples of locally-defined spline approximation schemes. We now define another such class. Let $B_{k,p}(x) := \sum_{j=0}^k b_{j,k,p} x^j := B_k(x) - \gamma_{k,p}$ be the vertically shifted Bernoulli polynomial of order $k+1$ satisfying

$$\|B_k(\cdot) - \gamma_{k,p}\|_{p,[0,1]} = \min_{\gamma \in R} \|B_k(\cdot) - \gamma\|_{p,[0,1]} =: C^*_{k,p}.$$

Note that for $p = 2$, $\gamma_{k,2} = 0$ and $C_{k,2} := C^*_{k,2}/k! = (|B_{2k}|/(2k)!)^{1/2}$, where B_{2k} is the $2k$th Bernoulli number. Since k and p remain fixed throughout this paper, we simply write $b_j = b_{j,k,p}$.

We modify slightly the definitions of the operators Q_N^α found in [2, p. 15]. For simplicity we only give the case corresponding to $\alpha = 0$, leaving the more general case to the interested reader. Let i be an integer, $0 \leqq i < N$, such that $i = mk$ for some integer m. On the interval $I_i := [t_i, t_{i+1})$, set

$$\bar{f}_i(\tau) := \sum_{j=0}^k f_i^j (\tau - t_i)^j := \sum_{j=0}^k \left[\frac{f^{(j)}(t_i)}{j!}\right](\tau - t_i)^j,$$

the Taylor expansion for f about t_i where defined. Denote

$$Q_N f := Q_N^{k,p}(t) f := \sum_l \lambda_{l,N}(f) N_{l,N} := \sum_l a_l N_l.$$

Each operator Q_N will be defined by the requirement that for each $i = 0 \pmod k$

$$(\bar{f}_i - Q_N f)(\tau) = f_i^k B_{k,p}\left(\frac{\tau - t_i}{h_i}\right) h_i^k \quad \text{for } \tau \in I_i.$$

More precisely, this means that

$$a_l = \sum_{j=0}^{k-1} \xi_{l,i}^{j+1} (f_i^j - f_i^k b_j h_i^{k-j}), \qquad l = i-k+1, \cdots, i, \tag{3.1}$$

where

$$\xi_{l,i}^j := (-1)^{j-1} \frac{(j-1)!}{(k-1)!} \phi_l^{(k-j)}(t_i) \quad \text{and} \quad \phi_l(\tau) := \prod_{s=1}^{k-1} (\tau - t_{l+s}). \tag{3.2}$$

Set $a_l = 0$ for any index not covered above (i.e., for l such that $mk < l < N \leqq (m+1)k$).

Each operator Q_N is well defined for a function $f \in C^k[0, 1]$. Each coefficient functional $\lambda_{l,N}$ is well defined at least when $f \in C^k[t_l, t_{l+k}]$.

LEMMA 3.3. *Let $f \in C^k[a, b]$, $t \in C^1[a, b]$ with $0 < \delta \leqq t' \leqq M < \infty$ on $[a, b]$ and $0 \leqq a < \bar{t} < b \leqq 1$. For each N, let j be chosen so that $t_j \leqq \bar{t} < t_{j+1}$. Thus, for N sufficiently large, $a < t_j \leqq \bar{t} < t_{j+1} < b$. With $h_j = t_{j+1} - t_j$ and $x = t^{-1}(\bar{t})$, let*

$$R_N(\tau; \bar{t}) = N^k (f - Q_N f)(t_j + \tau h_j), \qquad P(\tau; \bar{t}) = (t'(x))^k f^{(k)}(\bar{t}) \frac{B_{k,p}(\tau)}{k!}.$$

Then there exists a sequence of positive constants $\{\varepsilon_N\}$, tending to zero, which may be chosen independently of $\bar{t} \in (a, b)$ but which depend upon k, p, δ, M such that

$$\|R_N(\cdot; \bar{t}) - P(\cdot, \bar{t})\|_{\infty,[0,1]} < \varepsilon_N.$$

Proof outline. We can replace f by its Taylor expansion $\bar{f}(y;\bar{t}) := \sum_{j=0}^{k} f^{(j)}(\bar{t})(y-\bar{t})^j/j!$. Now

$$|f(y)-\bar{f}(y;\bar{t})| \leqq (y-\bar{t})^k \frac{\omega(f^{(k)}, y-\bar{t})}{k!}, \tag{3.3}$$

where ψ denotes the modulus of continuity. Also

$$\begin{aligned} N^k(f-Q_N f)(t_j+\tau h_j) &= N^k(\bar{f}-Q_N\bar{f})(t_j+\tau h_j)+N^k[(f-\bar{f})-Q_N(f-\bar{f})](t_j+\tau h_j) \\ &= N^k(\bar{f}-Q_N\bar{f})(t_j+\tau h_j)+o(1). \end{aligned} \tag{3.4}$$

This follows, using (3.1), (3.2), (3.3) and the hypotheses on t', as $N\to\infty$.

The rest of the argument follows exactly that of [2, Lemma 1, p. 106] with minor change to the vertically shifted Bernoulli polynomials.

The justification for (3.4) implies the following lemma which we record for later use.

LEMMA 3.4. *Under the hypotheses of Lemma* 3.3, *there exists a constant* K *independent of* N *but dependent upon the bounds on* t' *and dependent upon the bounds of the derivatives of* f *so that*

$$\|Q_N f\|_{p,[a,b]} \leqq K\|f\|_{p,[a,b]}.$$

4. Upper bounds. Using these locally-defined spline approximation operators, we can now obtain upper bounding results which complement the lower bounding results of § 2.

THEOREM 4.1. *Let* $f\in C^k[0,1]$, $t\in C^1[0,1]$, *and* $0<\delta\leqq t'\leqq M\leqq\infty$. *Then for* $1\leqq p<\infty$

$$\overline{\lim_{N\to\infty}} N^k\|f-P_N^{k,p}(t)f\|_{p,[0,1]} \leqq C_{k,p}(J_{k,p}(f,t))^{1/p}.$$

Proof. We can extend f and t to $[0,1+\varepsilon]$ and apply Lemma 3.3 to $[0,1]$ as follows:

$$\begin{aligned} N^{pk}\|f-P_N^{k,p}(t)f\|_{p,[0,1]}^p &\leqq N^{p,k}\|f-Q_N f\|_{p,[0,1]}^p \\ &= N^{pk}\int_0^1 |(f-Q_N f)(y)|^p\,dy \\ &= \sum_{i=0}^{N-1}\int_0^1 |R_N(\tau;t(\xi_i))|^p h_i\,d\tau \quad \text{for any } \frac{i}{N}\leqq\xi_i<\frac{i+1}{N}. \end{aligned}$$

Now choose ξ_i so that $t'(\xi_i)=N[t((i+1)/N)-t(i/N)]=Nh_i$. By Lemma 3.3

$$\sum_{i=0}^{N-1}\int_0^1 |R_N(\tau;t(\xi_i))|^p h_i\,d\tau = \frac{1}{N}\sum_{i=0}^{N-1}(t'(\xi_i))\int_0^1 \left|(t'(\xi_i))^k f^{(k)}(t(\xi_i))\frac{B_{k,p}(\tau)}{k!}-\varepsilon_{i,N}(\tau)\right|^p d\tau,$$

where $|\varepsilon_{i,N}(\tau)|<\varepsilon_N$. Further, since ε_N tends to zero as $N\to\infty$, this equals

$$\frac{1}{N}\sum_{i=0}^{N-1}(t'(\xi_i))^{kp+1}|f^{(k)}(t(\xi_i))|^p\int_0^1\left(\frac{B_{k,p}(\tau)}{k!}\right)^p d\tau+o(1).$$

The summation above is a Riemann sum for the pth power of the desired expression, i.e.,

$$C_{k,p}^p J_{k,p}(f,t) = C_{k,p}^p\int_0^1 |f^{(k)}(t(x))|^p (t'(x))^{kp+1}\,dx.$$

THEOREM 4.2. *Let* $I := [a, b] \subseteq [0, 1]$ *and* $J := [a-\eta, b+\eta] \cap [0, 1]$ *where* $\eta > 0$. *Define* $I^* := t(I)$ *and* $J^* := t(J)$. *Suppose that* $f \in C^k(J^*)$ *and* $t \in C^1(J)$ *with* $0 < \delta \leqq t' \leqq M < \infty$ *on* J. *Then for* $1 \leqq p < \infty$

$$\overline{\lim_{N\to\infty}} N^{kp} \|f - P_N^{k,p}(t) f\|_{p,I^*}^p \leqq C_{k,p}^p \int_I |f^{(k)}(t(x))|^p (t'(x))^{kp+1}\, dx.$$

Proof. This is a generalization of Theorem 4.1. The only modification in the proof given above is that the summations now should only run from $i_1 = \max\{i: (i/N) \leqq a\}$ to $i_2 = \min\{i: ((i+1)/N) \geqq b\}$. For N sufficiently large, the interval $[i_1/N, (i_2+1)/N] \subset J$ so that Lemma 3.3 applies.

THEOREM 4.3. *Suppose* $f \in C^k(0, 1) \cap L_p[0, 1]$, $|f^{(k)}|$ *is monotone near* 0 *and near* 1, $t \in C[0, 1]$ *with* t' *bounded and piecewise continuous, the set* $\bar{\mathbf{X}}$ *consisting of the points where* t' *is either zero or discontinuous has content zero, and neither* 0 *nor* 1 *is an accumulation point for* $\bar{\mathbf{X}}$. *Then for* $1 \leqq p < \infty$

$$\overline{\lim_{N\to\infty}} N^k \|f - P_N^{k,p}(t) f\|_{p,[0,1]} \leqq C_{k,p}(f, t))^{1/p}.$$

Proof. There is nothing to prove when $J_{k,p}(f, t)$ is infinite, so assume that it is finite. Let $\varepsilon > 0$ be given. Then there exist a finite number of relatively open subintervals $I_0, I_1, \cdots, I_M$ such that $I_0 = [0, \delta_0')$, $I_M = (1-\delta_1', 1]$, the sum of the lengths of all of the subintervals is less than ε, and $\bar{\mathbf{X}}$ is contained in their union. Let $J_0, J_1, \cdots, J_M$ denote the disjoint subintervals in the union of the sets

$$\{x: \exists y \in I_i \text{ with } |x-y| \leqq \eta\}, \qquad i = 1, \cdots, M,$$

where $\eta = \varepsilon/(2M)$. Thus the sum of the lengths of $J_0, \cdots, J_M$ will be less than 2ε. Let $J_i^* := t(J_i)$, $i = 0, 1, \cdots, M$, and $\delta_i^* := t(\delta_i) := t(\delta_i' + \eta)$, $i = 0, 1$.

We now define operators $\tilde{Q}_N$ which are mixtures of the operators $\not{s}_N$ and Q_n.

$$\tilde{Q}_N f := \sum \tilde{\lambda}_l(f) N_l,$$

where $\tilde{\lambda}_l = \lambda_l$ if the support $N_l = [t_l, t_{l+k}]$ intersects the complement of the union of $J_0^*, J_1^*, \cdots, J_M^*$ and where $\tilde{\lambda}_l = \mu_l$ otherwise. For N sufficiently large, $\tilde{Q}_N$ is well defined and $\tilde{Q}_N f$ agrees with $Q_N f$ on $(\cup J_i^*)^c$. The arguments used for Theorem 4.2 go through unchanged once we note that t' is continuous and bounded away from zero in $(\cup I_i)^c$. Thus

$$\begin{aligned}\overline{\lim_{N\to\infty}} N^{kp} \|f - P_N^{k,p} f\|_{p,[0,1]}^p &\leqq \overline{\lim_{N\to\infty}} N^{kp} \{\|f - \tilde{Q}_N f\|_{p,(\cup J_i^*)^c}^p + \|f - \tilde{Q}_N f\|_{p,(\cup J_i^*)}^p\} \\ &= C_{k,p}^p \int_{(\cup J_i)^c} |f^{(k)}(t(x))|^p t'(x)^{kp+1}\, dx + \overline{\lim_{N\to\infty}} N^{kp} \|f - \tilde{Q}_N f\|_{p,(\cup J_i^*)}^p. \end{aligned} \tag{4.1}$$

We now show that the last term in (4.1) can be made arbitrarily small with $\cup J_i^*$. Consider $J_0^* = [0, \delta_0^*]$. Let $L_N = \lceil N\delta_0 \rceil$, the number of knots located in J_0^*. Define p_{j+k} to be the Taylor polynomial for f of degree $(k-1)$ expanded about $t_{j+k} = t((j+k)/N)$. Then

$$\begin{aligned}\|f - \tilde{Q}_N f\|_{p,J_0^*}^p &\leqq \sum_{j=0}^{L_N} \|f - \tilde{Q}_N f\|_{p,[t_j, t_{j+1}]}^p \leqq \sum_{j=0}^{L_N} \|(f - p_{j+k}) - \tilde{Q}_N (f - p_{j+k})\|_{p,[t_j, t_{j+1}]}^p \\ &\leqq K_1 \sum_{j=0}^{L_N} \|f - p_{j+k}\|_{p,[t_{j-k+1}, t_{j+k}]}^p \end{aligned} \tag{4.2}$$

where the positive constant K_1 is guaranteed by the Lemma 3.2 and the fact that $\not{s}_N$

and Q_N are both locally defined bounded operators. Now

$$\|f-p_{j+k}\|^p_{p,[t_{j-k+1},t_{j+k}]}=\left[\frac{1}{(k-1)!}\right]^p\int_{t_{j-k+1}}^{t_{j+k}}\left|\int_x^{t_{j+k}}f^{(k)}(\tau)(x-\tau)^{k-1}\,d\tau\right|^p dx \tag{4.3}$$
$$\leqq\frac{1}{(k!)^p(kp+1)}\left[\int_{t_{j-k+1}}^{t_{j+k}}|f^{(k)}(\tau)|^{p/(kp+1)}\,d\tau\right]^{kp+1}$$

using the monoticity of $f^{(k)}$ (for sufficiently small ε) and a lemma of de Boor and Dodson (see [4], [5] or [8, p. 292]). We continue (4.3) by changing variables and then using Hölder's inequality.

$$\|f-p_{j+k}\|^p_{p,[t_{j-k+1},t_{j+k}]}$$
$$\leqq\frac{1}{(k!)^p(kp+1)}\left[\int_{\max[0,(j-k+1)/N]}^{(j+k)/N}|f^{(k)}(t(x))|^{p/(kp+1)}t'(x)\,dx\right]^{kp+1} \tag{4.4}$$
$$\leqq\frac{1}{(k!)^p(kp+1)}\left[\int_{\max[0,(j-k+1)/N]}^{(j+k)/N}|f^{(k)}(t(x))|^p(t'(x))^{kp+1}\,dx\right]\left[\frac{2k-1}{N}\right]^{kp}.$$

Combining (4.2) and (4.4), we have

$$N^{kp}\|f-\tilde{Q}_Nf\|^p_{p,J_0^*}\leqq K_2\int_{J_{0,N}}|f^{(k)}(t(x))|^p(t'(x))^{kp+1}\,dx, \tag{4.5}$$

where $J_{0,N}=[0,(L_N+k)/N]$.

A symmetric argument will show that (4.5) holds with J_0, J_0^* and $J_{0,N}$ replaced by J_M, J_M^* and $J_{M,N}$. Now taking the limit superior as $N\to\infty$ for $i=0,M$

$$\overline{\lim_{N\to\infty}}\,N^{kp}\|f-\tilde{Q}_Nf\|^p_{p,J_i^*}\leqq K_2\int_{J_i}|f^{(k)}(t(x))|^p(t'(x))^{kp+1}\,dx. \tag{4.6}$$

Consider an interior interval J_i, $0<i<M$. Let $F_i:=\max\{f^{(k)}(x):x\in J\}$. Then, as in (4.2),

$$\|f-\tilde{Q}_Nf\|^p_{p,J_i^*}\leqq K_1\sum_j\|f-p_{j+k}\|^p_{p,[t_{j-k+1},t_{j+k}]}\leqq K_1\sum_j F_i^p(k!)^{-p}\int_{t_{j-k+1}}^{t_{j+k}}|(x-t_{j+k})|^{kp}\,dx$$
$$\leqq K_1F_i^p[(k!)^p(kp+1)]^{-1}\sum_j(t_{j+k}-t_{j-k+1})^{kp+1} \tag{4.7}$$
$$\leqq K_3F_i^pN^{-kp}\left\{|J_i|+\frac{2k}{N}\right\}^{kp+1}.$$

The positive constant K_3 involves the upper bound on t', and the summations above run over the index set $\{j:t_j\in J_i\}$. Letting $N\to\infty$ in (4.7), we get

$$\overline{\lim_{N\to\infty}}\,N^{kp}\|f-\tilde{Q}_Nf\|^p_{p,J_i}\leqq K_3F_i^p|J_i|^{kp+1}. \tag{4.8}$$

Finally if we sum (4.8) for $0<i<M$ and add this to (4.6) for $i=0,M$, we get an upper bound which tends to the pth power of the desired result as $\varepsilon\to 0$.

Remark. The hypothesis on f in Theorem 4.3 can be weakened slightly. A close examination of the proof together with standard density arguments yield that it will suffice to require $f\in W^{k,p}_{\text{loc}}(0,1)$ with $|f^{(k)}|$ having a monotone majorant g near 0 and

near 1, where

$$\int_J |g(t(x))|^p (t'(x))^{kp+1}\, dx < \infty,$$

for some $J = [0, \delta] \cup [1-\delta, 1]$, $\delta > 0$.

This result we record in Theorem 4.4.

THEOREM 4.4. *Suppose* $f \in W^{k,p}_{\text{loc}}(0, 1)$ *and that there is a* $\delta > 0$ *and* g *which is monotone on* $J = (0, \delta] \cup [1-\delta, 1)$ *satisfying*

$$g(x) \geqq |f^{(k)}(x)|, \qquad \int_J |g(t(x))|^p t'(x)^{kp+1}\, dx < \infty.$$

In addition, we assume that $t \in C[0, 1]$ *with* t' *bounded and piecewise continuous, that the set* $\bar{\mathbf{X}}$ *consisting of points where* t' *is either zero or discontinuous has content zero, and that neither* 0 *or* 1 *is an accumulation point for* $\bar{\mathbf{X}}$. *Then for* $1 \leqq p < \infty$

$$\lim_{N\to\infty} N^k \|f - P_N^{k,p}(t) f\|_{p,[0,1]} = C_{k,p}(J_{k,p}(f, t))^{1/p}.$$

5. Best approximation with variable knots. Denote by S_N^k the closure of the set of all kth order splines having at most $N-1$ knots interior to $[0, 1]$ counting multiplicities. The variable knot problem then is: Given $f \in L_p[0, 1]$, find $s_N \in S_N^k$, called the best $L_p[0, 1]$ approximation to f such that

$$\|f - s_N\|_{p,[0,1]} = \inf_{s \in S_N^k} \|f - s\|_{p,[0,1]} =: \operatorname{dist}_p (f, S_N^k).$$

Existence of such a best approximation is assured because S_N^k is a closed set in $L_p[0, 1]$. Unicity, however, is a more difficult question which has not been completely answered.

In this section, we consider the implications of the upper bounds in § 4 to the quantity $\operatorname{dist}_p (f, S_N^k)$ as $N \to \infty$. Certainly, given N and f, the optimal location of $N-1$ interior knots for s_N must do better than or equal what would be achieved using any quantile function t. Thus:

LEMMA 5.1. *If* f *and* t *satisfy the hypotheses of Theorem* 4.3, *then for* $1 \leqq p < \infty$

$$\overline{\lim_{N\to\infty}} N^k \operatorname{dist}_p (f, S_N^k) \leqq C_{k,p}(J_{k,p}(f, t))^{1/p}. \tag{5.1}$$

Using calculus of variations, we can minimize the right-hand side of (5.1) over the class of all knot quantile functions. This turns out to be easier to do using the associated knot density function u, where u is related to t by $u(\tau) = (t^{-1})'(\tau)$. Thus $\int_0^1 u(\tau)\, d\tau = 1$ and $u(t(x)) > 0$ when $t'(x)$ is defined and bounded. Minimizing over the closure of this class of knot density functions, we find that the minimum of

$$\bar{J}(u) := \int_0^1 |f^{(k)}(\tau)|^p (u(\tau))^{-kp}\, d\tau = J_{k,p}(f, t)$$

is attained uniquely by the function $\bar{u}$, where $\sigma = p/(kp+1)$ and

$$\bar{u}(\tau) = |f^{(k)}(\tau)|^\sigma \Big/ \int_0^1 |f^{(k)}(\xi)|^\sigma\, d\xi.$$

The minimum value is

$$\bar{J}(\bar{u}) = \|f^{(k)}\|^p_{\sigma,[0,1]},$$

thus furnishing the proof of the following theorem.

THEOREM 5.1. *Suppose* $f \in C^k(0,1) \cup L_p[0,1]$ *and* $|f^{(k)}|$ *is monotone near* 0 *and near* 1. *Then for* $1 \leqq p < \infty$ *and* $\sigma = p/(kp+1)$

$$\overline{\lim_{N\to\infty}} N^k \operatorname{dist}_p (f, S_N^k) \leqq C_{k,p} \|f^{(k)}\|_{\sigma,[0,1]}. \tag{5.2}$$

It is interesting to note that when $|f^{(k)}|$ is never zero so that $\bar{u}$ is associated with $\bar{t} \in C^1(0,1)$ with $\bar{t}'$ bounded, then the integrand in

$$J_{k,p}(f, \bar{t}) = \int_0^1 |f^{(k)}(\bar{t}(x))|^p (\bar{t}'(x))^{kp+1}\, dx$$

is constant. Thus

$$J_{k,p}(f, \bar{t}) = \sum_{i=0}^{N-1} \int_{i/N}^{(i+1)/N} |f^{(k)}(\bar{t}(x))|^p (\bar{t}'(x))^{kp+1}\, dx = \|f^{(k)}\|_{\sigma,[0,1]}^{p-\sigma} \sum_{i=0}^{N-1} \int_{\bar{t}_i}^{\bar{t}_{i+1}} |f^{(k)}(\tau)|^\sigma\, d\tau,$$

where all of the terms in the last summation are equal. This can be interpreted as evidence supporting the notion, at least for large N, that "good" knot sequences are ones which "balance the error" in $\|f^{(k)}\|_{\sigma,[0,1]}^{\sigma}$ (see [4], [5]).

It is possible that (5.2) may not be sharp for a particular function f. Optimizing the knot locations for each N and then letting N tend to infinity may produce a lower limit than using the optimal knot quantile function $\bar{t}$ for locating simple knots. The following theorem describes when equality can be guaranteed in (5.2).

THEOREM 5.2. *Suppose* $f \in C^k[0,1]$. *Then*

$$\lim_{N\to\infty} N^k \operatorname{dist}_p (f, S_N^k) = C_{k,p} \|f^{(k)}\|_{\sigma,[0,1]},$$

where $\sigma = p/(kp+1)$ *and* $1 \leqq p < \infty$.

Proof. Appealing to Theorem 5.1, it suffices to prove that

$$\underline{\lim_{N\to\infty}} N^k \operatorname{dist}_p (f, S_N^k) \geqq C_{k,p} \|f^{(k)}\|_{\sigma,[0,1]}. \tag{5.3}$$

The proof follows closely the proof of [1, Thm. 2, pp. 300–302]. We look at special cases first.

Case (i). Let $f(x) = x^k$. We can replace f by $B_k(\cdot)$, the kth Bernoulli polynomial, without changing the distance because these two functions differ by something in S_N^k. Letting $\mathring{S}_{N+2k}^k$ denote the set of one-periodic splines of order k with at most $N+2k$ knots counting multiplicities, we note that S_N^k is a subset of $\mathring{S}_{N+2k}^k$. Theorem 2.1 provides the exact distance form $B_k(\cdot)$ to $\mathring{S}_{N+2k}^k$. Thus

$$\begin{aligned} N^k \operatorname{dist}_p (f, S_N^k) &= N^k \operatorname{dist}_p (B_k(\cdot), S_N^k) \geqq N^k \operatorname{dist}_p (B_k(\cdot), \mathring{S}_{N+2k}^k) \\ &= \left(\frac{N}{N+2k}\right)^k C_{k,p}^* = \left(\frac{N}{N+2k}\right)^k C_{k,p} \|f^{(k)}\|_{\sigma,[0,1]}. \end{aligned}$$

Taking the limit as $N \to \infty$ establishes (5.3) for $f(x) = x^k$.

Case (ii). Let $f \in C^k[0,1]$, $|f^{(k)}| \geqq \delta > 0$. First, suppose that (5.3) were false in this case. Then for some infinite subset of the positive integers, Z_1,

$$N^k \operatorname{dist}_p (f, S_N^k) =: N^k \|f - f_N\|_{p,[0,1]} < d \|f^{(k)}\|_{\sigma,[0,1]}, \tag{5.4}$$

where $0 < d < C_{k,p}$ and $N \in Z_1$. For each $m = 1, 2, \cdots$, and for N sufficiently large and in Z_1, subdivide the interval $[0,1]$ into finitely many closed subintervals

$\{I_r := [\alpha_r, \alpha_{r+1}]\}$ whose endpoints coincide with the knots of the optimal f_N, in such a way that each I_r contains in its interior $(m_r - 1)$ knots of f_N, where $m \leqq m_r \leqq m + k + 1$. Thus $\sum m_r =: \bar{N} < N$ with equality in case each α_r is a simple knot of f_N. The inequality (5.4) implies that

$$N^{k\sigma}\left(\int_0^1 |f - f_N|^p\right)^{1/(kp+1)} < d^\sigma \int_0^1 |f^{(k)}|^\sigma. \tag{5.5}$$

Secondly, suppose that

$$d^\sigma \int_{I_r} |f^{(k)}|^\sigma \leqq m_r^{k\sigma}\left(\int_{I_r} |f - f_N|^p\right)^{1/(kp+1)} \tag{5.6}$$

for all r. If so, then summing both sides over r and applying Hölder's inequality for finite sequences yields

$$\begin{aligned} d^\sigma \int_0^1 |f^{(k)}|^\sigma &\leqq \sum_r m_r^{k\sigma}\left(\int_{I_r} |f - f_N|^p\right)^{1/(kp+1)} \leqq \left\{\sum_r m_r\right\}^{k\sigma}\left\{\sum_r \int_{I_r} |f - f_N|^p\right\}^{1/(kp+1)} \\ &= \bar{N}^{k\sigma}\left(\int_0^1 |f - f_N|^p\right)^{1/(kp+1)}. \end{aligned} \tag{5.7}$$

However, (5.7) contradicts (5.5). Thus the second supposition (5.6) must be false for some $r = r_N$, or

$$m_{r_N}^k\left(\int_{I_{r_N}} |f - f_N|^p\right)^{1/p} < d\left(\int_{I_{r_N}} |f^{(k)}|^\sigma\right)^{1/\sigma}. \tag{5.8}$$

Since each m_{r_N} satisfies $m \leqq m_r \leqq m + k - 1$, some fixed integer $\bar{m}$ must be chosen by an infinite set $Z_2 \subset Z_1$. Let a be an accumulation point of the set of left endpoints of I_{r_N}, $N \in Z_2$. We pass again to an infinite subset Z_3, where a will be a limit point of appropriate left endpoints. Since $|f^{(k)}| \geqq \delta > 0$, the length of I_{r_N} must go to zero as N from Z_3 goes to infinity. Thus the right-hand side of (5.8) tends to $d|f^{(k)}(a)|$. Now divide both sides of (5.8) by $|f^{(k)}(a)|$, change variable from I_{r_N} to $[0, 1]$ on the left-hand side, and let N from Z_3 tend to infinity. We get

$$\bar{m}^k\left(\int_0^1 \left|\frac{x^k}{k!} - s_*(x)\right|^p dx\right)^{1/p} \leqq d, \tag{5.9}$$

where $s_* \in S_{\bar{m}}^k$. But (5.9) provides an upper bound on $\operatorname{dist}_p(x^k/k!, S_{\bar{m}}^k)$ which contradicts what we established in Case (i). Thus (5.3) must be true for the functions considered in this case.

Case (iii). Let $f \in C^k[0, 1]$ and $A = \{x \in [0, 1]: f^{(k)}(x) = 0\}$. Suppose (5.3) is false. Then for some infinite subset of the positive integers, Z_1,

$$N^{kp}\int_0^1 |f - f_N|^p < d^p\left(\int_{[0,1]\setminus A} |f^{(k)}|^\sigma\right)^{kp+1} \leqq \bar{d}^p\left(\sum_j \int_{I_j} |f^{(k)}|^\sigma\right)^{kp+1}$$

where $0 < d < \bar{d} < C_{k,p}$ and $\{I_j\}$ are the closed intervals where $|f^{(k)}| \geqq \delta > 0$ for some δ. Let M_j denote the number of knots of f_N interior to I_j so that

$$\sum M_j =: M \leqq N - 1 < N.$$

By an argument analogous to that given in Case (ii), roughly from equation (5.6) to

(5.8), we can conclude that for some index j,

$$M_j^{kp} \int_{I_j} |f - f_N|^p < \bar{d}^p \left(\int_{I_j} |f^{(k)}|^\sigma \right)^{kp+1}$$

We pass to an infinite subset $Z_2 \subset Z_1$ where the same index j is chosen for all $N \in Z_2$. Since $|f^{(k)}| \geqq \delta > 0$, M_j tends to infinity as $N \in Z_2$ tends to infinity. Now the argument given in Case (ii) can be carried out, simply by replacing N with M_j and replacing $[0, 1]$ with I_j. It leads to a contradiction analogous to (5.9).

Remark. One can also establish that the optimal knot density function $\bar{u}$ describes asymptotically where the knots will be located in the optimal f_N, as $N \to \infty$. This was done explicitly for $p = 2$ in [1, p. 302].

We conclude with a conjecture which is motivated by a comparison of the results of this paper with similar results for piecewise polynomials by Burchard and Hale [5] (see also [8, p. 295]). Let $S^k_{N,\nu}$ denote the closure of the space of polynomial splines of order k with $N-1$ distinct knots each of multiplicity ν, where ν is a fixed integer between 1 and k. Then when $\nu = k$, $S^k_{N,k}$ is the space of piecewise polynomials with no continuity between pieces. For $f \in C^k[0, 1]$, the conjecture is that

$$\lim_{N \to \infty} N^k \operatorname{dist}_p (f, S^k_{N,\nu}) = C_{k,p,\nu} \|f^{(k)}\|_{\sigma,[0,1]}, \tag{5.10}$$

where

$$C_{k,p,\nu} = \frac{1}{k!} \min_{a} \|B_k(x) - \{a_0 + a_1 x + \cdots + a_{\nu-1} x^{\nu-1}\}\|_{p,[0,1]}.$$

This paper establishes the conjecture for $\nu = 1$ and [5] establishes it for $\nu = k$. Thus the cases $\nu = 2, \cdots, k-1$ remain open.

REFERENCES

[1] D. L. BARROW AND P. W. SMITH, *Asymptotic properties of best $L_2[0, 1]$ approximation by splines with variable knots*, Quart. Appl. Math., 36 (1978), pp. 293–304.

[2] ———, *Efficient L_2 approximation by splines*, Numer. Math., 33 (1979), pp. 101–114.

[3] C. DE BOOR, *The quasi-interpolant as a tool in elementary polynomial spline theory*, in Approximation Theory, G. G. Lorentz, ed., Academic Press, New York, 1973, pp. 269–276.

[4] ———, *Good approximation by splines with variable knots*, in Spline Functions and Approximation Theory, A. Meir and A. Sharma, eds., Birkhouser, Basel, 1972, pp. 57–72.

[5] H. G. BURCHARD AND D. F. HALE, *Piecewise polynomial approximation on optimal meshes*, J. Approx. Theory, 14 (1975), pp. 128–147.

[6] M. GOLOMB, *Approximation by periodic spline interpolants on uniform meshes*, J. Approx. Theory, 1 (1968), pp. 62–65.

[7] I. J. SCHOENBERG, *Monosplines and quadrature formulae*, in Theory and Application of Spline Functions, T. N. E. Greville, ed., Academic Press, New York, 1969, pp. 157–207.

[8] L. L. SCHUMAKER, *Spline Functions: Basic Theory*, John Wiley, New York, 1981.

[9] A. A. ZHENSYKBAEV, *On the best quadrature formula on the class $W^r L_p$*, Soviet Math. Dokl., 17 (1976), pp. 377–380.

SIAM J. MATH. ANAL.
Vol. 13, No. 3, May 1982

0036-1410/82/1303-0005 $01.00/0

EQUILIBRIA OF THE CURVATURE FUNCTIONAL AND MANIFOLDS OF NONLINEAR INTERPOLATING SPLINE CURVES*

MICHAEL GOLOMB† AND JOSEPH JEROME‡

Abstract. A detailed global and local analysis is carried out of smooth solutions of the variational problem

$$\delta \int_0^s \kappa^2(s)\, ds = 0, \tag{1i}$$

subject to position function constraints

$$x(s_i) = p_i, \qquad 0 \leqq s_0 < s_1 < \cdots < s_m \leqq \bar{s}. \tag{1ii}$$

Here $\{p_i\}_0^m \subset \mathbb{R}^2$ is prescribed, x is a vector-valued function with curvature $\kappa(s)$ at arc length s and the interpolation nodes s_i are free. Problem (1) may be viewed as the mathematical formulation of the draftsman's technique of curve fitting by mechanical splines.

Although most of the basic equations satisfied by these nonlinear spline curves have been known for a very long time, calculation via elliptic integral functions has been hampered by a lack of understanding concerning what precise information must be specified for the stable determination of a smooth, unique interpolant modeling the thin elastic beam. In this report, sharp characterizations are derived for the extremal interpolants as well as structure theorems in terms of inflection point modes which guarantee uniqueness and well-posedness.

A certain type of stability is introduced and studied and shown to be related to (linearization) concepts associated with piecewise cubic spline functions, which have been studied for decades as a simplification of the nonlinear spline curves. Many examples are introduced and studied.

1. Introduction. Let $P = \{p_0, p_1, \cdots, p_m\}$ be an ordered set of points in the Euclidean plane (the p_i need not be distinct) and let it be required to pass a smooth curve through these points in the prescribed order. It is an old technique of draftsmen to use a mechanical spline to accomplish this. If the spline is considered as a thin elastic beam of uniform cross section with a central fiber that is inextensible, then the strain energy of the bent spline of length $\bar{s}$ is given by

$$A \int_0^{\bar{s}} \kappa^2(s)\, ds + B$$

where $\kappa(s)$ is the curvature of the fiber at arc length s and A, B are constants. An equilibrium position of the spline makes the energy functional stationary, hence satisfies

$$\delta \int_0^{\bar{s}} \kappa^2(s)\, ds = 0. \tag{1.1i}$$

This equation together with the interpolation conditions

$$x(s_i) = p_i, \qquad 0 \leqq s_0 < s_1 < \cdots < s_m \leqq \bar{s} \tag{1.1ii}$$

* Received by the editors October 30, 1980 and in revised form March 9, 1981. This research was sponsored by the U.S. Army under contract DAAG29-75-C-0024. This work was supported by the National Science Foundation under grant MPS 74-02292 A01.

† Division of Mathematical Sciences, Purdue University, West Lafayette, Indiana 47907.

‡ Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

for the position function x, constitute the mathematical formulation of the draftsman's technique. The present article deals with analytical (not graphical nor computational) problems arising from system (1.1). In elasticity theory the solutions of (1.1i) or of the more general equation

$$\delta\left(\int_0^{\bar{s}} [\lambda + \kappa^2(s)]\, ds\right) = 0, \tag{1.2}$$

where λ is a constant, are known as *elastica*. Their study dates back to the Bernoulli brothers, Euler and others (see Love [8, Ch. XIX] for classical results). The boundary conditions in traditional elasticity theory have little in common with the interpolation conditions (1.1ii). To materialize the latter ones in the beam model one may think of freely rotating small sleeves, anchored at the points $p_0, \cdots, p_m$, through which the spline can slide without friction. We refer to the solutions $x = \bar{x}$ of the variational problem (1.1i, ii) as *extremal P-interpolants*. In some parts of the present paper we deal with extremal *length-prescribed* P-interpolants, in which $\bar{s} = s_m - s_0$ is given in addition to P. To materialize this condition one replaces the sleeves at p_0, p_m by pins which allow no sliding. In other parts we consider extremal *angle-prescribed* P-interpolants, in which the angles that the spline makes at p_0 and p_m with a reference line are given. This situation prevails if the sleeves at p_0 and p_m are not allowed to rotate.

Stable equilibrium positions in mechanics are sought as positions that minimize the potential energy functional. In one of the earliest papers discussing nonlinear interpolating splines [2] it was pointed out that the infimum of $\int_0^{\bar{s}} \kappa^2(s)\, ds$ is 0 for any configuration P, hence can be attained only in the trivial case where P is interpolated by a straight segment. Lee and Forsythe [7], who make a substantial study of the variational problem (1.1i, ii), call the solutions (when existence is hypothesized) local minima. However, it will be proved in § 6 below that, in the simple case where P consists of 2 points, there are countably many nontrivial extremal P-interpolants; none of them constitutes a local minimum of the energy. This makes it evident that an extremal P-interpolant is, in general, not a local minimum (for a detailed discussion of the stability problem see [5]).

The existence questions for interpolating elastica are much more subtle. For length-constrained or length-prescribed interpolants one can prove existence of extremals (actually global minima) by the direct methods of the calculus of variations, because one has compactness in a suitably chosen function space (this was done in [3] and [6]; see also the Appendix of this paper). This is not the case for interpolants with no length restriction, and the existence of such extremals interpolating n points in general position, and whether they are local minima or not, remains an open question (some progress along these lines has been achieved by M. Golomb [4], [5]). Computational work on extremal interpolants is more advanced (cf. M. Malcolm [10]), although decisive progress in this area is also hampered by the lack of general existence and uniqueness theorems.

We now give a brief account of the content of this paper. In § 2 we define the function classes in which the extremal interpolants are sought. We also characterize them by Euler equations (for the Cartesian coordinates), boundary and regularity conditions. In § 3 we do the same for the "normal representation" of the extremals, by which we mean the function $s \to \theta(s)$, which is the angle that the extremal makes at arc length s with a reference line.

The normal representation θ of a length-prescribed extremal P-interpolant appears as the solution of a free multi-point boundary value problem for $s_1, \cdots, s_{m-1}, \theta$

with $s_m - s_0$ prescribed:

$$\text{(i)} \quad \ddot{\theta}(s) - \mu_i^1 \sin\theta(s) + \mu_i^2 \cos\theta(s) = 0, \qquad s_{i-1} < s < s_i,$$

$$\text{(1.3)} \quad \text{(ii)} \quad \dot{\theta}(s_0) = \dot{\theta}(s_m) = 0,$$

$$\text{(iii)} \quad (\mu_{i+1}^1 - \mu_i^1)\cos\theta(s_i) + (\mu_{i+1}^2 - \mu_i^2)\sin\theta(s_i) = 0, \qquad i = 1, \cdots, m-1.$$

For the general extremal P-interpolant, (1.3iii) holds also for $i = m$ and s_0 and s_m are free as well with $\mu_{m+1}^1 = \mu_{m+1}^2 = 0$. The function θ and the knot abscissas s_i are the unknowns; the multipliers μ_i^1, μ_i^2 are determined from the interpolation conditions. In § 4 it is shown how certain families $\mathscr{E}_{(k_1,\cdots,k_m)}$ of extremal interpolants with prescribed numbers $(k_1, \cdots, k_m)$ of inflection points between knots ("mode"), can be realized as smooth $2m$-dimensional manifolds $\mathscr{M}_{(k_1,\cdots,k_m)}$. The inverse mapping from $\mathscr{M}_{(k_1,\cdots,k_m)}$ into the position function space of the extremals is continuous when the latter is topologized by a suitable metric. Thus, the mode $(k_1, \cdots, k_m)$ of an extremal suitably delineates uniqueness and well-posedness. The union of the $\mathscr{E}_{(k_1,\cdots,k_m)}$ consists of only those extremals which have a genuine knot and no inflection point at each interior interpolation node. Points in the intersection of the boundaries of the manifolds $\mathscr{M}_{(k_1,\cdots,k_m)}$ correspond to singular points in position function space. For these boundary elements some interior knot is spurious (κ is not discontinuous) or is an inflection point. We give two examples to demonstrate this.

In § 5 we study the existence of elastica spline interpolation in the small. Does the set of configurations P for which extremal interpolants exist have nonempty interior in $\mathbb{R}^{2m}$? More specifically, which P in $\mathbb{R}^{2m}$ are interior points of this set? We show, by use of the implicit function theorem that near a given configuration $\bar{P}$ with extremal interpolant $\bar{E}$ there is a local diffeomorphism between configurations and extremal interpolants if $(\bar{E}, \bar{P})$ satisfies a certain hypothesis (A). It requires that a homogeneous linear differential equation with variable coefficients depending on $\bar{E}$ and with homogeneous linear side conditions has no nontrivial solution. Another formulation of this condition is that a computable function (involving many quadratures) be $\neq 0$ at the end point of $\bar{E}$. It is easily verified that the ray configuration P_0, with the trivial extremal interpolant E_0, satisfies (A), so that the existence of extremal P-interpolants for all configurations P in some Euclidean neighborhood of any ray configuration is thereby demonstrated. The differential equation problem of hypothesis (A) reduces to the natural cubic spline interpolation problem in the case (P_0, E_0). This demonstrates that cubic spline interpolation can be interpreted as the result of linearization of extremal interpolation (in the sense of making $\int \kappa^2\, ds$ stationary) near the trivial interpolant for the ray configuration. This proof makes precise the old idea that cubic splines are in some sense the "smoothest" interpolants. Of course, it has long been known that cubic spline functions arise from minimizing the quadratic functional $\int (D^2 f)^2$ among the interpolating functions f. Since the linear operator D^2 supposedly approximates the nonlinear curvature operator, the cubic splines recommend themselves as near optimally smooth interpolants. The "hairpin" configuration $\tilde{P}$ with a loop interpolant $\tilde{E}$ is given as an example where hypothesis (A) is not satisfied. There are configurations close to $\tilde{P}$ for which there exists no extremal interpolant near $\tilde{E}$ and there are other configurations close to $\tilde{P}$ for which there do exist extremal interpolants near $\tilde{E}$. This seems to be the first known example demonstrating singular behavior in nonlinear spline interpolation.

Section 6 contains an exhaustive study of extremal P-interpolants for the case where P consists of two points. It is shown that there exist, besides the trivial extremal, countably many nontrivial ones of distinct integral mode, that all of them are obtained

by simple transformations from a basic one, all have the same length and (cf. [5]) none makes the potential energy a local minimum. Composition of these 2-point extremals yields countably many extremal P-interpolants for various special configurations P. Section 6 also exhibits countably many angle-prescribed and countably many length-prescribed 2-point extremals.

In § 7 some special cases of closed extremal P-interpolants are considered. It is shown that the only closed length-prescribed extremal without knots are the repeatedly traversed circle and figure eight configurations. Formally, the Euler equation is the limiting case of the Euler equation for an elastic circular ring under hydrostatic pressure p as $p \to 0$ (cf. [1] and [12]). The ring, however, is not an elastica since its deformations satisfy stress-strain relationships. We also consider closed extremals which are not length-prescribed. Here the extremals in § 6 are used to construct infinitely many closed extremals for several special P-configurations, for example where P is the set of vertices of a regular polygon. In particular, if a regular m-gon, $m \geqq 2$, is inscribed in the unit circle, then a circumscribed extremal exists with length

$$s_m = \frac{2m \sin(\pi/m) F(\frac{1}{2}\sqrt{2}; \beta_m)}{\sqrt{2}[2E(\frac{1}{2}\sqrt{2}; \beta_m) - F(\frac{1}{2}\sqrt{2}; \beta_m)]},$$

where $\cos \beta_m = \sqrt{\cos(\pi/m)}$. There are similar formulas for the energy U_m and the arc length $s_m(\theta)$; of interest is the result that $s_m(\theta)/\theta \to 1$ as $m \to \infty$, so that the circumscribed extremals have the unit circle as a limiting configuration. These extremals are stable, i.e., they make the potential energy a local minimum, as proved in [5].

2. Regularity and characterizations of open extremals. For two points $p = (p^1, p^2)$ and $q = (q^1, q^2)$ in real Euclidean space $\mathbb{R}^2$ we employ the inner product $pq = p^1q^1 + p^2q^2$, the distance $|p-q| = [(p-q)(p-q)]^{1/2}$, and the exterior product $[p, q] = p^1q^2 - p^2q^1$ of such points. We consider mappings $x = (x^1, x^2)$ of the unit interval $I = [0, 1]$ to $\mathbb{R}^2$. We denote by $H_2(I)$ the real Hilbert space of those mappings x such that the derivative $\dot{x}$ is absolutely continuous and $\ddot{x} \in L_2(I)$, equipped with the inner product

$$(x, y)_{H_2} = \int_I (xy + \dot{x}\dot{y} + \ddot{x}\ddot{y}). \tag{2.1}$$

We say x is a *regular* element of H_2 if $|\dot{x}(t)| > 0$ for all $t \in I$. We observe that the regular elements of H_2 form an open subset H_2^{reg} of H_2.

For $x \in H_2$ we define the arc length map $s_x : I \to R_+$ by

$$s_x(t) = \int_0^t |\dot{x}|, \qquad t \in I. \tag{2.2}$$

If $x \in H_2^{\text{reg}}$ then s_x has an inverse $s_x^{-1} : [0, \bar{s}] \to [0, 1]$, where $\bar{s} = s_x(1)$, and in this case the function $x \circ s_x^{-1} : [0, \bar{s}] \to \mathbb{R}^2$ has an absolutely continuous derivative, and square-integrable second derivative. We identify x with the oriented curve C in the x^1x^2-plane which has parametric representation $x = x(t)$. Writing $\bar{x} = x \circ s_x^{-1}$, we say that $\bar{x}$ is the *arc length* parametrization of the curve C. Clearly $\bar{x} \in H_2(0, \bar{s})$ and we have:

$$\dot{\bar{x}} = \frac{\dot{x} \circ s_x^{-1}}{|\dot{x} \circ s_x^{-1}|}, \qquad \ddot{\bar{x}} = \frac{\ddot{x} \circ s_x^{-1} |\dot{x} \circ s_x^{-1}|^2 - \dot{x} \circ s_x^{-1} (\dot{x}\ddot{x} \circ s_x^{-1})}{|\dot{x} \circ s_x^{-1}|^4}.$$

If $x \in H_2^{\text{reg}}$, then its *curvature* $\kappa_x : I \to \mathbb{R}$ is defined by

$$\kappa_x(t) = [\dot{x}, \ddot{x}] \circ s_x(t) = [\dot{x}, \ddot{x}]|\dot{x}|^{-3}(t), \qquad t \in I. \tag{2.3}$$

Suppose $x \in H_2^{\text{reg}}$ and $s_x(1) = \bar{s}$. Then we define the *curvature functional*,

$$U(x) = \int_0^{\bar{s}} \kappa_x^2 = \int_0^{\bar{s}} [\dot{\bar{x}}, \ddot{\bar{x}}]^2 = \int_0^{\bar{s}} \ddot{\bar{x}}^2. \tag{2.4}$$

Note that (2.4) defines U as a mapping of H_2^{reg} into $\mathbb{R}_+$. The equivalent expression,

$$U(x) = \int_I [\dot{x}, \ddot{x}]^2 |\dot{x}|^{-5}, \tag{2.5}$$

is independent of the parametrization of x in the following sense. If u is a C^∞-map of I onto itself with $\dot{u} > 0$ and $x = y \circ u$ then

$$U(x) = \int_I [\dot{y}, \ddot{y}]^2 |\dot{y}|^{-5} = U(y).$$

If $x \in H_2^{\text{reg}}$ then U is Fréchet-differentiable at x and, for any increment $y \in H_2$,

$$U'(x)[y] = \int_I \{2[\dot{x}, \ddot{x}]([\dot{x}, \ddot{y}] + [\dot{y}, \ddot{x}])|\dot{x}|^{-5} - 5[\dot{x}, \ddot{x}]^2 \dot{x}\dot{y}|\dot{x}|^{-7}\}. \tag{2.6}$$

If the variable of integration is chosen to be s_x, (2.6) simplifies to

$$U'(x)[y] = \int_0^{\bar{s}} (2\ddot{\bar{x}}\ddot{\bar{y}} - 3\kappa_x^2 \dot{\bar{x}}\dot{\bar{y}})\, ds_x; \qquad \bar{x} = x \circ s_x,\ \bar{y} = y \circ s_x. \tag{2.7}$$

Let $0 = p_0, p_1, \cdots, p_m$ be fixed points in $\mathbb{R}^2$, not necessarily distinct, but $p_{i-1} \neq p_i$, $i = 1, \cdots, m$, and let P denote the ordered set $\{p_0, p_1, \cdots, p_m\}$. We refer to P as a *configuration* in $\mathbb{R}^2$. If $x \in H_2^{\text{reg}}$ is such that $x(t_i) = p_i$ $(i = 0, 1, \cdots, m)$ for some $0 \leqq t_0 < t_1 < \cdots < t_m \leqq 1$ we say the curve x is an *admissible P-interpolant*, with *knots* p_i $(i = 0, \cdots, m)$. The *terminals* $x(0), x(1)$ may or may not be coincident with the terminal knots p_0, p_m. The P-interpolants defined here are to be considered as *open* even if $x(0) = x(1)$. In the physical interpretation $p_i = p_j$ for some $i \neq j$ means that the beam is constrained to pass through two sleeves which are fixed at the same point p_i but can rotate independently of each other. The reason for the condition $p_{i-1} \neq p_i$ will become apparent from Remark 6.2, § 6.

Suppose x is a fixed admissible P-interpolant and $\bar{x} = x \circ s_x^{-1}$ is its arc length parametrization, $s_x(1) = \bar{s}$ its length, $\bar{x}(\bar{s}_i) = p_i$ $(i = 0, 1, \cdots, m)$ its knots. Given any $z \in H_2^{\text{reg}}$ let $\bar{z} = z \circ s_x^{-1}$, be the parametrization of z which uses the arc length of x as the parameter, and assume $\bar{z}(\bar{s}_i) = 0$ $(i = 0, 1, \cdots, m)$. For $|\varepsilon|$ sufficiently small, $x + \varepsilon z$ is an admissible P-interpolant and

$$U(x + \varepsilon z) - U(x) = \varepsilon U'(x)[z] + o(\varepsilon) \quad \text{as } \varepsilon \to 0. \tag{2.8}$$

This justifies the following.

DEFINITION 2.1. The admissible P-interpolant x, with arc length parametrization $\bar{x} = x \circ s_x^{-1}$, knots $p_i = \bar{x}(\bar{s}_i)$ $(i = 0, 1, \cdots, m)$, length $\bar{s} = s_x(1)$, is an *extremal P-interpolant* if

$$U'(x)[z] = 0, \tag{2.9}$$

i.e.,

$$\int_0^{\bar{s}} (2\ddot{\bar{x}}\ddot{\bar{z}} - 3\kappa_x^2 \dot{\bar{x}}\dot{\bar{z}}) = 0, \qquad \bar{z} = z \circ s_x^{-1},$$

for every $z \in H_2^{\text{reg}}$ satisfying $\bar{z}(\bar{s}_i) = 0$ $(i = 0, 1, \cdots, m)$.

The following proposition follows from (2.8) by the usual arguments of the calculus of variations. It helps to explain the interest in extremal P-interpolants.

PROPOSITION 2.1. *Suppose the admissible P-interpolant x minimizes the curvature function U locally, i.e.,*

$$U(x) \leqq U(y)$$

for every admissible P-interpolant y in a neighborhood of x in H_2. Then x is an extremal P-interpolant.

The three major propositions of this section follow. We use the notation y_J for the restriction of a map y to the interval J.

PROPOSITION 2.2. *The admissible P-interpolant x with arc length parametrization $\bar{x}$, knots $p_i = \bar{x}(\bar{s}_i)$, $i = 0, \cdots, m$, and length $\bar{s}$, is extremal if and only if the conditions*

$$\begin{aligned} &\text{(i)} \quad \bar{x} \in C^2[0, \bar{s}], \ddot{\bar{x}}(s) = 0, \quad \textit{for } 0 \leqq s \leqq \bar{s}_0 \textit{ and } \bar{s}_m \leqq s \leqq \bar{s}, \\ &\text{(ii)} \quad (2\dddot{\bar{x}} + 3\kappa_x^2 \dot{\bar{x}})(s) = c_i \in \mathbb{R}^2 \quad \textit{for } s \in (\bar{s}_{i-1}, \bar{s}_i),\ i = 1, \cdots, m \end{aligned} \tag{2.10}$$

hold with $\bar{x}_{(\bar{s}_{i-1}, \bar{s}_i)} \in C^\infty(\bar{s}_{i-1}, \bar{s}_i)$, $i = 1, \cdots, m$.

Remark 2.1. Throughout the paper we use the same symbol to denote regularity classes for both scalar and vector functions.

Proof. The implication (2.10)$\Rightarrow$(2.9) is routine and follows upon decomposing $[0, \bar{s}]$ into subintervals determined by the $\bar{s}_i$, dot-multiplying (2.10ii) by $\dot{z}$, integrating by parts and summing; the continuity of $\ddot{\bar{x}}\ddot{\bar{z}}$, the equations $\bar{z}(\bar{s}_i) = 0$, $i = 0, \cdots, m$ and the equations of (2.10) easily yield $U'(x)[z] = 0$.

Conversely, if (2.9) holds then, selecting $\bar{z} \in C^\infty[0, \bar{s}]$ with support in $[\bar{s}_i, \bar{s}_{i+1}]$, i fixed, we have

$$\int_{\bar{s}_i}^{\bar{s}_{i+1}} (2\ddot{\bar{x}} + F)\ddot{\bar{z}} = 0,$$

where $\dot{F} = 3\kappa_x^2 \dot{\bar{x}}$. By elementary distribution theory, $(2\ddot{\bar{x}} + F)_{(\bar{s}_i, \bar{s}_{i+1})}$ is in $C^\infty(\bar{s}_i, \bar{s}_{i+1})$ and

$$D^2(2\ddot{\bar{x}} + F)_{(\bar{s}_i, \bar{s}_{i+1})} = 0.$$

It follows that

$$(2\dddot{\bar{x}} + 3\kappa_x^2 \dot{\bar{x}})_{(\bar{s}_i, \bar{s}_{i+1})} = c_i$$

and, recursively, $\bar{x}_{(\bar{s}_i, \bar{s}_{i+1})} \in C^\infty(\bar{s}_i, \bar{s}_{i+1})$. To prove the continuity of $\ddot{\bar{x}}$ at an interior knot $\bar{s}_i$, select u in $C^\infty[0, \bar{s}]$ with support in $[\bar{s}_{i-1}, \bar{s}_{i+1}]$ satisfying $u(\bar{s}_i) = 0$, $u'(\bar{s}_i) = 1$ and put $z = (u, 0)$. Then, from (2.9) and integration by parts,

$$0 = \int_{\bar{s}_{i-1}}^{\bar{s}_{i+1}} \{(2\dddot{\bar{x}} + 3\kappa_x^2 \dot{\bar{x}})\dot{z}\} + 2(1, 0)(\ddot{\bar{x}}(\bar{s}_i + 0) - \ddot{\bar{x}}(\bar{s}_i - 0)).$$

Since the first term equals

$$c_i(z(\bar{s}_i) - z(\bar{s}_{i-1})) + c_{i+1}(z(\bar{s}_{i+1}) - z(\bar{s}_i)),$$

which is clearly zero, we conclude that $(\bar{x}^1)_{(\bar{s}_0, \bar{s}_m)}$ is in $C^2(\bar{s}_0, \bar{s}_m)$. A similar argument works for $\bar{x}^2$.

If $\bar{s}_i$ is either $\bar{s}_0 = 0$ or $\bar{s}_m = \bar{s}$, jumps are replaced by one-sided limits and one concludes $\ddot{\bar{x}}(\bar{s}_0 + 0) = \ddot{\bar{x}}(\bar{s}_m - 0) = 0$. Assume now $\bar{s}_m < \bar{s}$. One argues as above that

$$(2\dddot{\bar{x}} + 3\kappa_x^2 \dot{\bar{x}})_{(\bar{s}_m, \bar{s})} = c_m,$$

and that $\ddot{\bar{x}}$ is continuous at $\bar{s}_m$. We show that $c_m = 0$. Indeed, select $u \in C^\infty[0, \bar{s}]$, with $u \equiv 0$ for $0 \leqq s \leqq \bar{s}_m$, satisfying $u(\bar{s}) = 1, \dot{u}(\bar{s}) = 0$, and put $z = (u, 0)$. Then, from (2.9) and integration by parts over $[\bar{s}_m, \bar{s}]$,

$$0 = c_m(z(\bar{s}) - z(\bar{s}_m)) = c_m^1$$

and a similar result holds for c_m^2. Thus $c_m = 0$. By (2.15i) of Proposition 2.4 to follow, we conclude that $\kappa_x^2(s) = 0$ for $\bar{s}_m < s < \bar{s}$. In particular, $\ddot{\bar{x}}(s) = 0$ for $\bar{s}_m < s < \bar{s}$, and, by continuity, for $\bar{s}_m \leqq s \leqq \bar{s}$. A similar proof holds if $0 < \bar{s}_0$. This completes the proof of the proposition.

We introduce the following notation for the jump of the third derivative of the extremal $\bar{x}$ at the knot $\bar{s}_i$

$$\Delta_i \dddot{\bar{x}} = \dddot{\bar{x}}(\bar{s}_i + 0) - \dddot{\bar{x}}(\bar{s}_i - 0), \qquad i = 0, 1, \cdots, m. \tag{2.11}$$

If $\dddot{\bar{x}}(\bar{s}_0 - 0)$ and/or $\dddot{\bar{x}}(\bar{s}_m + 0)$ are not defined, they are to be replaced by 0.

The following proposition gives expressions for the extremal value $U(x)$ which do not involve quadratures.

COROLLARY 2.1. *If $\bar{x}$ is the extremal P-interpolant of Proposition* 2.2, *then*

$$U(\bar{x}) = \sum_{i=0}^{m-1} c_i(p_{i+1} - p_i), \tag{2.12 (i)}$$

and

$$U(\bar{x}) = -2 \sum_{i=0}^{m} p_i \Delta_i \dddot{\bar{x}}. \tag{2.12 (ii)}$$

Proof. If we dot-multiply (2.10ii) by $\dot{\bar{x}}$ and integrate over $[0, \bar{s}]$, using integration by parts, we obtain

$$-2U(\bar{x}) + 3U(\bar{x}) = \sum_{i=0}^{m-1} c_i(p_{i+1} - p_i),$$

which is (2.12i). Now use (2.10ii) at $\bar{s}_i + 0$ and $\bar{s}_i - 0$ and subtract to obtain $2\Delta \dddot{\bar{x}} = c_i - c_{i-1}$, which holds for $i = 0, 1, \cdots, m$ if we define $c_{-1} = c_m = 0$. By (2.12i) we have

$$U(\bar{x}) = \sum_{i=0}^{m} c_i(p_{i+1} - p_i) = -\sum_{i=0}^{m} p_i(c_i - c_{i-1}) = -2 \sum_{i=0}^{m} p_i \Delta_i \dddot{\bar{x}},$$

so (2.12ii) is also proved.

Remark 2.2. Since $U(x) = 0$ only if x is linear, it follows from (2.12ii) that an extremal P-interpolant $\bar{x}$ that is not linear must have a discontinuity of $\dddot{\bar{x}}$ at some of the knots (or else, $\dddot{\bar{x}}(p_0 + 0) \neq 0$ or $\dddot{\bar{x}}(p_m - 0) \neq 0$).

From an extremal x, as characterized in Proposition 2.2, one can obtain infinitely many other extremals by shifting the terminals $x(0)$ and $x(1)$ along the rays that are tangent to x at p_0 and p_m. The value of $U(x)$ is not changed by these variations. We wish to ignore these trivial portions of an extremal and will for this reason adopt the following convention. If we speak of an extremal P-interpolant x with arc parametrization $\bar{x}$, knots $p_i = \bar{x}(\bar{s}_i), i = 0, \cdots, m$ and length $\bar{s}$ then, unless stated otherwise, $\bar{s}_0 = 0$, $\bar{s}_m = \bar{s}$, and the terminals are $p_0 = \bar{x}(0) = 0$, $p_m = \bar{x}(\bar{s}_m) = \bar{x}(\bar{s})$.

For some applications one wishes to constrain the P-interpolants further by prescribing the length $\bar{s}$ of the arc between the terminals. Let the class of these P-interpolants be called *length-prescribed* (they differ from the "length-constrained" interpolants of [3]). The next definition deals with the extremals for U in this class.

DEFINITION 2.2. The admissible P-interpolant x, with arc length parametrization $\bar{x} = x \circ s_x^{-1}$, knots $p_i = \bar{x}(\bar{s}_i)$, $i = 0, \cdots, m$, length $\bar{s} = s_x(1)$, is a length-prescribed extremal P-interpolant if

$$\text{(2.13)} \quad \text{(i)} \qquad U'(x)[z] + \lambda S'(x)[z] = 0$$

for any $z \in H_2(I)$ for which $z \circ s_x^{-1}(\bar{s}_i) = 0$, $i = 0, \cdots, m$ and $\lambda \in \mathbb{R}$ determined so that

$$\text{(2.13)} \quad \text{(ii)} \qquad S(x) := \int_I |\dot{x}| = \bar{s}.$$

For these extremals we have a characterization similar to that of Proposition 2.2; note that $S'(x)$ is given by

$$S'(x)[y] = \int_0^{\bar{s}} \dot{\bar{x}}\dot{\bar{y}}, \quad \text{for all } y \in H_2(I), \quad \bar{x} = x \circ s_x^{-1}, \quad \bar{y} = y \circ s_x^{-1}.$$

PROPOSITION 2.3. *The admissible P-interpolant x with arc length parametrization* $\bar{x}$, *knots* $p_i = \bar{x}(\bar{s}_i)$, $i = 0, 1, \cdots, m$ *and length* $\bar{s} = \bar{s}_m$ *is a length-prescribed extremal if and only if*

$$\text{(i)} \qquad \bar{x} \in C^2[0, \bar{s}], \ddot{\bar{x}}(0) = 0, \qquad \ddot{\bar{x}}(\bar{s}) = 0,$$

$$\text{(2.14)} \quad \text{(ii)} \qquad (2\dddot{\bar{x}} + 3\kappa_x^2\dot{\bar{x}} - \lambda\dot{\bar{x}})_{(\bar{s}_{i-1}, \bar{s}_i)} = c_i \in \mathbb{R}^2, \qquad i = 1, \cdots, m,$$

$$\text{(iii)} \qquad \lambda\bar{s} = U(\bar{x}) - \sum_{i=1}^{m} c_i(p_i - p_{i-1}) = U(\bar{x}) + 2\sum_{i=0}^{m} p_i\Delta_i\ddot{\bar{x}},$$

hold with $\bar{x}_{(\bar{s}_{i-1}, \bar{s}_i)} \in C^\infty(\bar{s}_{i-1}, \bar{s}_i)$ $(i = 1, \cdots, m)$.

Proof. To prove the implication (2.14i, ii)$\Rightarrow$(2.13i) one proceeds as in the first part of the proof of Proposition 2.2. If, next, (2.14ii) is dot-multiplied by $\dot{\bar{x}}$ and integrated over $[0, \bar{s}]$ one obtains, using integration by parts,

$$-2U(\bar{x}) + 3U(\bar{x}) - \lambda\bar{s} = \sum_{i=1}^{m} c_i(p_i - p_{i-1}) = -2\sum_{i=0}^{m} p_i\Delta_i\ddot{\bar{x}},$$

which give (2.14iii), which is seen to be equivalent to (2.13ii).

Conversely, if (2.13i) holds then selecting $z \in C^\infty[0, \bar{s}]$ with support in $[\bar{s}_i, \bar{s}_{i+1}]$, i fixed, we have

$$\int_{\bar{s}_i}^{\bar{s}_{i+1}} (2\ddot{\bar{x}} + F - \lambda\bar{x})\ddot{\bar{z}} = 0,$$

where F is as in the proof of Proposition 2.2. It follows that

$$(2\dddot{\bar{x}} + 3\kappa_x^2\dot{\bar{x}} - \lambda\dot{\bar{x}})_{(\bar{s}_i, \bar{s}_{i+1})} = c_{i+1}.$$

The regularity properties of x are proved as before. The argument in the first part of the proof shows, that given (2.14i, ii), then (2.14iii) and (2.13ii) are equivalent.

Remark 2.3. For any configuration $P = \{p_0, p_1, \cdots, p_m\}$, there exists an extremal P-interpolant satisfying (2.14) with $\lambda \in \mathbb{R}$. Indeed, if

$$L_0 = \sum_{i=0}^{m-1} |p_{i+1} - p_i|,$$

then the length-prescribed extremal x, which minimizes $\int_0^L \kappa_x^2$ among all admissible P-interpolants with length equal to $L > L_0$ is guaranteed to exist [6] and satisfies (2.13i) [cf. Appendix].

Remark 2.4. Assume $\bar{x}_0$ is an (unconstrained) extremal P-interpolant and that its length is $\bar{s}_0$. Then $\bar{x}_0$ satisfies (2.10) and (2.12), hence $\bar{x}_0$ also satisfies (2.14) with $\lambda = 0$, i.e., $\bar{x}_0$ is also a length-prescribed extremal P-interpolant for the given length $\bar{s}_0$. We know that there exist length-prescribed P-interpolants of given length $\bar{s}, \bar{s}$ near $\bar{s}_0$. More precisely, let us assume that for every neighborhood N_ε of $\bar{x}_0$ in H_2 there exists some $\delta > 0$ such that if $|\bar{s} - \bar{s}_0| < \delta$ there is a unique solution $\bar{x}_{\bar{s}} \in N_\varepsilon$ of Equations (2.14) of length $\bar{s}$. Let $\bar{s} \to V(\bar{s}) = \int_0^{\bar{s}} |\ddot{\bar{x}}_s|^2$ be the function of $\bar{s}$ whose value is that of the curvature functional at $\bar{x}_s$. Then one can show that V has a critical point at $\bar{s}_0$, i.e., $V'(\bar{s}_0) = 0$. Moreover, if $V(\bar{s}_0)$ is a local minimum of V then $U(\bar{x}_0) \leqq U(\bar{x})$ for every admissible P-interpolant $\bar{x}$ in a neighborhood of $\bar{x}_0$ in H_2, i.e., the extremal $\bar{x}_0$ is locally minimizing.

Remark 2.5. Curves $\bar{x}: \mathbb{R} \to \mathbb{R}^2$, satisfying the equation

$$2\dddot{\bar{x}} + 3\kappa_x^2 \dot{\bar{x}} - \lambda \dot{\bar{x}} = c$$

are called *elastica* (cf. [8]), more specifically *inflexional* elastica if the curve has inflection points (which is the case if and only if $\lambda^2 \leqq c^2$). Curves for which $\lambda = 0$, the case of primary interest in this paper, will be referred to as *simple elastica*. Geometrically, simple elastica are characterized by the property that the angular variation between consecutive inflection points is exactly π (for all inflexional elastica the angular variation is $\geqq \pi$). A smooth oriented curve in $\mathbb{R}^2$ with continuous curvature which consists of finitely many subarcs of the simple elastica and has (possibly) discontinuities of the curvature derivative at the interpolation points $p_1, \cdots, p_{m-1}$ only is called an *interpolating elastica* (cf. [8]).

The next proposition deals with implications and equivalences of (2.14ii). It should be observed that these results apply to Equation (2.10ii) as well, since the latter is the special case of (2.14ii) with $\lambda = 0$.

PROPOSITION 2.4. *Condition* (2.14ii) *implies each of the following four conditions on* $(\bar{s}_{i-1}, \bar{s}_i)$ $(i = 1, \cdots, m)$:

$$\begin{aligned} &\text{(i)} && \kappa_x^2 = c_i \dot{\bar{x}} + \lambda, \\ &\text{(ii)} && \dot{\kappa}_x = \tfrac{1}{2}[\dot{\bar{x}}, c_i], \\ &\text{(iii)} && \kappa_x = \tfrac{1}{2}[\bar{x}, c_i] + \gamma_i, \qquad \gamma_i \in \mathbb{R}, \\ &\text{(iv)} && \ddot{\kappa}_x + \tfrac{1}{2}\kappa_x^3 - \left(\frac{\lambda}{2}\right)\kappa_x = 0. \end{aligned} \tag{2.15}$$

Moreover, (2.15i) *also implies* (2.14ii).

Proof. If (2.14ii) is dot-multiplied by $\dot{\bar{x}}$, one obtains, since $|\dot{\bar{x}}| = 1$, $|\ddot{\bar{x}}|^2 = \kappa_x^2$, $\dot{\bar{x}}\ddot{\bar{x}} = 0$, $\dot{\bar{x}}\dddot{\bar{x}} + |\ddot{\bar{x}}|^2 = 0$:

$$3\kappa_x^2 + 2\dot{\bar{x}}\dddot{\bar{x}} = \kappa_x^2 + 2|\ddot{\bar{x}}|^2 + 2\dot{\bar{x}}\dddot{\bar{x}} = \kappa_x^2 = c_i \dot{\bar{x}} + \lambda,$$

hence (2.15i). Differentiating (2.15i) we obtain

$$\kappa_x \dot{\kappa}_x = \tfrac{1}{2} c_i \ddot{\bar{x}}.$$

If $\kappa_x(s) \neq 0$ for some s, then $|\kappa_x(s)| = |\ddot{\bar{x}}(s)| > 0$ and $\ddot{\bar{x}}(s)/\kappa_x(s)$ is the unit vector $(-\dot{\bar{x}}^2(s), \dot{\bar{x}}^1(s))$. Therefore, $\dot{\kappa}_x(s) = \frac{1}{2}[\dot{\bar{x}}(s), c_i]$ and (2.15ii) holds in this case. If $\kappa_x(s) = 0$[1]

[1] The authors thank S. D. Fisher for a helpful suggestion here.

and s is a limit of s_n such that $\kappa_x(s_n) \neq 0$, then continuity of $\dot{x}$ and $\dot{\kappa}_x$ together with the previous argument gives the same equation. On a fixed interval $(\bar{s}_{i-1}, \bar{s}_i)$, let $\Gamma_i = \{s: \kappa_x(s) \neq 0\}$ and let (α, β) be any subinterval in the decomposition of $(\bar{s}_{i-1}, \bar{s}_i) \backslash \bar{\Gamma}_i$. We must show that (2.15ii) holds on (α, β). Now κ_x and $\dot{\kappa}_x$ are zero on (α, β). By the continuity of $\dot{\kappa}_x$ on $(\bar{s}_{i-1}, \bar{s}_i)$ it follows that $\dot{\kappa}_x(\beta) = 0 = [\dot{x}(\beta), c_i]$. Moreover, since $\kappa_x = 0$ on (α, β),

$$\bar{x}(s) = as + b, \quad s \in [\alpha, \beta], \quad a, b \in \mathbb{R}^2. \tag{2.16}$$

Thus, from (2.16),

$$[\dot{x}(s), c_i] = [a, c_i], \qquad s \in [\alpha, \beta].$$

In particular, $[a, c_i] = 0$ and thus (2.15ii) holds on (α, β). Next, (2.15iii) follows from (2.15ii) by integration. To show that (2.15iv) holds, let $\kappa_x(s) \neq 0$; then $\ddot{x}(s) = \kappa_x(s)(-\dot{x}^2(s), \dot{x}^1(s))$ and one obtains upon differentiating (2.15ii),

$$\dot{\kappa}_x(s) = \tfrac{1}{2}[\ddot{x}(s), c_i] = -\tfrac{1}{2}\kappa_x(s) c_i \dot{x}(s) = \tfrac{1}{2}[\lambda \kappa_x(s) - \kappa_x^3(s)].$$

Thus, (2.15iv) holds at every point s for which $\kappa_x(s) \neq 0$. The case when s is a limit point of s_n for which $\kappa_x(s_n) \neq 0$ then follows immediately. The case when s is not such a limit point is of course trivial.

To show that (2.15i) implies (2.14ii), assume $\ddot{x}(s) \neq 0$ for some $s \in (\bar{s}_i, \bar{s}_{i+1})$. Since $\dot{x}(s)$ and $\ddot{x}(s)/|\ddot{x}(s)|$ are orthogonal unit vectors in $\mathbb{R}^2$, we have, using (2.15i) and the fact that $\dot{x}(s)\dddot{\bar{x}}(s) + |\ddot{x}(s)|^2 = 0$,

$$\begin{aligned} c_i &= (c_i \dot{x}(s))\dot{x}(s) + (c_i \ddot{x}(s)) \frac{\ddot{x}(s)}{|\ddot{x}(s)|^2} \\ &= (\kappa_x^2(s) - \lambda)\dot{x}(s) + \frac{d(\kappa_x^2(s))}{ds} \frac{\ddot{x}(s)}{|\ddot{x}(s)|^2} \\ &= (3\kappa_x^2(s) - \lambda)\dot{x}(s) + (2\dddot{\bar{x}}(s)\dot{x}(s))\dot{x}(s) + (2\dddot{\bar{x}}(s)\ddot{x}(s)) \frac{\ddot{x}(s)}{|\ddot{x}(s)|^2} \\ &= (3\kappa_x^2(s) - \lambda)\dot{x}(s) + 2\dddot{\bar{x}}(s), \end{aligned}$$

which is just (2.14ii). If s is a limit point of s_n for which $\ddot{x}(s_n) \neq 0$, then the same equation holds by continuity. If s is not such a limit point, then (2.14ii) reduces to $c_i = -\lambda \dot{x}(s)$, which is clearly obtained by dot-multiplication of (2.15i) by $\dot{x}(s)$. This concludes the proof of Proposition 2.4.

At the end of this section we mention still another constraining condition for extremal interpolants. It consists in fixing the angle that the interpolant x makes with a fixed line at the terminal knot $p_0 = \bar{x}(\bar{s}_0)$ and/or $p_m = \bar{x}(\bar{s}_m)$. Thus, the condition is

$$\dot{x}(\bar{s}_0) = e_0 \quad \text{and/or} \quad \dot{x}(\bar{s}_m) = e_m \tag{2.17}$$

where e_0, e_m are unit vectors in $\mathbb{R}^2$. We refer to these extremals as *angle-constrained.*

If x is an admissible P-interpolant with knots $p_i = \bar{x}(\bar{s}_i)$ which satisfies the constraint (2.17) then any other P-interpolant in a sufficiently small H_2-neighborhood of x, satisfying the same condition, is of the form $x + \varepsilon z$ where $z \in H_2$, $z \circ s_x^{-1}(\bar{s}_i) = 0$, $i = 0, 1, \cdots, m$, $\dot{z} \circ s_x^{-1}(\bar{s}_0) = 0$ and/or $\dot{z} \circ s_x^{-1}(\bar{s}_m) = 0$. It is easily seen that if

$\bar{x}$ is an extremal P-interpolant with the added constraint $\dot{\bar{x}}(\bar{s}_0) = e_0$ then the "free boundary" condition $\ddot{\bar{x}}(\bar{s}_0) = 0$ of (2.10i) is replaced by $\dot{\bar{x}}(\bar{s}_0) = e_0$. Similarly if $\dot{\bar{x}}(\bar{s}_m) = e_m$ is a constraint then this condition replaces $\ddot{\bar{x}}(\bar{s}_m) = 0$. There is no other change in the conditions of Proposition 2.2.

3. Normal representation of extremals. In this section the dependent variable is an angle. Let $T(=T^1)$ denote the 1-dimensional torus. $\phi \in T$ is represented by a real number (also denoted as) ϕ, one of the set $\phi + 2k\pi$, $k = 0, \pm 1, \pm 2, \cdots$. A continuous function $\theta: (0, \bar{s}) \to T$ is represented by a continuous function (also denoted as) $\theta: (0, \bar{s}) \to \mathbb{R}$, one of the set $\theta + 2k\pi$. The derivative $\dot{\theta}$ is always a unique function $(0, \bar{s}) \to \mathbb{R}$. Let $\mathring{H}_1(0, \bar{s})$ denote the class of absolutely continuous functions $\theta: (0, \bar{s}) \to \mathbb{R}$ for which $\dot{\theta} \in L_2(0, \bar{s})$. Then the function $x = x_\theta: (0, \bar{s}) \to \mathbb{R}^2$, defined by

$$x_\theta^1(s) = \int_0^s \cos\theta, \qquad x_\theta^2(s) = \int_0^s \sin\theta \tag{3.1}$$

is in $H_2(0, \bar{s})$ and represents an oriented curve C in the x^1x^2-plane, parametrized with respect to arc length, $x(0) = 0$, $\dot{x}_2/\dot{x}_1 = \tan\theta$, $\theta(s)$ is the angle which the curve C makes with the x^1-axis at arc length s, and $\dot{\theta}_s$ is the curvature of C at s. Conversely, given an oriented curve C with cartesian representation $x \in H_2(0, \bar{s})$, parametrized with respect to arc length, there is a unique function $\theta_x \in \mathring{H}_1(0, \bar{s})$, representing a unique mapping $(0, \bar{s}) \to T$, such that

$$x^1(s) = x^1(0) + \int_0^s \cos\theta_x, \qquad x^2(s) = x^2(0) + \int_0^s \sin\theta_x. \tag{3.2}$$

We say that θ_x is the *normal representation* (n.r.) of C.

Curves C that differ by a translation have the same normal representation. If θ_x is the n.r. of C then $\theta_x + \text{const.}$ is the n.r. of a curve obtained from C by a rotation, and $-\theta_x$ is the n.r. of a curve obtained from C by a reflection at the x^1-axis. In a geometric setting we may identify curves C which differ only by a congruence, and each congruence class is represented by a single function θ with the specification $\theta(0) = 0$, $\dot{\theta}(0) > 0$ (or $\ddot{\theta}(0) > 0$ if $\dot{\theta}(0) = 0$).

In many cases it will be convenient to characterize extremal P-interpolants by their normal representation. For this purpose we replace Propositions 2.2 and 2.3 by the following propositions whose proofs we omit.

PROPOSITION 3.1. *The function $\bar{\theta} \in \mathring{H}_1(0, \bar{s})$ is the normal representation of an extremal P-interpolant with knots p_i, $i = 0, 1, \cdots, m$ at $0 = \bar{s}_0 < \bar{s}_1 < \cdots < \bar{s}_m$ and length $\bar{s} = \bar{s}_m$, if and only if the conditions*

$$\text{(i)} \qquad \bar{\theta} \in C^1[0, \bar{s}], \qquad \dot{\bar{\theta}}(0) = \dot{\bar{\theta}}(\bar{s}) = 0,$$

$$\text{(ii)} \qquad \ddot{\bar{\theta}}^2(s) = c_i^1 \cos\bar{\theta}(s) + c_i^2 \sin\bar{\theta}(s) \text{ for } s \in (\bar{s}_{i-1}, \bar{s}_i),\ c_i \in \mathbb{R}^2, \quad i = 1, \cdots, m, \tag{3.3}$$

$$\text{(iii)} \qquad \int_0^{\bar{s}_i} \cos\bar{\theta}(s)\, ds = p_i^1, \qquad \int_0^{\bar{s}_i} \sin\bar{\theta}(s)\, ds = p_i^2, \quad i = 1, \cdots, m$$

hold with $\bar{\theta}_{(\bar{s}_{i-1}, \bar{s}_i)} \in C^\infty(\bar{s}_{i-1}, \bar{s}_i)$, $i = 1, \cdots, m$.

PROPOSITION 3.2. *The function $\bar{\theta} \in \mathring{H}_1(0, \bar{s})$ is the normal representation of a length-prescribed extremal P-interpolant with knots p_i, $i = 0, 1, \cdots, m$ at $0 = \bar{s}_0 < \bar{s}_2 <$*

$\cdots < \bar{s}_m$ *and length* $\bar{s} = \bar{s}_m$ *if and only if the conditions*

$$
\begin{aligned}
&\text{(i)} && \bar{\theta} \in C^1[0, \bar{s}], \qquad \dot{\bar{\theta}}(0) = \dot{\bar{\theta}}(\bar{s}) = 0, \\
&\text{(ii)} && \dot{\bar{\theta}}^2(s) = c_i^1 \cos \bar{\theta}(s) + c_i^2 \sin \bar{\theta}(s) + \lambda \\
& && \qquad\qquad \text{for } s \in (\bar{s}_{i-1}, \bar{s}_i),\ c_i \in \mathbb{R}^2, \quad i = 1, \cdots, m, \\
&\text{(iii)} && \int_0^{\bar{s}_i} \cos \bar{\theta}(s)\, ds = p_i^1, \qquad \int_0^{\bar{s}_i} \sin \bar{\theta}(s)\, ds = p_i^2, \qquad i = 1, \cdots, m, \\
&\text{(iv)} && \lambda \bar{s} = \int_0^{\bar{s}} \dot{\bar{\theta}}^2 - \sum_{i=1}^{m} c_i (p_i - p_{i-1})
\end{aligned}
\tag{3.4}
$$

hold with $\bar{\theta}_{(\bar{s}_{i-1}, \bar{s}_i)} \in C^\infty(\bar{s}_{i-1}, \bar{s}_i)$, $i = 1, \cdots, m$.

Remark 3.1. The conditions $\dot{\bar{\theta}}(\bar{s}_i - 0) = \theta(\bar{s}_i + 0)$, $i = 1, \cdots, m-1$, $\theta(0) = 0$, and $\dot{\bar{\theta}}(\bar{s}_m) = 0$ result in $m+1$ conditions on the vector constants $c_1, \cdots, c_m$. Both in Proposition 3.1 and 3.2 we have

$$
\begin{aligned}
&\text{(i)} && (c_{i+1}^1 - c_i^1) \cos \bar{\theta}(\bar{s}_i) + (c_{i+1}^2 - c_i^2) \sin \bar{\theta}(\bar{s}_i) = 0, \quad i = 1, \cdots, m-1 \\
&\text{(ii)} && c_1^1 \cos \bar{\theta}(0) + c_1^2 \sin \bar{\theta}(0) + \lambda = 0, \\
&\text{(iii)} && c_m^1 \cos \bar{\theta}(\bar{s}_m) + c_m^2 \sin \bar{\theta}(\bar{s}_m) + \lambda = 0,
\end{aligned}
\tag{3.5}
$$

where λ is to be taken as 0 in the case of Proposition 3.1.

For use in §§ 4 and 5 we state and prove

PROPOSITION 3.3. *The function* $\bar{\theta} \in \mathring{H}_1(0, \bar{s})$ *is the normal representation of an extremal P-interpolant with knots* p_i, $i = 0, \cdots, m$ *at* $0 = \bar{s}_0 < \cdots < \bar{s}_m$ *and length* $\bar{s} = \bar{s}_m$ *if and only if interpolation conditions* (3.4iii) *and the conditions*

$$
\begin{aligned}
&\text{(i)} && 2\ddot{\bar{\theta}}(s) + c_i^1 \sin \bar{\theta}(s) - c_i^2 \cos \bar{\theta}(s) = 0 \\
& && \qquad\qquad \text{for } s \in (\bar{s}_{i-1}, \bar{s}_i), \quad c_i \in \mathbb{R}^2, \quad i = 1, \cdots, m, \\
&\text{(ii)} && (c_{i+1}^1 - c_i^1) \cos \bar{\theta}(\bar{s}_i) + (c_{i+1}^2 - c_i^2) \sin \bar{\theta}(\bar{s}_i) = 0, \qquad i = 1, \cdots, m, \\
&\text{(iii)} && \dot{\bar{\theta}}(0) = 0, \qquad \dot{\bar{\theta}}(\bar{s}) = 0
\end{aligned}
\tag{3.6}
$$

hold with $\bar{\theta}_{(\bar{s}_{i-1}, \bar{s}_i)} \in C^\infty(\bar{s}_{i-1}, \bar{s}_i)$, $i = 1, \cdots, m$ *and* $c_{m+1} = 0$.

Proof. The forward implication follows directly from Proposition 3.1 and Remark 3.1. The converse implication follows upon multiplying (3.6i) by $\dot{\bar{\theta}}(s)$ and integrating; if the integrated equation is evaluated at $s = \bar{s}_m$ and (3.6ii, iii) is used, the constant of integration is seen to be 0. Thus, (3.6) implies

$$
\begin{gathered}
\theta^2(s) - c_i^1 \cos \bar{\theta}(s) - c_i^2 \sin \bar{\theta}(s) = 0 \quad \text{for } s \in (\bar{s}_{i-1}, \bar{s}_i), \\
\bar{\theta} \in C^1[0, \bar{s}],\ i = 1, \cdots, m,
\end{gathered}
$$

and the result now follows from Proposition 3.1.

4. Manifold of extremals. As noted in the previous section, we may consider equivalence classes of curves differing by a congruence, with representer satisfying $x(0) = 0$, $\theta(0) = 0$, $\dot{\theta}(0) > 0$; or $\ddot{\theta}(0) > 0$ if $\dot{\theta}(0) = 0$.

DEFINITION 4.1. The extremal interpolant E is *proper* if:

(4.1)
(i) E has nonzero curvature κ_i at each internal interpolation node s_i, $i = 1, \cdots, m-1$.
(ii) Each internal interpolation node is a genuine knot, i.e., there is a discontinuity $\Delta_i \dot{\kappa} \neq 0$ in the derivative of the curvature at s_i, $i = 1, \cdots, m-1$.

DEFINITION 4.2. The m-tuple $(k_1, \cdots, k_m)$ of nonnegative integers is the *mode* of E if there are k_j inflection points strictly between the $(j-1)$th and jth interpolation nodes; here, an inflection point denotes a point of zero curvature and we note that $\dot{\theta}$ must change sign. For a fixed mode $(k_1, \cdots, k_m)$, $\mathscr{E} = \mathscr{E}_{(k_1, \cdots, k_m)}$ will denote the class of proper $(m+1)$-extremal interpolants E in the mode $(k_1, \cdots, k_m)$.

PROPOSITION 4.1. *$\mathscr{E}$ is a (finite-dimensional) metric space under the metric*

$$d(E_1, E_2) = \max_{0 \leqq t \leqq 1} \left| \frac{d}{dt} x_1(s_{E_1} t) - \frac{d}{dt} x_2(s_{E_2} t) \right|. \tag{4.2}$$

Here s_{E_j} and x_j represent the lengths and Cartesian representations (parametrized w.r.t. arc length) of E_j, $j = 1, 2$, respectively, where $x_j(0) = 0$, $j = 1, 2$.

Remark 4.1. We omit the routine proof. We observe that $\mathscr{E}$ is not complete. Indeed, if each $\mathscr{E}_{(k_1, \cdots, k_m)}$ (m fixed) is embedded in the space of all extremal interpolants, with metric described by (4.2), then the boundary $\partial\mathscr{E}$ of $\mathscr{E}$ may contain an extremal interpolant which is not proper. We also observe that if E_1 is close to E_2, then the configuration interpolated by E_1 is close to that interpolated by E_2. This would not be true if E_1, E_2 were not restricted to a class $\mathscr{E}_{(k_1, \cdots, k_m)}$ (see Example 4.2 at the end of this section).

Now let $\theta_E \in C^1[0, s_E]$ be the normal representation of some $E \in \mathscr{E}$. If $0 = s_0 < s_1 < \cdots < s_m = s_E$ are the interpolation nodes of E, put $\theta_E(s_i) = \alpha_i$, $i = 0, \cdots, m$ and $\alpha_E = (\alpha_1, \cdots, \alpha_{m-1}) \in T^{m-1}$. Setting $c_i = -2\mu_i$ we have from Propositions 3.1 and 3.3 the existence of a unique multiplier $\mu_E = (\mu_1, \cdots, \mu_m) \in (\mathbb{R}^2)^m$ such that, for $k = 1, \cdots, m$,

$$\begin{aligned}
&\text{(i)} \quad \tfrac{1}{2}\dot{\theta}_E^2(s) + \mu_k^1 \cos\theta_E(s) + \mu_k^2 \sin\theta_E(s) = 0, \qquad s_{k-1} < s < s_k, \\
&\text{(ii)} \quad \ddot{\theta}_E(s) - \mu_k^1 \sin\theta_E(s) + \mu_k^2 \cos\theta_E(s) = 0, \qquad s_{k-1} < s < s_k; \\
&\text{(iii)} \quad \dot{\theta}_E(0) = 0, \qquad \dot{\theta}_E(s_E) = 0.
\end{aligned} \tag{4.3}$$

By introducing the more convenient notation

$$B_k = (A_k, \beta_k), \quad A_k > 0, \quad \beta_k \in T^1, \quad k = 1, \cdots, m,$$

where

$$\mu_k^1 = -A_k \sin\beta_k, \qquad \mu_k^2 = A_k \cos\beta_k,$$

we may rewrite (4.3) in terms of the multipliers B_k for $k = 1, \cdots, m$:

$$\begin{aligned}
&\text{(i)} \quad \tfrac{1}{2}\dot{\theta}_E^2(s) + A_k \sin(\theta_E(s) - \beta_k) = 0, \qquad s_{k-1} < s < s_k, \\
&\text{(ii)} \quad \ddot{\theta}_E(s) + A_k \cos(\theta_E(s) - \beta_k) = 0, \qquad s_{k-1} < s < s_k; \\
&\text{(iii)} \quad \dot{\theta}_E(0) = 0, \qquad \dot{\theta}_E(s_E) = 0.
\end{aligned} \tag{4.4}$$

Since $\theta = 0$ is not a proper extremal we must have $\ddot{\theta}_E(0+) \neq 0$. We consider it as part of the definition of $\mathscr{E}$ that

$$\text{(iv)} \quad \ddot{\theta}_E(0+) > 0 \tag{4.4}$$

for all $E \in \mathscr{E}$. Geometrically speaking, $\mathscr{E}$ contains only extremals which turn counterclockwise near the initial point.

PROPOSITION 4.2. *$\theta_E \in C^1[0, s_E]$ is the normal representation of an extremal $E \in \mathscr{E}_{(k_1,\cdots,k_m)}$ with interpolation nodes at $0 = s_0 < s_1 < \cdots < s_m = s_E$ if and only if*

A. *θ_E satisfies* (4.4i–iv) *for some $A_k > 0$, $\beta_k \in T^1$.*

B. $\sin(\beta_i - \beta_{i+1}) \neq 0$.

C. $\operatorname{sgn} \dot\theta_E(s_i) = (-1)^{k_1+\cdots+k_i}$, $i = 1, \cdots, m-1$.

Proof. Since $\theta_E \in C^1$ we have, by (4.4i)

$$(4.5) \qquad \text{(i)} \quad A_i \sin(\alpha_i - \beta_i) - A_{i+1} \sin(\alpha_i - \beta_{i+1}) = 0, \qquad i = 1, \cdots, m-1,$$

where $\alpha_i = \theta_E(s_i)$. By (4.4ii), $\ddot\theta_E(s_i - 0) = \ddot\theta_E(s_i + 0)$ if and only if

$$(4.5) \qquad \text{(ii)} \quad A_i \cos(\alpha_i - \beta_i) - A_{i+1} \cos(\alpha_i - \beta_{i+1}) = 0.$$

The two equations (4.5i, ii) are equivalent to (4.5i) and $\sin(\beta_i - \beta_{i+1}) = 0$. Thus, the above condition B is equivalent to the condition that each interpolation node of E be a genuine knot (see 4.1ii). There are k_i inflection points of E between the $(i-1)$th and ith interpolation nodes if and only if $\dot\theta_E$ changes sign k_i times between s_{i-1} and s_i, i.e.,

$$(4.6) \qquad \operatorname{sgn} \kappa_{i-1} \cdot \operatorname{sgn} \kappa_i = (-1)^{k_i}, \qquad \kappa_i = \dot\theta_E(s_i).$$

Since by (4.4iv) $\dot\theta_E(s) > 0$ for all sufficiently small s, (4.6) is equivalent to the above condition C. Condition C also implies that $\kappa_i \neq 0$ for $i = 1, \cdots, m-1$, thus condition (4.1i) is also satisfied.

Remark 4.2. Condition B also implies

$$(4.7) \qquad A_i - A_{i+1} \neq 0, \qquad i = 1, \cdots, m-1.$$

Indeed if $A_i - A_{i+1} = 0$ then by (4.5i) $\sin(\alpha_i - \beta_i) = \sin(\alpha_i - \beta_{i+1})$, hence $\beta_i = \beta_{i+1}$ or $\beta_i = \beta_{i+1} + \pi \pmod{2\pi}$, which contradicts B.

We now define

$$\tilde{\mathbb{R}}_+^m = \{A \in \mathbb{R}_+^m : A_i - A_{i+1} \neq 0,\ i = 1, \cdots, m-1\},$$

$$(4.8) \qquad \tilde T^m = \{\beta \in T^m : \sin(\beta_i - \beta_{i+1}) \neq 0,\ i = 1, \cdots, m-1\},$$

$$B = B_E = (B_1, \cdots, B_m),\ \alpha = \alpha_E = (\alpha_1, \cdots, \alpha_m).$$

By Proposition 4.2 each extremal $E \in \mathscr{E}_{(k_1,\cdots,k_m)}$ determines a unique point $B_E \in \tilde{\mathbb{R}}_+^m \times \tilde T^m$. In the next three propositions we shall describe the mapping $E \to B_E$ via the composition of two mappings: the homeomorphism

$$(4.9) \qquad \text{(i)} \quad J : E \to (\alpha_E, B_E)$$

of $\mathscr{E}$ onto $J(\mathscr{E}) \subset T^{m-1} \times (\tilde{\mathbb{R}}_+^m \times \tilde T^m)$ and the projection

$$(4.9) \qquad \text{(ii)} \quad M : (\alpha, B) \to B$$

of $J(\mathscr{E})$ into $\tilde{\mathbb{R}}_+^m \times \tilde T^m$, which is a local diffeomorphism.[2] The composition $M \circ J$ is a global homeomorphism.

PROPOSITION 4.3. *The mapping J is a homeomorphism of $\mathscr{E}_{(k_1,\cdots,k_m)}$ onto its image.*

Proof. The continuity of J follows directly from (4.2) and (4.4); note that $B_1, \cdots, B_m$ can be expressed via (4.4) in terms of α_i, κ_i, $\Delta_i\dot\kappa$, thus also in terms of

[2] The homeomorphism M is a diffeomorphism if M and M^{-1} are continuously Fréchet differentiable on their domains. This is sometimes referred to as a C^1-diffeomorphism.

$\theta_E = \arctan \dot{x}_{E,2}/\dot{x}_{E,1}$. Suppose now that $J(E) = (\alpha_E, B_E)$. We show that $E \in \mathscr{E}$ is uniquely and continuously determined by its map (α_E, B_E). By (4.4)

$$\sin(\theta_E(0) - \beta_1) = 0, \qquad \cos(\theta_E(0) - \beta_1) < 0,$$

hence $\alpha_0 = \theta_E(0) = \beta_1 + \pi \pmod{2\pi}$. The restriction of θ_E to $[s_0, s_1]$ is now uniquely determined from

$$\text{(4.10)} \quad \begin{aligned} &\text{(i)} && \dot{\theta}_E(s) = -[-2A_1 \sin(\theta_E(s) - \beta_1)]^{1/2} < 0, \\ &\text{(ii)} && \theta_E(0) = \alpha_0, \end{aligned}$$

with s_1 uniquely determined from

$$\text{(4.10)} \quad \text{(iii)} \quad \theta_E(s_1) = \alpha_1, \quad \dot{\theta}_E(s) = 0 \quad \text{for } k_1 \text{ values of } s \text{ in } (0, s_1).$$

Indeed if there is an $s_1' > s_1$ for which $\theta_E(s_1) = \theta_E(s_1') = \alpha_1$ then $\dot{\theta}_E(\bar{s}) = 0$ for some $s_1 < \bar{s} < s_1'$, hence $\dot{\theta}_E(s) = 0$ for more than k_1 values of s in $(0, s_1')$. This clearly leads to an inductive process; indeed if θ_E is defined on $[0, s_i]$, one obtains the restriction of θ_E to $[s_i, s_{i+1}]$ from the initial value problem defined by (4.4ii) with initial values $\theta_E(s_i)$ and $\dot{\theta}_E(s_i)$. s_{i+1} is uniquely determined from

$$\theta_E(s_{i+1}) = \alpha_{i+1}, \quad \dot{\theta}_E(s) = 0 \quad \text{for } k_{i+1} \text{ values of } s \text{ in } (s_i, s_{i+1}).$$

The process is terminated at $i = m-1$ by replacing the condition $\theta_E(s_{i+1}) = \alpha_{i+1}$ by $\dot{\theta}_E(s_m) = 0$. Since the continuity of J^{-1} is an easy consequence of (4.4) the proof is complete.

We determine now the image set

$$\text{(4.11)} \qquad \mathscr{S} = \mathscr{S}_{(k_1, \cdots, k_m)} = J(\mathscr{E}_{(k_1, \cdots, k_m)}).$$

The following are necessary conditions for $(\alpha, B) \in \mathscr{S}$:

$$\text{(4.12)} \quad \begin{aligned} &\text{(i)} && \sin(\alpha_i - \beta_i) < 0, \qquad i = 1, \cdots, m-1. \\ &\text{(ii)} && \text{If } k_l = 0 \text{ for some } 2 \leqq l \leqq m-1 \\ &&& \qquad \text{then } (-1)^{k_1 + \cdots + k_{l-1}} \sin(\alpha_l - \alpha_{l-1}) > 0. \\ &\text{(iii)} && A_i \sin(\alpha_i - \beta_i) = A_{i+1} \sin(\alpha_i - \beta_{i+1}), \qquad i = 1, \cdots, m-1. \end{aligned}$$

Conditions (4.9i and iii) express that $\dot{\theta}_E^2$ is positive and continuous at $s_1, \cdots, s_{m-1}$. If some $k_l = 0$ then there must be no inflection point between s_{l-1} and s_l, hence $\sin(\theta_E - \alpha_{l-1})$ does not change sign, or

$$\sin(\theta_E(s) - \alpha_{l-1}) \cdot \dot{\theta}_E(s_{l-1}) > 0 \quad \text{for } s_{l-1} < s \leqq s_l.$$

Using C of Proposition 4.2, we obtain (4.12ii).

We now show that conditions (4.12) characterize the image set $\mathscr{S}$ completely. We observe that (4.12i, ii) define an open set in the $(3m-1)$-dimensional space $T^{m-1} \times (\tilde{\mathbb{R}}_+^m \times \tilde{T}^m)$, while equations (4.12iii) single out a $2m$-dimensional surface in the open set.

PROPOSITION 4.4. *The image set* $J(\mathscr{E}_{(k_1, \cdots, k_m)})$ *is*

$$\mathscr{S}_{(k_1, \cdots, k_m)} = \{(\alpha, B) \in T^{m-1} \times (\tilde{\mathbb{R}}_+^m \times \tilde{T}^m) : \textit{conditions } (4.12\text{i, ii, iii}) \textit{ hold}\}.$$

Proof. We need to show that if (α, B) is such that (4.12) holds, then there are numbers

$$0 = s_0 < s_1 < \cdots < s_m = s_E \text{ and a function } \theta_E \in C^1[0, s_E]$$

such that conditions A, C of Proposition 4.2 are satisfied and moreover,

$$\theta_E(s_i) = \alpha_i, \qquad i = 1, \cdots, m-1 \tag{4.13}$$

(condition B follows from the definition of $\tilde{T}^m$). s_1 and the restriction of θ_E to $[s_0, s_1]$ are determined as in the proof of Proposition 4.3. Next, the restriction of θ_E to $[s_1, s_2]$ (s_2 as yet unknown) is determined from the initial value problem

$$\dot{\theta}(s) + (-1)^{k_1}[-2A_2 \sin(\theta(s) - \beta_2)]^{1/2} = 0, \qquad \theta(s_1) = \alpha_1. \tag{4.14}$$

The solution is the normal representation of a simple elastica with inflection points at equally spaced abscissas σ_k where $\theta(\sigma_k) = \beta_2$ or $\beta_2 + \pi \pmod{2\pi}$. By (4.12i and iii) we have

$$\sin(\alpha_1 - \beta_2) < 0, \qquad \sin(\alpha_2 - \beta_2) < 0,$$

and, therefore, $\theta(s)$ attains the values α_1 and a_2 exactly once between any two consecutive σ_k. Thus if there are $k_2 \geqq 1$ inflection points between s_1 and s_2 there is exactly one s_2 for which $\theta(s_2) = \alpha_2$. If $k_2 = 0$ and, say $\dot{\theta}(s_1) > 0$, then by condition (4.12ii) $\sin(\alpha_2 - \alpha_1) > 0$, which implies that $\theta(s)$ attains the value α_2 for $s_2 > s_1$, with no inflection point between s_1 and s_2. By the same arguments the values of $s_3, \cdots, s_{m-1}$ and the restriction of θ_E to $[s_3, s_4], \cdots, [s_{m-2}, s_{m-1}]$ are determined. $s_m = s_E$ and θ_E on $[s_{m-1}, s_m]$ are similarly obtained, except that the condition $\theta(s_m) = \alpha_m$ is replaced by $\dot{\theta}(s_m) = 0$. The obtained function θ_E is in $C^1[0, s_E]$ because of condition (4.12iii), and it satisfies (4.13) and conditions A and C of Proposition 4.2 by construction.

PROPOSITION 4.5. *The projection $M_{|\mathcal{S}}$ is a local diffeomorphism onto an open subset $\mathcal{M}_{(k_1,\cdots,k_m)}$ of $\tilde{\mathbb{R}}^m_+ \times \tilde{T}^m$. Thus $\mathcal{S}$ is a $2m$-dimensional smooth (even analytic) manifold ($\mathcal{S}$ is not connected if $m \geqq 1$). The composition map $M \circ J$ is a (global) homeomorphism of $\mathcal{E}_{(k_1,\cdots,k_m)}$ onto $\mathcal{M}_{(k_1,\cdots,k_m)}$.*

Proof. Choose any $(\alpha^0, B^0) \in \mathcal{S}$, $\alpha^0 = (\alpha_1^0, \cdots, \alpha_m^0)$ and $B^0 = (A_1^0, \cdots, A_m^0, \beta_1^0, \cdots, \beta_m^0)$. Let U denote the open subset of $T^{m-1} \times (\tilde{\mathbb{R}}^m_+ \times \tilde{T}^m)$ satisfying conditions (4.12i, ii). Define a mapping $\varphi: U \to \mathbb{R}^{m-1}$ by

$$\varphi_i(\alpha, B) = A_i \sin(\alpha_i - \beta_i) - A_{i+1} \sin(\alpha_i - \beta_{i+1}), \qquad i = 1, \cdots, m-1. \tag{4.15}$$

With this notation, $(\alpha, B) \in U$ is in $\mathcal{S}$ if and only if $\varphi_i(\alpha, B) = 0$, $i = 1, \cdots, m-1$. Now $\varphi_i(\alpha^0, B^0) = 0$ and the Jacobian $[\partial\varphi_i/\partial\alpha_j]$ is nonsingular at (α^0, B^0), since it is a diagonal matrix with diagonal entries

$$\begin{aligned} \frac{\partial \varphi_i}{\partial \alpha_i}(\alpha^0, B^0) &= A_i^0 \cos(\alpha_i^0 - \beta_i^0) - A_{i+1}^0 \cos(\alpha_i^0 - \beta_{i+1}^0) \\ &= \Delta_i \kappa \neq 0. \end{aligned} \tag{4.16}$$

We conclude that, for every neighborhood $U_0 \subset U$ of (α^0, B^0), there is a neighborhood N_0 of B^0 in $\tilde{\mathbb{R}}^m_+ \times \tilde{T}^m$ and a C^1-mapping α of N_0 such that $\alpha(B^0) = \alpha^0$ and $\varphi_i(\alpha(B), B) = 0$ for all $B \in N_0$. This proves that M is a local diffeomorphism.

By Proposition 4.3 the mapping J is a homeomorphism of $\mathcal{E}_{(k_1,\cdots,k_m)}$ onto $\mathcal{S}$. If the composite map $M \circ J$ is not a homeomorphism, there must be (α, B), (α', B) in $\mathcal{S}$ with $\alpha \neq \alpha'$, say $\alpha_i \neq \alpha_i' \pmod{2\pi}$. By (4.12iii) we have $A_i \sin(\alpha_i - \beta_i) - A_{i+1} \sin(\alpha_i - \beta_{i+1}) = 0$ and $A_i \sin(\alpha_i' - \beta_i) - A_{i+1} \sin(\alpha_i' - \beta_{i+1}) = 0$. These equations imply $\alpha_i' = \alpha_i + \pi \pmod{2\pi}$. But by (4.12i), $\sin(\alpha_i - \beta_i) < 0$ and $\sin(\alpha_i' - \beta_i) < 0$, which contradicts the previous conclusion. Thus Proposition 4.5 is completely proved.

Remark 4.3. It seems to be difficult to give an intrinsic characterization of the set $\mathcal{M}_{(k_1,\cdots,k_m)}$. Examples show that it does not coincide with $\tilde{\mathbb{R}}_+^m \times \tilde{T}^m$. It certainly contains points $B=(A,\beta)$ for each combination of the inequalities $A_1 \gtrless A_2, A_2 \gtrless A_3, \cdots, A_{m-1} \gtrless A_m$. Thus, $\mathcal{M}_{(k_1,\cdots,k_m)}$ and $\mathcal{S}_{(k_1,\cdots,k_m)}$ have at least 2^{m-1} disjoint components.

Remark 4.4. The following examples show that the results of this section fail if in Definition 4.1 either (4.1i) or (4.1ii) is omitted.

Example 4.1. Consider the 3-point interpolant E_0 with normal representation

$$\dot{\theta}_0(s)-[-2\sin\theta_0(s)]^{1/2}=0, \qquad 0\leqq s\leqq s_1=\int_0^{\pi}\frac{du}{\sqrt{2\sin u}},$$

$$\dot{\theta}_0(s)+[-\sin\theta_0(s)]^{1/2}=0, \qquad s_1\leqq s\leqq s_1+\int_0^{\pi}\frac{du}{\sqrt{\sin u}}.$$

Here $B_{E_0}=(1,0,\frac{1}{2},0)$, $\alpha_0=\pi$, $\alpha_1=2\pi$, $\alpha_2=\pi$. The mode is (0,0). E_0 violates (4.1i) since $\dot{\theta}_0(s_1)=0$. For $\varepsilon>0$ let the extremal E_ε of the same mode $(0,0)$ be given by $B_{E_\varepsilon}=(1,0,\frac{1}{2},\varepsilon)$, so that

$$\dot{\theta}_\varepsilon(s)-[-2\sin\theta_\varepsilon(s)]^{1/2}=0, \qquad 0\leqq s\leqq s_{1,\varepsilon},$$
$$\dot{\theta}_\varepsilon(s)+[-\sin(\theta_\varepsilon(s)-\varepsilon)]^{1/2}=0, \qquad s_{1,\varepsilon}\leqq s\leqq s_{2,\varepsilon}.$$

If $\theta_\varepsilon(s_{1,\varepsilon})$ is close to $\alpha_1=2\pi$ then $\theta_\varepsilon(s_{1,\varepsilon})=2\pi-\delta$ for some $\delta>0$. Thus, for $s=s_{1,\varepsilon}$,

$$\dot{\theta}_\varepsilon(s_{1,\varepsilon})=[2\sin\delta]^{1/2}=-[\sin(\delta+\varepsilon)]^{1/2},$$

which is impossible. Thus, $\alpha_1(B)$ cannot be defined as a continuous function in a full neighborhood of B_0.

Example 4.2. Choose $\pi<\alpha_*<2\pi$ and consider the 3-point interpolant E_* with normal representation

$$\dot{\theta}_*(s)-[-2\sin\theta_*(s)]^{1/2}=0, \qquad 0\leqq s\leqq s_{1_*}=\int_0^{\alpha_*-\pi}\frac{du}{\sqrt{2\sin u}},$$

$$\dot{\theta}_*(s)-[-2\sin\theta_*(s)]^{1/2}=0, \qquad s_{1_*}\leqq s\leqq s_2=2\int_0^{\pi}\frac{du}{\sqrt{2\sin u}}.$$

Here $B_{E_*}=(1,0,1,0)$, $\alpha_0=\pi$, $\alpha_1=\alpha_*$, $\alpha_2=\pi$. The mode is $(0,1)$. E_* violates (4.1ii) since $A_1=A_2$. For $\pi<\tilde{\alpha}<2\pi$ let $\tilde{E}$ be defined by normal representation

$$\dot{\tilde{\theta}}(s)-[-2\sin\tilde{\theta}(s)]^{1/2}=0, \qquad 0\leqq s\leqq \tilde{s}_1=\int_0^{\tilde{\alpha}-\pi}\frac{du}{\sqrt{2\sin u}},$$

$$\dot{\tilde{\theta}}(s)-[-2\sin\tilde{\theta}(s)]^{1/2}=0, \qquad \tilde{s}_1\leqq s\leqq s_2=2\int_0^{\pi}\frac{du}{\sqrt{2\sin u}}.$$

Here $B_{\tilde{E}}=(1,0,1,0)$, $\alpha_0=\pi$, $\alpha_1=\tilde{\alpha}$, $\alpha_2=\pi$, and the mode is $(0,1)$ as before. Since there are extremals for all $\pi<\tilde{\alpha}<2\pi$, $\tilde{E}$ cannot be defined by B and its mode.

5. Perturbations of configurations. In the last section it was seen that the extremal interpolants with m variable interpolation nodes (more precisely, those that belong to a fixed class $\mathcal{E}_{(k_1,\cdots,k_m)}$) form a $2m$-dimensional manifold. One expects that an arbitrary configuration $P=\{0,p_1,\cdots,p_m\}$ can be interpolated by an extremal interpolant (possibly by one from each class $\mathcal{E}_{(k_1,\cdots,k_m)}$). No solution of any kind exists for

this existence problem. In this section we investigate existence in the small. Does the set of configurations P for which extremal interpolants exist have nonempty interior in $\mathbb{R}^{2m}$? More specifically, which P in $\mathbb{R}^{2m}$ are interior points of this set?

To attack this perturbation problem one is tempted to consider the mapping from the $2m$-dimensional set $\mathcal{M}_{(k_1,\cdots,k_m)}$ that coordinatizes the elements of $\mathcal{E}_{(k_1,\cdots,k_m)}$ (see § 4), or from another $2m$-dimensional set of parameters, to the configurations in $\mathbb{R}^{2m}$ which are interpolated. However, this mapping is so complicated—it involves the elliptic integrals which are the solutions of the extremal equations—that little insight is gained from its consideration. For this reason we start with the extremals themselves, as defined by their differential equations.

Let $\bar{\theta}:[0, \bar{s}_m]\to T^1$ be the normal representation of a given extremal interpolant $\bar{E}$, which interpolates the configuration

$$\bar{P}=\{0, \bar{p}_1, \cdots, \bar{p}_m\},$$

so that

$$\int_0^{\bar{s}_i}\cos\bar{\theta}=\bar{p}_i^1,\ \int_0^{\bar{s}_i}\sin\bar{\theta}=\bar{p}_i^2,\qquad i=0,1,\cdots,m,$$
$$0=\bar{s}_0<\bar{s}_1<\cdots<\bar{s}_m. \tag{5.1}$$

Since we consider only extremals E with n.r. θ near $\bar{\theta}$, hence with knots s_i near $\bar{s}_i$, we choose $\bar{\varepsilon}>0$, $\bar{\varepsilon}=\frac{1}{2}\min(\bar{s}_i-\bar{s}_{i-1})$ and extend $\bar{\theta}$ to the interval $[0,\bar{s}]$, $\bar{s}=\bar{s}_m+\bar{\varepsilon}$, by setting

$$\bar{\theta}(s)=\bar{\theta}(\bar{s}_m),\qquad \bar{s}_m<s\leqq\bar{s}.$$

We introduce two spaces of mappings from the interval $[0,\bar{s}]$:

NBV = space of functions $\kappa:[0,\bar{s}]\to\mathbb{R}$ of bounded variation $V(\kappa)$ and continuous from the right with $\kappa(0)=\kappa(\bar{s}-0)=0$ and norm $V(\kappa)$.

NBV_1 = space of functions $\theta:[0,\bar{s}]\to T^1$, which are locally absolutely continuous and have derivatives $\dot{\theta}\in\mathrm{NBV}$, with norm $\sup|\theta|+V(\dot{\theta})$.

Both NBV and NBV_1 are B-spaces. Clearly $\bar{\theta}$ as defined above is in NBV_1 and $\dot{\bar{\theta}}$ is in NBV.

If $\theta\in\mathrm{NBV}_1$ is the normal representation of an extremal E which interpolates the configuration $P=\{0,p_1,\cdots,p_m\}$ at the nodes $0=s_0<s_1<\cdots<s_m<\bar{s}$ (more precisely, we speak of the linear extension of E to length $\bar{s}$) then the following equations hold (see Proposition 2.4):

$$\text{(i)}\quad \ddot{\theta}(s)+\tfrac{1}{2}\dot{\theta}^3(s)=0,\quad \text{for } s_{i-1}<s<s_i,\quad i=1,\cdots,m,$$
$$\ddot{\theta}(s)=0,\quad \text{for } s_m<s<\bar{s}.$$
$$\text{(ii)}\quad \dot{\theta}(s_i-0)-\dot{\theta}(s_i)=0,\qquad i=1,\cdots,m. \tag{5.2}$$
$$\text{(iii)}\quad \int_0^{s_i}\cos\theta=p_i^1,\qquad \int_0^{s_i}\sin\theta=p_i^2,\qquad i=1,\cdots,m.$$

It is easy to show that these equations characterize the interpolant E completely.

We rewrite equations (5.2) by using the values

$$\dot{\theta}(s_i)=a_i,\qquad \ddot{\theta}(s_i+0)=b_i,\qquad i=1,\cdots,m \tag{5.3}$$

as parameters (but $a_0=0$, $b_m=0$ always).

$$
\begin{aligned}
&\text{(i)}\quad \dot\theta(s)+\frac{1}{2}\int_{s_{i-1}}^{s}(s-t)\dot\theta^3(t)\,dt = a_{i-1}+b_{i-1}(s-s_{i-1}), \qquad s_{i-1}\leqq s<s_i,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i=1,\cdots,m,\\
&\qquad \dot\theta(s)=0, \qquad s_m\leqq s\leqq \bar s.\\
&\text{(ii)}\quad a_i = a_{i-1}+b_{i-1}(s_i-s_{i-1})-\frac{1}{2}\int_{s_{i-1}}^{s_i}(s_i-t)\dot\theta^3(t)\,dt, \qquad i=1,\cdots,m.\\
&\text{(iii)}\quad \int_0^{s_i}\cos\theta = p_i^1, \qquad \int_0^{s_i}\sin\theta = p_i^2, \qquad i=1,\cdots,m.
\end{aligned}
\tag{5.4}
$$

Equations (5.4) define implicitly a mapping $\mathcal{G}$ from the space $\mathcal{P}\subset(\mathbb{R}^2)^m$ of configurations P to the space $\mathcal{E}$ of extremal interpolants E. To apply the implicit function theorem we introduce a mapping G on the product space $\mathcal{E}\times\mathcal{P}$ to $\mathrm{NBV}\times\mathbb{R}^m\times\mathbb{R}^{2m}$ as follows. We set

$$\Theta=(\theta;\, s_1,\cdots,s_m;\, a_1,\cdots,a_m;\, b_0,\cdots,b_{m-1}),$$

where

$$\theta\in\mathrm{NBV}_1, \quad s_i\in\mathbb{R}, \quad a_i\in\mathbb{R}, \quad b_i\in\mathbb{R},$$

$$D=\mathrm{NBV}_1\times\Pi(\bar s_i-\bar\varepsilon,\,\bar s_i+\bar\varepsilon)\times\mathbb{R}^m\times\mathbb{R}^m$$

and define a mapping $G=(g,\, r_i,\, q_i)$ with components $g\in\mathrm{NBV}$, $r_i\in\mathbb{R}$, $q_i\in\mathbb{R}^2$, as follows:

$$
\begin{aligned}
&\text{(i)}\quad g(s)=\dot\theta(s)+\frac{1}{2}\int_{s_{i-1}}^{s}(s-t)\dot\theta^3(t)\,dt-a_{i-1}-b_{i-1}(s-s_{i-1}), \qquad s_{i-1}\leqq s<s_i,\\
&\qquad g(s)=\dot\theta(s), \qquad s_m\leqq s\leqq\bar s.\\
&\text{(ii)}\quad r_i=a_i-a_{i-1}-b_{i-1}(s_i-s_{i-1})+\frac{1}{2}\int_{s_{i-1}}^{s_i}(s_i-t)\dot\theta^3(t)\,dt.\\
&\text{(iii)}\quad q_i^1=\int_0^{s_i}\cos\theta-p_i^1, \qquad q_i^2=\int_0^{s_i}\sin\theta-p_i^2.
\end{aligned}
\tag{5.5}
$$

Clearly, (5.4) are equivalent to

$$G(\Theta,P)=0.$$

In particular we have $G(\bar\Theta,\bar P)=0$, since we assume that $\bar\theta$ is the normal representation of the extremal interpolant $\bar E$ for the configuration $\bar P$. We need the Fréchet differential $G'_\Theta(\Theta,P)[\Psi]$, where

$$\Psi=(\psi;\, t_1,\cdots,t_m;\, \alpha_1,\cdots,\alpha_m;\, \beta_0,\cdots,\beta_{m-1})$$

is an increment to Θ. The components of $G'_\Theta(\Theta,P)[\Psi]$ are denoted by g', r_i', q_i'. One finds readily

$$
\begin{aligned}
&\text{(i)}\quad g'(s)=\dot\psi(s)+\frac{3}{2}\int_{s_{i-1}}^{s}(s-t)\kappa^2(t)\dot\psi(t)\,dt-\tfrac{1}{2}\kappa_{i-1}^3(s-s_{i-1})t_{i-1}\\
&\qquad\qquad -\alpha_{i-1}-\beta_{i-1}(s-s_{i-1})+b_{i-1}t_{i-1}, \qquad s_{i-1}\leqq s<s_i\\
&\qquad g'(s)=\dot\psi(s), \qquad s_m\leqq s<\bar s.\\
&\text{(ii)}\quad r_i'=\alpha_i-\alpha_{i-1}-\beta_{i-1}\Delta_i s+(b_{i-1}-\tfrac{1}{2}\Delta_i s\kappa_{i-1}^3)t_{i-1}-\dot\kappa(s_i-0)t_i\\
&\qquad\qquad +\frac{3}{2}\int_{s_{i-1}}^{s_i}(s_i-t)\kappa^2\dot\psi.\\
&\text{(iii)}\quad (q_i^1)'=-\int_0^{s_i}\psi\sin\theta+t_i\cos\theta_i, \quad (q_i^2)'=\int_0^{s_i}\psi\cos\theta+t_i\sin\theta_i.
\end{aligned}
\tag{5.6}
$$

Here we have used the notation

$$\kappa = \dot{\theta}, \quad \theta_i = \theta(s_i), \quad \kappa_i = \dot{\theta}(s_i), \quad \Delta_i s = s_i - s_{i-1}, \tag{5.7}$$

and the relation

$$\frac{1}{2}\int_{s_{i-1}}^{s_i} \dot{\theta}^3 = \ddot{\theta}(s_{i-1}+0) - \ddot{\theta}(s_i - 0) = b_{i-1} - \dot{\kappa}(s_i - 0),$$

which follows from (5.2i) and (5.3).

The continuity of G' near $(\bar{\Theta}, \bar{P})$ is readily ascertained from (5.6).

We can now state the main result of this section.

THEOREM 5.1. *$G'_\Theta(\bar{\Theta}, \bar{P})$ is an isomorphism of*

$$\mathrm{NBV}_1 \times \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m \quad \textit{onto} \quad \mathrm{NBV} \times \mathbb{R}^m \times (\mathbb{R}^2)^m$$

if and only if $(\bar{\Theta}, \bar{P})$ satisfies the following hypothesis:

(A) *The system for the unknown $\psi \in \mathrm{NBV}_1$*:

$$\text{(i)} \qquad \begin{aligned} &\dddot{\psi} + \tfrac{3}{2}\bar{\kappa}^2\dot{\psi} = 0 \quad \text{on } (\bar{s}_{i-1}, \bar{s}_i), \quad i = 1, \cdots, m, \\ &\dot{\psi} = 0 \quad \text{on } (\bar{s}_m, \bar{s}), \end{aligned}$$

$$\text{(ii)} \qquad \Delta_i\dot{\psi} + \Delta_i\dot{\bar{\kappa}} \int_0^{\bar{s}_i} \psi \sin(\bar{\theta} - \bar{\theta}_i) = 0, \qquad i = 1, \cdots, m, \tag{5.8}$$

$$\text{(iii)} \qquad \int_0^{\bar{s}_i} \psi \cos(\bar{\theta} - \bar{\theta}_i) = 0, \qquad i = 1, \cdots, m,$$

has only the trivial solution $\psi = 0$. If (A) *is satisfied then there are neighborhoods $\mathcal{N}_{\bar{\Theta}}$ in D and $\mathcal{N}_{\bar{P}}$ in $(\mathbb{R}^2)^m$ of $\bar{\Theta}$ and $\bar{P}$ respectively and a diffeomorphism Π,*

$$\Pi: \mathcal{N}_{\bar{\Theta}} \xrightarrow{\text{onto}} \mathcal{N}_{\bar{P}}$$

such that $\Theta \in N_\Theta$ defines an extremal P-interpolant E for $P = \Pi(\Theta)$.

Remark 5.1. Explanation of the notation used above and in the following:

$$\bar{\theta}_i = \bar{\theta}(\bar{s}_i), \quad \bar{\kappa}_i = \dot{\bar{\theta}}(\bar{s}_i), \quad \psi_i = \psi(\bar{s}_i),$$

$$\Delta_i\dot{\bar{\kappa}} = \dot{\bar{\kappa}}(\bar{s}_i + 0) - \dot{\bar{\kappa}}(\bar{s}_i - 0), \qquad \Delta_i\dot{\psi} = \dot{\psi}(\bar{s}_i + 0) - \dot{\psi}(\bar{s}_i - 0),$$

in particular $\Delta_0\dot{\bar{\kappa}} = \dot{\bar{\kappa}}(\bar{s}_0 + 0)$, $\Delta_m\dot{\bar{\kappa}} = -\dot{\bar{\kappa}}(\bar{s}_m - 0)$, etc.

Proof. We first demonstrate the injectivity of the bounded linear mapping $G'_\Theta(\bar{\Theta}, \bar{P})$ under hypothesis (A). Thus, assume that for some Ψ

$$G'_\Theta(\bar{\Theta}, \bar{P})[\Psi] = 0.$$

Then we have by (5.6i)

$$\text{(i)} \qquad \begin{aligned} &\dddot{\psi} + \tfrac{3}{2}\bar{\kappa}^2\dot{\psi} = 0 \quad \text{on } (\bar{s}_{i-1}, \bar{s}_i), \qquad i = 1, \cdots, m, \\ &\dot{\psi} = 0 \quad \text{on} \quad (\bar{s}_m, \bar{s}), \end{aligned} \tag{5.9}$$

$$\dot{\psi}(\bar{s}_i - 0) + \frac{3}{2}\int_{\bar{s}_{i-1}}^{\bar{s}_i} (\bar{s}_i - t)\bar{\kappa}^2\dot{\psi} - \tfrac{1}{2}\bar{\kappa}_{i-1}^3\Delta_i\bar{s}t_{i-1} - \alpha_{i-1} - \beta_{i-1}\Delta_i\bar{s} + b_{i-1}t_{i-1} = 0, \tag{5.10}$$

and using (5.6ii)

$$\text{(i)} \qquad \dot{\psi}(\bar{s}_i - 0) - \alpha_i + \dot{\bar{\kappa}}(\bar{s}_i - 0)t_i = 0, \qquad i = 1, \cdots, m. \tag{5.11}$$

Also by (5.6i)

$$\text{(5.11)} \quad \text{(ii)} \qquad \dot{\psi}(\bar{s}_i+0)-\alpha_i+\bar{b}_i t_i=0.$$

Since $\bar{b}_i=\dot{\kappa}(\bar{s}_i+0)$, the last two equations yield

$$\text{(5.12)} \qquad \Delta_i\dot{\psi}+t_i\Delta_i\dot{\kappa}=0, \qquad i=1,\cdots,m.$$

The remaining equations $(q_i^1)'=(q_i^2)'=0$ (see (5.6iii)) are equivalent to

$$\text{(5.13)} \qquad t_i=\int_0^{\bar{s}_i}\psi\sin(\bar{\theta}-\bar{\theta}_i), \qquad i=1,\cdots,m,$$

$$\text{(5.9)} \quad \text{(iii)} \qquad \int_0^{\bar{s}_i}\psi\cos(\bar{\theta}-\bar{\theta}_i)=0, \qquad i=1,\cdots,m.$$

When (5.13) is substituted in (5.12), one obtains

$$\text{(5.9)} \quad \text{(ii)} \qquad \Delta_i\dot{\psi}+\Delta_i\dot{\kappa}\int_0^{\bar{s}_i}\psi\sin(\bar{\theta}-\bar{\theta}_i)=0, \qquad i=1,\cdots,m.$$

By hypothesis (A) the equations (5.9i, ii, iii), which coincide with Equations (5.8), together with $\psi\in\mathrm{NBV}_1$ imply $\psi=0$ on $[0,\bar{s}_m]$. (5.9i) then implies $\psi=0$ also on $[\bar{s}_m,\bar{s}]$. Then by (5.13), $t_i=0$, $i=1,\cdots,m$, and by (5.11), $\alpha_i=0, i=1,\cdots,m$. Finally, by (5.10), $\beta_i=0$, $i=0,\cdots,m-1$, thus $\Psi=0$, hence $G'_\Theta(\bar{\Theta},\bar{P})$ is injective.

Conversely, if hypothesis (A) is not satisfied, i.e., system (5.9i, ii, iii) has a nontrivial solution $\psi\in\mathrm{NBV}_1$, then determine the t_i from (5.13), the α_i from (5.11) and the β_i from (5.10). $\Psi=(\psi;t_1,\cdots,t_m;\alpha_1,\cdots,\alpha_m;\beta_0,\cdots,\beta_{m-1})$ is then a nontrivial solution of $G'_\Theta(\bar{\Theta},\bar{P})[\Psi]=0$, thus $G'_\Theta(\bar{\Theta},\bar{P})$ is not an isomorphism.

To show surjectivity of $G'_\Theta(\bar{\Theta},\bar{P})$ onto $\mathrm{NBV}\times\mathbb{R}^m\times\mathbb{R}^{2m}$, assume we are given $h\in\mathrm{NBV}, u_i\in\mathbb{R}, v_i\in\mathbb{R}^2, i=1,\cdots,m$. We must find $\psi\in\mathrm{NBV}_1$, $t_i\in\mathbb{R}$, $\alpha_i\in\mathbb{R}$, $\beta_i\in\mathbb{R}$, such that (see 5.6)

$$\text{(5.14)} \quad \text{(i)} \qquad g'(s)=h(s), \qquad 0\leqq s<\bar{s},$$

$$\text{(ii)} \qquad r_i'=u_i, \qquad i=1,\cdots,m,$$

$$\text{(iii)} \qquad (q_i^1)'=v_i^1, \qquad (q_i^2)'=v_i^2, \qquad i=1,\cdots,m.$$

In particular, (5.14i) requires (see (5.6i))

$$\text{(5.15)} \quad \text{(i)} \quad \dot{\psi}(\bar{s}_i-0)+\frac{3}{2}\int_{\bar{s}_{i-1}}^{\bar{s}_i}(\bar{s}_i-t)\bar{\kappa}^2\dot{\psi}-\tfrac{1}{2}\bar{\kappa}_{i-1}^3\Delta_i\bar{s}t_{i-1}-\alpha_{i-1}-\beta_{i-1}\Delta_i\bar{s}$$
$$+\bar{b}_{i-1}t_{i-1}=h(\bar{s}_i-0),$$

and (5.14ii) requires (see (5.6ii))

$$\alpha_i-\alpha_{i-1}-\beta_{i-1}\Delta_i\bar{s}+(\bar{b}_{i-1}-\tfrac{1}{2}\Delta_i\bar{s}\bar{\kappa}_{i-1}^3)t_{i-1}-\dot{\kappa}(\bar{s}_i-0)t_i+\frac{3}{2}\int_{\bar{s}_{i-1}}^{\bar{s}_i}(\bar{s}-t)\bar{\kappa}^2\dot{\psi}=u_1.$$

The last two equations imply

$$\text{(5.16)} \quad \text{(i)} \qquad \dot{\psi}(\bar{s}_i-0)-\alpha_i+\dot{\kappa}(\bar{s}_i-0)t_i=h(\bar{s}_i-0)-u_i.$$

(5.14i) also requires

$$\text{(5.16)} \quad \text{(ii)} \qquad \dot{\psi}(\bar{s}_i+0)-\alpha_i+\dot{\kappa}(\bar{s}_i+0)t_i=h(\bar{s}_i+0).$$

We therefore must have

$$(5.17) \qquad \text{(i)} \quad \dot{\psi}(\bar{s}_i+0)=\dot{\psi}(\bar{s}_i-0)-\Delta_i\dot{\kappa}t_i+\Delta_i h+u_i, \qquad i=1,\cdots,m.$$

The last equations (5.14iii) are (see (5.6iii))

$$(5.17) \qquad \text{(ii)} \quad -\int_0^{\bar{s}_i}\psi\sin\bar{\theta}+t_i\cos\bar{\theta}_i=v_i^1, \qquad \int_0^{\bar{s}_i}\psi\cos\bar{\theta}+t_i\sin\bar{\theta}_i=v_i^2.$$

The general solution $\psi\in\mathrm{NBV}_1$ of (5.14i) is the sum of a particular solution and the linear combination of $3m$ functions with the coefficients $\alpha_1,\cdots,\alpha_m,\beta_0,\cdots,\beta_{m-1},t_1,\cdots,t_m$. When this ψ is substituted in (5.17i and ii) a nonhomogeneous system of $3m$ equations for the unknowns α_i,β_i,t_i is obtained. The homogeneous part of this system corresponds to the case $h=0$, $u_i=0$, $v_i=0$, and it has only the trivial solution $\alpha_i=0,\beta_i=0,t_i=0$, as shown in the first part of the proof. This demonstrates the surjectivity of $G'_\Theta(\bar{\Theta},\bar{P})$ and finishes the proof of Theorem 5.1.

The utility of this theorem is illustrated by the fact that it readily implies the following important result.

COROLLARY 5.2. *Suppose* $\bar{P}=\{0,\bar{p}_1,\cdots,\bar{p}_m\}$ *is the ray configuration* $\bar{p}_i=(\bar{s}_i,0)$, $i=1,\cdots,m$ *with the trivial interpolant* $\bar{E}$. *Then the hypothesis* (A) *is satisfied, hence the conclusion of Theorem* 5.1 *holds.*

Proof. $\bar{\theta}=0$ in this case. If we put

$$(5.18) \qquad s=x, \quad \bar{s}_i=x_i, \quad \bar{s}=\bar{x}, \quad y(x)=\int_0^x\psi,$$

then $y(0)=0$ and (5.8) become

$$(5.19) \qquad \begin{aligned} &y^{(4)}=0 \quad \text{on } (x_{i-1},x_i), \qquad i=1,\cdots,m,\\ &y''=0 \quad \text{on } (x_m,\bar{x}),\\ &y''(x_i+0)-y''(x_i-0)=0, \qquad i=1,\cdots,m,\\ &y(x_i)=0, \qquad i=1,\cdots,m, \end{aligned}$$

while $y'\in\mathrm{NBV}_1$, i.e., $y''\in\mathrm{NBV}$, in particular $y''(0)=0$. These are exactly the equations for a natural cubic spline that interpolates the points $(x_i,0)$, $i=0,\cdots,m$. It follows that $y(x)\equiv 0$, and the corollary is proved.

Remark 5.2. We briefly draw a connection between the above corollary and natural cubic spline interpolation. The mapping Γ from the space of configurations P to the space of extremal interpolants E is implicitly defined by $G(\Theta,P)=0$. Perturbation theory looks for a pair $\Theta=\bar{\Theta}+\Psi$, $P=\bar{P}+Z$ ($Z=\{z_1,\cdots,z_m\}$) close to the initial pair $\bar{\Theta},\bar{P}$ for which

$$(5.20) \qquad G'_\Theta(\bar{\Theta},\bar{P})[\Psi]+G'_P(\bar{\Theta},\bar{P})[Z]=0.$$

It is readily seen that this is system (5.9i, ii, iii) except that (5.9ii) is replaced by

$$(5.21) \qquad \text{(i)} \qquad \int_0^{\bar{s}_i}\psi\cos(\bar{\theta}-\bar{\theta}_i)=z_i^2\cos\bar{\theta}_i-z_i^1\sin\bar{\theta}_i.$$

Also, (5.13) is replaced by

$$(5.21) \qquad \text{(ii)} \qquad t_i=\int_0^{\bar{s}_i}\psi\sin(\bar{\theta}-\bar{\theta}_i)+z_i^1\cos\bar{\theta}_i+z_i^2\sin\bar{\theta}_i.$$

Now suppose the initial configuration $\bar{P}$ and the interpolant $\bar{E}$ are the ray configuration

and trivial interpolant, as in Corollary 5.2. Further suppose P is the configuration

$$P = \{(x_0, y_0), (x_1, y_1). \cdots, (x_m, y_m)\}, x_0 = y_0 = 0$$

with the $x_i - \bar{s}_i = z_i^1$ and $y_i = z_i^2$ small. Then (5.21i, ii) become (in the perturbation approximation)

$$\int_0^{x_i} \psi \approx y_i, \qquad \bar{s}_i + t_i \approx x_i; \tag{5.22}$$

moreover,

$$x(s) = \int_0^s \cos\theta = \int_0^s \cos\psi \approx s, \qquad y(s) = \int_0^s \sin\theta = \int_0^s \sin\psi \approx \int_0^s \psi. \tag{5.23}$$

The perturbed extremal E is now the graph of the function $x \to y(x)$, which satisfies (5.19), except that the last equation is replaced by $y(x_i) = y_i$. Thus y is the natural cubic spline that interpolates the data (x_i, y_i), $i = 0, 1, \cdots, m$.

We have shown that the cubic spline interpolant is the result of linearization of extremal interpolation (in the sense of making $\int \kappa^2\, ds$ stationary) near the trivial interpolant for the ray configuration.

Remark 5.3. The differential equation that appears in hypothesis (A) of Theorem 5.1,

$$\dddot{\psi} + \tfrac{3}{2}\kappa^2\dot{\psi} = 0, \tag{5.24 (i)}$$

where $\kappa = \dot{\theta}$ is defined by the equation

$$\ddot{\theta} + \tfrac{1}{2}\dot{\theta}^3 = 0, \tag{5.24 (ii)}$$

can be completely integrated by quadratures alone. Indeed $\psi = 1$ is clearly one integral;

$$\psi_1 = \kappa \tag{5.25 (i)}$$

is another; for $\dddot{\psi}_1 = \theta^{(4)}$ and $\theta^{(4)} + \frac{3}{2}\dot{\theta}^2\ddot{\theta} = 0$. The third one is

$$\psi_2(s) = s\kappa(s). \tag{5.25 (ii)}$$

In the special case where $\kappa(s) \equiv cs$, three linearly independent integrals are

$$1, s, s^2. \tag{5.25 (iii)}$$

In the remainder of this section we discuss the replacement of hypothesis (A) by a simpler condition, which requires only that the value of an explicitly given function (involving many quadratures) at $\bar{s}_m$ be $\neq 0$. For this purpose we introduce a condition on arcs of simple elastica. The arc of $\bar{E}$ from $\bar{s}_{i-1}$ to $\bar{s}_i$ is said to be *ordinary* if it is either straight or

$$\begin{aligned}\delta_i :=\ & (\bar{p}_i^1 - \bar{p}_{i-1}^1)[(\bar{\kappa}_{i-1})^2 \cos\bar{\theta}_i + \dot{\bar{\kappa}}_{i-1}\sin\bar{\theta}_i] \\ & + (\bar{p}_i^2 - \bar{p}_{i-1}^2)[(\bar{\kappa}_{i-1})^2 \sin\bar{\theta}_i - \dot{\bar{\kappa}}_{i-1}\cos\bar{\theta}_i] + \dot{\bar{\kappa}}_{i-1}\sin(\bar{\theta}_{i-1} - \bar{\theta}_i) \neq 0.\end{aligned} \tag{5.26}$$

One sees readily that only exceptionally is such an arc not ordinary. For example, the arc from $\bar{s}_0 = 0$ to $\bar{s}_1$, where $\bar{\kappa}_0 = 0$, $\dot{\bar{\kappa}}_0 \neq 0$, is not ordinary if and only if

$$\bar{p}_1^1 \sin\bar{\theta}_1 - \bar{p}_1^2 \cos\bar{\theta}_1 = 0,$$

i.e., the chord $\overline{p_0 p_1}$ is tangent to E at $\bar{p}_1$.

THEOREM 5.3. *Suppose the extremal $\bar{E}$ consists of ordinary arcs only. Let ψ_0, $\psi_0(0) = 1$, be the solution of system* (5.8), *exclusive of the condition on* $\Delta_m\dot{\psi}$ (ψ_0 *is*

explicitly constructed below). *Then hypothesis* (A) *of Theorem* 5.1 *is satisfied if and only if*

$$\dot{\psi}_0(\bar{s}_m - 0) + \dot{\bar{\kappa}}(\bar{s}_m - 0) \int_0^{\bar{s}_m} \psi_0 \sin(\bar{\theta} - \bar{\theta}_m) \neq 0. \tag{5.27}$$

Proof. Let ϕ_i, χ_i be the integrals of $\ddot{\psi} + \frac{3}{2}\bar{\kappa}^2\psi = 0$ (case $\kappa \not\equiv 0$) for which

$$\begin{aligned} &\phi_i(\bar{s}_{i-1}) = \dot{\phi}_i(\bar{s}_{i-1} + 0) = 0, \qquad \ddot{\phi}_i(\bar{s}_{i-1} + 0) = 1, \\ &\chi_i(\bar{s}_{i-1}) = \ddot{\chi}_i(\bar{s}_{i-1} + 0) = 0, \qquad \dot{\chi}_i(\bar{s}_{i-1} + 0) = 1. \end{aligned} \tag{5.28}$$

These are linear combinations of $1, \psi_1, \psi_2$. One finds easily

$$\begin{aligned} \phi_i &= \rho_i[(\bar{\kappa}_{i-1})^2 - \bar{\kappa}_{i-1}\bar{\kappa} + \dot{\bar{\kappa}}_{i-1}(s - \bar{s}_{i-1})\bar{\kappa}], \\ \chi_i &= \rho_i[-2\bar{\kappa}_{i-1}\dot{\bar{\kappa}}_{i-1} + 2\dot{\bar{\kappa}}_{i-1}\bar{\kappa} + \tfrac{1}{2}(\bar{\kappa}_{i-1})^3(s - \bar{s}_{i-1})\bar{\kappa}], \end{aligned} \tag{5.29 (i)}$$

where we have used the abbreviation

$$\frac{1}{\rho_i} = 2(\dot{\bar{\kappa}}_{i-1})^2 + \tfrac{1}{2}(\bar{\kappa}_{i-1})^4. \tag{5.29 (ii)}$$

We construct ψ_0 successively on the intervals $[\bar{s}_{i-1}, \bar{s}_i]$, $i = 1, \cdots, m$. On $[0, \bar{s}_1]$ we have since $\psi_0(0) = 1$, $\dot{\psi}_0(0) = 0$:

$$\psi_0 = 1 + \ddot{\psi}_0(0)\phi_1 \tag{5.30 (i)}$$

with ψ_0 satisfying condition (5.9iii):

$$\int_0^{\bar{s}_1} \cos(\bar{\theta} - \bar{\theta}_0) + \ddot{\psi}_0(0) \int_0^{\bar{s}_1} \phi_1 \cos(\bar{\theta} - \bar{\theta}_0) = 0. \tag{5.30 (ii)}$$

One finds the last integral to be $\rho_1\delta_1$. Thus, $\ddot{\psi}_0(0)$ is uniquely determined from (5.30ii), and (5.30i) gives ψ_0 on $(0, \bar{s}_1)$. (If $\bar{\kappa} \equiv 0$ on $[0, \bar{s}_1]$ then one finds $\psi_0(s) = 1 - 3(s/\bar{s}_1)^2$). Assume ψ_0 has been determined on $[0, \bar{s}_{i-1}] (i < m)$. Then $\psi_0(\bar{s}_{i-1})$ and $\dot{\psi}_0(\bar{s}_{i-1} - 0)$ are known, hence $\dot{\psi}_0(\bar{s}_{i-1})$ can be found from condition (5.8ii):

$$\dot{\psi}_0(\bar{s}_{i-1}) = \dot{\psi}_0(\bar{s}_{i-1} - 0) + \Delta_{i-1}\dot{\bar{\kappa}} \int_0^{\bar{s}_{i-1}} \psi_0 \sin(\bar{\theta} - \bar{\theta}_{i-1}). \tag{5.31}$$

Then if $\kappa \not\equiv 0$ on $[\bar{s}_{i-1}, \bar{s}_i]$ we have

$$\psi_0 = \psi_0(\bar{s}_{i-1}) + \dot{\psi}_0(\bar{s}_{i-1})\chi_i + \ddot{\psi}_0(\bar{s}_{i-1})\phi_i, \tag{5.32 (i)}$$

and to satisfy condition (5.9iii):

$$\begin{aligned} &\int_0^{\bar{s}_{i-1}} \psi_0 \cos(\bar{\theta} - \bar{\theta}_i) + \int_{\bar{s}_{i-1}}^{\bar{s}_i} [\psi_0(\bar{s}_{i-1}) + \dot{\psi}_0(\bar{s}_{i-1})\chi_i] \cos(\bar{\theta} - \bar{\theta}_i) \\ &\qquad + \ddot{\psi}_0(\bar{s}_{i-1}) \int_{\bar{s}_{i-1}}^{\bar{s}_i} \phi_i \cos(\bar{\theta} - \bar{\theta}_i) = 0. \end{aligned} \tag{5.32 (ii)}$$

One finds, using (5.29i) and

$$\int_{\bar{s}_{i-1}}^{\bar{s}_i} \cos\bar{\theta} = \bar{p}_i^1 - \bar{p}_{i-1}^1, \int_{\bar{s}_{i-1}}^{\bar{s}_i} \sin\bar{\theta} = \bar{p}_i^2 - \bar{p}_{i-1}^2$$

that the last integral in (5.32ii) is $\rho_i\delta_i \neq 0$. Thus $\ddot{\psi}_0(\bar{s}_{i-1})$ is uniquely determined from (5.32ii), and (5.32i) gives ψ_0 on $[\bar{s}_{i-1}, \bar{s}_i]$.

In the omitted case $\bar{\kappa} \equiv 0$ on $[\bar{s}_{i-1}, \bar{s}_i]$, (5.32iii) are replaced by

$$(5.33)\qquad \begin{aligned} &\text{(i)}\quad \psi_0(s) = \psi_0(\bar{s}_{i-1}) + \dot{\psi}_0(\bar{s}_{i-1})(s - \bar{s}_{i-1}) + \tfrac{1}{2}\ddot{\psi}_0(\bar{s}_{i-1})(s - \bar{s}_{i-1})^2, \\ &\text{(ii)}\quad \int_0^{\bar{s}_{i-1}} \psi_0 \cos(\bar{\theta} - \bar{\theta}_i) + \psi_0(\bar{s}_{i-1})(\bar{s}_i - \bar{s}_{i-1}) + \tfrac{1}{2}\dot{\psi}_0(\bar{s}_{i-1})(\bar{s}_i - \bar{s}_{i-1})^2 \\ &\qquad + \tfrac{1}{6}\ddot{\psi}_0(\bar{s}_{i-1})(\bar{s}_i - \bar{s}_{i-1})^2 = 0. \end{aligned}$$

The conclusions remain the same as before.

With ψ_0 found on $[0, \bar{s}_m]$ there remains condition (5.9ii) to be satisfied:

$$(5.34)\qquad \dot{\psi}_0(\bar{s}_m - 0) + \dot{\bar{\kappa}}(\bar{s}_m - 0) \int_0^{\bar{s}_m} \psi_0 \sin(\bar{\theta} - \bar{\theta}_m) = 0.$$

System (5.6) has a nontrivial solution if and only if the constructed integral ψ_0 satisfies (5.34). This proves the theorem.

Example 5.1. To illustrate the utility of the preceding theorem consider the configuration

$$P = \{(0, 0), (a, 0), (a, b)\}$$

with the extremal interpolant $\bar{E}$ whose normal representation $\bar{\theta}$ is defined by

$$(5.35)\qquad \dot{\bar{\theta}}(s) = \begin{cases} 0, & 0 \leqq s \leqq a, \\ \left(\dfrac{\beta}{b}\right)[\sin \bar{\theta}(s)]^{1/2}, & a \leqq s \leqq \bar{s}. \end{cases}$$

Here β and $\bar{s}$ are the definite integrals

$$\beta = \int_0^{\pi} \sin^{1/2}, \qquad \bar{s} = a + \left(\frac{b}{\beta}\right)\int_0^{\pi} \sin^{-1/2}.$$

That $\bar{E}$ does indeed interpolate the point (a, b) follows from

$$\int_a^{\bar{s}} \cos \bar{\theta}\, ds = \int_0^{\pi} \left(\frac{1}{\dot{\bar{\theta}}}\right) \cos t\, dt = \left(\frac{b}{\beta}\right)\int_0^{\pi} \sin^{-1/2} \cos = 0,$$

$$\int_a^{\bar{s}} \sin \bar{\theta}\, ds = \left(\frac{b}{\beta}\right)\int_0^{\pi} \sin \cdot \sin^{-1/2} = b.$$

Using the construction of the preceding theorem, one finds by straightforward computation

$$\psi_0(s) = 1 - 3\left(\frac{s}{a}\right)^2, \qquad 0 \leqq s \leqq a,$$

$$\phi_2(s) = \left(\frac{b}{\beta}\right)^2 (s - a)\dot{\bar{\theta}}(s), \qquad \chi_2(s) = 2\left(\frac{b}{\beta}\right)^2 \dot{\bar{\theta}}(s),$$

$$\psi_0(s) = -2 - \left(\frac{12}{a}\right)\left(\frac{b}{\beta}\right)^2 \dot{\bar{\theta}}(s), \qquad a \leqq s \leqq \bar{s},$$

$$\dot{\psi}_0(\bar{s} - 0) + \dot{\bar{\kappa}}(\bar{s} - 0) \int_0^{\bar{s}} \psi_0 \sin(\bar{\theta} - \bar{\theta}_2) = -\left(\frac{6}{a} + \frac{\beta^2}{b}\right) < 0.$$

By Theorem 5.2 unrestricted perturbation of the 3-point rectangular configuration $\bar{P}$ is possible.

Even for this simple example we were not able to prove this result in the more direct way, by expressing the parameters of the elastica spline in terms of the coordinates of the interpolated configuration.

Example 5.2. Let $\tilde{P}$ be the "hairpin configuration"

$$\tilde{P}=\{(0,0),(0,\delta),(0,0)\},\qquad 2\delta=\int_0^{\pi}(2\sin)^{1/2},$$

with the extremal interpolant $\tilde{E}$ whose normal representation $\tilde{\theta}$ is defined by:

$$\tilde{\theta}(0)=0$$

$$\dot{\tilde{\theta}}(s)=\begin{cases}[2\sin\tilde{\theta}(s)]^{1/2}, & 0\leqq s\leqq\sigma:=\int_0^{\pi}(2\sin)^{-1/2},\\ [-2\sin\tilde{\theta}(s)]^{1/2}, & \sigma\leqq s\leqq 2\sigma.\end{cases}$$

Then $\tilde{\kappa}(0)=\tilde{\kappa}(\sigma)=\tilde{\kappa}(2\sigma)=0$, $\Delta\dot{\tilde{\kappa}}_1=\dot{\tilde{\kappa}}(\sigma+0)-\dot{\tilde{\kappa}}(\sigma-0)=2$. One finds readily:

$$\phi_1(s)=\frac{s}{2}\tilde{\kappa}(s),\qquad \chi_1(s)=\tilde{\kappa}(s),\qquad 0\leqq s\leqq\sigma,$$

$$\phi_2(s)=\frac{s-\bar{s}}{2}\tilde{\kappa}(s),\qquad \chi_2(s)=\tilde{\kappa}(s),\qquad \sigma\leqq s\leqq 2\sigma,$$

$$\psi_0(s)=\begin{cases}1, & 0\leqq s\leqq\sigma,\\ 1+2\delta\tilde{\kappa}(s), & \sigma\leqq s\leqq 2\sigma.\end{cases}$$

Then

$$\begin{aligned}\dot{\psi}_0(2\sigma-0)+\dot{\tilde{\kappa}}(2\sigma-0)\int_0^{2\sigma}\psi_0\sin(\tilde{\theta}-2\pi)&\\ =-2\delta-\left[\int_0^{\sigma}\sin\tilde{\theta}+\int_{\sigma}^{2\sigma}(1+2\delta\tilde{\kappa})\sin\tilde{\theta}\right]&\\ =-2\delta-[\delta-\delta-2\delta]=0.&\end{aligned}$$

By Theorem 5.3 hypothesis (A) is not satisfied, thus it cannot be concluded that the hairpin configuration with the extremal interpolant $\tilde{E}$ permits perturbation. Indeed, one can show directly that if $\tilde{P}$ is replaced by the perturbed configuration $P_\varepsilon=\{(-\varepsilon,0),(0,\delta),(\varepsilon,0)\}$ there exists no extremal interpolant close to $\tilde{E}$ no matter how small $\varepsilon\neq 0$ is. On the other hand, if $\tilde{P}$ is replaced by ${}_\varepsilon P=\{(0,0),(0,\delta),(0,\varepsilon)\}$, which is also close to $\tilde{P}$, then there is an extremal interpolant ${}_\varepsilon E$ near $\tilde{E}$, ${}_\varepsilon E$ coincides with E for the arc from $(0,0)$ to $(0,\delta)$, the remaining arc is the simple elastica joining $(0,\delta)$ and $(0,\varepsilon)$. Thus, we have an example of singular behavior taking place in the perturbation from $\tilde{P}$ to ${}_\varepsilon P$.

6. Special cases of open extremals. We study in some detail in this section extremal P-interpolants for some special configurations.

A. Two-point extremal interpolants. Let P be the configuration $P=\{p_0,p_1\}$, $p_0=0$. For an extremal E with normal representation $\bar{\theta}$, $\bar{\theta}(0)=0$, we have by (3.3):

$$\text{(i)}\qquad \bar{\theta}\in C^1[0,\bar{s}],\qquad \bar{\theta}(0)=\dot{\bar{\theta}}(0)=\dot{\bar{\theta}}(\bar{s})=0,$$

$$\text{(6.1)}\quad\text{(ii)}\qquad \ddot{\bar{\theta}}^2(s)=c^1\cos\bar{\theta}(s)+c^2\sin\bar{\theta}(s),\qquad 0<s<\bar{s},$$

$$\text{(iii)}\qquad \int_0^{\bar{s}}\cos\bar{\theta}=p_1^1,\qquad \int_0^{\bar{s}}\sin\bar{\theta}=p_1^2.$$

(6.1i) implies $c^1 = 0$ and $c^2 \sin \bar{\theta}(\bar{s}) = 0$. Clearly, either $c^2 = 0$, yielding the extremal $\bar{\theta}(s) \equiv 0$, $p_0 = 0$, $p_1 = (\bar{s}, 0)$, or, $c^2 \neq 0$. In this case we write $c^2 = -2/l$. Differentiation of (6.1ii) gives by Proposition 2.4 the differential equation,

$$l\ddot{\bar{\theta}}(s) + \cos \bar{\theta}(s) = 0, \qquad 0 < s < \bar{s}. \tag{6.2}$$

When this equation is integrated over $(0, \bar{s})$ and $\dot{\bar{\theta}}(0) = \dot{\bar{\theta}}(\bar{s}) = 0$ is used, one obtains the first of equations (6.1iii) with $p_1^1 = 0$. From the equation,

$$\left(\frac{l}{2}\right)\dot{\bar{\theta}}^2 = -\sin \bar{\theta},$$

it follows that $-\pi \leqq \bar{\theta} \leqq 0$ if $l > 0$ and $0 \leqq \bar{\theta} \leqq \pi$ if $l < 0$. Since the choice $l < 0$ amounts to a rotation through π of the extremal corresponding to $l > 0$ or to a change in orientation transforming $\bar{\theta}$ into $-\bar{\theta}$, we may assume $l > 0$ and $-\pi \leqq \bar{\theta} \leqq 0$. In this case $p_1^2 = -d := -|p_0 - p_1|$ in the second equation of (6.1iii). We may rewrite (6.1iii):

$$\int_0^{\bar{s}} \cos \bar{\theta} = 0, \qquad \int_0^{\bar{s}} \sin \bar{\theta} = -d. \tag{6.1iii'}$$

From (6.1ii), $\dot{\bar{\theta}}(\bar{s}) = 0$ is equivalent to $\sin \bar{\theta}(\bar{s}) = 0$. Since $\sin \bar{\theta}(s) \leqq 0$ for $0 \leqq s \leqq \bar{s}$ and $\bar{\theta}$ is continuous, there are only two possibilities: $\bar{\theta}(\bar{s}) = 0$ or $\bar{\theta}(\bar{s}) = -\pi$.

Altogether we have shown that (6.1) may be replaced by the simpler system

(6.3)

$$\text{(i)} \qquad \frac{l}{2}\dot{\bar{\theta}}^2(s) = -\sin \bar{\theta}(s), \quad 0 \leqq s \leqq \bar{s}; \quad \bar{\theta}(0) = 0, \quad \bar{\theta}(\bar{s}) = 0 \text{ or } -\pi,$$

$$\text{(ii)} \qquad \int_0^{\bar{s}} \sin \bar{\theta} = -d.$$

If s is interpreted as physical time, $l\theta$ as the displacement along a circle of radius l, then (6.3i) represents the pendulum equation with pendulum length l, unit mass and unit force downward, starting from horizontal position with velocity 0 at $s = 0$ and reaching velocity 0 again at $s = \bar{s}$ when $\bar{\theta} = 0$ or $-\pi$. The pendulum swings from horizontal position $\bar{\theta} = 0$ at $s = 0$ through one or more half-swings to horizontal position at $s = \bar{s}$. The kinetic analogue of the elastica equation was discovered by G. Kirchhoff; see [8, p. 399].

One interpretation of the interpolation condition (6.3ii) is that the time integral of the kinetic energy $\frac{1}{2}(l\dot{\theta})^2$ divided by the maximum kinetic energy, l, is the prescribed "minimum time" d. The length of the pendulum is the main unknown of the problem.

The solution of (6.3i) for various values of l can be derived from the solutions of the same system for $l = 2$. Indeed, the transformation

$$\bar{\theta}(s) = \tilde{\theta}\left(\sqrt{\frac{2}{l}}s\right), \qquad \bar{s} = \sqrt{\frac{l}{2}}\tilde{s} \tag{6.4}$$

converts (6.3i) to the "normalized system"

$$\ddot{\tilde{\theta}}(s) = -\sin \tilde{\theta}(s), \qquad 0 \leqq s \leqq \tilde{s}, \quad \tilde{\theta}(0) = 0, \quad \tilde{\theta}(\tilde{s}) = 0 \text{ or } -\pi. \tag{6.5}$$

The solution of the pendulum equation (6.5) with $\tilde{\theta}(\tilde{s}) = -\pi$ is well known. It is explicitly given by

$$\tilde{\theta}(s) = -\frac{\pi}{2} - 2 \operatorname{arc\,sin}\left[2^{-1/2}\, sn\left(2^{-1/2}\left(s - \frac{\tilde{s}}{2}\right)\right)\right], \qquad 0 \leqq s \leqq \tilde{s}, \tag{6.6}$$

where $2^{-1/2}\tilde{s}$ is the half-period of the Jacobi function $sn(u) = sn(u; 2^{-1/2})$ and arc sin is the branch of the inverse of sin with range $[-\pi/2, \pi/2]$. For $\tilde{s}$ and $\tilde{U} = U(\tilde{\theta}) = \int_0^{\tilde{s}} \dot{\tilde{\theta}}^2$ we find in terms of the complete elliptic integrals of the first and second kind [13]:

$$\text{(6.7)}\quad \begin{aligned} &\text{(i)} \quad \tilde{s} = \int_0^{\pi} \frac{d\theta}{\sqrt{\sin\theta}} = 2\sqrt{2}\int_0^1 \frac{dt}{\sqrt{(1-t^2)\left(1-\frac{t^2}{2}\right)}} = 2\sqrt{2}K(2^{-1/2}), \\ &\text{(ii)} \quad \tilde{U} = \int_0^{\pi} \sqrt{\sin\theta}\, d\theta = 2\sqrt{2}\int_0^1 \sqrt{\frac{1-t^2}{1-\frac{t^2}{2}}} = 2\sqrt{2}[2E(2^{-1/2}) - K(2^{-1/2})]. \end{aligned}$$

The analytic continuation of $\tilde{\theta}$ (also denoted as $\tilde{\theta}$) to all of $\mathbb{R}$ is given by $\tilde{\theta}(s) = -\tilde{\theta}(-s)$ for $s < 0$ and

$$\text{(6.8)}\quad \tilde{\theta}(s) = \begin{cases} \tilde{\theta}(2\tilde{s} - s) & \text{for } \tilde{s} \leqq s \leqq 2\tilde{s}, \\ \tilde{\theta}(s - 2k\tilde{s}) & \text{for } 2k\tilde{s} \leqq s \leqq 2(k+1)\tilde{s}, \quad k = 1, 2, \cdots. \end{cases}$$

It is seen that $\tilde{\theta}_{[0,2\tilde{s}]}$ also solves (6.5), the value at the new boundary point $2\tilde{s}$ being 0. In general $\tilde{\theta}_{[0,k\tilde{s}]}$ $(k = 1, 2, \cdots)$ solves (6.5), with $\tilde{\theta}(k\tilde{s}) = -\pi$ or 0 depending on whether k is odd or even. These are all the solution of (6.5) with free right end-point.

We now use the solution $\tilde{\theta}$ of the normalized problem to express the general solution of the boundary value problem (6.3i) for fixed $\bar{s} > 0$. It is given by

$$\text{(6.9)}\quad \begin{aligned} &\text{(i)} \quad \bar{\theta}(s) = \tilde{\theta}\left(k\frac{\tilde{s}}{\bar{s}}s\right), \qquad 0 \leqq s \leqq \bar{s}, \\ &\text{(ii)} \quad l = 2\left(\frac{\bar{s}}{k\tilde{s}}\right)^2, \end{aligned}$$

where k is any positive integer. To satisfy the remaining condition (6.3ii) we must have

$$-d = \int_0^{\bar{s}} \sin\bar{\theta}(s)\, ds = \int_0^{\bar{s}} \sin\tilde{\theta}\left(k\frac{\tilde{s}}{\bar{s}}s\right) ds = \frac{\bar{s}}{k\tilde{s}}\int_0^{k\tilde{s}} \sin\tilde{\theta}(s)\, ds = -\frac{\bar{s}}{\tilde{s}}\tilde{U};$$

hence

$$\text{(6.10)}\quad \bar{s} = \left(\frac{\tilde{s}}{\tilde{U}}\right)d,$$

independent of k.

We write $\bar{\theta}_k$ for the solution corresponding to $k = 1, 2, \cdots, l_k$ for the corresponding parameter value in (6.3ii) and $\bar{U}_k$ for the value of the curvature functional for $\bar{\theta}_k$. Thus,

$$\text{(6.11)}\quad \begin{aligned} &\text{(i)} \quad \bar{\theta}_k(s) = \tilde{\theta}\left(k\frac{\tilde{U}}{d}s\right), \qquad 0 \leqq s \leqq \bar{s} = \left(\frac{\tilde{s}}{\tilde{U}}\right)d, \\ &\text{(ii)} \quad l_k = 2\left(\frac{d}{k\tilde{U}}\right)^2 = 2\left(\frac{\bar{s}}{k\tilde{s}}\right)^2, \qquad k = 1, 2, \cdots, \\ &\text{(iii)} \quad \bar{U}_k = \int_0^{\bar{s}} \dot{\bar{\theta}}_k^2(s)\, ds = k^2\frac{\tilde{U}^2}{d} = k^2\left(\frac{\tilde{s}}{\bar{s}}\right)\tilde{U}. \end{aligned}$$

All these solutions have the same arc length $\bar{s} = d\tilde{s}/\tilde{U} \approx 2.2d$. Also $\bar{\theta}_k(s) = \bar{\theta}_1(ks)$ for

$0 \leqq s \leqq \bar{s}/k$ and $\bar{\theta}_k(s) = \bar{\theta}_1(ks - \bar{s})$ for $\bar{s}/k \leqq s \leqq 2\bar{s}/k$, etc. Thus the curve represented by $\bar{\theta}_k$ consists of k congruent arcs, all similar (contracted by the factor $1/k$) to $\bar{\theta}_1$. The curve whose normal representation is $\bar{\theta}_k$ is characterized as that arc of the simple elastica whose endpoints are inflection points and which has $k-1$ internal inflection points (it belongs to class $\mathscr{E}_{(k-1)}$ in the notation of § 4). For the above physical interpretation the result means: If the ratio of the time integral of the kinetic energy to the maximum kinetic energy (in a motion from the horizontal to the horizontal position) is to be the fixed number d then the pendulum must have one of the lengths $l_k = 2(d/k\tilde{U})^2$, $k = 1, 2, \cdots$ and the pendulum makes k half-swings in total time $\bar{s} = (\tilde{s}/\tilde{U})d$.

We summarize the results in

PROPOSITION 6.1. *There are countably many extremal P-interpolants E_k interpolating a two-point configuration with $|p_0 - p_1| = d > 0$, one for each $k = 0, 1, 2, \cdots$. E_0 is the trivial ray interpolant; E_k is an arc of the simple elastica with inflection points at the terminals and $k-1$ internal inflection points. The normal representation $\bar{\theta}_k$ of E_k is given by* (6.11), *where $\tilde{\theta}$ is the elliptic function* (6.6). *Each of the E_k, $k = 1, 2, \cdots$ has the same length $\bar{s} = (\tilde{s}/\tilde{U})d$, where $\tilde{s}$ and $\tilde{U}$ are given by* (6.7). *The extremal value of the curvature functional for E_k is $\bar{U}_k = k^2\bar{U}_1$, $k = 0, 1, 2, \cdots$, where $\bar{U}_1 = \tilde{U}^2/d$.*

Remark 6.1. It is shown in [5] that none of the extremals $E_1, E_2, \cdots$ provides a local minimum.

Remark 6.2. Since the simple elastica has no double point, it follows that there exists no extremal interpolant for the "loop" configuration $\{p_0, p_1\}$ with $p_0 = p_1$.

B. Ray and rectangular configurations.

COROLLARY 6.1. *Suppose $P = \{0, p_1, \cdots, p_m\}$ is the "ray" configuration where $p_i = (s_i, 0)$ and $0 < s_1 < \cdots < s_m$. Countably many nontrivial extremal P-interpolants are obtained from the 2-point extremals of Proposition* 6.1 *as follows. Let $\bar{x}_{[0,\bar{s}_1]}$ represent any of the nontrivial extremals for the configuration $\{0, p_1\}$. Define $\bar{x}_{[\bar{s}_1,\bar{s}_2]}$ as one of the $\{p_1, p_2\}$ interpolants of Proposition* 6.1 *or the negative of it so that $\dot{\bar{x}}(\bar{s}_1 - 0) = \dot{\bar{x}}(\bar{s}_1 + 0)$. Continue in this way to the intervals $[\bar{s}_2, \bar{s}_3], \cdots, [\bar{s}_{m-1}, \bar{s}_m]$. The obtained curve represented by $\bar{x}$ is an extremal P-interpolant.*

COROLLARY 6.2. *Suppose $P = \{p_0, p_1, \cdots, p_m\}$ is the "rectangular" configuration where the angle between $\overline{p_{i-1}p_i}$ and $\overline{p_ip_{i+1}}$, $i = 1, \cdots, m-1$ is either* 0 *or* $\pm\pi/2$. *Let $\bar{x}$ represent a P-interpolant such that the segment from p_i to p_{i+1}, $i = 0, \cdots, m-1$, is any of the extremals of Proposition* 6.1 *(including the trivial one) and so that $\dot{\bar{x}}$ is continuous. In this way countably many extremal P-interpolants are obtained for any rectangular configuration P.*

C. Angle-constrained two point extremal interpolants. The boundary conditions

$$\dot{\bar{x}}(0) \cdot (p_1 - p_0) = \gamma_0, \quad \dot{\bar{x}}(\bar{s}) \cdot (p_1 - p_0) = \gamma_1, \qquad 0 \leqq \gamma_i \leqq |p_1 - p_0| \tag{6.12}$$

are added to the problem of § 6A, replacing the zero curvature endpoint conditions.

PROPOSITION 6.2. *There are countably many extremal $\{p_0, p_1\}$-interpolants constrained by conditions* (6.12) *if $\gamma_0 = \gamma_1$.*

Proof. Suppose $\bar{\theta}$ is the normal representation of the sought extremal. It is easily seen that one can normalize $\bar{\theta}$ so that $\bar{\theta}$ satisfies (6.3i), except that the values of $\bar{\theta}$ at 0 and $\bar{s}$ are not 0 and $-\pi$, but given numbers $\theta_0, \theta_1, 0 \geqq \theta_0 \geqq -\pi/2$, $\theta_1 = \theta_0$ or $\theta_1 = \theta_0 - \pi$. It follows that the solution of the problem is a symmetric arc of the curve (6.9i), with k even if $\theta_0 = \theta_1$, k odd if $\theta_0 = \theta_1 + \pi$, scaled and moved so that the terminals are p_0, p_1.

D. Regular configurations. Suppose $P=\{p_0, p_1, \cdots, p_m\}$ is a "regular" configuration, by which we mean each segment $\overline{p_i p_{i+1}}$, $i=0, \cdots, m-1$ is of the length d and makes the same (exterior) angle α, $\pi<\alpha\leqq 2\pi$, with the following segment $\overline{p_{i+1}p_{i+2}}$. We seek an angle-constrained P-interpolant $\bar{x}$ for which

$$\dot{\bar{x}}(0)\cdot(p_1-p_0)=\dot{\bar{x}}(\bar{s}_m)\cdot(p_m-p_{m-1})=d\sin\frac{\alpha}{2}. \tag{6.13}$$

COROLLARY 6.3. *For each regular configuration $P=\{p_0, p_1, \cdots, p_m\}$ there are infinitely many extremal P-interpolants constrained by the condition* (6.13), *which consist of m congruent segments.*

Proof. Let $\bar{z}$ represent one of the infinitely many extremal $\{p_0, p_1\}$-interpolants E_0 of length $\bar{s}_1$, constrained by

$$\dot{\bar{z}}(0)\cdot(p_1-p_0)=\dot{\bar{z}}(\bar{s}_1)\cdot(p_1-p_0)=d\sin\frac{\alpha}{2}$$

and such that $\dot{\bar{z}}(0)\times(p_1-p_0)=-\dot{\bar{z}}(\bar{s}_1)\times(p_1-p_0)$. Let $\bar{x}$ represent the uniquely defined P-interpolant E that extends E_0 by congruent pieces. Then $\bar{x}$ has continuous slope and curvature and satisfies (6.13); hence E is one of the sought extremals.

E. Length-prescribed two-point extremal interpolants. We assume now the length $\bar{s}$ of the extremal $\{p_0, p_1\}$-interpolant $\bar{E}$ is prescribed. Suppose $\bar{\theta}$ is the normal representation of $\bar{E}$. With the proper choice of the coordinate system we conclude, using Proposition 3.2, that $\bar{\theta}$ must satisfy the following conditions (assuming $\bar{s}>d$)

$$\text{(i)}\qquad \dot{\bar{\theta}}(0)=\dot{\bar{\theta}}(\bar{s})=0,$$

$$\text{(ii)}\qquad \frac{l}{2}\dot{\bar{\theta}}^2(s)=-\sin\bar{\theta}(s)+\lambda,\qquad 0\leqq s\leqq\bar{s},\quad l>0,\quad \lambda\in\mathbb{R}, \tag{6.14}$$

$$\text{(iii)}\qquad \int_0^{\bar{s}}\cos\bar{\theta}(s)\,ds=0,\qquad \int_0^{\bar{s}}\sin\bar{\theta}(s)\,ds=-d.$$

Set $\bar{\theta}(0)=\theta_0$ $(-\pi/2\leqq\theta_0<\pi/2)$; then

$$\lambda=\sin\theta_0. \tag{6.15}$$

In particular, $-1\leqq\lambda<1$. We may again interpret (6.14) as describing the motion of a pendulum swinging from some position θ_0 where $\dot{\theta}=0$ in fixed time $\bar{s}$ to another position where $\dot{\theta}=0$ and so that the time integral of the kinetic energy satisfies a certain condition.

By (6.14ii), $\dot{\bar{\theta}}(\bar{s})=0$ if and only if $\bar{\theta}(\bar{s})$ is θ_0 or $-\pi-\theta_0$. The condition $\int_0^{\bar{s}}\cos\bar{\theta}=0$ follows from (6.14ii). Thus, (6.14) may be replaced by the simpler system

$$\text{(i)}\qquad \frac{l}{2}\dot{\bar{\theta}}^2(s)=\sin\theta_0-\sin\bar{\theta}(s),\quad 0\leqq s\leqq\bar{s},\qquad \bar{\theta}(0)=\theta_0,$$

(6.16)

$$\text{(ii)}\qquad \bar{\theta}(\bar{s})=\theta_0 \text{ or } -\pi-\theta_0,\qquad \int_0^{\bar{s}}\sin\bar{\theta}=-d.$$

We make the transformation (6.4) again, to obtain the normalized system

$$\dot{\tilde{\theta}}^2(s)=\sin\theta_0-\sin\tilde{\theta}(s),\quad 0\leqq s\leqq\tilde{s},\quad \tilde{\theta}(0)=\theta_0,\quad \tilde{\theta}(\tilde{s})=-\pi-\theta_0. \tag{6.17}$$

The solution is given by

$$\tilde{\theta}(s) = \tilde{\theta}(s;\theta_0) = -\frac{\pi}{2} - 2\arcsin\left[qsn\left(2^{-1/2}\left(s-\frac{\tilde{s}}{2}\right);q^2\right)\right],$$

(6.18) (i)

$$q = \left|\sin\left(\frac{\pi}{4}+\frac{\theta_0}{2}\right)\right|,$$

where $2^{-1/2}\tilde{s}$ is the half-period of the Jacobi function $sn(u;q)$:

$$\tilde{s} = \tilde{s}(\theta_0) = -\int_{\theta_0}^{-\pi-\theta_0} \frac{d\theta}{\sqrt{\sin\theta_0 - \sin\theta}} = \int_{-\theta_0}^{\pi+\theta_0} \frac{d\theta}{\sqrt{\sin\theta_0+\sin\theta}}$$

(6.18) (ii)

$$= 2\int_{-\theta_0}^{\pi/2} \frac{d\theta}{\sqrt{\sin\theta_0+\sin\theta}}.$$

We observe that $\tilde{s}(\theta_0)$ increases monotonically from 0 to ∞ as θ_0 varies from $-\pi/2$ to $\pi/2$. For $\tilde{d} = -\int_0^{\tilde{s}} \sin\tilde{\theta}$ and $U = \int_0^{\tilde{s}} \tilde{\theta}^2$ we have

(6.18) (iii)
$$\tilde{d} = \tilde{d}(\theta_0) = 2\int_{-\theta_0}^{\pi/2} \frac{\sin\theta}{\sqrt{\sin\theta_0+\sin\theta}}\,d\theta,$$

(6.18) (iv)
$$\tilde{U} = \tilde{U}(\theta_0) = 2\int_{-\theta_0}^{\pi/2} \sqrt{\sin\theta_0+\sin\theta}\,d\theta.$$

We also observe the identity

(6.19)
$$\tilde{s}(\theta_0)\sin\theta_0 + \tilde{d}(\theta_0) = \tilde{U}(\theta_0).$$

The analytic continuation of $\tilde{\theta}$ (also denoted as $\tilde{\theta}$) to all of $\mathbb{R}$ is given, as before, by (6.8). Then we put

(6.20) (i)
$$\bar{\theta}_k(s) = \tilde{\theta}\left(k\frac{\tilde{s}}{\bar{s}}s\right), \qquad 0 \leqq s \leqq \bar{s}, \quad k = 1, 2, \cdots,$$

(6.20) (ii)
$$l_k = 2\left(\frac{\bar{s}}{k\tilde{s}}\right)^2.$$

These quantities still depend on θ_0. θ_0 must be determined so that the last condition

(6.20) (iii)
$$-d = \int_0^{\bar{s}} \sin\bar{\theta}_k(s)\,ds = \int_0^{\bar{s}} \sin\tilde{\theta}\left(\frac{k\tilde{s}s}{\bar{s}}\right)ds = \frac{\bar{s}}{\tilde{s}}\int_0^{\tilde{s}} \sin\tilde{\theta}(s)\,ds = -\frac{\bar{s}}{\tilde{s}(\theta_0)}\tilde{d}(\theta_0)$$

is satisfied. Thus θ_0 is determined from the equation

(6.21)
$$\frac{\tilde{d}(\theta_0)}{\tilde{s}(\theta_0)} = \frac{d}{\bar{s}},$$

where $\tilde{d}$ and $\tilde{s}$ are given in (6.18). As θ_0 varies from $-\pi/2$ to 0, $\tilde{d}(\theta_0)$ increases from 0 to $2\int_0^{\pi/2} \sin^{1/2}\theta\,d\theta$ and then decreases from this value to $-\infty$ as θ_0 varies from 0 to $\pi/2$. Clearly $\tilde{d}(\theta_*) = 0$ for some θ_* between 0 and $\pi/2$ ($\theta_* \approx 40°$). The ratio $\tilde{d}(\theta_0)/\tilde{s}(\theta_0)$ can be seen to decrease monotonically on $[-\pi/2, \theta_*]$ with values in the entire interval $[0, 1]$. Therefore, for any $0 < d < \bar{s}$, there is a unique θ_0 in $[-\pi/2, \theta_*)$ such that (6.21) holds. For this unique θ_0, $\tilde{s} = \tilde{s}(\theta_0)$ is determined according to (6.18ii), then $\bar{\theta}_k$ and l_k from (6.20). Together with $\lambda = \sin\theta_0$, these quantities satisfy the original

system. $\tilde{U}(\theta_0)$ is given by (6.18iv) and the value of U for $\bar{\theta}_k$ by

$$\bar{U}_k := \int_0^{\bar{s}} \dot{\bar{\theta}}_k^2(s)\, ds = k^2 \frac{\tilde{s}(\theta_0)}{\bar{s}} \tilde{U}(\theta_0). \tag{6.22}$$

Altogether we have proved

PROPOSITION 6.3. *There are countably many extremal $\{p_0, p_1\}$-interpolants $\bar{E}_1$, $\bar{E}_2, \cdots$ of prescribed length $\bar{s} > |p_0 - p_1|$. Their normal representations are given explicitly by equations* (6.18) *and* (6.20), *with the angle of inclination θ_0 at p_0 determined from* (6.21). *For the value of the functional U, the relation $\bar{U}_k = k^2 U_1$, $k = 1, 2, \cdots$ holds.*

Remark 6.3. The curves of Proposition 6.3 are subarcs of inflexional elastica (cf. Remark 2.4). For illustrations see [8, p. 404, Figs. 48–53].

Remark 6.4. In the beam interpretation the joint at p_0 exerts a force R whose tangential component is $\sin \theta_0/l$ and whose normal component is $\cos \theta_0/l$. Thus R acts along the line joining p_0 to p_1. The magnitude of the force is $1/l$. For fixed $\bar{s}$ and d, the force R_k in the mode $\bar{E}_k$ has magnitude $k^2 R_1$. For $\theta_0 < 0$ the tangential force on the joints is a pressure, for $\theta_0 > 0$ it is a pull.

7. Examples of closed extremals. Let $P = \{p_0, p_1, \cdots, p_m\}$ be a configuration as in § 2. If $x \in H_2^{\text{reg}}$ is such that $x(t_i) = p_i$, $i = 0, \cdots, m$ for some $0 < t_0 < \cdots < t_m < 1$, and besides $x^{(k)}(0+) = x^{(k)}(1-)$ for $k = 0, 1, 2$, we say x represents an *admissible closed P-interpolant* with knots p_i. Suppose $\bar{x}$ represents an extremal closed P-interpolant, i.e., $\bar{x}$ makes the curvature functional U stationary in the family of admissible closed P-interpolants. Then the graph of $\bar{x}$ is a closed curve which has continuous curvature everywhere, and the curvature is continuously differentiable at all points other than the knots.

In this paragraph we examine four classes of closed extremals:

A. Closed extremals of prescribed length with no knots.

B. Closed symmetric extremals with two knots.

C. Closed extremals for rectangular configurations.

D. Closed extremals for regular polygons.

A. If $\bar{x}$ represents a closed extremal with no knots, of prescribed length $\bar{s} > 0$, parametrized with respect to arc length, $\kappa(s)$ its curvature at s, then one finds, as in Proposition 2.3, that

$$\text{(i)} \qquad \bar{x} \in C^\infty[0, \bar{s}], \quad \bar{x}^{(k)}(0) = \bar{x}^{(k)}(\bar{s}), \qquad k = 0, 1, 2,$$

$$\text{(ii)} \qquad 2\dddot{\bar{x}} + 3\kappa^2 \dot{\bar{x}} - \lambda \dot{\bar{x}} = c, \qquad c \in \mathbb{R}^2, \tag{7.1}$$

$$\text{(iii)} \qquad \lambda = \frac{1}{\bar{s}} \int_0^{\bar{s}} \kappa^2.$$

Conversely if $\bar{x}$ satisfies (7.1) then $\bar{x}$ represents a closed extremal with no knots, of prescribed length $\bar{s}$, parametrized by arc length. For the normal representation $\bar{\theta}$, where $\bar{x}^1(s) = \int_0^s \cos \bar{\theta}$, $\bar{x}^2(s) = \int_0^s \sin \bar{\theta}$, (7.1) gives

$$\text{(i)} \qquad \bar{\theta} \in C^\infty[0, \bar{s}], \qquad \int_0^{\bar{s}} \cos \bar{\theta} = \int_0^{\bar{s}} \sin \bar{\theta} = 0,$$

$$\bar{\theta}(\bar{s}) = \bar{\theta}(0) + 2k\pi, \qquad k = 0, \pm 1, \cdots; \quad \dot{\bar{\theta}}(0) = \dot{\bar{\theta}}(\bar{s}), \tag{7.2}$$

$$\text{(ii)} \qquad \dot{\bar{\theta}}^2 = c^1 \cos \bar{\theta} + c^2 \sin \bar{\theta} + \lambda,$$

$$\text{(iii)} \qquad \lambda = \frac{1}{\bar{s}} \int_0^{\bar{s}} \dot{\bar{\theta}}^2.$$

PROPOSITION 7.1. *For each $k = 1, 2, \cdots$ there exist exactly two closed extremals with no knots of prescribed length $\bar{s} > 0$. These are the circle of radius $\bar{s}/(2k\pi)$ traversed k times and a contracted figure eight configuration traversed k times.*

Proof. We can omit (iii) in (7.2) since it follows from (i) and (ii). We write $A \sin(\bar{\theta} + \alpha)$ with $A \geqq 0$, $\alpha \in T$ for $c^1 \cos\bar{\theta} + c^2 \sin\bar{\theta}$. If $\bar{\theta}$ represents an extremal then $\bar{\theta} - \alpha$ represents the same extremal rotated by the angle α. Therefore, (7.2ii) may be replaced by $\dot{\bar{\theta}}^2 = A \sin\bar{\theta} + \omega_0^2$, where $\omega_0^2 > 0$. ($\omega_0^2 = 0$ means $\lambda = 0$, which is impossible by (7.2iii).) We may also assume $k = 0, 1, \cdots$, in (7.2i). Thus, (7.2) is replaced by

$$\bar{\theta} \in C^\infty[0, \bar{s}], \qquad \int_0^{\bar{s}} \cos\bar{\theta} = \int_0^{\bar{s}} \sin\bar{\theta} = 0,$$

(7.3) (i) $$\bar{\theta}(\bar{s}) = \bar{\theta}(0) + 2k\pi, \qquad k = 0, 1, 2, \cdots; \quad \dot{\theta}(0) = \dot{\bar{\theta}}(\bar{s}),$$

(ii) $$\dot{\bar{\theta}}^2 = A \sin\bar{\theta} + \omega_0^2, \qquad A \geqq 0, \omega_0 > 0.$$

Case 1. $A = 0$. In this case we may assume $\bar{\theta}(0) = 0$. Then $\bar{\theta}(s) = \omega_0 s$ and $\omega_0 \bar{s} = 2k\pi$, hence $\omega_0 = 2k\pi/\bar{s}$. For $k = 1, 2, \cdots$, $\bar{\theta}_k(s) = 2k\pi s/\bar{s}$ satisfies all conditions. $\bar{\theta}_k$ represents a circle of radius $\bar{s}/2k\pi$, traversed k times. The value of U for $\bar{\theta}_k$ is $(2k\pi/\bar{s})^2$.

Case 2. $A > 0$, $\omega_0^2 > A$. We may assume $A = 1$ since if $\bar{\theta}$ is a solution of (7.3) for $A > 0$, of length $\bar{s} > 0$, then $\bar{\theta}$ defined by $\bar{\theta}(A^{-1/2}s)$ is a solution for $A = 1$, $\omega_0^2 = A^{-1}\omega_0^2 > 1$, of length $\tilde{s} = A^{1/2}\bar{s}$. If $\bar{\theta}$ satisfies (7.3) then $\bar{\theta}(s)$ is uniquely defined by

$$s = \int_{\theta_0}^{\bar{\theta}(s)} \frac{d\varphi}{\sqrt{\omega_0^2 + \sin\varphi}}, \tag{7.4}$$

k in (7.3i) must be positive, and ω_0^2 is uniquely defined by

$$\bar{s} = \int_0^{2k\pi} \frac{d\varphi}{\sqrt{\omega_0^2 + \sin\varphi}}.$$

For $\bar{\theta}$ defined in this way we have, after a change of variable,

$$\int_0^{\bar{s}} \sin\bar{\theta}(s)\, ds = \int_{-k\pi}^{k\pi} \frac{\sin\varphi}{\sqrt{\omega_0^2 + \sin\varphi}}\, d\varphi$$

$$= k\left[\int_0^{\pi} \frac{\sin\varphi}{\sqrt{\omega_0^2 + \sin\varphi}}\, d\varphi - \int_0^{\pi} \frac{\sin\varphi}{\sqrt{\omega_0^2 - \sin\varphi}}\, d\varphi\right] < 0,$$

which contradicts (7.3i). Thus no solution exists for $\omega_0^2 > A$.

Case 3. $A > 0$, $\omega_0^2 < A$. We may again assume $A = 1$, thus $\omega_0 < 1$. Conveniently replace θ by $\theta + \pi$, and write $\sin\theta_*$ for ω_0^2, with $0 < \theta_* < \pi/2$. Thus (7.3ii) is replaced by

(7.3) (ii′) $$\dot{\bar{\theta}}^2 = \sin\theta_* - \sin\bar{\theta}.$$

In this case $-\pi - \theta_* \leqq \bar{\theta}(s) \leqq \theta_*$, thus we must have $k = 0$ in (7.3i). Since $\bar{\theta}(s)$ cannot be monotone, we must have $\dot{\bar{\theta}}(s) = 0$ and $\sin\bar{\theta}(s) = \sin\theta_*$ for some s; it is no restriction to assume that this happens for $s = 0$, and $\bar{\theta}(0) = \theta_*$, $\dot{\bar{\theta}}(0) = 0$. As s increases from 0 to some s_*, $\bar{\theta}(s)$ decreases from θ_* to $-\pi - \theta_*$, when $\dot{\bar{\theta}}(s_*) = 0$. Since $-\pi - \theta_* \neq \theta_*$ the curve cannot be closed yet, and as s increases further, $\bar{\theta}(s)$ increases up to the value θ_*, which is attained for some $s = s_{**}$, and $\dot{\bar{\theta}}(s_{**}) = 0$. If we set $\theta_1(t) = \bar{\theta}(2s_* - t)$, we see that θ_1 satisfies (7.3ii′) and $\theta_1(s_*) = \bar{\theta}(s_*)$, hence $\theta_1 = \bar{\theta}$, i.e.,

$$\bar{\theta}(s_* + t) = \bar{\theta}(s_* - t)$$

and, in particular, $s_{**}=2s_*$. Thus the curve obtained is symmetric w.r.t. the point $s=0$. If we put $\theta_2(s)=-\pi-\bar{\theta}(s_*-s)$ then we see that θ_2 satisfies (7.3ii') and $\theta_2(0)=\bar{\theta}(0)=\theta_*$, hence

$$\bar{\theta}(s_*-t)=-\pi-\bar{\theta}(t),$$

the curve obtained is symmetric with respect to the line $\theta=0$ and $\bar{\theta}(s_*/2)=-\pi/2$.

The curve will be closed iff the conditions $\int_0^{s_{**}}\cos\bar{\theta}=0$, $\int_0^{s_{**}}\sin\bar{\theta}=0$ are satisfied. The first equation follows directly from the symmetry of the curve. We are left with

$$(7.3)\qquad \text{(i')}\qquad 0=\int_0^{s_{**}}\sin\bar{\theta}(s)\,ds=4\int_{-\pi/2}^{\theta_*}\frac{\sin\varphi}{\sqrt{\sin\theta_*-\sin\varphi}}\,d\varphi=0.$$

Let the last integral be denoted as $I(\theta_*)$. Clearly $I(0)=-\frac{1}{2}\tilde{U}<0$ and $I(\pi/2)=+\infty$. Thus, there is a value θ_* between 0 and $\pi/2$ for which (7.3i') holds, and it is easily seen that there is only one such value (approximately $\theta_*=40°$). With this value of θ_* we have obtained a closed extremal E^* of length s_{**}. It is an analytic curve, crossing itself at $s=\frac{1}{2}s_{**}$, and consists of 2 congruent loops, each symmetric w.r.t. the same axis (an illustration appears in [8, p. 404] as an example of an inflexional elastica). By proper scaling the curve will have the prescribed length $\bar{s}$. The differential equation for the normal representation of the curve E_1^* is

$$a^2\dot{\theta}_1^{*2}=\sin\theta_*-\sin\theta_1^*,\quad \theta_1^*(0)=\theta_*,\quad \theta_1^*\left(\frac{\bar{s}}{4}\right)=\frac{-\pi}{2}.$$

Thus, the inverse function $\theta\to s(\theta)$ is given by

$$s(\theta)=a\int_\theta^{\theta_*}\frac{d\varphi}{\sqrt{\sin\theta_*-\sin\varphi}},\qquad -\frac{\pi}{2}\leqq\theta\leqq\theta_*$$

where the constant a is determined from

$$\frac{\bar{s}}{4}=a\int_{-\pi/2}^{\theta_*}\frac{d\varphi}{\sqrt{\sin\theta_*-\sin\varphi}},$$

The other extremals $E_2^*, E_3^*, \cdots$ in this sequence are obtained by traversing E_1^* $2, 3, \cdots$ times with scale factor $\frac{1}{2}, \frac{1}{3}, \cdots$, thus their normal representations satisfy

$$\theta_k^*(s)=\theta_1^*(ks),\qquad 0\leqq s\leqq\bar{s}.$$

Case 4. $A_0>0$, $\omega_0^2=A$. In this case the solution of (7.3ii) is monotonically increasing in s but does not attain $\theta_0+2\pi$ for finite s.

Remark 7.1. The restriction of $\theta_1^*(s)$ to $[0,\bar{s}/2]$ represents a length-prescribed extremal (length $=\bar{s}/2$) interpolating the "*loop configuration*" $\{p_0,p_1\}$ with $p_0=p_1$. This is not a closed extremal, although it is a closed curve. Each $\theta_1^*(ks)$, $k=1,2,\cdots;0\leqq s\leqq\bar{s}/2$ can also be considered as the normal representation of such an interpolating extremal. The curvature functional (potential energy) for this extremal is seen to have the value

$$\frac{4k^2}{\bar{s}}\left[\int_{-\pi/2}^{\theta_*}\frac{d\varphi}{\sqrt{\sin\theta_*-\sin\varphi}}\right]^2\sin\theta_*.$$

The k-times traversed circles and contracted figures eight with the point $p_0=p_1$ anywhere on these curves are also length-prescribed loop interpolants but have no knot at the interpolation point.

Remark 7.2. The length-prescribed extremal interpolating the loop configuration with the knot at the interpolation point can also be obtained as the limiting case of the extremal of § 6. E as $d \to 0$. There it was pointed out that $\tilde{d}(\theta_*) = 0$ for θ_* satisfying (see 6.18iii)

$$0 = \int_{-\theta_*}^{\pi/2} \frac{\sin\theta}{\sqrt{\sin\theta_* + \sin\theta}}\, d\theta.$$

Clearly, this is the above condition (7.3i').

B. We turn to the problem of closed extremals $\mathring{E}$ with two knots. We consider only extremals that are symmetric with respect to the line joining the two given knots. We assume $p_0 = (0, 0), p_1 = (0, -d)$ with $d > 0$ are the knots and that $\mathring{x} = (\mathring{x}^1, \mathring{x}^2)$ represents the extremal $\mathring{E}$, parametrized by arc length. If $\mathring{s}$ is the length of $\mathring{E}$ and $\mathring{x}(0) = p_0$, then $\mathring{x}^{(k)}(\mathring{s}) = \mathring{x}^{(k)}(0)$ for $k = 0, 1, 2$, and because of the symmetry, $\mathring{x}(\mathring{s}/2) = p_1$. Thus, we may assume

$$\mathring{x}^1(s) = -\mathring{x}^1(\mathring{s} - s), \quad \mathring{x}^2(s) = \mathring{x}^2(\mathring{s} - s), \qquad 0 \leqq s \leqq \mathring{s},$$

$$\text{(7.5)} \qquad \mathring{x}^1(0) = \mathring{x}^1\left(\frac{\mathring{s}}{2}\right) = \mathring{x}^1(\mathring{s}) = 0, \quad \mathring{x}^2(0) = \mathring{x}^2(\mathring{s}) = 0, \quad \mathring{x}^2\left(\frac{\mathring{s}}{2}\right) = -d,$$

$$\mathring{x}'^1(0) = \mathring{x}'^1(\mathring{s}) = 1, \quad \mathring{x}'^2\left(\frac{\mathring{s}}{2}\right) = \pm 1, \quad \mathring{x}'^2(0) = \mathring{x}'^2\left(\frac{\mathring{s}}{2}\right) = \mathring{x}'^2(\mathring{s}) = 0.$$

If $\mathring{\theta}$ is the normal representation of $\mathring{E}$ then by (7.5) and Proposition 3.1,

$$\text{(i)} \quad \cos\mathring{\theta}(s) = \cos\mathring{\theta}(\mathring{s} - s), \qquad \sin\mathring{\theta}(s) = -\sin\mathring{\theta}(\mathring{s} - s), \qquad 0 \leqq s \leqq \mathring{s},$$

$$\text{(ii)} \quad \int_0^{\mathring{s}/2} \cos\mathring{\theta} = 0, \quad \int_0^{\mathring{s}/2} \sin\mathring{\theta} = -d, \quad \int_0^{\mathring{s}} \cos\mathring{\theta} = 0, \quad \int_0^{\mathring{s}} \sin\mathring{\theta} = 0,$$

$$\text{(7.6)} \quad \text{(iii)} \qquad \mathring{\theta}(0) = 0, \quad \mathring{\theta}\left(\frac{\mathring{s}}{2}\right) = j\pi, \quad \mathring{\theta}(\mathring{s}) = 2j\pi, \quad j = 0 \text{ or } -1,$$

$$\text{(iv)} \qquad \mathring{\theta}'(0) = \mathring{\theta}'\left(\frac{\mathring{s}}{2}\right) = \mathring{\theta}'(\mathring{s}) = 0,$$

$$\text{(v)} \qquad \mathring{\theta}'^2(s) = \begin{cases} \lambda_1^1 \cos\mathring{\theta}(s) + \lambda_1^2 \sin\mathring{\theta}(s), & 0 \leqq s \leqq \dfrac{\mathring{s}}{2}, \\ \lambda_2^1 \cos\mathring{\theta}(s) + \lambda_2^2 \sin\mathring{\theta}(s), & \dfrac{\mathring{s}}{2} \leqq s \leqq \mathring{s}. \end{cases}$$

$s = \mathring{s}/2$ in (v) gives $\lambda_1^1 = \lambda_2^1 = 0$; substitution of (i) in (v) gives $\lambda_1^2 = -\lambda_2^2 = -\lambda$. Thus (7.6v) becomes

$$\text{(7.6)} \quad \text{(v')} \qquad \mathring{\theta}'^2(s) = \begin{cases} -\lambda \sin\mathring{\theta}(s), & 0 < s \leqq \dfrac{\mathring{s}}{2}, \\ +\lambda \sin\mathring{\theta}(s), & \dfrac{\mathring{s}}{2} \leqq s \leqq \mathring{s}. \end{cases}$$

When the conditions $\mathring{\theta}(0) = 0$, $\mathring{\theta}(\mathring{s}/2) = 0$ or $-\pi$, $\mathring{\theta}(0) = \mathring{\theta}(\mathring{s}/2) = 0$ are taken together with (7.6v'), it is seen that $\mathring{\theta}_{[0,\bar{s}/2]}$ is one of the functions θ_k of § 6, with $\bar{s}$ replaced by $\mathring{s}/2$. The symmetry condition gives for $\mathring{\theta}_{[\mathring{s}/2,\mathring{s}]}$:

$$\mathring{\theta}(s) = -2j\pi - \mathring{\theta}(\mathring{s} - s), \qquad \frac{\mathring{s}}{2} \leqq s \leqq \mathring{s}.$$

Thus we have found all solutions of system (7.6) and have proved

PROPOSITION 7.2. *There are countably many closed extremals* $\mathring{E}_1, \mathring{E}_2, \cdots$ *with normal representation* $\mathring{\theta}_1, \mathring{\theta}_2, \cdots$ *with two knots which are symmetric with respect to the line joining the knots.* $\mathring{E}_k$ *is obtained from the open extremal* E_k *of Proposition* 6.1 *by reflection at the line joining the knots. Each* $\mathring{E}_k$ *has the same length* $2\mathring{s}$, *where* $\mathring{s}$ *is the length of the open* E_k. *The value of the curvature functional* U *for the extremal* $\mathring{E}_k$ *is* $2k^2\bar{U}_1$ *where* $\bar{U}_1$ *is the value for the open* E_1.

C. In this section we consider rectangular configurations as defined in Corollary 6.1.

PROPOSITION 7.3. *Let* $P=\{p_0, \cdots, p_m, p_0\}$ *be a rectangular configuration as in Corollary* 6.1. *There exist countably many closed extremals with knots at* $p_0, \cdots, p_m$.

Proof. Since P is closed there must be an even number of right angles between consecutive segments $\overline{p_{i-1}p_i}, \overline{p_ip_{i+1}}$. To connect p_i to p_{i+1} we use either the trivial extremal or one of the 2-point open extremals of Proposition 6.1, with the proviso that we switch from one class of extremals to the other if the angle at p_i is a right angle, otherwise (if the angle is 0) no switch is made. It is easy to see that infinitely many closed P-interpolants with continuous curvature everywhere can be obtained in this way.

D. Let $p_1, \cdots, p_m$ be the vertices of a regular polygon ordered as they come when the polygon is traversed counterclockwise. Define p_i for $i>m$ by periodicity: $p_i = p_{i-m}$. Let $P_{m,k}=\{p_1, p_{1+k}, \cdots, p_{1+mk}\}$ for $k=1, 2, \cdots$. $P_{m,k}$ is a configuration of the kind that Corollary 6.3 applies to, and if the construction used there is applied to $P_{m,k}$ one obtains closed extremals. Thus we have

PROPOSITION 7.4. *For each regular configuration* $P_{m,k}$ *as described above there are infinitely many closed extremal P-interpolants, each composed of congruent segments. For each* k, *there is precisely one such extremal whose intersection with the polygonal path connecting the points of* $P_{m,k}$ *is precisely* $P_{m,k}$. *For* $k=1$, *this extremal* E_m^* *circumscribes the polygon counterclockwise and its representation* x_m^* *satisfies*

$$(p_{i+1}-p_i)\cdot \dot{x}_m^*(s_i) = |p_{i+1}-p_i| \cos\frac{\pi}{m}, \qquad i=1, \cdots, m.$$

If the polygon is inscribed in a unit circle, each of the m *arcs of* E_m^* *may be expressed in terms of the inverse of its normal representation*:

$$(7.7)\quad \text{(i)}\quad s_m^*(\theta) = \frac{2\sin(\pi/m)F(\frac{1}{2}\sqrt{2}, \psi)}{2E(\frac{1}{2}\sqrt{2}, \beta_m) - F(\frac{1}{2}\sqrt{2}, \beta_m)}, \qquad \frac{\pi}{2}-\frac{\pi}{m} \leqq \theta \leqq \frac{\pi}{2}+\frac{\pi}{m},$$

where $\cos\psi = \sqrt{\sin\theta}$ *and* $\cos\beta_m = \sqrt{\cos\pi/m}$. *In this case, the length* s_m^* *and the energy* U_m^* *of* E_m^* *are given explicitly by*

$$(7.7)\quad \text{(ii)}\qquad s_m^* = \frac{2m\sin(\pi/m)F(\frac{1}{2}\sqrt{2}, \beta_m)}{2E(\frac{1}{2}\sqrt{2}, \beta_m) - F(\frac{1}{2}\sqrt{2}, \beta_m)},$$

and

$$(7.7)\quad \text{(iii)}\qquad U_m^* = \frac{4m}{\sin(\pi/m)}[2E(\tfrac{1}{2}\sqrt{2}, \beta_m) - F(\tfrac{1}{2}\sqrt{2}, \beta_m)]^2.$$

Finally,

$$(7.8)\qquad \frac{s_m^*(\theta)}{\theta} \to 1 \quad \text{as } m\to\infty,$$

so that the extremals have the unit circle as limit.

Proof. We sketch the verification of (7.7). Starting from the differential equation

$$\gamma_m^2 \dot{\theta}^2 = \sin\theta,$$

we obtain

$$s_m^*(\theta) = \gamma_m \int_{\alpha_m}^{\theta} \frac{d\varphi}{\sqrt{\sin\varphi}}, \qquad \alpha_m \leqq \theta \leqq \pi - \alpha_m, \quad \alpha_m = \frac{\pi}{2} - \frac{\pi}{m}.$$

Since the distance between adjacent points of $P_{m,1}$ is $2\cos\alpha_m$, the constant γ_m is determined from

$$\gamma_m = \frac{\cos\alpha_m}{\int_{\alpha_m}^{\pi/2} \sqrt{\sin\varphi}\, d\varphi}$$

$$= \frac{\sin(\pi/m)}{\sqrt{2}[2E(\frac{1}{2}\sqrt{2}, \beta_m) - F(\frac{1}{2}\sqrt{2}, \beta_m)]}.$$

This gives (7.7i). (7.7ii) is immediate and (7.7iii) follows from

$$U_m^* = 2m \int_0^{s_m^*/2m} \dot{\theta}^2\, ds = \frac{2m}{\gamma_m} \int_{\alpha_m}^{\pi/2} \sqrt{\sin\varphi}\, d\varphi.$$

The calculation (7.8) is routine.

Remark 7.3. It is shown in [5] that the extremals E_m^* are stable, i.e., they provide a local minimum for the curvature functional.

Appendix. Let $P = \{p_0, \cdots, p_m\}$, be given and let $L_0 = \sum_{i=1} |p_{i+1} - p_i|$. We assume that P is not collinear.

THEOREM. *For every $L > L_0$ there exists a length-prescribed extremal P-interpolant of length L satisfying Definition* 2.2.

Proof. The existence of a function $\bar{x}$, parametrized with respect to arc length, for which

$$U(\bar{x}) = \min\{U(x): x \text{ is an admissible } P\text{-interpolant of length } L\}$$

is demonstrated in [6]. (The modification required for the ordering of the points in P is trivial.) If X denotes the closed subspace of $H_2[0, L]$ consisting of those functions vanishing at the knots,

$$0 \leqq \bar{s}_0 < \bar{s}_1 < \cdots < \bar{s}_m \leqq L,$$

of $\bar{x}$, i.e., at those points $\bar{s}_i$ for which $\bar{x}(\bar{s}_i) = p_i$, $i = 0, \cdots, m$, define f to be the mapping of X into $\mathbb{R}$ obtained by $f(y) = U(\bar{x} + y)$ and let H be the function such that $H(y) = S(\bar{x} + y) - L$, where S is the usual length functional (cf. (2.13ii)). Clearly,

$$f(0) = \min\{f(y): H(y) = 0\},$$

and $H'(0)$ is surjective, since P is assumed noncollinear. The result follows from the Lagrange multiplier rule (see, e.g., [9, Thm. 1, p. 243]).

REFERENCES

[1] S. ANTMANN, *The shape of buckled nonlinearly elastic rings*, ZAMP, 21 (1970), pp. 422–438.

[2] G. BIRKHOFF AND C. DE BOOR, *Piecewise polynomial interpolation and approximation*, Approximation of Functions, H. L. Garabedian, editor, Elsevier, New York and Amsterdam, 1965, pp.164–190.

[3] S. FISHER AND J. JEROME, *Stable and unstable elastica equilibrium and the problem of minimum curvature*, J. Math. Anal. Appl., 53 (1976), pp. 367–376.
[4] M. GOLOMB, *Stability of interpolating elastica*, Transactions of the 24th Conference of Army Mathematicians. ARO-Report 1, 1979, pp. 301-350.
[5] ———, *Stability of interpolating elastica*, MRC Technical Summary Report #1852, University of Wisconsin, Madison, May 1978.
[6] J. JEROME, *Smooth interpolating curves of prescribed length and minimum curvature*, Proc. Amer. Math. Soc., 51 (1975), pp. 62–66.
[7] E. LEE AND G. FORSYTHE, *Variational study of nonlinear spline curves*, SIAM Rev., 15 (1973), pp. 120–133.
[8] A. LOVE, *The Mathematical Theory of Elasticity*, 4th ed., Cambridge Univ. Press, London, 1927.
[9] D. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[10] M. MALCOLM, *Nonlinear spline functions*, SIAM J. Numer. Anal., 14 (1977), pp. 254–279.
[11] L. NIRENBERG, *Topics in Nonlinear Functional Analysis*, Courant Institute of Mathematical Sciences, New York University, New York, 1974.
[12] I. TADJBAKHSH AND F. ODEH, *Equilibrium states of elastic rings*, J. Math. Anal. Appl., 18 (1967), pp. 59–74.
[13] *Handbook of Mathematical Functions*, National Bureau of Standards Applied Mathematics Series, 55, 1964.

SIAM J. MATH. ANAL.
Vol. 13, No. 3, May 1982

0036-1410/82/1303-0006 $01.00/0

RESOLVENT FORMULAS FOR A VOLTERRA EQUATION IN HILBERT SPACE*

RALPH W. CARR† AND KENNETH B. HANNSGEN‡

Abstract. Let $\mathbf{y}(t, \mathbf{x}, \mathbf{f})$ denote the solution of the Cauchy problem

$$\mathbf{y}'(t)+\int_0^t [d+a(t-s)]\mathbf{L}\mathbf{y}(s)\,ds=\mathbf{f}(t), \qquad t\geqq 0, \qquad \mathbf{y}(0)=\mathbf{x},$$

where $d\geqq 0$ and $\mathbf{L}$ is a self-adjoint densely defined linear operator on a Hilbert space $\mathscr{H}$ with $\mathbf{L}\geqq\lambda_1\mathbf{I}$. Let $\mathbf{U}(t)\mathbf{x}=\mathbf{y}(t, \mathbf{x}, \mathbf{0})$, $\mathbf{V}=\mathbf{U}'$. By analyzing a related scalar equation with parameter, we find sufficient conditions on the kernel a in order that $\int_0^\infty \|\mathbf{V}(t)L^{-\gamma}\|\,dt<\infty$ $(\gamma>0)$. These results and certain resolvent formulas can be used to study the asymptotic behavior of the solution $\mathbf{y}(t, \mathbf{x}, \mathbf{f})$ as $t\to\infty$. An application to a semilinear integro-partial differential equation is presented.

1. Introduction. We continue our study, begun in [2], of the nonhomogeneous linear equation

$$\mathbf{y}'(t)+\int_0^t [d+a(t-s)]\mathbf{L}\mathbf{y}(s)\,ds=\mathbf{f}(t), \qquad t\geqq 0,$$
$$\mathbf{y}(0)=\mathbf{y}_0\in\mathscr{H}, \qquad '=\frac{d}{dt}, \tag{1.1}$$

where $\mathbf{L}$ is a positive self-adjoint linear operator defined on a dense subspace $\mathscr{D}$ of the Hilbert space $\mathscr{H}$. The kernel $d+a(t)$ satisfies

$$a\in L^1_{\mathrm{loc}}(\mathbb{R}^+, \bar{\mathbb{R}}^+)(\mathbb{R}^+=(0,\infty), \bar{\mathbb{R}}^+=[0,\infty)); \ a \text{ is nonincreasing and convex}$$
$$\text{with } a(\infty)=0<a(0+)\leqq\infty, \text{ and } d\geqq 0, \tag{1.2}$$

and $\mathbf{f}$ belongs to $\mathscr{B}^1_{\mathrm{loc}}(\bar{\mathbb{R}}^+, \mathscr{H})$, the class of locally Bochner integrable functions from $\bar{R}^+$ to $\mathscr{H}$.

Let $u(t, \lambda)$ denote the solution of the real equation

$$u'(t)+\lambda\int_0^t [d+a(t-s)]u(s)\,ds=0, \qquad u(0)=1; \tag{1.3}$$

define $v=\partial u/\partial t$,

$$\mathbf{U}(t)=\int_{\mathbb{R}} u(t,\lambda)\,d\mathbf{E}_\lambda, \qquad \mathbf{V}(t)=\int_{\mathbb{R}} v(t,\lambda)\,d\mathbf{E}_\lambda,$$

where $\{\mathbf{E}_\lambda\}$ is the spectral family corresponding to $\mathbf{L}$. In [2] we established the resolvent formula

$$\mathbf{y}(t)=\mathbf{U}(t)\mathbf{y}_0+\int_0^t \mathbf{U}(t-s)\mathbf{f}(s)\,ds \tag{1.4}$$

* Received by the editors June 8, 1981. The work of these authors was sponsored in part by the United States Army under Contract No. DAAG29-80-C-0041 and by the National Science Foundation under Grant Nos. MCS77-28436 and MCS78-27618.

† Department of Mathematics and Computer Science, St. Cloud State University, St. Cloud, Minnesota 56301.

‡ Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

for the solution of (1.1), and we gave sufficient conditions for

$$\int_0^\infty \|\mathbf{U}(t)\|\,dt<\infty. \tag{1.5}$$

In particular, (1.5) holds if $-a'$ is convex. (See Theorem A after Theorem 2.4 below; here and below we use the norm symbol for a space to indicate the operator norm for linear operators from that space to itself.)

We are principally concerned here with $\mathbf{V}$, the formal derivative of $\mathbf{U}$. $\mathbf{V}$ can be used with (1.4) to express $\mathbf{y}'(t)$, and it appears in the alternate resolvent formula

$$\mathbf{y}(t)=\mathbf{F}(t)+\int_0^t \mathbf{V}(t-s)\mathbf{F}(s)\,ds \tag{1.6}$$

for the integrated version of (1.1), that is,

$$\mathbf{y}(t)+\int_0^t [(t-s)\,d+A(t-s)]\mathbf{L}\mathbf{y}(s)\,ds=\mathbf{F}(t), \tag{1.7}$$

where $A(t)=\int_0^t a(s)\,ds$, $\mathbf{F}(t)=\int_0^t \mathbf{f}(s)\,ds+\mathbf{y}_0$.

Estimate (1.5), with $\mathbf{V}$ in place of $\mathbf{U}$, is always false (see Corollary 2.1 below). Our main results, Theorems 2.3 and 2.4, contain the following:

THEOREM 1.1. *Let* (1.2) *hold, and assume that* $-a'$ *is convex. Then*

$$t\|\mathbf{V}(t)\mathbf{L}^{-1/2}\| \text{ is bounded on } \mathbb{R}^+, \text{ and } \int_0^\infty \|\mathbf{V}(t)\mathbf{L}^{-1/2}\|\,dt<\infty. \tag{1.8}$$

The conditions of Theorem A for (1.5) do imply

$$\int_0^\infty \|\mathbf{V}(t)\mathbf{L}^{-1}\|\,dt<\infty. \tag{1.9}$$

Estimates (1.8) and (1.9) can be used with (1.4) and (1.6) to study the asymptotic behavior of $\mathbf{y}(t)$ under various assumptions on the forcing term.

A variant of (1.1) is

$$\begin{aligned}&\mathbf{z}'(t)+\int_0^t [d+a(t-s)][\mathbf{L}\mathbf{z}(s)+\mathbf{g}(s)]\,ds=\mathbf{f}(t), \qquad t\geqq 0,\\ &\mathbf{z}(0)=\mathbf{z}_0,\end{aligned} \tag{1.10}$$

with $\mathbf{g}:\mathbb{R}^+\to\mathscr{H}$. Proceeding formally from (1.4) and the formal identity

$$\mathbf{V}(t)=-\int_0^t [d+a(t-s)]\mathbf{L}\mathbf{U}(s)\,ds,$$

we obtain

$$\mathbf{z}(t)=\mathbf{U}(t)\mathbf{z}_0+\int_0^t \mathbf{U}(t-s)\mathbf{f}(s)\,ds+\int_0^t \mathbf{V}(t-s)\mathbf{L}^{-1}\mathbf{g}(s)\,ds. \tag{1.11}$$

In § 3 we state a theorem justifying (1.11), and we use it to study the semilinear equation

$$\begin{aligned}&\mathbf{y}'(t)+\int_0^t [d+a(t-s)][\mathbf{L}\mathbf{y}(s)+\mathbf{N}\mathbf{y}(s)]\,ds=\mathbf{f}(t),\\ &\mathbf{y}(0)=\mathbf{y}_0.\end{aligned} \tag{1.12}$$

Here $\mathbf{N}$ is a nonlinear operator with

$$\mathbf{N}(\mathbf{0}) = \mathbf{0}, \tag{1.13}$$

$$\sup_{\|\mathbf{x}_1\|_{\mathscr{D}_1}, \|\mathbf{x}_2\|_{\mathscr{D}} \leqq \Delta} \|\mathbf{N}\mathbf{x}_1 - \mathbf{N}\mathbf{x}_2\|_{\mathscr{D}_1} \leqq \varepsilon(\Delta)\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathscr{D}},$$

where $\varepsilon: (0, \alpha) \to \mathbb{R}^+$ and $\varepsilon \to 0$ as $\Delta \to 0$,

$$\|\mathbf{x}\|_{\mathscr{D}}^2 = \|\mathbf{x}\|^2 + \|\mathbf{L}\mathbf{x}\|^2, \qquad \|\mathbf{x}\|_{\mathscr{D}_1}^2 = \|\mathbf{x}\|^2 + \|\mathbf{L}^{1/2}\mathbf{x}\|^2.$$

We also give an example of an integro-partial differential equation of the form (1.12), to which our result applies.

The spectrum of $\mathbf{L}$ is contained in a closed subinterval of $\mathbb{R}^+$; without loss of generality, we take this interval to be $[1, \infty)$. Then for $0 \leqq \gamma < \infty$,

$$\begin{aligned} \|\mathbf{V}(t)\mathbf{L}^{-\gamma}\| &\leqq \sup_{1 \leqq \lambda < \infty} |v(t, \lambda)|\lambda^{-\gamma} \equiv v_\gamma(t), \\ \|\mathbf{V}(t)\mathbf{L}^{-\gamma}\|_{\mathscr{D}} &\leqq v_\gamma(t). \end{aligned} \tag{1.14}$$

We shall develop estimates for v_γ from (1.3) and deduce estimates such as (1.9) from (1.14).

In § 2 we state our results from v_γ; they are proved in §§ 4 through 8. In particular, § 8 contains a correction for the proof of [2, Lemma 5.2]. We discuss the operator $\mathbf{V}$ and (1.11) and (1.12) in § 3; proofs follow in § 9.

2. Statement of results for v_γ. Throughout this paper it is assumed that $d + a(t)$ satisfies (1.2). We define

$$A(t) = \int_0^t a(s)\, ds, \qquad A_1(t) = \int_0^t s\, a(s)\, ds,$$

$$\hat{a}(\tau) = \int_0^\infty a(t)\, e^{-i\tau t}\, dt \equiv \varphi(\tau) - i\tau\theta(\tau), \qquad \tau > 0$$

(with φ and θ real; note that $\hat{a}$ is continuous),

$$D(\tau) = D(\tau, \infty) = \hat{a}(\tau) - i\, d\tau^{-1}, \qquad D(\tau, \lambda) = D(\tau) + i\tau\lambda^{-1}.$$

Formally, the Fourier transform of $v(t, \lambda)$ (defined to be zero for $t < 0$) is given by

$$\hat{v}(\tau, \lambda) = \frac{-D(\tau)}{D(\tau, \lambda)}, \tag{2.1}$$

so $v(\cdot, \lambda) \notin L^1(\mathbb{R}^+)$ if $D(\tau, \lambda) = 0$ for some τ. By [4], $\varphi(\tau) \geqq 0$; moreover, $\varphi(\tau) > 0$ $(\tau > 0)$ unless $a(t)$ is piecewise linear with changes of slope only at integral multiples of a fixed number t_0 (taken as large as possible) and τ is an integral multiple of $2\pi/t_0$. In all other cases, $D(\tau, \lambda) \neq 0$ $(\tau > 0)$; then the hypotheses of [15, Theorem 2] hold, and $v(\cdot, \lambda) \in L^1(\mathbb{R}^+)$ and (2.1) holds. Throughout this paper, we restrict ourselves to this case by assuming

$$\varphi(\tau) > 0, \qquad \tau > 0. \tag{2.2}$$

Estimates for v_γ depend crucially on the size of $\hat{v}(\tau, \lambda)$ when $\operatorname{Im} D(\tau, \lambda) = \tau[\lambda^{-1} - \theta(\tau) - d\tau^{-2}]$ is zero. Choose and fix $t_1 > 0$ with $a(t_1) > 0$, and let $\rho = 6/t_1$. We showed in [2] that $\theta \downarrow 0 (\tau \uparrow \infty)$ and that the equation

$$\lambda^{-1} - \theta(\omega) - d\omega^{-2} = 0 \tag{2.3}$$

defines a continuous, strictly increasing function $\omega(\lambda)$ on the interval $[\lambda_0, \infty)$, where

$$\lambda_0 = \max\{1, [\theta(\rho) + d\rho^{-2}]^{-1}\}.$$

We extend ω to $[1, \infty)$ if necessary by defining $\omega(\lambda) = \rho (1 \leqq \lambda \leqq \lambda_0)$. We showed in [2, (4.3), (4.24), (4.27)] that

$$\tfrac{1}{5}A_1(\tau^{-1}) \leqq \theta(\tau) \leqq 12A_1(\tau^{-1}), \qquad \tau > 0, \tag{2.4}$$

$$10\omega^2 \geqq a(t_1)\lambda, \qquad \lambda \geqq 1, \tag{2.5}$$

$$\tfrac{1}{5}A_1(\omega^{-1}) \leqq \lambda^{-1} \leqq C_1 A_1(\omega^{-1}), \qquad \lambda \geqq 1, \tag{2.6}$$

where $C_1 = \lambda_0[12 + (2d/a(t_1))] \geqq 12$. (We shall often suppress λ as in (2.5) and (2.6).)

For $\lambda \geqq \lambda_0$, we then have

$$\lambda^{-\gamma}\int_0^\infty |v(t,\lambda)|\,dt \geqq \left(\frac{\theta(\omega)}{60}\right)^\gamma |\hat{v}(\omega,\lambda)| \geqq \left(\frac{\theta(\omega)}{60}\right)^\gamma \frac{\omega\theta(\omega)}{\varphi(\omega)}.$$

This proves our first result.

THEOREM 2.1. *Let* (1.2) *and* (2.2) *hold, and let* $\gamma > 0$. *If* $v_\gamma \in L^1(\mathbb{R}^+)$, *then*

$$\sup_{\rho \leqq \tau < \infty} \frac{\tau[\theta(\tau)]^{1+\gamma}}{\varphi(\tau)} < \infty. \tag{2.7}$$

Suppose, in particular, that $a(0+) < \infty$. From (2.6) we see that

$$\tfrac{1}{10}a(t_1)\lambda \leqq \omega^2 \leqq \tfrac{1}{2}C_1 a(0+)\lambda.$$

In this case, for $\gamma = \frac{1}{2}$, (2.7) is equivalent to

$$\sup_{0 < \tau < \infty} \frac{1+\tau^2}{\varphi(\tau)} < \infty; \tag{2.8}$$

that is, a is strongly positive.

To find upper bounds for v_γ, we first define $\sigma = \sigma(\lambda)$ to be the unique solution of

$$\sigma^{-1}A(\sigma^{-1}) = \lambda^{-1}. \tag{2.9}$$

Then $\sigma : \mathbb{R}^+ \to \mathbb{R}^+$ is strictly increasing, since $\alpha(t) \equiv tA(t)$ is strictly increasing. Using (2.6), we see that for $\lambda \geqq 1$,

$$\alpha\left(\frac{C_1}{\omega}\right) \geqq \frac{C_1}{\omega}A\left(\frac{1}{\omega}\right) \geqq C_1 A_1\left(\frac{1}{\omega}\right) \geqq \frac{1}{\lambda} = \alpha\left(\frac{1}{\sigma}\right).$$

Therefore, since (2.5) holds,

$$\omega \leqq C_1\sigma \qquad \text{and} \qquad \lambda \leqq C_2\sigma^2, \qquad \lambda \geqq 1, \tag{2.10}$$

with $C_2 = 10C_1^2/a(t_1)$. σ can grow faster than ω; for example, if $a(t) = t^{-1}(-\log t)^{-3/2}$ for small t, one shows from (2.6) and (2.9) that

$$K_1\omega \log\omega \leqq K_2\lambda\,(\log\lambda)^{-1/2} \leqq \sigma \leqq K_3\lambda(\log\lambda)^{-1/2} \leqq K_4\omega\log\omega,$$

where the K_j are positive constants. Note, however, that

$$\lim_{\lambda\to\infty}\frac{\sigma}{\lambda} = \lim_{\lambda\to\infty} A\left(\frac{1}{\sigma}\right) = 0. \tag{2.11}$$

The next result relates σ to v.

THEOREM 2.2. *If* (1.2) *holds, then*

$$\frac{\sigma}{8(8+dC_2)} \leqq \sup_{t\geqq 0} |v(t,\lambda)| \leqq (8+dC_2)\sigma, \qquad (\lambda \geqq 1). \tag{2.12}$$

The proof of Theorem 2.2 contains the following:

COROLLARY 2.1. *Let* (1.2) *hold. There exist* $\varepsilon, K>0$ *such that* $v_0(t)\geqq K/t$ $(0<t<\varepsilon)$; *in particular,* $\int_0^1 v_0(t)\,dt=\infty$.

By (2.6) and (2.10), (2.12) shows that $\lambda^{-1/2}v(t,\lambda)$ is not bounded if $a(0+)=\infty$. If $a(0+)<\infty$, (2.9) shows that $\sigma^2 \leqq a(0+)\lambda$, so $\lambda^{-1/2}v(t,\lambda)$ is bounded. The latter conclusion strengthens [6, Lemma 5.2]; it thus improves Theorems 1 and 2 of that paper by showing that one may omit the term $\log(\lambda/\Lambda)$ from the definition of u^1 without changing the conclusions. Our main results, Theorems 2.3 and 2.4, generalize this part of [6] to cases where $a(0+)=\infty$. As in [2], we shall need the technical hypothesis

$$\begin{aligned} &a(t)=b(t)+c(t), \text{ where } b \text{ and } c \text{ each satisfy (1.2),} \\ &\text{except that either } b(0+)=0 \text{ or } c(0+)=0 \text{ is permitted.} \\ &\text{Moreover, } \int_1^\infty t^{-1}b(t)\,dt<\infty \text{ and } -c' \text{ is convex.} \end{aligned} \tag{2.13}$$

THEOREM 2.3. *Suppose* (1.2) *and* (2.2) *hold, and let* $0\leqq\gamma<\infty$. (i) *If*

$$\sup_{1/2\rho\leqq\tau<\infty} \frac{\tau[\theta(\tau)]^{1+\gamma}}{\varphi(\tau)}<\infty, \tag{2.14}$$

then $\sup_{t\geqq 0} tv_\gamma(t)<\infty$. (ii) *If* (2.13) *holds and either*

$$\sup_{1/2\rho\leqq\tau<\infty} \frac{\tau[\theta(\tau)]^{1+\gamma-\varepsilon}}{\varphi(\tau)}<\infty \quad \textit{for some } \varepsilon, \quad 0<\varepsilon<\gamma, \tag{2.15}$$

or $\gamma\geqq 1$ *and*

$$\sup_{1/2\rho\leqq\tau<\infty} \frac{\tau[\theta(\tau)]^{2+\gamma}}{\varphi^2(\tau)}<\infty, \tag{2.16}$$

then

$$\int_0^\infty v_\gamma(t)\,dt<\infty. \tag{2.17}$$

When $\gamma=\frac{1}{2}$, the following criterion is sometimes weaker than (2.15).

THEOREM 2.4. *If* (1.2), (2.2), *and* (2.13) *hold, and if*

$$\sup_{1/2\rho\leqq\tau<\infty} \frac{\tau^2\theta^2(\tau)}{\varphi(\tau)}<\infty, \tag{2.18}$$

then $\int_0^\infty v_{1/2}(t)\,dt<\infty$.

For purposes of comparison, we restate our conditions for (1.5) from [2].

THEOREM A. *Suppose* (1.2), (2.2) *and* (2.13) *hold. Then*

$$\int_0^\infty \sup_{1\leqq\lambda<\infty} |u(t,\lambda)|\,dt<\infty \tag{2.19}$$

if and only if

$$\sup_{1/2\rho\leqq\tau<\infty} \frac{\theta(\tau)}{\varphi(\tau)}<\infty. \tag{2.20}$$

The hypotheses in these results satisfy the following implications:

$$(2.18)\Rightarrow(2.14)(\gamma=\tfrac{1}{2})\Rightarrow(2.20)\Rightarrow(2.16)(\gamma\geqq 1) \tag{2.21}$$

(see (2.4)). If $a(0+)<\infty$, (2.4) gives us

$$\tfrac{1}{10}a(t_1)\tau^{-2}\leqq\theta(\tau)\leqq 6a(0+)\tau^{-2},$$

so (2.18), (2.14) ($\gamma=\frac{1}{2}$) and (2.20) all are equivalent to strong positivity. Thus, while (2.15) obviously implies (2.14), the kernel $a(t)=e^{-t}$ provides an example where (2.18) holds but (2.15) ($\gamma=\frac{1}{2}$) is false.

If $0<\beta<1$, the example $a(t)=t^{-\beta}$ satisfies (2.14) ($\gamma=0$) and, hence, satisfies (2.15) for all positive γ.

By considering a certain family of piecewise linear kernels, we can demonstrate other differences among our hypotheses. We defer the proof to § 7.

THEOREM 2.5. *There are kernels* a_1, $a_{2,\gamma}(\gamma=\frac{1}{2}, 1, \frac{3}{2}, \cdots)$ a_3 *and* a_4 *satisfying* (1.2), (2.2) *and* (2.13) *and such that*

(i) a_1 *satisfies* (2.15) ($\gamma=\frac{1}{2}$) *but not* (2.18).

(ii) *For each fixed* γ, $a_{2,\gamma}$ *satisfies* (2.14), *but neither* (2.15) *nor* (2.18) *nor* (2.16) *when* $\gamma\geqq 1$ *holds.*

(iii) a_3 *satisfies* (2.20) *but not* (2.14) ($\gamma=\frac{1}{2}$).

(iv) a_4 *satisfies* (2.16) ($\gamma=1$) *but not* (2.20).

By (2.21), Theorem 2.3 and (1.14), the sufficient condition (2.20) of Theorem A implies (1.9), as asserted in § 1. The following corollary shows that Theorems 2.3 and 2.4 contain Theorem 1.1.

COROLLARY 2.2. *If* (1.2), (2.2) *and* (2.13) *hold, and if*

$$\limsup_{t\to 0+}\frac{\int_0^t b(s)\,ds}{\int_0^t c(s)\,ds}<\infty, \tag{2.22}$$

then (2.18) *holds, so* (*by* (2.21) *and Theorems* A *and* 2.3) $\sup_{t\geqq 0} tv_{1/2}(t)<\infty$, *and* (2.17) ($\gamma=\frac{1}{2}$) *and* (2.19) *are valid.*

3. Statement of results for equations in $\mathcal{H}$. A solution of (1.1) (or (1.10) or (1.12)) is a continuously differentiable function $\mathbf{y}:\bar{\mathbb{R}}^+\to\mathcal{H}$ such that $\mathbf{Ly}:\bar{\mathbb{R}}^+\to\mathcal{H}$ is defined and continuous (in brief, $\mathbf{y}\in C(\bar{\mathbb{R}}^+,\mathcal{D})$) and (1.1) (or (1.10) or (1.12)) holds. Unless otherwise specified, integrals $\int_a^b$ of $\mathcal{H}$-valued functions are Bochner integrals in $\mathcal{B}^1((a,b),\mathcal{H})$; Hille and Phillips [7, pp. 59–89] give the theory of this integral. We recall from [2, Theorem 2.1(i)] that if (1.2) holds, then $\mathbf{U}(t)$ is strongly continuous on $\mathcal{H}$ and $\|\mathbf{U}(t)\|\leqq 1$ $(t\in\bar{\mathbb{R}}^+)$.

Our first result concerns $\mathbf{V}(t)$ as an operator from $\mathcal{D}_1$ to $\mathcal{H}$. The results of § 2 can also be used to study $\mathbf{V}(t)\mathbf{L}^{-\gamma}(\gamma\neq\frac{1}{2})$.

THEOREM 3.1. (i) *Suppose* (1.2) *and* (2.2) *hold and*

$$\sup_{1/2\rho\leqq t<\infty}\frac{\tau[\theta(\tau)]^{3/2}}{\varphi(\tau)}<\infty. \tag{3.1}$$

Then for $t>0$, $\mathbf{V}(t)\mathbf{L}^{-1/2}$ *is a bounded operator on* $\mathcal{H}$, *strongly continuous on* $\mathbb{R}^+$. *Moreover,*

$$\mathbf{V}(t)\mathbf{y}=\frac{d}{dt}\mathbf{U}(t)\mathbf{y},\qquad (t>0,\quad \mathbf{y}\in\mathcal{D}_1). \tag{3.2}$$

(ii) *If* $a(0+)<\infty$, *we may omit* (2.2) *and* (3.1) *in* (i); *moreover,* $\mathbf{V}(t)\mathbf{L}^{-1/2}$ *is strongly continuous and uniformly bounded on* $\bar{\mathbb{R}}^+$.

Next we state a representation theorem for solutions of (1.10).

THEOREM 3.2. (i) *Let the hypotheses of Theorem* 2.3 (ii) ($\gamma = \frac{1}{2}$) *or of Theorem* 2.4 *hold. Let* $\mathbf{z}_0 \in \mathscr{D}$, *let* $\mathbf{f} \in C(\bar{\mathbb{R}}^+, \mathscr{H})$ *with* $\mathbf{f}(t) \in \mathscr{D}(t \geqq 0)$ *and* $\mathbf{Lf} \in \mathscr{B}^1_{\text{loc}}(\bar{\mathbb{R}}^+, \mathscr{H})$. *Assume that* $\mathbf{g} \in \mathscr{B}^\infty_{\text{loc}}(\bar{\mathbb{R}}^+, \mathscr{D}_1)$. *Then the function* $\mathbf{z}(t)$ *given by* (1.11) *is the unique solution of* (1.10).

(ii) *Let* (1.2) *hold with* $a(0+) < \infty$. *Let* $\mathbf{z}_0$ *and* $\mathbf{f}$ *satisfy the hypotheses of* (i), *and let* $\mathbf{g} \in \mathscr{B}^1_{\text{loc}}(\bar{\mathbb{R}}, \mathscr{D}_1)$. *Then the conclusion of* (i) *is valid.*

Remark. In (i) above, we need $\|\mathbf{V}(\cdot)\mathbf{L}^{-1/2}\| \in L^1_{\text{loc}}(\bar{\mathbb{R}}^+)$, and by (1.14), the conclusions of Theorems 2.3(ii) ($\gamma = \frac{1}{2}$) and 2.4 imply this.

Miller [13] shows how to combine the resolvent formula for Volterra equations with fixed point theorems in order to prove global existence theorems for nonlinear equations. We use this method and Theorem 3.2 to obtain a result for (1.12).

THEOREM 3.3. *Let the hypotheses of Theorem* 2.3(ii) ($\gamma = \frac{1}{2}$) *or of Theorem* 2.4 *hold, and let* $\mathbf{y}_0 \in \mathscr{D}$. *Let f satisfy the hypotheses of Theorem* 3.2(i) *with* $\mathbf{f} = \mathbf{f}_1 + \mathbf{f}_2$, $\mathbf{f}_1 \in \mathscr{B}^1(\mathbb{R}^+, \mathscr{D})$, $\mathbf{f}_2 \in \mathscr{B}^\infty(\mathbb{R}^+, \mathscr{D})$. *Let*

$$\mathbf{N}: \{\mathbf{x} \in \mathscr{D} \mid \|\mathbf{x}\|_{\mathscr{D}} < \alpha\} \to \mathscr{D}_1$$

satisfy conditions (1.13). *Then if* $\mu \equiv \|\mathbf{y}_0\|_{\mathscr{D}} + \|\mathbf{f}_1\|_{\mathscr{B}^1(\mathbb{R}^+, \mathscr{D})} + \|\mathbf{f}_2\|_{\mathscr{B}^\infty(\mathbb{R}^+, \mathscr{D})}$ *and* $\Delta > 0$ *are sufficiently small,* (1.12) *has one and only one solution* $\mathbf{y}$ *such that* $\|\mathbf{y}(t)\|_{\mathscr{D}} \leqq \Delta(t \in \bar{\mathbb{R}}^+)$.

A simple example illustrating Theorem 3.3 is the problem

$$\begin{gathered} u_t(t, x) = \int_0^t a(t-s)[u_{xx}(s, x) + u(s, x)u_x(s, x)]\, ds + F(t, x), \\ u(t, 0) = u(t, \pi) = 0, \quad t \geqq 0, \qquad u(0, x) = u_0(x). \end{gathered} \tag{3.3}$$

We take $\mathscr{H} = L^2(0, \pi)$, $\mathbf{Ly} = -y''$ on $\mathscr{D}$, the space of differentiable functions $\mathbf{y}$ on $[0, \pi]$ with $\mathbf{y}(0) = \mathbf{y}(\pi) = 0$, $\mathbf{y}'$ absolutely continuous and $\mathbf{y}'' \in \mathscr{H}$. $\mathscr{D}_1$ consists of absolutely continuous functions which vanish at 0 and π and have square integrable first derivatives.

In terms of Fourier sine series

$$\mathbf{y}(x) = \sum_{n=1}^{\infty} c_n \sin nx,$$

$\mathscr{D}$ and $\mathscr{D}_1$ are characterized respectively by the conditions $\sum n^4 c_n^2 < \infty$ and $\sum n^2 c_n^2 < \infty$ and

$$\mathbf{L}^{1/2}\mathbf{y}(x) = \sum_{n=1}^{\infty} nc_n \sin nx.$$

Thus $\|\mathbf{L}^{1/2}\mathbf{y}\| = \|\mathbf{y}'\|$ ($\mathbf{y} \in \mathscr{D}_1$). Note also that, if $\mathbf{y} \in \mathscr{D}$,

$$|\mathbf{y}'(x)|^2 \leqq \Big(\sum_{n=1}^{\infty} n|c_n|\Big)^2 \leqq \sum_{n=1}^{\infty} n^{-2} \sum_{n=1}^{\infty} n^4 c_n^2 = B^2 \|\mathbf{Ly}\|^2$$

($0 \leqq x \leqq \pi$), so also $|y(x)| \leqq \frac{1}{2} B\pi \|\mathbf{Ly}\| (0 \leqq x \leqq \pi)$. Using these facts, one easily shows that $\mathbf{Ny} = \mathbf{yy}'$ satisfies (1.13).

The nonlinearity uu_x in (3.3) could be generalized, but our theorem does not cover such nonlinearities as u_x^2 or $\mathbf{N}_1\mathbf{u} = [h(u_x)]_x$; $\mathbf{N}_1$ is important in viscoelasticity theory.

MacCamy [11], [12], Dafermos and Nohel [3] and Staffans [17] have established global existence results for (3.3) with $\mathbf{N}$ replaced by $\mathbf{N}_1$ and $a(0+) < \infty$. Londen's global existence results [10] deal with (1.1) with $\mathbf{L}$ replaced by a maximal monotone (nonlinear) operator and $a(0+) < \infty$, $a'(0+) = -\infty$. Travis and Webb [18] prove a

general local existence result for hyperbolic semilinear equations, including (1.12) when $a(0+)<\infty$.

4. Proofs of Theorem 2.2 and Corollary 2.1. We redefine a', b', c' where necessary to make them continuous from the left on $\mathbb{R}^+$, da' denotes the Lebesgue-Stieltjes measure on $\mathbb{R}^+$. We adopt the conventions

$$\int_0^y f(t)\,da'(t) \equiv \int_{(0,y)} F\,da', \qquad \int_x^y f(t)\,da'(t) \equiv \int_{[x,y)} f\,da', \qquad 0<x<y.$$

For this proof, we define $\delta = \sigma^{-1}$.

Recall that when (1.2) holds,

$$|u(t,\lambda)| \leqq 1, \qquad t \geqq 0, \quad \lambda > 0 \tag{4.1}$$

(see [5], [2, p. 965]). Then (1.3), (4.1) and (2.10) imply

$$|v(t,\lambda)| \leqq \lambda(td + A(t)) \leqq \sigma + \lambda\,d\delta \leqq \sigma(1+dC_2)(0 \leqq t < \delta). \tag{4.2}$$

For $\delta \leqq t < \infty$, we make the change of variable $s \to t-s$ in (1.3) and integrate by parts to obtain the identity

$$\begin{aligned} v(t,\lambda) = \lambda \int_0^\delta a'(s) \int_{t-s}^t u(r,\lambda)\,dr\,ds + \lambda \int_\delta^t a'(s) \int_{t-s}^t u(r,\lambda)\,dr\,ds \\ -\lambda(d+a(t)) \int_0^t u(s,\lambda)\,ds \\ \equiv v_1(t,\lambda) + v_2(t,\lambda) + v_3(t,\lambda). \end{aligned}$$

Clearly,

$$|v_1(t,\lambda)| \leqq -\lambda \int_0^\delta s\,a'(s)\,ds \leqq \lambda A(\delta) = \sigma. \tag{4.3}$$

Since a' is monotone, we can use Fubini's theorem to see (with λ suppressed) that

$$\begin{aligned} &\int_\delta^t \int_\delta^s a'(r)[u(s)-u(s-r)]\,dr\,ds \\ &\quad = \int_\delta^t a'(r) \int_r^t [u(s)-u(s-r)]\,ds\,dr \\ &\quad = \int_\delta^t a'(s) \left[\int_s^t - \int_0^{t-s}\right] u(r)\,dr\,ds \\ &\quad = \lambda^{-1} v_2(t,\lambda) - \int_\delta^t a'(s) \int_0^s u(r)\,dr\,ds, \qquad t \geqq \delta. \end{aligned}$$

Therefore, $v_2(t,\lambda)$ is locally absolutely continuous in t and

$$\frac{1}{\lambda}\frac{\partial v_2}{\partial t} = a'(t) \int_0^t u(s,\lambda)\,ds + \int_\delta^t a'(s)[u(t,\lambda) - u(t-s,\lambda)]\,ds$$

a.e. ($t \geqq \delta$). Integration by parts then yields

$$\frac{1}{\lambda}\frac{\partial v_2}{\partial t} = u(t,\lambda)[a(t)-a(\delta)] + a'(\delta)\int_{t-\delta}^{t} u(r,\lambda)\,dr + \int_{\delta}^{t}\left[\int_{t-s}^{t} u(r,\lambda)\,dr\right] da'(s) \quad \text{a.e.}, \quad t \geqq \delta,$$

so

$$\frac{1}{\lambda}\left|\frac{\partial v_2}{\partial t}\right| \leqq 2a(\delta) - 2\delta a'(\delta) - 2a(t) + ta'(t) \quad \text{a.e.} \tag{4.4}$$

Since

$$\frac{1}{\lambda}\frac{\partial v_3}{\partial t} = -a'(t)\int_0^t u(s,\lambda)\,ds - (d+a(t))u(t,\lambda) \quad \text{a.e.},$$

$$\frac{1}{\lambda}\left|\frac{\partial v_3}{\partial t}\right| \leqq -ta'(t) + d + a(t) \quad \text{a.e.}$$

Adding this to (4.4) yields

$$\frac{1}{\lambda}\left|\frac{\partial(v_2+v_3)}{\partial t}\right| \leqq 2a(\delta) - 2\delta a'(\delta) + d \quad \text{a.e.} \tag{4.5}$$

Suppose there exists $t^* > \delta$ such that

$$|v(t^*,\lambda)| > (8 + dC_2)\sigma. \tag{4.6}$$

Let $I = [t^*-\delta, t^*+\delta]$, and observe that if $s \in I$,

$$\begin{aligned} |v(s,\lambda)| &\geqq |v(t^*,\lambda)| - 2\sup_{r\in I}|v_1(r,\lambda)| - \delta \operatorname*{ess\,sup}_{r\in I}\left|\frac{\partial(v_2+v_3)}{\partial t}(r,\lambda)\right| \\ &> (8+dC_2)\sigma - 2\sigma - 2\lambda(\delta a(\delta) - \delta^2 a'(\delta)) - \lambda d\delta; \end{aligned} \tag{4.7}$$

here (4.3), (4.4) and the absolute continuity of v_2+v_3 have been used. Integration by parts shows that

$$0 \leqq \int_0^{\delta} t^2\,da'(t) = 2A(\delta) - 2\delta a(\delta) + \delta^2 a'(\delta).$$

Combining this with (4.7), we obtain

$$|v(s,\lambda)| > (6+dC_2)\sigma - 4\lambda A(\delta) - \lambda d\delta, \qquad s \in I.$$

But $\delta^{-1} = \sigma = \lambda A(\delta)$, and since (2.10) holds,

$$|v(s,\lambda)| > 2\delta^{-1}, \qquad s \in I.$$

Thus, by (4.1) and the Mean Value Theorem, (4.6) has led us to the contradiction

$$2 \geqq |u(t^*,\lambda) - u(t^*-\delta,\lambda)| > \delta \cdot 2\delta^{-1} = 2.$$

Since (4.2) holds, the second inequality in (2.12) is established. It follows that

$$u(t,\lambda) \geqq 1 - (8+dC_2)\sigma t, \qquad t \geqq 0,$$

so $u(t,\lambda)\geqq\frac{1}{2}$ for $0\leqq t\leqq[2\sigma(8+dC_2)]^{-1}\equiv 2T$. Then by (1.2) and (1.3),

$$|v(t,\lambda)|\geqq\tfrac{1}{2}\lambda A(T)\geqq\frac{\lambda A(\sigma^{-1})}{8(8+dC_2)}=\frac{\sigma}{8(8+dC_2)},\qquad T\leqq t\leqq 2T. \tag{4.8}$$

This proves Theorem 2.2.

If $a(0+)<\infty$, the second inequality in (2.12) is essentially contained in Levin [8]. The idea of writing $v=v_1+v_2+v_3$ in case $a(0+)=\infty$ was introduced by Londen [9, Lemma 2].

For Corollary 2.1, let $T=T(\lambda)$, as in (4.8). If $t>0$ is sufficiently small, we can find $\lambda=\lambda_t$ such that $T(\lambda)\leqq t\leqq 2T(\lambda)$. Then by (4.8)

$$v_0(t)\geqq v(t,\lambda)\geqq\frac{\sigma}{8(8+dC_2)}\geqq\frac{1}{16t(8+dC_2)^2},$$

as asserted.

5. Proof of Theorem 2.3. Throughout this paper, the symbol M denotes a finite positive constant, independent of $\lambda(1\leqq\lambda<\infty)$; the numerical value of M can change each time M appears. We assume (1.2) and (2.2).

(2.11) and (2.12) immediately yield

$$v_\gamma(t)\leqq M,\qquad \gamma\geqq 1,\quad t\geqq 0. \tag{5.1}$$

Choose $\omega^*=\omega^*(\lambda)$ so that

$$\tfrac{1}{2}\omega\leqq\tau\leqq 2\omega\quad\text{and}\quad\varphi(\omega^*)=\min_{\omega/2\leqq\tau\leqq 2\omega}\varphi(\tau):$$

for instance, ω^* could be the smallest such number.

We shall establish the following estimates;

$$|v(t,\lambda)|\leqq M\left(1+\frac{\omega^*\theta(\omega^*)}{\varphi(\omega^*)}\right)t^{-1},\qquad t>0. \tag{5.2}$$

If (2.13) holds,

$$|v(t,\lambda)|\leqq M\left[\left(1+\frac{\omega^*\theta(\omega^*)}{\varphi(\omega^*)}\right)Q(t)+\left(1+\frac{\omega^*\theta^2(\omega^*)}{\varphi^2(\omega^*)}\right)t^{-2}\right],\qquad t\geqq 1, \tag{5.3}$$

where $Q\in L^1(1,\infty)$.

Before proving (5.2) and (5.3), we show that they imply the conclusions of Theorem 2.3. Note that

$$\int_t^{2t} sa(s)\,ds\leqq a(t)\int_t^{2t} s\,ds=3a(t)\int_0^t s\,ds\leqq 3A_1(t).$$

Therefore,

$$A_1(2t)\leqq 4A_1(t),\qquad t>0. \tag{5.4}$$

Using (5.4), we can combine (2.4) and (2.6) to see that

$$\frac{1}{M}\leqq\lambda\theta(\tau)\leqq M,\qquad \tfrac{1}{2}\omega\leqq\tau\leqq 2\omega. \tag{5.5}$$

Then if (2.14) holds, (5.2) gives us the conclusion of Theorem 2.3(i).

If $\gamma \geqq 1$ and (2.13) and (2.16) hold, we use the algebraic inequality

$$\frac{2\theta}{\varphi} \leqq 1 + \left(\frac{\theta}{\varphi}\right)^2 \tag{5.6}$$

to deduce from (5.3) that

$$|v(t,\lambda)| \leqq M\left[Q(t)(1+\omega^*) + \frac{\omega^*\theta^2(\omega^*)}{\varphi^2(\omega^*)}t^{-2}\right], \qquad t \geqq 1.$$

Then by (5.5), (2.10), (2.11) and (2.16),

$$\lambda^{-\gamma}|v(t,\lambda)| \leqq M(Q(t)+t^{-2}), \qquad t \geqq 1.$$

Since (5.1) holds, (2.17) is valid.

Now assume (2.13) and (2.15). If $0<\gamma<1$, we deduce from (2.15), (5.2) and (5.5) that

$$|v(t,\lambda)| \leqq Mt^{-1}\lambda^{\gamma-\varepsilon}, \qquad t>0. \tag{5.7}$$

If $p=(1-\gamma)/(1-\gamma+\varepsilon)$, then $0<p<1$ and $p(\gamma-\varepsilon)+(1-p)=\gamma$, so (2.11), (2.12) and (5.7) tell us that

$$|v(t,\lambda)| = |v(t,\lambda)|^{p+(1-p)} \leqq M\lambda^{\gamma}t^{-p}, \qquad t>0, \tag{5.8}$$

if $\gamma<1$. We conclude from (5.1) and (5.8) that

$$\int_0^1 v_\gamma(t)\,dt < \infty \quad \text{if (2.15) holds.} \tag{5.9}$$

Choose $\delta < \varepsilon/(\gamma-\varepsilon)$, $0<\delta<1$. If $1 \leqq t^\delta \leqq \lambda^\varepsilon$, (2.15), (5.2) and (5.5) imply that

$$\begin{aligned} |v(t,\lambda)| &\leqq M\left(1+\frac{\omega^*\theta^{1+\gamma-\varepsilon}(\omega^*)}{\varphi(\omega^*)}\right)\lambda^{\gamma-\varepsilon}\left(\frac{\lambda^\varepsilon}{t^{1+\delta}}\right) \\ &\leqq M\lambda^\gamma t^{-1-\delta}. \end{aligned} \tag{5.10}$$

If $\lambda^\varepsilon < t^\delta$, then $\lambda^{-\varepsilon+\varepsilon/\delta} < t^{1-\delta}$, so (2.15), (5.3) and (5.5) yield

$$|v(t,\lambda)| \leqq M\left[\lambda^{\gamma-\varepsilon}Q(t) + \left(\frac{\omega^*\theta^{1+\gamma-\varepsilon}(\omega^*)}{\varphi(\omega^*)}\right)^2 \frac{\lambda^{2\gamma-2\varepsilon}t^{1-\delta}}{t^2\lambda^{-\varepsilon+\varepsilon/\delta}}\right].$$

Since $\gamma-\varepsilon < \varepsilon/\delta$, another application of (2.15) shows that

$$|v(t,\lambda)| \leqq M\lambda^\gamma[Q(t)+t^{-1-\delta}], \qquad \lambda^\varepsilon < t^\delta.$$

This inequality, taken together with (5.9) and (5.10), gives us (2.17).

We have shown that Theorem 2.3 is a consequence of (5.2) and (5.3), which we prove next.

When (1.2) and (2.2) hold, one has the inversion formula

$$\pi v(t,\lambda) = \operatorname{Re}\left\{\frac{1}{t\lambda}\int_0^\infty e^{i\tau t}\left(\frac{\tau D'(\tau)-D(\tau)}{D^2(\tau,\lambda)}\right)d\tau\right\}, \qquad t>0, \tag{5.11}$$

where the integral is absolutely convergent at both $\tau=0$ and $\tau=\infty$. This was established in [1].

The next lemmas will enable us to estimate D and D'.

LEMMA 5.1. *If* (1.2) *holds, then*

$$\varphi(\tau) \geqq \tfrac{1}{2}[A(\tau^{-1}) - 3\tau A_1(\tau^{-1})], \qquad \tau>0. \tag{5.12}$$

Proof. Two integrations by parts yield

$$\begin{aligned}\varphi(\tau) &= \tau^{-2}\int_0^\infty (1-\cos\tau t)\,da'(t)\\ &\geqq \frac{1}{4}\int_0^{1/\tau} t^2\,da'(t)\\ &\geqq \frac{1}{4}\int_0^{1/\tau} (t^2-\tau t^3)\,da'(t)\\ &= \frac{1}{2}[A(\tau^{-1})-3\tau A_1(\tau^{-1})]+\frac{1}{4\tau}a(\tau^{-1}).\end{aligned}$$

Here we have used $1-\cos x \geqq \frac{1}{4}x^2 (0\leqq x\leqq 1)$ and the fact that da' is a positive measure. Since $a\geqq 0$, the lemma is proved.

LEMMA 5.2. *If* (1.2) *holds, then*

$$\varphi^2(\tau)+\left(\frac{\tau-\omega}{\lambda}\right)^2 \leqq M|D(\tau,\lambda)|^2, \qquad \tau\geqq\frac{1}{2}\rho, \tag{5.13}$$

$$A(\tau^{-1})\leqq M|D(\tau,\lambda)|, \qquad \tau\in[\tfrac{1}{2}\rho,\tfrac{1}{2}\omega]\cup[2\omega,\infty). \tag{5.14}$$

Proof. [2, Lemma 5.2] states that (1.2) implies

$$|\tau-\omega|\leqq M\lambda|D(\tau,\lambda)|, \qquad \tau\geqq\tfrac{1}{2}\omega, \tag{5.15}$$

$$\tau A_1(\tau^{-1})\leqq M|D(\tau,\lambda)|, \qquad \tfrac{1}{2}\rho\leqq\tau\leqq\tfrac{1}{2}\omega. \tag{5.16}$$

In § 8 below we give a corrected proof of this lemma.

For $\tau\geqq\frac{1}{2}\omega$, (5.13) is a trivial consequence of (5.15). (2.6) and (5.15) show that

$$\begin{aligned}\frac{1}{10}\tau A_1(\tau^{-1}) &\leqq \frac{1}{10}\tau A_1(\omega^{-1}) \leqq \frac{1}{2}\frac{\tau}{\lambda}\leqq\frac{\tau-\omega}{\lambda}\\ &\leqq M|D(\tau,\lambda)|, \qquad \tau\geqq 2\omega.\end{aligned} \tag{5.17}$$

Thus, if $\tau\in[\frac{1}{2}\rho,\frac{1}{2}\omega]\cup[2\omega,\infty)$, (5.12), (5.16) and (5.17) imply that

$$A(\tau^{-1})\leqq 2\varphi(\tau)+3\tau A_1(\tau^{-1})\leqq M|D(\tau,\lambda)|,$$

as asserted in (5.14).

If $\frac{1}{2}\rho\leqq\tau\leqq\frac{1}{2}\omega$, by (2.6),

$$\frac{\omega-\tau}{\lambda}\leqq\frac{\omega}{\lambda}\leqq C_1\omega A_1\left(\frac{1}{\omega}\right)\leqq C_1 A\left(\frac{1}{\tau}\right).$$

By (5.14), this implies (5.13) for such τ, and our proof is complete.

Recall from [2, Lemma 4.1] that when (1.2) holds we have

$$2^{-3/2}A(\tau^{-1})\leqq|\hat{a}(\tau)|\leqq 4A(\tau^{-1}), \quad |\hat{a}'(\tau)|\leqq 40A_1(\tau^{-1}), \qquad \tau>0. \tag{5.18}$$

We now deduce (5.2) from (5.11). If $d>0$, (5.18) shows that

$$|\tau D'(\tau)-D(\tau)|\leqq M\tau^{-1}, \qquad 0<\tau\leqq\rho,$$

while (2.2) gives

$$|D(\tau,\lambda)|\geqq\max\left\{\varphi(\tau),\frac{(d-\tau^2)}{\tau}\right\}\geqq 1/M\tau, \qquad 0<\tau\leqq\rho. \tag{5.19}$$

Thus,

$$\int_0^{\rho/2} \frac{|\tau D'(\tau) - D(\tau)|}{\lambda |D(\tau, \lambda)|^2}\, d\tau \leqq M. \tag{5.20}$$

On the other hand, if $d = 0$, (5.18) implies that $|\tau D'(\tau) - D(\tau)| \leqq MA(\tau^{-1})$ and

$$|D(\tau, \lambda)| \geqq \max\{2^{-3/2} A(\tau^{-1}) - \tau, \varphi(\tau)\}, \tag{5.21}$$

so (5.20) is again valid.

By (2.10), (5.18) and (5.14),

$$\left(\int_{\rho/2}^{\omega/2} + \int_{2\omega}^{2C_1\sigma}\right) \frac{|\tau D'(\tau) - D(\tau)|}{\lambda |D(\tau, \lambda)|^2}\, d\tau \leqq M \int_{\rho/2}^{2C_1\sigma} \frac{A(\tau^{-1})\, d\tau}{\lambda A^2(\tau^{-1})} \leqq \frac{M\sigma}{\lambda A(\sigma^{-1})} = M. \tag{5.22}$$

Next we use (5.18) and (5.13) to obtain

$$\int_{2C_1\sigma}^{\infty} \frac{|\tau D'(\tau) - D(\tau)|}{\lambda |D(\tau, \lambda)|^2}\, d\tau \leqq M\lambda \int_{2C_1\sigma}^{\infty} \frac{A(\tau^{-1})}{\tau^2}\, d\tau \leqq M\lambda A(\sigma^{-1})\sigma^{-1} \leqq M. \tag{5.23}$$

Before estimating the final piece in (5.11), note that (5.18) implies

$$MA(\tau^{-1}) \leqq \varphi(\omega^*) + \omega^* \theta(\omega^*), \qquad \tfrac{1}{2}\omega \leqq \tau \leqq 2\omega. \tag{5.24}$$

Now (5.13), (5.18) and (5.24) give us

$$\begin{aligned} \int_{\omega/2}^{2\omega} \frac{|\tau D'(\tau) - D(\tau)|}{\lambda |D(\tau, \lambda)|^2}\, d\tau &\leqq M\lambda A(2\omega^{-1}) \int_{\omega/2}^{2\omega} \frac{d\tau}{[\lambda\varphi(\omega^*)]^2 + |\tau - \omega|^2} \\ &\leqq \frac{MA(2\omega^{-1})}{\varphi(\omega^*)} \leqq M\left(1 + \frac{\omega^*\theta(\omega^*)}{\varphi(\omega^*)}\right). \end{aligned} \tag{5.25}$$

Thus, from (5.22), (5.23) and (5.25), we obtain (5.2).

Next we turn to (5.3). Assume (2.13), in addition to (1.2) and (2.2), and write (5.11) as

$$v(t, \lambda) = \operatorname{Re}\{\lambda^{-1} v_1(t) + i\lambda^{-2} v_2(t) + \lambda^{-3} v_3(t) - v_4(t, \lambda) - v_5(t, \lambda)\}, \tag{5.26}$$

where (these v_j are unrelated to those of § 4)

$$\begin{aligned} tv_1(t) &= \int_0^{\rho} e^{i\tau t} \frac{\tau D'(\tau)}{D^2(\tau)}\, d\tau, \\ tv_2(t) &= \int_0^{\rho} \frac{\tau e^{i\tau t}}{D^2(\tau)} \left[1 - \frac{2\tau D'(\tau)}{D(\tau)}\right] d\tau, \\ tv_3(t) &= \int_0^{\rho} e^{i\tau t} \frac{2\tau^2}{D^3(\tau)}\, d\tau, \\ t\lambda v_4(t, \lambda) &= \int_0^{\rho} e^{i\tau t} \left\{\left(\frac{\tau^3 D_\tau(\tau, \lambda)}{\lambda^2 D^3(\tau) D(\tau, \lambda)}\right)\left(\frac{2}{D(\tau)} + \frac{1}{D(\tau, \lambda)}\right) + \frac{1}{D(\tau, \lambda)}\right\} d\tau, \\ t\lambda v_5(t, \lambda) &= \int_{\rho}^{\infty} e^{i\tau t} \left(\frac{\tau D'(\tau) - D(\tau)}{D^2(\tau, \lambda)}\right) d\tau. \end{aligned}$$

We shall show that

$$|v_4(t, \lambda)| + |v_5(t, \lambda)| \leqq M\left\{\left[1 + \frac{\omega^*\theta(\omega^*)}{\varphi(\omega^*)}\right] q(t) + \frac{\omega^*\theta^2(\omega^*)}{\varphi^2(\omega^*) t^2}\right\}, \qquad t \geqq 1, \tag{5.27}$$

where

$$q(t) = t^{-2} + t^{-2}\int_0^t b(s)\,ds + t^{-1}b(t) - b'(t), \qquad t \geqq 1.$$

We know from [2, p. 972] that $q \in L^1(1, \infty)$. Moreover, from [15, Theorem 2] and the fact that

$$\operatorname{Re}\int_0^\infty e^{-st}a(t)\,dt > 0, \qquad \operatorname{Re} s \geqq 0,\ s \neq 0,$$

under our hypotheses [4], it follows that $v(\cdot, \lambda) \in L^1(\mathbb{R}^+)$. Then by (5.26) and (5.27), each of v_1, v_2, v_3 belongs to $L^1(1, \infty)$. (5.3) now follows from (5.26) and (5.27) with $Q = |v_1| + |v_2| + |v_3| + q$. We have reduced (5.3) to (5.27).

Let $J(u) = iu(1 - e^{iu}) - 2(1 - iu - e^{iu})$, and recall from [2, (4.9)] that

$$b'(\tau) = \tau^{-3}\int_0^\infty J(-\tau s)\,db'(s), \qquad \tau > 0.$$

For $t, \tau > 0$ define

$$\beta^0(t, \tau) = \tau^{-3}\int_0^t J(-\tau s)\,db'(s),$$

$$\beta^\infty(t, \tau) = \tau^{-3}\int_t^\infty J(-\tau s)\,db'(s)$$

$$\Delta(t, \tau) = \beta^0(t, \tau) + \hat{c}'(\tau) + i\,d\tau^{-2} = D'(\tau) - \beta^\infty(t, \tau). \tag{5.28}$$

In [2, Lemma 5.1] we proved by direct estimates that $\hat{c} \in C^2(\mathbb{R}^+)$, $\partial\beta^0/\partial\tau \in C(\mathbb{R}^+ \times \mathbb{R}^+)$ and

$$|\hat{c}''(\tau)| \leqq 6000\int_0^{1/\tau} s^2 c(s)\,ds, \qquad \tau > 0, \tag{5.29}$$

$$|\beta^\infty(t, \tau)| \leqq 40\tau^{-2}(b(t) - tb'(t)), \qquad t, \tau > 0, \tag{5.30}$$

$$\left|\frac{\partial\beta^0}{\partial\tau}(t, \tau)\right| \leqq 500\tau^{-2}\int_0^t b(s)\,ds, \qquad t, \tau > 0, \tag{5.31}$$

$$|\beta^0(t, \tau)| \leqq 40\int_0^{1/\tau} sb(s)\,ds, \qquad t, \tau > 0, \tag{5.32}$$

$$|\hat{c}'(\tau)| \leqq 40\int_0^{1/\tau} sc(s)\,ds, \qquad \tau > 0. \tag{5.33}$$

Write $v_4 = v_{41} + v_{42}$, where

$$t\lambda v_{41}(t, \lambda) = \int_0^\rho e^{i\tau t}\left[\left(\frac{\tau^3[\Delta(t, \tau) + i\lambda^{-1}]}{\lambda^2 D^3(\tau)D(\tau, \lambda)}\right)\left(\frac{2}{D(\tau)} + \frac{1}{D(\tau, \lambda)}\right) + \frac{1}{D(\tau, \lambda)}\right] d\tau, \tag{5.34}$$

$$t\lambda v_{42}(t, \lambda) = \int_0^\rho e^{i\tau t}\left[\frac{\tau^3\beta^\infty(t, \tau)}{\lambda^3 D^3(\tau)D(\tau, \lambda)}\left(\frac{2}{D(\tau)} + \frac{1}{D(\tau, \lambda)}\right)\right] d\tau. \tag{5.35}$$

Likewise, let $v_5 = v_{51} + v_{52}$, where

$$t\lambda v_{51}(t, \lambda) = \int_\rho^\infty e^{i\tau t}\frac{\tau\Delta(t, \tau) - D(\tau)}{D^2(\tau, \lambda)}\,d\tau, \tag{5.36}$$

$$t\lambda v_{52}(t,\lambda)=\int_\rho^\infty e^{i\tau t}\frac{\tau\beta^\infty(t,\tau)}{D^2(\tau,\lambda)}\,d\tau. \tag{5.37}$$

Now integrate by parts in (5.34) and (5.36) to obtain

$$\begin{aligned} i\lambda t^2 v_{41}(t,\lambda)=e^{i\rho t}&\left\{\frac{\rho^3[\Delta(t,\rho)+i\lambda^{-1}]}{\lambda^2 D^3(\rho)D(\rho,\lambda)}\left(\frac{2}{D(\rho)}+\frac{1}{D(\rho,\lambda)}\right)+\frac{1}{D(\rho,\lambda)}\right\}-\frac{1}{D(0+)}\\ &-\frac{1}{\lambda^2}\int_0^\rho e^{i\tau t}\left\{\frac{3\tau^2[\Delta(t,\tau)+i\lambda^{-1}]+\tau^2\Delta_\tau(t,\tau)}{D^3(\tau)D(\tau,\lambda)}\left[\frac{2}{D(\tau)}+\frac{1}{D(\tau,\lambda)}\right]\right.\\ &\qquad-\frac{\tau^3[\Delta(t,\tau)+i\lambda^{-1}][3D'(\tau)D(\tau,\lambda)+D(\tau)D_\tau(\tau,\lambda)]}{D^4(\tau)D^2(\tau,\lambda)}\\ &\qquad\qquad\cdot\left[\frac{2}{D(\tau)}+\frac{1}{D(\tau,\lambda)}\right]\\ &\qquad\left.-\frac{\tau^3[\Delta(t,\tau)+i\lambda^{-1}]}{D^3(\tau)D(\tau,\lambda)}\left[\frac{2D'(\tau)}{D^2(\tau)}+\frac{D_\tau(\tau,\lambda)}{D^2(\tau,\lambda)}\right]-\frac{\lambda^2 D_\tau(\tau,\lambda)}{D^2(\tau,\lambda)}\right\}d\tau, \end{aligned} \tag{5.38}$$

$$\begin{aligned} -i\lambda t^2 v_{51}(t,\lambda)=e^{i\rho t}&\left(\frac{\rho\Delta(t,\rho)-D(\rho)}{D^2(\rho,\lambda)}\right)\\ &+\int_\rho^\infty e^{i\tau t}\left(\frac{\Delta(t,\tau)+\tau\Delta_\tau(t,\tau)-D'(\tau)}{D^2(\tau,\lambda)}-2\left(\frac{D_\tau(\tau,\lambda)(\tau\Delta(t,\tau)-D(\tau))}{D^3(\tau,\lambda)}\right)\right)d\tau. \end{aligned} \tag{5.39}$$

Here (5.18), (5.32) and (5.33) have been used to simplify the boundary terms. In (5.38), $1/D(0+)$ is zero unless $d=0$ and $a\in L^1(\mathbb{R}^+)$. Our estimates (5.18), (5.19), (5.21) and (5.40) below show that the integrals in (5.38) converge absolutely.

By (5.18),

$$\left|\frac{D_\tau(\tau,\lambda)}{D^2(\tau,\lambda)}\right|\leqq M\frac{A_1(\tau^{-1})+d\tau^{-2}+\lambda^{-1}}{|D(\tau,\lambda)|^2}.$$

If $d>0$, (5.19) shows that

$$\int_0^\rho\left|\frac{D_\tau(\tau,\lambda)}{D^2(\tau,\lambda)}\right|d\tau\leqq M. \tag{5.40}$$

If $d=0$, we recall from [15, (1.21)] that

$$\int_0^1\frac{A_1(\tau^{-1})}{A^2(\tau^{-1})}\,d\tau<\infty.$$

Thus, by (5.21), (5.40) holds in this case as well. It is now a straightforward matter to use (5.18), (5.19), (5.21), (5.40) and (5.29) through (5.33) to estimate the terms in (5.35) and (5.38) and deduce

$$|v_4(t,\lambda)|\leqq Mq(t),\qquad t\geqq 1. \tag{5.41}$$

We turn now to v_5. The following estimates, direct consequences of (2.6), (5.18) and (5.29) through (5.33), will be used without explicit mention for estimates of the numerators.

$$
\begin{gathered}
|\Delta(t,\tau)|+|\tau\Delta_\tau(t,\tau)|+|D'(\tau)|\leqq M[A_1(\tau^{-1})+t^2q(t)\tau^{-1}], \qquad t\geqq 1, \quad \tau\geqq \tfrac{1}{2}\rho,\\
|D_\tau(\tau,\lambda)|\leqq MA_1(\tau^{-1})\leqq M\tau^{-1}A(\tau^{-1}), \qquad \tau\geqq\tfrac{1}{2}\rho,\\
\tau^2|\beta^\infty(t,\tau)|\leqq Mtq(t), \qquad t\geqq 1, \quad \tau>0,\\
\tau|\Delta(t,\tau)|+|D(\tau)|\leqq MA(\tau^{-1}), \qquad t\geqq 1, \quad \tau\geqq\tfrac{1}{2}\rho,\\
|\Delta(t,\tau)|+|\tau\Delta_\tau(t,\tau)|+|D'(\tau)|\leqq M(\lambda^{-1}+\tau^{-1}t^2q(t)), \qquad t\geqq 1, \quad \tau\geqq\tfrac{1}{2}\omega,\\
|D_\tau(\tau,\lambda)|\leqq M\lambda^{-1}, \qquad \tau\geqq\tfrac{1}{2}\omega.
\end{gathered}
\tag{5.42}
$$

We recall as well that

$$
A(\tau^{-1})\geqq A\left(\frac{1}{2C_1\sigma}\right)\geqq\frac{A(\sigma^{-1})}{2C_1}=\frac{\sigma}{2C_1\lambda}, \qquad \tau\leqq 2C_1\sigma
\tag{5.43}
$$

and

$$
\tau A(\tau^{-1})\geqq a(\tau^{-1})\geqq a(t_1), \qquad \tau\geqq\tfrac{1}{2}\rho.
$$

We use Lemma 5.2 and its simple consequence

$$
\tau\lambda^{-1}\leqq M|D(\tau,\lambda)|, \qquad 2\omega\leqq\tau<\infty
\tag{5.44}
$$

to get

$$
\begin{aligned}
&\left(\int_{\rho/2}^{\omega/2}+\int_{2\omega}^{2C_1\sigma}\right)\frac{|\Delta(t,\tau)+\tau\Delta_\tau(t,\tau)-D'(\tau)|}{\lambda|D(\tau,\lambda)|^2}\,d\tau\\
&\qquad\leqq\frac{M}{\lambda}\int_{\rho/2}^{2C_1\sigma}\frac{A_1(\tau^{-1})+\tau^{-1}t^2q(t)}{A^2(\tau^{-1})}\,d\tau\\
&\qquad\leqq\frac{M}{\lambda}\int_{\rho/2}^{2C_1\sigma}\frac{1}{\tau A(\tau^{-1})}+\frac{t^2q(t)}{\tau A^2(\tau^{-1})}\,d\tau\\
&\qquad\leqq\frac{M\sigma}{\lambda a(t_1)}\left(1+\frac{t^2q(t)}{A(\sigma^{-1})}\right)\leqq Mt^2q(t)
\end{aligned}
\tag{5.45}
$$

and (here (2.6) is used as well)

$$
\begin{aligned}
&\int_{2C_1\sigma}^{\infty}\frac{|\Delta(t,\tau)+\tau\Delta_\tau(t,\tau)-D'(\tau)|}{\lambda|D(\tau,\lambda)|^2}\,d\tau\\
&\qquad\leqq\frac{M}{\lambda}\int_{2C_1\sigma}^{\infty}\frac{\lambda^{-1}+\tau^{-1}t^2q(t)}{(\tau/\lambda)^2}\,d\tau\leqq Mt^2q(t).
\end{aligned}
\tag{5.46}
$$

Similarly,

$$
\begin{aligned}
&\left(\int_{\rho/2}^{\omega/2}+\int_{2\omega}^{2C_1\sigma}+\int_{2C_1\sigma}^{\infty}\right)\frac{|D_\tau(\tau,\lambda)\|\tau\Delta(t,\tau)-D(\tau)|}{\lambda|D(\tau,\lambda)|^3}\,d\tau\\
&\qquad\leqq M\left(1+\lambda A(\sigma^{-1})\int_{2C_1\sigma}^{\infty}\frac{d\tau}{\tau^3}\right)\leqq M.
\end{aligned}
\tag{5.47}
$$

On $[\frac{1}{2}\omega, 2\omega]$, we use (5.13) to estimate the denominator. This yields

$$\begin{aligned}\int_{\omega/2}^{2\omega} \frac{|\Delta(t,\tau)+\tau\Delta_\tau(t,\tau)-D'(\tau)|}{\lambda|D(\tau,\lambda)|^2}\,d\tau &\leqq M\lambda\int_{\omega/2}^{2\omega}\frac{\lambda^{-1}+\tau^{-1}t^2q(t)}{[\lambda\varphi(\omega^*)]^2+|\tau-\omega|^2}\,d\tau\\ &\leqq M(1+t^2q(t)\lambda\omega^{-1})\int_0^\infty \frac{ds}{[\lambda\varphi(\omega^*)]^2+s^2}\\ &\leqq \frac{Mt^2q(t)\lambda\omega^{-1}}{\lambda\varphi(\omega^*)} \leqq Mt^2q(t)\frac{\omega^*\theta(\omega^*)}{\varphi(\omega^*)}.\end{aligned} \tag{5.48}$$

The last inequality above used (2.5) and (5.5). Similarly, using (5.24), (5.25) and (5.6), we obtain

$$\begin{aligned}&\int_{\omega/2}^{2\omega}\frac{|D_\tau(\tau,\lambda)\|\tau\Delta(t,\tau)-D(\tau)|}{\lambda|D(\tau,\lambda)|^3}\,d\tau\\ &\quad\leqq M\int_{\omega/2}^{2\omega}\frac{\lambda A(2\omega^{-1})\,d\tau}{[(\lambda\varphi(\omega^*))^2+|\tau-\omega|^2]^{3/2}}\\ &\quad\leqq \frac{MA(2\omega^{-1})}{\lambda\varphi^2(\omega^*)} \leqq M\frac{[\varphi(\omega^*)+\omega^*\theta(\omega^*)]\theta(\omega^*)}{\varphi^2(\omega^*)}\\ &\qquad= M\left(\frac{\theta(\omega^*)}{\varphi(\omega^*)}+\frac{\omega^*\theta^2(\omega^*)}{\varphi^2(\omega^*)}\right)\leqq M\left(1+\frac{\omega^*\theta^2(\omega^*)}{\varphi^2(\omega^*)}\right).\end{aligned} \tag{5.49}$$

Thus, the representation (5.39), along with the estimates (5.45) through (5.49), gives us

$$|v_{51}(t,\lambda)|\leqq M\left\{q(t)\left[1+\frac{\omega^*\theta(\omega^*)}{\varphi(\omega^*)}\right]+t^{-2}\left[1+\frac{\omega^*\theta^2(\omega^*)}{\varphi^2(\omega^*)}\right]\right\}. \tag{5.50}$$

As in (5.45) through (5.48), we derive

$$\left[\int_{\rho/2}^{\omega/2}+\int_{2\omega}^{2C_1\sigma}+\int_{2C_1\sigma}^{\infty}\right]\frac{\tau|\beta^\infty(t,\tau)|}{\lambda|D(\tau,\lambda)|^2}\,d\tau\leqq Mtq(t). \tag{5.51}$$

Again, we use (5.13) on $[\frac{1}{2}\omega, 2\omega]$. This gives us

$$\int_{\omega/2}^{2\omega}\frac{\tau|\beta^\infty(t,\tau)|}{\lambda|D(\tau,\lambda)|^2}\,d\tau\leqq\frac{Mtq(t)}{\omega^*\varphi(\omega^*)}\leqq Mtq(t)\left[\frac{\omega^*\theta(\omega^*)}{\varphi(\omega^*)}\right], \tag{5.52}$$

where the last inequality invokes (2.5) and (5.5).

Then (5.37), (5.51) and (5.52) imply

$$|v_{52}(t,\lambda)|\leqq Mq(t)\left[1+\frac{\omega^*\theta(\omega^*)}{\varphi(\omega^*)}\right]. \tag{5.53}$$

But $v_5 = v_{51}+v_{52}$, so (5.41), (5.50) and (5.53) give us (5.27). This, in turn, gives us (5.3).

This completes the proof of Theorem 2.3.

6. Proofs of Theorem 2.4 and Corollary 2.2. To prove Theorem 2.4 we need (6.1), (6.2) and (6.3) below, which are consequences of (2.18), (5.18) and (5.24).

$$\frac{A^2(\tau^{-1})}{\varphi(\omega^*)}\leqq M\frac{|\hat a(\omega^*)|^2}{\varphi(\omega^*)}\leqq M\left(\varphi(\omega^*)+\frac{(\omega^*\theta(\omega^*))^2}{\varphi(\omega^*)}\right)\leqq M,\qquad \tfrac{1}{2}\omega\leqq\tau\leqq 2\omega. \tag{6.1}$$

Thus,

$$1+\frac{\omega^*\theta(\omega^*)}{\varphi(\omega^*)}\leqq 2\frac{|\hat{a}(\omega^*)|}{\varphi(\omega^*)}\leqq M\frac{A(2/\omega)}{\varphi(\omega^*)}\leqq\frac{M}{A(2/\omega)} \tag{6.2}$$
$$\leqq\frac{M}{A(1/C_1\sigma)}\leqq\frac{C_1M}{A(\sigma^{-1})}\leqq M\lambda\sigma^{-1}.$$

Furthermore, by (2.21), (2.18) implies (2.20), so

$$1+\frac{\omega^*\theta^2(\omega^*)}{\varphi^2(\omega^*)}\leqq\left(1+\frac{\omega^*\theta(\omega^*)}{\varphi(\omega^*)}\right)\left(1+\frac{\theta(\omega^*)}{\varphi(\omega^*)}\right)\leqq M\lambda\sigma^{-1}. \tag{6.3}$$

Comparing (6.2) and (6.3) with (5.3) shows that

$$|v(t,\lambda)|\leqq MQ(t)\lambda\sigma^{-1}\leqq MQ(t)\lambda^{1/2},\qquad \lambda\geqq 1,\quad t\geqq 1. \tag{6.4}$$

Using (5.2) and (6.2) it follows that

$$|v(t,\lambda)|\leqq M\lambda\sigma^{-1}t^{-1},\qquad \lambda\geqq 1,\quad t>0. \tag{6.5}$$

Combining (2.12) and (6.5) yields

$$|v(t,\lambda)|=|v(t,\lambda)|^{1/2+1/2}\leqq M\sigma^{1/2}(\lambda\sigma^{-1}t^{-1})^{1/2}=M\lambda^{1/2}t^{-1/2},\quad \lambda\geqq 1,\quad 0<t\leqq 1. \tag{6.6}$$

Theorem 2.4 is an easy consequence of (6.4) and (6.6).

Proof of Corollary 2.2. If $a(0+)<\infty$, then [14, Corollaries 2.1 and 2.2] imply that c is strongly positive. Then a is strongly positive. As noted in § 2, strong positivity implies (2.18) (which in turn implies (2.14)) in this case, so our assertion follows from Theorems 2.3(i) and 2.4.

If $a(0+)=\infty$, we follow the proof of [2, Cor. 2.1(ii)] for this case. There we invoked [16, Thm. 2(iii)] to obtain

$$\varphi(\tau)\geqq\frac{\alpha}{8\beta^2}A^2(\tau^{-1}),\qquad \tau\geqq\max\{\rho, x_0^{-1}\} \tag{6.7}$$

(α, β, x_0 are positive constants whose values are irrelevant here) at an intermediate stage of the proof. Since $A(\tau^{-1})\geqq\tau A_1(\tau^{-1})$ and (2.4) holds, we deduce from (6.7) that

$$\varphi(\tau)\geqq\frac{\alpha}{8\beta^2}\left[\frac{\tau\theta(\tau)}{12}\right]^2,\qquad \tau\geqq\max\{\rho, x_0^{-1}\}.$$

But φ and θ are continuous, so (2.18) holds, and our conclusions follow as before. This completes the proof.

7. Proof of Theorem 2.5. Each example has the form

$$a(t)=\sum_{k=0}^{\infty}c_kb_k(t), \tag{7.1}$$

where

$$b_k(t)=(1-2^{\beta^k}t)\chi_k(t);$$

χ_k is the characteristic function of the interval $[0, 2^{-\beta^k}]$ and β is an integer greater than or equal to 2.

Each c_k will be positive, and we shall have $2A(\infty)=\sum_{k=0}^{\infty}c_k2^{-\beta^k}<\infty$; then (1.2) and (2.2) hold. (2.13) is clear because $a(\frac{1}{2})=0$.

For any kernel of the form (7.1),

$$\varphi(\tau)=\sum_{k=0}^{\infty} c_k 2^{\beta^k}\frac{(1-\cos 2^{-\beta^k}\tau)}{\tau^2}. \tag{7.2}$$

Note that

$$\tfrac{1}{4}u^2\leqq 1-\cos u\leqq \tfrac{1}{2}u^2,\qquad 0\leqq u\leqq 1. \tag{7.3}$$

Therefore,

$$\varphi(\tau)\geqq\frac{c_{m+1}}{4}2^{-\beta^{m+1}}\geqq\frac{c_{m+1}}{4\tau^{\beta}},\qquad 2^{\beta^m}\leqq\tau\leqq 2^{\beta^{m+1}}. \tag{7.4}$$

On the other hand, if we let $\tau_n=2^{\beta^n}(2\pi)$, (7.2) and (7.3) show that

$$\varphi(\tau_n)\leqq\frac{1}{2}\sum_{k=n+1}^{\infty} c_k 2^{-\beta^k}. \tag{7.5}$$

From (2.4), we get

$$\begin{aligned}\frac{1}{12}\theta(\tau)&\leqq\sum_{k=0}^{m} c_k\int_0^{1/\tau} t(1-2^{\beta^k}t)\,dt+\sum_{k=m+1}^{\infty} c_k\int_0^{2^{-\beta^k}} t(1-2^{\beta^k}t)\,dt\\&\leqq\frac{1}{2\tau^2}\sum_{k=0}^{n} c_k+\frac{1}{6}\sum_{k=m+1}^{\infty} c_k 2^{-2\beta^k},\qquad 2^{\beta^m}\leqq\tau\leqq 2^{\beta^{m+1}},\end{aligned} \tag{7.6}$$

$$5\theta(\tau_n)\geqq\sum_{k=0}^{n} c_k\int_0^{1/\tau_n} t(1-\tau_n t)\,dt\geqq\frac{1}{6\tau_n^2}\sum_{k=0}^{n} c_k. \tag{7.7}$$

Now we need only choose β and c_k appropriately.
For a_1, let $\beta=2$, $c_k=k2^{2^{k-2}}$, $0<\varepsilon<\frac{1}{14}$. By (7.4), if $2^{2^m}\leqq\tau\leqq 2^{2^{m+1}}$,

$$\varphi(\tau)\geqq\frac{m+1}{4}2^{-3(2)^{m/2}}\geqq\log_2\log_2\tau/4\tau^{3/2}.$$

Since (7.6) holds and

$$(x+y)^r\leqq(2x)^r+(2y)^r,\qquad x,y,r>0, \tag{7.8}$$

$$\frac{1}{150}[\theta(\tau)]^{3/2-\varepsilon}\leqq\left(\frac{1}{2}\tau^{-2}\sum_{k=0}^{m} c_k\right)^{3/2-\varepsilon}+\left(\frac{1}{6}\sum_{k=m+1}^{\infty} k2^{-7(2)^{k-2}}\right)^{3/2-\varepsilon},\qquad 2^{2^m}\leqq\tau\leqq 2^{2^{m+1}}.$$

The first sum on the right is dominated by c_m; the second sum is dominated by its first term. Thus

$$\begin{aligned}\frac{1}{150}[\theta(\tau)]^{3/2-\varepsilon}&\leqq(\tau^{-2}c_m)^{3/2-\varepsilon}+((m+1)2^{-7(2)^{m-1}})^{3/2-\varepsilon}\\&\leqq(1+\log_2\log_2\tau)^{3/2}(\tau^{-21/8+7\varepsilon/4}+2^{-5(2)^m}),\qquad 2^{2^m}\leqq\tau\leqq 2^{2^{m+1}}.\end{aligned}$$

Since θ and φ are continuous, we deduce that (2.15) holds with $\gamma=\frac{1}{2}$. By (7.5) and (7.7),

$$\varphi(\tau_n)\leqq(n+1)2^{-3(2)^{n-1}},$$

$$900\tau_n^2\theta^2(\tau_n)\geqq\frac{n^2}{4\pi^2}2^{-3(2)^{n-1}},$$

so $\tau_n^2\theta^2(\tau_n)/\varphi(\tau_n)\to\infty(n\to\infty)$, and (2.18) does not hold.

For $a_{2,\gamma}$, let $\beta = 1+2\gamma$, $c_k = 2^{(1+\gamma)^k}$. Note that

$$\frac{c_{k+1}2^{-2\beta^{k+1}}}{c_k 2^{-2\beta^k}} < 2^{-3\gamma} < 1, \qquad \frac{c_{k+1}}{c_k} > 2^{\gamma} > 1.$$

Therefore, (7.6) implies

$$\theta(\tau) \leqq K(\gamma)\left[\frac{c_m}{\tau^2} + c_{m+1}2^{-2\beta^{(m+1)}}\right], \qquad 2^{\beta^m} \leqq \tau \leqq 2^{\beta^{m+1}} \tag{7.9}$$

for some number $K(\gamma) < \infty$. Using (7.4), (7.8) and (7.9), we get

$$\begin{aligned}\frac{\tau\theta^{1+\gamma}(\tau)}{\varphi(\tau)} &\leqq 2^{1+\gamma}K(\gamma)\left(\frac{c_{m+1}\tau^{-2(\gamma+1)+1}}{c_{m+1}\tau^{-(1+2\gamma)}} + \frac{c_{m+1}^{1+\gamma}2^{(-1-2\gamma)\beta^{m+1}}}{c_{m+1}2^{-\beta^{(m+1)}}}\right)\\ &\leqq 2^{1+\gamma}K(\gamma)[1+2^{\gamma[(1+\gamma)^{m+1}-2(1+2\gamma)^{m+1}]}]\\ &\leqq 2^{2+\gamma}K(\gamma), \qquad 2^{\beta^m} \leqq \tau \leqq 2^{\beta^{m+1}}.\end{aligned}$$

Thus, (2.14) holds. On the other hand, (7.5) and (7.7) yield

$$\varphi(\tau_n) \leqq \frac{1}{2}\left(1+\frac{1}{1-2^{-\gamma}}\right)c_{n+1}\left(\frac{2\pi}{\tau_n}\right)^{\beta}, \qquad 30\theta(\tau_n) \geqq c_n\tau_n^{-2}. \tag{7.10}$$

From (7.10), our final conclusions about $a_{2,\gamma}$ follow easily.

For a_3, we take $\beta = 2$, $c_k = 2^k$, and for a_4, we take $\beta = 2$, $c_k = 1$. The estimates are similar to those given above, so we omit them. Example a_4 appeared in [2].

8. A lemma. In this section we prove:

LEMMA 8.1. *If* (1.2) *holds, then*

$$|D(\tau,\lambda)| \geqq M\tau A_1(\tau^{-1}), \qquad \tfrac{1}{2}\rho \leqq \tau \leqq \tfrac{1}{2}\omega, \tag{8.1}$$

$$|D(\tau,\lambda)| \geqq M\frac{|\tau-\omega|}{\lambda}, \qquad \tau \geqq \tfrac{1}{2}\omega. \tag{8.2}$$

This is the same as [2, Lemma 5.2], but our proof in [2] contains an error.

Proof. When $\frac{1}{2}\rho \leqq \tau < \rho = \omega$ or $\tau \geqq \omega$, the proof in [2] is valid, so we exclude those cases here. [2, (5.11)] is not correct when $\tau < \omega$. When we integrate the inequality

$$-\theta'(\tau) \geqq \frac{\tau}{5}\int_0^{1/\tau} r^3 a(r)\,dr \geqq \frac{1}{80\tau^3}\,a(\tau^{-1})$$

[2,(4.4)] from τ to ω, we obtain

$$\begin{aligned}|\operatorname{Im} D(\tau,\lambda)| \geqq{}& \frac{\tau(\omega-\tau)(\omega+\tau)}{20}\int_0^{1/\omega} r^3 a(r)\,dr\\ &+\frac{\tau}{160}\int_{1/\omega}^{1/\tau} ra(r)\,dr, \qquad \rho < \omega, \quad \frac{1}{2}\rho \leqq \tau < \omega.\end{aligned} \tag{8.3}$$

Since

$$\int_{1/\omega}^{2/\omega} r^3 a(r)\,dr \leqq 15\int_0^{1/\omega} r^3 a(r)\,dr,$$

we have

$$16\int_0^{1/\omega} r^3 a(r)\,dr \geqq \int_0^{2/\omega} r^3 a(r)\,dr \geqq \int_0^{1/\tau} r^3 a(r)\,dr, \qquad \tau \geqq \frac{1}{2}\omega. \tag{8.4}$$

By (8.3) and (8.4),

$$|\operatorname{Im} D(\tau,\lambda)| \geqq \frac{\tau(\omega-\tau)(\tau+\omega)}{320}\int_0^{1/\tau} r^3 a(r)\,dr + \frac{\tau}{160}\int_{1/\omega}^{1/\tau} ra(r)\,dr \tag{8.5}$$

for $\rho<\omega$, $\frac{1}{2}\omega \leqq \tau < \omega$. Except for a constant, this is the same as [2, (5.11)] for these τ, ω, so the remainder of the proof of (8.2) as given in [2] is valid. We need only establish (8.1).

Note that

$$|D(\tau,\lambda)| \geqq \varphi(\tau) \geqq M \geqq M\tau A_1(\tau^{-1}), \qquad \tfrac{1}{2}\rho \leqq \tau \leqq \rho. \tag{8.6}$$

If $\frac{1}{2}\omega \leqq \rho$, then (8.6) implies (8.1). Otherwise we consider two cases.

Case 1. If $\rho \leqq \tau \leqq \frac{1}{2}\omega$ and $A(\omega^{-1}) \geqq 6\omega A_1(\omega^{-1})$, then, as in the proof of Lemma 5.1,

$$\begin{aligned} |\operatorname{Re} D(\tau,\lambda)| = \varphi(\tau) &\geqq \frac{1}{4}\int_0^{1/\tau} t^2\,da'(t) \\ &\geqq \frac{1}{4}\int_0^{1/\omega} t^2\,da'(t) \\ &\geqq \frac{1}{2}[A(\omega^{-1}) - 3\omega A_1(\omega^{-1})] + \frac{1}{4\omega}a(\omega^{-1}) \\ &\geqq \tfrac{1}{4}A(\omega^{-1}) \geqq \tfrac{1}{4}\omega A_1(\omega^{-1}) \geqq \tfrac{1}{4}\tau A_1(\omega^{-1}). \end{aligned} \tag{8.7}$$

Thus, (8.3) and (8.7) imply

$$\begin{aligned} \sqrt{2}\,|D(\tau,\lambda)| &\geqq \frac{\tau}{160}\int_{1/\omega}^{1/\tau} ra(r)\,dr + \frac{\tau}{4}\int_0^{1/\omega} ra(r)\,dr \\ &\geqq \frac{\tau}{160}A_1(\tau^{-1}) \quad \text{in Case 1.} \end{aligned} \tag{8.8}$$

Case 2. If $\rho \leqq \tau \leqq \frac{1}{2}\omega$ and $A(\omega^{-1}) < 6\omega A_1(\omega^{-1})$, then let $g(t) = (6\omega t - 1)a(t)$. In Case 2, then, $\int_0^{1/\omega} g(t)\,dt > 0$.

It is easy to see that $(6\omega t)^n g(t) \geqq g(t)$, $(t>0,\ n=1,2)$, so we conclude that

$$(6\omega)^n \int_0^{1/\omega} t^n g(t)\,dt > 0, \qquad n = 1, 2. \tag{8.9}$$

From (8.9) it follows that

$$36\omega^2 \int_0^{1/\omega} t^3 a(t)\,dt > \int_0^{1/\omega} ta(t)\,dt. \tag{8.10}$$

Now (8.3) implies

$$|\operatorname{Im} D(\tau,\lambda)| \geqq \frac{\tau\omega^2}{40}\int_0^{1/\omega} r^3 a(r)\,dr + \frac{\tau}{160}\int_{1/\omega}^{1/\tau} ra(r)\,dr, \qquad \rho \leqq \tau \leqq \frac{1}{2}\omega. \tag{8.11}$$

(8.10) and (8.11) combine to yield

$$\begin{aligned} |\operatorname{Im} D(\tau,\lambda)| &\geqq \frac{\tau}{1440}\int_0^{1/\omega} ra(r)\,dr + \frac{\tau}{160}\int_{1/\omega}^{1/\tau} ra(r)\,dr \\ &\geqq \frac{\tau}{1440}A_1(\tau^{-1}) \quad \text{in Case 2.} \end{aligned} \tag{8.12}$$

Finally, (8.6), (8.8) and (8.12) establish (8.1) in all cases. This completes the proof of Lemma 8.1.

9. Proofs of Theorems 3.1, 3.2 and 3.3. For Theorem 3.1(i), first observe that Theorem 2.3(i) implies

$$|v(t,\lambda)| \leqq Mt^{-1}\lambda^{1/2}, \qquad t>0. \tag{9.1}$$

By (1.14), $\mathbf{V}(t)\mathbf{L}^{-1/2}$ is bounded, for each $t>0$. Moreover, if $t, s>0$ and $\mathbf{y}\in\mathscr{H}$,

$$\|[\mathbf{V}(t)-\mathbf{V}(s)]\mathbf{L}^{-1/2}\mathbf{y}\|^2 = \int_1^\infty [v(t,\lambda)-v(s,\lambda)]^2\lambda^{-1}\, d(\mathbf{E}_\lambda\mathbf{y},\mathbf{y}).$$

Since $v(t,\lambda)$ is continuous in t and (9.1) holds, Lebesgue's Dominated Convergence Theorem shows that $\mathbf{V}(t)\mathbf{L}^{-1/2}\mathbf{y}\to\mathbf{V}(s)\mathbf{L}^{-1/2}\mathbf{y}(t\to s)$. $v(t,\lambda)$ is differentiable in t, so the Mean Value Theorem implies

$$\|h^{-1}[\mathbf{U}(t+h)-\mathbf{U}(t)-h\mathbf{V}(t)]\mathbf{y}\|^2 = \int_1^\infty \left|\frac{v(t+\eta,\lambda)-v(t,\lambda)}{\lambda^{1/2}}\right|^2 \lambda\, d(\mathbf{E}_\lambda\mathbf{y},\mathbf{y}) \tag{9.2}$$

($\mathbf{y}\in\mathscr{D}_1$), where $\eta=\eta(t,\lambda,h)$ is between 0 and h. For $\mathbf{y}\in\mathscr{D}_1$, $\lambda d(\mathbf{E}_\lambda\mathbf{y},\mathbf{y})$ is a finite measure, so by (9.1) and dominated convergence, the integral in (9.2) tends to zero as $h\to 0$. Therefore, $\mathbf{U}(t)\mathbf{y}$ is differentiable ($t>0$) and (3.1) holds. This proves Theorem 3.1(i).

Under the hypotheses of Theorem 3.1(ii),

$$\sup_{t\geqq 0}|v(t,\lambda)| \leqq M\sigma \leqq Ma(0+)\lambda^{1/2} \tag{9.3}$$

(see Theorem 2.2 and the remarks following it). Using (9.3) in place of (9.1), we can argue as above and prove the results on the closed interval $\bar{\mathbb{R}}^+$. This completes the proof of Theorem 3.1.

Proof of Theorem 3.2. To simplify formulas, we take $d=0$, since this does not change the argument. For (i), the uniqueness assertion and the special case $\mathbf{g}\equiv\mathbf{0}$ are just Theorem 2.1(ii) of [2]. Therefore, it suffices to establish (1.11) when $\mathbf{f}\equiv\mathbf{0}$ and $\mathbf{z}_0=0$.

Let n be a positive integer, and let $\mathbf{g}_n=\mathbf{E}_n\mathbf{g}$,

$$\mathbf{h}_n(t)=\int_0^t a(t-s)\mathbf{g}_n(s)\, ds.$$

Then $\mathbf{g}_n\in\mathscr{B}_{\mathrm{loc}}(\bar{\mathbb{R}}^+,\mathscr{H})$. Since $\|\mathbf{L}\mathbf{E}_n\|\leqq n$ and $\mathbf{L}\mathbf{g}_n=\mathbf{E}_n\mathbf{L}\mathbf{g}$ is measurable, $\mathbf{g}_n$ belongs to $\mathscr{B}^\infty_{\mathrm{loc}}(\bar{\mathbb{R}}^+,\mathscr{D})$. Therefore, $\mathbf{h}_n:\bar{\mathbb{R}}^+\to\mathscr{D}$ is continuous. By [2, Thm. 2.1], the unique solution of

$$\mathbf{z}'(t)+\int_0^t a(t-s)[\mathbf{L}\mathbf{z}(s)+\mathbf{g}_n(s)]\, ds=\mathbf{0}, \qquad \mathbf{z}(0)=\mathbf{0},$$

is

$$\mathbf{z}_n(t)=-\int_0^t \mathbf{U}(t-s)\mathbf{h}_n(s)\, ds.$$

Then $\mathbf{z}_n\in C(\bar{\mathbb{R}}^+,\mathscr{D})$ and

$$\mathbf{z}_n(t)=\int_0^t\int_0^s a(s-r)[\mathbf{L}\mathbf{z}_n(r)+\mathbf{g}_n(r)]\, dr\, ds, \qquad t\geqq 0. \tag{9.4}$$

But for $\mathbf{y}_0 \in \mathscr{D}$, $\mathbf{y}(t) = \mathbf{U}(t)\mathbf{y}_0$ is the solution of

$$\mathbf{y}'(t) + \int_0^t a(t-s)\mathbf{L}\mathbf{y}(s)\,ds = 0, \qquad \mathbf{y}(0) = \mathbf{y}_0$$

[2, Theorem 2.1(i)]. Since $\mathbf{L}$ is closed and (3.1) holds, this means (see Theorem 3.1)

$$\begin{aligned}\mathbf{L}^{-1}\mathbf{V}(t)\mathbf{y}_0 = \mathbf{V}(t)\mathbf{L}^{-1}\mathbf{y}_0 &= -\int_0^t a(t-s)\mathbf{U}(s)\mathbf{y}_0\,ds \\ &= -\int_0^t \mathbf{U}(t-s)a(s)\mathbf{y}_0\,ds, \qquad \mathbf{y}_0 \in \mathscr{D}.\end{aligned}$$

Therefore,

$$\begin{aligned}\int_0^t \mathbf{V}(t-s)\mathbf{L}^{-1}\mathbf{g}_n(s)\,ds &= -\int_0^t \left[\int_0^{t-s} a(t-s-r)\mathbf{U}(r)\mathbf{g}_n(s)\,dr\right] ds \\ &= -\int_0^t \mathbf{U}(r)\int_0^{t-r} a(t-r-s)\mathbf{g}_n(s)\,ds\,dr \\ &= \mathbf{z}_n(t).\end{aligned}$$

Since $\mathbf{z}_0 = \mathbf{f} = \mathbf{0}$, (1.11) reduces to

$$\mathbf{z}(t) = \int_0^t \mathbf{V}(t-s)\mathbf{L}^{-1}\mathbf{g}(s)\,ds,$$

but $\mathbf{V}(\cdot)\mathbf{L}^{-1/2}$ is strongly continuous on $\mathbb{R}^+$ and $\|\mathbf{V}(\cdot)\mathbf{L}^{-1/2}\| \in L^1(\mathbb{R}^+)$, while $\mathbf{g} \in \mathscr{B}^\infty_{\mathrm{loc}}(\bar{\mathbb{R}}^+, \mathscr{D}_1)$. Therefore,

$$\mathbf{V}(t-s)\mathbf{L}^{-1}\mathbf{g}(s) \quad \text{and} \quad \mathbf{L}\mathbf{V}(t-s)\mathbf{L}^{-1}\mathbf{g}(s) = \mathbf{V}(t-s)\mathbf{L}^{-1/2}\cdot\mathbf{L}^{1/2}\mathbf{g}(s)$$

are strongly measurable in s (a modified version of [2, Lemma 3.1] shows this), and standard estimates show that $\mathbf{z} \in C(\bar{\mathbb{R}}^+, \mathscr{D})$. Then by (1.14),

$$\begin{aligned}\|L[\mathbf{z}(t) - \mathbf{z}_n(t)]\| &\leqq \int_0^t \|\mathbf{V}(t-s)\mathbf{L}^{-1/2}\|\,\|\mathbf{L}^{1/2}[\mathbf{g}(s) - \mathbf{g}_n(s)\|\,ds \\ &\leqq \int_0^t v_{1/2}(s)\|(\mathbf{I} - \mathbf{E}_n)\mathbf{L}^{1/2}\mathbf{g}(t-s)\|\,ds.\end{aligned}$$

But $\mathbf{E}_n \to \mathbf{I}$ strongly ($n \to \infty$), and the integrand here is dominated by the L^1 function

$$w(s) = v_{1/2}(s) \operatorname*{ess\,sup}_{0 \leqq r \leqq t} \|\mathbf{L}^{1/2}\mathbf{g}(r)\|,$$

so

$$\|\mathbf{L}[\mathbf{z}(t) - \mathbf{z}_n(t)]\| \leqq \int_0^T w(s)\,ds, \qquad 0 \leqq t \leqq T < \infty,$$

$$\mathbf{L}\mathbf{z}_n(t) \to \mathbf{L}\mathbf{z}(t) \quad \text{in} \quad \mathscr{H}, \qquad n \to \infty, \quad t \geqq 0.$$

Similarly, $\mathbf{z}_n(t) \to \mathbf{z}(t)$ ($n \to \infty$) and $\mathbf{z} - \mathbf{z}_n$ is bounded on finite intervals. Therefore, we can let $n \to \infty$ in (9.4), using dominated convergence, and deduce that

$$\mathbf{z}(t) = \int_0^t \int_0^s a(s-r)[\mathbf{L}\mathbf{z}(r) + \mathbf{g}(r)]\,dr\,ds.$$

Therefore, $\mathbf{z}(t)$ is a solution of (1.10) with $\mathbf{z}_0 = \mathbf{f} = \mathbf{0}$, as asserted. For (ii), the hypotheses imply $v_{1/2}(t) \leqq M$ (see Theorem 2.2), so the proof of (i) can be repeated with minor changes. This proves Theorem 3.2.

Proof of Theorem 3.3. By (4.1), Theorem A of § 2, and the fact that (2.14) ($\gamma = \frac{1}{2}$) implies (2.21), our hypotheses yield

$$\|\mathbf{U}(t)\|_{\mathscr{D}} \leqq 1 (t \geqq 0), \qquad \int_0^\infty \|\mathbf{U}(t)\|_{\mathscr{D}}\, dt \equiv \nu < \infty. \tag{9.5}$$

Let $\mathbf{T}: \mathbf{g} \to \mathbf{z}$ be the operator defined formally by the right-hand side of (1.11) with $\mathbf{y}_0$ in place of $\mathbf{z}_0$, but interpret the integrals as Bochner integrals in $\mathscr{B}^1((0, t), \mathscr{D})$. If $\mathbf{g} \in \mathscr{B}^\infty(\mathbb{R}, \mathscr{D}_1)$, Theorem 3.2(i) shows that $\mathbf{Tg} \in C(\bar{\mathbb{R}}^+, \mathscr{D})$. Moreover, by (9.5) and (1.14),

$$\|\mathbf{Tg}(t)\|_{\mathscr{D}} \leqq \|\mathbf{y}_0\|_{\mathscr{D}} + \|\mathbf{f}_1\|_{\mathscr{B}^1(\mathbb{R}^+, \mathscr{D})} + \nu \|\mathbf{f}_2\|_{\mathscr{B}^\infty(\mathbb{R}^+, \mathscr{D})} + \|\mathbf{g}\|_{\mathscr{B}^\infty(\mathbb{R}^+, \mathscr{D}_1)} \int_0^\infty v_{1/2}(t)\, dt, \qquad t \in \bar{\mathbb{R}}^+.$$

With $K = 1 + \nu + \|v_{1/2}\|_{L^1}$,

$$\|\mathbf{Tg}(t)\|_{\mathscr{D}} \leqq K(\mu + \|\mathbf{g}\|_{\mathscr{B}^\infty(\mathbb{R}^+, \mathscr{D}_1)}), \qquad t \in \bar{\mathbb{R}}^+.$$

Referring to our hypotheses, we choose Δ, $0 < \Delta < \alpha$, so small that $K \in (\Delta) < \frac{1}{2}$, and choose $\mu \leqq \Delta / 2K$. Then $\mathbf{TN}$ maps the ball

$$S_\Delta = \{\mathbf{y} \mid \|y(t)\|_{\mathscr{D}} \leqq \Delta,\ t \in \bar{\mathbb{R}}^+\}$$

in the Banach space $C(\bar{\mathbb{R}}^+, \mathscr{D})$ into itself. $\mathbf{y} \in S_\Delta$ is a fixed point of $\mathbf{TN}$ if and only if $\mathbf{y}$ is a solution of (1.12) in S_Δ.

We complete the proof by showing that $\mathbf{TN}$ is a contraction on S_Δ. For $\mathbf{z}_1, \mathbf{z}_2 \in S_\Delta$,

$$\begin{aligned}
\|\mathbf{TNz}_1(t) - \mathbf{TNz}_2(t)\|_{\mathscr{D}} &= \left\| \int_0^t \mathbf{V}(t-s)\mathbf{L}^{-1}[\mathbf{Nz}_1(s) - \mathbf{Nz}_2(s)]\, ds \right\|_{\mathscr{D}} \\
&\leqq K \|\mathbf{Nz}_1 - \mathbf{Nz}_2\|_{\mathscr{B}^\infty(\mathbb{R}^+, \mathscr{D}_1)} \\
&\leqq K \in (\Delta) \|z_1 - z_2\|_{\mathscr{B}^\infty(\mathbb{R}^+, \mathscr{D})} \\
&< \tfrac{1}{2} \|\mathbf{z}_1 - \mathbf{z}_2\|_{C(\bar{\mathbb{R}}^+, \mathscr{D})}.
\end{aligned}$$

This proves Theorem 3.3.

REFERENCES

1. R. W. CARR, *Uniform L^p estimates for a linear integrodifferential equation with a parameter*, Ph.D. Thesis, University of Wisconsin-Madison, 1977.
2. R. W. CARR AND K. B. HANNSGEN, *A nonhomogeneous integrodifferential equation in Hilbert space*, this Journal, 10 (1979), pp. 961–984.
3. C. M. DAFERMOS AND J. A. NOHEL, *Energy methods for nonlinear hyperbolic Volterra integrodifferential equations*, Comm. Partial Differential Equations, 4 (1979), pp. 219–278.
4. K. B. HANNSGEN, *Indirect Abelian theorems and a linear Volterra equation*, Trans. Amer. Math. Soc., 142 (1969), pp. 539–555.
5. ———, *A Volterra equation with parameter*, this Journal, 4 (1973), pp. 22–30.
6. ———, *Uniform L^1 behavior for an integrodifferential equation with parameter*, this Journal, 8 (1977), pp. 626–639.
7. E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, American Mathematical Society, Providence, RI, 1957.
8. J. J. LEVIN, *The asymptotic behavior of the solution of a Volterra equation*, Proc. Amer. Math. Soc., 14 (1963), pp. 534–541.

9. S.-O. LONDEN, *The qualitative behavior of the solutions of a nonlinear Volterra equation*, Michigan Math. J., 18 (1971), pp. 321–330.
10. ———, *An existence result on a Volterra equation in Banach space*, Trans. Amer. Math. Soc., 235 (1978), pp. 285–305.
11. R. C. MACCAMY, *An integro-differential equation with applications in heat flow*, Quart. Appl. Math., 35 (1977), pp. 1–19.
12. ———, *A model for one-dimensional, nonlinear viscoelasticity*, Quart. Appl. Math., 35 (1977), pp. 21–33.
13. R. K. MILLER, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.
14. J. A. NOHEL AND D. F. SHEA, *Frequency domain methods for Volterra equations*, Adv. Math., 22 (1976), pp. 278–304.
15. D. F. SHEA AND S. WAINGER, *Variants of the Wiener–Lévy theorem, with applications to stability problems for some Volterra integral equations*, Amer. J. Math., 97 (1975), pp. 312–343.
16. O. J. STAFFANS, *An inequality for positive definite Volterra kernels*, Proc. Amer. Math. Soc., 58 (1976), pp. 205–210.
17. ———, *On a nonlinear hyperbolic Volterra equation*, this Journal, 11 (1980), pp. 793–812.
18. C. C. TRAVIS AND G. F. WEBB, *An abstract second order semilinear Volterra integrodifferential equation*, this Journal, 10 (1979), pp. 412–424.

SIAM J. MATH. ANAL.
Vol. 13, No. 3, May 1982

0036-1410/82/1303-0007 $01.00/0

A THEORY FOR THE APPROXIMATION OF SOLUTIONS OF BOUNDARY VALUE PROBLEMS ON INFINITE INTERVALS*

PETER A. MARKOWICH†

Abstract. An ad hoc method to solve boundary value problems which are posed on infinite intervals is to reduce the infinite interval to a finite but large one and to impose additional boundary conditions at the far end. These boundary conditions should be posed in a way so that they express the asymptotic behavior of the actual solution well. In this paper a theory is derived which defines classes of appropriate additional boundary conditions. Appropriate is to be understood in the sense that the solutions of the approximate problems converge to the actual solution of the "infinite" problem as the length of the finite interval tends to infinity. Moreover, boundary conditions which produce convergence with the largest expectable order are devised.

1. Introduction. Boundary value problems on infinite intervals, which are posed in the following way:

$$y' = t^{\alpha} f(t, y), \qquad 1 \leqq t < \infty, \quad \alpha \in \mathbb{N}_0, \tag{1.1}$$

$$y \in C([1, \infty]) :\Leftrightarrow y \in C([1, \infty)) \text{ and } \lim_{t \to \infty} y(t) \text{ exists}, \tag{1.2}$$

$$b(y(1)) = 0 \tag{1.3}$$

where $f: \mathbb{R}^{n+1} \to \mathbb{R}^n$ and $\mathbb{N}_0$ is the set of nonnegative integers, are often solved numerically by restricting the infinite interval to a finite but large one and by imposing additional suitable boundary conditions at the right end. The resulting two-point boundary value problem has the following form:

$$x_T' = t^{\alpha} f(t, x_T), \quad 1 \leqq t \leqq T, \quad T \gg 1, \tag{1.4}$$

$$b(x_T(1)) = 0, \tag{1.5}$$

$$S(x_T(T), T) = 0, \tag{1.6}$$

and can be solved by any appropriate code. The questions this paper answers are the following:

1) What class of asymptotic boundary conditions $S(x_T(T), T) = 0$ implies convergence in the following sense

$$\|x_T - y\|_{[1,T]} \to 0 \quad \text{as } T \to \infty \tag{1.7}$$

where $\|z\|_{[a,b]} := \sup_{t \in [a,b]} |z(t)|$?

2) Which asymptotic boundary conditions yield a reasonably fast order of convergence?

It will be shown that the admissible boundary conditions have to be constructed with regard to the invariant subspaces and eigenspaces of the matrix

$$A_0(y_\infty) := f_y(\infty, y(\infty)). \tag{1.8}$$

* Received by the editors December 11, 1980, and in revised form June 8, 1981. This research was supported by the U.S. Army under contract DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under grant MCS-79-27062.

† Department of Mathematics, University of Texas at Austin, Austin, Texas 78712.

2. Linear constant coefficient problems. The problem

$$y' - t^{\alpha} A y = t^{\alpha} f(t), \quad 1 \leqq t < \infty, \quad \alpha \in \mathbb{R}, \quad \alpha > -1, \tag{2.1}$$

$$y \in C([1, \infty]), \tag{2.2}$$

$$By(1) = \beta \tag{2.3}$$

shall be approximated by the "finite" problem

$$x_T' - t^{\alpha} A x_T = t^{\alpha} f(t), \qquad 1 \leqq t \leqq T, \tag{2.4}$$

$$Bx_T(1) = \beta, \tag{2.5}$$

$$S(T) x_T(T) = \gamma(T) \tag{2.6}$$

as T approaches infinity. A is assumed to be a real $n \times n$ matrix with the Jordan form J:

$$A = EJE^{-1} \neq \theta \tag{2.7}$$

(from here on θ stands for a zero matrix of appropriate dimension) and J has the block diagonal form

$$J = \operatorname{diag}(J^{+}, J^{0}, J^{-}) \tag{2.8}$$

where J^{+} contains the eigenvalues of A with positive real part, J^{0} the eigenvalues of A with a zero real part and J^{-} the eigenvalues of A with a negative real part. The dimensions of these three matrices are r_{+}, r_0, r_{-} respectively, and the geometrical multiplicity of the eigenvalue zero will be called $\bar{r}_0$ (geometrical multiplicity refers to the number of independent eigenvectors).

The projection-like matrices G_{+}, G_0, G_{-} and $\bar{G}_0$ are obtained by taking the matrices D_{+}, D_0, D_{-} and $\bar{D}_0$, which are the projections onto the direct sums of invariant subspaces of J belonging to eigenvalues with positive, zero, negative real part respectively, and onto the direct sums of eigenspaces belonging to zero eigenvalues of J^{0}, and by cancelling all columns of these matrices which have only zero entries. So G_{+} is $n \times r_{+}$, G_0 is $n \times r_0$, G_{-} is $n \times r_{-}$ and $\bar{G}_0$ is $n \times \bar{r}_0$.

By substituting

$$u = E^{-1} y \tag{2.9}$$

we get the problem

$$u' - t^{\alpha} J u = t^{\alpha} E^{-1} f(t), \qquad 1 \leqq t < \infty, \tag{2.10}$$

$$u \in C([1, \infty]), \tag{2.11}$$

$$BEu(1) = \beta. \tag{2.12}$$

The general solution of (2.10), (2.11) is

$$u(t) = [\phi(t)\bar{G}_0, \phi(t) G_{-}]\eta + (Hf)(t), \qquad \eta \in C^{\bar{r}_0 + r_{-}} \tag{2.13}$$

where

$$\phi(t) = \exp\left(\frac{J}{\alpha + 1} t^{\alpha+1}\right) \tag{2.14}$$

and H is a solution operator for the inhomogeneous problem:

$$Hf = H_{+} f + H_0 f + H_{-} f,$$

where

$$(H_+f)(t)=\phi(t)\int_\infty^t D_+\phi^{-1}(s)E^{-1}f(s)s^\alpha\,ds, \tag{2.15}$$

$$(H_0f)(t)=\phi(t)\int_\infty^t D_0\phi^{-1}(s)E^{-1}f(s)s^\alpha\,ds, \tag{2.16}$$

$$(H_-f)(t)=\phi(t)\int_\delta^t D_-\phi^{-1}(s)E^{-1}f(s)s^\alpha\,ds \tag{2.17}$$

holds for $\delta\geqq 1$. f is assumed to be in $C([1,\infty])$ and in order to make the integral in (2.16) exist we assume that

$$\|D_0E^{-1}f(t)\|=O(t^{-(\alpha+1)r-\varepsilon}),\qquad \varepsilon>0. \tag{2.18}$$

Here r is the dimension of the largest Jordan block of J associated with an eigenvalue with zero real part. We assume that $r>0$ because the case $r=0$ has been treated by de Hoog and Weiss (1980b). An analysis of the operator H can be found in de Hoog and Weiss (1980a) and Lentini and Keller (1980a). Markowich (1980b) has shown the following estimates, which hold for $t\geqq\delta\geqq 1$:

$$\|(H_+f)(t)\|\leqq\text{const.}\,\|D_+E^{-1}f\|_{[t,\infty]}, \tag{2.19}$$

$$\|(H_0f)(t)\|\leqq\text{const.}\,t^{-\varepsilon}\max_{s\geqq t}\|s^{(\alpha+1)r+\varepsilon}D_0E^{-1}f(s)\|, \tag{2.20}$$

$$\|(H_-f)(t)\|\leqq\text{const.}\,t^{-\gamma}\max_{\delta\leqq s\leqq t}\|s^\gamma D_-E^{-1}f(s)\|\quad\text{for }\gamma\geqq 0. \tag{2.21}$$

All constants are independent of f and δ. Moreover we assume that B is a $(\bar r_0+r_-)\times n$ matrix, that $\beta\in\mathbb{R}^{\bar r_0+r_-}$ and that the $(\bar r_0+r_-)\times(\bar r_0+r_-)$ matrix

$$BE[\phi(1)\bar G_0,\phi(1)G_-] \tag{2.22}$$

is nonsingular so that (2.3) defines $\bar r_0+r_-$ independent boundary conditions. According to Markowich (1980a, b) these propositions are necessary and sufficient for the unique solvability of the problems (2.1), (2.2) and (2.3) for all appropriate f's and $\beta\in\mathbb{R}^{\bar r_0+r}$. Therefore $S(T)$ has to be a $(r_++(r_0-\bar r_0))\times n$ matrix and $\gamma(T)\in\mathbb{R}^{r_++(r_0-\bar r_0)}$ so that (2.5) and (2.6) set up n boundary conditions.

At first we prove a stability estimate for (2.4), (2.5), (2.6):

THEOREM 2.1. *We assume that* (2.22) *holds and that* (A), (B), (C) *which are defined as follows, are fulfilled.*

$$\text{(A)}\qquad \|S(T)\|\leqq\text{const.}\quad \textit{for } T\to\infty,$$

$$\text{(B)}\qquad \|S(T)E\bar G_0\|=o(T^{-(\alpha+1)(r-1)})\quad \textit{for } T\to\infty,$$

$$\text{(C)}\qquad \|[S(T)EG_+,S(T)E\bar{\bar G}_0]^{-1}\|\leqq\text{const.}\quad \textit{for } T\to\infty$$

where $\bar{\bar G}_0$ is the $n\times(r_0-\bar r_0)$ matrix which is obtained by cancelling the columns of the matrix $\bar{\bar D}_0=D_0-\bar D_0$ which have only zero entries.

Then the problem (2.4), (2.5), (2.6) *has a unique solution x_T for all T sufficiently large and x_T fulfills the stability estimate*

$$\|x_T\|_{[1,T]}\leqq\text{const.}\,(\|\beta\|+T^{(\alpha+1)(r-1)}\|\gamma(T)\|+T^{(\alpha+1)r}\|f\|_{[1,T]}) \tag{2.23}$$

if $f\in C([1,T])$, $\beta\in R^{\bar r_0+r_-}$, $\gamma(T)\in\mathbb{R}^{n-(\bar r_0+r_-)}$.

Proof. In order to prove this we first reorder J^0 by permuting its lines and columns so that

$$RJ^0R^{-1} = \begin{bmatrix} J_1^0 & \theta \\ & 0 \ddots 0 \\ J_2^0 & \underbrace{\qquad}_{\bar{r}_0} \end{bmatrix} \Big\}\bar{r}_0 = \tilde{J}^0 \tag{2.24}$$

where R is an appropriate permutation matrix. This corresponds to a reordering of the columns of E. The reordered matrix will be called $\tilde{E}$. More information on this reordering is given by Lentini and Keller (1980a). Now

$$\text{(a) } \bar{G}_0 = \begin{bmatrix} \theta \\ I_{\bar{r}_0} \\ \underbrace{\theta}_{\bar{r}_0} \end{bmatrix} \begin{matrix} \}r_+ + (r_0 - \bar{r}_0) \\ \\ \}r_- \end{matrix}, \qquad \text{(b) } \bar{\bar{G}}_0 = \begin{bmatrix} \theta \\ I_{r_0-\bar{r}_0} \\ \underbrace{\theta}_{r_0-\bar{r}_0} \end{bmatrix} \begin{matrix} \}r_+ \\ \\ \}\bar{r}_0 + r_- \end{matrix} \tag{2.25}$$

holds. From (2.24) it is easily concluded that

$$\exp\left(\frac{\tilde{J}^0}{\alpha+1}(t^{\alpha+1}-1)\right) = \begin{bmatrix} e_1(t^{\alpha+1}-1) & \theta \\ \underbrace{e_2(t^{\alpha+1}-1)}_{r_0-\bar{r}_0} & \underbrace{I_{\bar{r}_0}}_{\bar{r}_0} \end{bmatrix} \begin{matrix} \}r_0-\bar{r}_0 \\ \}\bar{r}_0 \end{matrix}. \tag{2.26}$$

Obviously

$$\exp\left(\frac{\tilde{J}^0}{\alpha+1}\tau\right) = R\exp\left(\frac{J^0}{\alpha+1}\tau\right)R^{-1}, \qquad e_1(\tau) = \exp\left(\frac{J_1^0}{\alpha+1}\tau\right)$$

and

$$e_2(\tau) = J_2^0(J_1^0)^{-1}\left(\exp\left(\frac{J_1^0}{\alpha+1}\tau\right) - I_{r_0-\bar{r}_0}\right)$$

holds for all $\tau \in \mathbb{R}$. (Note that J_1^0 is nonsingular because it has no zero eigenvalue.)

We substitute

$$\tilde{v}_T = \tilde{E}^{-1}x_T \tag{2.27}$$

and get the problem

$$\tilde{v}_T' - t^\alpha \tilde{J}v_T = t^\alpha \tilde{E}^{-1}f(t), \qquad 1 \leqq t \leqq T, \tag{2.28}$$

$$B\tilde{E}v_T(1) = \beta, \tag{2.29}$$

$$S(T)\tilde{E}v_T(T) = \gamma(T) \tag{2.30}$$

where $\tilde{J}$ has the block structure

$$\tilde{J} = \operatorname{diag}(J^+, \tilde{J}^0, J^-). \tag{2.31}$$

We write the general solution of (2.28) as follows:

$$\tilde{v}_T(t) = A(t, T)\xi_1 + C(t)\xi_2 + \tilde{v}_p(t, T) \tag{2.32}$$

where

$$A(t, T) = \begin{bmatrix} \exp\left(\frac{J^+}{\alpha+1}(t^{\alpha+1}-T^{\alpha+1})\right) & \theta \\ \theta & e_1(t^{\alpha+1}-1) \\ \theta & e_2(t^{\alpha+1}-1) \\ \underbrace{\theta}_{r_+} & \underbrace{\theta}_{r_0-\bar{r}_0} \end{bmatrix} \Big\} n \tag{2.33}$$

and

$$C(t)=\left.\begin{bmatrix} \theta & \theta \\ \theta & \theta \\ I & \theta \\ \underbrace{\theta}_{\bar{r}_0} & \underbrace{\exp\left(\frac{J^-}{\alpha+1}t^{\alpha+1}\right)}_{r_-} \end{bmatrix}\right\}^n, \tag{2.34}$$

so that $\xi_1 \in \mathbb{C}^{r_+ + (r_0 - \bar{r}_0)}$ and $\xi_2 \in \mathbb{C}^{\bar{r}_0 + r_-}$ hold. $\tilde{v}_p(t, T)$ is an appropriate particular solution which will be defined later.

From (2.26) we easily derive the following properties of e_1, e_2:

$$e_1(0) = I_{r_0 - \bar{r}_0}, \qquad e_2(0) = \theta, \tag{2.35a, b}$$

$$e_1(t^{\alpha+1}-1)^{-1} = e_1(1-t^{\alpha+1}), \tag{2.36a}$$

$$e_2(t^{\alpha+1}-1)e_1(1-t^{\alpha+1}) = -e_2(1-t^{\alpha+1}) \tag{2.36b}$$

for all $t \in R$. A more general statement than (2.36) is

$$e_1(t_0^{\alpha+1}-1)e_1(1-t_1^{\alpha+1}) = e_1(t_0^{\alpha+1}-t_1^{\alpha+1}), \tag{2.37a}$$

$$e_2(t_0^{\alpha+1}-1)e_1(1-t_1^{\alpha+1}) = e_2(t_0^{\alpha+1}-t_1^{\alpha+1}) - e_2(1-t_1^{\alpha+1}) \tag{2.37b}$$

for all $t_0, t_1 \in \mathbb{R}$.

$$\|e_1(t^{\alpha+1}-1)\| = O(t^{(\alpha+1)(r-1)}) \quad \text{for } t \to \infty, \tag{2.38a}$$

$$\|e_2(t^{\alpha+1}-1)\| = O(t^{(\alpha+1)(r-1)}) \quad \text{for } t \to \infty. \tag{2.38b}$$

The estimate (2.38) is derived by using that

$$\exp\left(\frac{J_k}{\alpha+1}t^{\alpha+1}\right) = \exp\left(i\frac{\gamma}{\alpha+1}t^{\alpha+1}\right)F(t), \tag{2.39}$$

where J_k is an r_k-dimensional Jordan block with the imaginary eigenvalue $i\gamma$ and $F(t)$ is a real matrix whose entries are polynomials of maximal degree $(r_k - 1)(\alpha + 1)$.

By inserting (2.32) into the boundary conditions (2.29), (2.30) we get the linear block system

$$\begin{bmatrix} A_1(T) & A_2 \\ A_3(T) & A_4(T) \end{bmatrix}\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \beta - B\tilde{E}\tilde{v}_p(1, T) \\ \gamma(T) - S(T)\tilde{E}\tilde{v}_p(T, T) \end{pmatrix} \tag{2.40}$$

where

$$A_1(T) = \left[B\tilde{E}G_+ \exp\left(\frac{J^+}{\alpha+1}(1-T^{\alpha+1})\right), B\tilde{E}\bar{G}_0\right], \tag{2.41a}$$

$$A_2 = \left[B\tilde{E}\bar{G}_0, B\tilde{E}G_- \exp\left(\frac{J^-}{\alpha+1}\right)\right], \tag{2.41b}$$

$$A_3(T) = [S(T)\tilde{E}G_+, S(T)\tilde{E}\bar{G}_0 e_1(T^{\alpha+1}-1) + S(T)\tilde{E}\bar{G}_0 e_2(T^{\alpha+1}-1)], \tag{2.41c}$$

$$A_4(T) = \left[S(T)\tilde{E}\bar{G}_0, S(T)\tilde{E}G_- \exp\left(\frac{J^-}{\alpha+1}T^{\alpha+1}\right)\right]. \tag{2.41d}$$

As we will see in (2.52), the system (2.40) is soluble if and only if the matrices A_2 and $A_3(T) - A_4(T)A_2^{-1}A_1(T)$ are invertible. A_2 is invertible by assumption (2.22) and

the existence of $(A_3 - A_4 A_2^{-1} A_1)^{-1}$ has to be proven. We will show that

(2.42)

$$(A_3 - A_4 A_2^{-1} A_1)^{-1} = \begin{bmatrix} I & \theta \\ \theta & e_1(1 - T^{\alpha+1}) \end{bmatrix} [S(T)\tilde{E}G_+, S(T)\tilde{E}\bar{\tilde{G}}_0]^{-1}(I + o(1)) \quad \text{for } T \to \infty.$$

The existence of the right-hand side of (2.42) is assured by proposition (C) of Theorem 2.1.

We split $A_3(T)$ into

(2.43)
$$A_3(T) = [\underbrace{S(T)\tilde{E}G_+, S(T)\tilde{E}\bar{\tilde{G}}_0 e_1(T^{\alpha+1} - 1)}_{A_3^1(T)}] + [\underbrace{\theta, S(T)\tilde{E}\bar{G}_0 e_2(T^{\alpha+1} - 1)}_{A_3^2(T)}],$$

and get

(2.44)
$$A_3^1(T) = [S(T)\tilde{E}G_+, S(T)\tilde{E}\bar{\tilde{G}}_0] \begin{bmatrix} I & \theta \\ \theta & e_1(T^{\alpha+1} - 1) \end{bmatrix}.$$

From (C) and from (2.36a) we conclude

(2.45)
$$(A_3^1(T))^{-1} = \begin{bmatrix} I & \theta \\ \theta & e_1(1 - T^{\alpha+1}) \end{bmatrix} [S(T)\tilde{E}G_+, S(T)\tilde{E}\bar{\tilde{G}}_0]^{-1}.$$

Moreover

$$\begin{aligned} A_3^2(T)(A_3^1(T))^{-1} &= [\theta, S(T)\tilde{E}\bar{G}_0 e_2(T^{\alpha+1} - 1)e_1(1 - T^{\alpha+1})][S(T)\tilde{E}G_+, S(T)\tilde{E}\bar{\tilde{G}}_0]^{-1} \\ &= [\theta, -S(T)\tilde{E}\bar{G}_0 e_2(1 - T^{\alpha+1})][S(T)\tilde{E}G_+, S(T)\tilde{E}\bar{\tilde{G}}_0]^{-1} \end{aligned} \tag{2.46}$$

holds because of (2.36b). The proposition (B) and (2.38b) assure that

(2.47)
$$\|A_3^2(T)(A_3^1(T))^{-1}\| = o(1) \quad \text{for } T \to \infty.$$

Therefore $A_3(T)^{-1}$ exists and can be written as

$$\begin{aligned} (A_3(T))^{-1} &= (A_3^1(T))^{-1}\left(I + \sum_{i=1}^{\infty} (-1)^i (A_3^2(T)(A_3^1(T))^{-1})^i\right) \\ &= (A_3^1(T))^{-1}(I + o(1)) \quad \text{for } T \to \infty. \end{aligned} \tag{2.48}$$

Moreover

(2.49a)
$$\|A_4(T)\| = o(T^{-(\alpha+1)(r-1)}) \quad \text{for } T \to \infty,$$

(2.49b)
$$\|A_1(T)\| = O(1) \quad \text{for } T \to \infty$$

and therefore

(2.50)
$$\|A_4(T)A_2^{-1}A_1(T)(A_3(T))^{-1}\| = o(1) \quad \text{for } T \to \infty$$

because of (2.48), (2.49), (2.45). So

$$\begin{aligned} (A_3 - A_4 A_2^{-1} A_1)^{-1} &= (A_3(T))^{-1}\left(I + \sum_{i=1}^{\infty} (A_4(T)A_2^{-1}A_1(T)(A_3(T))^{-1})^i\right) \\ &= (A_3^1(T))^{-1}(I + o(1)) \quad \text{for } T \to \infty \end{aligned} \tag{2.51}$$

and (2.42) follows immediately. The linear equation (2.40) can now be solved:

$$\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}=\begin{bmatrix}-(A_3-A_4A_2^{-1}A_1)^{-1}A_4A_2^{-1} & (A_3-A_4A_2^{-1}A_1)^{-1}\\ A_2^{-1}+A_2^{-1}A_1(A_3-A_4A_2^{-1}A_1)^{-1}A_4A_2^{-1} & -A_2^{-1}A_1(A_3-A_4A_2^{-1}A_1)^{-1}\end{bmatrix}$$

$$\cdot\begin{pmatrix}\beta-B\tilde{E}\tilde{v}_p(1,T)\\ \gamma(T)-S(T)\tilde{E}\tilde{v}_p(T,T)\end{pmatrix}. \tag{2.52}$$

Inserting ξ_1 into (2.33) the term $A(t,T)(A_3(T)-A_4(T)A_2^{-1}A_1(T))^{-1}$ appears. From (2.33) and (2.42) we get

$$A(t,T)(A_3-A_4A_2^{-1}A_1)^{-1}$$

$$=\begin{bmatrix}\exp(J^+(t^{\alpha+1}-T^{\alpha+1})/(\alpha+1)) & \theta\\ \theta & e_1(t^{\alpha+1}-1)e_1(1-T^{\alpha+1})\\ \theta & e_2(t^{\alpha+1}-1)e_1(1-T^{\alpha+1})\\ \theta & \theta\end{bmatrix}O(1). \tag{2.53}$$

Using (2.37) we conclude that

$$\|A(\cdot,T)(A_3(T)-A_4(T)A_2^{-1}A_1(T))^{-1}\|_{[1,T]}=O(T^{(\alpha+1)(r-1)}) \tag{2.54}$$

holds.

From (2.52), (2.33) and (A) we derive:

$$\begin{aligned}\|\tilde{v}_T\|_{[1,T]}\leqq\text{const.}\,(\|\beta\|+T^{(\alpha+1)(r-1)}\|\gamma(T)\|+\|\tilde{v}_p(1,T)\|\\ +T^{(\alpha+1)(r-1)}\|\tilde{v}_p(T,T)\|+\|\tilde{v}_p(\cdot,T)\|_{[1,T]}).\end{aligned} \tag{2.55}$$

Now $\tilde{v}_p(\cdot,T)$ has to be defined. We set

$$\hat{f}(t,T)=\begin{cases}f(t), & 1\leqq t\leqq T,\\ f(T), & T\leqq t\leqq\infty\end{cases} \tag{2.56}$$

and

$$({}_TH_+f)(t):=\int_T^t\exp\left(\frac{\tilde{J}}{\alpha+1}(t^{\alpha+1}-s^{\alpha+1})\right)D_+\tilde{E}^{-1}s^\alpha f(s)\,ds, \tag{2.57}$$

$$({}_TH_0f)(t):=\int_T^t\exp\left(\frac{\tilde{J}}{\alpha+1}(t^{\alpha+1}-s^{\alpha+1})\right)D_0\tilde{E}^{-1}s^\alpha f(s)\,ds, \tag{2.58}$$

$$({}_TH_-f)(t)=(H_-E\tilde{E}^{-1}f)(t),\qquad \delta\leqq t\leqq T,\quad 1\leqq\delta<T, \tag{2.59}$$

so that we can define:

$$\tilde{v}_p(\cdot,T)=H_Tf={}_TH_+f+{}_TH_0f+{}_TH_-f. \tag{2.60}$$

From the estimates (2.19) and (2.21) we conclude that

$$\|{}_TH_+f+{}_TH_-f\|_{[\delta,T]}\leqq\text{const.}\,\|f\|_{[\delta,T]} \tag{2.61a}$$

because

$$\|{}_TH_+f\|_{[\delta,T]}\leqq\text{const.}\,\|H_+\|_{[\delta,\infty]}\|\hat{f}(\cdot,T)\|_{[\delta,\infty]} \tag{2.61b}$$

holds. Moreover (2.58) can be estimated as follows

$$\|({}_TH_0f)(t)\|\leqq\text{const.}\,\|f\|_{[t,T]}\int_t^T(t^{\alpha+1}-s^{\alpha+1})^{(r-1)}s^\alpha\,ds\leqq\text{const.}\,T^{(\alpha+1)r}\|f\|_{[t,T]}. \tag{2.62}$$

Altogether we get

$$\|H_T f\|_{[\delta,T]} \leqq c \cdot T^{(\alpha+1)r} \|f\|_{[\delta,T]}, \tag{2.63}$$

and

$$\|\tilde{v}_p(T,T)\| = \|({}_T H_- f)(T)\| \leqq \text{const.}\, \|f\|_{[1,T]} \tag{2.64}$$

holds because $({}_T H_0 f)(T) = 0$. From these estimates and from (2.55) the stability estimate (2.23) follows. □

In order to derive a convergence statement we write the problem (2.1), (2.2), (2.3) as follows

$$y' - t^\alpha A y = t^\alpha f(t), \qquad 1 \leqq t \leqq T, \quad y \in C([1,\infty]), \tag{2.65}$$

$$By(1) = \beta, \tag{2.66}$$

$$S(T)y(T) = S(T)y(T) \tag{2.67}$$

and subtract (2.4), (2.5), (2.6) from (2.65), (2.66), (2.67). We get

$$(y - x_T)' - t^\alpha A(y - x_T) = 0, \tag{2.68}$$

$$B(y - x_T)(1) = 0, \tag{2.69}$$

$$S(T)(y - x_T)(T) = S(T)y(T) - \gamma(T). \tag{2.70}$$

If $S(T)$ fulfills the assumptions (A), (B), (C), Theorem 2.1 can be applied giving:

$$\|y - x_T\|_{[1,T]} \leqq \text{const.}\, T^{(\alpha+1)(r-1)} \|S(T)y(T) - \gamma(T)\|. \tag{2.71}$$

Setting $\gamma(T) \equiv 0$ and using (2.9), (2.13) we get

$$\|y - x_T\|_{[1,T]} \leqq \text{const.}\, T^{(\alpha+1)(r-1)} \|S(T)E[\phi(T)\bar{G}_0, \phi(T)G_-]\eta + S(T)E(Hf)(T)\| \tag{2.72}$$

for some $\eta \in \mathbb{C}^{r_- + \tilde{r}_0}$. Assumption (B) guarantees that all columns of the fundamental matrix which are constant are dampened by an $o(T^{-(\alpha+1)(r-1)})$. All other appearing columns decay exponentially. Therefore the term which originates from the solution of the homogeneous problem converges as an $o(1)$ as $T \to \infty$. So assumption (B) is necessary for convergence for general β and f. Now let

$$\|f(t)\| = O(t^{-(\alpha+1)(2r-1)-\varepsilon}), \qquad \varepsilon > 0. \tag{2.73}$$

From the estimates (2.19), (2.20), (2.21) we conclude that

$$\|(Hf)(T)\| = O(T^{-(\alpha+1)(r-1)-\varepsilon}) \tag{2.74}$$

holds. Altogether we get

$$\|y - x_T\|_{[1,T]} \leqq \text{const.}\, (T^{(\alpha+1)(r-1)} (\|S(T)E\bar{G}_0\| + \|S(T)E\phi(T)G_-\|) + T^{-\varepsilon}). \tag{2.75}$$

If $f(t)$ contains an exponentially decreasing factor so that it has the asymptotic behavior

$$\|f(t)\| = O\left(t^\beta \exp\left(-\frac{\omega}{\alpha+1} t^{\alpha+1}\right)\right), \qquad \omega > 0 \tag{2.76}$$

then there is an operator $\bar{H}$ so that $y_p = E\bar{H}f$ is a particular solution of (2.65) and

$$\|(\bar{H}f)(T)\| = O\left(T^{\beta+(\alpha+1)r} (\ln T) \exp\left(-\frac{\omega}{\alpha+1} T^{\alpha+1}\right)\right) \tag{2.77}$$

holds. A proof for this can be found in Markowich (1980b). In this case $T^{-\varepsilon}$ in (2.75) has to be substituted by $T^{\beta+(\alpha+1)(2r-1)}(\ln T)\exp\left(-\frac{\omega}{\alpha+1}T^{\alpha+1}\right)$.

An optimal choice $S(T)=S_D(T)$ would be so that

$$[S_D(T)E\bar{G}_0, S_D(T)EG_-]=\theta \tag{2.78}$$

holds. Equation (2.78) is fulfilled for

$$S_D(T)E\equiv S_DE=\begin{bmatrix}(G_+)^T\\(\bar{\bar{G}}_0)^T\end{bmatrix}\Leftrightarrow E^TS_D^T=[G_+,\bar{\bar{G}}_0]. \tag{2.79}$$

These are linear equations for the rows of S_D which can be chosen independently of T. The asymptotic boundary condition (2.79), called a projection condition, fulfills (A), (B), (C) in Theorem 2.1 and is optimal in the sense that it makes the first two terms on the right-hand side of (2.75) vanish.

3. Linear variable coefficient problems: distinct eigenvalues. In this section we analyze the problem

$$y'-t^\alpha A(t)y=t^\alpha f(t),\quad \alpha\in N_0,\quad 1\leqq t<\infty, \tag{3.1}$$

$$y\in C([1,\infty]), \tag{3.2}$$

$$By(1)=\beta \tag{3.3}$$

and we require the $n\times n$ matrix $A(t)$ to fulfill

$$A\in C([1,\infty]),\qquad A(\infty)\neq 0, \tag{3.4}$$

$$A(t)=\sum_{i=0}^{\infty}A_it^{-i}\quad \text{for } t \text{ sufficiently large.} \tag{3.5}$$

Moreover let J_0 be the Jordan form of A_0 obtained by

$$A_0=EJ_0E^{-1}. \tag{3.6}$$

The following assumption is basic for this section:

$$J_0=\operatorname{diag}(\lambda_1,\cdots,\lambda_n),\qquad \lambda_i\neq\lambda_j\quad \text{for } i\neq j. \tag{3.7}$$

The substitution

$$u=E^{-1}y \tag{3.8}$$

gives the problem

$$u'-t^\alpha J(t)u=t^\alpha E^{-1}f(t),\qquad 1\leqq t<\infty, \tag{3.9}$$

$$u\in C([1,\infty]) \tag{3.10}$$

where

$$J(t)=E^{-1}A(t)E=\sum_{i=0}^{\infty}J_it^{-i},\qquad J_i=E^{-1}A_iE. \tag{3.11}$$

The fundamental matrix of the homogeneous problem (3.9) can be represented as an asymptotic series (see Wasow (1965) and Coddington and Levinson (1955)):

$$\phi(t)=P(t)t^De^{Q(t)} \tag{3.12}$$

where

$$P(t) \sim I + \sum_{i=1}^{\infty} P_i t^{-i}, \tag{3.13}$$

$$D = \operatorname{diag}(d_1, \cdots, d_n), \tag{3.14}$$

$$Q(t) = J_0 \frac{t^{\alpha+1}}{\alpha+1} + Q_1 \frac{t^{\alpha}}{\alpha} + \cdots + Q_\alpha t, \; Q_i \quad \text{are diagonal matrices} \tag{3.15}$$

hold. The unknown coefficients P_i, Q_i and D can be calculated by algebraic operations from the J_i's. An algorithm for that is given in Markowich (1980b), and therefore the asymptotic behavior of the basic solution can be determined knowing the Q_i's and D.

Let $\tilde{D}_0$ be the projection onto the direct sum of eigenspaces of J_0 associated with those eigenvalues with a real part zero which produce a basic solution which is in $C([1, \infty])$ and let $\tilde{G}_0$ be the projection-like matrix which is obtained by cancelling those columns of $\tilde{D}_0$ which have only zero entries. Let $\tilde{G}_0$ be a $n \times \tilde{r}_0$ matrix. Then the general solution of the problem (3.9), (3.10) is

$$u(t) = [\phi(t)\tilde{G}_0, \phi(t)G_-]\eta + (Hf)(t), \quad \eta \in \mathbb{C}^{\tilde{r}_0 + r_-} \tag{3.16}$$

(if $\tilde{r}_0 = 0$ then $[\phi(t)\tilde{G}_0, \phi(t)G_-]$ has to be substituted by $[\phi(t)G_-]$ in (3.16) and in the sequel) where G_- is defined as in § 2 and $u_p(t) = (Hf)(t)$ is an appropriate particular solution, which has been described by Markowich (1980b). The operator H operates on the space of all functions fulfilling

$$f \in C([\delta, \infty]), \quad \delta \geqq 1, \qquad \|f(t)\| = O(t^{-\alpha-1-\varepsilon}), \quad \varepsilon > 0. \tag{3.17}$$

Then the estimate

$$\|(Hf)(t)\| \leqq \text{const. } t^{-\varepsilon} \ln t \cdot \max_{s \geqq \delta} \|s^{\alpha+1+\varepsilon} f(s)\| \tag{3.18}$$

has been proven by Markowich (1980b). The particular solution on $[1, \infty]$ is obtained by continuation. The boundary value problems (3.1), (3.2), (3.3) is—under the given assumption on $A(t)$ and f—for all $\beta \in \mathbb{R}^{\tilde{r}_0 + r_-}$ uniquely soluble if and only if the $(\tilde{r}_0 + r_-) \times (\tilde{r}_0 + r_-)$ matrix

$$BE[\phi(1)\tilde{G}_0, \phi(1)G_-] \quad \text{is nonsingular.} \tag{3.19}$$

Of course, B has to be a $(\tilde{r}_0 + r_-) \times n$ matrix. We consider the approximating problems

$$x_T' - t^{\alpha} A(t) x_T = t^{\alpha} f(t), \qquad 1 \leqq t \leqq T, \quad \alpha \in N_0, \tag{3.20}$$

$$Bx_T(1) = \beta, \tag{3.21}$$

$$S(T)x_T(T) = \gamma(T). \tag{3.22}$$

$S(T)$ is a $(n - (\tilde{r}_0 + r_-)) \times n$ matrix and $\gamma(T) \in \mathbb{R}^{n-(\tilde{r}_0 + r_-)}$. For the following, G_+ is defined as in § 2 and $\tilde{\tilde{G}}_0$ is the $n \times (r_0 - \tilde{r}_0)$ matrix which is obtained by cancelling the zero columns of $D_0 - \tilde{D}_0$. Then the following stability theorem, which is analogous to Theorem 2.1, holds:

THEOREM 3.1. *Assume that* (3.19) *and* (A_1), (B_1), (C_1) *which are defined as follows, hold*:

(A_1) $$\|S(T)\| \leqq \text{const.} \quad \textit{as } T \to \infty,$$

(B_1) $$\|S(T)E\tilde{G}_0\| = o(1) \quad \textit{as } T \to \infty,$$

(C_1) $$\|[S(T)EG_+, S(T)E\tilde{\tilde{G}}_0]^{-1}\| \leqq \text{const.} \quad \textit{as } T \to \infty.$$

Then the problem (3.19), (3.21), (3.22) *has a unique solution* x_T *for sufficiently large* T. x_T *fulfills the estimate*

$$\|x_T\|_{[1,T]} \leqq \text{const.}\ (\|\beta\| + \|\gamma(T)\| + T^{\alpha+1} \ln T \|f\|_{[1,T]}) \tag{3.23}$$

for $f \in C([1, T])$, $\beta \in \mathbb{R}^{\tilde{r}_0 + r_-}$, $\gamma(T) \in \mathbb{R}^{n-(\tilde{r}_0+r_-)}$.

The substitution

$$v_T = E^{-1} x_T \tag{3.24}$$

gives the new problem

$$v_T' - t^\alpha J(t) v_T = t^\alpha E^{-1} f(t), \qquad 1 \leqq t \leqq T, \tag{3.25}$$

$$BE v_T(1) = \beta, \tag{3.26}$$

$$S(T) E v_T(T) = \gamma(T). \tag{3.27}$$

As the general solution of (3.25) we take for convenience

$$v_T(t) = \phi(t) e^{-Q(T)} T^{-D} [G_+, \tilde{\tilde{G}}_0] \xi_1 + \phi(t) [\tilde{G}_0, G_-] \xi_2 + v_p(t, T), \tag{3.28}$$

where $\xi_1 \in \mathbb{C}^{r_+ + (r_0 - \tilde{r}_0)}$, $\xi_2 \in \mathbb{C}^{\tilde{r}_0 + r_-}$ hold and $v_p(\cdot, T)$ is an appropriate particular solution which will be defined later. $\phi(t)$ is the fundamental matrix as of (3.12). Evaluation of the boundary conditions (3.26), (3.27) gives the linear block system

$$\begin{bmatrix} BE\phi(1) e^{-Q(T)} T^{-D} [G_+, \tilde{\tilde{G}}_0] & BE\phi(1)[\tilde{G}_0, G_-] \\ S(T) E P(T) [G_+, \tilde{\tilde{G}}_0] & S(T) E \phi(T) [\tilde{G}_0, G_-] \end{bmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \beta - BE v_p(1, T) \\ \gamma(T) - S(T) E v_p(T, T) \end{pmatrix}. \tag{3.29}$$

The matrix in the (1, 1) position is bounded because of the definition of G_+, $\tilde{\tilde{G}}_0$ and because of the diagonal form of $Q(T)$ and D. The matrix in the (2, 2) position is bounded too, because $\phi(t)[\tilde{G}_0, G_-]$ is the matrix whose columns are the basic solution of the homogeneous problem which are in $C([1, \infty])$ and because (A_1) holds. Moreover

$$\|S(T) E \phi(T) [\tilde{G}_0, G_-]\| = o(1) \quad \text{as } T \to \infty \tag{3.30}$$

because of (B_1) and (3.12). The matrix in the (1, 2) position is invertible because of (3.19) and its inverse is, as the matrix, independent of T. Finally

$$S(T) E P(T) [G_+, \tilde{\tilde{G}}_0] = S(T) E [G_+, \tilde{\tilde{G}}_0] + O(T^{-1}) \tag{3.31}$$

because of the asymptotic expansion for $P(t)$, (C_1) assures the bounded invertibility of the matrix in the (2, 1) position.

From (2.40) and (2.52) we conclude immediately that the system (3.29) has a unique solution $(\xi_1, \xi_2) \in \mathbb{C}^n$ and the estimate

$$\|v_T\|_{[1,T]} \leqq \text{const.}\ (\|\beta\| + \|\gamma(T)\| + \|v_p(\cdot, T)\|_{[1,T]}) \tag{3.32}$$

follows.

The particular solution $v_p(\cdot, T)$ has to be defined now. We set

$$v_p(\cdot, T) = H_T^{(1)} f = {}_T H_+^{(1)} f + {}_T H_0^{(1)} f + {}_T H_-^{(1)} f, \tag{3.33a}$$

$${}_T H_0^{(1)} f = \sum_{i=1}^{r_0} {}_T H_{0i}^{(1)} f, \tag{3.33b}$$

where

$$(3.34)\qquad ({}_TH_+^{(1)}f)(t)=\phi(t)\int_T^t D_+\phi^{-1}(s)s^{\alpha}E^{-1}f(s)\,ds,$$

$$(3.35)\qquad ({}_TH_-^{(1)}f)(t)=\phi(t)\int_\delta^t D_-\phi^{-1}(s)s^{\alpha}E^{-1}f(s)\,ds,\qquad \delta\geqq 1$$

and

$$(3.36)\qquad ({}_TH_{0i}^{(1)}f)(t)=\begin{cases}\phi(t)\displaystyle\int_T^t D_{0i}\phi^{-1}(s)s^{\alpha}E^{-1}f(s)\,ds, & \text{if (I) holds,}\\ \phi(t)\displaystyle\int_\delta^t D_{0i}\phi^{-1}(s)s^{\alpha}E^{-1}f(s)\,ds, & \text{if (II) holds.}\end{cases}$$

D_{0i} is the projection onto that eigenspace of J_0 which belongs to the ith eigenvalue with real part zero and (I), (II) are defined as

(I) $\operatorname{Re}(Q(t)D_{0i})\to+\infty$ or $(\operatorname{Re}(Q(t)D_{0i})\equiv 0$ and $\operatorname{Re}(DD_{0i})\geqq 0)$,

(II) $\operatorname{Re}(Q(t)D_{0i})\to-\infty$ or $(\operatorname{Re}(Q(t)D_{0i})\equiv 0$ and $\operatorname{Re}(DD_{0i})<0)$.

From the considerations in Markowich (1980b, § 3), we immediately conclude that

$$(3.37)\qquad \|H_T^{(1)}f\|_{[\delta,T]}\leqq T^{\alpha+1}\ln T\|f\|_{[\delta,T]}.$$

Therefore the estimate (3.23) follows and Theorem 3.1 is proven. □

As in § 2 the convergence estimate follows

$$(3.38)\qquad \|y-x_T\|_{[1,T]}\leqq \text{const.}\,\|S(T)y(T)-\gamma(T)\|$$

for all $(r_+ + (r_0-\tilde r_0))\times n$ matrices $S(T)$ which fulfill (A_1), (B_1) and (C_1).

Setting $\gamma(T)\equiv 0$ and inserting (3.16) we conclude

$$(3.39)\qquad \|y-x_T\|_{[1,T]}\leqq \text{const.}\,(\|S(T)E\phi(T)[\tilde G_0, G_-]\|+\|S(T)E(Hf)(T)\|).$$

The assumptions (A_1) and (B_1) guarantee convergence for all f fulfilling (3.17) because (3.18) holds. If

$$(3.40)\qquad S(T)E\tilde G_0\equiv 0,$$

convergence of the order $T^{-1}+T^{-\varepsilon}\ln T$ follows. In many practical cases all eigenvalues with real part zero produce exponentially decaying solutions and f also decays exponentially. The operator H can be changed to an operator $\bar H$, so that $(\bar Hf)(t)$ decays with the same exponential factor (see Markowich (1980b)). In this case exponential convergence follows from (3.39).

The optimal boundary condition is again the projection condition and it has to be calculated from the equation

$$(3.41)\qquad S_D(T)EP(T)\equiv\begin{bmatrix} G_+^T\\ (\tilde G_0)^T\end{bmatrix}$$

which is uniquely soluble because of the regularity of E and $P(T)$.

This immediately yields

$$(3.42)\qquad S_D(T)E\phi(T)[\tilde G_0, G_-]\equiv\theta$$

because of the form of $\phi(T)$ and because of (3.41). However, we do not know $P(T)$, but we can calculate the coefficients P_i of its expansion recursively (for the algorithm

see Markowich (1980b)). Having calculated $P_1, P_2, \cdots, P_k$ we set

$$\bar{P}(T) = I + \sum_{i=1}^{k} P_i T^{-i} \tag{3.43}$$

and solve

$$\bar{S}_D(T) E \bar{P}(T) = \begin{bmatrix} G_+^T \\ (\tilde{G}_0)^T \end{bmatrix} \tag{3.44}$$

instead of (3.41). Because

$$P(T) = \bar{P}(T) + O(T^{-k-1}) \quad \text{as } T \to \infty \tag{3.45}$$

holds, we get by a simple perturbation analysis

$$\bar{S}_D(T) = S_D(T) + O(T^{-k-1}) \quad \text{as } T \to \infty. \tag{3.46}$$

Therefore

$$\|\bar{S}_D(T) E \phi(T)[\tilde{G}_0, G_-]\| \leqq \text{const. } T^{-k-1} \|\phi(T)[\tilde{G}_0, G_-]\| \tag{3.47}$$

holds and the boundary condition $\bar{S}_D(T) x_T(T) = 0$ implies at least convergence of the order T^{-k-1} for $\|y - x_T\|_{[1,T]}$ if $f \equiv 0$ holds. More generally speaking, the order of convergence is determined by inserting (3.47) into (3.39).

However, this rather work-intensive procedure does only give significant improvement if some columns of $\phi(t)[\tilde{G}_0, G_-]$ do not converge exponentially. Only in this case the projection conditions imply a significant improvement of the order of convergence.

4. Linear problems—the general case. In this section we admit a general Jordan form of A_0. So we deal with problems of the form (3.1), (3.2), (3.3) with the assumption (3.4), (3.5), (3.6).

Again we perform the substitution (3.8) and get (3.9), (3.10).

The fundamental matrix $\phi(t)$ of the homogeneous problem (3.9) can now be represented as an asymptotic log-exponential power series of the following form

$$\phi(t) = P(t) t^D e^{Q(t)} \tag{4.1}$$

(see Wasow (1965)) where

$$P(t) \sim \sum_{i=0}^{\infty} P_i t^{-i/p} \quad \text{as } t \to \infty, \qquad p \in N_0, \tag{4.2}$$

$$D \text{ is a constant matrix in Jordan form,} \tag{4.3}$$

$$Q(t) = \text{diag}\,(J_0) \frac{t^{\alpha+1}}{\alpha+1} + Q_1 \frac{t^{\alpha+1-1/p}}{p(\alpha+1)-1} + \cdots + Q_{p(\alpha+1)-1} t^{1/p} \tag{4.4}$$

and the Q_i are diagonal matrices. diag (J_0) is the matrix which has the same diagonal entries as J_0 and all other entries zero. The matrices t^D and $e^{Q(t)}$ commute because the diagonal elements of $Q(t)$ which belong to a particular Jordan block of D are equal. Wasow (1965, Thm. 19.1) only says that D is a constant matrix, but from the proof of that theorem it is easily concluded that D can be taken in Jordan form and $Q(t)$ is still diagonal (see Markowich (1980b)). Moreover, $P(t)$ can be split up into:

$$P(t) = P_{(1)}(t) \cdot P_{(2)}(t) \cdot P_{(3)}(t), \tag{4.5}$$

where

$$P_{(1)}(t) \sim I + \sum_{i=1}^{\infty} P_{(1)i} t^{-i}, \tag{4.6}$$

$$P_{(k)}(t) \sim \sum_{i=0}^{\infty} P_{(k)i} t^{-i/p}, \qquad k = 2, 3. \tag{4.7}$$

$P_{(2)}(t)$, $P_{(3)}(t)$ are in block diagonal form, too. The ith diagonal block of $P_{(2)}(t)$ corresponds to that block in J_0 which is obtained by gathering all Jordan blocks belonging to the ith eigenvalue of J_0 and the jth diagonal block of $P_{(3)}(t)$ corresponds to the jth eigenvalue of $Q(t)$, where in both cases only different eigenvalues are counted. Markowich (1980b) has shown that

$$\|(P_{(2)}(t))^{-1} D_i\| \leqq \text{const. } t^{(\alpha+1)(r_i-1)} \tag{4.8}$$

holds, where D_i is the projection onto the direct sum of invariant subspaces associated with the ith eigenvalue of J_0 and r_i is the algebraic multiplicity of that ith eigenvalue. The statement (4.8) holds for the matrix $P_{(2)}$ derived as in Markowich (1980b).

The matrix $P_{(3)}(t)t^D$ is the fundamental matrix of the system

$$z' = \left(\frac{1}{x} B + \frac{1}{x^2} \tilde{B}(x)\right) z, \qquad \tilde{B} \in C([1, \infty]) \tag{4.9}$$

where

$$u = P_{(1)}(t) P_{(2)}(t)\, e^{Q(t)} z \tag{4.10}$$

has been set. The system (4.9) has a singularity of the first kind of $t = \infty$. Obviously $P_{(3)}(t)$ and D are not uniquely defined; only their product is unique (neglecting multiplication with a constant matrix from the right side). $\tilde{P}_{(3)}(t)t^{\tilde{D}} = (P_{(3)}(t)t^{-1})t^{D+I}$ would also be a way of splitting the product. The algorithm given by Wasow (1965) establishes a matrix $\tilde{P}_{(3)}(t)$ which has a convergent power series expansion, but $\tilde{P}_{(3)}(\infty)$ is not regular. We will show now that a representation can be given, so that $\tilde{P}_{(3)}(\infty)$ is regular. Therefore we assume that B is in Jordan-canonical form:

$$B = \text{diag}\,(B_1, \cdots, B_q) \tag{4.11}$$

and B_i has the only eigenvalue b_i, where $\text{Re}\,(b_i) \leqq 0$ for $1 \leqq i \leqq s$ and $\text{Re}\,(b_j) > 0$ for $s+1 \leqq j \leqq n$. We write (4.9) as

$$z' = \frac{1}{x} Bz + \frac{1}{x}(\bar{B}(x)z), \qquad \bar{B}(x) = \frac{1}{x}\tilde{B}(x) \tag{4.12}$$

and set for $1 \leqq i \leqq s$

$$z_i(t) = \begin{bmatrix} \theta \\ \cdot \\ \theta \\ t^{B_i} \\ \theta \\ \cdot \\ \theta \end{bmatrix} + (G\bar{B}z_i)(t), \qquad \delta \leqq t < \infty, \tag{4.13}$$

where G is the operator defined in (4.46), (4.47) in Section 4 of Markowich (1980b)

which applied to a function g defines an appropriate particular solution of the problem

$$z' = \frac{1}{x}Bz + \frac{1}{x}g, \qquad \delta \leqq x < \infty, \tag{4.14}$$

where $\delta \geqq 1$. Markowich (1980b) showed that

$$\|(Gg)(t)\| \leqq \text{const. } (\ln t)^{j+k_i-\varepsilon}\|\bar{g}\|_{[\delta,\infty]}, \qquad 0 \leqq j \leqq \max(\dim(B_i)) \tag{4.15}$$

holds if

$$g(t) = t^{-\varepsilon}(\ln t)^k \bar{g}(t), \qquad g, \bar{g} \in C_b([\delta, \infty]), \qquad \varepsilon > 0, \tag{4.16}$$

where $C_b([\delta, \infty))$ is the space of all functions $f \in C([\delta, \infty))$ which are bounded as $t \to \infty$. Set

$$\sigma_i(t) = \|t^{B_i}\| = c \cdot t^{\mathrm{Re}(b_i)}(\ln t)^{\dim(B_i)-1}, \qquad 1 \leqq i \leqq S. \tag{4.17}$$

We want to show that (4.13) establishes a fixed-point equation for $z_i \in A_{\sigma_i,\delta}$, where

$$A_{\sigma_i,\delta} = \{u \mid u(t) = U(t)\sigma_i(t),\ U \in C_b([\delta, \infty))\}, \qquad \|u\|_i = \|U\|_{[\delta,\infty]}. \tag{4.18}$$

We show that the operator

$$(\psi_i(z_i))(t) = \begin{bmatrix} \theta \\ \cdot \\ \theta \\ t^{B_i} \\ \theta \\ \cdot \\ \theta \end{bmatrix} + (G\bar{B}z_i)(t) \tag{4.19}$$

is a contraction on $A_{\sigma_i,\delta}$ for δ sufficiently large. From (4.13), (4.15), (4.17) we conclude that ψ_i maps $A_{\sigma_i,\delta}$ into $A_{\sigma_i,\delta}$. Moreover,

$$\|\psi_i(z_{i1}) - \psi_i(z_{i2})\|_i = \|G\bar{B}(z_{i1} - z_{i2})\|_i \leqq \text{const. } \delta^{-1}(\ln \delta)^j \|z_{i1} - z_{i1}\|_i \tag{4.20}$$

holds where $0 \leqq j \leqq \max(\dim(B_i))$, and therefore ψ_i is a contraction on $A_{\sigma_i,\delta}$ for δ sufficiently large. From (4.13) we conclude

$$z_i(t) = (I + O(t^{-1}(\ln t)^{2j+\dim(B_i)})) \begin{bmatrix} \theta \\ \cdot \\ \theta \\ t^{B_i} \\ \theta \\ \cdot \\ \theta \end{bmatrix} \quad \text{as } t \to \infty. \tag{4.21}$$

Now let $s+1 \leqq i \leqq n$ hold, so that $\mathrm{Re}(b_i) > 0$. We substitute

$$z_i = \tilde{z}_i \cdot t^{b_i+1} \tag{4.22}$$

and (4.12) becomes

$$\tilde{z}_i' = \frac{1}{x}(B - (b_i+1)I)\tilde{z}_i + \frac{1}{x}\bar{B}(x)\tilde{z}_i. \tag{4.23}$$

Now we set

$$\tilde{z}_i(t) = \begin{bmatrix} \theta \\ \cdot \\ \theta \\ t^{B_i-(b_i+1)I} \\ \theta \\ \cdot \\ \theta \end{bmatrix} + (G\tilde{B}\tilde{z}_i)(t) \tag{4.24}$$

so that

$$\sigma_i = \|t^{B_i-(b_i+1)I}\| = c \cdot t^{-1}(\ln t)^{\dim(B_i)-1} \tag{4.25}$$

and (4.13), (4.21) implies that

$$\tilde{z}_i(t) = (I + O(t^{-1}(\ln t)^{2_i+\dim(B_i)})) \begin{bmatrix} \theta \\ \cdot \\ \theta \\ t^{B_i-(b_i+1)I} \\ \theta \\ \cdot \\ \theta \end{bmatrix} \tag{4.26}$$

holds and from (4.22) we conclude

$$z_i(t) = (I + O(t^{-1}(\ln t)^{2_i+\dim(B_i)})) \begin{bmatrix} \theta \\ \cdot \\ \theta \\ t^{B_i} \\ \theta \\ \cdot \\ \theta \end{bmatrix} \quad \text{as } t \to \infty. \tag{4.27}$$

Obviously the matrix

$$z(t) = [z_1(t), \cdots, z_n(t)] = [I + O(t^{-1}(\ln t)^m)]t^B, \qquad m \in N_0 \tag{4.28}$$

is a fundamental matrix of the system (4.9). Therefore $P_{(3)}(t)$ and D in (4.5) and (4.1) can be chosen so that

$$\|P_{(3)}(t)^{-1}\| = \|I + o(1)\| \leqq \text{const.} \tag{4.29}$$

holds.

Knowing the fundamental matrix asymptotically we can sort out the basic solution φ_i fulfilling $\varphi_i \in C([1, \infty])$ so that the general solution of (3.9), (3.10) is

$$u(t) = \phi(t)[\tilde{G}_0, G_-]\eta + (Hf)(t), \qquad \eta \in \mathbb{C}^{\tilde{r}_0+r_-} \tag{4.30}$$

where H defines an appropriate particular solution Hf on $[\delta, \infty]$ if

$$f(t) = t^{-(\alpha+1)\bar{r}-\varepsilon}(\ln t)^l F(t), \qquad F \in C_b([\delta, \infty)) \tag{4.31}$$

where $\bar{r}$ is the maximal algebraic multiplicity of eigenvalues of J_0 which have real part zero. We assume that $\bar{r} > 0$ because $\bar{r} = 0$ is covered by the theory given in de Hoog and Weiss (1980a, b). Moreover

$$\|(Hf)(t)\| \leqq \text{const. } t^{-\varepsilon}(\ln t)^{j+l}\|F\|_{[\delta,\infty]}, \qquad 0 \leqq j \leqq n. \tag{4.32}$$

The particular solution on $[1, \infty]$ is obtained by continuation.

Again, the boundary value problem on the infinite interval is uniquely soluble for all $\beta \in \mathbb{R}^{\tilde{r}_0+r_-}$ and f's which fulfill (4.31) if and only if the $(\tilde{r}_0+r_-)\times(\tilde{r}_0+r_-)$ matrix

$$BE[\phi(1)\tilde{G}_0, \phi(1)G_-] \quad \text{is nonsingular.} \tag{4.33}$$

Of course, B is an $(\tilde{r}_0+r_-)\times n$ matrix.

The approximating problems have the form (3.20), (3.21), (3.22). We will again prove a stability theorem.

THEOREM 4.1. *Assume that* (4.33) *and* (A_2), (B_2), (C_2) *which are defined as follows, hold*:

$$\|S(T)\| \leqq \text{const.} \quad \text{as } T \to \infty, \tag{A_2}$$

$$\|S(T)EP(T)\tilde{G}_0\| = o(T^{-(\alpha+1)(\bar{r}-1)}), \tag{B_2}$$

$$\|(S(T)EP_{(1)}(T)[G_+, P_{(2)}(T)P_{(3)}(T)\tilde{\tilde{G}}_0])^{-1}\| = O(T^{(\alpha+1)(\bar{r}-1)}). \tag{C_2}$$

Then there is a unique solution x_T of the problem (3.20), (3.21), (3.22) *for T sufficiently large and the following estimate holds for all $\beta \in \mathbb{R}^{\tilde{r}_0+r_-}$, $\gamma(T) \in \mathbb{R}^{(r_0-\tilde{r}_0)+r_+}$, $f \in C([1, T])$:*

$$\|x_T\|_{[1,T]} \leqq \text{const.}\,(\|\beta\| + T^{(\alpha+1)(\bar{r}-1)}\|\gamma(T)\| + T^{(\alpha+1)\bar{r}}(\ln T)^j \|f\|_{[1,T]}), \qquad 0 \leqq j \leqq n. \tag{4.34}$$

Proof. We substitute

$$x_T(t) = EP_{(1)}(t)w_T(t), \qquad w_T = \begin{pmatrix} w_T^+ \\ w_T^0 \\ w_T^- \end{pmatrix} \tag{4.35}$$

and get three separate problems

$$\begin{pmatrix} w_T^+(t) \\ w_T^0(t) \\ w_T^-(t) \end{pmatrix}' = t^\alpha \begin{bmatrix} J^+(t) & & \\ & J^0(t) & \\ & & J^-(t) \end{bmatrix} \begin{pmatrix} w_T^+(t) \\ w_T^0(t) \\ w_T^-(t) \end{pmatrix} + \underbrace{t^\alpha P_{(1)}^{-1}(t)E^{-1}f(t)}_{\tilde{f}(t)} \tag{4.36}$$

where

$$\begin{bmatrix} J^+(\infty) & & \\ & J^0(\infty) & \\ & & J^-(\infty) \end{bmatrix} = \begin{bmatrix} J_0^+ & & \\ & J_0^0 & \\ & & J_0^- \end{bmatrix} = J_0 \tag{4.37}$$

and the eigenvalues of J_0^+ have a positive real part, the eigenvalues of J_0^0 have a zero real part and the eigenvalues of J_0^- have a negative real part. This structure can always be obtained by reordering the columns of E. Now we rewrite the equation for w_T^+:

$$w_T^+(t)' = t^\alpha J_0^+ w_T^+(t) + (J^+(t) - J_0^+)w_T^+(t) + t^\alpha \tilde{f}_+(t). \tag{4.38}$$

We define the general solution of (4.38) as

$$w_T^+(t) = \exp\left(\frac{J_0^+}{\alpha+1}(t^{\alpha+1} - T^{\alpha+1})\right)\xi^+ + ({}_TH_+(J^+ - J_0^+)w_T^+)(t) + ({}_TH_+\tilde{f}_+)(t), \tag{4.39}$$

where ${}_TH_+$ is defined in (2.57) with $E = I$ and $J^0 = J_0^+$. We derive

$$((I - {}_TH_+(J^+ - J_0^+))w_T^+)(t) = \exp\left(\frac{J_0^+}{\alpha+1}(t^{\alpha+1} - T^{\alpha+1})\right)\xi^+ + ({}_TH_+\tilde{f}_+)(t). \tag{4.40}$$

De Hoog and Weiss (1980b) have shown that $(I - {}_T H_+(J^+ - J_0^+))$ is invertible as operator on $C([\delta, T])$ with δ and T sufficiently large, so that

$$w_T^+(t) = \psi_+(t, T)\xi^+ + \psi_+(\tilde{f}_+)(t) \in C([\delta, T]), \tag{4.41}$$

where

$$\psi_+(\cdot, T) = (I - {}_T H_+(J^+ - J_0^+))^{-1} \exp\left(\frac{J_0^+}{\alpha+1}(h - T^{\alpha+1})\right), \qquad h(t) = t^{\alpha+1} \tag{4.42}$$

and

$${}_T\psi_+(\tilde{f}_+) = (I - {}_T H_+(J^+ - J_0^+))^{-1} {}_T H_+ \tilde{f}_+. \tag{4.43}$$

Moreover, they have shown that

$$\|\psi_+(\cdot, T)\|_{[\delta,T]} \leqq \text{const.}, \tag{4.44a}$$

$$\|{}_T\psi_+(\tilde{f}_+)\|_{[\delta,T]} \leqq \text{const.}\, \|f\|_{[\delta,T]} \tag{4.44b}$$

and from (2.57) and (4.42) we derive that

$$\psi_+(T, T) = I. \tag{4.45}$$

Now we define the general solution of (4.36) as

$$\begin{aligned} w_T(t) = &\left[\begin{bmatrix} \psi_+(t, T) & \theta \\ \theta & \theta \end{bmatrix} G_+, P_{(2)}(t)P_{(3)}(t)e^{Q(t)-Q(T)}\left(\frac{t}{T}\right)^D \tilde{\tilde{G}}_0\right]\xi_1 \\ &+ P_{(2)}(t)P_{(3)}(t)e^{Q(t)}t^D[\tilde{G}_0, G_-]\xi_2 + \begin{pmatrix} \psi_+(\tilde{f}_+)(t) \\ w_p^0(t, T) \\ w_p^-(t, T) \end{pmatrix}, \qquad \xi_1 \in \mathbb{C}^{r_+ + (r_0 - \tilde{r}_0)}, \quad \xi_2 \in \mathbb{C}^{\tilde{r}_0 + r_-} \end{aligned} \tag{4.46}$$

where w_p^0, w_p^- are appropriate particular solutions. This solution is defined on $[\delta, T]$ and the corresponding solution on $[1, T]$ is obtained by continuing $\psi_+(\cdot, T)$. Resubstituting in (4.35) and evaluating at the boundaries sets up the linear block system for ξ_1, ξ_2:

$$\begin{bmatrix} BEP_{(1)}(1)\left[\begin{bmatrix} \psi_+(1, T) & \theta \\ \theta & \theta \end{bmatrix} G_+, P_{(2)}(1)P_{(3)}(1)e^{Q(1)-Q(T)}{}_T{}^{-D}\tilde{\tilde{G}}_0\right] & BE\phi(1)[\tilde{G}_0, G_-] \\ S(T)EP_1(T)[G_+, P_{(2)}(T)P_{(3)}(T)\tilde{\tilde{G}}_0] & S(T)E\phi(T)[\tilde{G}_0, G_-] \end{bmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$$

$$= \begin{bmatrix} \beta - BEP_{(1)}(1)\begin{pmatrix} {}_T\psi_+(\tilde{f}_+)(1) \\ w_p^0(1, T) \\ w_p^-(1, T) \end{pmatrix} \\ \gamma(T) - S(T)EP_{(1)}(T)\begin{pmatrix} {}_T\psi_+(\tilde{f}_+)(T) \\ w_p^0(T, T) \\ w_p^-(T, T) \end{pmatrix} \end{bmatrix} \tag{4.47}$$

The matrix in the $(1, 1)$ position is bounded, $BE\phi(1)[\tilde{G}_0, G_-]$ is independent of T and invertible because of (4.33), the matrix in the $(2, 1)$ position fulfills (C_2) and

$$\|S(T)E\phi(T)[\tilde{G}_0, G_-]\| \leqq o(1)T^{-(\alpha+1)(\tilde{r}-1)} \tag{4.48}$$

holds because of (B_2). From (2.40), (2.52) we conclude that

$$\begin{aligned}\|x_T\|_{[1,T]} \leqq \text{const.}\ (&\|\beta\| + T^{(\alpha+1)(\bar{r}-1)}\|\gamma(T)\| + \|v_p(1,T)\| \\ &+ T^{(\alpha+1)(\bar{r}-1)}\|S(T)Ev_p(T,T)\| + \|v_p(\cdot,T)\|_{[1,T]})\end{aligned} \tag{4.49}$$

where

$$v_p(t,T) := P_{(1)}(t)\begin{pmatrix} {}_T\psi_+(\tilde{f}_+)(t) \\ w_p^0(t,T) \\ w_p^-(t,T)\end{pmatrix} = (H_T^{(2)}f)(t) \tag{4.50}$$

has been used. Splitting $H_T^{(2)}$ into ${}_TH_+^{(2)}$, ${}_TH_0^{(2)}$, ${}_TH_-^{(2)}$ where ${}_TH_+^{(2)}$ is already defined by ${}_T\psi_+(\tilde{f}_+)$, we can define ${}_TH_0^{(2)}$ as we have defined ${}_TH_0^{(1)}$ in (3.36), only the D_{0i} have to be substituted by the projections onto the invariant subspaces of D. The estimate

$$\|{}_TH_0^{(2)}f\|_{[1,T]} \leqq \text{const.}\ T^{(\alpha+1)\bar{r}}(\ln T)^i\|f\|_{[1,T]} \tag{4.51}$$

results as in § 3.

${}_TH_-^{(2)}f$ can be constructed by the same perturbation approach we used for the construction of ${}_TH_+^{(2)}f$. We set

$${}_T\psi_-(\tilde{f}_-) = (I - {}_TH_-(J^- - J_0^-))^{-1}{}_TH_-\tilde{f}_- = w_p^-(\cdot,T) \tag{4.52}$$

where ${}_TH_-$ is defined in (2.59) with $E = I$ and $J = J_0^-$ and then

$$({}_TH_-^{(2)}f)(t) = P_{(1)}(t)\begin{pmatrix} \theta \\ \theta \\ {}_T\psi_-(\tilde{f}_-)(t)\end{pmatrix} \tag{4.53}$$

holds. Moreover the estimate

$$\|{}_TH_-^{(2)}f\|_{[1,T]} \leqq \text{const.}\ \|f\|_{[1,T]} \tag{4.54}$$

is fulfilled. Because of (B_2) and (3.36) we get

$$\|S(T)E({}_TH_0^{(2)}f)(T)\| \leqq \text{const.}\ T^{\alpha+1}(\ln T)^i\|f\|_{[1,T]} \tag{4.55}$$

so that the estimate (4.34) follows. □

Again the convergence estimate follows:

$$\|y - x_T\|_{[1,T]} \leqq \text{const.}\ T^{(\alpha+1)(\bar{r}-1)}\|S(T)y(T) - \gamma(T)\| \tag{4.56}$$

for all matrices $S(T)$ fulfilling (A_2), (B_2), (C_2). Setting $\gamma(T) \equiv 0$ and using (4.30) we get the order of convergence as follows:

$$\|y - x_T\|_{[1,T]} \leqq \text{const.}\ T^{(\alpha+1)(\bar{r}-1)}(\|S(T)E\phi(T)[\tilde{G}_0, G_-]\| + \|(Hf)(T)\|). \tag{4.57}$$

Assumption (C_2) guarantees convergence for all f's which fulfill

$$\|f(t)\| = O(t^{-(\alpha+1)(2\bar{r}-1)-\varepsilon}), \qquad \varepsilon > 0 \tag{4.58}$$

because the columns in $\phi(T)$ which may be constant as $T \to \infty$ are dampened by the factor $o(T^{-(\alpha+1)(\bar{r}-1)})$. Again if all columns of $\phi(t)$ and f decay exponentially, the convergence is exponential, too.

Still the question has to be answered whether there is a matrix $S(T)$ fulfilling the assumption of Theorem 4.1 and how it can be constructed. We set

$$S(T)EP_{(1)}(T) = \tilde{S}(T) \tag{4.59}$$

and choose $\tilde{S}(T)$ so that:

$$\tilde{S}(T)[G_+, \theta, \theta] = \begin{bmatrix} (G_+)^T \\ \theta \end{bmatrix}, \tag{4.60}$$

$$\tilde{S}(T)[\theta, \theta, G_-] = \theta \tag{4.61}$$

and

$$\tilde{S}(T)P_{(2)}(T)P_{(3)}(T)[\theta, G_0, \theta] = T^{-(\alpha+1)(\bar{r}-1)} \begin{bmatrix} \theta \\ (\tilde{\tilde{G}}_0)^T \end{bmatrix}. \tag{4.62}$$

Because of the block structure of $P_{(2)}$, $P_{(3)}$ (4.62) is equivalent to

$$\tilde{S}(T)[\theta, G_0, \theta] = T^{-(\alpha+1)(\bar{r}-1)} \begin{bmatrix} \theta \\ (\tilde{\tilde{G}}_0)^T \end{bmatrix} P_{(3)}^{-1}(T)P_{(2)}^{-1}(T). \tag{4.63}$$

The equations (4.60), (4.61), (4.63) determine $\tilde{S}(T)$ and $S(T)$ can be calculated from

$$S(T) = \tilde{S}(T)P_{(1)}^{-1}(T)E^{-1}. \tag{4.64}$$

$S(T)$ fulfills (B_2) because of (4.62), the proposition (A_2) follows from (4.8) and (4.29), and (C_2) is implied by (4.60), (4.62). This asymptotic boundary condition (with $\gamma(T) \equiv 0$) is the projection condition fulfilling

$$S_D(T)E\phi(T)[\tilde{G}_0, G_-] = \theta. \tag{4.65}$$

In general, we only can determine a finite number of coefficients of the expansion of $P_{(1)}$, $P_{(2)}$, $P_{(3)}$. An algorithm is given in Markowich (1980b), and it is shown that

$$P_{(2)}(t) = \prod_{i=1}^{m} S_i(t)E_iP_{(2)}^i(t) \tag{4.66}$$

where the matrices S_i are in block diagonal form and their diagonal blocks are

$$S_{ij}(t) = \operatorname{diag}(1, t^{-g_{ij}}, \cdots, t^{-(k_{ij}-1)g_{ij}}). \tag{4.67}$$

The E_i's are regular and

$$P_{(2)}^i(t) \sim I + \sum_{j=1}^{\infty} P_{(2)}^{ij} t^{-j/p} \quad \text{as } t \to \infty \tag{4.68}$$

holds.

We denote by

$$P_{(2)0}(t) = \prod_{i=1}^{m} S_{i0}(t)E_iP_{(2)0}^i(t), \qquad P_{(2)0}^i(t) \sim I + \sum_{j=1}^{\infty} P_{(2)0}^{ij} t^{-j/p} \tag{4.69}$$

that diagonal block of $P_{(2)}$ which is associated with the zero real part eigenvalues of J_0. Assume now that we know $P_{(2)0}^{i1}, \cdots, P_{(2)0}^{ik_2}$ for $i = 1(1)m$ and $k_2 > 2p(\alpha+1)(\bar{r}-1)-1$. Then we set

$$\bar{P}_{(2)0}(t) = \prod_{i=1}^{m} S_{i0}(t)E_i\bar{P}_{(2)0}^i(t), \qquad \bar{P}_{(2)0}^i = I + \sum_{j=1}^{k_2} P_{(2)0}^{ij} t^{-j/p} \tag{4.70}$$

and assume that we know

$$\bar{P}_{(3)0}(t) = I + \sum_{j=1}^{k_3} P_{(3)0}^j t^{-j/p}, \qquad k_3 > p(\alpha+1)(\bar{r}-1)-1. \tag{4.71}$$

$\bar{P}_{(3)0}(t)$ consists of the first k_3+1 summands of $P_{(3)0}(t)$ which is that block of $P_{(3)}(t)$ associated with the real part zero eigenvalues of J_0. Moreover we have to know

$$\bar{P}_{(1)}(t) = I + \sum_{j=1}^{k_1} P_{(1)j} t^{-j}, \qquad k_1 > (\alpha+1)(\bar{r}-1)-1. \tag{4.72}$$

Using $\bar{P}_{(1)}$, $\bar{P}_{(2)0}$, $\bar{P}_{(3)0}$ instead of $P_{(1)}$, $P_{(2)0}$, $P_{(3)0}$ we calculate a matrix $\bar{S}(T)$ instead of $S(T)$; using (4.60), (4.61), (4.63), (4.64) and a perturbation analysis shows that

$$\bar{S}(T) = S(T) + O(T^{-k-1}) \tag{4.73}$$

where

$$k = \min\left(k_1, \frac{k_2+1}{p} - (\alpha+1)(\bar{r}-1) - 1, \frac{k_3+1}{p} - 1\right) \geqq (\alpha+1)(\bar{r}-1). \tag{4.74}$$

Therefore

$$\bar{S}(T)\phi(T)[\tilde{G}_0, G_-] = O(T^{-k-1})\phi(T)[\tilde{G}_0, G_-] \tag{4.75}$$

and the order of convergence for homogeneous problems is at least $T^{(\alpha+1)(r-1)-k-1}$.

The requirement that $A(t)$ is analytical in $t=\infty$ is very restricting; therefore, we will now admit matrices $A(t)$ fulfilling

$$A \in C([1,\infty]), \quad A\left(\frac{1}{\cdot}\right) \in C^{(\alpha+1)\bar{l}+1}\left(\left[0, \frac{1}{\delta}\right]\right), \quad \delta \geqq 1, \tag{4.76}$$

where $\bar{l}$ is the maximal algebraic multiplicity of an eigenvalue of $A(\infty)$ with nonpositive real part. Therefore A can be expanded:

$$A(t) = A_0 + t^{-1}A_1 + \cdots + t^{-(\alpha+1)\bar{l}} A_{(\alpha+1)\bar{l}} + \tilde{A}(t) \tag{4.77}$$

where

$$\tilde{A}(t) = \bar{A}(t) t^{-(\alpha+1)\bar{l}-1-\varepsilon}, \qquad \varepsilon > 0, \quad \bar{A} \in C_b([1,\infty)). \tag{4.78}$$

The problem (3.1), (3.2) can now be rewritten as

$$y' - t^{\alpha} \sum_{i=0}^{(\alpha+1)\bar{l}} A_i t^{-i} y = t^{\alpha}\tilde{A}(t)y + t^{\alpha} f(t), \tag{4.79}$$

$$y \in C([1,\infty]), \tag{4.80}$$

and can be regarded as a perturbed system of

$$\tilde{y}' - t^{\alpha} \sum_{i=0}^{(\alpha+1)\bar{l}} A_i t^{-i} \tilde{y} = t^{\alpha} f(t), \tag{4.81}$$

$$\tilde{y} \in C([1,\infty]). \tag{4.82}$$

Markowich (1980b) has proven that the $n \times (\tilde{r}_0 + r_-)$ solution matrix $E\psi_-^0$ of (4.79), (4.80) for $f \geqq 0$ fulfills for large t:

$$\|\psi_-^0(t) - \phi(t)[\tilde{G}_0, G_-]\| \leqq \text{const. } t^{-1-\varepsilon}(\ln t)^{2j} \|\phi(t)[\tilde{G}_0, G_-]\|, \qquad \varepsilon > 0 \tag{4.83}$$

where $0 \leqq j \leqq n$ holds and $\phi(t)[\tilde{G}_0, G_-]$ is the general solution of (4.81), (4.82). Moreover, a particular solution $E\psi(f)$ of (4.79), (4.80) can be constructed if f fulfills (4.31) and

$$\|\psi(f)(t) - (Hf)(t)\| \leqq \text{const. } t^{-1-\varepsilon}(\ln t)^{2j} \|(Hf)(t)\| \tag{4.84}$$

holds. The problem (3.1), (3.2), (3.3) is for all $\beta \in \mathbb{R}^{\tilde{r}_0+r_-}$ and f fulfilling (4.31) uniquely soluble if and only if the $(\tilde{r}_0+r_-)\times(\tilde{r}_0+r_-)$-matrix

$$BE\psi_-^0(1) \quad \text{is nonsingular.} \tag{4.85}$$

Of course, B is a $(\tilde{r}_0+r_-)\times n$ matrix.

For the existence theory and for the following stability theorem it is sufficient to require that (4.77) holds with $\bar{l}=\bar{r}$ and with $\tilde{A}(t)=t^{-(\alpha+1)r-\varepsilon}\bar{A}(t)$ for some $\varepsilon>0$. This implies that the right-hand sides of (4.83) and (4.84) evaluated as $t=T$ equal const. $(\ln T)^j T^{-\varepsilon}$. Note that $\bar{r}$, which is the largest algebraic multiplicity of an eigenvalue of A_0 with real part zero, is always larger than or equal to r (defined in § 2), which is the dimension of the largest Jordan block which has an eigenvalue with real part zero.

We consider the asymptotic boundary value problem

$$x_T' - t^\alpha A(t)x_T = t^\alpha f(t), \qquad 1 \leqq t \leqq T, \tag{4.86}$$

$$Bx_T(1)=\beta, \tag{4.87}$$

$$S(T)x_T(T)=\gamma(T) \tag{4.88}$$

and show that the construction of $S(T)$ and the stability estimate (4.34) depend only on the validity of (A_2), (B_2), (C_2) for the perturbed problem

$$\tilde{x}_T' - t^\alpha\Big(\sum_{i=0}^{(\alpha+1)\bar{r}} A_i t^{-1}\Big)\tilde{x}_T = t^\alpha f(t), \tag{4.89}$$

$$S(T)\tilde{x}_T(T)=\gamma(T) \tag{4.90}$$

if (4.85) holds.

THEOREM 4.2. *Let A fulfill* (4.76) *and let* (4.85), (A_2), (B_2), (C_2) *of Theorem* 4.1 *hold where*

$$\phi(t)=P_{(1)}(t)P_{(2)}(t)P_{(3)}(t)t^R e^{Q(t)}$$

is the fundamental matrix of the homogeneous problem (4.81). *Then there is a unique solution x_T of the problem* (4.86), (4.87), (4.88) *for T sufficiently large and for all $\beta \in \mathbb{R}^{\tilde{r}_0+r_-}$, $\gamma(T)\in\mathbb{R}^{r_++(r_0-\tilde{r}_0)}$, $f\in C([1,T])$. This solution x_T fulfills the estimate* (4.34) *with j substituted by $2j$.*

Proof. We write (4.87)

$$x_T' - t^\alpha\Big(\sum_{i=0}^{(\alpha+1)\bar{r}} A_i t^{-i}\Big)x_T = t^\alpha(\tilde{A}(t)x_T + f(t)), \qquad 1 \leqq t \leqq T \tag{4.91}$$

and write the general solution after having set $x_T = Ev_T$ as

$$\begin{aligned} v_T(t) = {} & P_{(1)}(t)\left[\begin{bmatrix}\psi_+(t,T) & \theta\\ \theta & \theta\end{bmatrix}G_+, P_{(2)}(t)P_{(3)}(t)e^{Q(t)-Q(T)}\left(\frac{t}{T}\right)^D \tilde{\tilde{G}}_0\right]\xi_1 \\ & + P(t)e^{Q(t)}t^D[\tilde{G}_0, G_-]\xi_2 + (H_T^{(2)}\tilde{A}Ev_T)(t) + (H_T^{(2)}f)(t). \end{aligned} \tag{4.92}$$

We restrict t to the interval $[\delta, T]$ where δ is sufficiently large, so that

$$H_T^{(2)}\tilde{A}: C([\delta,T]\to C([\delta,T]) \tag{4.93}$$

holds. Then the following estimate is fulfilled:

$$\|H_T^{(2)}\tilde{A}\|_{[\delta,T]} = \max_{\|u\|_{[\delta,T]}\leqq 1} \|H_T^{(2)}\tilde{A}u\|_{(\delta,T]}$$

$$\leqq \max_{\|u\|_{[\delta,T]}\leqq 1} (\|{}_TH_+^{(2)}\tilde{A}u\|_{[\delta,T]} + \|{}_TH_0^{(2)}\tilde{A}u\|_{[\delta,T]} + \|{}_TH_-^{(2)}\tilde{A}u\|_{[\delta,T]}) \tag{4.94}$$

$$\leqq \text{const.}\,(\|\tilde{A}\|_{[\delta,T]} + \delta^{-\varepsilon}(\ln\delta)^j\|a\|_{[\delta,T]}) \leqq \frac{1}{2\|E\|}$$

for all $T > \delta \geqq \bar{\delta}$ sufficiently large. This follows from the estimates for ${}_TH_+^{(1)}$, ${}_TH_-^{(1)}$ given in § 2 and from the estimates from H_0 given in Markowich (1980b). Therefore, $(I - H_T\tilde{A}E)$ is invertible on $C([\delta, T])$ and we get:

$$v_T(t) = {}_T\psi_+^0(t)\xi_1 + {}_T\psi_-^0\xi_2 + {}_T\psi(f)(t), \qquad t \in [\delta, T] \tag{4.95}$$

where ${}_T\psi_+^0$, ${}_T\psi_-^0$ fulfill the equations

(4.96)

$${}_T\psi_+^0 - H_T^{(2)}\tilde{A}E_T\psi_+^0 = P_{(1)}(\cdot)\left[\begin{bmatrix}\psi_+(\cdot, T) & \theta\\ \theta & \theta\end{bmatrix}G_+, P_{(2)}(\cdot)P_{(3)}(\cdot)e^{Q(\cdot)-Q(T)}\left(\frac{\cdot}{T}\right)^D \tilde{\tilde{G}}_0\right],$$

$${}_T\psi_-^0 - H_T^{(2)}\tilde{A}E_T\psi_-^0 = \phi[\tilde{G}_0, G_-] \tag{4.97}$$

and ${}_T\psi(f)$ fulfills

$${}_T\psi(f) - H_T^{(2)}\tilde{A}_T\psi(f) = H_T^{(2)}f. \tag{4.98}$$

By evaluating at the boundaries the following block system is generated:

$$\begin{bmatrix} BE_T\psi_+^0(1) & BE_T\psi_-^0(1)\\ S(T)E_T\psi_+^0(T) & S(T)E_T\psi_-^0(T)\end{bmatrix}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix} = \begin{pmatrix}\beta - BE_T\psi(f)(1)\\ \gamma(T) - S(T)E_T\psi(f)(T)\end{pmatrix}. \tag{4.99}$$

Obviously, the matrix in the (1,1) position is bounded as $T \to \infty$. From (4.97) we derive

$$S(T)E_T\psi_-^0(T) = S(T)E\phi(T)[\tilde{G}_0, G_-] + S(T)E(H_T^{(2)}\tilde{A}E_T\psi_-^0)(T). \tag{4.100}$$

From the definition of $H_T^{(2)}$ we conclude that

$$S(T)E(H_T^{(2)}\tilde{A}E_T\psi_-^0)(T) = S(T)EP(T)\tilde{D}_0e^{Q(T)}T^D\int_\delta^T \bar{\tilde{D}}_0e^{-Q(s)}s^{-D}P^{-1}(s) \cdot E^{-1}\tilde{A}E_T\psi_-^0(s)\,ds + S(T)E({}_TH_-^{(2)}\tilde{A}_T\psi_-^0)(T), \tag{4.101}$$

where $\bar{\tilde{D}}_0$ is the projection onto the direct sum of invariant subspaces of D belonging to case (II) as in (3.36).

Now from Markowich (1980b, § 4) we conclude

$$\|({}_TH_-^{(2)}\tilde{A}u)(T)\| = \|(H_-\tilde{A}u)(T)\| = O(T^{-(\alpha+1)\bar{r}-\varepsilon/2}), \qquad u \in C([\delta, T]). \tag{4.102}$$

This and (B_2) guarantee that

$$\|S(T)E_T\psi_-^0(T)\| = o(T^{-(\alpha+1)(\bar{r}-1)}) \tag{4.103}$$

holds. Similarly, (4.96) implies

$$S(T)E_T\psi_+^0(T) = S(T)EP_{(1)}(T)[G_+, P_{(2)}(T)P_{(3)}(T)\tilde{\tilde{G}}_0] + o(T^{-(\alpha+1)(\bar{r}-1)}). \tag{4.104}$$

Markowich (1980b) has proven that

$$\psi_-^0C - \tilde{H}\tilde{A}E\psi_-^0C = \phi[\tilde{G}_0, G_-] \tag{4.105}$$

where $\tilde{H}$ is defined as $H_T^{(2)}$, except that the integrals are stretched to ∞ instead of T and C is a regular $(\tilde{r}_0+r_-)\times(\tilde{r}_0+r_-)$ matrix. We subtract (4.97) from (4.105), getting

$$(\psi_-^0 C - {}_T\psi_-^0) - \tilde{H}\tilde{A}E\psi_-^0 C + H_T^{(2)}\tilde{A}E_T\psi_-^0 = 0 \tag{4.106}$$

or

$$(\psi_-^0 C - {}_T\psi_-^0) = H_T^{(2)}\tilde{A}E(\psi_-^0 C - {}_T\psi_-^0) + (\tilde{H}\tilde{A}E\psi_-^0 C - H_T^{(2)}\tilde{A}E\psi_-^0 C), \quad \delta \leqq t \leqq T \tag{4.107}$$

and therefore

$$\|\psi_-^0 C - {}_T\psi_-^0\|_{[\delta,T]} \leqq \text{const.}\, \|\tilde{H}\tilde{A} - H_T^{(2)}\tilde{A}\|_{[\delta,T]}. \tag{4.108}$$

By continuation to $[1, T]$ we get

$$\lim_{T\to\infty} \|\psi_-^0 C - {}_T\psi_-^0\|_{[1,T]} = 0 \tag{4.109}$$

so that

$${}_T\psi_-^0(1) = \psi_-^0(1)C + o(1) \quad \text{as } T\to\infty. \tag{4.110}$$

Also we derive

$$S(T)E_T\psi(f)(T) = S(T)E(H_T^{(2)}f)(T) + o(T^{-(\alpha+1)(\tilde{r}-1)})\|f\|_{[\delta,T]}. \tag{4.111}$$

Theorem 4.2 follows now from (4.103), (4.104), (4.110), (4.111) by considering the system (4.99) as in the proof of Theorem 4.1. The convergence results change correspondingly to (4.83), (4.84):

$$\begin{aligned}\|y - x_T\|_{[1,T]} \leqq \text{const.}\, (T^{(\alpha+1)(\tilde{r}-1)}(&\|S(T)E\phi(T)[\tilde{G}_0, G_-]\| \\ &+ \|(Hf)(T)\|) + T^{-\varepsilon}(\ln T)^j).\end{aligned} \tag{4.112}$$

5. Nonlinear problems. Now we deal with problems of the following form:

$$y' = t^\alpha f(t, y), \qquad 1 \leqq t < \infty, \quad \alpha \in \mathbb{N}. \tag{5.1}$$

$$y \in C([1, \infty]), \tag{5.2}$$

$$b(y(1)) = 0. \tag{5.3}$$

$f: \mathbb{R}^{n+1} \to \mathbb{R}^n$ is supposed to be continuous in $(\infty, y(\infty))$. Equations (5.1) and (5.2) imply that

$$f(\infty, y(\infty)) = 0 \tag{5.4}$$

holds. So $y(\infty)$ can be calculated a priori as a solution of a system of n nonlinear equations. If

$$\operatorname{rank}\left(\frac{\partial f(\infty, y(\infty))}{\partial y}\right) = n \tag{5.5}$$

then the solution manifold is discrete so that the possible values of $y(\infty)$ are known a priori. This case has been treated by de Hoog and Weiss (1980b). We will assume that the rank of this matrix is smaller than n, so that we have to expect a continuous solution manifold $y_\infty(\mu)$ with $\mu \in S \subset \mathbb{R}^{n_1}$, $n_1 \leqq n$. We assume that we have determined such a n_1-dimensional manifold and that $f(t, \cdot) \in C^1(\mathbb{R}^n)$ for all $t \in [1, \infty]$ and

$$A(t, \mu) = \frac{\partial f(t, y_\infty(\mu))}{\partial y} = \sum_{i=0}^{\infty} A_i(\mu) t^{-i} \quad \text{for } t \geqq \tilde{t} \tag{5.6}$$

holds. We calculate the fundamental matrix $E(\mu)\phi(t, \mu)$ of the linearized system

$$\tilde{y}' = t^{\alpha} A(t, \mu)\tilde{y} \tag{5.7}$$

as an asymptotic series. $E(\mu)$ transforms $A_0(\mu)$ to its Jordan canonical form $J_0(\mu)$.

Now we restrict μ to subsets $\tilde{S} \subset S$ so that r_+, r_0, r_-, $\bar{r}$ which are defined for $J_0(\mu)$ as in the last sections are independent of μ in $\tilde{S}$. Moreover, we require that there is an $n \times \hat{r}_0$ projection-like matrix $\hat{G}_0$ independent of $\mu \in \tilde{S}$ so that

$$\|\phi(t, \mu)[\hat{G}_0, G_-]\| \leqq C(\mu) t^{-(\alpha+1)\bar{r}-\varepsilon_1(\mu)} (\ln t)^j, \qquad \varepsilon_1(\mu) > 0, \quad 0 \leqq j \leqq n \tag{5.8}$$

and that

$$\|f(t, y_\infty(\mu))\| \leqq C(\mu) t^{-2(\alpha+1)\bar{r}-\varepsilon_2(\mu)}, \qquad \varepsilon_2(\mu) > 0, \quad \mu \in \tilde{S} \tag{5.9}$$

holds. Assuming that f_y is locally uniformly Lipschitz continuous around $y_\infty(\mu)$, Markowich (1980b) showed that there are solutions $y = y(\cdot, \mu, \eta)$ in the space of functions in $C([\delta, \infty])$ which decay to a finite limit at least as fast as $t^{-(\alpha+1)\bar{r}-\varepsilon} \cdot (\ln t)^{2j}$ where $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$ and δ is sufficiently large. These solutions fulfill the estimate:

$$\|y(t, \mu, \eta) - y_\infty(\mu) - E(\mu)\phi(t, \mu)[\hat{G}_0, G_-]\eta\| \leqq \text{const.}\,(\mu) \cdot (\ln t)^{2j} t^{-(\alpha+1)\bar{r}-\varepsilon} \tag{5.10}$$

for $\eta \in \mathbb{R}^{\hat{r}_0 + r}$. For many important applications

$$f(t, y_\infty(\mu)) \equiv 0 \tag{5.11}$$

and $\phi(t, \mu)[\hat{G}_0, G_-]$ decays exponentially. In this case the right-hand side of (5.10) contains the exponential factor $\|\phi(t, \mu)[\hat{G}_0, G_-]\|^2$ and the algebraic and logarithmic factors may be different. It follows from this analysis that the boundary value problem (5.1), (5.2), (5.3) is soluble if the equation

$$b(y(1, \mu, \eta)) = 0 \tag{5.12}$$

is soluble where $b: \mathbb{R}^n \to \mathbb{R}^{n_1 + \hat{r}_0 + r_-}$ and $y(t, \mu, \eta)$ denotes the continuation to $[1, \infty]$ (if it exists). We assume that $b \in C^1(\mathbb{R}^n)$.

The approximating problems have the form

$$x_T' = t^{\alpha} f(t, x_T), \qquad 1 \leqq t \leqq T, \tag{5.13}$$

$$b(x_T(1)) = 0, \tag{5.14}$$

$$S(x_T(T), T) = 0 \tag{5.15}$$

and $S: \mathbb{R}^{n+1} \to \mathbb{R}^{n-(n_1+\hat{r}_0+r_-)}$. We assume that we have obtained a solution $y^* \equiv y(\cdot, \mu^*, \eta^*)$ fulfilling (5.10) and that this solution is isolated, i.e., the linearized problem

$$w' = t^{\alpha} \frac{\partial f(t, y^*(t))}{\partial y} w, \tag{5.16}$$

$$w \in C([1, \infty]), \tag{5.17}$$

$$\frac{\partial b}{\partial y}(y^*(1)) w(1) = 0 \tag{5.18}$$

has only the trivial solution $w \equiv 0$. Using (5.10) we get for f_y Lipschitz continuous

$$\frac{\partial f(t, y^*(t))}{\partial y} = A(t, \mu^*) + O(t^{-(\alpha+1)\bar{r}-\varepsilon(\mu^*)/2}). \tag{5.19}$$

From Markowich (1980b, § 4) we derive that the general solution of (5.16), (5.17) is

$$w = E(\mu^*)\psi_-^0(t, \mu^*, \eta^*)\xi, \qquad \xi \in \mathbb{C}^{\tilde{r}_0 + r_-}. \tag{5.20}$$

ψ_-^0 is an $n \times (\tilde{r}_0 + r_-)$ matrix.

For the following we assume that

$$\tilde{r}_0 = n_1 + \hat{r}_0 \tag{5.21}$$

holds, which means that the nonlinear problem and the linearized problem have solution manifolds of the same dimension. The isolatedness of y^* now implies that the $(\tilde{r}_0 + r_-) \times (\tilde{r}_0 + r_-)$ matrix

$$\frac{\partial b}{\partial y}(y^*(1))E(\mu^*)\psi_-^0(1, \mu^*, \eta^*) \tag{5.22}$$

is regular. Now we define:

$$(Fy)(t) = (t^{-\alpha}y' - f(t, y), b(y(1)), S(y(T), T) - S(y^*(T), T)), \tag{5.23}$$

$$\begin{aligned} F: (C([1, \infty]) \cap C^1([1, \infty)) \cap \{y \mid t^{-\alpha}y' \in C([1, \infty])\}, \|\cdot\|_{[1,\infty]}) \\ \to (C([1, \infty]) \times R^n, \|(v, a)\| = \|v\|_{[1,\infty]} + \|a\|) \end{aligned} \tag{5.24}$$

and

$$(F_T x)(t) = (t^{-\alpha}x' - f(t, x), b(x(1)), S(x(T), T)), \tag{5.25}$$

$$\begin{aligned} F_T: (C^1([1, T]), \|x\| = \|x\|_{[1,T]} + T^{-\alpha}\|x'\|_{[1,T]}) \\ \to (C([1, T]) \times \mathbb{R}^n, \|(w, b)\| = \|w\|_{[1,T]} + \|b\|). \end{aligned} \tag{5.26}$$

All involved spaces are linear normed spaces and the space on which F_T is defined is a Banach space.[1] We calculate the Frechet derivative of $F_T(y^*)$ where y^* is an (isolated) solution of $F(y^*) = 0$ assuming that $S \in C^1(\mathbb{R}^n)$:

$$((F_T'(y^*))z)(t) = \left(t^{-\alpha}z' - \frac{\partial f(t, y^*(t))}{\partial y}z, \frac{\partial b}{\partial y}(y^*(1)), \frac{\partial S(y^*(T), T)}{\partial y}\right). \tag{5.27}$$

Assuming that

$$\begin{aligned} &\frac{\partial f}{\partial y}(t, \cdot) \text{ is locally Lipschitz continuous in the } \|\cdot\|_{[1,\infty]}\text{-norm} \\ &\qquad \text{around } y^* \text{ uniformly in } t \in [1, \infty], \end{aligned} \tag{5.28}$$

$$\begin{aligned} &\frac{\partial b}{\partial y}(y(1)), \frac{\partial S}{\partial y(T)}(y(T)) \text{ are locally Lipschitz continuous} \\ &\qquad \text{around } y^*(1), y^*(T) \text{ respectively} \end{aligned} \tag{5.29}$$

hold, we derive

$$\|F_T'(y_1) - F_T'(y_2)\| \leqq \text{const.} \|y_1 - y_2\| \tag{5.30}$$

for

$$\|y^* - y_i\|_{[1,T]} + T^{-\alpha}\|y^{*\prime} - y_i'\|_{[1,T]} \leqq \text{const.}, \qquad i = 1, 2. \tag{5.31}$$

Moreover Theorem 4.2 assures that the problem

$$z' = t^{\alpha}\frac{\partial f(t, y^*(t))}{\partial y}z + t^{\alpha}f(t), \tag{5.32}$$

[1] Norms are always assumed to be taken in the appropriate spaces.

$$\frac{\partial b}{\partial y}(y^*(1))z(1)=\beta, \tag{5.33}$$

$$\frac{\partial S(y^*(T),T)}{\partial y}z(T)=\gamma(T) \tag{5.34}$$

is for all $f\in C([1,T])$, $\beta\in\mathbb{R}^{\tilde{r}_0+r_-}$, $\gamma(T)\in\mathbb{R}^{n-(\tilde{r}_0+r_-)}$ uniquely soluble if

$$\frac{\partial S(y^*(T),T)}{\partial y} \quad \text{fulfills } (A_2),\ (B_2),\ (C_2) \tag{5.35}$$

where

$$E(\mu^*)\phi(t)=E(\mu^*)\phi(t,\mu^*)=E(\mu^*)P(t,\mu^*)t^{D(\mu^*)}e^{Q(t,\mu^*)} \tag{5.36}$$

is the fundamental matrix of the problem

$$\tilde{w}'=t^{\alpha}\frac{\partial f(t,y_\infty(\mu^*))}{\partial y}\tilde{w}. \tag{5.37}$$

It also implies that the unique solution fulfills

$$\|x\|_{[1,T]}+\frac{1}{T^{\alpha}}\|x'\|_{[1,T]}\leqq\text{const. } T^{(\alpha+1)\bar{r}}(\ln T)^{2j}\|(f,\beta,\gamma(T))\| \tag{5.38}$$

so that $F_T'(y^*)$ is invertible and

$$\|(F_T'(y^*))^{-1}\|\leqq\text{const. } T^{(\alpha+1)\bar{r}}(\ln T)^{2j}. \tag{5.39}$$

We have used the fact that (5.19) holds. From the nonlinear stability-consistency concept in Spijker (1971) we conclude that

$$\|x_T-y^*\|_{[1,T]}+T^{-\alpha}\|x_T'-(y^*)'\|_{[1,T]}\leqq\text{const. } T^{(\alpha+1)\bar{r}}(\ln T)^{2j}\|S(y^*(T),T)\| \tag{5.40}$$

if

$$\|S(y^*(T),T)\|\leqq\rho_1 T^{-2(\alpha+1)\bar{r}}(\ln T)^{-4j} \tag{5.41}$$

holds, where ρ_1 is sufficiently small. x_T is a solution of $F_T(x)=0$ which is unique in a sphere whose center is the restriction of y^* to $[1,T]$ and whose radius is smaller than $\rho_2 T^{-(\alpha+1)\bar{r}}(\ln T)^{-2j}$ with ρ_2 sufficiently small. This holds in the $\|x\|_{[1,T]}+T^{-\alpha}\|x'\|_{[1,T]}$-norm. From (5.40) we conclude

$$\|x_T-y^*\|_{[1,T]}\leqq\text{const. } T^{(\alpha+1)\bar{r}}(\ln T)^{2j}\|S(y^*(T),T)\|, \qquad 0\leqq j\leqq n \tag{5.42}$$

if (5.35) and (5.41) hold.

Because of (5.10) it is sufficient to require that (A_2), (B_2), (C_2) hold for the matrix $\partial S/\partial y(y_\infty(\mu^*),T)$ instead for $\partial S/\partial y(y(T,\mu^*,\eta^*),T)$. Moreover, (5.41) is fulfilled if

$$S(y_\infty(\mu^*),T)\equiv 0 \quad \text{for } T \text{ sufficiently large} \tag{5.43}$$

and

$$\left\|\frac{\partial S}{\partial y}(y_\infty(\mu^*),T)(y(T,\mu^*,\eta^*)-y_\infty(\mu^*))\right\|=O(T^{-2(\alpha+1)\bar{r}-\varepsilon}), \qquad \varepsilon>0. \tag{5.44}$$

In most cases of physical interest $y^*(T)$ converges exponentially so that (5.44) is fulfilled automatically. Therefore, if (A_2), (B_2), (C_2) hold for $\partial S/\partial y(y_\infty(\mu^*),T)$ and if (5.43) is fulfilled, convergence follows at isolated solutions and the order of convergence can be estimated by (5.42).

For the case when f is independent of t, Lentini and Keller (1980) have generalized the projection condition and an example for the construction of an appropriate asymptotic boundary condition in the other case will be presented in § 6.

6. A case study. The problem we analyze is a similarity equation for a combined formed and free convection flow over a horizontal plate (see Schneider (1979)). The governing equations are

$$y' = x \begin{pmatrix} \frac{y_2}{x} \\ \frac{y_3}{x} \\ -\frac{1}{2}\left(1+\frac{y_1}{x}\right)y_2 - \frac{k}{2}y_4 \\ -\frac{1}{2}\left(1+\frac{y_1}{x}\right)y_4 \end{pmatrix} = xf(x, y), \qquad 0 \leqq x < \infty, \tag{6.1}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} y(0) = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}, \tag{6.2}$$

$$y \in C([0, \infty]). \tag{6.3}$$

In this section x instead of t and X instead of T are used. From (6.1) we conclude that

$$y_\infty = y_\infty(\mu) = (\mu, 0, 0, 0)^T, \qquad \mu \in \mathbb{R} \tag{6.4}$$

and

$$\frac{\partial f(x, y_\infty(\mu))}{\partial y} = \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & -k/2 \\ 0 & 0 & 0 & -\frac{1}{2} \end{bmatrix}}_{A_0} + \frac{1}{x}\underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\mu/2 & 0 \\ 0 & 0 & 0 & -\mu/2 \end{bmatrix}}_{A_1(\mu)}. \tag{6.5}$$

For this problem $\alpha = 1$, $\bar{r} = 2$ hold. We calculate:

$$J_0 = E^{-1}A_0E = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & 1 \\ 0 & 0 & 0 & -\frac{1}{2} \end{bmatrix}, \qquad E = \operatorname{diag}\left(1, 1, 1, -\frac{2}{k}\right), \tag{6.6}$$

$$J_1(\mu) = E^{-1}A_1(\mu)E = A_1(\mu). \tag{6.7}$$

Markowich (1980b) calculated an asymptotic expression for the fundamental matrix $\phi(x, \mu)$ of the system

$$\tilde{w}' = x\left(J_0 + \frac{1}{x}J_1(\mu)\right)\tilde{w}, \tag{6.8}$$

(6.9)

$$\phi(x,\mu)=\underbrace{\begin{bmatrix}1 & 1 & O(x^{-2}) & O(x^{-1})\\ 0 & x^{-1} & O(x^{-1}) & O(x^{-1})\\ 0 & 0 & 1 & 1-x^{-2}\\ 0 & 0 & 0 & x^{-2}\end{bmatrix}}_{P(x,\mu)}\underbrace{\operatorname{diag}(1,x,1,x^2)}_{x^D}\underbrace{\operatorname{diag}(1,1,e^{-x^2/4-\mu x/2},e^{x^2/4-\mu x/2})}_{e^{Q(x,\mu)}}.$$

Markowich (1980b) showed that the problem (6.1) has solutions $y(\cdot,\xi_1,\xi_2,\mu)$ which fulfill

$$\|y(x,\xi_1,\xi_2,\mu)-y_\infty(\mu)\|\leqq \text{const. } x^2e^{-x^2/4-\mu x/2}, \tag{6.10}$$

for x sufficiently large, where the constant depends linearly on ξ_1, ξ_2. These solutions are in $A_{\varepsilon,\delta}+\{y_\infty(\mu)\}$ where

$$A_{\varepsilon,\delta}=\{u\,|\,u(x)=x^{-4-\varepsilon}U(x),\ U\in C_b([\delta,\infty))\},\qquad \varepsilon>0,\quad \delta\geqq 0. \tag{6.11}$$

From (6.9) we conclude that

$$\tilde{G}_0=[1,0,0,0]^T,\qquad \tilde{\tilde{G}}_0=[0,1,0,0]^T. \tag{6.12}$$

The simplest choice of S is a linear function so we set

$$S(X)=[s_1(X),s_2(X),s_3(X),s_4(X)]. \tag{6.13}$$

Condition (B_2) of Theorem 4.1 applied to our problem gives

$$s_1(X)=O(X^{-2}). \tag{6.14}$$

We choose $s_1(X)\equiv 0$. Condition (C_2) gives

$$(s_2(X)X^{-1})^{-1}=O(X^2). \tag{6.15}$$

Therefore any matrix $S(X)$ of the form

$$S(X)=[0,s_2(X),s_3(X),s_4(X)], \tag{6.16}$$

where

$$s_2(X)=\text{const.}\neq 0,\quad s_3(X)=O(1),\quad s_4(X)=O(1) \tag{6.17}$$

fulfills (A_2), (B_2), (C_2). A natural choice is the following asymptotic boundary condition

$$[0,1,0,0]v_X(X)=0, \tag{6.18}$$

which assures convergence of the order

$$\|v_X-y\|_{[1,X]}\leqq \text{const. } X^6\exp\left(-\frac{X^2}{4}-\frac{\mu^*}{2}X\right)(\ln X)^2 \tag{6.19}$$

where μ^* is the parameter value of the actual solution $y(\cdot,\mu^*,\xi_1^*,\xi_2^*)$ of (6.1), (6.2), (6.3) which is assumed to be isolated. Inequality (6.19) holds because

$$[0,1,0,0]y_\infty(\mu)\equiv 0\quad \text{for } \mu\in\mathbb{R} \tag{6.20}$$

and because of (5.41).

Numerical calculations can be found in Schneider (1978).

REFERENCES

E. CODDINGTON AND N. LEVINSON (1955), *Theory of Ordinary Differential Equations*, McGraw-Hill, New York.

F. R. DE HOOG AND R. WEISS (1980a), *On the boundary value problem for systems of ordinary differential equations with a singularity of the second kind*, this Journal, 11, pp. 41–60.

———, (1980b), *An approximation method for boundary value problems on infinite intervals*, Computing, 24, pp. 227–239.

M. LENTINI AND H. B. KELLER (1980a), *Boundary value problems over semi-finite intervals and their numerical solution*, SIAM J. Numer. Anal., 17, pp. 577–604.

———, (1980b), *The von Karman swirling flows*, SIAM J. Appl. Math., 38, pp. 52–64.

P. A. MARKOWICH (1980a), *Randwertprobleme auf unendlichen Intervallen*, Dissertation, Technische Universität Wien, Vienna.

———, (1980b), *Analysis of boundary value problems on infinite intervals*, Tech. Summary Rep. 2138, Mathematics Research Center, University of Wisconsin, Madison.

W. SCHNEIDER (1979), *A similarity solution for combined forced and free convection flow over a horizontal plate*, Int. J. Heat Mass Transfer, 22, pp. 1401–1406.

M. N. SPIJKER (1971), *Equivalence theorems for nonlinear finite difference methods*, in Numerische Lösung nichtlinearer Partieller Differential- und Integrodifferentialgleichungen, Lecture Notes in Mathematics 267, Springer-Verlag, Berlin.

W. WASOW (1965), *Asymptotic Expansion for Ordinary Differential Equations*, Series in Pure and Applied Mathematics, vol. XIV, John Wiley, New York.

PROFESSIONAL BIOGRAPHY: MICHAEL GOLOMB

Michael Golomb was born in Munich, Germany, on May 3, 1909. His family soon moved to Würzburg, where Michael received most of his schooling. He moved to Berlin for doctoral research, and was awarded the Ph.D. degree in 1933, under the supervision of A. Hammerstein and Erhard Schmidt [1]. His dissertation and early mathematical work were in the field of nonlinear integral and general functional equations.

There was no possibility of a career under the Nazi regime, and so it was necessary to leave Germany. Europe and the United States were in a deep economic depression and had closed their borders to anyone seeking employment. Michael moved to Yugoslavia where he had been offered an assistantship at the University of Belgrade, but the state authorities intervened and annulled the offer. He remained there during the next several years, living under constant threat of deportation. Michael gave some lectures at the University of Zagreb and other institutions, but had no chance for employment (academic or otherwise). He had all but given up hope of a career in mathematics when in 1939 he received an immigration visa from the United States; by this time he was a stateless person. Most of Michael's closest family perished during the last year of World War II in the Nazi extermination camps.

Shortly before leaving Europe, he married Dagmar Račić, a mathematics student at the University of Zagreb. They have two children, Miriam and Deborah.

Upon arrival in the United States, it was necessary to find a position, and through the help of Hermann Weyl he obtained an appointment as research fellow in the Electrical Engineering School of Cornell and one year later an additional instructorship in the Mathematics Department. From 1942 until his retirement in 1975 Michael was on the faculty of Purdue University, although he spent numerous periods away during World War II (war-related research) and on academic leaves. His activities during these leaves have influenced many mathematicians. Lectures at the Argonne National Laboratory were collected in the influential publication *Lectures on Theory of Approximation* [24], and the effect of a similar series delivered in Bréau (France) is affectionately confirmed in the dedication and preface to Laurent's *Approximation et optimisation* [La]. Michael was a Member-at-Large of the Council of the American Mathematical Society (1968–1971) and has been a Fellow of the American Association for the Advancement of Science for over thirty years. He is at present Member-at-Large of the Mathematics Section Committee of the AAAS.

Michael has been a respected and admired teacher and colleague. His wide knowledge and experience in mathematics and applications were recognized by Purdue University, which gave him a joint appointment in the School of Engineering. The strengthening of the Purdue Mathematics Department in the 1950's and 1960's owes much to Michael's experience, insight and counsel. Michael also has coached the Purdue team in the Putnam competition for several years. The Mathematics Department showed its appreciation for his contribution to the teaching program by instituting the Michael Golomb Award in 1975. This award is given annually to the undergraduate with the best performance in this competition. Michael has directed several doctoral dissertations. The second author of this biographical note is among this group, and remains grateful for Michael's guidance.

Michael Golomb's research interests have included nonlinear analysis, ordinary differential equations, numerical analysis, approximation theory and extremal problems in analysis; these contributions are discussed in the next three sections. Lettered

references refer to the References which immediately follow the text, while numbered citations are listed in Michael Golomb's professional bibliography, which immediately follows the References. The concluding entry is a list of his doctoral students.

I. Nonlinear Analysis and Dynamical Systems

These topics presently correspond to the properties of mappings in infinite dimensional spaces (or manifolds) and to the study of flows on manifolds, respectively. There is a tendency today (cf. L. Nirenberg [N]) to categorize nonlinear methods via global topological methods, variational methods and local perturbation study. Gradient mappings, monotone operators (for generalizations, cf. H. Brezis [Br]) and their perturbations play a decisive role in much of this work.

Golomb's early work touches on all of these topics. Item [1] is remarkably modern in outlook, though it is an early study of problems in nonlinear functional equations, mostly integral equations, especially of Hammerstein type. Existence and uniqueness of solutions are proved by successive approximation, fixed point theorems and direct methods of the variational calculus. Several concepts that proved to be fundamental in the later development of functional analysis were introduced in this paper. These include: gradient mappings in Hilbert space, later studied with R. A. Tapia in [40] as "metric gradient" in normed linear spaces; also, the splitting $K = H^*H$ of a linear integral operator, involving what is known in the Soviet literature as the Golomb–Hammerstein functional; and, finally, strongly monotone mappings, though the name is not used. Both G. J. Minty [M] and F. E. Browder and C. P. Gupta [BG] have recognized [1] as the historical antecedent of monotonicity in nonlinear functional analysis. Item [2] is a sequel to [1] with essentially the same methods applied to some more general equations and systems. Many of the problems of these early papers and the methods developed there were later extended and systematized by Soviet mathematicians, especially M. Vainberg, N. Nazarov, M. Krasnosel'skii and Ya. Rutickii, and by American mathematicians such as C. L. Dolph, G. J. Minty, F. E. Browder, C. P. Gupta, I. Kolodner, S. Weiss et al.

Items [18], [21] and [23] deal with the existence of periodic and almost periodic solutions of ordinary linear or nonlinear differential equations and systems, and with the stability of such solutions. Constructive existence proofs, based on novel kinds of expansions, are provided; new expansions are discovered even for the solutions of the classical equations of Mathieu and Hill. Also, some stability results, related to those of L. Cesari and J. Hale, are obtained. Recent work by these authors is described in [CKW].

Golomb's publications in this area ended with [2], but several of his students were introduced to it under Michael's direction.

II. Numerical Analysis and Spline Functions

Golomb's major contribution to Numerical Analysis is the joint paper [19] with H. F. Weinberger *Optimal Approximation and Error Bounds.* It is a basic paper on what is called "optimal recovery" these days; see [MR]. It deals with the basic computational task of evaluating or estimating a linear functional μ on some function g from information about g. This information consists of the numbers $\lambda_1 g, \cdots, \lambda_n g$ for some known linear functionals $\lambda_1, \cdots, \lambda_n$ and, in addition, some bound on some seminorm νg of g. The authors rightly take issue with the traditional approach in which one is content to show that the error in some rule for μg based on $\lambda_1, \cdots, \lambda_n$ is $O(n^{-a}\nu g)$; the authors' characterization of this O as the "fig leaf [which] covers

the offending member" cannot be improved upon. The authors propose to find the best estimate for μg based on the available information about g. They show how to provide this best estimate in case the seminorm ν derives from a quadratic form. In that case, the best estimate is of the form μPg, with Pg a "spline interpolant" to g.

The relationship of [19] to spline theory is a curious one. It serves to illustrate the importance of using the "right" terminology. For, while [19] contains in considerable generality various aspects of the variational approach to splines in a Hilbert space setting, the word "spline" never occurs. As a consequence, others are now usually credited with certain basic spline results which first appeared in [19].

It pays to be a bit more explicit, also in order to prepare for the discussion of other papers. The abstract variational approach to splines (as first proposed in a Hilbert space setting by Atteia in [A]) produces splines as the solution to the following problem: A collection Λ of continuous linear functionals on some linear space X is given. Since the problem will only involve

$$\Lambda_\perp := \{x \in X: \lambda x = 0 \text{ for all } \lambda \in \Lambda\},$$

we might as well assume that Λ is a closed linear subspace. Further, a bounded linear map T on X to some normed linear space Y is given. For given $g \in X$, determine, if possible, an element g^* at which the map

$$g + \Lambda_\perp \to \mathbf{R}_+: y \mapsto \|Ty\|$$

takes on its minimum. Such a minimizer g^* is called a (T, Λ)-spline interpolant to g. The word "interpolant" is appropriate since g^* agrees with g on Λ. Now [19] deals with the special case: $X = Y$ is a Hilbert space and $T = 1$. In this setting, as can be inferred from [19], the map $g \mapsto g^*$ is just the orthoprojector P_Λ onto Λ considered as a subspace of X. Further—and this constitutes the main result of [19]—if we only know g on Λ together with the fact that $\|g\| \leqq r$, then, for any linear functional μ,

$$|\mu g - \mu P_\Lambda g| \leqq (\|\mu\|^2 - \|\mu P_\Lambda\|^2)^{1/2}(r^2 - \|P_\Lambda g\|^2)^{1/2}$$

is the sharpest possible estimate for μg. It is instructive to relate this to the bound

$$|\mu g - \mu P g| \leqq \operatorname{dist}(\mu, \operatorname{ran} P')\|1 - P\| \operatorname{dist}(g, \operatorname{ran} P)$$

obtainable for any bounded linear projector P on any normed linear space. Incidentally, Sard's treatment of the problem [Sa2] only provides the bound

$$|\mu g - \mu P_\Lambda g| \leqq \operatorname{dist}(\mu, \Lambda) r.$$

It should be stressed that $P_\Lambda g$ is constructible entirely in terms of the given data, viz., the values of g on Λ. This requires, though, the imbedding of Λ in X, i.e., the construction of the representer for any $\lambda \in \Lambda$. Unless Λ is already introduced into the problem in terms of its representers, this task can, in general, be accomplished only with the aid of a reproducing kernel, a fact rightly stressed in [19]. The work of [19] was continued in [22].

Several of Golomb's later papers take up again the variational approach to splines, in the form of an extension problem: The function g is only known on Λ (as a linear function, of course) and one wonders whether there is some function $G \in X$ for which $G_{|\Lambda} = g$. Formulated this way, one recognizes this, of course, as a problem which can be solved with the aid of the Hahn-Banach theorem (as is the case with the problem posed in [19]). It should be clear, though, that there will be additional difficulties in case X is not reflexive. These extension problems are usually formulated more concretely: X is a space of functions on some domain and Λ is spanned by the linear

functionals $g \mapsto g(a)$ with a taken from a closed set A. In this form, such extension theorems go back at least to Hassler Whitney who in [W] established a simple necessary and sufficient condition for a function on some closed set A in R to be the restriction to A of some $C^{(m)}(R)$ function.

Golomb's interest in such problems seems to have been sparked by Schoenberg. According to [30], I. J. Schoenberg posed this extension problem for the function space $X = H^m = L_2^{(m)}$ of m-fold integrals of functions in $L_2(R)$. According to [32], this problem is attacked in [31] (a paper which only exists by implication, i.e., as a reference in [30], [32], [37] and [39] with slightly differing titles) by considering a minimal H^m-extension of g, i.e., an extension for which $\int |G^{(m)}|^2$ is as small as possible. This is, of course, entirely in the spirit of splines and the extensions so obtainable make up the class $S^m(A)$ of *H^m-splines with knots in A*, i.e., the class of functions in H^m which, on any open interval in $R \backslash A$, agree with some polynomial of degree $<2m$ and which are in C^{2m-2} on every open interval in $R \backslash A'$. Here, A' is the set of limit points of A. Further (now quoting [30] for some of the details), an extension (and therefore such a minimal extension) is shown in [31] to exist if and only if the sequence (s_n) is H^m-bounded, with s_n the natural $2m$-order spline interpolant to g at the points $x_1, \cdots, x_n$ and $(x_1, x_2, \cdots)$ a dense sequence in A. To this, [32] adds a more detailed analysis: The construction of a minimal extension for $g = f|_A$ for $g \in H^m$ is shown to be the orthogonal projection of f onto $S^m(A)$, thus $S^m(A)^\perp = \{f \in H^m : f|_A = 0\}$. It is also shown that the extension problem decomposes into essentially disjoint problems, one for each of the countably many disjoint open intervals whose union makes up $R \backslash A'$. Special attention is therefore paid to the special case when A consists of the entries of an infinite strictly increasing sequence.

In [39], paper [32] is extended in the following directions: H^m is replaced by $H^{m,p} = L_p^{(m)}$ for some $1 < p < \infty$. Points in A may be multiple, i.e., on some points of A, the first so many derivatives of g may be prescribed along with the function value. In other words, a "Taylor field" is given on A. In the case of particular interest that A consists of the entries of an infinite nondecreasing sequence (x_i) with finite global mesh ratio $\sup_{i,j} \Delta x_i / \Delta x_j$, the necessary and sufficient condition for $H^{m,p}$-extendability patterned after [31] and [32] is replaced by the more concise one that

$$\sum (x_{i+m} - x_i) |[x_i, \cdots, x_{i+m}]g|^p < \infty$$

with $[x_i, \cdots, x_{i+m}]g$ the mth divided difference of g at $x_i, \cdots, x_{i+m}$. Subsequent work in [B] has shown this to be correct for arbitrary sequences (x_i).

Golomb also contributed substantially to the constructive approach to splines. In [33], the reader has a chance to share with the author the delight that comes with the discovery that the polynomial spline interpolant on a uniform mesh to the complex exponential function shares so many properties with that function. The circulant character of the matrices involved permits explicit calculation of error expansions for periodic spline interpolation on uniform meshes and Golomb makes full use of that. These explicit error expansions are put to good use in [34] (as corrected by [44]) where the effect of jump discontinuities in some derivative of the function to be interpolated is gauged so precisely that one could use the information to obtain good estimates of the location and size of the discontinuity from the behavior of the interpolant.

An alternative approach to Golomb's results in [33] is provided by Schoenberg's elegant analysis of cardinal spline interpolation, detailed in [S].

Finally, in [46] and [47], Golomb successfully tackles, jointly in part with J. Jerome, the difficult analysis of true splines, i.e., of curves modelling a flexible rod.

III. Approximation Theory and Optimization

Modern approximation theory proceeds by a reduction of complexity often involving well-defined extremal characterizations and the associated measures of deviation. Representative surveys of the field may be found in M. Golomb [24], G. G. Lorentz [Lo], A. Sard [Sa1] and L. L. Schumaker [Sc]. One of the central issues in the field is that of understanding, preferably by means of algorithmic procedures, the complexity of functions of several variables.

Item [20] is a survey of approximations, optimal in some metric, of functions of several variables by finite aggregates of functions of fewer variables of prescribed structure. Except for a few classical results due to E. Schmidt, the results presented there are due to Golomb. One of them (in Section 7) is related to his early work: If the function $f \in L^2(\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n)$ is to be approximated by a product $u_1 \cdots u_n$ with $u_i \in L^2(\Omega_i)$ $(i = 1, 2, \cdots, n)$ so that the squared distance $\int_{\Omega_1} \cdots \int_{\Omega_n} |f - u_1 \cdots u_n|^2$ is minimized, the u_i are solutions of the nonlinear eigenvalue problem

$$\int_{\Omega_1} \cdots \int_{\Omega_{i-1}} \int_{\Omega_{i+1}} \cdots \int_{\Omega_n} f \prod_{j \neq i} u_j = \lambda u_i \qquad (i = 1, \cdots, n)$$

belonging to largest possible "eigenvalue" λ. Solutions can be found by the Newton-Kantorovich method. An appraisal of this work and its significance can be found in the detailed review by E. Stiefel [St].

In [28], the distances from an ellipsoid determined by a nonnegative definite self-adjoint operator R to manifolds determined by the associated spectral projection family were computed, and shown to satisfy an optimality criterion. It is straightforward to use these results to characterize the n-dimensional diameters of the ellipsoids in terms of reciprocal square roots of positive eigenvalues bounded away from the infimum of the essential spectrum of R. In the case that R is an elliptic operator on a smoothly bounded Euclidean domain, these results can be combined with known results for asymptotic eigenvalue estimates to yield the asymptotic distribution of the diameters. This project was the basis of the second author's dissertation, directed by Golomb.

Another major task of approximation theory is that of data fitting subject to an optimality criterion, e.g., smoothest point or spectral interpolation. We have already discussed this to a certain extent in the preceding section. We briefly mention here item [36] concerning the estimation of uniform norms of periodic functions, with certain zero Fourier coefficients, in terms of higher order derivatives. Golomb's student P. W. Smith in his dissertation [Sm] did additional work in uniform norm optimization, with auxiliary studies in L^p, which gave several insights on earlier work of J. Favard [F] and G. Glaeser [G]. A substantial amount of related activity followed the appearance of [Sm] (cf. especially the seminal paper of S. Karlin [K]).

Golomb's contributions to the area of approximation and optimization have been both synthetic and original. As a final example of his unifying observations, we cite his identification in [30] of the spline as a distribution of finite order, hence a classical solution of the standard direct Hilbert space minimization problem, whose prior existence is established by direct methods of the calculus of variations. This seemingly innocuous observation served to distinguish questions of existence and uniqueness and served to embed a part of the spline theory firmly in the calculus of variations, thus effectively giving substance and order to the various generalized spline theories developed in the 1960's. An application of this perspective is furnished in [37].

CARL DE BOOR
JOSEPH W. JEROME

REFERENCES

[A] M. ATTEIA, *Généralisation de la définition et de propriétés des "spline" fonctions*, Comptes Rendus, Acad. Sci. Paris, 260 (1965), pp. 2550–3553.

[B] C. DE BOOR, *How small can one make the derivatives of an interpolating function?* J. Approximation Theory, 13 (1975), pp. 105–116.

[Br] H. BRÉZIS, *Équations et inéquations non-linéaires dans les espaces vectoriels en dualité*, Ann. Inst. Fourier (Grenoble), 18 (1968), pp. 115–175.

[BG] F. E. BROWDER AND C. P. GUPTA, *Monotone operators and nonlinear integral equations of Hammerstein type*, Bull. Amer. Math. Soc., 75 (1969).

[CKW] L. CESARI, R. KANNAN AND H. F. WEINBERGER, eds., *Nonlinear Analysis*, Academic Press, New York, 1978.

[F] J. FAVARD, *Sur l'interpolation*, J. Math. Pures Appl., 19 (1940), pp. 281–306.

[G] G. GLAESER, *Prolongement extrémal de fonctions différentiables d'une variable*, J. Approximation Theory, 8 (1973), pp. 249–261.

[K] S. KARLIN, *Some variational problems on certain Sobolev spaces and perfect splines*, Bull. Amer. Math. Soc. 79 (1973), pp. 124–128.

[La] P. J. LAURENT, *Approximation et optimisation*, Hermann, Paris, 1972.

[Lo] G. G. LORENTZ, *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1966.

[MR] C. A. MICCHELLI AND T. RIVLIN, eds., *Optimal Estimation in Approximation Theory*, Plenum, New York, 1977.

[M] G. J. MINTY, *On some aspects of the theory of monotone operators*, in Proceedings, NATO Advanced Study Institute on Theory and Applications of Monotone Operators, Edizioni Oderisi, Gubbio, Italy, 1969.

[N] L. NIRENBERG, *Variational and topological methods in nonlinear problems*, Bull. Amer. Math. Soc., 4 (1981), pp. 267–302.

[Sa1] A. SARD, *Linear Approximation*, AMS Surveys, Vol. 9, American Mathematical Society, Providence, RI, 1963.

[Sa2] A. SARD, Best approximate integration formulas, best approximation formulas, Amer. J. of Math., 71 (1949), pp. 80–91.

[Sc] L. SCHUMAKER, *Spline Functions: Basic Theory*, Wiley-Interscience, New York, 1981.

[S] I. J. SCHOENBERG, *Cardinal Spline Interpolation*, Society for Industrial and Applied Mathematics, Philadelphia, 1973.

[Sm] P. SMITH, $W^{r,p}(R)$*-splines*, Ph.D. dissertation, Purdue University, W. Lafayette, IN, 1972.

[St] E. STIEFEL, Math. Rev., 21 (1960), #962.

[W] H. WHITNEY, *On the extension of differentiable functions*, Bull. Amer. Math. Soc., 50 (1934), pp. 76–81.

LIST OF PUBLICATIONS OF MICHAEL GOLOMB

[1] *Zur Theorie der nichtlinearen Integralgleichungen, Integralgleichungssysteme und allgemeinen Funktionalgleichungen*, Math. Z., 39, pp. 45–75 (1934).

[2] *Über Systeme von nichtlinearen Integralgleichungen*, Publications Mathématiques de l'Université de Belgrade, 5, pp. 52–83 (1936).

[3] *Electric propagation on long lines terminated by lumped networks. I*, J. Franklin Institute, 235, pp. 41–73 (1943).

[4] *Electric propagation on long lines terminated by lumped networks. II*, J. Franklin Institute, 235, pp. 101–117 (1943).

[5] *Zeros and poles of functions defined by Taylor series*, Bull. Amer. Math. Soc., 49, pp. 581–592 (1943).

[6] *The convergence of sequences of Hadamard determinants*, Duke Math. J., 11, pp. 759–777 (1944).

[7] *Elements of Ordinary Differential Equations* (with M. E. Shanks), McGraw-Hill, New York, 365 pp. (1950).

[8] *The mathematical theory of multidimensional servomechanisms*, J. Franklin Institute, 253, pp. 29–57 (1952).

[9] *Critical speeds of uniform shafts under axial torque* (with R. M. Rosenberg). Proceedings of the first U.S. National Congress of Applied Mathematics, pp. 103–110 (1952).

[10] *Theoretical Mechanics*. Lecture Notes, Dept. of Math., Purdue Univ., W. Lafayette, IN, 172 pp. (1952).

[11] *A note on linear vector spaces of mappings with positive Jacobians*, Proc. Amer. Math. Soc., 5, pp. 536–538 (1954).

[12] *On the polynomial solutions of a Riccati equation* (with J. G. Campbell), Amer. Math. Monthly, 61, pp. 402–404 (1954).
[13] *Functions which are symmetric about several points* (with P. Erdös). Nieuw Archief voor Wiskunde (1), 3, pp. 13–19 (1955).
[14] *On the initial value problem for non-normal systems of partial differential equations*, Tech. Report 1, OOR Project no. Ord-1045, 16 pp. (1955).
[15] *Systems of partial differential equations with several time-like variables* (with D. L. Phillips and R. G. Segers), Tech report 2, OOR Project no. ORD-1045, 48 pp. (1955).
[16] *Bounds for solutions of nonlinear differential systems*, Arch. Rational Mech. Anal., 1, pp. 272–282 (1958).
[17] *Approximate method for calculations using concentration-dependent diffusion coefficients* (with A. G. Guy and A. S. Yue), Trans. Amer. Inst. Met. Eng., 211, pp. 1204–1206 (1957).
[18] *Expansion and boundedness theorems for solutions of linear differential systems with periodic or almost periodic coefficients*, Arch. Rational Mech. Analysis, 2, pp. 284–308 (1958).
[19] *Optimal approximation and error bounds* (with H. F. Weinberger), Symposium on Numerical Approximation, R. Langer, ed., Univ. of Wisconsin Press, Madison, pp. 117–190 (1959).
[20] *Approximation by functions of fewer variables*, Symposium on Numerical Approximation, R. Langer, ed., Univ. of Wisconsin Press, Madison, pp. 275–327 (1959).
[21] *Solution of certain non-autonomous differential equations by series of exponential functions*, Illinois J. Math., 3, pp. 45–65 (1959).
[22] *On the uniformly best approximation of functions given by incomplete data*, MRC Report 121, Univ. of Wisconsin, Madison (1959).
[23] *On the reducibility of certain linear differential systems*, J. Reine Ange. Math., 205, pp. 171–185 (1961).
[24] *Lectures on Theory of Approximation*, Argonne National Laboratory, Applied Mathematics Division, Argonne, IL, 289 pp. (1962).
[25] *Elementary proofs for the equivalence of Fermat's Principle and Snell's Law*, Amer. Math. Monthly, 71, pp. 541–543 (1964).
[26] *Elements of Ordinary Differential Equations*, Second revised edition (with M. E. Shanks), 410 pp. (1965).
[27] *An algebraic method in differential equations*, Amer. Math. Monthly, 72, pp. 1107–1110 (1965).
[28] *Optimal approximation manifolds in L_2-spaces*, J. Math. Anal. Appl., 12, pp. 505–512 (1965).
[29] *Optimal and nearly-optimal linear approximations*, in Approximation of Functions, H. L. Garabedian, ed., Elsevier, Amsterdam, pp. 83–100 (1965).
[30] *Splines, n-widths and optimal approximations*, MRC Technical Report 784, Univ. of Wisconsin, Madison (1967).
[31] *On H^m-extension of functions and spline interpolation. I: Existence and uniqueness* (with I. J. Schoenberg). Unpublished manuscript, 11 pp. (1966).
[32] *On H^m-extension of functions and spline interpolation. II: Representation of the extremal H^m-extensions.* MRC Technical Report 807, Univ. of Wisconsin, Madison, 23 pp. (1967).
[33] *Approximation by periodic spline interpolants on uniform meshes*, J. Approximation Theory, 1, pp. 26–65 (1968).
[34] *Spline interpolation near discontinuities*, in Approximation with Special Emphasis on Spline Functions, I. J. Schoenberg, ed., Academic Press, New York, pp. 51–74 (1969).
[35] *Rates of convergence of L-splines which interpolate H_2-functions*, MRC Technical Summary Report 1048, Univ. of Wisconsin, Madison, 24 pp. (1970).
[36] *Some extremal problems for differentiable periodic functions in $L_\infty(R)$*, MRC Technical Summary Report 1069, Univ. of Wisconsin, Madison, 32 pp. (1970).
[37] *Linear differential equations with boundary conditions on arbitrary point sets* (with J. Jerome), Trans. Amer. Math. Soc., 153, pp. 235–264 (1971).
[38] *Spline approximations to the solutions of two-point boundary value problems*, MRC Technical Summary Report 1066, Univ. of Wisconsin, Madison, 49 pp. (1971).
[39] *$H^{m,p}$-extensions by $H^{m,p}$-splines*, J. Approximation Theory, 5, pp. 238–275 (1972).
[40] *The metric gradient in normed linear spaces* (with R. A. Tapia), Numer. Math., 20, pp. 115–124 (1972).
[41] *Complete orthonormal systems in pre-Hilbert spaces*, Amer. Math. Monthly, 79, pp. 263–267 (1972).
[42] *An inequality for elliptic and hyperbolic segments* (with Hiroshi Haruki), Math. Magazine, 46, pp. 152–155 (1973).
[43] *Variations on a theorem by Archimedes*, Amer. Math. Monthly, 81, pp. 138–145 (1974).
[44] Erratum to *Spline interpolation near discontinuities*, J. Approximation Theory, 10, p. 101 (1974).

[45] *Interpolation operators as optimal recovery schemes for classes of analytic functions*, in Optimal Estimation in Approximation Theory, Ch. A. Micchelli and T. J. Rivlin, eds., Plenum, New York, pp. 93–138 (1977).

[46] *Stability of interpolating elastica*, MRC Technical Summary Report 1852, Univ. of Wisconsin, Madison, 49 pp. (1978).

[47] *Equilibria of the curvature function and manifolds of nonlinear interpolating spline curves* (with J. Jerome), MRC Technical Summary Report 2024, Univ. of Wisconsin, Madison, 64 pp. (1979), also SIAM J. Math. Anal., this issue, pp. 421–458.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0001 $01.00/0

RAYLEIGH–SCHRÖDINGER PERTURBATION OF SEMIGROUPS*

HAROLD E. BENZINGER†

Abstract. A second order differential operator with abstract operator coefficients is considered in the spaces $L^p(0,1)$, $1\le p<\infty$, where the operator coefficient and the boundary conditions are analytic functions of a finite number of complex parameters. Asymptotic expressions for the eigenvalues and eigenfunctions are obtained, which are uniform over all eigenvalues. These are applied to develop a structure theory for the operators and the semigroups they generate. An application to the problem of determining the extent to which the semigroup approximates the identity operator is given.

1. Introduction. Let X denote any one of the Banach spaces $L^p(0,1)$, $1\le p<\infty$. Let $\varepsilon=(\varepsilon_0,\cdots,\varepsilon_r)$ and $\alpha=(\alpha_0,\alpha_1)$ be vectors of complex constants. Let $V=V(\varepsilon):X\to X$ be a bounded linear operator, defined for ε in a neighborhood of 0 in complex $(r+1)$-space such that

$$\|V(\varepsilon)\|\le K|\varepsilon| \tag{1.1}$$

and assume that V is analytic as a function of each ε_j, $0\le j\le r$. Also, for suitable functions $u(x)$, $0\le x\le 1$, let

$$U_ju=U_j(\alpha)u=u^{(l_j)}(j)+\alpha_ju^{(l_j-1)}(j), \tag{1.1.1}$$

where $j=0,1$, $l_j=0,1$ and $\alpha_j=0$ if $l_j=0$. We consider the second order ordinary differential operator defined by

$$Hu=H(\varepsilon,\alpha)u=u^{(2)}+V(\varepsilon)u \tag{1.2}$$

with domain of definition

$$\mathcal{D}(\alpha)=\left\{u\in C^1[0,1]:u^{(1)}\in AC[0,1],\ u^{(2)}\in X,\ U_j(\alpha)u=0,j=0,1\right\}. \tag{1.3}$$

In the special case that $\varepsilon=0$, $\alpha=0$, the operator H is a simple Sturm–Liouville operator whose eigenvalues and eigenfunctions can be computed explicitly, along with the semigroup thereby generated. Our purpose in this paper is to obtain asymptotic estimates for the eigenvalues $\lambda_k=\lambda_k(\varepsilon,\alpha)$ and eigenfunctions $u_k=u_k(x;\varepsilon,\alpha)$ of $H(\varepsilon,\alpha)$, and to use these to obtain structural relationships between the operators H and $H_0=H(0,0)$ as well as the corresponding semigroups $U(t)=U(t;\varepsilon,\alpha)$ and $U_0(t)=U(t;0,0)$. An indication of the structural results which are possible is given in [3] for $r=0$, $\varepsilon=1$, $\alpha=0$, $l_j=0$ and V of the special form $Vf(x)=q(x)f(x)$, for $q\in L^\infty(0,1)$.

Rayleigh–Schrödinger perturbation theory arises in the context of two Hamiltonian operators H,H_0 and a potential V, related by $H=H_0+\varepsilon V$, where ε is a small, nonnegative constant and the operators are acting on a Hilbert space. If λ^0 is an isolated eigenvalue of H_0, along with normalized eigenvector u^0, then it is assumed that the perturbation εV introduces perturbations in the eigenvalue and eigenvector which are analytic in ε. Then one obtains formulae

$$\lambda(\varepsilon)=\lambda^0+\varepsilon(Vu^0,u^0)+\text{“corrections to order }\varepsilon^2\text{”}, \tag{1.4}$$

$$u(\varepsilon)=u^0+\varepsilon u^1+\text{“corrections to order }\varepsilon^2\text{”}, \tag{1.5}$$

* Received by the editors November 4, 1980, and in revised form, April 23, 1981.
† Department of Mathematics, University of Illinois, Urbana, Illinois 61801.

where u^1 can be expressed in terms of V and objects related to H_0. A derivation of (1.4), (1.5) can be found in [5, pp. 686–689]. Conditions permitting rigorous justification are given in [8, pp. 10–18].

The Rayleigh–Schrödinger method is useful for determining the influence of V on a given energy level and stationary state. However, no information is provided on how the "corrections to order ε^2" vary from eigenvalue to eigenvalue. Consequently, the overall influence of the perturbation on the structure of H, as well as the time evolution of the system, cannot be determined from this method. For the differential operators considered in this paper, it is possible to estimate the error terms so as to obtain a structural relationship between H and H_0 which can be used to express the semigroup (or unitary group) generated by H in terms of the similar object generated by H_0. This can be done without resorting to time-ordered integrals as in time-dependent perturbation theory. See [5, pp. 722–728] and [7, pp. 282–292]. Additionally, the error estimates and structural properties incorporate the dependence of the domain on the vector parameter α.

Sections 2 and 3 contain asymptotic estimates for solutions of the differential equation

$$u^{(2)}+V(\varepsilon)u=-\rho^2 u,$$

as well as for the eigenvalues and eigenfunctions of H. These generalize results obtained in the classical case that V is determined by a bounded function: $Vf=qf$. In §4 it is shown that the eigenfunctions form a basis for $L^p(0,1)$, $1<p<\infty$. In §§4 and 5, with additional assumptions on V, it is shown that the eigenfunctions of H form a basis equivalent to the basis of eigenfunctions of H_0, and a structural relationship between H and H_0 is obtained. In the case that $X=L^2(0,1)$, the additional assumptions are not needed. In §6 a structural relationship between the semigroups generated by H and H_0 is obtained, and in §7 this is applied to the problem of estimating the extent to which the semigroup approximates the identity operator for small values of time.

2. Solutions of the differential equation. Our purpose is to obtain asymptotic estimates for a fundamental set of solutions of the differential equation

$$u^{(2)}+V(\varepsilon)u=-\rho^2 u, \tag{2.1}$$

where ρ is a complex parameter. In the "classical" case that $(Vf)(x)=q(x)f(x)$ for some $q\in L^\infty(0,1)$, the asymptotic estimates are obtained for a half-plane $\operatorname{Im}\rho>-K$ (some $K>0$). We shall see that for general V asymptotic estimates are obtained only in a strip $|\operatorname{Im}\rho|<K$. Because of the theorem given below, this is adequate. If $\lambda=-\rho^2$ is an eigenvalue of a linear operator, we shall refer to ρ as an eigenvalue also, relying on the context to avoid confusion. A linear operator is *discrete* if it has compact resolvent, and consequently has a discrete spectrum consisting only of eigenvalues and the origin.

THEOREM 2.2. *Let $\delta,\tilde{\varepsilon},\tilde{\alpha},K$, be preassigned positive constants, with $|\varepsilon|<\tilde{\varepsilon}$, $|\alpha|<\alpha$. Then $H(\varepsilon,\alpha)$ is a discrete operator whose eigenvalues $\rho_k(\varepsilon,\alpha)$ are ultimately in the strip $S:|\operatorname{Im}\rho|<K$.*

Proof. $H(0,\alpha)$ is a classical Sturm–Liouville operator with compact resolvent represented by a Green function:

$$R(\lambda;0,\alpha)f=-\int_0^1 G(x,t,\rho;0,\alpha)f(t)\,dt.$$

Since the eigenvalues of $H(0,\alpha)$ are asymptotically equal to those of $H(0,0)$, for $|\rho|$ sufficiently large and $|\rho-\rho_k(0,0)|\geq\delta$, ρ is in the resolvent set of $H(0,\alpha)$. Then there

exists a constant $K(\delta)>0$ such that

$$|G(x,t,\rho;0,\alpha)|\leq\frac{K(\delta)}{|\rho|}$$

[6, pp. 73–74]. Consequently,

$$\|R(\lambda;0,\alpha)\|\leq\frac{K(\delta)}{|\rho|}. \tag{2.3}$$

Since $H(\varepsilon,\alpha)=H(0,\alpha)+V(\varepsilon)$, we have

$$R(\lambda;\varepsilon,\alpha)=[I-R(\lambda;0,\alpha)V(\varepsilon)]^{-1}R(\lambda;0,\alpha), \tag{2.3′}$$

provided λ is in the resolvent set of $H(0,\alpha)$ and $\|R(\lambda;0,\alpha)V(\varepsilon)\|<1$. Using (1.1) and (2.3), it suffices to have $|\rho-\rho_k(0,0)|\geq\delta$ and $|\rho|>KK(\delta)\tilde{\varepsilon}$. Thus $R(\lambda;\varepsilon,\alpha)$ is a compact operator, and the eigenvalues of $H(\varepsilon,\alpha)$ ultimately lie in the disks $|\rho-\rho_k(0,0)|<\delta$. In particular, since $\delta>0$ was arbitrary, the eigenvalues ultimately lie in any preassigned strip $|\mathrm{Im}\,\rho|<K$.

To find a fundamental set of solutions to (2.1), we proceed as in the classical case and convert (2.1) into a first order system

$$\phi^{(1)}=\begin{bmatrix}0 & 1\\ -\rho^2-V(\varepsilon) & 0\end{bmatrix}\phi,\qquad \phi=\begin{bmatrix}\phi_1\\ \phi_2\end{bmatrix},$$

being careful to treat V as an operator rather than as a scalar. Note that the success of this procedure implies that the notion of a fundamental set is in fact valid for these operator differential equations.

For $\rho\neq0$, let

$$T(\rho)=\begin{bmatrix}1 & 1\\ i\rho & -i\rho\end{bmatrix},\qquad T^{-1}(\rho)=\frac{1}{2i\rho}\begin{bmatrix}i\rho & 1\\ i\rho & -1\end{bmatrix}.$$

$T(\rho)$ generates the similarity transform which diagonalizes $\left[\begin{smallmatrix}0 & 1\\ -\rho^2 & 0\end{smallmatrix}\right]$. Introducing a new dependent variable y through $\phi=T(\rho)y$, the differential equation for y is

$$y^{(1)}=\left[i\rho\Omega+\frac{1}{2i\rho}A_0V\right]y, \tag{2.4}$$

where

$$\Omega=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix},\qquad A_0=\begin{bmatrix}-1 & -1\\ 1 & 1\end{bmatrix}.$$

Let

$$E(x,\rho)=\begin{bmatrix}e^{i\rho x} & 0\\ 0 & e^{-i\rho x}\end{bmatrix}=e^{i\rho\Omega x}.$$

If the matrix integral equation

$$Y(x,\rho,\varepsilon)=E(x,\rho)+\frac{1}{2i\rho}\int^x E(x-t,\rho)A_0V(Y(t,\rho,\varepsilon))\,dt$$

has an invertible solution Y, then Y is a fundamental matrix for (2.4). We make the further substitution $Y(x,\rho,\varepsilon)=Z(x,\rho,\varepsilon)E(x,\rho)$, obtaining for Z the integral equation

$$Z(x,\rho,\varepsilon)=I+\frac{1}{2i\rho}\int^x E(x-t,\rho)A_0V(Z(t,\rho,\varepsilon)E(t,\rho))E^{-1}(x,\rho)\,dt. \tag{2.5}$$

It is this equation which we actually solve. Let $\omega_1=i$, $\omega_2=-i$, and let P_{kj} denote the 2×2 matrix with 1 in the k,j position and zeros elsewhere. Then

$$E(x,\rho)=\sum_{k=1}^{2} e^{\rho\omega_k x}P_{kk},$$

and (2.5) becomes

$$Z(x,\rho,\varepsilon)=I+\frac{1}{2i\rho}\sum_{k=1}^{2}\sum_{j=1}^{2}\sum_{l=1}^{2}\int_{x_{kjl}}^{x} e^{\rho\omega_k(x-t)}P_{kk}A_0V\big(Z(t,\rho,\varepsilon)e^{\rho\omega_l t}P_{ll}\big)e^{-\rho\omega_j x}P_{jj}\,dt.$$

Since P_{ll} is a scalar matrix, it can be removed from the parentheses. Using $P_{ll}P_{jj}=\delta_{lj}P_{jj}$, we have

$$Z(x,\rho,\varepsilon)=I+\frac{1}{2i\rho}\sum_{k=1}^{2}\sum_{j=1}^{2}\int_{x_{kj}}^{x} e^{\rho(\omega_k-\omega_j)(x-t)}P_{kk}A_0\big[e^{-\rho\omega_j t}V\big(e^{\rho\omega_j t}Z(t,\rho,\varepsilon)\big)\big]P_{jj}\,dt.$$

For any matrix $M(t)$ with components in $L^p(0,1)$, let

$$\|M\|_p=\max\|M_{kj}\|,\qquad 1\le k,j\le 2.$$

Then the linear map $M\to A_0[e^{-\rho\omega_j t}V(e^{\rho\omega_j t}M(t))]$ has norm bounded by

$$2e^{|\mathrm{Im}\,\rho|}\|V\|\le 2e^{|\mathrm{Im}\,\rho|}K|\varepsilon|,$$

where we have used (1.1). Using K as a generic constant, we see that, in any strip $|\mathrm{Im}\,\rho|\le K$,

$$\|A_0\big[e^{-\rho\omega_j t}V\big(e^{\rho\omega_j t}M(t)\big)\big]\|\le K|\varepsilon|\,\|M\|.$$

Let $W_j(z)=e^{-\rho\omega_j t}V(e^{\rho\omega_j t}Z)$. Then with the choice

$$x_{11}=x_{12}=0,\qquad x_{21}=x_{22}=1,$$

we have

$$Z(x,\rho,\varepsilon)=I+\frac{1}{2i\rho}H(Z),$$

where H is the linear operator defined by

$$H(Z)=\begin{bmatrix}\displaystyle\int_0^x (A_0W_1(Z))_{11}\,dt & \displaystyle\int_0^x e^{2i\rho(x-t)}(A_0W_2(Z))_{12}\,dt\\ \displaystyle-\int_x^1 e^{-2i\rho(x-t)}(A_0W_1(Z))_{21}\,dt & \displaystyle-\int_x^1 (A_0W_2(Z))_{22}\,dt\end{bmatrix}.$$

For $|\mathrm{Im}\,\rho|\le K$, we have $\|H\|\le K|\varepsilon|$, uniformly in ρ. For $|\rho|$ sufficiently large, $\frac{1}{2i\rho}H$ is a contraction mapping on matrix L^p, uniformly in ε, $|\varepsilon|\le\tilde{\varepsilon}$, as well as a contraction mapping of L^p into L^∞.

THEOREM 2.6. *For $|\varepsilon|\le\tilde{\varepsilon}$, $|\mathrm{Im}\,\rho|\le K$, $|\rho|$ sufficiently large,* (2.5) *has an absolutely continuous, nonsingular solution such that*

$$Z(x,\rho,\varepsilon)=I+\frac{1}{2i\rho}H(I)+O\left(\left(\frac{|\varepsilon|}{\rho}\right)^2\right), \tag{2.7}$$

uniformly in x, $0\le x\le 1$, and ε, $|\varepsilon|\le\tilde{\varepsilon}$, as $|\rho|\to\infty$. Additionally, Z is an analytic function of ε for fixed x,ρ.

Proof. Since $\frac{1}{2i\rho}H$ is a contraction of L^p into L^∞, the successive approximations $Z_0=I$,

$$Z_{n+1}=I+\frac{1}{2i\rho}H(Z_n),$$

are uniformly convergent to a continuous function $Z(x,\rho,\varepsilon)$, which is a solution of (2.5) and therefore absolutely continuous. Relation (2.7) is an immediate consequence of

$$Z(x,\rho,\varepsilon)=\sum_{l=0}^{\infty}\left(\frac{1}{2i\rho}H\right)^l I.$$

Since the successive approximations are analytic and the convergence is uniform, Z is analytic in ε.

COROLLARY 2.8. *For ε,ρ as above,* (2.1) *has a fundamental matrix*

$$\Phi(x,\rho,\varepsilon)=T(\rho)\left[I+\frac{1}{2i\rho}H(I)+O\left((|\varepsilon|/\rho)^2\right)\right]E(x,\rho).$$

More precise information can be obtained by evaluating $H(I)$. Let $e_j(t,\rho,\varepsilon)=e^{-\rho\omega_j t}V(\varepsilon)(e^{\rho\omega_j t})$. Then

$$H(I)=-\begin{bmatrix}\int_0^x e_1(t,\rho,\varepsilon)\,dt & \int_0^x e^{2i\rho(x-t)}e_2(t,\rho,\varepsilon)\,dt\\ \int_x^1 e^{-2i\rho(x-t)}e_1(t,\rho,\varepsilon)\,dt & \int_x^1 e_2(t,\rho,\varepsilon)\,dt\end{bmatrix}$$

$$=-\begin{bmatrix}h_{11}(x,\rho,\varepsilon) & h_{12}(x,\rho,\varepsilon)\\ h_{21}(x,\rho,\varepsilon) & h_{22}(x,\rho,\varepsilon)\end{bmatrix}.$$

THEOREM 2.9. *For ε,ρ as above,* (2.1) *has a fundamental matrix $\Phi(x,\rho,\varepsilon)$ of the form*

$$\Phi(x,\rho,\varepsilon)=\begin{bmatrix}\left[1-\frac{1}{2i\rho}(h_{11}+h_{21})+b\right]e^{i\rho x} & \left[1-\frac{1}{2i\rho}(h_{22}+h_{12})+b\right]e^{-i\rho x}\\ i\rho\left[1-\frac{1}{2i\rho}(h_{11}-h_{21})+b\right]e^{i\rho x} & -i\rho\left[1-\frac{1}{2i\rho}(h_{22}-h_{12})+b\right]e^{-i\rho x}\end{bmatrix},$$

where $b=b(x,\rho,\varepsilon)$ is generic notation for a function which is $O((|\varepsilon|/\rho)^2)$.

COROLLARY 2.10. *For ε,ρ as above,* (2.1) *has two linearly independent solutions $u_k(x,\rho,\varepsilon)$, $k=1,2$, which, along with their derivatives $u_k^{(j)}$, $j=0,1$, are analytic in ρ,ε and are expressed as*

$$u_k^{(j)}(x,\rho,\varepsilon)=(\rho\omega_k)^j\left[1-\frac{1}{2i\rho}H_{kj}(x,\rho,\varepsilon)+b(x,\rho,\varepsilon)\right]e^{\rho\omega_k x},$$

where H_{kj} are linear combinations of the entries in the matrix $H(I)$.

Proof. The functions $u_k^{(j)}$ are the entries in the matrix $\Phi(x,\rho,\varepsilon)$. Note that $|H_{kj}(x,\rho,\varepsilon)|\le K|\varepsilon|$, uniformly in x,ρ.

3. Eigenvalues and eigenfunctions of $H(\varepsilon,\alpha)$. Using the fundamental set $u_k(x,\rho,\varepsilon)$, $k=1,2$ as above, let

$$u(x,\rho,\varepsilon,\alpha)=U_0(u_2)u_1-U_0(u_1)u_2. \tag{3.1}$$

This solution of (2.1) satisfies the boundary condition at $x=0$. Thus the eigenvalues are those values of $\rho_k=\rho_k(\varepsilon,\alpha)$ satisfying

$$U_1(u)=0, \tag{3.2}$$

and the eigenfunctions are $u(x,\rho_k,\varepsilon,\alpha)$.

To lighten the notation, in (1.1.1) let $l=l_0$, $m=l_1$. Also let $\beta_j=\alpha_j/|\alpha|$ ($\beta_j=0$ if $|\alpha|=0$).

LEMMA 3.3. *For $k=1,2$,*

$$U_0(u_k)=(\rho\omega_k)^l\left[1-\frac{1}{2i\rho}H_{kl}(0,\rho,\varepsilon)+\frac{|\alpha|}{\rho\omega_k}\beta_0+C(\rho,\varepsilon,\alpha)\right],$$

$$U_1(u_k)=(\rho\omega_k)^m e^{\rho\omega_k}\left[1-\frac{1}{2i\rho}H_{km}(1,\rho,\varepsilon)+\frac{|\alpha|}{\rho\omega_k}\beta_1+C(\rho,\varepsilon,\alpha)\right],$$

where $C(\rho,\varepsilon,\alpha)$ is generic notation for a function which is $O((|\varepsilon|/\rho)^2)+O((|\alpha|/\rho)^2)$, uniformly in ρ in the strip, $|\varepsilon|\le\tilde{\varepsilon}$, $|\alpha|\le\tilde{\alpha}$.

Proof. Consider U_0. Using (1.1.1) and the estimates in (2.10), we have

$$U_0(u_k)=(\rho\omega_k)^l\left\{\left[1-\frac{1}{2i\rho}H_{kl}(0,\rho,\varepsilon)+b(\rho,\varepsilon)\right]\right.$$
$$\left.+\frac{|\alpha|}{\rho\omega_k}\beta_0\left[1-\frac{1}{2i\rho}H_{kl-1}(0,\rho,\varepsilon)+b(\rho,\varepsilon)\right]\right\}.$$

Now

$$\left|\frac{|\alpha|}{\rho\omega_k}\beta_0\frac{1}{2i\rho}H_{kl-1}(0,\rho,\varepsilon)\right|\le K\frac{|\alpha|}{|\rho|}\frac{|\varepsilon|}{|\rho|}\le K'\left|\left(\frac{|\varepsilon|}{|\rho|}\right)^2+\left(\frac{|\alpha|}{|\rho|}\right)^2\right|.$$

Thus

$$U_0(u_k)=(\rho\omega_k)^l\left[1-\frac{1}{2i\rho}H_{kl}(0,\rho,\varepsilon)+\frac{|\alpha|}{\rho\omega_k}\beta_0+C(\rho,\varepsilon,\alpha)\right].$$

The proof for U_1 is similar.

Let $\kappa\equiv l+m \pmod 2$. Using (3.3) a direct computation shows that

$$U_1(u)=(-1)^m(i\rho)^{l+m}e^{-i\rho}$$
$$\times\left\{\left\{1-\frac{1}{2i\rho}[H_{1l}(0,\rho,\varepsilon)+H_{2m}(1,\rho,\varepsilon)]+\frac{|\alpha|}{i\rho}[\beta_0-\beta_1]+C(\rho,\varepsilon,\alpha)\right\}\right.$$
$$\left.-(-1)^\kappa e^{2i\rho}\left\{1-\frac{1}{2i\rho}[H_{2l}(0,\rho,\varepsilon)+H_{1m}(1,\rho,\varepsilon)]-\frac{|\alpha|}{i\rho}[\beta_0-\beta_1]+C(\rho,\varepsilon,\alpha)\right\}\right\}.$$

Thus the zeros of $U_1(u)$ are characterized by

$$e^{2i\rho-\pi i\kappa}=\frac{1-\frac{1}{2i\rho}[H_{1l}(0,\rho,\varepsilon)+H_{2m}(1,\rho,\varepsilon)]+(|\alpha|/i\rho)[\beta_0-\beta_1]+C(\rho,\varepsilon,\alpha)}{1-\frac{1}{2i\rho}[H_{2l}(0,\rho,\varepsilon)+H_{1m}(1,\rho,\varepsilon)]-(|\alpha|/i\rho)[\beta_0-\beta_1]+C(\rho,\varepsilon,\alpha)}.$$

With $|\varepsilon|$ and $|\alpha|$ bounded, we can assume $|\rho|$ is sufficiently large, so that we can use the geometric series. Then

$$e^{2i\rho-\pi i\kappa}=1+\frac{1}{2i\rho}\Lambda(\rho,\varepsilon)+\frac{2|\alpha|}{i\rho}\beta+C(\rho,\varepsilon,\alpha), \tag{3.4}$$

where

$$-\Lambda(\rho,\varepsilon)=H_{1l}(0,\rho,\varepsilon)+H_{2m}(1,\rho,\varepsilon)-H_{2l}(0,\rho,\varepsilon)-H_{1m}(1,\rho,\varepsilon),$$
$$\beta=\beta_0-\beta_1.$$

We introduce the further notation

$$\rho_k^0=\rho_k(0,0) \qquad \left(=k\pi+\frac{\pi}{2}\kappa\right),$$
$$\rho_k=\rho_k(\varepsilon,\alpha),$$
$$\Lambda_k(\varepsilon)=\Lambda(\rho_k^0,\varepsilon),$$
$$C_k(\varepsilon,\alpha)=C(\rho_k,\varepsilon,\alpha).$$

THEOREM 3.5. *The zeros* ρ_k *of* (3.4) *satisfy*

$$\rho_k=\rho_k^0-\frac{1}{4\rho_k^0}\Lambda_k(\varepsilon)-\frac{|\alpha|}{\rho_k^0}\beta+C_k(\varepsilon,\alpha).$$

Proof. Again using the assumption that $|\rho|$ is sufficiently large, we can use the power series expansion of the logarithm on the right of (3.4) to obtain

$$2i\rho_k-\pi i\kappa=2k\pi i+\frac{1}{2i\rho_k}\Lambda(\rho_k,\varepsilon)+\frac{2|\alpha|}{i\rho_k}\beta+C(\rho_k,\varepsilon,\alpha). \tag{3.6}$$

It suffices then to show that ρ_k can be replaced by ρ_k^0 on the right, with errors which can be incorporated in C. From (3.8) we have

$$\rho_k=\rho_k^0+O\left(\frac{|\varepsilon|}{k}\right)+O\left(\frac{|\alpha|}{k}\right).$$

Thus,

$$\frac{1}{\rho_k}=\frac{1}{\rho_k^0}+O\left(\frac{|\varepsilon|}{(\rho_k^0)^2}\right)+O\left(\frac{|\alpha|}{(\rho_k^0)^2}\right).$$

Since the analytic functions $H_{kj}(x,\rho,\varepsilon)$ are analytic with respect to ρ, uniformly with respect to x,ε, in the strip $|\mathrm{Im}\,\rho|\le K$, all derivatives with respect to ρ are uniformly bounded, perhaps in a slightly smaller strip. Since $\Lambda(\rho,\varepsilon)$ is a linear combination of the H_{kj}'s, we see that

$$|\Lambda(\rho_k,\varepsilon)-\Lambda(\rho_k^0,\varepsilon)|\le K|\rho_k-\rho_k^0|$$

and

$$\frac{1}{\rho_k}\Lambda(\rho_k,\varepsilon)=\frac{1}{\rho_k^0}\Lambda(\rho_k^0,\varepsilon)+C_k(\varepsilon,\alpha),$$
$$\frac{2|\alpha|}{i\rho_k}\beta=\frac{2|\alpha|}{i\rho_k^0}\beta+C_k(\varepsilon,\alpha).$$

COROLLARY 3.7. *With* $\lambda=-\rho^2$,

$$\lambda_k=\lambda_k^0+\frac{1}{2}\Lambda_k(\varepsilon)+2|\alpha|\beta+O\left(\frac{|\varepsilon|^2}{k}\right)+O\left(\frac{|\alpha|^2}{k}\right).$$

The expressions for the eigenvalues, as well as for the eigenfunctions, can be made more enlightening by introducing the notation

$$T_l(z)=\sqrt{2}\,\frac{e^{iz}-(-1)^l e^{-iz}}{2i(-i)^l}. \tag{3.8}$$

It is then a matter of direct computation to verify that

$$\Lambda_k(\varepsilon)=2\int_0^1 T_l(\rho_k^0 t)V(\varepsilon)\big(T_l(\rho_k^0 s)\big)\,dt+O\left(\frac{|\varepsilon|^2}{k}\right). \tag{3.9}$$

To obtain expressions for the eigenfunctions, we use (3.1) along with (2.10) and (3.3). Performing these computations, multiplying by the inessential factor $(\frac{i}{\rho})^l$, and thereby redefining $u(x,\rho,\varepsilon,\alpha)$, we have

$$\begin{aligned} u(x,\rho,\varepsilon,\alpha)=&\left[e^{i\rho x}-(-1)^l e^{-i\rho x}\right]\\ &-\frac{1}{2i\rho}\left\{\left[H_{2l}(0,\rho,\varepsilon)+H_{10}(x,\rho,\varepsilon)\right]e^{i\rho x}\right.\\ &\qquad\left.-(-1)^l\left[H_{1l}(0,\rho,\varepsilon)+H_{20}(x,\rho,\varepsilon)\right]e^{-i\rho x}\right\}\\ &-\frac{|\alpha|}{i\rho}\beta_0\left[e^{i\rho x}+(-1)^l e^{-i\rho x}\right]+C(x,\rho,\varepsilon,\alpha). \end{aligned} \tag{3.10}$$

Again redefining u, after multiplying (3.10) by $\sqrt{2}\,/2i(-i)^l$, we have

$$u(x,\rho,\varepsilon,\alpha)=T_l(\rho x)+\frac{1}{\rho}Y_l(x,\rho,\varepsilon)+\frac{|\alpha|\beta_0}{\rho}T_{l+1}(\rho x)+C(x,\rho,\varepsilon,\alpha). \tag{3.11}$$

Performing another lengthy computation, we have

$$\begin{aligned} Y_l(x,\rho,\varepsilon)=&T_{l+1}(\rho x)\int_0^x T_l(\rho t)V(T_l(\rho s))\,dt\\ &-T_l(\rho x)\int_0^x T_{l+1}(\rho t)V(T_l(\rho s))\,dt+iT_l(\rho x)\int_0^1 e^{i\rho t}V(e^{-i\rho s})\,dt. \end{aligned} \tag{3.12}$$

The boundary value problem adjoint to (1.2), (1.3), on the space X^*, is obtained by replacing $V(\varepsilon)$ with $(V(\varepsilon))^*$ and α_j with $\bar{\alpha}_j$. Note that the parameters l,m are unchanged. Thus, corresponding to (3.11), a solution of

$$v^{(2)}+(V(\varepsilon))^*v=-\sigma^2 v, \tag{3.13}$$

satisfying the adjoint boundary condition at $x=0$, is

$$v(x,\sigma,\varepsilon,\alpha)=T_l(\sigma x)+\frac{1}{\sigma}Y_l^+(x,\sigma,\varepsilon)+\frac{|\alpha|\bar{\beta}_0}{\sigma}T_{l+1}(x)+C(x,\sigma,\varepsilon,\alpha), \tag{3.14}$$

where Y_l^+ is obtained from (3.12) by replacing $V(\varepsilon)$ with $(V(\varepsilon))^*$.

Let $E_k(x,\varepsilon,\alpha)$ denote any function which is

$$O\left(\frac{|\varepsilon|}{k^2}\right)+O\left(\frac{|\alpha|}{k^2}\right),$$

uniformly in x,ε,α.

THEOREM 3.15. *A normalized biorthogonal system of eigenfunctions for H and H^* is given by*

$$u_k(x,\varepsilon,\alpha)-T_l(\rho_k^0 x)=\frac{1}{\rho_k^0}\left[A_k(x,\varepsilon)T_l(\rho_k^0 x)+B_k(x,\varepsilon,\alpha)T_{l+1}(\rho_k^0 x)\right]+E_k(x,\varepsilon,\alpha),$$

$$v_k(x,\varepsilon,\alpha)-T_l(\rho_k^0 x)=\frac{1}{\rho_k^0}\left[C_k(x,\varepsilon)T_l(\rho_k^0 x)+D_k(x,\varepsilon,\alpha)T_{l+1}(\rho_k^0 x)\right]+E_k(x,\varepsilon,\alpha),$$

where

$$A_k(x,\varepsilon)=i\int_0^1 e^{i\rho_k^0 t}V\left(e^{-i\rho_k^0 s}\right)dt-\int_0^x T_{l+1}(\rho_x^0 t)V\left(t_l(\rho_k^0 s)\right)dt$$
$$+\frac{1}{2}\int_0^1(1-t)T_{l+1}(\rho_k^0 t)\left[V\left(T_l(\rho_k^0 s)\right)+V^*\left(T_l(\rho_k^0 s)\right)\right]dt,$$

$$B_k(x,\varepsilon,\alpha)=\int_0^x T_l(\rho_k^0 t)V\left(T_l(\rho_k^0 s)\right)dt-\frac{x}{2}\int_0^1 T_l(\rho_k^0 t)V\left(T_l(\rho_k^0 s)\right)dt+|\alpha|(\beta_0-\beta x),$$

$$C_k(x,\varepsilon,\alpha)=i\int_0^1 e^{i\rho_k^0 t}V^*\left(e^{-i\rho_k^0 s}\right)dt-\int_0^x T_{l+1}(\rho_k^0 t)V^*\left(T_l(\rho_k^0 s)\right)dt$$
$$+\frac{1}{2}\int_0^1(1-t)T_{l+1}(\rho_k^0 t)\left[\overline{V\left(T_l(\rho_k^0 s)\right)}+\overline{V^*\left(T_l(\rho_k^0 s)\right)}\right]dt,$$

$$D_k(x,\varepsilon,\alpha)=\int_0^x T_l(\rho_k^0 t)V^*\left(T_l(\rho_k^0 s)\right)dt-\frac{x}{2}\int_0^1 T_l(\rho_k^0 t)V^*\left(T_l(\rho_k^0 s)\right)dt+|\alpha|(\bar\beta_0-\bar\beta x).$$

Proof. Eigenfunctions for H are obtained from (3.11) with $\rho=\rho_k$. We want to replace ρ_k with ρ_k^0. Using the identity

$$T_l(z+u)=T_l(z)\cos u+T_{l+1}(z)\sin u,$$

we have

$$T_l(\rho_k x)=T_l(\rho_k^0 x)-\left[\frac{x\Lambda_k(\varepsilon)}{4\rho_k^0}+\frac{x|\alpha|\beta}{\rho_k^0}\right]T_{l+1}(\rho_k^0 x)+C_k(x,\varepsilon,\alpha).$$

For the remaining terms on the right of (3.11), the presence of the factor $\frac{1}{\rho_k}$ means that directly replacing ρ_k with ρ_k^0 introduces errors which can be incorporated into C_k. Similar considerations hold for (3.14). Using (3.9) for $\Lambda_k(\varepsilon)$, we have

$$\begin{aligned}
&u(x,\rho_k,\varepsilon,\alpha)-T_l(\rho_k^0 x)\\
&\quad=\frac{1}{\rho_k^0}\left[i\int_0^1 e^{i\rho_k^0 t}V\left(e^{-i\rho_k^0 s}\right)dt-\int_0^x T_{l+1}(\rho_k^0 t)V\left(T_l(\rho_k^0 s)\right)dt\right]T_l(\rho_k^0 x)\\
&\quad+\frac{1}{\rho_k^0}\left[\int_0^x T_l(\rho_k^0 t)V\left(T_l(\rho_k^0 s)\right)dt\right.\\
&\qquad\left.-\frac{x}{2}\int_0^1 T_l(\rho_k^0 t)V\left(T_l(\rho_k^0 s)\right)dt+|\alpha|(\beta_0-\beta x)\right]T_{l+1}(\rho_k^0 x)\\
&\quad+C_k(x,\varepsilon,\alpha),
\end{aligned}\tag{3.16}$$

along with a similar expression for $v(x,\bar\rho_k,\varepsilon,\bar\alpha)$. Since u,v come from adjoint problems, we have

$$\int_0^1 u(x,\rho_k,\varepsilon,\alpha)\bar v(x,\bar\rho_k,\varepsilon,\bar\alpha)\,dx=0,\qquad k\neq j.$$

It remains to scale these functions so that $(u,v)=1$ if $k=j$. We note that the leading terms $T_l(\rho_k^0 x)$ are already normalized:

$$\int_0^1 T_l^2(\rho_k^0 x)\,dx=1.$$

Using (3.16) and the corresponding expression for v, we have

$$\int_0^1 u(x,\rho_k,\varepsilon,\alpha)\bar{v}(x,\bar{\rho}_k,\varepsilon,\bar{\alpha})\,dx$$

$$=1-\frac{1}{\rho_k^0}\int_0^1\int_0^x T_{l+1}(\rho_k^0 t)\left[V\left(T_l(\rho_k^0 s)\right)+V^*\left(T_l(\rho_k^0 s)\right)\right]dt\,dx+E_k.$$

Thus, multiplying u by

$$1+\frac{1}{2\rho_k^0}\int_0^1(1-t)T_{l+1}(\rho_k^0 t)\left[V\left(T_l(\rho_k^0 s)\right)+V^*\left(T_l(\rho_k^0 s)\right)\right]dt+E_k$$

and multiplying v by the conjugate, we have u_k, v_k as given in the statement of the theorem.

In carrying out this computation, one must use the relation

$$\int_0^1 e^{i\rho_k^0 t}V\left(e^{-i\rho_k^0 s}\right)dt-\int_0^1 e^{-i\rho_k^0 t}\overline{V^*\left(e^{-i\rho_k^0 s}\right)}\,dt=0.$$

This holds since the left side is

$$\left(V\left(e^{-i\rho_k^0 s}\right),e^{-i\rho_k^0 t}\right)-\left(e^{-i\rho_k^0 t},V^*\left(e^{-i\rho_k^0 t}\right)\right),$$

which is zero by the property of V^*.

4. Basis properties of the eigenfunctions.

THEOREM 4.1. *The eigenfunctions of $H(\varepsilon,\alpha)$ are complete in $L^p(0,1)$, $1\le p<\infty$.*

Proof. We shall show that the hypotheses of [2, Theorem 1.1] are satisfied. By Theorem 2.2, $H=H(\varepsilon,\alpha)$ is a discrete operator. Assume $|\rho|$ is sufficiently large, and ρ is bounded away from the eigenvalues of the classical operator $H(0,\alpha)$. Then from (2.3) and (2.3.1) we see that $\|R(\lambda;\varepsilon,\alpha)\|$ also satisfies (2.3). Thus [2, Theorem 1.1] applies.

Let $\tilde{H}=H(0,\alpha)$ and let $\{\tilde{u}_k\}$, $\{\tilde{v}_k\}$ denote a normalized system of eigenfunctions for $\tilde{H}$ and $\tilde{H}^*$. The partial sum operators for H and $\tilde{H}$ are

$$\mathcal{S}_N f=\sum_1^N (f,v_k)u_k,\qquad \tilde{S}_N f=\sum_1^N (f,\tilde{v}_k)\tilde{u}_k.$$

LEMMA 4.2. *In each space $L^p(0,1)$, $1<p<\infty$,*

$$\|\mathcal{S}_N-\tilde{S}_N\|\le M$$

for some M independent of N.

Proof. Let $\{C_N\}$ be a family of circles centered at the origin of the λ-plane, enclosing the first N eigenvalues of H and $\tilde{H}$, and uniformly bounded away from both sets of eigenvalues. Then

$$\left(\mathcal{S}_N-\tilde{S}_N\right)f=\frac{1}{2\pi i}\int_{C_N}\left[R(\lambda;\varepsilon,\alpha)-R(\lambda;0,\alpha)\right]f\,d\lambda.$$

Using (2.3.1) gives

$$R(\lambda;\varepsilon,\alpha)-R(\lambda;0,\alpha)=\left(\sum_1^\infty (R(\lambda;0,\alpha)V(\varepsilon))^k\right)R(\lambda;0,\alpha),$$

provided $\|R(\lambda;0,\alpha)V(\varepsilon)\|<1$. From (1.1) and (2.3), we can have $\|R(\lambda;0,\alpha)V(\varepsilon)\|\le\frac{1}{2}$ for N sufficiently large, as well as

$$\|R(\lambda;0,\alpha)V(\varepsilon)\|<\frac{k}{|\rho|}\lambda \quad \text{on } C_N.$$

Thus we have

$$\|\mathcal{R}(\lambda;\varepsilon,\alpha)-R(\lambda;0,\alpha)\|\le\left(\sum_1^\infty\left(\frac{K}{|\rho|}\right)^k\right)\frac{K}{|\rho|}=\frac{1}{1-\frac{K}{|\rho|}}\left(\frac{K}{|\rho|}\right)^2\le 2\left(\frac{K}{|\rho|}\right)^2=\frac{M}{|\lambda|}.$$

Thus

$$\|(\mathcal{S}_N-S_N)f\|\le M\|f\|.$$

THEOREM 4.3. *The eigenfunctions of H form a Schauder basis for $L^p(0,1)$, $1<p<\infty$.*

Proof. A complete eigenfunction system is a basis if and only if its partial sum operators are uniformly bounded. Since the eigenfunctions of the classical operator $\tilde{H}$ are a basis for $1<p<\infty$ [1], there is a uniform bound for the family $\{\tilde{S}_N\}$, $1<p<\infty$. From Lemma 4.2, there is a uniform bound for $\{\mathcal{S}_N\}$, $1<p<\infty$.

DEFINITION 4.4. Two bases $\{u_k\}$ and $\{u_k^0\}$ of a Banach space X are *equivalent* if there is a bicontinuous linear map $B: X\to X$ such that

$$Bu_k^0=u_k.$$

If $\{u_k\}$, $\{u_k^0\}$ are equivalent, then the infinite series

$$\sum_1^\infty (f,v_k^0)u_k=f+Lf=f+\sum_1^\infty (f,v_k^0)(u_k-u_k^0)=Bf, \tag{4.5}$$

$$\sum_1^\infty (f,v_k)u_k^0=f+Jf=f+\sum_1^\infty (f,v_k-v_k^0)u_k^0=B^{-1}f \tag{4.6}$$

converge and define bounded linear operators on X, where $\{v_k\}$, $\{v_k^0\}$ are the dual bases. Conversely, if the series in (4.5), (4.6) converge for all f in X, they represent closed and therefore bounded linear operators which are necessarily inverses of each other.

Let $\{u_k\}$ denote the eigenfunctions of $H(\varepsilon,\alpha)$, and $\{u_k^0\}$ the eigenfunctions of $H(0,0)$, with $\{v_k\}$, $\{u_k^0\}$ for the adjoint problems.

Assumption 4.7. There exists a constant $M>0$ such that, for all f in $L^\infty(0,1)$, $|V^*f(x)|\le M\|f\|_\infty$.

Remark 4.8. Assumption 4.7 is satisfied for operators of the form

$$V(\varepsilon)f=q(x,\varepsilon)f(x),$$

where $q(\cdot,\varepsilon)$ is in L^∞ and is analytic in ε, $q(\cdot,0)=0$, and for

$$V(\varepsilon)f=\int_0^1 K(x,t,\varepsilon)f(t)\,dt,$$

where K is bounded uniformly in all variables and analytic in ε.

THEOREM 4.9. *The eigenfunction systems* $\{u_k\}$, $\{u_k^0\}$ *are equivalent bases in* $X = L^p(0,1)$ *provided either*

(i) $1<p<\infty$ *and* V^*k *satisfies* (4.7) *or*

(ii) $p=2$ *and* V *is just bounded on* L^2.

Proof. Using (3.15) in (4.5), we consider the series

$$\sum_1^\infty \frac{1}{\rho_k^0} A_k(x,\varepsilon)\big(f,v_k^0\big)u_k^0(x)+\sum_1^\infty \frac{1}{\rho_k^0} B_k(x,\varepsilon,\alpha)\big(f,v_k^0\big)T_{l+1}\big(\rho_k^0 x\big) + \sum_1^\infty E_k(x,\varepsilon,\alpha)\big(f,v_k^0\big).$$

We show that each of these terms represents a bounded linear operator on X. Since $E_k = O(k^{-2})$, this is the case for the third term above. Consider the second term. If $1<p\le 2$, then $\{(f,v_k^0)\}$ is in l^q, $pq=p+q$, by the Hausdorff–Young inequality, and

$$\left(\sum_1^\infty |\big(f,v_k^0\big)|^q\right)^{1/q} \le K\|f\|_p.$$

Since $\{1/\rho_k^0\}$ is in l^p for any p, $1<p<\infty$, we see that the second term converges absolutely and uniformly to a bounded function, depending continuously on f. If $2<p<\infty$, then f in L^p implies f is in L^2, so $\{(f,v_k^0)\}$ is in l^2 and

$$\left(\sum_1^\infty |\big(f,v_k^0\big)|^2\right)^{1/2} \le K\|f\|_2 \le K\|f\|_p.$$

Considerations for the term involving $A_k(x,\varepsilon)$ are similar. Thus, B is a bounded linear operator.

Next we consider (4.6). Using (3.15), we obtain

$$\sum_1^\infty \frac{1}{\rho_k^0}\big(f,C_k(x,\varepsilon)v_k^0\big)u_k^0+\sum_1^\infty \frac{1}{\rho_k^0}\big(f,D_k(x,\varepsilon,\alpha)T_{l+1}\big(\rho_k^0 x\big)\big)u_k^0 + \sum_1^\infty \big(f,E_k(x,\varepsilon,\alpha)\big)u_k^0.$$

Again, the series involving E_k presents no difficulties. Consider the expression for D_k in the statement of Theorem 3.15. Let

$$X_k(x,\varepsilon)=\int_0^x T_l\big(\rho_k^0 t\big)V^*\big(T_l\big(\rho_k^0 s\big)\big)\,dt,$$

$$Y_k=-|\alpha|\bar{\beta}-\frac{1}{2}\int_0^1 T_l\big(\rho_k^0 t\big)V^*\big(T_l\big(\rho_k^0 s\big)\big)\,dt.$$

Then

$$D_k(x,\varepsilon,\alpha)=X_k(x,\varepsilon)+xY_k(\varepsilon,\alpha)+|\alpha|\bar{\beta}_0. \tag{4.10}$$

We see that

$$\sum_1^\infty \frac{1}{\rho_k^0}\big(f,\big(xY_k+|\alpha|\bar{\beta}_0\big)T_{l+1}\big(\rho_k^0 x\big)\big)u_k^0$$

represents a bounded linear operator on X, using the same arguments as in the consideration of B. Next we consider

$$\sum_1^\infty \frac{1}{\rho_k^0}\big(f, X_k(x,\varepsilon)T_{l+1}(\rho_k^0 x)\big)u_k^0(x). \tag{4.11}$$

Note that

$$\int_0^1 f(x)\bar{X}_k(x,\varepsilon)T_{l+1}(\rho_k^0 x)\,dx = \int_0^1 T_l(\rho_k^0 t)\overline{V^*\big(T_l(\rho_k^0 s)\big)}\left(\int_t^1 f(x)T_{l+1}(\rho_k^0 x)\,dx\right)dt.$$

Using Assumption 4.7 and the uniform bound on $T_l(\rho_k^0 t)V^*(T_l(\rho_k^0 s))$, we have

$$\sum_1^\infty \frac{1}{\rho_k^0}|(f, X_k(x,\varepsilon)T_{l+1}(\rho_k x))| \le M\sum_1^\infty \frac{1}{\rho_k^0}\int_0^1 \left|\int_t^1 f(x)T_{l+1}(\rho_k^0 x)\,dx\right| dt.$$

Let

$$f_t(x) = \begin{cases} 0, & 0 \le x < t, \\ f(x), & t < x < 1. \end{cases}$$

Then

$$\sum_1^\infty \frac{1}{\rho_k^0}|\big(f, X_k(x,\varepsilon)T_{l+1}(\rho_k^0 x)\big)| \le M\sum_1^\infty \frac{1}{\rho_k^0}\int_0^1 |\big(f_t, T_{l+1}(\rho_k^0 x)\big)|\,dt$$

$$= M\int_0^1 \left|\sum_1^\infty \left[\frac{1}{\rho_k^0}|\big(f_t, T_{l+1}(\rho_k^0 x)\big)|\right]\right| dt.$$

For $1 < p \le 2$, we again use the Hausdorff–Young inequality

$$\left(\sum_1^\infty |\big(f_t, T_{l+1}(\rho_k^0 x)\big)|^q\right)^{1/q} \le K\|f_t\|_p \le K\|f\|_p$$

and proceed as before. For $2 < p < \infty$, f_t as in L^2 and

$$\left(\sum_1^\infty |\big(f_t, T_{l+1}(\rho_k^0 x)\big)|^2\right)^{1/2} \le K\|f_t\|_2 \le K\|f\|_2 \le K\|f\|_p.$$

This completes the proof of part (i) of the theorem.

For $p = 2$ we can show that (4.11) represents a bounded linear operator without using Assumption 4.7. We have

$$\left|\int_0^1 f(x)\bar{X}_k(x,\varepsilon)T_{l+1}(\rho_k^0 x)\,dx\right|$$

$$\le \left[\int_0^1 |T_{l+1}(\rho_k^0 x)\overline{V^*\big(T_l(\rho_k^0 s)\big)}|^2\,dt\right]^{1/2}\left[\int_0^1 \left|\int_0^1 f_t(x)T_{l+1}(\rho_k^0 x)\,dx\right|^2 dt\right]^{1/2}$$

Since $V: L^2 \to L^2$ and all functions T_l are uniformly bounded, there is a constant $M > 0$ such that

$$\int_0^1 |T_l(\rho_k^0 t)\overline{V^*\big(T_l(\rho_k^0 s)\big)}|^2\,dt \le M.$$

Thus

$$\left|\sum_1^\infty \frac{1}{\rho_k^0}\left(f, X_k(x,\varepsilon)T_{l+1}(\rho_k^0 x)\right)\right| \le \left(\sum_1^\infty \left(\frac{1}{\rho_k^0}\right)^2\right)^{1/2}\left[\sum_1^\infty |(f, X_k)T_{l+1}(\rho_k^0 x)|^2\right]^{1/2}$$

$$\le M\left[\sum_1^\infty \int_0^1 \left|\int_0^1 f_t(x)T_{l+1}(\rho_k^0 x)\,dx\right|^2 dt\right]^{1/2}$$

$$= M\left|\int_0^1 \sum_1^\infty \left|\int_0^1 f_t(x)T_{l+1}(\rho_k^0 x)\,dx\right|^2 dt\right|^{1/2}$$

$$\le M\left|\int_0^1 \| f_t\|_2^2\, dt\right|^{1/2} \le M\| f\|_2.$$

THEOREM 4.12. *Under the hypotheses of Theorem* 4.9, *the operators* $L=L(\varepsilon,\alpha)$ *and* $J=J(\varepsilon,\alpha)$ *are compact and analytic in each component of* ε *and* α.

Proof. Since the partial sums of the infinite series defining L, J converge in the uniform operator topology, L and J are approximated by operators of finite rank and are consequently compact.

If f is in X, g is in X^*, then (Lf, g) and (Jf, g) are clearly analytic in ε and α.

5. The structure of $H(\varepsilon,\alpha)$. Let

$$V_k(\varepsilon) = \left(VT_l(\rho_k^0 x), T_l(\rho_k^0 x)\right). \tag{5.1}$$

Assumption 5.2. The sequence $\{V_k(\varepsilon)\}$ defines a multiplier transform

$$M_0(\varepsilon)f = \sum_1^\infty V_k(\varepsilon)\left(f, v_k^0\right)u_k^0$$

for each ε, $|\varepsilon| \le \tilde{\varepsilon}$ and

$$\| M_0(\varepsilon)\| \le K\,\| V(\varepsilon)\|.$$

Remark 5.3. The operators introduced in Remark 4.8 also satisfy Assumption 5.2. For $V(\varepsilon)f = q(x,\varepsilon)f(x)$,

$$V_k(\varepsilon) = \int_0^1 q(x,\varepsilon)T_l^2(\rho_k^0 x)\,dx$$

$$= \int_0^1 q(x,\varepsilon)\,dx - 2(-1)^l \int_0^1 q(x,\varepsilon)\cos 2k\pi X\,dx$$

$$= Q(\varepsilon) - 2(-1)^l q_k(\varepsilon),$$

where $Q(\varepsilon)$ defines the multiplier $Q(\varepsilon)I$, and since $\{q_k(\varepsilon)\}$ is in l^2, $\{q_k\}$ defines a multiplier on each $L^p(0,1)$, $1 \le p < \infty$. For

$$V(\varepsilon)f = \int_0^1 K(x,t,\varepsilon)f(t)\,dt,$$

we have

$$V_k(\varepsilon) = \int_0^1 \int_0^1 K(x,t,\varepsilon)T_l(\rho_k^0 x)T_l(\rho_k^0 t)\,dx\,dt.$$

With K uniformly bounded, $\{V_k(\varepsilon)\}$ is in l^2.

For $X = L^2(0,1)$, Assumption 5.2 is no restriction on $V(\varepsilon)$.

THEOREM 5.4. *Let $X=L^p(0,1)$, $1<p<\infty$. If $V(\varepsilon)$ satisfies Assumptions* 4.7, 5.2, *then there exists an analytic, compact multiplier transform $M_1(\varepsilon,\alpha)$ on X, such that*

$$\|M_1(\varepsilon,\alpha)\|=O\big(|\varepsilon|^2+|\alpha|^2\big)$$

and

$$H(\varepsilon,\alpha)=B(\varepsilon,\alpha)\big[H_0+M_0(\varepsilon)+2|\alpha|\beta I+M_1(\varepsilon,\alpha)\big]B^{-1}(\varepsilon,\alpha). \tag{5.5}$$

If $X=L^2(0,1)$, this result holds with no restriction on $V(\varepsilon)$.

Proof. With Assumption 4.7 or $X=L^2$, Theorem 4.9 is valid. Thus, for any f in X,

$$\begin{aligned}Hf&=\sum_1^\infty (Hf,v_k)u_k=B\sum_1^\infty \lambda_k\big(f,B^{-1*}v_k^0\big)u_k^0\\&=B\left[\sum_1^\infty \lambda_k^0\big(B^{-1}f,v_k^0\big)u_k^0+\sum_1^\infty\big(\lambda_k-\lambda_k^0\big)\big(B^{-1}f,v_k^0\big)u_k^0\right]\\&=B\big[H_0+M(\varepsilon,\alpha)\big]B^{-1}f,\end{aligned}$$

provided the infinite series

$$M(\varepsilon,\alpha)g=\sum_1^\infty\big(\lambda_k-\lambda_k^0\big)\big(g,v_k^0\big)u_k^0$$

converges. Using Corollary 3.7 and (3.9), we have

$$M(\varepsilon,\alpha)g=M_0(\varepsilon)g+2|\alpha|\beta g+M_1(\varepsilon,\alpha)g,$$

provided the series in Assumption 5.2 defining $M_0(\varepsilon)$ and the series

$$M_1(\varepsilon,\alpha)g=|\varepsilon|^2\sum_1^\infty a_k(\varepsilon,\alpha)\frac{(g,v_k^0)}{k}u_k^0+|\alpha|^2\sum_1^\infty b_k(\varepsilon,\alpha)\frac{(g,v_k^0)}{k}u_k^0 \tag{5.6}$$

converge. These latter series converge due to the $\frac{1}{k}$, and clearly M_1 satisfies the conditions stated. From Assumption 5.2 the series defining $M_0(\varepsilon)$ also converges.

6. The semigroup generated by $H(\varepsilon,\alpha)$. Assumptions 4.7, 5.2 hold throughout this section. Let $U(t)$ denote the semigroup generated by H, and let $U_0(t)$ denote the semigroup generated by H_0. Since M_0, M_1 in (5.5) commute with H_0, we have using (5.5),

$$U(t)=BU_0(t)e^{M_0(\varepsilon)t}e^{2|\alpha|\beta t}e^{M_1(\varepsilon,\alpha)t}B^{-1}. \tag{6.1}$$

Since the exponential of a bounded operator can be evaluated directly from the power series, we have

$$U(t)=BU_0(t)\big[I+t\big(M_0(\varepsilon)+2|\alpha|\beta I+M_1(\varepsilon,\alpha)\big)+W(\varepsilon,\alpha,t)\big]B^{-1}, \tag{6.2}$$

where

$$\|W\|\le t^2K\big(|\varepsilon|^2+|\alpha|^2\big)$$

uniformly on any interval $0\le t\le T$ for given $T>0$. We summarize in a form slightly less detailed than (6.2):

THEOREM 6.3. *Let $T>0$ be given. There exists an operator $Y(\varepsilon,\alpha,t)$, analytic in ε and α, $|\varepsilon|\le\tilde{\varepsilon}$, $|\alpha|\le\tilde{\alpha}$, such that*

$$\|Y(\varepsilon,\alpha,t)\|\le Kt(|\varepsilon|+|\alpha|),\qquad 0\le t\le T,$$

and such that

$$U(t)=B(\varepsilon,\alpha)\left[U_0(t)+tY(\varepsilon,\alpha,t)\right]B^{-1}(\varepsilon,\alpha).$$

7. The saturation problem and approximation to the identity. A general survey of the saturation problem for semigroups is given in [4, pp. 83–92]. For a semigroup $U(t)$ and a vector f, we consider the problem of estimating $\|U(t)f-f\|$ as $t\to 0^+$. Unless $U(t)f$ is constant, $\|U(t)f-f\|$ can be no smaller than $O(t)$. If X is reflexive, this optimal bound is achieved if and only if f is in the domain of the infinitesimal generator of $U(t)$.

Referring to the semigroups generated by H and H_0, let $\phi(s)$ be a positive, nonincreasing function on $(0,\infty)$ such that $\phi(s)\to 0$ as $s\to\infty$. Let $\Lambda(\phi)$ denote the collection of all f in X such that

(i) $$\|U(t)f-f\|=O\big(\phi(t^{-1})\big), \qquad t\to 0,$$

(ii) $$\limsup_{t\to 0}\frac{\|U(t)f-f\|}{\phi(t^{-1})}>0.$$

Let $\Lambda_0(\phi)$ denote the similar class for $U_0(t)$.

THEOREM 7.1. *If $s\phi(s)\to\infty$ as $s\to\infty$, then $\Lambda(\phi)=B(\varepsilon,\alpha)\Lambda_0(\phi)$.*

Proof. Using (6.3), we have

$$U(t)f-f=B\big(U_0(t)-I\big)B^{-1}f+tBYB^{-1}f. \tag{7.2}$$

Let g be in $\Lambda_0(\phi)$ and let $f=Bg$. Then

$$\|U(t)f-f\|\le K\big(\phi(t^{-1})+t\|g\|\big)=K\phi(t^{-1})\left[1+\frac{t}{\phi(t^{-1})}\|g\|\right].$$

Since $s\phi(s)\to\infty$, we have $t/\phi(t^{-1})\to 0$, so

$$\|U(t)f-f\|\le K\phi(t^{-1}).$$

For the same reason,

$$\limsup_{t\to 0}\frac{\|U(t)f-f\|}{\phi(t^{-1})}>0,$$

so $f\in\Lambda(\phi)$. Solving (7.2) for $U_0(t)-I$, we show in the same way the converse: that if f is in $\Lambda(\phi)$, then $B^{-1}f$ is in $\Lambda_0(\phi)$.

THEOREM 7.3. *If $\varphi(s)=\frac{1}{s}$ and zero is an eigenvalue of neither H nor H_0, then $\Lambda(\varphi)=B(\varepsilon,\alpha)\Lambda_0(\varphi)$.*

Proof. In a reflexive Banach space, for $\phi(s)=\frac{1}{s}$, condition (i) is satisfied if and only if f is in the domain of the infinitesimal generator [4, Corollary 2.1.3, p. 90]. If zero is not an eigenvalue of H, then $\Lambda(\phi)=\mathfrak{D}(\alpha)$, and similarly for H_0. Thus it suffices to show that $\mathfrak{D}(\alpha)=B(\varepsilon,\alpha)\mathfrak{D}(0)$.

If f is in $\mathfrak{D}(0)$, then $(f,v_k^0)=(H_0f,v_k^0)/\lambda_k^0$, so

$$Bf=\sum_1^\infty\frac{1}{\lambda_k^0}\big(H^0f,v_k^0\big)u_k.$$

Since $\lambda_k^0\sim k^2$, this series and its term by term derivative converge uniformly, so Bf satisfies the boundary conditions of $\mathfrak{D}(\alpha)$. Using the differential equation and 3.7 for λ_k, we see that the second term by term derivative converges in the norm of X. Thus

$B\mathcal{D}(0)\subset\mathcal{D}(\alpha)$. For the converse, we note that for f in $\mathcal{D}(\alpha)$, $(f,v_k)=(Hf,v_k)/\lambda_k$, so

$$B^{-1}f=\sum_1^\infty \frac{1}{\lambda_k}(Hf,v_k)u_k^0.$$

We then proceed as above.

REFERENCES

[1] H. E. BENZINGER, *The L^p behavior of eigenfunction expansions*, Trans. Amer. Math. Soc., 174 (1972), pp. 333–344.

[2] _____, *Completeness of eigenvectors in Banach spaces*, Proc. Amer. Math. Soc. 38 (1973), pp. 319–324.

[3] _____, *Perturbation of the heat equation*, J. Differential Equations 32 (1979), 398–419.

[4] P. L. BUTZER AND H. BERENS, *Semi-groups of Operators and Approximation*, Springer-Verlag, New York, 1967.

[5] A. MESSIAH, *Quantum Mechanics*, Vol. II, John Wiley, New York, 1962.

[6] M. A. NEUMARK, *Lineare Differentialoperatoren*, Akademie-Verlag, Berlin, 1960.

[7] M. REED AND B. SIMON, *Methods of Mathematical Physics, Vol.* II, *Fourier Analysis, Self-Adjointness*, Academic Press, New York, 1975.

[8] _____, *Methods of Mathematical Physics, Vol.* IV, *Analysis of Operators*, Academic Press, New York, 1975.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0002 $01.00/0

EXPONENTIAL LEVELING FOR STOCHASTICALLY PERTURBED DYNAMICAL SYSTEMS*

MARTIN DAY†

Abstract. This paper considers solutions of $0=\varepsilon\sum_{i,j}a^{\varepsilon}_{i,j}(x)u^{\varepsilon}_{x_ix_j}+\sum_i b^{\varepsilon}_i(x)u^{\varepsilon}_i$ in a bounded domain Ω for which $\sup_\Omega|u^\varepsilon|$ is bounded in $\varepsilon>0$. We assume that $a^\varepsilon\to a^0$, $b^\varepsilon\to b^0$ and that all solutions of the ODE $\dot{x}=b^0(x)$, $x(0)\in\Omega$ converge to a single linearly asymptotically stable critical point in Ω without leaving Ω. We give a proof, based on the standard probabilistic interpretation of u^ε, of an exponential leveling property: $\sup_{x,y\in K}|u^\varepsilon(x)-u^\varepsilon(y)|\le e^{-\delta/\varepsilon}$ for some $\delta>0$ which depends on the compact set $K\subseteq\Omega$.

1. Introduction. Consider a deterministic system described by an ordinary differential equation in $\mathbb{R}^d$:

$$dx^0(t)=b(x^0(t))\,dt. \tag{1.1}$$

A natural model for the behavior of this system, when subjected to a small stochastic perturbation, is the diffusion process described by the Ito equation:

$$dx^\varepsilon(t)=b(x^\varepsilon(t))\,dt+\sqrt{\varepsilon}\,\sigma(x^\varepsilon(t))\,d\omega_t, \tag{1.2}$$

ω_t being a Brownian motion in $\mathbb{R}^d$. Applications of this type of model can be found in Ludwig [6], Schuss [10] and Matkowsky and Schuss [9]. Several aspects of the asymptotic behavior of $x^\varepsilon(\cdot)$ as $\varepsilon\to 0$ are of interest. Consider in particular a bounded domain Ω. If τ_Ω denotes the exit time of $x^\varepsilon(\cdot)$ from Ω and E^ε_x the expectation for the solution of (1.2) subject to $x^\varepsilon(0)=x$, then (under appropriate regularity assumptions)

$$u^\varepsilon(x)=E^\varepsilon_x[f(x^\varepsilon(\tau_\Omega))]$$

is the solution of the Dirichlet problem

$$0=\mathcal{L}^\varepsilon[u]=\frac{\varepsilon}{2}\sum_{i,j}a_{ij}(x)u_{x_ix_j}+\sum_i b_i(x)u_{x_i}\quad\text{in }\Omega\text{ with }u|_{\partial\Omega}=f. \tag{1.3}$$

(Here $a=\sigma\sigma^T$.) The behavior of u^ε as $\varepsilon\downarrow 0$ depends, of course, on the nature of the trajectories of (1.1) which start in Ω. One of the more interesting cases is when all deterministic trajectories starting in Ω remain in Ω and approach a unique stable point, at the origin, say. Because all continuous solutions of the reduced equation

$$0=\sum_i b_i(x)u^0_{x_i}$$

in Ω are constant, one expects that u^ε approaches a constant function, or at least somehow "levels out". We prove here, under modest assumptions, that this leveling does occur, and at an exponential rate:

$$\sup_{x,y\in K}|u^\varepsilon(x)-u^\varepsilon(y)|\le e^{-\delta/\varepsilon}$$

for any compact $K\subseteq\Omega$, some $\delta>0$ and all sufficiently small ε. In many cases much more is known. Matkowsky and Schuss [8] presented a formal calculation to show that

---

* Received by the editors August 4, 1980. This research was supported by the U.S. Air Force Office of Scientific Research under grant AF-AFOSR 76-3063C.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. Now at Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

u^ε converges to a constant function and derived a formula for what this constant should be. Kamin [5] and Devinatz and Friedman [1] gave rigorous proofs of this in cases where $\mathcal{L}^\varepsilon$ has a selfadjoint form,

$$\mathcal{L}^\varepsilon[u]=e^{-\psi/\varepsilon}\sum_{i,j}\frac{\partial}{\partial x_i}\left(e^{\psi/\varepsilon}a_{ij}u_{x_j}\right). \tag{1.4}$$

In [4] Kamin showed that the formal calculation of Matkowsky and Schuss for (1.3) is correct provided the solutions of certain auxiliary first order PDE's exist and are sufficiently smooth. The fundamental work of Vent-tsel and Freidlin [12] also establishes that u^ε converges to a constant for (1.3) in the case that the variational distance $V(0,y)$ which is central to their treatment attains its minimum over $y\in\partial\Omega$ at a unique place.

Actually, the $\mathcal{L}^\varepsilon$ in (1.4) is of a more general form than (1.3):

$$\mathcal{L}^\varepsilon[u]=\frac{\varepsilon}{2}\sum_{i,j}a_{ij}u_{x_ix_j}+\sum_i b_i^\varepsilon u_{x_i} \tag{1.5}$$

with $b^\varepsilon\to b^0$ as $\varepsilon\to 0$. In this context the solutions u^ε may not converge to a constant. Indeed, in [1] the authors presented the two examples:

$$\varepsilon(x+2)u''-x(x+2)u'=0, \tag{1.6a}$$

$$\varepsilon(x+2)u''+(\varepsilon-x(x+2))u'=0 \tag{1.6b}$$

on $[-1,1]$, both with $u(-1)=0$, $u(1)=1$. They observed that $u^\varepsilon\to\frac{1}{2}$ for (1.6a) and $u^\varepsilon\to\frac{3}{4}$ for (1.6b). If we combine these two examples as

$$\varepsilon(x+2)u''+\left(\varepsilon\sin\left(\frac{1}{\varepsilon}\right)-x(x+2)\right)u'=0, \tag{1.6c}$$

we get an example of the type (1.5) for which u^ε does not converge.

The result proved here is for $\mathcal{L}^\varepsilon$ of the form

$$\mathcal{L}^\varepsilon[u]=\frac{\varepsilon}{2}\sum_{i,j}a_{i,j}^\varepsilon u_{x_ix_j}+\sum_i b_i^\varepsilon u_{x_i}.$$

The a^ε, b^ε are required to converge to a^0, b^0 as $\varepsilon\downarrow 0$. This form of $\mathcal{L}^\varepsilon$ encompasses all the cases (1.3)–(1.6) mentioned. The boundary function $u^\varepsilon|_{\partial\Omega}=f^\varepsilon$ is allowed to be ε-dependent $u^\varepsilon|_{\partial\Omega}$ and is required only to be bounded (in both x and ε) and measurable (Borel), but need not converge with ε.

Section 2 contains the technical assumptions and the statement of the main theorem. Sections 3 and 4 are devoted to a bound on hitting probabilities which is the cornerstone of our proof. The proof of the theorem is given in §4 also. Section 5 contains two additional remarks.

2. Technical assumptions and statement of the main result. The domain $\Omega\subseteq\mathbb{R}^d$ is assumed to be bounded. To treat u^ε as the solution to an elliptic boundary value problem with $u^\varepsilon|_{\partial\Omega}=f^\varepsilon$, a specified continuous function, one might also want to impose the requirement that $\partial\Omega$ be C^2. The probabilistic definition (2.1) of u^ε renders this unnecessary, however. The assumptions on the coefficients are as follows:

(a) $a^\varepsilon(x)$, $a^0(x)$ are Lipschitz (or just Hölder) continuous in x, uniformly in ε, positive definite symmetric $d\times d$ matrices on $\bar{\Omega}$ and $|a_{ij}^\varepsilon-a_{ij}^0|\to 0$ uniformly on $\bar{\Omega}$ as $\varepsilon\downarrow 0$;

(b) $b^\varepsilon(x)$ and $b^0(x)$ are in $C^1(\bar{\Omega})$, $|b^\varepsilon-b^0|$ and $|b_{x_i}^\varepsilon-b_{x_i}^0|$, $i=1,\cdots,d$ all converge to 0 uniformly on $\bar{\Omega}$ as $\varepsilon\downarrow 0$;

(c) for any solution to $x^{0\prime}(t)=b^0(x^0(t))$ with $x^0(0)\in\Omega$, $x^0(t)\in\Omega$ for all $t\geq 0$ and $\lim_{t\to+\infty} x^0(t)=0$;

(d) the matrix $B=[\partial b_i^0(0)/\partial x_j]$ is stable, i.e., all itseigenvalues have negative real parts.

For a specified $x\in\Omega$, $x^\varepsilon(t)$ is a Markov diffusion process with $x^\varepsilon(0)=x$ and differential generator

$$\mathcal{L}^\varepsilon=\frac{\varepsilon}{2}\sum_{i,j} a_{i,j}^\varepsilon \frac{\partial}{\partial x_i \partial x_j}+\sum_i b_i^\varepsilon \frac{\partial}{\partial x_i}.$$

For definiteness one can think of a^ε, b^ε as being extended to all of $\mathbb{R}^d$, the process $x^\varepsilon(\cdot)$ then being obtained as below for all $t<+\infty$. We are only concerned with $x^\varepsilon(t)$ for $t\leq\tau_\Omega$, however, which does not depend on this extention. If one likes, $x^\varepsilon(t)$ can be considered as the solution to a stochastic differential equation (see [7] or [11])

$$dx^\varepsilon(t)=b^\varepsilon(x^\varepsilon(t))\,dt+\sqrt{\varepsilon}\,\sigma^\varepsilon(x^\varepsilon(t))\,d\omega_t, \qquad x^\varepsilon(0)=x$$

if $a^\varepsilon=\sigma^\varepsilon(\sigma^\varepsilon)^T$, where σ^ε is Lipschitz. Alternately, $x^\varepsilon(t)$ can be discussed directly via the martingale problem associated with $\mathcal{L}^\varepsilon$ [11]. (Continuity of coefficients is sufficient for that treatment.)

The boundary functions $f^\varepsilon(x)$ are assumed to be bounded in x and $\varepsilon>0$ and measurable on $\partial\Omega$. The $u^\varepsilon(x)$ are now defined by

$$u^\varepsilon(x)=E_x^\varepsilon[f^\varepsilon(x^\varepsilon(\tau_\Omega))]. \tag{2.1}$$

It can be shown that $u^\varepsilon\in C^2(\Omega)$ and satisfies

$$\mathcal{L}^\varepsilon[u^\varepsilon]=0 \quad \text{in } \Omega.$$

Indeed, on any ball B with $\bar{B}\subseteq\Omega$ it is true that u^ε is the Perron solution corresponding to the boundary data $u^\varepsilon|_{\partial B}$. Since Perron solutions are C^2 on the interior of their domains, for bounded measurable data [3, Thm. 6.11], it follows that $u^\varepsilon\in C^2(\Omega)$. The boundary behavior of u^ε does not concern us; only the boundedness in ε, x is necessary for our arguments below.

There is one more condition that we will need when proving Theorem 2 below. In deriving (4.6) we will use

$$\frac{b^\varepsilon(x)-b^0(x)}{|x|}=o(1) \quad \text{as } \varepsilon\downarrow 0 \text{ uniformly in } \Omega. \tag{2.2}$$

We know that $b^0(0)=0$, so the above will follow from the convergence of $b_{x_i}^\varepsilon$ to $b_{x_i}^0$ if the further condition $b^\varepsilon(0)=0$ is true. Once we restrict our attention to x in a compact $K\subseteq\Omega$, however, we can achieve $b^\varepsilon(0)=0$ without imposing any further assumptions. The following argument accomplishes this: from the stability of 0 with respect to b^0 as in (d) above, and the uniform convergence of b^ε to b^0, one can deduce that (for sufficiently small ε) b^ε has a critical point ζ^ε such that $\zeta^\varepsilon\to 0$ as $\varepsilon\downarrow 0$. Change variables to $y=x-\zeta^\varepsilon$. The new coefficients $\tilde{a}^\varepsilon(y)=a^\varepsilon(y+\zeta^\varepsilon)$ and $\tilde{b}^\varepsilon(y)=b^\varepsilon(y+\zeta^\varepsilon)$ satisfy all of our assumptions above as well as $\tilde{b}^\varepsilon(0)=0$. The only difficulty is that the domains $\Omega-\zeta^\varepsilon$ are ε-dependent. We can pass to a subdomain Ω' so that $y\in\Omega'$ implies $x=y+\zeta^\varepsilon\in\Omega$ and a compact subset $K'\subseteq\Omega'$ so that $x\in K$ implies $y=x-\zeta^\varepsilon\in K'$, for sufficiently small ε. Applying Theorem 1 to Ω', K', we get the same result as for Ω, K. Taking the details of this argument for granted, we assume in the following that $b^\varepsilon(0)=0$, and consequently that (2.2) is true.

Here is our main theorem.

THEOREM 1. *Under the assumptions described above, for any compact* $K\subseteq\Omega$ *there exist* $\delta>0$ *and* $\varepsilon_0>0$ *so that for all* $0<\varepsilon<\varepsilon_0$

$$\sup_{x,y\in K}|u^\varepsilon(x)-u^\varepsilon(y)|\leq e^{-\delta/\varepsilon}.$$

Roughly, the reasoning behind the proof is that for small ε $x^\varepsilon(t)$ should, with high probability, follow the deterministic trajectory $x^0(t)$ into the vicinity of the origin before making its first excursion to the boundary $\partial\Omega$. A precise probabilistic estimate along these lines is established in the next two sections. To apply the probabilistic estimate, we need to know a modulus of continuity for u^ε. The following lemma establishes the modulus that we need; the rescaling argument is the same one used by Kamin [5].

LEMMA 1. *Let* $K\subseteq\Omega$ *be compact. Then there is a constant* C *so that* $|\nabla u^\varepsilon(x)|\leq C\varepsilon^{-1/2}$ *for all* $x\in K$ *and* $\varepsilon<1$.

Proof. Make the change of variables $y=\varepsilon^{-1/2}x$. Then $v^\varepsilon(y)=u^\varepsilon(\varepsilon^{1/2}y)$ satisfies

$$\frac{1}{2}\sum_{i,j}\tilde{a}^\varepsilon_{i,j}(y)v_{y_iy_j}+\sum_i\tilde{b}^\varepsilon_i(y)v_{y_i}=0\quad\text{for } y\in\varepsilon^{-1/2}\Omega.$$

The coefficients $\tilde{a}^\varepsilon(y)=a^\varepsilon(\varepsilon^{1/2}y)$, $\tilde{b}^\varepsilon(y)=\varepsilon^{-1/2}b^\varepsilon(\varepsilon^{1/2}y)$ are Hölder continuous with respect to y uniformly in ε. If we take r so that $B_r(x)=\{z:|x-z|<r\}\subseteq\Omega$ whenever $x\in K$, then $B_r(y)\subseteq\varepsilon^{-1/2}\Omega$ whenever $y\in\varepsilon^{-1/2}K$ and $\varepsilon<1$. We can apply the basic Scahuder interior estimate [3, Thm. 6.2] to $B_r(y)$ for any $y\in\varepsilon^{-1/2}K$ to conclude that $|\nabla v(y)|\leq C$ for all $y\in K$ for some constant C. This implies the lemma after changing back to the original variable x.

3. A prototype: An Ornstein–Uhlenbeck process. Before proving the estimate on hitting probabilities of the next section, it is convenient to look at the special case of an Ornstein-Uhlenbeck process with generator as in (3.1) below. The proof of the general case rests on a comparison with the function described by (3.4) and analyzed below.

In $\mathbb{R}^d$, $d\geq 2$, suppose that $\alpha>0$ is a constant and $\zeta^\varepsilon(t)$ is a diffusion process with differential generator

$$\mathcal{G}^\varepsilon[u](x)=\frac{\varepsilon}{2}\Delta u(x)-\alpha x\cdot\nabla u(x). \tag{3.1}$$

Let $\tau(r)$ be the hitting time of the sphere of radius r;

$$\tau(r)=\inf\{t\geq 0:|\zeta^\varepsilon(t)|=r\}. \tag{3.2}$$

Take a fixed $R>0$ and, for $r_0<|x|<R$, define the hitting probability

$$Q^\varepsilon_{r_0}(x)=P^\varepsilon_x[\tau(r_0)<\tau(R);\tau(r_0)<\infty].$$

What we show is that there exists a positive constant $\delta_1>0$ so that

$$Q^\varepsilon_{r_0}(x)\geq 1-e^{-\delta_1/\varepsilon}\quad\text{whenever } r_0\geq e^{-\delta_1/\varepsilon}\text{ and }|x|<\tfrac{1}{3}R. \tag{3.3}$$

In words, for $|\zeta^\varepsilon(0)|<\frac{1}{3}R$ we can let $r_0\downarrow 0$ exponentially in ε^{-1} and at the same time have the probability that $\tau(r_0)<\tau(R)$ converging to 1 exponentially fast. To prove this we calculate $Q^\varepsilon_{r_0}(\cdot)$. By symmetry $Q^\varepsilon_{r_0}$ depends only on $r=|x|$. Thus $Q^\varepsilon_{r_0}(x)=Q(r)$, where $\mathcal{G}^\varepsilon[Q(r)]=0$ with $Q(r_0)=1$, $Q(R)=0$.

$$0=\mathcal{G}^\varepsilon[Q(|x|)]=\frac{\varepsilon}{2}Q''(r)+\left[\frac{\varepsilon}{2}\frac{d-1}{r}-\alpha r\right]Q'(r)$$

or

$$Q''(r)+\left[\frac{\beta}{r}-\frac{2\alpha}{\varepsilon}r\right]Q'(r)=0,\quad Q(r_0)=1,\quad Q(R)=0. \tag{3.4}$$

For the above, $\beta=d-1$, but we will carry out the calculation for arbitrary positive constants α,β. Solving (3.4) gives

$$Q(r)=1-\frac{\int_{r_0}^{r}s^{-\beta}e^{(\alpha/\varepsilon)s^2}\,ds}{\int_{r_0}^{R}s^{-\beta}e^{(\alpha/\varepsilon)s^2}\,ds}.$$

If $r_0<r\leq\frac{1}{3}R$, then

$$\frac{\int_{r_0}^{r}s^{-\beta}e^{(\alpha/\varepsilon)s^2}\,ds}{\int_{r_0}^{R}s^{-\beta}e^{(\alpha/\varepsilon)s^2}\,ds}\leq\frac{\int_{r_0}^{R/3}}{\int_{2R/3}^{R}}\leq\frac{\frac{R}{3}e^{(\alpha/\varepsilon)(R/3)^2}r_0^{-\beta}}{\frac{R}{3}e^{(\alpha/\varepsilon)(2R/3)^2}(R)^{-\beta}}=e^{-(\alpha/\varepsilon)R^2/3}r_0^{-\beta}R^{\beta}.$$

If $\log(r_0)\geq-\frac{1}{\varepsilon}\cdot(\frac{\alpha R^2}{6\beta})$, then the preceding is $\leq e^{-(\alpha/\varepsilon)R^2/6}R^{\beta}$. Consequently, if δ_1 is slightly less than the minimum of $\frac{\alpha R^2}{6}$ and $\frac{\alpha R^2}{6\beta}$ (slightly less so as to absorb the R^{β}), then, for $r_0\geq e^{-\delta_1/\varepsilon}$ and $r\leq\frac{R}{3}$, we have

$$Q(r)\geq 1-e^{-\delta_1/\varepsilon}\quad\text{for sufficiently small }\varepsilon>0. \tag{3.5}$$

4. The hitting probabilities in the general case. Next, we prove that an estimate like (3.3) holds for the hitting probabilities of the process $x^{\varepsilon}(t)$. ($\tau(r)$ now denotes the time of first contact with the ball of radius r about the origin for $x^{\varepsilon}(\cdot)$.)

THEOREM 2. *For any compact $K\subseteq\Omega$, there exists $\delta>0$ so that for some $\varepsilon_0>0$ and all $0<\varepsilon<\varepsilon_0$,*

$$P_x^{\varepsilon}[\tau(r_0)<\tau_{\Omega}]\geq 1-e^{-\delta/\varepsilon}\quad\textit{whenever } x\in K \textit{ and } r_0\geq e^{-\delta/\varepsilon}.$$

Proof. We will first make an argument for an appropriate neighborhood of the origin. The key to the proof is to use not the standard Euclidean norm $|x|$ but a different symmetric positive definite quadratic form. By hypothesis, the matrix $B=[\partial b_i^0(0)/\partial x_j]$ is stable. Lyapunov's theorem on matrices implies that there exists a unique symmetric positive definite matrix V which solves

$$B^TV+VB=-I.$$

Define $\rho(x)=[x^TVx]^{1/2}$. For $f\in C^2(\mathbb{R})$, a computation gives that

$$\begin{aligned}&\mathcal{L}^{\varepsilon}[f(\rho(x))]=f''(\rho)\cdot\frac{\varepsilon}{2}\sum_{i,j}a_{ij}^{\varepsilon}\rho_{x_i}\rho_{x_j}+f'(\rho)\cdot\left[\frac{\varepsilon}{2}\sum_{i,j}\alpha_{ij}^{\varepsilon}\rho_{x_ix_j}+\sum_i b_i\rho_{x_i}\right],\\ &\nabla\rho(x)=\frac{x^TV}{\rho},\qquad \rho_{x_ix_j}=\frac{V_{ij}}{\rho}+\frac{1}{\rho^3}\sum_{l,k}V_{il}x_lx_kV_{kj}.\end{aligned} \tag{4.1}$$

The idea is to effect a comparison of each of the terms of $\mathcal{L}^{\varepsilon}[f(\rho)]$ with those of $\mathcal{G}^{\varepsilon}[Q(r)]$ in (3.4). First,

$$\sum_{i,j}\alpha_{ij}^{\varepsilon}\rho_{x_i}\rho_{x_j}=\frac{x^TVa^{\varepsilon}Vx}{x^TVx},$$

which is bounded above and below, away from 0, on $\Omega-\{0\}$. Moreover, these bounds can be taken to be independent of ε sufficiently small since $a^\varepsilon \to a^0$ uniformly. Thus, there exists a constant $A>0$ so that

$$A^{-1} \le \sum_{i,j} a_{ij}^\varepsilon \rho_{x_i}\rho_{x_j} \le A \quad \text{in } \Omega-\{0\}, \quad \text{all sufficiently small } \varepsilon. \tag{4.2}$$

Secondly,

$$\sum_{i,j} a_{ij}^\varepsilon \rho_{x_i x_j} = \frac{1}{\rho}\sum_{i,j} a_{ij}^\varepsilon V_{ij} + \frac{1}{\rho}\frac{x^T V a^\varepsilon V x}{\rho^2} \le \frac{C}{\rho} \tag{4.3}$$

for a positive constant C (again uniformly in ε sufficiently small). Thirdly,

$$\sum_i b_i^\varepsilon \rho_{x_i} = \frac{x^T V b(x)}{\rho(x)} = \frac{x^T V b^0(x)}{\rho(x)} + \frac{x^T V(b^\varepsilon - b^0)}{\rho}.$$

Using $b^0(x)=Bx+o(|x|)$ and $b^\varepsilon(x)-b^0(x)=|x|o(1)$ from (2.2), we find that

$$\sum_i b_i^\varepsilon \rho_{x_i} = \rho\cdot\left[\frac{x^T V B x}{\rho^2} + o(|x|) + o(1)\right] = \rho\cdot\left[-\frac{1}{2}\frac{|x|^2}{\rho^2} + o(|x|) + o(1)\right].$$

(The $o(|x|)$ is as $|x|\to 0$ and is independent of ε. The $o(1)$ is as $\varepsilon\to 0$ and is uniform in x.) The second equality is a consequence of our choice of V. It follows that, for some D, R and ε_0 all positive,

$$\sum_i b_i^\varepsilon \rho_{x_i} \le -D\rho(x) \quad \text{if } \rho(x)\le R \text{ and } \varepsilon<\varepsilon_0. \tag{4.4}$$

(Also, restrict R so that $x\in\Omega$ whenever $\rho(x)<R$.) Take $\alpha=DA^{-1}$, $\beta=AC$ and then $Q(\cdot)$ as in (3.4). Since $Q'\le 0$, (4.1)–(4.4) combined imply that, for $\varepsilon<\varepsilon_0$ and $\rho(x)\le R$,

$$\mathcal{L}^\varepsilon[Q(\rho(x))] \ge \frac{\varepsilon}{2}A^{-1}\cdot\left\{Q''(\rho)+Q'(\rho)\left[\frac{AC}{\rho}-\frac{2}{\varepsilon}A^{-1}D\rho\right]\right\} = 0. \tag{4.5}$$

Using $\tilde\tau(r)$ for the hitting time of the set $\{x:\rho(x)=r\}$ by $x^\varepsilon(\cdot)$, (4.5) implies, either via the maximum principle or the fact that $Q(\rho(x^\varepsilon(t)))$ is a submartingale, that

$$P_x^\varepsilon[\tilde\tau(r_0)<\tilde\tau(R)] \ge Q(\rho(x)).$$

If $\gamma>0$ is a constant so that $\gamma\le\rho(x)/|x|\le\gamma^{-1}$, then $\tilde\tau(\gamma r_0)>\tau(r_0)$, provided $|x^\varepsilon(0)|>r_0$. If $|x^\varepsilon(0)|<\gamma R$, then $\rho(x^\varepsilon(0))<R$ and $\tilde\tau(R)\le\tau_\Omega$. Consequently, for $r_0<|x^\varepsilon(0)|<\gamma R$, we have

$$P_x^\varepsilon[\tau(r_0)<\tau_\Omega] \ge P_x^\varepsilon[\tilde\tau(\gamma r_0)<\tilde\tau(R)] \ge Q(\rho(x)).$$

This is trivially true also if $|x^\varepsilon(0)|\le r_0$. The calculation of §2 now implies the existence of $\delta_1>0$ so that for all $0<\varepsilon<\varepsilon_0$ and $|x|\le\frac{\gamma}{3}R$,

$$P_x^\varepsilon[\tau(r_0)<\tau_\Omega] \ge 1-e^{-\delta_1/\varepsilon} \quad \text{if } \gamma r_0 \ge e^{-\delta_1/\varepsilon}. \tag{4.6}$$

The last step of the proof is to show that such an estimate remains true for all $x\in K$. By the strong Markov property,

$$\begin{aligned} P_x^\varepsilon[\tau(r_0)<\tau_\Omega] &= E_x^\varepsilon\left[P_{x^\varepsilon(\tau(\gamma R/3))}^\varepsilon[\tau(r_0)<\tau_\Omega], \tau\left(\frac{\gamma}{3}R\right)<\tau_\Omega\right] \\ &\ge (1-e^{-\delta_1/\varepsilon})\cdot P_x^\varepsilon\left[\tau\left(\frac{\gamma}{3}R\right)<\tau_\Omega\right]. \end{aligned}$$

It is sufficient, therefore, to prove that for some $\delta_2>0$ and all $x\in K$

$$P_x^\varepsilon\left[\tau\left(\frac{\gamma}{3}R\right)<\tau_\Omega\right]\geq 1-e^{-\delta_2/\varepsilon}. \tag{4.7}$$

Let $\phi^\varepsilon(t;x),\varepsilon\geq 0$, denote the solution of the deterministic equation $\phi'(t)=b^\varepsilon(\phi(t))$ with $\phi(0)=x$. Since K is contained in the domain of attraction of the stable point 0, there exist $T,\eta>0$ so that if $x\in K$ then

$$y\in\Omega \quad \text{whenever } |y-\phi^0(s;x)|<2\eta \quad \text{for some } 0\leq s\leq T,$$

and

$$|y|<\frac{\gamma}{3}R \quad \text{whenever } |y-\phi^0(T;x)|<2\eta.$$

As $\varepsilon\to 0$, b^ε converges to b^0 uniformly in Ω, and consequently $\phi^\varepsilon(t;x)$ converges to $\phi^0(t;x)$ uniformly for $t\in[0,T]$ and $x\in K$. Therefore, if ε is sufficiently small and $x\in K$, it will be true that

$$y\in\Omega \quad \text{whenever } |y-\phi^\varepsilon(s;x)|<\eta \quad \text{for some } 0<s<T,$$

and

$$|y|<\frac{\gamma}{3}R \quad \text{whenever } |y-\phi^\varepsilon(T;x)|<\eta.$$

For such ε and $x\in K$,

$$P_x^\varepsilon\left[\tau_\Omega\leq\tau\left(\frac{\gamma}{3}R\right)\right]\leq P_x^\varepsilon\left[\sup_{0\leq s\leq T}|\phi^\varepsilon(s,x)-x^\varepsilon(s)|\geq\eta\right].$$

Define

$$\theta^\varepsilon(t)=x^\varepsilon(t)-x-\int_0^t b^\varepsilon(x^\varepsilon(s))\,ds$$

$(=\sqrt{\varepsilon}\int_0^t\sigma^\varepsilon(x^\varepsilon(s))\,dw_s$ if x^ε is obtained from an Ito equation). Gronwall's inequality implies that

$$\sup_{[0,T]}|\phi^\varepsilon(s;x)-x^\varepsilon(s)|\leq e^{MT}\cdot\sup_{0\leq s\leq T}|\theta^\varepsilon(s)|,$$

where M is the Lipschitz constant for the $b^\varepsilon(\cdot)$ (uniform in ε). It is a standard argument, using exponential martingales, that

$$P_x^\varepsilon\left[\sup_{[0,T]}|\theta^\varepsilon(s)|\geq l\right]\leq(2d)\exp\left[\frac{-l^2}{2d\varepsilon AT}\right],$$

where $x^Ta^\varepsilon x\leq A\|x\|^2$; see [11, eq. (2.1), p. 87 and proof], for instance. Combining these facts, for all $x\in K$ and ε sufficiently small, we have

$$P_x^\varepsilon\left[\tau_\Omega\leq\tau\left(\frac{\gamma}{3}R\right)\right]\leq(2d)\exp\left[-\frac{1}{\varepsilon}\left(\frac{\eta^2e^{-2MT}}{2dAT}\right)\right].$$

This shows (4.7) and completes the proof.

Theorem 1 is now simple.

Proof of Theorem 1. Take any $x\in K$ and set $r_0=e^{-\delta/\varepsilon}$ ($\delta>0$ as in Theorem 2),

$$u^\varepsilon(x)=E_x^\varepsilon[u^\varepsilon(x^\varepsilon(\tau_\Omega(r_0)));\tau(r_0)<\tau_\Omega]+E_x^\varepsilon[f^\varepsilon(x^\varepsilon(\tau_\Omega));\tau_\Omega\leq\tau(r_0)].$$

Therefore,

$$u^\varepsilon(x)-u^\varepsilon(0)=E_x^\varepsilon\left[u^\varepsilon(x^\varepsilon(\tau_\Omega(r_0)))-u^\varepsilon(0);\tau(r_0)<\tau_\Omega\right]$$
$$+E_x^\varepsilon\left[f^\varepsilon(x^\varepsilon(\tau_\Omega))-u^\varepsilon(0);\tau_\Omega\le\tau(r_0)\right],$$
$$|u^\varepsilon(x)-u^\varepsilon(0)|\le\sup_{|y|\le r_0}|u^\varepsilon(y)-u^\varepsilon(0)|+2\sup_{\partial\Omega}|f^\varepsilon|\cdot P_x^\varepsilon\left[\tau_\Omega\le\tau(r_0)\right]$$
$$\le\sup_K|\nabla u^\varepsilon|\cdot r_0+2\sup_{\partial\Omega}|f^\varepsilon|e^{-\delta/\varepsilon}$$
$$\le C\varepsilon^{-1/2}e^{-\delta/\varepsilon}+2\sup_{\partial\Omega}|f^\varepsilon|e^{-\delta/\varepsilon}.$$

The theorem now follows (with a new slightly smaller δ).

5. Concluding remarks. We have two simple observations to make in closing. The first is regarding the case in which Ω contains several critical points of (1.1). This has been discussed in the literature [9], [12]. Both $u^\varepsilon\to$ constant and $u^\varepsilon\to$ a piecewise constant function are possibilities now, depending on the Vent-tsel–Freidlin variational distances between the critical points and $\partial\Omega$. If x^* is an asymptotically stable critical point (replacing the origin in (d) of §2) and $\Omega^*\subseteq\Omega$ is its domain of attraction, then by taking $f^\varepsilon=u^\varepsilon$ on $\partial\Omega^*$ we can apply Theorem 1 to see that leveling takes place exponentially fast in each such domain of attraction.

Finally, we observe that the specification of boundary data f^ε is actually superfluous. All that matters in the proof is the availability of a bound in ε for the u^ε. Theorem 1 could be formulated as follows:

For $K\subseteq\Omega$ compact there exist $\delta>0$ and $\varepsilon_0>0$ so that whenever $\mathcal{L}^\varepsilon[u]=0$ in Ω and $\varepsilon<\varepsilon_0$,

$$\sup_{x,y\in K}|u(x)-u(y)|\le e^{-\delta/\varepsilon}\sup_\Omega|u|. \tag{5.1}$$

Define the exit measures on the Borel subsets of $\partial\Omega$ by

$$\pi_x^\varepsilon(B)=P_x^\varepsilon\left[x^\varepsilon(\tau_\Omega)\in B\right].$$

The strong maximum principle implies that π_x^ε and π_y^ε are mutually absolutely continuous for $x,y\in\Omega$. Equation (6.1) implies that

$$\left|\int_{\partial\Omega}f(s)\left(1-\frac{d\pi_y^\varepsilon}{d\pi_x^\varepsilon}\right)\pi_x^\varepsilon\,ds\right|\le e^{-\delta/\varepsilon}\|f\|_{L^\infty(\pi_x^\varepsilon)}$$

for all f bounded and measurable on $\partial\Omega$. This is equivalent to

$$\left\|1-\frac{d\pi_y^\varepsilon}{d\pi_x^\varepsilon}\right\|_{L^1(\pi_x^\varepsilon)}\le e^{-\delta/\varepsilon}\quad\text{for } x,y\in K,\varepsilon<\varepsilon_0. \tag{5.2}$$

In cases for which a Green's function exists (if $\partial\Omega$ and all the coefficients are C^2, for instance) so that u^ε can be expressed as

$$u^\varepsilon(x)=\int_{\partial\Omega}G^\varepsilon(x,s)f^\varepsilon(s)\,ds,$$

then $\pi_x^\varepsilon(ds)=G^\varepsilon(x,s)\,ds$ on $\partial\Omega$ and (6.2) becomes, for $x,y\in K$,

$$\int_{\partial\Omega}|G^\varepsilon(x,s)-G^\varepsilon(y,s)|\,ds\le e^{-\delta/\varepsilon}. \tag{5.3}$$

REFERENCES

[1] A. DEVINATZ AND A. FRIEDMAN, *The asymptotic behavior of the solution of a singularly perturbed Dirichlet problem*, Indiana Univ. Math. J., 27 (1978), pp. 527–537.

[2] _____, *Asymptotic behavior of the principal eigenfunction for a singularly perturbed Dirichlet problem*, Indiana Univ. Math. J., 27 (1978), pp. 143–157.

[3] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer, New York, 1977.

[4] S. KAMIN, *Elliptic perturbation of a first order operator with a singular point of attracting type*, Indiana Univ. Math. J., 27 (1978), pp. 935–952.

[5] _____, *On elliptic singular perturbation problems with turning points*, this Journal, 10 (1979), pp. 447–455.

[6] D. LUDWIG, *Persistence of dynamical systems under random perturbations*, SIAM Rev., 17 (1975), pp. 605–640.

[7] H. P. MCKEAN, JR., *Stochastic Integrals*, Academic Press, New York, 1969.

[8] B. J. MATKOWSKY AND Z. SCHUSS, *The exit problem for randomly perturbed dynamical systems*, SIAM J. Appl. Math., 33 (1977), pp. 365–382.

[9] _____, *The exit problem: A new approach to diffusion across potential barriers*, SIAM J. Appl. Math, 36 (1979), pp. 604–623.

[10] Z. SCHUSS, *Singular perturbation methods in stochastic differential equations of mathematical physics*, SIAM Rev., 22 (1980), pp. 119–155.

[11] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer, Berlin, 1979.

[12] A. D. VENT-TSEL AND M. I. FREIDLIN, *On small random perturbations of dynamical systems*, Uspekhi Mat. Nauk., 25 (1970), pp. 1–56; =Russian Math. Surveys, 25 (1970), pp. 1–55.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0003 $01.00/0

TURNING-POINT CONNECTION AT CLOSE QUARTERS*

J. F. PAINTER† AND R. E. MEYER‡

Abstract. WKB-connection theory across singular and turning points of linear second order differential equations is extended from Langer's [1931], [1932] class of fractional turning points to a larger one including logarithmic turning points and even more irregular, singular points. Uniform approximations of solutions then become necessarily much less tractable and useful, but this is shown not to affect the first order WKB-connection formulae of primary practical importance. The extension is achieved by a new connection method which abandons the reliance on progressive paths to control the rapidly varying exponentials.

Introduction. A central problem for the general linear second-order equation

$$\varepsilon^2 \frac{d^2w}{dz^2} + q^2 w(z) = 0, \qquad q = q(z), \quad \varepsilon = \text{const} \downarrow 0 \tag{1.1}$$

has long been known to be that of connection across roots and singular points of $q(z)$ (which is independent of ε and is assumed analytic on an appropriate domain, see §2). The solutions have asymptotic "WKBJ" or "Liouville–Green" approximations (Olver [1974])

$$\begin{aligned} q^{1/2}w &\sim A\exp\left(\frac{i}{\varepsilon}\int^z q(s)\,ds\right) + B\exp\left(\frac{-i}{\varepsilon}\int^z q(s)\,ds\right), \\ -i\varepsilon q^{-1/2}w' &\sim A\exp\left(\frac{i}{\varepsilon}\int^z q(s)\,ds\right) - B\exp\left(\frac{-i}{\varepsilon}\int^z q(s)\,ds\right) \end{aligned} \tag{1.2}$$

with constants A, B in appropriate regions of the z-plane, but such regions do not include roots or singular points of $q(z)$. The constant (A, B) may therefore have different values on different sides of such a transition point, and the connection problem may be phrased as that of determining the value on the right, given the value on the left.

Langer [1931], [1932], [1935] solved this problem for the class of "fractional turning points" (Olver [1977]) for which $z^{-\nu}q(z)$ is analytic and nonzero at the transition point (taken as $z=0$) with real constant $\nu > -1$. But, if branch points of $q(z)$, even of infinite order, be thus admitted, why not logarithmic branch points? The limitation of Langer's "central-connection method" is intimately related to its triumph: he showed that Bessel functions provide approximations to the solutions of (1.1) as $\varepsilon \to 0$ that are *uniformly* valid both near the transition point and away from it. This provides full insight and information, and (1.2) emerges from the asymptotic properties of those Bessel functions. The fractional turning points, however, stand in one–one correspondence to the Bessel functions of real order, and logarithmic turning points, e.g., do not (Painter [1979]) admit uniform approximations of such type because the solutions have

* Received by the editors June 18, 1980, and in revised form August 3, 1981. This research was sponsored by the U.S. Army under contract No. DAAG29-80-C-0041 and partially supported by the National Science Foundation under grant MCS77-00097 and by the Wisconsin Alumni Research Foundation. It is published by the U.S. Department of Energy under contract W-7405-Eng-48.

† Lawrence Livermore Laboratory, Livermore, California 94550.

‡ Mathematics Department and Research Center, University of Wisconsin, Madison, Wisconsin, 53706.

branch points of different structure. Uniform approximands must then be anticipated to be much less familiar, tractable and useful, and no attempt to find them is made in the following because the most important applications, such as in scattering, do not require any of the details of information provided by uniform approximation. What they do require is connection, and especially, a first approximation to it, and this will be shown not to depend on the details.

From (1.2), the WKB (or Langer) variable

$$\xi=\int^z q(s)\,ds$$

is seen to be a more natural, independent variable than z for a study of (1.1). When (1.1) is a time-reduced wave or Schrödinger equation, $\varepsilon^{-1}\xi$ measures distance in local wavelengths; for an oscillator equation, it measures time in local periods. For the connection problem to be fully meaningful, $q(z)$ should be absolutely integrable, even at the transition point $z=0$, and this will be assumed here. Then the intrinsic variable can be defined as

$$\xi=\int_0^z q(s)\,ds. \tag{1.3}$$

The new connection method to be presented starts from the replacement of (1.2) by the definition (Kemble [1935], Fröman and Fröman [1965])

$$\begin{aligned} q^{1/2}w(z)&=A(\xi)e^{i\xi/\varepsilon}+B(\xi)e^{-i\xi/\varepsilon},\\ -i\varepsilon q^{-1/2}\frac{dw}{dz}&=A(\xi)e^{i\xi/\varepsilon}-B(\xi)e^{-i\xi/\varepsilon} \end{aligned} \tag{1.4}$$

of "modulation coefficients" $A(\xi)$ and $B(\xi)$ for (1.1) of which the circuits as $\varepsilon\to 0$ are sought which are the WKB-constants in (1.2). Equations (1.1) and (1.4) imply

$$\frac{dA}{d\xi}=\phi(\xi)B(\xi)e^{-2i\xi/\varepsilon}, \tag{1.5}$$

$$\frac{dB}{d\xi}=\phi(\xi)A(\xi)e^{2i\xi/\varepsilon}, \tag{1.6}$$

where the "modulation function" is

$$\phi(\xi)=\frac{1}{2}q^{-2}\frac{dq}{dz}. \tag{1.7}$$

This canonical formulation helps to identify the parameter really governing connection: all commonly encountered roots or (integrably) singular points of $q(z)$ satisfy

$$\phi(\xi)=\xi^{-1}(\gamma+o(1)) \quad \text{as } |\xi|\to 0 \tag{A}$$

uniformly in $\arg\xi$ with real constant $\gamma<\frac{1}{2}$ which will therefore be adopted as the main assumption here. The fractional turning points (Olver [1977]), for example, satisfy (A) with $\gamma=\frac{1}{2}\nu/(1+\nu)$ and so do "logarithmic turning points" for which $q(z)\sim z^{\nu}(\log z)^{\mu}$ with $\nu>-1$, $\mu>0$. (Details of this and other relatively simple technical points may be found in Painter and Meyer [1980].) The canonical formulation (1.4) is thus seen to relegate to an error term in (A) logarithms and other multivalued functions that had seemed fundamentally different in the central-connection version (Olver, [1977]) of the irregular singular point of (1.1). Of course, the error term in (A) may be disagreeable, it is only $O(|\log\xi|^{-1})$ in the logarithmic case, and many technical difficulties in the following stem from the fact that (A) admits errors tending to zero much more slowly still and barely characterized.

To solve connection in such generality, neither the lateral-connection method (Zwaan [1929]) nor the central-connection method (Langer [1931]) can serve, and a main objective of the following is to show that the arsenal of methods is not nearly as exhausted by them as the literature of the last three decades might seem to indicate. Lateral connection (Evgrafov and Fedoryuk [1966]) operates too far from the transition point to make good use of the singularity structure of the modulation function (1.7), on which connection depends. Central connection, on the other hand, gets too intimately involved in the local solution properties at the nonintegrable, poorly defined singularity of $\phi(\xi)$, where the good properties of the exponentials in (1.4) cannot be used for estimates. The new method presented below compromises by connecting along a semi-circle

$$|\xi| = \text{const} = \delta(\varepsilon)$$

such that

$$\delta(\varepsilon) \to 0, \text{ but } \delta(\varepsilon)\varepsilon^{-1} \to \infty \text{ as } \varepsilon \to 0 \tag{1.8}$$

in order to profit from both singularity structure and WKB-exponentials. The price is abandonment of the traditional key feature of turning point proofs to employ only "progressive paths" (Olver [1974]) on which the exponential kernels can be controlled in terms of their magnitude at an endpoint. A first departure from this absolute rule occurs in Olver [1978], where a nonprogressive path's length and distance from the turning point are made so small that the exponentials are nicely bounded. But in the following, control of the nonmonotone and wild variation of the exponentials in (1.5), (1.6) on the nonprogressive path $|\xi| = \delta$ becomes the central feature of the connection proof.

The method of proof will be traditional in rephrasing (1.1) as a Volterra integral equation to be contracted, which is also the normal method to generate error bounds (Olver [1974]), but we feel it is premature to consider this aspect here.

2. Integral equation. The connection problem arises in many equivalent or related forms. For definiteness, let us here take it to be that of connecting the limits of $A(\xi)$ and $B(\xi)$ at fixed points $\xi = -\Xi$ on the left with those at $\xi = \Xi$ on the right of the singular point $\xi = 0$, when the branch of $\phi(\xi)$ in (A) is defined by a cut along the negative imaginary axis of ξ so that connection must be made through the upper half-plane. Then our main result is the

THEOREM. *Let* (1.1) *hold on an open connected simply connected set* $R_z \subset C$, *with* $q(z)$ *analytic and nonzero on* R_z *and on its boundary except at a boundary point* $z = 0$. *At* $z = 0$, *let* $q(z)$ *be absolutely integrable. Let the image set* $R_\xi = \xi(R_z)$ *under the map* (1.3) *contain an* ε*-independent rectangular neighborhood of* $z = 0$ (*Fig.* 1), *except for a branch cut along the negative imaginary axis. If* $\operatorname{Im}\xi$ *has no lower bound on* R_ξ, *then let* $\phi(\xi) = \frac{1}{2}q^{-2}\frac{dq}{dz}$ *be of exponential order as* $\xi \to -i\infty$.

Suppose that $\phi(\xi)$ *satisfies*

$$\phi(\xi) = \xi^{-1}\left[\gamma + O\left(1/\log^{(n)}|\xi|^{-1}\right)\right] \quad \text{as } |\xi| \to 0, \tag{A'}$$

where $\log^{(n)}$ *denotes any finite number of iterations of the logarithm. Then as* $\varepsilon \to 0$,

$$\begin{aligned} &\text{(a)} \quad A(\xi_r) = A(\xi_l) - 2i\sin(\gamma\pi)B(\xi_l) + o(1), \\ &\text{(b)} \quad B(\xi_r) = B(\xi_l) + o(1), \end{aligned} \tag{2.1}$$

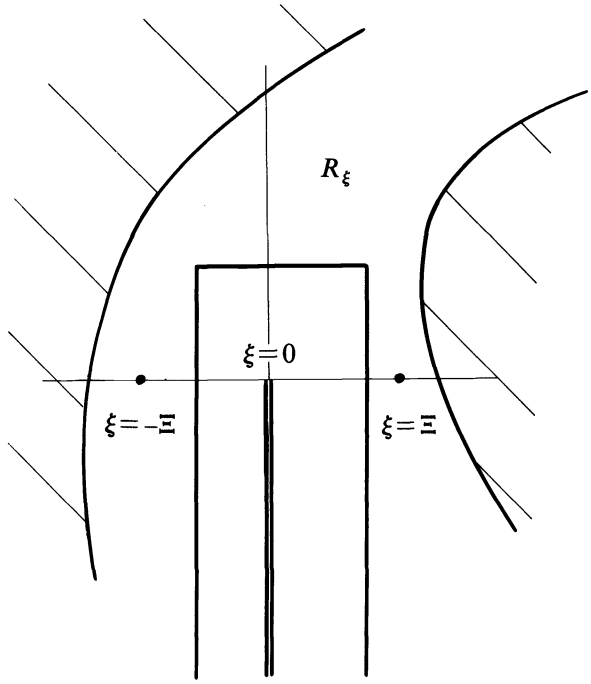

FIG. 1

where γ *is given by* (A′) *or* (A) (§1), ξ_r *is any point between* $\delta(\varepsilon)$ *and* Ξ, ξ_l *is any point between* $-\delta(\varepsilon)$ *and* $-\Xi$, $\delta(\varepsilon)$ *is an arbitrary function satisfying* (1.8) *and* (3.4), *and* Ξ *is an arbitrary point* (*fixed as* $\varepsilon\to 0$) *for which the intervals* $(-\Xi,0)$ *and* $(0,\Xi)$ *lie in* R_ξ. *If* $\phi(\xi)$ *satisfies* (A) *but not* (A′), *then* (2.1) *holds with* $\xi_r=\delta(\varepsilon)$ *and* $\xi_l=-\delta(\varepsilon)$.

A simple example of an equation to which this theorem applies is

$$\varepsilon^2 w''+(z\log z)^2 w=0.$$

Suppose that we know w and w' at $z=\frac{1}{2}\exp(i0)$ and need to compute them at $z=\frac{1}{2}\exp(-i\pi)(=-\frac{1}{2}$, briefly in what follows). In terms of the WKBJ-variable $\xi(z)=\frac{1}{2}z^2(\log z-\frac{1}{2})$, we are connecting from $\xi(\frac{1}{2})=\frac{1}{8}(\frac{1}{2}+\log 2)\exp(i\pi)$ to $\xi(-\frac{1}{2})=\frac{1}{8}(\frac{1}{2}+\log 2+i\pi)\exp(-i\pi)$.

Since $\xi(-\frac{1}{2})$ has a negative imaginary part, (1.2) implies that as $\varepsilon\to 0$,

$$\begin{aligned} w\left(-\frac{1}{2}\right)&=(A+o(1))q^{-1/2}\left(-\frac{1}{2}\right)\exp\left(\frac{i\xi(-1/2)}{\varepsilon}\right),\\ w'\left(-\frac{1}{2}\right)&=(A+o(1))i\varepsilon^{-1}q^{1/2}\left(-\frac{1}{2}\right)\exp\left(\frac{i\xi(-1/2)}{\varepsilon}\right). \end{aligned} \tag{2.2}$$

The approximation (1.2) holds in a ξ-region that includes both $\xi(-\frac{1}{2})$ and positive real values of ξ such as $\xi=\frac{1}{8}$, so we may take $A=A(\frac{1}{8})$ in (2.2) and it remains to find $A(\frac{1}{8})$. But it is easy to verify that the hypotheses of the theorem hold with $\xi_r=\frac{1}{8}$, $\xi_l=\xi(\frac{1}{2})$ and

$$R_\xi=\left\{\xi(z):0<|z|<\frac{3}{4},-\frac{\pi}{2}<\arg\xi<\frac{3\pi}{2}\right\}.$$

Therefore,

$$A\left(\frac{1}{8}\right)=A\left(\xi\left(\frac{1}{2}\right)\right)-i\sqrt{2}\,B\left(\xi\left(\frac{1}{2}\right)\right)+o(1)$$

and A, B at $\xi(\frac{1}{2})$ can be computed with (1.4).

The theorem will be proven in the rest of this paper. Assumption (A) rather than (A′) will suffice until the end of §4. Equation (2.1)(b) presents no problem: in the upper half-plane where connection is made, B is the coefficient of the dominant term in (1.4), and a well-known asymptotic argument (Evgrafov and Fedoryuk [1966]) gives

$$B(1)=B(-1)+o(1) \quad \text{as } \varepsilon\to 0, \tag{2.3}$$

the difficulty lies entirely in relating $A(\Xi)$ to $A(-\Xi)$ and $B(-\Xi)$. To attack it, integrate (1.5) by parts and use (1.6) to get

$$A(\xi)-A_l=B_l j(\xi,\xi_l)e^{-\rho\xi_l}+\int_{\xi_l}^{\xi}A(s)j(\xi,s)\phi(s)\,ds, \tag{2.4}$$

where

$$\rho=\frac{2i}{\varepsilon}, \quad A_l=A(\xi_l), \quad B_l=B(\xi_l),$$
$$j(\xi,s)=\int_s^{\xi}e^{\rho(s-t)}\phi(t)\,dt, \tag{2.5}$$

and the path of integration is chosen along the upper semicircle $|\xi|=\delta(\varepsilon)$ from $\xi_l=-\delta$; on it, $|\rho\xi|=|\rho s|=\frac{2\delta}{\varepsilon}\to\infty$ by (1.8).

The computation of the kernel $j(\xi,s)$ on this nonprogressive path is one of the main hurdles, and the Appendix summarizes a method of estimating it adequately by computing (2.5) on paths depending on $\arg\xi$ and $\arg s$ and chosen carefully to afford simultaneous control of $\exp(-\rho t)$ and $\phi(t)$. The result, valid uniformly as $\varepsilon\to 0$ for ξ and s on respective arcs of the circle $|\xi|=|s|=\delta(\varepsilon)$ (Fig. 2), is

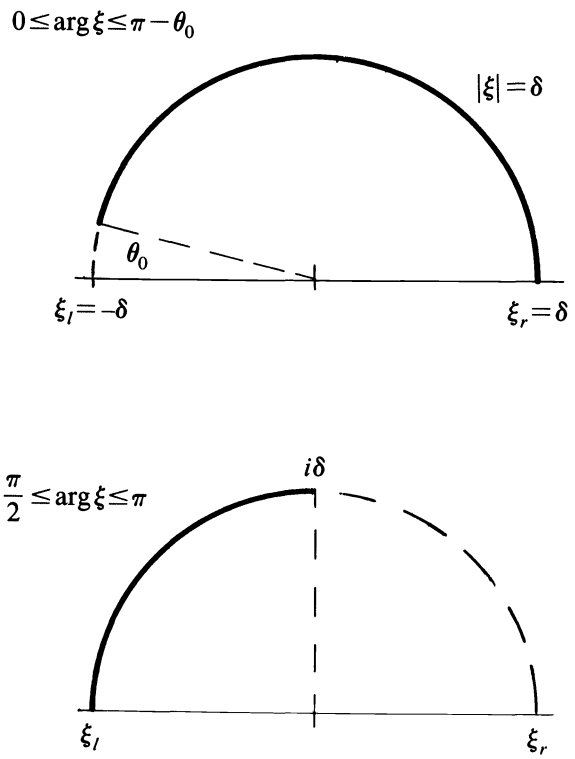

FIG. 2

$$j(\xi,s)e^{-\rho s}=(\gamma+o(1))\left[\frac{e^{-\rho s}}{\rho s}-\frac{e^{-\rho\xi}}{\rho\xi}\right] \tag{2.6}$$

for $0\le\arg\xi$, $\arg s\le\pi-\theta_0$, where $\theta_0>0$ is an arbitrary constant

$$j(\xi,\xi_l)\exp(-\rho\xi_l)=(\gamma+o(1))\left[\frac{e^{-\rho\xi_r}}{\rho\xi_r}-\frac{e^{-\rho\xi}}{\rho\xi}-2\pi i\right] \tag{2.7}$$

for $0\le\arg\xi\le\pi-\theta_0$, and

$$j(\xi,s)e^{-\rho s}=O\left(\frac{e^{-\rho\xi}}{(\rho\xi)}\right) \tag{2.8}$$

for $\frac{\pi}{2}\le\arg\xi\le\arg s\le\pi$.

To use these estimates in the integral equation, it helps to split off the main exponential factor in $A(\xi)$ by defining

$$a(\xi)=A(\xi)e^{\rho\xi},$$

so that (2.4) becomes

$$a(\xi)-a_l e^{\rho\xi-\rho\xi_l}=B_l j(\xi,\xi_l)e^{\rho\xi-\rho\xi_l}+\int_{\xi_l}^{\xi}a(s)j(\xi,s)e^{\rho\xi-\rho s}\phi(s)\,ds. \tag{2.9}$$

For ξ in the upper left quarter-circle (Fig. 2), where $\mathrm{Re}(\rho\xi)\le\mathrm{Re}(\rho s)\le\mathrm{Re}(\rho\xi_l)=0$, (2.6) and (1.8) show the integral operator in (2.7) to contract and thereby to assure a bounded solution $a(\xi)$ of (2.9) given the constants A_l and B_l. Therefore,

$$A(\xi)-O(e^{-\rho\xi})=O(e^{2\delta/\epsilon}), \tag{2.10}$$

and by full use of (2.7) to (2.9), the contraction argument and bound (2.10) can be extended to the entire semicircle $|\xi|=\delta$, $0\le\arg\xi\le\pi$ (Fig. 2) (Painter & Meyer [1980]).

This existence proof for a nonprogressive path, even if constructive, is disappointing because the potentially large bound (2.8) is insufficient to support a sequence of asymptotic approximations for $A(\delta)=A_r$ at the right end of the semicircle. The essential fault stems from the lack of definite information on the error term in (A); by the Dubois–Reymond lemma, (A) sets no definite limit to how slowly that term can decrease with $|\xi|$.

3. Bessel comparison. To refine the proof, recourse may be had to Langer's [1931] idea that turning-point theory concerns generalizations of Bessel functions. In the simplest special case, $q(z)=z^\nu$, the modulation function (1.7) is $\phi^\gamma(\xi)=\gamma/\xi$ exactly and (1.1) is a form of Bessel's equation with solutions

$$w^\gamma(z)=\xi^\lambda H_\lambda^{(i)}\left(\frac{\xi}{\epsilon}\right),\qquad \lambda=\frac{1}{2}-\gamma, \tag{3.1}$$

in terms of Hankel functions $H_\lambda^{(i)}$, $i=1,2$. From their asymptotic properties (Olver [1974]), the limit A_r^γ of $A(\xi)$ at $\xi_r=\delta$ for this special case can be computed straightforwardly to be

$$A_r^\gamma=A_l-2iB_l\sin(\gamma\pi)+o(1). \tag{3.2}$$

This special case $A^\gamma(\xi)$ of the modulation coefficient $A(\xi)$ satisfies (2.4) with ϕ and j replaced by ϕ^γ and

$$j^\gamma(\xi,s)=\gamma\int_s^\xi e^{\rho s-\rho t}t^{-1}\,dt,$$

respectively, and (2.4) can be rephrased as an integral equation for $A(\xi)-A^\gamma(\xi)$,

$$A(\xi)-A^{\gamma}(\xi)=B_l\big[j(\xi,\xi_l)-j^{\gamma}(\xi,\xi_l)\big]e^{-\rho\xi_l}$$
$$+\int_{\xi_l}^{\xi}A^{\gamma}(s)\big[j(\xi,s)\phi(s)-\gamma s^{-1}j^{\gamma}(\xi,s)\big]\,ds \tag{3.3}$$
$$+\int_{\xi_l}^{\xi}\big[A(s)-A^{\gamma}(s)\big]\,j(\xi,s)\phi(s)\,ds.$$

Now, $A(\xi)$ and $A^{\gamma}(\xi)$ are not generally close, e.g., they have different types of branch points at $\xi=0$, but there is a small subfamily of the semicircles (1.8) characterized by the condition

$$\varepsilon\delta^{-1}e^{2\delta/\varepsilon}\sup\{\xi\phi(\xi)-\gamma:|\xi|\le\delta\}\to 0 \quad \text{as } \varepsilon\to 0 \tag{3.4}$$

on which the estimates can be improved. They are very small semicircles (typically we might have $\delta(\varepsilon)\sim\varepsilon\log|\log\varepsilon|$ as $\varepsilon\to 0$), but whenever $\phi(\xi)$ satisfies (A), the subfamily is not empty. On them, $j-j^{\gamma}$ can be computed by a modification of the calculation outlined in the Appendix to obtain

$$j(\xi,s)\sim j^{\gamma}(\xi,s) \tag{3.5}$$

for ξ and s in the same arc as in (2.4) (Fig. 2),

$$j(\xi,\xi_l)\sim j^{\gamma}(\xi,\xi_l) \tag{3.6}$$

for ξ in the same arc as in (2.5) (Fig. 2), and

$$j(\xi,s)=j^{\gamma}(\xi,s)+o(1) \tag{3.7}$$

for ξ and s in the upper left quarter-circle (Fig. 2).

With these estimates, together with (2.6)–(2.8) and the bound (2.10) on $A^{\gamma}(\xi)$, (3.3) can be reduced to

$$A(\xi)-A^{\gamma}(\xi)=o(1)+\int_{\xi_l}^{\xi}\big[A(s)-A^{\gamma}(s)\big]\,j(\xi,s)\phi(s)\,ds \tag{3.8}$$

on the semicircle $|\xi|=\delta$, $\operatorname{Im}\xi>0$ (Fig. 2) subject to (1.8) and (3.4). Since the integral operator is the same as in (2.4), (3.8) is solved by the same contraction (§2) in the upper left quarter-circle (Fig. 2) and there yields

$$A(\xi)-A^{\gamma}(\xi)\to 0 \quad \text{as } \varepsilon\to 0. \tag{3.9}$$

For the upper right quarter circle, (3.9) and the equations (A), (2.6)–(2.8) and (3.5)–(3.7) may be used to rewrite (3.8) as

$$A(\xi)-A^{\gamma}(\xi)=o(1)+\int_{i\delta}^{\xi}\big[A(s)-A^{\gamma}(s)\big]\,j(\xi,s)\phi(s)\,ds.$$

On this part of the path, the integrand is susceptible to (2.6) so that (A) and (3.4) show the integral operator to be a contraction and thereby to assure (3.9) uniformly for ξ in the whole semicircle $|\xi|=\delta$, $\operatorname{Im}\xi\ge 0$ (Fig. 2).

It now follows that the connection formula (3.2) carries over to any $A(\xi)$ subject to (A), on these semicircles. A further corollary is that, while the solutions of (1.1) cannot be approximated uniformly by Bessel functions in the framework of (A), they can be so approximated in the annulus specified by (1.8) and (3.4).

4. Connection gap. For practical usefulness connection is needed between points bounded from the singular one independently of ε, say between $\xi=\Xi=-1$ and $\xi=\Xi=1$

and that between $\xi_l=-\delta(\varepsilon)$ and $\xi_r=\delta(\varepsilon)$ achieved in §3 falls seriously short of that goal. The gap from $\xi=-1$ to $\xi=-\delta(\varepsilon)$, subject to (1.8) and (3.4), still needs to be covered; the dual one from $\delta(\varepsilon)$ to 1 then poses no new problem.

We begin by showing that $A(\xi)$ is nearly constant on sufficiently short ξ-intervals $[-\delta_2(\varepsilon), -\delta_1(\varepsilon)]$ with $\varepsilon<\delta(\varepsilon)\le\delta_1\le\delta_2\le 1$. The integral equation (2.4) applies here with $-\delta_2$ in place of ξ_l and with $\xi\in(-\delta_2, -\delta_1)$. Integration by parts of (2.5) gives

$$j(\xi,s)e^{-\rho s}=\rho^{-1}\left[e^{-\rho s}\phi(s)-e^{-\rho\xi}\phi(\xi)\right]+\rho^{-1}\int_s^{\xi}e^{-\rho t}\phi'(t)\,dt,$$

and by (A) the first term on the right-hand side is bounded by $\frac{\varepsilon}{2\delta}(\gamma+o(1))$. Since $\phi(\xi)$ is holomorphic near zero except for the branch cut, (A) also implies

$$\phi'(t)=O(t^{-2}) \quad \text{for real } t\to 0, \tag{4.1}$$

and therefore

$$j(\xi,s)e^{-\rho s}=O\left(\frac{\varepsilon}{\delta_1}\right).$$

It follows from (2.4) that, for some constant C and the supremum norm on $(-\delta_2, -\delta_1)$,

$$|A(\xi)-A(-\delta_2)|\le C\varepsilon\delta_1^{-1}\left[1+\|A\|\int_{-\delta_2}^{\xi}|s^{-1}\,ds|\right]\le C\varepsilon\delta_1^{-1}\left[1+\|A\|\log\left(\frac{\delta_2}{\delta_1}\right)\right].$$

That implies a bound on $A(\xi)$ between $-\delta_2$ and $-\delta_1$ independent of ε, provided δ_1 and δ_2 are so close that

$$\varepsilon\delta_1^{-1}\log\left(\frac{\delta_2}{\delta_1}\right)\to 0 \quad \text{as } \varepsilon\to 0, \tag{4.2}$$

and it then follows that A is approximately constant on $[-\delta_2, -\delta_1]$.

To determine intervals over which $B(\xi)$ will also be nearly constant, integrate (1.6) by parts,

$$B(\xi)-B(-\delta_2)=\int_{-\delta_2}^{\xi}\left[A(s)-A(-\delta_2)\right]e^{\rho s}\phi(s)\,ds$$
$$+\rho^{-1}A(-\delta_2)\left\{e^{\rho\xi}\phi(\xi)-e^{-\rho\delta_2}\phi(-\delta_2)-\int_{-\delta_2}^{\xi}e^{\rho s}\phi'(s)\,ds\right\},$$

and apply (A), (4.1) and (4.2) to obtain

$$|B(\xi)-B(-\delta_2)|\le C\varepsilon\delta_1^{-1}\left[1+\left\{\log\left(\frac{\delta_2}{\delta_1}\right)\right\}^2\right].$$

Approximate constancy of B on $[-\delta_2, -\delta_1]$ is therefore assured by a strengthening of (4.2) to

$$\varepsilon\delta_1^{-1}\left\{\log\left(\frac{\delta_2}{\delta_1}\right)\right\}^2\to 0 \quad \text{as } \varepsilon\to 0. \tag{4.3}$$

Unfortunately, (4.3) is too restrictive for a direct bridge across the gap from -1 to $-\delta(\varepsilon)$. The estimates just sketched will, however, bridge the gap from -1 to $-\delta_1(\varepsilon)$ if a

finite sequence

$$\delta_1(\varepsilon)<\delta_2(\varepsilon)<\delta_3(\varepsilon)<\cdots<\delta_N(\varepsilon)=1 \tag{4.4}$$

can be found for which each pair δ_i, δ_{i+1} satisfies

$$\varepsilon\delta_i^{-1}\left\{\log\left(\frac{\delta_{i+1}}{\delta_i}\right)\right\}^2\to 0 \quad \text{as } \varepsilon\to 0. \tag{4.5}$$

It is not plausible that (A), which admits an error term decreasing arbitrarily slowly, can assure existence of such a sequence with $\delta_1(\varepsilon)=\delta(\varepsilon)$ subject to (3.4), and a counterexample with error term in (A) approaching zero more slowly then the reciprocal of any finite iterate $\log\log\cdots\log(|\xi|^{-1})$ has been constructed (Painter [1979]). Accordingly, (A) must be slightly strengthened to (A′) (§2).

This new assumption does not exclude any common examples; for a logarithmic turning point $q(z)\sim z^\nu(\log z)^\mu$, e.g., (A′) holds with $n=1$. It is also sufficient: (3.4) may be satisfied with

$$\delta(\varepsilon)=\delta_1(\varepsilon)=\varepsilon\log^{(n+2)}\left(\frac{1}{\varepsilon}\right),$$

and the sequence (4.4) may be chosen as

$$\begin{aligned}
\delta_2(\varepsilon)&=\varepsilon\exp\left\{\log^{(n+2)}\left(\frac{1}{\varepsilon}\right)\right\}^{1/4},\\
\delta_3(\varepsilon)&=\varepsilon\left\{\log^{(n+2)}\left(\frac{1}{\varepsilon}\right)\right\}^{3},\\
\delta_4(\varepsilon)&=\varepsilon\exp\left\{\log^{(n+1)}\left(\frac{1}{\varepsilon}\right)\right\}^{1/4},\\
&\vdots\\
\delta_{2n+3}(\varepsilon)&=\varepsilon\left\{\log\left(\frac{1}{\varepsilon}\right)\right\}^{3},\\
\delta_{2n+4}&=1,
\end{aligned}$$

which satisfies (4.5). Hence (A′) suffices to support the connection formula

$$A_r=A_l-2iB_l\sin(\gamma\pi)+o(1), \qquad B_r=B_l+o(1).$$

Appendix. Properties and error bounds of the kernel. From (2.5) and (A),

$$j(\xi,s)e^{-\rho s}=j(s)e^{-\rho s}-j(\xi)e^{-\rho\xi}, \tag{A1}$$

where

$$j(\xi)e^{-\rho\xi}=\int_\xi^{-i\infty}e^{-\rho s}\phi(s)\,ds \tag{A2}$$

and

$$\phi(\xi)=\xi^{-1}(\gamma+\tilde g(\xi))$$

with an "error seed" function $\tilde g$ specified only to the extent that it is analytic, except for the branch cut, and of exponential order as $\xi\to -i\infty$ and that

$$g(\delta)=\sup\{|\tilde g(\xi)|:|\xi|\le\delta\}\to 0 \quad \text{as } \delta\to 0. \tag{A3}$$

To derive the bounds (2.6)–(2.8) resolves itself largely into estimating the contributions of $\tilde{g}(s)$ in (A2) in the presence of the volatile factor $\exp(-\rho s)$.

If $\tilde{g}$ could be ignored, $j(\xi)\exp(-\rho\xi)$ would be the incomplete gamma function (Olver [1974]) which has a Stokes phenomenon at $\arg\xi=\pi$ so that different calculations are needed for ξ close to ξ_l and for ξ bounded away from the negative real axis, say $0\leq\arg\xi\leq\pi-\theta_0$ with any fixed $\theta_0>0$ (Fig. 2). In either case only points

$$\xi=\xi_R+i\xi_i$$

with real ξ_R, ξ_i need be considered that lie on the semicircle $|\xi|=\delta(\varepsilon)\to 0$, $\xi_i\geq 0$.

(i) The case $0\leq\arg\xi\leq\pi-\theta_0$ (Fig. 2) admits a "progressive" path Λ of integration on which the exponential is monotone in magnitude, but which keeps its distance from the origin, where $\phi(s)$ is not integrable, and yet stays largely within the circle $|s|=\delta(\varepsilon)\to 0$ where (A) applies. For definiteness, it will be chosen to run (Fig. 3) from ξ vertically down to

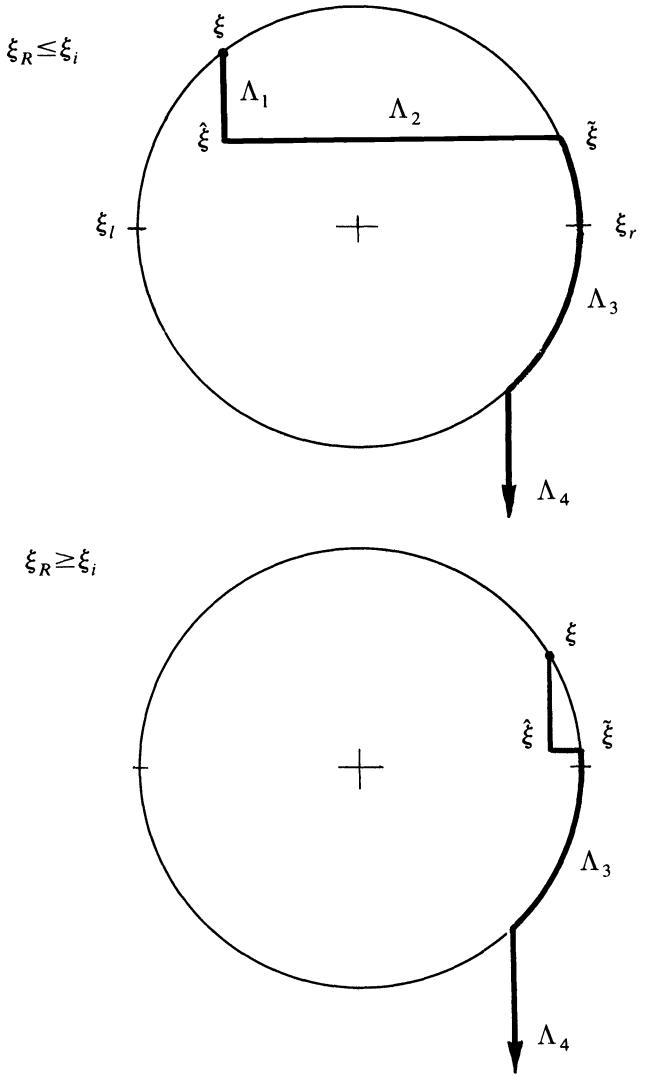

FIG. 3

(A4) $$\hat{\xi}=\xi_R+i\hat{\xi}_i \quad \text{with } \hat{\xi}_i=\min\left(\frac{1}{2}\hat{\xi}_i,\xi_i-\frac{1}{2}\xi_R\right),$$

then horizontally to the circle at $\tilde{\xi}=\tilde{\xi}_R+i\tilde{\xi}_i$, $|\tilde{\xi}|=\delta$, then around the circle to the lower half-plane and finally down to $-i\infty$. Thus,

(A5) $$\left|j(\xi)e^{-\rho\xi}-\gamma\int_{\rho\xi}^{-i\infty}e^{-u}\frac{du}{u}\right|\le g(\delta)\sum_{j=1}^{3}\int_{\Lambda_j}\left|e^{-\rho s}\frac{ds}{s}\right|+\int_{\Lambda_4}\left|e^{-\rho s}\tilde{g}(s)\frac{ds}{s}\right|,$$

where $\Lambda_1,\cdots,\Lambda_4$ are the four segments of Λ (Fig. 3). It will be shown now that the "error terms" on the right are smaller than the incomplete gamma function, approximable (Olver [1974, p. 110]) by

(A6) $$\int_{\rho\xi}^{-i\infty}e^{-u}\frac{du}{u}\sim(\rho\xi)^{-1}e^{-\rho\xi}$$

as $\varepsilon\to 0$, uniformly for ξ in the arc $|\xi|=\delta$, $0\le\arg\xi\le\pi-\theta_0$ (Fig. 2). Λ_1 is a path of steepest descent of the exponential factor and integration by parts gives

$$\int_{\Lambda_1}\left|e^{-\rho s}\frac{ds}{s}\right|=|e^{-\rho\xi}|\left||\rho\xi|^{-1}-|\rho\hat{\xi}|^{-1}\exp\left[i\rho\left(\xi_i-\hat{\xi}_i\right)\right]+\frac{1}{|\rho|}\int_0^{\xi_i-\hat{\xi}_i}e^{i\rho t}\frac{(\xi_i-t)\,dt}{|\xi-it|^3}\right|.$$

Since $\hat{\xi}/\delta$ is bounded from zero, the first two terms in the last sum are $O(\varepsilon/\delta)$, and since $\delta^2(\xi_i-t)/|\xi-it|^3$ is bounded on Λ_1, the last term is $O(\varepsilon^2/\delta^2)$. Therefore,

$$\int_{\Lambda_1}\left|e^{-\rho s}\frac{ds}{s}\right|\le C|e^{-\rho\xi}|\varepsilon/\delta$$

for some constant C, which is a bound of the magnitude of the incomplete gamma function (A6), and the factor $g(\delta)$ on the right-hand side of (A5) makes the contribution from Λ_1 small by comparison.

That from Λ_2 (Fig. 3) can be evaluated explicitly as

$$\int_{\Lambda_2}\left|e^{-\rho s}\frac{ds}{s}\right|=e^{-\rho i\hat{\xi}_i}\left|\log\left[\frac{\tilde{\xi}_R+|\tilde{\xi}|}{\xi_R+|\hat{\xi}|}\right]\right|$$

and Λ_2 has been chosen far enough from both ξ and 0 so that the logarithmic term is uniformly bounded and

(A7) $$\xi_i-\hat{\xi}_i\ge 2^{-3/2}\delta.$$

This integral is therefor small even by comparison with $\frac{\varepsilon}{\delta}\exp(-\rho\xi)$, when $|\xi|=\delta$ and $0\le\arg\xi\le\pi-\theta_0$, and the same follows from (A7) for

$$\int_{\Lambda_3}\left|e^{-\rho s}\frac{ds}{s}\right|.$$

For the last term in (A5), the assumption that $\phi(s)$ grows at most exponentially as $s\to -i\infty$ assures convergence and descent, and for sufficiently small ε, the upper endpoint of Λ_4 (Fig. 3) becomes decisive and, also,

$$\int_{\Lambda_4}|e^{-\rho s}\tilde{g}(s)\,ds|\ll\left(\frac{\varepsilon}{\delta}\right)e^{-\rho\xi}.$$

Hence,

$$j(\xi)e^{-\rho\xi}=\frac{e^{-\rho\xi}}{\rho\xi}(\gamma+o(1)) \tag{A8}$$

as $\varepsilon\to 0$, uniformly for ξ on the arc $0\leq\arg\xi\leq\pi-\theta_0$ (Fig. 2), and (2.6) follows from (A1).

(ii) Further information is needed for ξ and s near ξ_l (Fig. 2), and this cannot be obtained by integration on progressive paths. In particular,

$$j(\xi_r,\xi_l)e^{-\rho\xi_l}=\int_{\xi_l}^{\xi_r}e^{-\rho s}\phi(s)\,ds$$

must be estimated, on a path avoiding the branch cut (Fig. 1), say $\Gamma+L$ (Fig. 4). An analogue of Jordan's lemma applies to the integral along Γ: the exponential keeps the integrand small except on the path segments close to the real axis, where integration by parts leads to a bound

$$\int_\Gamma e^{-\rho s}\phi(s)\,ds=O\left(\frac{\varepsilon}{\delta}\right).$$

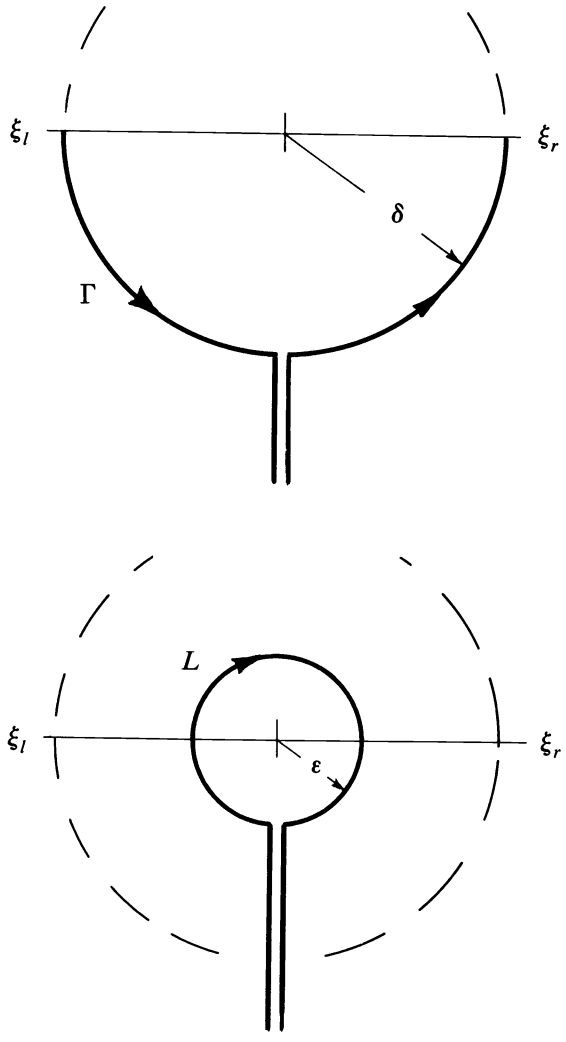

FIG. 4

By the residue theorem,

$$\int_L e^{-\rho s}\frac{ds}{s}=-2\pi i,$$

so that

$$j(\xi_r,\xi_l)e^{-\rho\xi_l}=-2\pi i\gamma+O\left(\frac{\varepsilon}{\delta}\right)+\int_L e^{-\rho s}\tilde{g}(s)\frac{ds}{s}.$$

To avoid the large values of the exponential in the upper half-plane, the curved part of L follows the circle $|s|=\varepsilon$, and if L' denotes the straight part (Fig. 4),

$$\int_L e^{-\rho s}\tilde{g}(s)\frac{ds}{s}=\int_{L'} e^{-\rho s}\tilde{g}(s)\frac{ds}{s}+O(g(\varepsilon)).$$

Since L' lies in the lower half-plane, the exponential is bounded on L' and the segment $\varepsilon\le|s|\le\delta$ contributes $O(g(\delta))$. On the remaining segment, $|s|>\delta$ and $\exp(-\rho s)\to 0$, and since the exponential order assumption on ϕ carries to $\tilde{g}$, the dominant contribution comes from the neighborhood of $|s|=\delta$ and is of the order of $\frac{\varepsilon}{\delta}\exp(-\frac{2\sigma}{\varepsilon})$. Therefore

$$\begin{aligned}\int_{L'} e^{-\rho s}\tilde{g}(s)\frac{ds}{s}&=O(g(\delta))+O\left(\frac{\varepsilon}{\delta}e^{-2\delta/\varepsilon}\right)\\&=o(1)\quad\text{as }\varepsilon\to 0\end{aligned}$$

and

$$j(\xi_r,\xi_l)e^{-\rho\xi_l}=-2\pi i\gamma+o(1), \tag{A9}$$

and (2.7) now follows from adding (2.6) and (A9).

Finally, $j(\xi,s)$ for both ξ and s close to ξ_l may be estimated by computing (2.5) along an arc of the circle $|s|=\delta$ (Fig. 5). Here $|\phi(s)|=O(s^{-1})$ by (A), and with $t=\mathrm{Im}(s)$

$$|j(\xi,\xi_l)e^{-\rho\xi_l}|\le C\int_{\xi_l}^{\xi}\left|e^{-\rho s}\frac{ds}{s}\right|\le C\int_0^{\mathrm{Im}(\xi)}e^{-i\rho t}|\delta^2-t^2|^{-1/2}\,dt$$

for some constant C, and the last integral can be computed to obtain

$$j(\xi,\xi_l)e^{-\rho\xi_l}=O\left(\frac{\varepsilon}{\delta}e^{-\rho\xi}\right). \tag{A10}$$

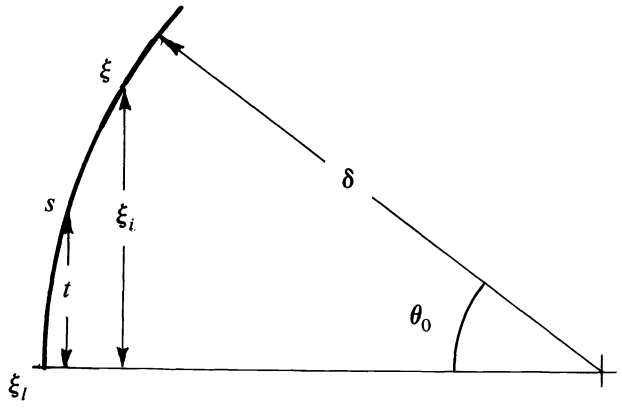

FIG. 5

This applies to ξ near ξ_l, in the first place, but (2.7) extends it to the arc $|\xi|=\delta, \theta_0 \leq \arg\xi \leq \pi$ (Fig. 6). It now follows from (2.5) or (A1), (A2) that for

$$\frac{\pi}{2} \leq \arg\xi \leq \arg s \leq \pi$$

on the left upper quarter-circle (Fig. 2)

$$j(\xi,s)e^{-\rho s}=j(\xi,\xi_l)e^{-\rho\xi_l}-j(s,\xi_l)e^{-\rho\xi_l}=O\left(\frac{\varepsilon}{\delta}e^{-\rho\xi}\right),$$

which is (2.8).

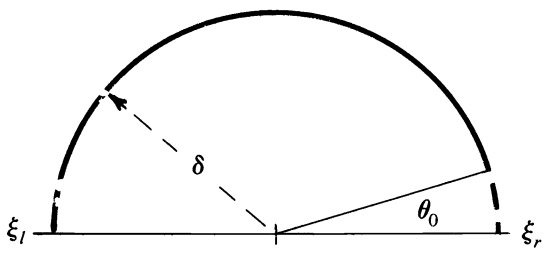

FIG. 6

REFERENCES

[1] M. A. EVGRAFOV AND M. V. FEDORYUK, *Asymptotic behaviour as* $\lambda \to \infty$ *of the solution of the equatin* $w''(z)-p(z,\lambda)w(z)=0$ *in the complex z-plane*, Usp. Mat. Nauk, 21, pp. 3–50; Russian Math Surveys, 21 (1966), pp. 1–48.

[2] P. O. FRÖMAN AND N. FRÖMAN, JWKB-*Approximation, Contributions to the Theory*, North-Holland, Amsterdam, 1965.

[3] E. C. KEMBLE, *A contribution to the theory of the* B.W.K. *method*, Phys. Rev., 48 (1935), pp. 549–561.

[4] R. E. LANGER, *On the asymptotic solutions of ordinary differential equations, with an application to Bessel functions of large order*, Trans. Amer. Math. Soc., 33 (1931), pp. 23–64.

[5] _____, *On the asymptotic solutions of differential equations, with an applicatio to Bessel functions of large complex order*, Trans. Amer. Math. Soc., 34 (1932), pp. 447–464.

[6] _____, *On the asymptotic solutions of ordinary differential equations, with reference to Stokes' phenomenon about a singular point*, Trans. Amer. Math. Soc., 37 (1935), pp. 397–416.

[7] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York and London, 1974.

[8] _____, *Second-order differential equations with fractional transition points*, Trans. Amer. Math. Soc., 226, (1977), pp. 227–241.

[9] _____, *General connection formulae for Liouville–Green approximations in the complex plane*, Phil. Trans. Roy. Soc. London A, 289 (1978), pp. 501–548.

[10] J. F. PAINTER, *Connection at close quarters to generalized turning points*, Ph.D. Thesis, University of Wisconsin-Madison, 1979.

[11] J. F. PAINTER AND R. E. MEYER, *Connection at close quarters to generalized turning points*, Tech. Sum. Rep. 2068, Mathematics Research Center, University of Wisconsin, Madison, 1980.

[12] A. ZWAAN, *Intensitaten im Ca-Funkenspektrum*, Arch. Neerlandaises Sci. Exactes Natur. Ser. 3A, 12 (1929), pp. 1–76.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0004 $01.00/0

GLOBAL STRUCTURE OF BIFURCATING SOLUTIONS OF SOME REACTION-DIFFUSION SYSTEMS*

Y. NISHIURA†

Abstract. We study the global structure of the set of bifurcating solutions of a class of coupled nonlinear reaction-diffusion systems. Our main result is that when one of the diffusion coefficients is sufficiently large, the bifurcating branch emanating from a uniform state continues to exist until it is connected to the singularly perturbed solutions which contain interior transition layers (Theorem 6.1).

We also present a global branching theorem for the bifurcating branch which shows that in the general situation it does not fall entirely on the trivial branch (Theorem 2.2).

0. Introduction. Many of the mathematical models that have been proposed for the study of population dynamics, biochemistry, morphogenesis, plasma physics and other fields, take the form of a system of coupled nonlinear diffusion equations:

$$(\tilde{\mathrm{P}}) \qquad u_t = d_1 \Delta u + f(u,v), \qquad v_t = d_2 \Delta v + g(u,v),$$

subject to

$$\frac{\partial u}{\partial n} = \frac{\partial v}{\partial n} = 0 \quad \text{on the boundary.}$$

We assume that $(\tilde{\mathrm{P}})$ has a constant solution $\bar{U}=(\bar{u},\bar{v})$ which is stable to spatially homogeneous perturbations, and that $f_u(\bar{u},\bar{v})>0$, $g_v(\bar{u},\bar{v})<0$. One of the most important and interesting problems is to find nonconstant stationary (especially stable) solutions of $(\tilde{\mathrm{P}})$ when d_1 and d_2 are taken as adjustable parameters, i.e., the study of the structure of the set of solutions of the following problem:

$$(\tilde{\mathrm{SP}}) \qquad 0 = d_1 \Delta u + f(u,v), \qquad 0 = d_2 \Delta v + g(u,v),$$

in the space $\mathbb{R}^2_+ \times X$, where $\mathbb{R}^2_+ = \{(d_1,d_2) \mid d_1>0,\ d_2>0\}$ is a parameter space and X is a suitable function space. It is obvious that $\{(D,\bar{U}) \mid D=(d_1,d_2)\in\mathbb{R}^2_+\}$ is a solution sheet of $(\tilde{\mathrm{SP}})$ which we call *the trivial branch* of $(\tilde{\mathrm{SP}})$.

Two powerful analytical methods for $(\tilde{\mathrm{SP}})$ have been developed: bifurcation theory and the singular perturbation method. The former method is closely related to the phenomenon of diffusion driven instability, which can be traced back to Turing's work [26]. When d_1 and d_2 are sufficiently large, the constant state $\bar{U}$ is locally stable (it is globally and asymptotically stable if $(\tilde{\mathrm{P}})$ has an invariant rectangle, see, e.g., Conway, Hoff and Smoller [3]). As d_1 becomes smaller, $\bar{U}$ loses its stability at some value $D=D_c=(d_1^c,d_2^c)$ (see Fig. 6 below), and new nonconstant solutions appear in place of $\bar{U}$. We call $(D_c,\bar{U})$ *the bifurcation point* with respect to the trivial branch. The bifurcation method gives us precise pictures of the solution set of $(\tilde{\mathrm{SP}})$ near the bifurcation points, and there is extensive literature such as [1], [23], [20], [17]. However, solutions obtained by this method are of small amplitude, and when the parameters leave the critical point D_c we usually have little information about the behavior of the bifurcating solutions.

On the other hand, the latter method, when one of the diffusion coefficients d_1 is sufficiently small, enables us to obtain large amplitude solutions which contain interior transition layers. In fact, Fife [7] proved the existence of singularly perturbed solutions

*Received by the editors October 27, 1980, and in final form June 22, 1981.
†Department of Computer Sciences, Kyoto Sangyo University, Kyoto 603, Japan.

of (SP̃) in one space dimension under Dirichlet boundary conditions and, using Fife's method, Mimura, Tabata and Hosono [19] proved analogous results for Neumann boundary conditions.

The main purpose of this paper is to integrate the results obtained separately by the above two methods; i.e., the bifurcating branch continues to exist with respect to d_1 until it is connected to the singularly perturbed solutions when d_2 is sufficiently large. The asymptotic analysis as $d_2 \uparrow \infty$ for fixed d_1 was done by Keener [14], which shows the existence of large amplitude solutions and suggests the relation between bifurcating solutions and large amplitude ones. However, this does not cover the problem of the global structure of bifurcating branches and their asymptotic behavior as $d_1 \downarrow 0$. We note that the largeness of d_2 appears naturally in some applications (see the examples below).

In the case where d_2 is not large, the situation is more complicated. The bifurcating branches of different modes can intersect each other, and therefore secondary and tertiary bifurcations necessarily occur. To understand the mechanism of the global behavior of these branches, we need to perform more systematic analytical and numerical studies. In this regard, see, e.g., Fujii, Mimura and Nishiura [8].

In order to prove the global existence of the bifurcating branch with respect to d_1, the assumption of large d_2 is unnecessary. This is discussed more precisely in §2. Rabinowitz's alternative theorem [22] plays an important role in the proof of Theorem 2.2.

Throughout this paper, when d_2 is fixed, we use the space $\mathfrak{S}=\mathbb{R}^+\times X$, where $\mathbb{R}^+=\{d_1 \mid d_1>0\}$.

Now we restate our problem and assumptions. We treat the system of equations in one-dimensional space:

$$\text{(SP-1)} \qquad 0=d_1u_{xx}+f(u,v), \qquad x\in I=(0,1),$$

$$\text{(SP-2)} \qquad 0=d_2v_{xx}+g(u,v),$$

subject to zero flux boundary conditions

$$\text{(SP-3)} \qquad u_x(0)=u_x(1)=0, \qquad v_x(0)=v_x(1)=0.$$

In vector notation we write (SP) as

$$\text{(SP)} \qquad 0=DU_{xx}+F(U),$$

where

$$D=\begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}, \qquad F(U)=\begin{pmatrix} f(u,v) \\ g(u,v) \end{pmatrix}.$$

We use the symbol D for the diagonal matrix or a point $(d_1,d_2)\in\mathbb{R}^2_+$, as the case may be.

We impose the following conditions on the nonlinearities of (SP):

(A-0) The nonlinear functions f and g are smooth (C^∞-class) on some open set $\Omega\subset\mathbb{R}^2$ which contains in its interior a closed rectangle R defined below by (0.1).

(A-1) The zero level curve $f(u,v)=0$ is S-shaped in Ω. When it is solved with respect to u, it consists of three branches $u=h_i(v)$ $(i=0,1,2)$ such that

$h_i(v)\in C^0(I_i)\cap C^\infty(\mathring{I}_i)$ $(i=0,1,2)$ with $I_0=[\underline{\xi},\bar{\xi}]$ and $I_0\subset I_i$ $(i=1,2)$ (each I_i is a closed interval); they satisfy

$$h_1(v)\le h_0(v)\le h_2(v) \quad \text{for } v\in I_0,$$

where equalities hold if and only if $v=\underline{\xi}$ for $h_1(v)=h_0(v)$ and $v=\bar{\xi}$ for $h_0(v)=h_2(v)$ (see Fig. 1). Moreover we assume that $f_u(h_i(v),v)\neq 0$ for $v\in\mathring{I}_i$ $(i=0,1,2)$. The positive-valued region of f corresponds to the upper side of the graph $u=h_0(v)$. We define the closed rectangle R by

(0.1) $$R=[u_{\min},u_{\max}]\times I_0,$$

where

$$u_{\min}=\min_{v\in I_0} h_1(v), \qquad u_{\max}=\max_{v\in I_0} h_2(v).$$

Hereafter we denote the open interval $\mathring{I}_0=(\underline{\xi},\bar{\xi})$ by Ξ.

Remark 0.1. It follows from (A-1) that, for each fixed $v=\xi\in\Xi$, $f(u,\xi)$ has just three zeros $u_l(\xi)<u_c(\xi)<u_r(\xi)$ in Ω (see Fig. 1).

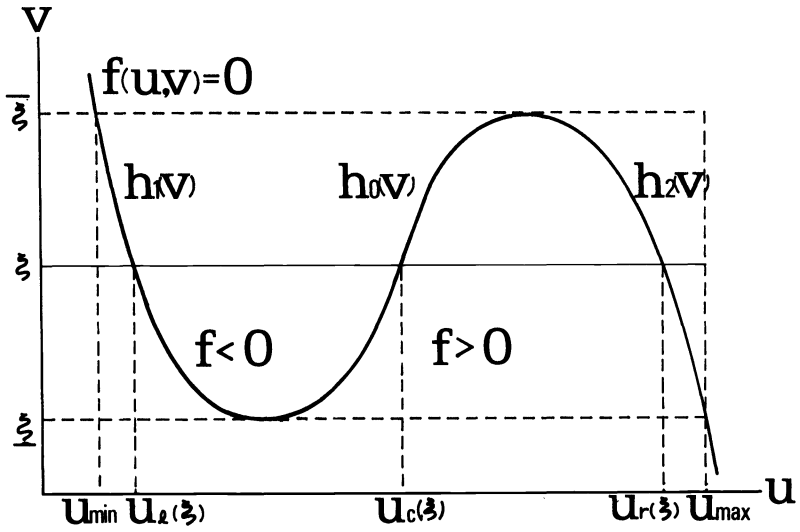

FIG. 1. *Typical functional form of* $f(u,v)=0$.

(A-2) Define $J(\xi)$ by $J(\xi)=E_r^\infty(\xi)-E_l^\infty(\xi)$ for $\xi\in\Xi$, where

$$E_r^\infty(\xi)=\int_{u_c(\xi)}^{u_r(\xi)} f(s,\xi)\,ds, \qquad E_l^\infty(\xi)=\int_{u_c(\xi)}^{u_l(\xi)} f(s,\xi)\,ds.$$

Then $J(\xi)=0$ holds if and only if $\xi=\xi^*\in\Xi$ and

$$\frac{d}{d\xi}E_r^\infty(\xi)<0, \qquad \frac{d}{d\xi}E_l^\infty(\xi)>0 \quad \text{at } \xi=\xi^*$$

(which obviously imply that $dJ(\xi)/d\xi<0$ at $\xi=\xi^*$).

(A-3) In Ω there exists a unique constant solution $\bar{U}=(\bar{U},\bar{v})$ of $f(u,v)=g(u,v)=0$.

(A-4) At $(u,v)=(\bar{u},\bar{v})$, $f_u>0$, $f_v<0$, $g_u>0$, $g_v<0$, $f_u+g_v<0$ and $f_ug_v-f_vg_u>0$ hold.

Remark 0.2. It follows from (A-0), (A-1), (A-3) and (A-4) that $(\bar{u},\bar{v})$ must lie on the branch $u=h_0(v)$ and the zero level curve of g intersects it transversally at $(\bar{u},\bar{v})$. Moreover, the positive-valued region of g is on the right-hand side of the zero level curve of g (see Fig. 2).

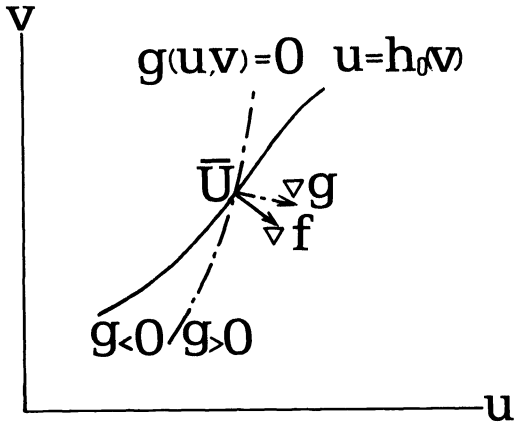

FIG. 2. *Functional forms of* $f=0$ *and* $g=0$ *near* $U=\bar{U}$.

We show several examples which satisfy the above assumptions.

Example 1 (Diffusive prey-predator model [17]). First we give a model system of a prey and its predator with diffusion:

(E-1) $$0=d_1u_{xx}+f_0(u)u-kuv, \qquad 0=d_2v_{xx}-g_0(v)v+kuv,$$

where k is a positive constant; $f_0(u)$ is a smooth function such that

(i) $$f_0(0)\geq 0,$$

(ii) $$\frac{d}{du}f_0(u)\begin{cases}>0, & 0\leq u<c,\\ =0, & u=c,\\ <0, & u>c\end{cases}$$

for some positive constant c; and $g_0(v)=c_0+c_1v^m(c_0,c_1,m>0)$.

In this model, u and v denote the population densities of a prey and its predator, respectively. If the predator's diffusivity is very high, we can assume $d_2\gg 1$. The zero level curves of nonlinear terms are drawn in Fig. 3.

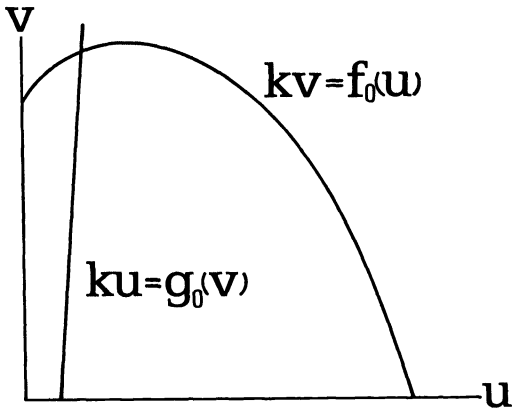

FIG. 3. *Functional forms of prey-predator model.*

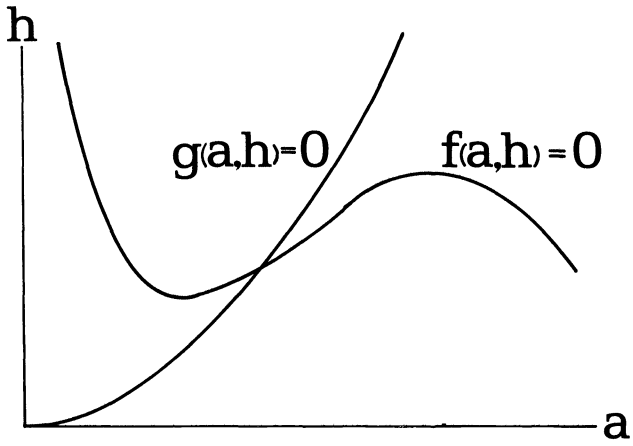

FIG. 4. *Functional forms of Gierer–Meinhardt model with saturation.*

Example 2 (Gierer-Meinhardt model with saturation). The following model was proposed by Gierer and Meinhardt [10] for the study of morphogenesis:

$$\text{(E-2)} \qquad \begin{aligned} 0 &= d_a a_{xx} + \rho\rho_0 + \frac{c\rho a^2}{h(1+\kappa a^2)} - \mu a, \\ 0 &= d_h h_{xx} + c'\rho' a^2 - \mu a, \end{aligned}$$

where $0 < d_a \ll d_h$ and ρ, ρ', ρ_0, c, c', κ, μ and ν are all positive constants. For the zero level curves $f(a,h)=0$ and $g(a,h)=0$, see Fig. 4.

Example 3 (Seelig's model with diffusion [15]). Finally we give a model of a substrate-inhibitaion reaction diffusion system:

$$\text{(E-3)} \qquad \begin{aligned} 0 &= d_1 u_{xx} + j_1 - u - \beta r(u,v), \\ 0 &= d_2 v_{xx} + j_2 - \gamma r(u,v), \end{aligned}$$

where $r(u,v)=uv/(1+u+v+Ku^2)$ and j_1, j_2, β, γ and K are all positive constants. This model without diffusion was originally proposed by Seelig [25]. The zero level curves of f and g are drawn in Fig. 5.

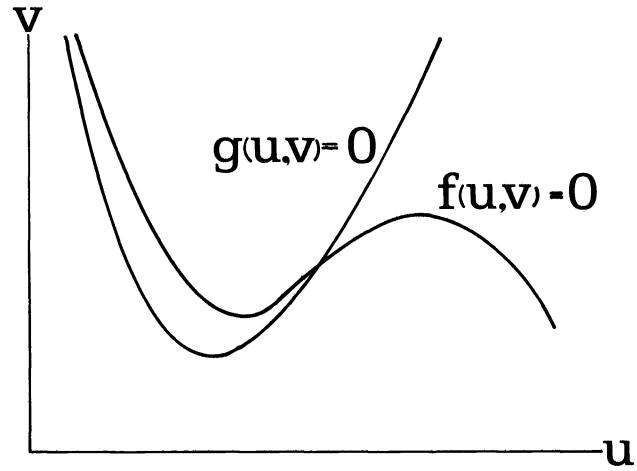

FIG. 5. *Functional forms of Seelig's model.*

This paper is divided into seven parts. In §1, we summarize the results of a local bifurcation analysis, especially near the simple critical case. In §2, we prove the general global branching theorem under the assumptions (A-0), (A-3) and (A-4). In §3, we derive a limit system of (SP) as $d_2 \uparrow \infty$, which we call the *shadow system*, and study its global structure in $\mathfrak{E}$ of the set of bifurcating solutions from the trivial branch. We also prove that there are generically no secondary bifurcations along the solution branch of the shadow system. In §4, we study the relation between the structure of the shadow system and that of the original system (SP). We show that the former is a nice approximation to the latter when d_2 is sufficiently large. In §5, we summarize the results of singular perturbation analysis with respect to d_1 studied by [7] and [19], and derive some lemmas to be used later. In §6, using the results of previous sections, we prove the main theorem: The bifurcating branch continues to be extended with respect to d_1 until it is connected to the singularly perturbed solutions of §5. Finally we conclude with some comments on our results.

Some of the results in §§4 and 5 were reported at an International Symposium on Mathematical Topics in Biology at the Research Institute for Mathematical Sciences, Kyoto University, 1978 [16] and the main theorem was announced in [21].

Throughout this paper, we use the following notation:

$\mathbb{R}^2_+ = \{(d_1, d_2) \in \mathbb{R}^2 \mid d_1 > 0,\ d_2 > 0\}$.
$\mathbb{R}^+ = \{d_1 \in \mathbb{R} \mid d_1 > 0\}$.
$\mathfrak{E} = \mathbb{R}^+ \times X$, where X is a suitable function space.
$R = [u_{\min}, u_{\max}] \times I_0$. See (0.1).
$\Xi = (\xi, \bar{\xi})$. See (A-1).
$E_r^\infty(\xi)$, $E_l^\infty(\xi)$: See (A-2).
C_n: See (1.4).
$\mathfrak{B} = \bigcup_{n=1}^\infty C_n$.
$d^n_{c,d_2} = d^n_{c,\alpha}$, where $\alpha = d_2^{-1}$. See Remark 1.3.
$d^n_{c,0} = \lim_{\alpha \downarrow 0} d^n_{c,\alpha} = \lim_{d_2 \uparrow \infty} d^n_{c,d_2} = b_{11}/(n\pi)^2$.
$\mathbb{H}^2_N(I) = (H^2_N(I))^2$ with norm $\|\cdot\|_2$, where $H^2_N(I)$ = closure of $\{\cos(n\pi x)\}_{n=0}^\infty$ in $H^2(I)$.
$V^{[n]} = (H^{2,n}_N)^2$, where $H^{2,n}_N$ = closure of $\{\cos(mn\pi x)\}_{m=0}^\infty$ in $H^2(I)$.
$\mathfrak{S}$ = closure of the set of nonconstant solutions of (SP) in $\mathfrak{E}$.
$\mathcal{C}^n_{d_2} = \mathcal{C}^n_\alpha\ (\alpha = d_2^{-1})$ = component in $\mathfrak{S}$ to which $(d^n_{c,d_2}, \bar{U})$ belongs.
$\mathfrak{S}_1$: See Theorem 3.3.
$\mathrm{Sec}_k(W) = \{w \mid (k, w) \in W\}$, where W is a subset of $\mathfrak{E}$ and $k \in \mathbb{R}^+$.
$\mathcal{C}^1_{0,\pm}$: See (4.1), (4.2).
$\mathcal{C}^1_0 = \mathcal{C}^1_{0,+} \cup \mathcal{C}^1_{0,-}$.
$\mathcal{C}^n_0$: See Remark 4.1.
$A \setminus B = \{x \mid x \in A \text{ and } x \notin B\}$.

1. Local bifurcation analysis. Throughout this section and the next one, we only assume (A-0), (A-3) and (A-4). It follows from (A-4) that the uniform state $\bar{U}$ is a stable equilibrium point of the evolution system (P) with $d_1 = d_2 = 0$. Moreover, even if the diffusion coefficients d_1 and d_2 are not zero, it is a stable solution of (P) when they are sufficiently large. Therefore the instability of $\bar{U}$ occurs when at least one of the diffusion coefficients is not so large.

To find out the critical points $D_c=(d_1^c,d_2^c)$ where $\bar{U}$ becomes unstable, we have to solve the following linear eigenvalue problem associated with the linearized operator of (SP) at $\bar{U}$:

$$D\Psi_{xx}+B\Psi=\lambda\Psi, \tag{EP-1}$$

$$\Psi_x(0)=\Psi_x(1)=0, \tag{EP-2}$$

where $\Psi={}^t(\psi_1,\psi_2)$ and $B=\{b_{ij}\}_{i,j=1,2}$ is a Jacobi matrix of $F={}^t(f,g)$ at $\bar{U}$, i.e., $b_{11}=f_u|_{\bar{U}}$, $b_{12}=f_v|_{\bar{U}}$, $b_{21}=g_u|_{\bar{U}}$, $b_{22}=g_v|_{\bar{U}}$. The Fourier cosine expansion $\Psi=\sum_{n=0}^{\infty}\Psi_n\cos(n\pi x)$ reduces the problem (EP) to the following system:

$$(B_n-\lambda E)\Psi_n=0, \qquad n=0,1,2,\cdots, \tag{1.1}$$

where $B_n=B-(n\pi)^2D$ and E is a unit matrix. The characteristic equation of (1.1) is

$$\begin{aligned}\lambda^2-\{(b_{11}+b_{22})-(d_1+d_2)(n\pi)^2\}\lambda \\ +\left(b_{11}-(n\pi)^2d_1\right)\left(b_{22}-(n\pi)^2d_2\right)-b_{12}b_{21}=0.\end{aligned} \tag{1.2}$$

It follows from (A-4) that the coefficient of λ is strictly positive for any n. Therefore a pair of complex conjugate roots of (1.2) never crosses the imaginary axis when D varies in $\mathbb{R}^2_+$. Hence the instability of $\bar{U}$ occurs if and only if some real root of (1.2) crosses the origin. It is easily seen that (1.2) has a zero root if and only if D satisfies the following:

$$\left(b_{11}-(n\pi)^2d_1\right)\left(b_{22}-(n\pi)^2d_2\right)-b_{12}b_{21}=0. \tag{1.3}$$

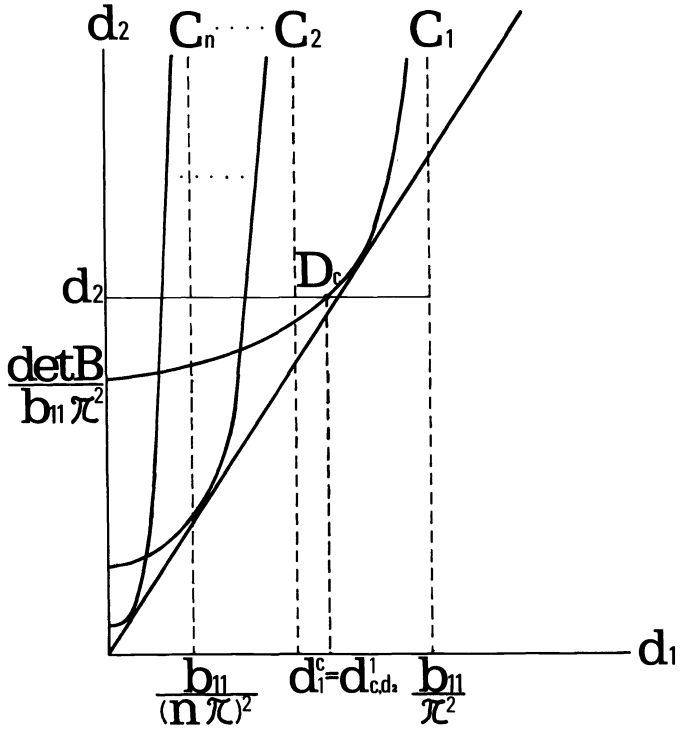

FIG. 6. *Schematic bifurcation set in* (d_1,d_2)*-space.*

Thus the set of values of D where (EP) has a zero eigenvalue is given by the hyperbolic curves $\{C_n\}_{n=1}^{\infty}$ in $\mathbb{R}_+^2$ (see Fig. 6):

$$C_n: \quad d_2=\frac{b_{12}b_{21}/(n\pi)^4}{d_1-b_{11}/(n\pi)^2}+\frac{b_{22}}{(n\pi)^2}, \qquad n=1,2,\cdots. \tag{1.4}$$

Remark 1.1. All the curves $\{C_n\}_{n=1}^{\infty}$ are tangent to the straight line $d_2=md_1$, where

$$m=\frac{\det B-b_{12}b_{21}+2\sqrt{-b_{12}b_{21}\det B}}{b_{11}^2}.$$

Remark 1.2. It follows from (1.4) that for any m and n, we have the following:

$$\text{If } (d_1,d_2)\in C_m, \quad \text{then } \left(m^2d_1/n^2, m^2d_2/n^2\right)\in C_n.$$

We call $\cup_{n=1}^{\infty} C_n$ the *bifurcation set* $\mathfrak{B}$ with respect to $\bar{U}$. We define the *primary bifurcation curve* Γ (Fig. 6) by

$$\Gamma=\bigcup_{n=1}^{\infty}\hat{C}_n, \qquad \hat{C}_n=\{(d_1,d_2)\in C_n \mid P_n\le d_1<P_{n-1}\},$$

where $P_0=b_{11}/\pi^2$ and P_n $(n\ge 1)$ is an abscissa of the intersecting point of C_n and C_{n+1}. Note that $\hat{C}_n\neq\varnothing$ and $\hat{C}_n\cap\hat{C}_m=\varnothing$ when $n\neq m$. We denote by $\mathfrak{B}^*$ the countable set of intersection points of two curves of $\{C_n\}_{n=1}^{\infty}$, and $\mathfrak{B}'=\mathfrak{B}\backslash\mathfrak{B}^*$. It is apparent that, if $D_c\in\mathfrak{B}'$, there exists a unique n such that $D_c\in C_n$. Every point $D_c\in\mathfrak{B}'$ is a *nondegenerate simple critical point* in the sense that [4, Thm. 1.7] is applicable to our case when the bifurcation parameter D crosses $\mathfrak{B}'$ transversally at D_c. Here, for simplicity, we take d_1 as a bifurcation parameter with d_2 being fixed.

THEOREM 1.1 ([17]). *For any $D_c=(d_1^c,d_2)\in\mathfrak{B}'\cap C_n$ with d_2 fixed, there exists a positive constant ε_0 such that* (SP) *has a unique one-parameter family of solutions $(d_1(\varepsilon),U(\varepsilon))=\mathbb{R}^+\times\mathbb{H}_N^2(I)$ for $|\varepsilon|<\varepsilon_0$. Here $d_1(\varepsilon)$ is a smooth function of ε with $d_1(0)=d_1^c$ and*

$$U(\varepsilon)=\bar{U}+\varepsilon\Phi_n+o(|\varepsilon|),$$

where Φ_n is a normalized eigenfunction corresponding to the zero eigenvalue of (EP) *with $D=D_c$.*

Moreover the above bifurcation is one-sided (i.e., $d_1(\varepsilon)<$ (or $>$) d_1^c for $\varepsilon\neq 0$) and, if D_c lies on the primary bifurcation curve Γ, the bifurcating solutions are stable when they appear above criticality (i.e., $d_1(\varepsilon)<d_1^c$) and unstable when they occur below criticality (i.e., $d_1(\varepsilon)>d_1^c$).

Remark 1.3. For each fixed d_2, we denote the countable set $\{d_1^c \mid (d_1^c,d_2)\in\mathfrak{B}\}$ by $\mathfrak{B}_{d_2}$. We can label each element d_1^c of $\mathfrak{B}_{d_2}$ in the following manner:

$$d_1^c=\begin{cases} d_{c,d_2}^n & \text{if } (d_1^c,d_2)\in C_n \text{ and } \notin C_m \text{ for } m\neq n,\\ d_{c,d_2}^{n,m} & \text{if } (d_1^c,d_2)\in C_n\cap C_m,\ n<m.\end{cases}$$

We also note that

$$\lim_{d_2\uparrow\infty} d_{c,d_2}^n=\frac{b_{11}}{(n\pi)^2}. \tag{1.5}$$

If D_c is a point of double multiplicity, the local structure of solutions of (SP) near $D=D_c$ is more complicated than the above theorem (see [8], [13]). However, using

Theorem 1.1, we can always pick up bifurcating solutions from a double critical point by restricting the solution space to the subspace of $\mathbb{H}_N^2$ which has some symmetric property. In fact we have the following:

PROPOSITION 1.2. *$V^{[n]}$ is an invariant subspace of $\mathbb{H}_N^2$ under the operation $\mathcal{G}$, where $V^{[n]}=(H_N^{2,n})^2$, $H_N^{2,n}=$closure of $\{\cos(mn\pi x)\}_{m=0}^{\infty}$ in $H^2(I)$ and $\mathcal{G}$ is a compact operator form of* (SP) *which is defined by* (2.1) *in* §2. *When we consider the problem* (SP) *in $V^{[n]}$, the corresponding bifurcation set $\mathcal{B}^n$ consists of $\{C_{mn}\}_{m=1}^{\infty}$. Therefore, if D_c is a point of double multiplicity such that $D_c\in C_k\cap C_l(0<k<l)$, D_c becomes a nondegenerate simple critical point if we restrict the solution space to $V^{[l]}$ if l/k is integral, and to $V^{[k]}$ or $V^{[l]}$ if l/k is nonintegral.*

Proof. We prove only the invariant property of $V^{[n]}$. The remaining part is easy to prove, so we leave it to the reader. In order to prove that $\mathcal{G}$ maps any element of $V^{[n]}$ into itself, it suffices to show the following:

$$\int_0^1 F(U(x))\cdot\cos(k\pi x)\,dx=0 \tag{1.6}$$

for any positive integer $k\not\equiv n \pmod n$ and $U(x)\in V^{[n]}$.

After even extension of the integrand to $(-1,0)$, the left-hand side of (1.6) is equal to

$$\frac{1}{2}\int_{-1}^1 F(U(x))\cos(k\pi x)\,dx.$$

Since $F(U(x))$ is an even function at $x=0$, we can rewrite this as follows:

$$=\frac{1}{2}\int_{-1}^1 F(U(x))\exp(ik\pi x)\,dx.$$

It is clear that $F(U(x))$ and $\exp(ik\pi x)$ can be extended periodically to $\mathbb{R}$. Then, transforming x to $x-2/n$, we obtain

$$=\frac{1}{2}\exp\left(-\frac{2k\pi i}{n}\right)\int_{-1}^1 F(U(x))\exp(ik\pi x)\,dx.$$

Here we use the fact that

$$U\left(x-\frac{2}{n}\right)=U(x)\quad\text{for } U(x)\in V^{[n]}.$$

Thus we obtain

$$\left\{1-\exp\left(-\frac{2k\pi i}{n}\right)\right\}\int_{-1}^1 F(U(x))\exp(ik\pi x)\,dx=0,$$

which implies (1.6) because $k\not\equiv n \pmod n$.

Remark 1.4. The space $V^{[n]}$ is characterized by the following property of symmetry: $U(x)(\in\mathbb{H}_N^2)$ belongs to $V^{[n]}$ if and only if $\tilde{U}(x)=\tilde{U}(-x)$ and $\tilde{U}(x-2/n)=\tilde{U}(x)$, where $\tilde{U}(x)$ is defined by the periodic extension of $U_e(x)$ to $\mathbb{R}$ and $U_e(x)$ denotes the even extension of $U(x)$ to $(-1,1)$.

The above result is a consequence of our homogeneous Neumann boundary conditions. In fact, for Dirichlet boundary conditions, for example, we cannot obtain the corresponding result because the even power of the sine function cannot be expressed by sine functions only. As is seen from Remark 1.4, the group-theoretic method is useful for our problem. For results in this direction, see [8].

2. Global existence theorem of the bifurcating branch with respect to the parameter d_1. It is natural to ask how the bifurcating branch in the previous section behaves when d_1 leaves the critical point d_1^c. Rabinowitz's alternative theorem [20] says that the component in $\mathfrak{S}=\mathbb{R}^+\times\mathbb{H}_N^2$ which contains the bifurcating branch of Theorem 1.1 *exists globally* in the sense that either there is no closed bounded set M in the interior of $\mathfrak{S}$ which contains the component, or else such a set M must contain the bifurcation points different from $(d_1^c, \bar{U})$. We will show in this section that the former case occurs for our system. Here we assume, for simplicity, that the nonlinear term F is globally defined, i.e., $\Omega=\mathbb{R}^2$.

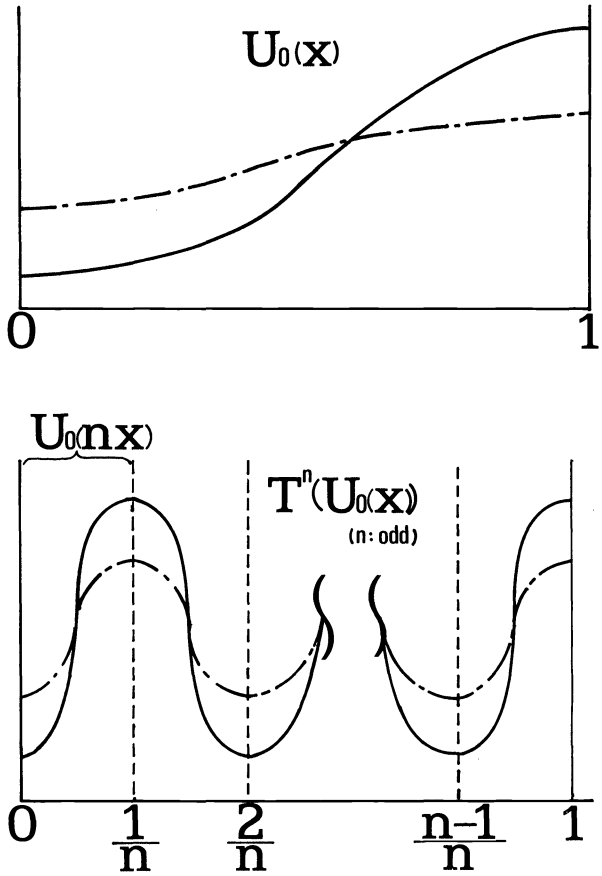

FIG. 7. *Transformation T^n.*

First we convert (SP) into compact operator form by operating on it with

$$K=\left(-D\frac{d^2}{dx^2}+\hat{I}\right)^{-1},$$

where $\hat{I}$ denotes the identity operator. Then we obtain

$$U=\mathcal{G}(d_1, U), \tag{2.1}$$

where $\mathcal{G}(d_1, U)=K(\hat{I}+F(U))$. Here we consider d_2 to be fixed. It is clear that $\mathcal{G}$ is a compact continuous operator from $\mathbb{R}^+\times\mathbb{H}_N^2$ into $\mathbb{H}_N^2$. The known curve of solutions

$\{(d_1, \bar{U}) \mid d_1 \in \mathbb{R}^+\}$ will be referred to as the *trivial branch*. The closure of the set of nonconstant solutions of (SP) in $\mathfrak{E}$ will be denoted by $\mathfrak{S}$, and $\mathcal{C}^n_{d_2}$ (resp. $\mathcal{C}^{n,m}_{d_2}$) denotes the component in $\mathfrak{S}$ to which $(d^n_{c,d_2}, \bar{U})$ (resp. $d^{n,m}_{c,d_2}, \bar{U}$) belongs. The following lemma is a consequence of our boundary conditions (SP-3).

LEMMA 2.1. *Let $U_0(x)$ be any smooth solution of* (SP) *with $D = D_0$. Then $T^n(U_0(x))$ defined by*

$$T^n(U_0(x)) = \begin{cases} U_0\left(n\left(x - \frac{i}{n}\right)\right) & \text{if } i \text{ is even}, \\ U_0\left(n\left(\frac{1}{n} - \left(x - \frac{i}{n}\right)\right)\right) & \text{if } i \text{ is odd} \end{cases} \tag{2.2}$$

for $i/n \le x \le (i+1)/n$ and $i = 0, 1, \cdots, n-1$ is a solution of (SP) *with $D = D_0/n^2$ (see Fig. 7).*

Proof. $T^n(U_0(x))$ is constructed by means of successive $(n-1)$-times reflections of $U_0(nx)$ $(0 \le x \le 1/n)$ at i/n $(i = 1, 2, \cdots, n-1)$. Since (SP) is an autonomous system and $T^n(U_0(x))$ is matched at i/n $(i = 1, \cdots, n-1)$ in the C^1-sense, the conclusion follows easily.

Now we can prove

THEOREM 2.2. *Under the assumptions* (A-0), (A-3) *and* (A-4), $\mathcal{C}^n_{d_2}$ *(resp. $\mathcal{C}^{n,m}_{d_2}$) exists globally in $\mathfrak{E}$ for any $d_2 \in \mathbb{R}^+$ and $n \ge 1$ (resp. n, m $(1 \le n < m)$).*

Proof. Let us assume the contrary. It then follows from Rabinowitz's alternative theorem [22] (and Proposition 1.2) that $\mathcal{C}^n_{d_2}$ (resp. $\mathcal{C}^{n,m}_{d_2}$) contains a finite subset $A = \{d^c_1 \mid (d^c_1, \bar{U}) \in \mathcal{C}^n_{d_2}$ (resp. $\mathcal{C}^{n,m}_{d_2})\}$ of $\mathfrak{B}_{d_2}$. Let $\bar{m} = \max\{m \mid$ there exists an element d^c_1 of A such that $d^c_1 \in C^m\}$ and denote the value of $d^c_1 \in A$ which attains $\bar{m}$ by $d^{\bar{m}}_1$. It follows from Lemma 2.1 and Remark 1.2 that if $(\bar{m}^2 d^{\bar{m}}_1, \bar{m}^2 d_2) \in C_1$ is simple we have

$$T^{\bar{m}}\left(\mathcal{C}^1_{\bar{m}^2 d_2}\right) \subset \mathcal{C}^n_{d_2} \quad \left(\text{resp. } \mathcal{C}^{n,m}_{d_2}\right), \tag{2.3}$$

and if $(\bar{m}^2 d^{\bar{m}}_1, \bar{m}^2 d_2)$ is of double multiplicity, i.e., there exists some integer $l \ge 2$ such that it belongs to $C_1 \cap C_l$, we have

$$T^{\bar{m}}\left(\mathcal{C}^{1,l}_{\bar{m}^2 d_2}\right) \subset \mathcal{C}^n_{d_2} \quad \left(\text{resp. } \mathcal{C}^{n,m}_{d_2}\right).$$

It suffices to consider the simple critical case, since restricting the solution space to $V^{[l]}$, we can apply the following discussions to $\mathcal{C}^{1,l}_{\bar{m}^2 d_2}$ in virtue of Proposition 1.2.

Suppose that $\mathcal{C}^1_{\bar{m}^2 d_2}$ exists globally; then $T^{\bar{m}}(\mathcal{C}^1_{\bar{m}^2 d_2})$ also exists globally, which leads to a contradiction from (2.3). If $\mathcal{C}^1_{\bar{m}^2 d_2}$ does not exist globally, it must contain a bifurcation point $(d^k_{c,\bar{m}^2 d_2}, U)$ for $k \ge 2$ (see [22, Thm. 1.16]). But this contradicts the fact that $\bar{m}$ is the maximum value, because applying $T^{\bar{m}}$ to $\mathcal{C}^1_{\bar{m}^2 d_2}$, we can see that $\mathcal{C}^n_{d_2}$ (resp. $\mathcal{C}^{n,m}_{d_2}$) must contain $(d^{\bar{m}k}_{c,d_2}, \bar{U})$. In either case we have a contradiction. Thus the proof is completed.

If we have some a priori estimate for the bifurcating branch $\mathcal{C}^n_{d_2}$ (resp. $\mathcal{C}^{n,m}_{d_2}$), much more can be said about its global behavior. We say that (SP) has the *property* (B) if and only if

(B) There exists a positive continuous function $M(d_1, d_2)$ on $\mathbb{R}^2_+$ such that for any $(d_1, U) \in \mathcal{C}^n_{d_2}$ (resp. $\mathcal{C}^{n,m}_{d_2}$) the estimate $\|U\|_\infty \le M(d_1, d_2)$ holds, where $\|\cdot\|_\infty$ denotes the L^∞-norm.

One sufficient condition for (SP) to satisfy (B) is that ($\tilde{P}$) has a *strictly contracting rectangle* (see [18, Remark 5.2]). In this case we can take M to be independent of d_1 and d_2. It is known that examples (E-2) and (E-3) have such rectangles (see [18] and [15]).

We say that $\mathcal{C}^n_{d_2}$ (resp. $\mathcal{C}^{n,m}_{d_2}$) *exists globally with respect to* d_1 if and only if $\overline{\mathrm{Proj}(\mathcal{C}^n_{d_2})}$ $\left(\textit{resp.}\ \overline{\mathrm{Proj}(\mathcal{C}^{n,m}_{d_2})}\right)$ contains 0, where Proj denotes the projection operator from $\mathfrak{S}$ to d_1-space and the upper bar denotes the closure operator in $\mathbb{R}$.

THEOREM 2.3. *Suppose that* (SP) *satisfies the property* (B). *Then* $\mathcal{C}^n_{d_2}$ (*resp.* $\mathcal{C}^{n,m}_{d_2}$) *exists globally with respect to* d_1 *in* $\mathfrak{S}$ *for any* $d_2\in\mathbb{R}^+$ *and* $n\geq 1$ (*resp.* n,m $(1\leq n<m)$).

Proof. Using the equations of (SP), we can see that property (B) leads to the estimate

$$\|U\|_2\leq kM(d_1,d_2)\left(\frac{1}{d_1}+\frac{1}{d_2}\right), \tag{2.4}$$

for any solution $(d_1,U)\in\mathcal{C}^n_{d_2}$ (resp. $\mathcal{C}^{n,m}_{d_2}$), where k is a positive constant independent of d_1 and d_2. Combining (2.4) and Theorem 2.2, we have the required result.

3. Shadow system. We start by giving a heuristic derivation of a limit system of (SP) as $d_2\uparrow\infty$. First we rewrite (SP) as follows:

$$\text{(SP)}\qquad \begin{aligned} 0&=du_{xx}+f(u,v),\\ 0&=v_{xx}+\alpha g(u,v),\\ u_x(0)&=u_x(1)=v_x(0)=v_x(1)=0, \end{aligned}$$

where $d=d_1$ and $\alpha=d_2^{-1}$. Let us suppose that, for some fixed d, solutions of (SP) remain to be bounded as $\alpha\uparrow 0$. The second equation of (SP) implies that the limit of v satisfies

$$v_{xx}=0. \tag{3.1}$$

It follows from (3.1) and the boundary conditions that v must be a constant. On the other hand, integrating the second equation of (SP) from 0 to 1, we can see that

$$\int_0^1 g(u,v)\,dx=0 \tag{3.2}$$

holds independently of α. Thus we obtain a limit system of (SP) as $\alpha\downarrow 0$:

$$\text{(SS-1)}\qquad 0=du_{xx}+f(u,\xi),$$

$$\text{(SS-2)}\qquad \int_0^1 g(u,\xi)\,dx=0,\qquad x\in I,$$

$$\text{(SS-3)}\qquad u_x(0)=u_x(1)=0,$$

where $v=\xi$ is a constant function. In (SS), ξ is considered to be a parameter which is contained in both (SS-1) and (SS-2). We call (SS) the *shadow system* of (SP). In this section, we study the global structure of the set of solutions of (SS) in $\mathfrak{S}$. We will see in the next section that its structure is a nice approximation to that of (SP) for small $\alpha>0$.

3.1. Global branching theorem for the shadow system. First let us consider the problem (SS-1) and (SS-3). It follows from Remark 0.1 that $f(u,\xi)$ has three zeros $u_l(\xi)<u_c(\xi)<u_r(\xi)$ if and only if $\xi\in\Xi$. Note that $U=\bar{U}$ is the only constant solution of (SS), and that (SS-1) and (SS-3) has no nonconstant solutions for $\xi\notin\Xi$. Therefore,

without loss of generality, we can assume that ξ moves only in Ξ. The energy form of (SS-1) is given by

$$\frac{d}{2}(u_x)^2+F(u,\xi)=E, \tag{3.3}$$

where

$$F(u,\xi)=\int_{u_c(\xi)}^{u} f(s,\xi)\,ds \tag{3.4}$$

and E is a nonnegative real parameter of energy level (for a typical figure of F, see Fig. 8). All nonconstant solutions of (SS-1) and (SS-3) for fixed ξ are given by the E-parametrizable family of solutions in the following. From (3.3) we have

$$\left(\frac{du}{dx}\right)^2=\frac{2}{d}(E-F(u,\xi)). \tag{3.5}$$

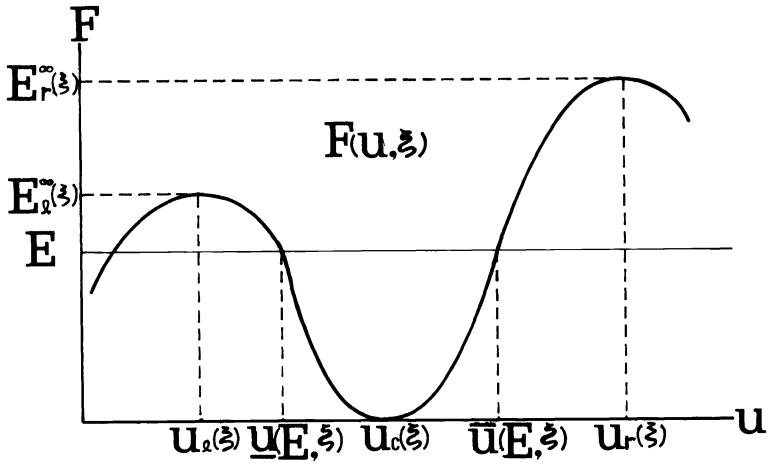

FIG. 8. *Typical functional form of* $F(u,\xi)$.

Since (SS-1) is scalar and autonomous, it is sufficient to construct the strictly increasing solutions; in fact, all the other solutions can be obtained by Lemma 2.1, if we replace U_0, D_0 by u,d. Hence we consider the equation

$$\frac{du}{dx}=\left\{\frac{2}{d}(E-F(u,\xi))\right\}^{1/2}. \tag{3.6}$$

Equation (3.6) has a nonconstant solution satisfying (SS-3) if and only if $0<E<E^{\infty}(\xi)$, where

$$E^{\infty}(\xi)=\min\{E_l^{\infty}(\xi), E_r^{\infty}(\xi)\} \qquad \text{(see (A-2))}. \tag{3.7}$$

By using separation of variables, it is easily seen that the strictly monotone increasing solution $u=u(x;E,\xi)$ of (3.6) and (SS-3) is given by the inverse function of

$$\sqrt{d(E,\xi)}\int_{\underline{u}(E,\xi)}^{u}\frac{du}{\sqrt{2(E-F(u,\xi))}}=x \tag{3.8}$$

for each fixed $0<E<E^{\infty}(\xi)$, where $d=d(E,\xi)$ is determined by the relation

$$\sqrt{\frac{1}{d(E,\xi)}}=\int_{\underline{u}(E,\xi)}^{\bar{u}(E,\xi)}\frac{du}{\sqrt{2(E-F(u,\xi))}}, \tag{3.9}$$

and $\underline{u}(E,\xi)$ and $\bar{u}(E,\xi)$ are two consecutive zeros of $E-F(u,\xi)=0$ with $u_l(\xi)<\underline{u}(E,\xi)<\bar{u}(E,\xi)<u_r(\xi)$ (see Fig. 8). The definition domain of $d=d(E,\xi)$ is

$$T=\bigcup_{\xi\in\Xi}(0,E^{\infty}(\xi))\times\{\xi\}. \tag{3.10}$$

It follows from (A-1) and (A-2) that the boundary of T consists of three parts, i.e., $E=0$, $E=E_l^{\infty}(\xi)$ for $\underline{\xi}<\xi\leq\xi^*$ and $E=E_r^{\infty}(\xi)$ for $\xi^*\leq\xi<\bar{\xi}$ (see Fig. 9).

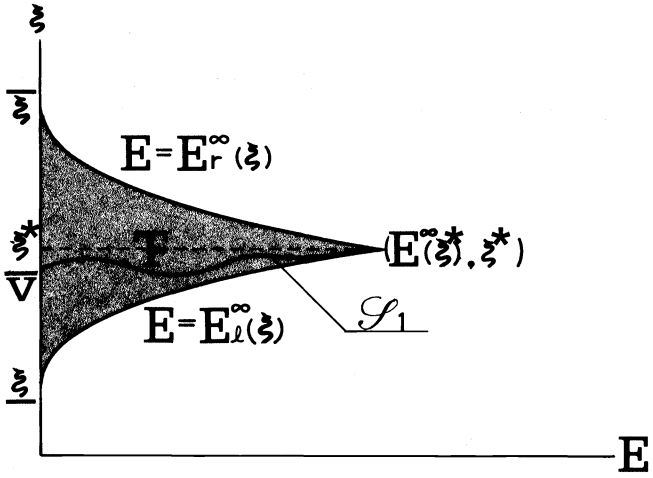

FIG. 9. *Shadow branch in T.*

We use the following notation in the next lemma:

$\underline{T}=\bigcup_{\xi\in\Xi}[0,E^{\infty}(\xi))\times\{\xi\}$,

$T_\kappa=\bigcup_{\xi\in\Xi_\kappa}(0,E^{\infty}(\xi))\times\{\xi\}$,

$\underline{T}_\kappa=\bigcup_{\xi\in\Xi_\kappa}[0,E^{\infty}(\xi))\times\{\xi\}$,

$\alpha_l(\xi)=|F_{uu}(u_l(\xi),\xi)|\quad(=|f_u(u_l(\xi),\xi)|)$,

$\alpha_r(\xi)=|F_{uu}(u_{\underline{r}}(\xi),\xi)|\quad(=|f_u(u_r(\xi),\xi)|)$,

where $\Xi_\kappa=(\underline{\xi}+\kappa,\bar{\xi}-\kappa)$ with $\kappa>0$. We note from (A-1) that $\alpha_l(\xi)\neq0$, $\alpha_r(\xi)\neq0$ and $\alpha_c(\xi)\neq0$ for $\xi\in\Xi$.

LEMMA 3.1. *For each $\xi\in\Xi$, all nonconstant strictly monotone increasing solutions of* (SS-1) *and* (SS-3) *in* $\mathfrak{S}$ *are given by the E-parameter family of solutions* $(d(E,\xi), u(x;E,\xi))$ *for* $0<E<E^{\infty}(\xi)$, *where* $d(E,\xi)$ *is defined by* (3.9) *and* $u(x;E,\xi)$ *is the inverse function of* (3.8). *Moreover, the following properties hold*:

1) $d(E,\xi)\in C^0(\underline{T}_\kappa)\cap C^{\infty}(T)$,
$\dfrac{\partial}{\partial\xi}d(E,\xi)\in C^0(\underline{T}_\kappa)$ *for any* $\kappa>0$.

2) $\lim_{E\downarrow0}d(E,\xi)=d_c(\xi)=f_u(u_c(\xi),\xi)/\pi^2$
uniformly in $\xi\in\Xi_\kappa$ *for any* $\kappa>0$.

3) $\lim_{E\uparrow E^{\infty}(\xi)}d(E,\xi)=0$
uniformly in $\xi\in\Xi_\kappa$ *for any* $\kappa>0$.

4) $u(x;E,\xi)\in C^0(\bar{I}\times\underline{T}_\kappa)\cap C^{\infty}(\bar{I}\times T)$,
$\dfrac{\partial}{\partial\xi}u(x;E,\xi)\in C^0(\bar{I}\times\underline{T}_\kappa)$ *for any* $\kappa>0$.

5) $\lim_{E\downarrow0}u(x;E,\xi)=u_c(\xi)$
uniformly in $x\in\bar{I}$ *and* $\xi\in\Xi_\kappa$ *for any* $\kappa>0$.

6)

$$\lim_{E\uparrow E^{\infty}(\xi)} u(x;E,\xi)$$

$$= \begin{cases} u_l(\xi) & \text{uniformly in } x\in[0,1-\kappa] \text{ for any } \kappa>0 \text{ if } E_l^{\infty}(\xi)<E_r^{\infty}(\xi); \\ u_{\xi^*}(x)=\begin{cases} u_l(\xi^*) & \text{for } 0\le x<\alpha_r^*/(\alpha_l^*+\alpha_r^*), \\ u_r(\xi^*) & \text{for } \alpha_r^*/(\alpha_l^*+\alpha_r^*)<x\le 1 \end{cases} \\ \qquad \text{uniformly in } x\in\bar{I}\setminus I_\kappa \text{ for any } \kappa>0 \text{ if } E_l^{\infty}(\xi)=E_r^{\infty}(\xi) \\ \qquad (i.e.,\ \xi=\xi^*,\ \text{see (A-2)}),\ \text{where } I_\kappa=(\alpha_r^*/(\alpha_l^*+\alpha_r^*)-\kappa, \\ \qquad \alpha_r^*/(\alpha_l^*+\alpha_r^*)+\kappa) \text{ and } \alpha_l^*=\alpha_l(\xi^*),\ \alpha_r^*=\alpha_r(\xi^*); \\ u_r(\xi) & \text{uniformly in } x\in[\kappa,1] \text{ for any } \kappa>0, \text{ if } E_l^{\infty}(\xi)>E_r^{\infty}(\xi). \end{cases}$$

Proof. We will only prove properties 1), 3), 4) and 6). The others are easy to prove, so we leave them to the reader.

First we change variables from u to v:

$$v=\tilde{F}(u,\xi)=\begin{cases} \{F(u,\xi)\}^{1/2}, & u_c(\xi)\le u\le \bar{u}(E), \\ -\{F(u,\xi)\}^{1/2}, & \underline{u}(E)\le u\le u_c(\xi). \end{cases} \tag{3.11}$$

It follows from the generalized Morse lemma (see, e.g., [5]) that $F(u,\xi)$ can be written as

$$F(u,\xi)=\frac{\alpha_c(\xi)}{2}(u-u_c(\xi))^2\{1+h(u,\xi)\},$$

where $\alpha_c(\xi)=F_{uu}(u_c(\xi),\xi)>0$ and $h(u,\xi)=O(|u-u_c(\xi)|)$. Therefore we have

$$\tilde{F}(u,\xi)=\sqrt{\frac{\alpha_c(\xi)}{2}}\,(u-u_c(\xi))\{1+h(u,\xi)\}^{1/2}$$

near $u=u_c(\xi)$. Thus the above transformation (3.11) is a diffeomorphism from $[\underline{u}(E), \bar{u}(E)]$ to $[-\{F(\underline{u}(E),\xi)\}^{1/2}, \{F(\bar{u}(E),\xi)\}^{1/2}]$. The right-hand side of (3.9) becomes

$$\int_{\underline{u}(E)}^{\bar{u}(E)} \frac{du}{\sqrt{2(E-F(u,\xi))}}=\int_{-\sqrt{E}}^{\sqrt{E}} \frac{1}{\sqrt{2(E-v^2)}}\frac{\partial u}{\partial v}\,dv,$$

where $\partial u/\partial v=\tilde{F}_v^{-1}(v,\xi)$. Here we note that $F(\bar{u}(E),\xi)=F(\underline{u}(E),\xi)=E$. We set $v=\sqrt{E}\,x$; then

$$\begin{aligned} &=\int_{-1}^{1}\frac{1}{\sqrt{2E(1-x^2)}}\frac{\partial \tilde{F}^{-1}\left(\sqrt{E}\,x,\xi\right)}{\partial x}dx \\ &=\int_{-1}^{1}\frac{1}{\sqrt{2(1-x^2)}}\tilde{F}_v^{-1}\left(\sqrt{E}\,x,\xi\right)dx. \end{aligned} \tag{3.12}$$

If $\tilde{D}$ represents a general differential operator with respect to E and ξ, then

$$\sup_{x\in[-1,1]} |\tilde{D}\tilde{F}_v^{-1}\left(\sqrt{E}\,x,\xi\right)|$$

is locally bounded in T. This implies from Lebesgue's theorem that $d(E,\xi)\in C^\infty(T)$. Next, noting that

$$\limsup_{E\downarrow 0,\, x\in[-1,1]} |\tilde{F}_v^{-1}(\sqrt{E}\,x,\xi)-\tilde{F}_v^{-1}(0,\xi)|=0$$

uniformly in Ξ_κ for any $\kappa>0$, and

$$\limsup_{E\downarrow 0,\, x\in[-1,1]} |\tilde{F}_{v\xi}^{-1}(\sqrt{E}\,x,\xi)-\tilde{F}_{v\xi}^{-1}(0,\xi)|=0$$

uniformly in Ξ_κ for any $\kappa>0$, we can see that $d(E,\xi)\in C^0(\underline{T}_\kappa)$ and $\partial d(E,\xi)/\partial\xi\in C^0(\underline{T}_\kappa)$.

Converting (SS-1) into the first order system of equations, we can see from 1) that 4) is a direct consequence of smooth dependence of solutions on parameters (cf. [2]).

In the proof of properties 3) and 6), the following asymptotic formula plays an essential role:

$$\int_0^\delta \frac{(\alpha x)^{-1/2}}{(x+t)^{1/2}}\,dx=-\alpha^{-1/2}\log t+O(1) \quad \text{as } t\downarrow 0, \tag{3.13}$$

for any fixed positive constant δ (see [6, p. 138]). First let us consider the case $E_l^\infty(\xi)\geq E_r^\infty(\xi)$. We will prove the formula: As $E\uparrow E_r^\infty(\xi)$,

$$\int_{\bar{u}(E)-\delta}^{\bar{u}(E)} \frac{du}{\sqrt{2(E-F(u,\xi))}}=-\frac{1}{2}\alpha_r(\xi)^{-1/2}\log\left(\frac{E_r^\infty(\xi)-E}{2E}\right)+O(1). \tag{3.14}$$

First we note from (3.12) that the left-hand side of (3.14) can be written in the form

$$\int_{1-\delta}^1 \frac{1}{\sqrt{2(1-x^2)}}\tilde{F}_v^{-1}(\sqrt{E}\,x,\xi)\,dx. \tag{3.15}$$

Here we alter the term $O(1)$ of the right-hand side of (3.14) appropriately. Since

$$\tilde{F}_u(u,\xi)=\frac{F_u(u,\xi)}{2\tilde{F}(u,\xi)},$$
$$F_u(u,\xi)=-\alpha_r(\xi)(u-u_r(\xi))(1+O(|u-u_r(\xi)|))$$

and

$$u=\tilde{F}^{-1}(v,\xi)=u_r(\xi)-\sqrt{\frac{2}{\alpha_r(\xi)}}\,\{E_r^\infty(\xi)-v^2\}^{1/2}\left(1+O\left(|E_r^\infty(\xi)-v^2|^{1/2}\right)\right),$$

we obtain

$$\begin{aligned}\tilde{F}_v^{-1}(v,\xi)&=\frac{2\tilde{F}(\tilde{F}^{-1}(v,\xi),\xi)}{F_u(\tilde{F}^{-1}(v,\xi),\xi)}\\&=\frac{2v}{F_u(\tilde{F}^{-1}(v,\xi),\xi)}\\&=\frac{\sqrt{2}\,v}{\sqrt{\alpha_r(\xi)}\,\{(E_r^\infty(\xi)-v^2)\}^{1/2}}\left(1+O\left(|E_r^\infty(\xi)-v^2|^{1/2}\right)\right).\end{aligned}$$

Therefore

$$(3.16)\quad \tilde{F}_v^{-1}\left(\sqrt{E}\,x,\xi\right)=\frac{\sqrt{2}\,x}{\sqrt{\alpha_r(\xi)}\sqrt{(1-x^2)+\dfrac{E_r^\infty(\xi)-E}{E}}}\left(1+O\left(|E_r^\infty(\xi)-Ex^2|^{1/2}\right)\right).$$

Substituting (3.16) into (3.15), we have

$$(3.15)=\int_{1-\delta}^{1}\frac{x}{\sqrt{\alpha_r(\xi)(1-x^2)}\sqrt{(1-x^2)+\dfrac{E_r^\infty(\xi)-E}{E}}}(1+O(\cdot))\,dx$$

$$=\frac{1}{2\sqrt{\alpha_r(\xi)}}\int_0^\delta\frac{y^{-1/2}dy}{\sqrt{y+\dfrac{E_r^\infty(\xi)-E}{2E}}}+O(1).$$

Here we set $y=1-x$ and, using (3.13), we have

$$=-\frac{1}{2}\alpha_r(\xi)^{-1/2}\log\left(\frac{E_r^\infty(\xi)-E}{2E}\right)+O(1).$$

Thus (3.14) is established. In the same way, if $E_l^\infty(\xi)\le E_r^\infty(\xi)$, we obtain the formula

$$(3.17)\quad \int_{\underline{u}(E)}^{\underline{u}(E)+\delta}\frac{du}{\sqrt{2(E-F(u,\xi))}}=-\frac{1}{2}\alpha_l(\xi)^{-1/2}\log\left(\frac{E_l^\infty(\xi)-E}{2E}\right)+O(1)$$

as $E\uparrow E^\infty(\xi)$. Property 3) follows directly from (3.14) and/or (3.17).

Now we are ready to prove property 6). In the case of $E_l^\infty(\xi)<E_r^\infty(\xi)$, the integral

$$\int_{\underline{u}(E)+\delta}^{\bar{u}(E)}\frac{du}{\sqrt{2(E-F(u,\xi))}}$$

is bounded as $E\uparrow E^\infty(\xi)(=E_l^\infty(\xi))$ for any small positive δ, since the integrand has a singularity of order $-\frac{1}{2}$ at $\bar{u}(E)$ as $E\uparrow E^\infty(\xi)$. Noting this fact and (3.17), we can see that, for any $\kappa>0$ and $\varepsilon>0$, there exists $E_0=E_0(\kappa,\varepsilon)$ such that for any E with $E_0<E<E^\infty(\xi)$, the following inequalities hold:

$$(3.18)\quad \begin{aligned}&|\underline{u}(E)-u_l(\xi)|<\frac{\varepsilon}{2},\\ &\sqrt{d(E,\xi)}\int_{\underline{u}(E)}^{\underline{u}(E)+\varepsilon/2}\frac{du}{\sqrt{2(E-F(u,\xi))}}>1-\kappa.\end{aligned}$$

This implies that $u(E,\xi)$ converges uniformly to $u_l(\xi)$ on any compact set $[0,1-\kappa]$ as $E\uparrow E^\infty(\xi)$. We can prove the corresponding result analogously in the case of $E_l^\infty(\xi)>E_r^\infty(\xi)$. Finally, we consider the case $E_l^\infty(\xi)=E_r^\infty(\xi)$. It follows from (A-2) that this case occurs if and only if $\xi=\xi^*$. Both asymptotic formulas (3.14) and (3.17) are valid in this case. Since the integral

$$\int_{\underline{u}(E)+\delta}^{\underline{u}(E)-\delta}\frac{du}{\sqrt{2(E-F(u,\xi^*))}}$$

is bounded as $E \uparrow E^{\infty}(\xi^*)$ for any small positive constant δ, we can see from (3.9), (3.14) and (3.17) that the following two limits hold:

$$\lim_{E\uparrow E^{\infty}(\xi^*)} \sqrt{d(E,\xi^*)} \int_{\underline{u}(E)}^{\underline{u}(E)+\delta} \frac{du}{\sqrt{2(E-F(u,\xi^*))}} = \frac{\alpha_r(\xi^*)^{1/2}}{\alpha_l(\xi^*)^{1/2}+\alpha_r(\xi^*)^{1/2}},$$

(3.19)

$$\lim_{E\uparrow E^{\infty}(\xi^*)} \sqrt{d(E,\xi^*)} \int_{\bar{u}(E)-\delta}^{\bar{u}(E)} \frac{du}{\sqrt{2(E-F(u,\xi^*))}} = \frac{\alpha_l(\xi^*)^{1/2}}{\alpha_l(\xi^*)^{1/2}+\alpha_r(\xi^*)^{1/2}}.$$

These imply the required result. Thus the proof is completed.

In order to solve (SS), all we have to do is to pick out all solutions from $\{u(x;E,\xi)\}_{(E,\xi)\in T}$ which satisfy (SS-2); i.e., we must find all solutions of the scalar equation

$$B(E,\xi)=\int_0^1 g(u(x;E,\xi),\xi)\,dx=0. \tag{3.20}$$

From lemma 3.1 we can see that $B(E,\xi)$ satisfies the following properties:

LEMMA 3.2.

1) $B(E,\xi)\in C^0(\underline{T})\cap C^{\infty}(T)$, $B_{\xi}(E,\xi)\in C^0(\underline{T}_{\kappa})$.

2) $B(0,\xi)=g(u_c(\xi),\xi)$.

3) $\lim_{(E,\xi)\to(E_l(\hat{\xi}),\hat{\xi})} B(E,\xi)=g(u_l(\hat{\xi}),\hat{\xi})$ *for* $\underline{\xi}<\hat{\xi}<\xi^*$,

$\lim_{(E,\xi)\to(E_r(\hat{\xi}),\hat{\xi})} B(E,\xi)=g(u_r(\hat{\xi}),\hat{\xi})$ *for* $\xi^*<\hat{\xi}<\bar{\xi}$.

Remark 3.1. $B(E,\xi)$ cannot be continuously extended to $\overline{T}$, since $B(E_l^{\infty}(\xi),\xi)$ and $B(E_r^{\infty}(\xi),\xi)$ do not coincide with each other at $\xi=\xi^*$. However, noting that for $\xi\neq\xi^*$ the convergence in property 6) of Lemma 3.1 is a uniform one for an appropriate neighborhood of ξ, we can see that it can be continuously extended to $\overline{T}\backslash\{(E^{\infty}(\xi^*),\xi^*)\}$.

Remark 3.2. It follows from (A-1), (A-3) and Remark 0.2 that there exists a positive constant δ independent of ξ such that $g(u_l(\xi),\xi)<-\delta$ and $g(u_r(\xi),\xi)>\delta$ hold for $\xi\in\Xi$. We extend the domain of definition of $B(E,\xi)$ for $E<0$ as follows:

$$B(E,\xi)=B(0,\xi) \quad \text{for } E<0, \quad \xi\in\Xi. \tag{3.21}$$

This extension preserves the continuity of $B_{\xi}(E,\xi)$ for $E\leq 0$.

Now we can prove the main theorem in this section.

THEOREM 3.3. *The component (i.e., maximal closed connected set)* $\mathbb{S}_1$ *in* $\underline{T}$ *of solutions of* (3.20), *which contains* $(0,\bar{v})$, *exists globally with respect to E in the sense that* $\overline{\mathbb{S}}_1\cap T^{\infty}\neq\emptyset$ *where* $T^{\infty}=\{(E^{\infty}(\xi),\xi)\,|\,\xi\in\Xi\}$, *and its intersection consists of only one element*:

$$\overline{\mathbb{S}}_1\cap T^{\infty}=\{(E^{\infty}(\xi^*),\xi^*)\}. \tag{3.22}$$

Here we note that $\overline{\mathbb{S}}_1$ *denotes the closure of* $\mathbb{S}_1$ *in* $\mathbb{R}^2$. *Moreover, there exists a positive constant* δ *such that* $\mathrm{Sec}_E(\mathbb{S}_1)$ *consists of only one element for* $0\leq E<\delta$, *where* $\mathrm{Sec}_E(\mathbb{S}_1)=\{\xi\,|\,(E,\xi)\in\mathbb{S}_1\}$ *for fixed E. Therefore* $\overline{\mathbb{S}}_1$ *is wholly contained in T except for the starting*

point $(0,\bar{v})$ *and the terminal point* $(E^{\infty}(\xi^*), \xi^*)$. *Finally,* (3.22) *means that a pair of solutions* $(u(x;E,\xi), \xi)$ *of* (SS) *corresponding to* $\mathfrak{S}_1$ *satisfies the property*

$$u(x;E,\xi)\to u^*(x)=\begin{cases} u_l(\xi^*), & 0\le x<x^*, \\ u_r(\xi^*), & x^*<x\le 1 \end{cases} \tag{3.23}$$

uniformly in $[0,x^*-\kappa]$ *and* $[x^*+\kappa,1]$ *for any* $\kappa>0$,

and $\xi\to\xi^*$ *as* $(E,\xi)\to(E^{\infty}(\xi^*),\xi^*)$, *where* x^* *is uniquely determined by the relation*

$$0=\int_0^1 g(u^*(x),\xi^*)\,dx. \tag{3.24}$$

Remark 3.3. It follows from Lemma 3.1, property 3) that "globally exist with respect to E" is equivalent to "globally exist with respect to d_1" as stated in §2.

Remark 3.4. Given an arbitrary path in T which goes into the point $(E^{\infty}(\xi^*),\xi^*)$, the position of an interior transition layer of the step function obtained as the limit of the given path as $E\uparrow E^{\infty}(\xi^*)$ is closely related to the asymptotic form of its path as $E\uparrow E^{\infty}(\xi^*)$. For the details, see Appendix 1.

Proof of Theorem 3.3. First we construct the local solutions near $(E,\xi)=(0,\bar{v})$. Let us differentiate (3.20) by ξ,

$$B_{\xi}(E,\xi)=\int_0^1\left\{g_u(u(x;E,\xi),\xi)\frac{\partial}{\partial\xi}u(x;E,\xi)+g_{\xi}(u(x;E,\xi),\xi)\right\}dx.$$

At $(E,\xi)=(0,\bar{v})$, we have

$$B_{\xi}(0,\bar{v})=\int_0^1\left\{g_u(\bar{u},\bar{v})\frac{du_c(\xi)}{d\xi}\bigg|_{\xi=\bar{v}}+g_{\xi}(\bar{u},\bar{v})\right\}dx. \tag{3.25}$$

On the other hand, $f(u_c(\xi),\xi)=0$ holds identically in the vicinity of $\xi=\bar{v}$. Differentiating this equation by ξ, we have, at $\xi=\bar{v}$,

$$f_u(\bar{u},\bar{v})\frac{du_c(\xi)}{d\xi}\bigg|_{\xi=\bar{v}}+f_{\xi}(\bar{u},\bar{v})=0. \tag{3.26}$$

Substituting (3.26) into (3.25), we have

$$B_{\xi}(0,\bar{v})=\{f_u(\bar{u},\bar{v})g_{\xi}(\bar{u},\bar{v})-f_{\xi}(\bar{u},\bar{v})g_u(\bar{u},\bar{v})\}/f_u(\bar{u},\bar{v}).$$

Therefore, by (A-4),

$$B_{\xi}(0,\bar{v})\neq 0. \tag{3.27}$$

Thus, by the implicit function theorem, we obtain the unique solution $\xi=\xi(E)$ of (3.20) in some neighborhood of $(0,\bar{v})$. Here we note that (A-3) and Lemma 3.1, property 5) imply that $\xi=\bar{v}$ is the unique solution of (3.20) when $E=0$, and that there exists a positive constant δ_1 such that we have no solutions of (3.20) for $\xi\in(\bar{\xi}-\delta_1,\bar{\xi})$ or $(\bar{\xi},\bar{\xi}+\delta_1)$. Therefore the continuity of $u(x;E,\xi)$ implies that $\mathrm{Sec}_E(\mathfrak{S}_1)$ consists of only one element for $0\le E\le\delta$ with δ an appropriate positive constant.

Next we show the global existence of $\mathfrak{S}_1$ with respect to E. When we consider $B(E,\xi)$ to be a function of ξ, a finite dimensional degree

$$\deg(B(E,\xi),\mathrm{Sec}_E(T),0),$$

where $\mathrm{Sec}_E(T)=\{\xi\,|\,(E,\xi)\in T\}$ for fixed E, is well defined. It is clear from the above discussion that

$$\deg\left(B(0,\xi),\,\mathrm{Sec}_0(T),\,0\right)\neq 0$$

holds. Since solutions of (3.20) never lie on the boundary of $\mathrm{Sec}_E(T)$ (see Lemma 3.2, property 3)), it follows from the homotopy invariance property of degree that

$$\deg\left(B(E,\xi),\,\mathrm{Sec}_E(T),\,0\right)\neq 0 \tag{3.28}$$

for $0\leq E<E^{\infty}(\xi^*)$. Hence the required result follows from (3.28).

Lastly we prove the final statement. It follows from the asymptotic formulas (3.14) and (3.17) that, for any $\varepsilon>0$, there exist a positive constant $\delta(<\varepsilon)$ and a closed ε-interval $I_\varepsilon=[x_c-\varepsilon/2,\,x_c+\varepsilon/2]$ $(\subset I)$ such that the values of $u(x;E,\xi)\in\mathfrak{S}_1$ belong to $(u_r(\xi^*)-\varepsilon,\,u_r(\xi^*)]$ or $[u_l(\xi^*),\,u_l(\xi^*)+\varepsilon)$ for $x\in I\setminus I_\varepsilon$ and $|E-E^{\infty}(\xi^*)|+|\xi-\xi^*|<\delta$. Since g is smooth, there exist positive constants K_1 and K_2 such that, for $|E-E^{\infty}(\xi^*)|+|\xi-\xi^*|<\delta$, we have

$$|g(u_r(\xi^*),\xi^*)-g(u(x;E,\xi),\xi)|\leq K_1\varepsilon \quad \text{for } x\in I_r, \tag{3.29}_a$$

$$|g(u_l(\xi^*),\xi^*)-g(u(x;E,\xi),\xi)|\leq K_1\varepsilon \quad \text{for } x\in I_l, \tag{3.29}_b$$

$$\left|\int_{I_\varepsilon} g(w(x),\xi)\,dx\right|\leq K_2\varepsilon, \tag{3.29}_c$$

for any function $w(x)$ with values in $[u_{\min},u_{\max}]$, where $I_r=(x_c+\varepsilon/2,1]$ and $I_l=[0,x_c-\varepsilon/2)$. We define the step function $u^*(x;x_c)$ by

$$u^*(x;x_c)=\begin{cases} u_l(\xi^*), & 0\leq x<x_c,\\ u_r(\xi^*), & x_c<x\leq 1.\end{cases} \tag{3.30}$$

It follows from (3.29) that

$$\left|\int_0^1 \{g(u^*(x;x_c),\xi^*)-g(u(x;E,\xi),\xi)\}\,dx\right|\leq K_3\varepsilon \tag{3.31}$$

holds for an appropriate constant $K_3>0$. Since $u(x;E,\xi)$ is a solution of (SS), (3.31) becomes

$$\left|\int_0^1 g(u^*(x;x_c),\xi^*)\,dx\right|\leq K_3\varepsilon. \tag{3.32}$$

On the other hand, the function $G(t)$ defined by

$$\begin{aligned} G(t)&=\int_0^1 g(u^*(x;t),\xi^*)\,dx\\ &=\{g(u_l(\xi^*),\xi^*)-g(u_r(\xi^*),\xi^*)\}t+g(u_r(\xi^*),\xi^*)\end{aligned}$$

is a strictly monotone decreasing linear function of t, and has a unique zero at $t=x^*$. Therefore it follows from (3.32) that

$$|x_c-x^*|\leq K_4\varepsilon \tag{3.33}$$

for some positive constant K_4 independent of ε. The estimate (3.33) leads to the conclusion. Thus Theorem 3.3 is established.

The above theorem states that the bifurcating branch starts from $(d_c(\bar v),\,\bar U)$ and goes into $(0,\,(u^*(x),\,\xi^*))$ in the space $\mathfrak{S}$. In the vicinity of $(d_c(\bar v),\,\bar U)$, its branch is a

smooth curve. Then what is the structure of the component $\mathfrak{S}_1$ between the two ends $(0,\bar{v})$ and $(E^\infty(\xi^*),\xi^*)$, and how many branches hit the terminal point $(E^\infty(\xi^*), \xi^*)$? We shall consider these problems in the following two subsections.

3.2. Uniqueness of the shadow branch near $(E^\infty(\xi^*), \xi^*)$. In the previous subsection we have shown that the shadow branch exists globally with respect to E and its starting point and terminal point are uniquely determined. Moreover, near the starting point, $\mathfrak{S}_1$ consists of a unique curve. Then what is the structure of $\mathfrak{S}_1$ near the terminal point $(E^\infty(\xi^*), \xi^*)$? Since $d(E,\xi)$ becomes zero as (E,ξ) tends to this point, this problem is related to the uniqueness of the singularly perturbed solutions of (SS). We have already reduced (SS) to the problem of finding the zeros of $B(E,\xi)$. Therefore the problem is how many zeros are there in the vicinity of $(E^\infty(\xi^*), \xi^*)$ for fixed E? The answer is only one; i.e., $\mathrm{Sec}_E(\mathfrak{S}_1)$ consists of only one element in some small neighborhood of $(E^\infty(\xi^*), \xi^*)$. In this subsection we will prove this result, which plays an essential role in §6.

First we rewrite the right-hand side of (3.20).

LEMMA 3.4.

$$\int_0^1 g(u(x;E,\xi),\xi)\,dx = \frac{\displaystyle\int_{\underline{u}(E,\xi)}^{\bar{u}(E,\xi)} \frac{g(u,\xi)}{\sqrt{2(E-F(u,\xi))}}\,du}{\displaystyle\int_{\underline{u}(E,\xi)}^{\bar{u}(E,\xi)} \frac{1}{\sqrt{2(E-F(u,\xi))}}\,du}.$$

Proof. E and ξ are fixed throughout the proof, so we write $u(x)$ instead of $u(x;E,\xi)$, and so on. We assume, for simplicity, that g is monotonic with respect to u for fixed ξ (if g is not monotonic, we can prove the above formula by dividing the interval into several parts where in each subinterval g is monotonic).

First we set $z(x)=g(u(x))$. Since u and g are monotonic, $z(x)$ is a monotonic function (see Fig. 10). Denoting the inverse function of $z(x)$ by $x(z)$, we have

$$\begin{aligned}\int_0^1 g(u(x))\,dx &= g(\bar{u}) - \int_0^{g(\bar{u})} x(z)\,dz - \int_{g(\underline{u})}^0 x(z)\,dz \\ &= g(\bar{u}) - \int_{g(\underline{u})}^{g(\bar{u})} x(z)\,dz.\end{aligned} \tag{3.34}$$

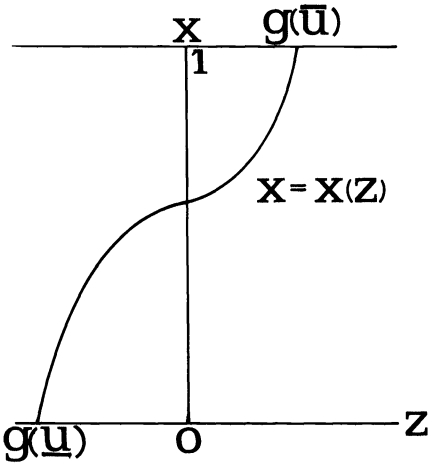

FIG. 10. *A graph of* $x=x(z)$.

If we change variables from z to u, the second term of (3.34) (which we denote by J) becomes

$$J=\int_{g(\underline{u})}^{g(\bar{u})}x(z)\,dz=\int_{\underline{u}}^{\bar{u}}x(g(u))g'(u)\,du,$$

and noting that (see (3.8))

$$x(z)=\sqrt{d}\int_{\underline{u}}^{g^{-1}(z)}\frac{dv}{\sqrt{2(E-F(v))}},$$

we have

$$J=\int_{\underline{u}}^{\bar{u}}\sqrt{d}\int_{\underline{u}}^{u}\frac{dv}{\sqrt{2(E-F(v))}}g'(u)\,du.$$

Integrating by parts, we obtain

$$\begin{aligned}&=\sqrt{d}\left[(g(u)-g(\bar{u}))\int_{\underline{u}}^{u}\frac{dv}{\sqrt{2(E-F(v))}}\right]_{\underline{u}}^{\bar{u}}-\sqrt{d}\int_{\underline{u}}^{\bar{u}}\frac{(g(u)-g(\bar{u}))}{\sqrt{2(E-F(u))}}du\\ &=\sqrt{d}\int_{\underline{u}}^{\bar{u}}\frac{g(\bar{u})-g(u)}{\sqrt{2(E-F(u))}}du.\end{aligned}\tag{3.35}$$

Substituting (3.35) into (3.34) and noting that

$$\sqrt{d}\int_{\underline{u}}^{\bar{u}}\frac{du}{\sqrt{2(E-F(u))}}=1,$$

we obtain the conclusion.

We will show that $B(E,\xi)$ is strictly monotone increasing with respect to $\xi\in \mathrm{Sec}_E(T)$ when (E,ξ) is near $(E^{\infty}(\xi^*),\xi^*)$, which implies the unique existence of the solution of $B(E,\xi)=0$ for fixed E. For this purpose it suffices to prove the following:

LEMMA 3.5. *The derivative of $B(E,\xi)$ with respect to ξ is strictly positive when (E,ξ) $(\in T)$ belongs to the sufficiently small neighborhood of $(E^{\infty}(\xi^*),\xi^*)$. More precisely, we have*

$$\begin{aligned}B_{\xi}(E,\xi)=&\{K(E,\xi)E_r^{\infty\prime}(\xi)R(E,\xi)^{-1}\log L(E,\xi)\\ &-K(E,\xi)E_l^{\infty\prime}(\xi)L(E,\xi)^{-1}\log R(E,\xi)\}(1+o(1))\Big/\\ &\left\{\frac{1}{2\sqrt{\alpha_r(\xi)}}\log R(E,\xi)+\frac{1}{2\sqrt{\alpha_l(\xi)}}\log L(E,\xi)\right\}^2\\ >0\quad &\text{for } \xi\in \mathrm{Sec}_E(T),\end{aligned}\tag{3.36}$$

where

$$\begin{aligned}&K(E,\xi)=g(\bar{u}(E,\xi),\xi)-g(\underline{u}(E,\xi),\xi),\\ &R(E,\xi)=\{E_r^{\infty}(\xi)-E\}/2E,\\ &L(E,\xi)=\{E_l^{\infty}(\xi)-E\}/2E\end{aligned}$$

and $'$ denotes $d/d\xi$.

Proof. Using Lemma 3.4, we have

$$B_\xi(E,\xi)=\left\{\frac{\partial I_2}{\partial \xi}I_1-\frac{\partial I_1}{\partial \xi}I_2\right\}\Big/ I_1^2, \tag{3.37}$$

where

$$I_1=\int_{\underline{u}}^{\bar{u}}\frac{du}{\sqrt{2(E-F(u,\xi))}},\qquad I_2=\int_{\underline{u}}^{\bar{u}}\frac{g(u,\xi)}{\sqrt{2(E-F(u,\xi))}}du.$$

Let us calculate the asymptotic expansions of the four terms I_1, I_2, $\partial I_1/\partial\xi$ and $\partial I_2/\partial\xi$ as $(E,\xi)\sim(E^\infty(\xi^*),\xi^*)$. First we note that $|E-E_l^\infty(\xi)|+|E-E_r^\infty(\xi)|$ becomes zero when (E,ξ) tends to $(E^\infty(\xi^*),\xi^*)$ (see Fig. 9 above). Therefore it follows from (3.14) and (3.17) that I_1 is expanded as

$$I_1=-\frac{1}{2}\left\{\alpha_r(\xi)^{-1/2}\log R(E,\xi)+\alpha_l(\xi)^{-1/2}\log L(E,\xi)\right\}+O(1) \tag{3.38}$$

when (E,ξ) is near $(E^\infty(\xi^*),\xi^*)$. Analogously, I_2 can be expanded as

$$I_2=-\frac{1}{2}\left\{\alpha_r(\xi)^{-1/2}g(\bar{u},\xi)\log R(E,\xi)+\alpha_l(\xi)^{-1/2}g(\underline{u},\xi)\log L(E,\xi)\right\}+O(1) \tag{3.39}$$

when (E,ξ) is in the vicinity of $(E^\infty(\xi^*),\xi^*)$.

Next we consider the term $\partial I_1/\partial\xi$. It follows from (3.12) that

$$\frac{\partial I_1}{\partial\xi}=\frac{\partial}{\partial\xi}\int_{-1}^{1}\frac{1}{\sqrt{2(1-x^2)}}\tilde{F}_v^{-1}\left(\sqrt{E}\,x,\xi\right)dx.$$

We divide the interval $[-1,1]$ into three parts,

$$\frac{\partial I_1}{\partial\xi}=\frac{\partial}{\partial\xi}\left(\int_{-1}^{-1+\delta}+\int_{-1+\delta}^{1-\delta}+\int_{1-\delta}^{1}\right),$$

where δ is some small positive constant. Since $(\partial/\partial\xi)\int_{-1+\delta}^{1-\delta}=O(1)$, we must expand the other two terms $(\partial/\partial\xi)\int_{-1}^{-1+\delta}$, $(\partial/\partial\xi)\int_{1-\delta}^{1}$. First let us consider the integral $(\partial/\partial\xi)\int_{1-\delta}^{1}$. Noting that

$$\tilde{F}_v^{-1}\left(\sqrt{E}\,x,\xi\right)=2\sqrt{E}\,x/F_u\left(\tilde{F}^{-1}\left(\sqrt{E}\,x,\xi\right),\xi\right)$$

(see the proof of Lemma 3.1), we obtain

$$\begin{aligned}\frac{\partial}{\partial\xi}\int_{1-\delta}^{1}&=\int_{1-\delta}^{1}\frac{\sqrt{2E}\,x}{\sqrt{(1-x^2)}}\frac{\partial}{\partial\xi}\left(\frac{1}{F_u\left(\tilde{F}^{-1}\left(\sqrt{E}\,x,\xi\right),\xi\right)}\right)dx\\&=-\int_{1-\delta}^{1}\frac{\sqrt{2E}\,x}{\sqrt{1-x^2}}\left\{F_{uu}\tilde{F}_\xi^{-1}+F_{u\xi}\right\}/F_u^2\,dx.\end{aligned} \tag{3.40}$$

Let w and z be

$$\begin{aligned}&w=\left\{F(u_r(\xi),\xi)-v^2\right\}^{1/2}=\left\{E(1-x^2)+\left(E_r^\infty(\xi)-E\right)\right\}^{1/2},\\&z=u-u_r(\xi),\end{aligned} \tag{3.41}$$

where $v=\sqrt{E}\,x$. Using Taylor's expansion of $F(u,\xi)$ and (3.11), we can see that

$$
\begin{aligned}
&z=-\sqrt{\frac{2}{\alpha_r(\xi)}}\,w(1+O(|w|)),\\
&F_u=-\sqrt{2\alpha_r(\xi)}\,w(1+O(|w|)),\\
&F_\xi=E_r^{\infty\prime}(\xi)-\sqrt{2\alpha_r(\xi)}\,u_r'(\xi)w(1+O(|w|)),\\
&F_{uu}=-\alpha_r(\xi)(1+O(|w|)),\\
&F_{u\xi}=\alpha_r(\xi)u_r'(\xi)(1+O(|w|)),
\end{aligned}
\tag{3.42}
$$

and noting that $\tilde{F}_\xi^{-1}=F_\xi/F_u$ $(u\neq u_r(\xi))$, we have

$$
\begin{aligned}
\{F_{uu}\tilde{F}_\xi^{-1}+F_{u\xi}\}/F_u^2=\frac{1}{2\alpha_r(\xi)w^2}\Bigg\{&\sqrt{\frac{\alpha_r(\xi)}{2}}\,E_r^{\infty\prime}(\xi)w^{-1}(1+O(|w|))\\
&+\alpha_r(\xi)u_r'(\xi)(1+O(|w|))\Bigg\}(1+O(|w|)).
\end{aligned}
\tag{3.43}
$$

Here the following asymptotic formulas are needed.

LEMMA 3.6. *For a positive constant* δ, *we have as* $t\downarrow 0$

1) $$\int_0^\delta \frac{x^{-1/2}}{(x+t)}\,dx=\pi t^{-1/2}(1+o(1)).$$

2) $$\int_0^\delta \frac{x^{-1/2}}{(x+t)^{3/2}}\,dx=2t^{-1}(1+o(1)).$$

Proof. Taking $y=x/t$ as a new variable, we can easily derive the above formulas, so the details are left to the reader (see [6, p. 138]).

Combining (3.40), (3.43) with Lemma 3.6, we obtain

$$
(3.40)=-\frac{E_r^{\infty\prime}(\xi)}{4E\sqrt{\alpha_r(\xi)}}R(E,\xi)^{-1}+O\big(R(E,\xi)^{-1/2}\big).
\tag{3.44}
$$

For $(\partial/\partial\xi)\int_{-1}^{-1+\delta}$, we obtain the same result as (3.44) replacing r by l. Since $g(u,\xi)$ has no singular points, $\partial I_2/\partial\xi$ can be expanded in an analogous way. Thus we obtain

$$
\begin{aligned}
\frac{\partial I_1}{\partial\xi}=-\frac{1}{4E}\Bigg\{&\frac{E_r^{\infty\prime}(\xi)}{\sqrt{\alpha_r(\xi)}}R(E,\xi)^{-1}+O\big(R(E,\xi)^{-1/2}\big)\\
&+\frac{E_l^{\infty\prime}(\xi)}{\sqrt{\alpha_l(\xi)}}L(E,\xi)^{-1}+O\big(L(E,\xi)^{-1/2}\big)\Bigg\}+O(1),
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial I_2}{\partial\xi}=-\frac{1}{4E}\Bigg\{&\frac{g(\bar{u},\xi)E_r^{\infty\prime}(\xi)}{\sqrt{\alpha_r(\xi)}}R(E,\xi)^{-1}+O\big(R(E,\xi)^{-1/2}\big)\\
&+\frac{g(\underline{u},\xi)E_l^{\infty\prime}(\xi)}{\sqrt{\alpha_l(\xi)}}L(E,\xi)^{-1}+O\big(L(E,\xi)^{-1/2}\big)\Bigg\}+O(1).
\end{aligned}
\tag{3.45}
$$

Substituting (3.38), (3.39) and (3.45) into (3.37) and taking the principal part, we obtain (3.36). Since $g(\bar{u},\xi)>0$, $g(\underline{u},\xi)<0$, $E_r^{\infty\prime}(\xi)<0$ and $E_l^{\infty\prime}(\xi)>0$ (see (A-2)), the right-hand side of (3.36) is strictly positive. Thus the proof of Lemma 3.5 is completed.

As a direct consequence of Lemma 3.5, we obtain

THEOREM 3.7. *There exists some neighborhood U^* of $(E^\infty(\xi^*), \xi^*)$ and $\bar{\delta}>0$ such that $\mathrm{Sec}_E(\mathcal{S}_1)\cap U^*$ consists of only one element for $E^\infty(\xi^*)-\bar{\delta}<E<E^\infty(\xi^*)$.*

3.3. Local arcwiseness of the shadow branch. So far, we have studied the unique existence of the shadow branch $\mathcal{S}_1$ when (E,ξ) is near $(0,\bar{v})$ or $(E^\infty(\xi^*), \xi^*)$. In this subsection, we reveal the structure of $\mathcal{S}_1$ between these two ends by using Sard's lemma: $\mathcal{S}_1$ is generically a one-dimensional submanifold in T, i.e., a locally smooth curve. The definite meaning of the genericity will be given in Theorem 3.9.

First we state the well-known Sard's lemma:

LEMMA 3.8 (Sard's lemma). *Let $\mathcal{O}$ be an open set in $\mathbb{R}^k$ $(k\geq 1)$ and let $f\colon \mathcal{O}\to\mathbb{R}^1$ be a function of C^∞-class. Then the set C of critical values of f has measure zero in $\mathbb{R}^1$. Moreover, if $\bar{\mathcal{O}}$ is compact and f is real analytic in $\bar{\mathcal{O}}$, C is a finite set.*

Proof. For the proof, see, e.g., [11].

We recall that $\mathring{\mathcal{S}}_1=\mathcal{S}_1\cap T$ is the preimage of a zero of $B(E,\xi)$ in T. We note that $\mathring{\mathcal{S}}_1=\mathcal{S}_1\setminus\{(0,\bar{v})\}$ (see Theorem 3.3) and $B(E,\xi)$ is of C^∞-class from Lemma 3.2. Therefore whether or not $\mathring{\mathcal{S}}_1$ is a one-dimensional submanifold of T depends on whether or not zero is a regular value of B. This is what is called the "preimage theorem" (see [12]). On the other hand, Lemma 3.8 says that the set of critical values (i.e., nonregular values) has measure zero. Hence it is plausible that zero is not a critical value in almost all cases. In fact we know:

DENSITY THEOREM. *In any small neighborhood in the Whitney C^∞-topology of the given function B, there exists a function $\bar{B}$ such that zero is a regular value.*

This is a special case of Thom's transversality theorem (for the details, see, e.g., [11]). However, in practical applications, this theorem is not tractable because, in general, all small perturbations for the nonlinearities f and g which do not destroy the functional form required by the given model do not correspond to all small perturbations for B. For example, if f and g are of polynomial type, we can only vary their coefficients. However, in the next theorem we give a useful result that some perturbations for g which preserve the functional form bring B into the general situs.

THEOREM 3.9. *Suppose that there exist an admissible perturbation $\varepsilon h(u,\xi)$ with ε a small real parameter, and a positive constant ε_0 such that for $\varepsilon\in[-\varepsilon_0,\varepsilon_0]$ $g_\varepsilon(u,\xi)=g(u,\xi)+\varepsilon h(u,\xi)$ has the functional form required by the given model and that*

$$\int_0^1 h(u(x;E,\xi),\xi)\,dx\neq 0 \quad \text{for } (E,\xi)\in T. \tag{3.46}$$

Then, for almost everywhere $\varepsilon\in[-\varepsilon_0,\varepsilon_0]$, zero is a regular value of $B_\varepsilon(E,\xi)$, where

$$B_\varepsilon(E,\xi)=\int_0^1 g_\varepsilon(u(x;E,\xi),\xi)\,dx. \tag{3.47}$$

In other words, if we replace g in (SP) *by g_ε, its shadow branch in T is a one-dimensional submanifold of T for almost every ε. Therefore, it has no secondary bifurcation points.*

Proof. First we note that the preimage of zero of $B(E,\xi)$ is the same as that of the following $\hat{B}(E,\xi)$:

$$\hat{B}(E,\xi)=\frac{B(E,\xi)}{\int_0^1 h(u(x;E,\xi),\xi)\,dx}.$$

Analogously we can define $\hat{B}_\varepsilon(E,\xi)$ for $B_\varepsilon(E,\xi)$ which has the same preimage of zero as that of B_ε. On the other hand, we have

$$\hat{B}_\varepsilon(E,\xi)=\frac{B(E,\xi)+\varepsilon\int_0^1 h(u(x;E,\xi),\xi)\,dx}{\int_0^1 h(u(x;E,\xi),\xi)\,dx}=\hat{B}(E,\xi)+\varepsilon.$$

Therefore whether or not zero is a regular value of $\hat{B}_\varepsilon$ depends on whether or not $-\varepsilon$ is a regular value of $\hat{B}$. Hence, applying Lemma 3.8 to $\hat{B}$, we obtain the result that zero is a regular value of $\hat{B}_\varepsilon$ for almost every ε. This implies the conclusion.

Remark 3.5. If f and g are real analytic in R, $B(E,\xi)$ is also real analytic in any $\mathcal{O}$ with $\bar{\mathcal{O}}\subset T$. It follows from Lemma 3.8 that zero is a regular value of $B_\varepsilon(E,\xi)|_{\mathcal{O}}$ (restriction to $\mathcal{O}$) for "all small ε except possibly $\varepsilon=0$". Combining Theorem 3.3, Theorem 3.7 and this result, we can see that Theorem 3.9 holds with "almost everywhere ε" replaced by "all ε except possibly $\varepsilon=0$" if f and g are real analytic.

Since the nonlinearity of polynomial type frequently appears in applications, the next corollary is useful.

COROLLARY 3.10.

1) *If $g(u,\xi)$ has a constant term, one can take $h(u,\xi)=1$.*

2) *If $g(u,\xi)$ has a monomial ξ^m $(m\geq 1)$ and $0\notin\Xi$, one can take $h(u,\xi)=\xi^m$.*

3) *If $g(u,\xi)$ has a monomial u^n $(n\geq 1)$ and $0\notin[u_{\min},u_{\max}]$, one can take $h(u,\xi)=u^n$.*

We give several examples which serve to illustrate the above results. First let us consider the *prey-predator model* given by Example 1 in §0. $B(E,\xi)$ becomes

$$B(E,\xi)=\int_0^1\{g_0(\xi)-ku(x;E,\xi)\}\xi\,dx,$$

where $g_0(\xi)=c_0+c_1\xi^m(c_0,c_1,m>0)$. Noting that $\xi>0$, we can take $h(u,\xi)=\xi$ or ξ^{m+1}.

Next we consider the *Gierer-Meinhardt model* (Example 2 in §0). In this case

$$B(E,\xi)=\int_0^1\left\{c'\rho'u(x;E,\xi)^2-\nu\xi\right\}dx.$$

One can take $h(u,\xi)=\xi$, since $\underline{\xi}>0$. It is also possible to set $h(u,\xi)=u^2$ since $u_{\min}=h_1(\underline{\xi})>0$ (see Fig. 4).

Finally for *Seelig's model* (Example 3 in §0), we can apply, for example, Corollary 3.10, part 1) to it, since g has a constant term j_2.

4. Approximation theorem for large d_2. Let us bring the shadow branch $\mathfrak{S}_1$ of Theorem 3.3 into the space $\mathfrak{E}$. Using the mapping $d_1=d(E,\xi)$ in Lemma 3.1, we have the corresponding shadow branches $\mathcal{C}^1_{0,+}$ and $\mathcal{C}^1_{0,-}$ in $\mathfrak{E}$:

$$\mathcal{C}^1_{0,+}=\{(d(E,\xi),(u(x;E,\xi),\xi))\,|\,(E,\xi)\in\mathfrak{S}_1\},\tag{4.1}$$

$$\mathcal{C}^1_{0,-}=\{(d(E,\xi),(u(1-x;E,\xi),\xi))\,|\,(E,\xi)\in\mathfrak{S}_1\}.\tag{4.2}$$

$\mathcal{C}^1_{0,-}$ is the reflection image of $\mathcal{C}^1_{0,+}$. We set

$$\mathcal{C}^1_0=\mathcal{C}^1_{0,+}\cup\mathcal{C}^1_{0,-}.$$

$\mathcal{C}^1_{0,+}$ (resp. $\mathcal{C}^1_{0,-}$) connects $(d_c(\bar{v}),\overline{U})$ to $(0,(u^*(x),\xi^*))$ (resp. $(0,(u^*(1-x),\xi^*))$) and, if $\mathfrak{S}_1$ is locally one-parametrizable, $\mathcal{C}^1_0$ is also locally one-parametrizable.

Remark 4.1. Using the transformation T^n in Lemma 2.1 and replacing U_0, D_0 by u,d_1, we can construct any other solution branch with n modes of (SS) (which we denote by $\mathcal{C}^n_0$). Roughly speaking, $\mathcal{C}^n_0$ is similar to $\mathcal{C}^1_0$ with ratio $1:1/n^2$ in $\mathfrak{E}$.

We use the notation $\mathcal{C}_\alpha^n$ and $d_{c,\alpha}^n$ instead of $\mathcal{C}_{d_2}^n$ and d_{c,d_2}^n and denote the value $b_{11}/(n\pi)^2=(\lim_{d_2\uparrow\infty} d_{c,d_2}^n)$ by $d_{c,0}^n$. In what follows we only discuss $\mathcal{C}_\alpha^1$. However, the corresponding results for $\mathcal{C}_\alpha^n (n\geq 2)$ can be obtained in the same way.

Now we will show in the next theorem that $\mathcal{C}_0^1$ is a good approximation of $\mathcal{C}_\alpha^1$ for small α. We denote by $\mathcal{C}_{\alpha,\delta}^1$ the restriction of $\mathcal{C}_\alpha^1$ to $[\delta, +\infty)\times\mathbb{H}_N^2$.

THEOREM 4.1. *For any positive δ_1 and ε, there exists a positive constant α_1 such that*

$$\mathcal{C}_{\alpha,\delta_1}^1 \subset U_\varepsilon^{\delta_1} \quad \text{for } 0\leq\alpha<\alpha_1,$$

where $U_\varepsilon^{\delta_1}$ is the restriction of the ε-neighborhood of $\mathcal{C}_0^1$ in $\mathbb{R}^+\times\mathbb{H}_N^2$ to $[\delta_1, +\infty)\times\mathbb{H}_N^2$.

Proof. First, it follows easily from (1.5) and Lemma 3.1, property 2) that the bifurcation point $(d_{c,\alpha}^1, \bar{U})$ converges to $(d_{c,0}^1, \bar{U})$ as $\alpha\downarrow 0$. The boundedness of $\mathcal{C}_{0,\delta_1}^1$ and compactness of $\mathcal{G}$ imply that $\mathcal{C}_{0,\delta_1}^1$ is compact in $[\delta_1, +\infty)\times\mathbb{H}_N^2$. On the other hand, it follows from Remark 4.1 that only finitely many $\mathcal{C}_{0,\delta_1}^n$ $(n=2,\cdots,l)$ are not empty for fixed δ_1 and $\mathcal{C}_{0,\delta_1}^1\cap\{\bigcup_{n=2}^l \mathcal{C}_{0,\delta_1}^n\}=\varnothing$. Since $\mathcal{C}_{0,\delta_1}^1$ and $\bigcup_{n=2}^l \mathcal{C}_{0,\delta_1}^n$ are compact in $\mathcal{G}$, the distance between these two sets is positive. Therefore, for all sufficiently small positive ε, we can take the 3ε-neighborhood $U_{3\varepsilon}^{\delta_1}$ of $\mathcal{C}_0^1$ in $[\delta_1, +\infty)\times\mathbb{H}_N^2$ such that $U_{3\varepsilon}^{\delta_1}\setminus\mathcal{C}_{0,\delta_1}^1$ does not contain any other solutions of (SS) except the trivial solutions $(d_1, \bar{U})$ with $d_{c,0}^1-\gamma(\varepsilon)<d_1<d_{c,0}^1+\gamma(\varepsilon)$ and $d_1\neq d_{c,0}^1$, where $\lim_{\varepsilon\downarrow 0}\gamma(\varepsilon)=0$. There may be a solution set M_{δ_1} of (SS) in $[\delta_1, +\infty)\times\mathbb{H}_N^2$ which does not belong to $\bigcup_{n=1}^l \mathcal{C}_{0,\delta_1}^n$. However, since M_{δ_1} is closed, the distance between the compact set $\mathcal{C}_{0,\delta_1}^1$ and M_{δ_1} is positive. Note that none of the bifurcation points $(d_{c,0}^n, \bar{U})$ $(n\in N)$ belong to the set $U_{3\varepsilon}^{\delta_1}\setminus\mathcal{C}_{0,\delta_1}^1$ for small ε.

Now let us assume the contrary: There exist some positive constants δ_1 and ε such that, for any small α, $\mathcal{C}_{\alpha,\delta_1}^1$ is not contained in $U_\varepsilon^{\delta_1}$. Then we can extract the sequence $(d_{\alpha_n}^*, U_n^*)$ $(n\in N)$ of solutions of (SP) with $\alpha=\alpha_n$ such that $\lim_{n\uparrow\infty}\alpha_n=0$, $(d_{\alpha_n}^*, U_n^*)\notin U_\varepsilon^{\delta_1}$ but $(d_{\alpha_n}^*, U_n^*)\in U_{2\varepsilon}^{\delta_1}$ (see the remark at the beginning of this proof). Since the set $\{(d_{\alpha_n}^*, U_n^*)\}_{n=1}^\infty$ is bounded in $[\delta_1, +\infty)\times\mathbb{H}_N^2$ and each element is a solution of (SP) with $d_1\geq\delta_1$, we can extract the strong convergent subsequence $\{(d_{\alpha_l}^*, U_l^*)\}_{l=1}^\infty$ in $[\delta_1, +\infty)\times\mathbb{H}_N^2$. We set $(d^*, U^*)=\lim_{l\uparrow\infty}(d_{\alpha_l}^*, U_l^*)$. The strong limit (d^*, U^*) is a solution of (SS) and belongs to $U_{3\varepsilon}^{\delta_1}\setminus\mathcal{C}_{0,\delta_1}^1$. Moreover, due to the following lemma, we can see that $U^*\neq\bar{U}$.

LEMMA 4.2. *For any fixed small $\varepsilon>0$, there exists a positive constant $\alpha=\alpha(\varepsilon)$ such that the operator norm of*

$$\left(D\frac{d^2}{dx^2}+B\right)^{-1}$$

from $(L^2(I))^2$ to $\mathbb{H}_N^2$ with $\varepsilon\leq|d-d_{c,0}^1|\leq 2\varepsilon$ is uniformly bounded for $d_2>\alpha(\varepsilon)^{-1}$.

Proof. By using the proof of [16, Lemma 5.1], we can verify this lemma in a straightforward way, so we omit the details.

Thus we have a contradiction and Theorem 4.1 is established.

5. Singularly perturbed solutions. We know from Theorem 4.1 that when α is small the branch $\mathcal{C}_\alpha^1$ is approximated well by the shadow branch $\mathcal{C}_0^1$ for $d_1\geq\delta_1>0$. Then how does $\mathcal{C}_\alpha^1$ behave as $d_1\downarrow 0$? To answer this question we need more precise analysis of (SP) with d_1 sufficiently small, usually called singular perturbation analysis. Such study for a coupled nonlinear diffusion system was done by Fife [7] for Dirichlet boundary conditions, and recently Mimura, Tabata and Hosono [19] used Fife's method to resolve the problem for Neumann boundary conditions. In this section, we summarize the results of [19] and derive some lemmas from them to be used later.

In order to construct the singularly perturbed solutions, we make the following assumption in addition to (A-0)–(A-4):

(A-5) $G_i(v)\in C^1(\mathring{I}_i)$ and $dG_i(v)/dv<0$ in $\mathring{I}_i$ $(i=1,2)$, where $G_i(v)=g(h_i(v), v)$ $(i=1,2$ and each $h_i(v)$ is defined as in (A-1)).

Note that Examples 1–3 in §0 satisfy (A-5).

Now we rewrite (SP) as

$$(\mathrm{SP})_\varepsilon \qquad 0=\varepsilon^2 u_{xx}+f(u,v), \qquad 0=v_{xx}+\alpha g(u,v),$$

where $\varepsilon=\sqrt{d_1}$. First we solve the reduced problem of $(\mathrm{SP})_\varepsilon$:

$$(\mathrm{SP})_0 \qquad 0=f(u,v), \qquad 0=v_{xx}+\alpha g(u,v).$$

It follows from Remark 0.1 that $f(u,v)=0$ has three solution branches $u=h_i(v)$ $(i=0,1,2)$. Therefore there are many ways of taking the solution $u=h(v)$ of the first equation of $(\mathrm{SP})_0$ which may be discontinuous. Therefore we seek a solution pair of $(\mathrm{SP})_0$ in a weak sense; i.e., we call $(U(x), V(x))$ a solution of the problem $(\mathrm{SP})_0$ if (U, V) satisfies

$$\begin{aligned}&(U,V)\in L^2(I)\times H^1(I),\\ &f(U,V)=0 \quad \text{almost everywhere in } I,\\ &\langle V_x,\phi_x\rangle=\langle \alpha g(U,V),\phi\rangle \quad \text{for all } \phi\in H^1(I),\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ is the inner product on $L^2(I)$.

Here we take the following special solution $u=h^*(v)$ of $f(u,v)=0$:

$$u=h^*(v)=\begin{cases} h_1(v) & \text{for } v\in\{v<\xi^*\}\cap\mathring{I}_1,\\ h_2(v) & \text{for } v\in\{v>\xi^*\}\cap\mathring{I}_2,\end{cases} \tag{5.1}$$

where ξ^* is the value of v which appears in (A-2). We define $G^{\xi^*}(v)$ corresponding to (5.1) by

$$G^{\xi^*}(v)=\begin{cases} G_1(v) & \text{for } v\in\{v<\xi^*\}\cap\mathring{I}_1,\\ G_2(v) & \text{for } v\in\{v>\xi^*\}\cap\mathring{I}_2.\end{cases}$$

We can see from Remark 0.2 that

$$G_1(v)<0 \quad \text{on } I_1, \qquad G_2(v)>0 \quad \text{on } I_2. \tag{5.2}$$

It is found that $G^{\xi^*}(v)$ has a discontinuity of the first kind at $v=\xi^*$. Thus the problem $(\mathrm{SP})_0$ is reduced to the two-point boundary value problem for a single equation:

$$(5.3)_a \qquad V_{xx}+\alpha G^{\xi^*}(V)=0, \qquad x\in I,$$
$$(5.3)_b \qquad V_x(0)=V_x(1)=0.$$

For this problem, a solution V is defined by

$$(5.4)_a \qquad V\in H^1(I),$$
$$(5.4)_b \qquad \langle V_x,\phi_x\rangle=\langle \alpha G^{\xi^*}(V),\phi\rangle \quad \text{for all } \phi\in H^1(I).$$

We have

LEMMA 5.1 ([19]). *Under assumptions* (A-0)–(A-5), *there exists a positive constant* α_1 *such that* (5.4) *has a unique strictly increasing solution* $V_{1,0}^{\xi^*,\alpha}(x) \in C^1(I)$ *for* $0<\alpha<\alpha_1$.

Remark 5.1. Since (5.4) is reflection-invariant, $V_{1,1}^{\xi^*,\alpha}(x)=V_{1,0}^{\xi^*,\alpha}(1-x)$ is also a solution of (5.4), which is strictly decreasing.

Remark 5.2. It follows from (5.2) and the boundary conditions that $V_{1,0}^{\xi^*,\alpha}(x)$ (resp. $V_{1,1}^{\xi^*,\alpha}(x)$) crosses the line $V=\xi^*$ at a unique point x_1^0 (resp. $x_1^1=1-x_1^0$) with $0<x_1^0<1$.

Defining $U_{1,0}^{\xi^*,\alpha}(x)$ by

$$U_{1,0}^{\xi^*,\alpha}(x)=h^*\big(V_{1,0}^{\xi^*,\alpha}(x)\big), \tag{5.5}$$

we obtain a pair of solutions $(U_{1,0}^{\xi^*,\alpha}(x),\ V_{1,0}^{\xi^*,\alpha}(x))$ of $(\mathrm{SP})_0$. Then the following theorem holds.

THEOREM 5.2 [19]. *Suppose that* (A-0)–(A-5) *hold. Let* $(U_{1,0}^{\xi^*,\alpha}(x),\ V_{1,0}^{\xi^*,\alpha}(x))$ *be a solution with one mode of the reduced problem* $(\mathrm{SP})_0$. *Then there exist some positive constants* ε_0 *and* α_0 *such that, for each fixed* $\alpha\in(0,\alpha_0)$, *an* ε*-family of solutions* $U_\alpha(x;\varepsilon)=(u_\alpha(x;\varepsilon),\ v_\alpha(x;\varepsilon))\in(C^2(\bar I))^2$ *of the problem* $(\mathrm{SP})_\varepsilon$ *exists, for* $0<\varepsilon<\varepsilon_0$, *which satisfies*

$$\lim_{\varepsilon\downarrow 0} u_\alpha(x;\varepsilon)=U_{1,0}^{\xi^*,\alpha}(x) \quad \textit{uniformly in } x\in\bar I-I_\kappa,$$

$$\lim_{\varepsilon\downarrow 0} v_\alpha(x;\varepsilon)=V_{1,0}^{\xi^*,\alpha}(x) \quad \textit{uniformly in } x\in\bar I,$$

for any $\kappa>0$, *where* $I_\kappa=(x_1^0-\kappa,x_1^0+\kappa)$ *and* x_1^0 *is the point defined in Remark* 5.2. *Moreover* $(u_\alpha(x;\varepsilon),\ v_\alpha(x;\varepsilon))$ *is a continuously differentiable mapping from* $(0,\varepsilon_0)\times(0,\alpha_0)$ *to* $(C^2(\bar I))^2$.

Remark 5.3. The image by reflection $(u_\alpha(1-x;\varepsilon),\ v_\alpha(1-x;\varepsilon))$ is another family of one-mode solutions of $(\mathrm{SP})_\varepsilon$, which converges to $(U_{1,1}^{\xi^*,\alpha}(x),\ V_{1,1}^{\xi^*,\alpha}(x))$ as $\varepsilon\downarrow 0$. However it suffices to consider the family of solutions $(u_\alpha(x;\varepsilon),\ v_\alpha(x;\varepsilon))$ in the following discussions.

When we consider $U_\alpha(x;\varepsilon)$ to be an α-family of solutions for fixed ε, we can obtain the following result.

LEMMA 5.3. *For any fixed small* ε, $U_\alpha(x;\varepsilon)=(u_\alpha(x;\varepsilon),\ v_\alpha(x;\varepsilon))$ *converges to the unique solution* $\mathrm{Sec}_{\varepsilon^2}(\mathcal{C}^1_{0,+})$ *in the* $(C^2(\bar I))^2$*-topology as* $\alpha\downarrow 0$.

Proof. It suffices to show that any sequence $U_{\alpha_n}(x;\varepsilon)(\alpha_n\downarrow 0)$ of $U_\alpha(x;\varepsilon)$ has a convergent subsequence which converges to the unique limit $\mathrm{Sec}_{\varepsilon^2}(\mathcal{C}^1_{0,+})$ for small ε (see Theorem 3.7, or Theorem 3.7′ in §6). It follows from our construction of singularly perturbed solutions that $U_{\alpha_n}(x;\varepsilon)$ is uniformly C^0-bounded. Therefore $U_{\alpha_n}(x;\varepsilon)$ is uniformly bounded in C^2-norm because it satisfies $(\mathrm{SP})_\varepsilon$. Applying Ascoli's theorem to this, we can extract a C^1-strong convergent subsequence $U_{\alpha_{n(k)}}(x;\varepsilon)$. By using the equations $(\mathrm{SP})_\varepsilon$ again, we can see that this subsequence is a C^2-strong convergent sequence whose limit is clearly a solution of the shadow system. Since the strictly monotone increasing solution of the shadow system is unique for small ε, its limit solution must coincide with $\mathrm{Sec}_{\varepsilon^2}(\mathcal{C}^1_{0,+})$.

In order to prove the next lemma, we need the following technical assumption:

(A-6) There exists a positive constant δ such that $f_v(u,\xi)<0$ for $u\in(h_1(\xi),h_2(\xi))$ and $|\xi-\xi^*|<\delta$.

Remark 5.4. All examples in §0 satisfy (A-6).

LEMMA 5.4. *For any fixed* ε *with* $0<\varepsilon<\varepsilon_0$ (*if necessary, we take* ε_0 *in Theorem* 5.2 *smaller*), *there exists a positive constant* $\hat{\alpha}=\hat{\alpha}(\varepsilon)$ $(\leq\alpha_0)$ *such that the inverse operator* $(\mathfrak{L}_{\varepsilon,\alpha})^{-1}$ *of the following Fréchet derivative of* $(\mathrm{SP})_\varepsilon$ *at* $(u_\alpha(x;\varepsilon), v_\alpha(x;\varepsilon))$:

$$\mathfrak{L}_{\varepsilon,\alpha}\begin{pmatrix} w \\ z \end{pmatrix} = \begin{pmatrix} \varepsilon^2\frac{d^2}{dx^2}+f_u(u_\alpha,v_\alpha) & f_v(u_\alpha,v_\alpha) \\ g_u(u_\alpha,v_\alpha) & \frac{1}{\alpha}\frac{d^2}{dx^2}+g_v(u_\alpha,v_\alpha) \end{pmatrix}\begin{pmatrix} w \\ z \end{pmatrix}, \tag{5.6}$$

$$ {}^t(w,z)\in\left(C^2(\bar{I})\right)^2,$$

exists and is uniformly bounded from $(C^0(\bar{I}))^2$ *to* $(C^2(\bar{I}))^2$ *on* $(\varepsilon,\varepsilon_0)\times(0,\hat{\alpha}(\varepsilon))$.

Proof. See Appendix 2.

LEMMA 5.5. *For any* $\varepsilon\in(0,\varepsilon_0)$, *there exists a positive constant* $\gamma_0=\gamma_0(\varepsilon)$ *such that* $(\mathrm{SP})_\varepsilon$ *has no solutions in the* γ_0*-neighborhood of* $(u_\alpha(x;\varepsilon), v_\alpha(x;\varepsilon))$ *in* $(C^2(\bar{I}))^2$ *other than itself for* $0<\alpha<\hat{\alpha}(\varepsilon)$.

Proof. Applying the transformation $w_1=u-u_\alpha(x;\varepsilon)$, $w_2=v-v_\alpha(x;\varepsilon)$ and using Lemma 5.4, we can easily prove the above lemma by contradiction. Therefore we leave the details to the reader.

It follows from (5.4) and Remark 5.2 that

$$\lim_{\alpha\downarrow 0}\lim_{\varepsilon\downarrow 0} v_\alpha(x;\varepsilon)=\lim_{\alpha\downarrow 0} V_{1,0}^{\xi^*,\alpha}(x)=\xi^*.$$

Therefore, for any $\kappa>0$,

$$\lim_{\alpha\downarrow 0}\lim_{\varepsilon\downarrow 0} u_\alpha(x;\varepsilon)=\lim_{\alpha\downarrow 0} U_{1,0}^{\xi^*,\alpha}(x)=u^*(x) \qquad \text{(see (3.23))}$$

uniformly in $x\in\bar{I}\setminus[x^*-\kappa, x^*+\kappa]$, because

$$\int_0^1 g\left(U_{1,0}^{\xi^*,\alpha}(x), V_{1,0}^{\xi^*,\alpha}(x)\right)dx=0$$

holds for any α. Thus combining the above results with Theorem 3.3 and Lemma 5.3, we obtain the following commutative relation.

Remark 5.5.

$$\lim_{d_1\downarrow 0}\lim_{\alpha\downarrow 0}\left(u_\alpha\left(x;\sqrt{d_1}\right), v_\alpha\left(x;\sqrt{d_1}\right)\right)$$
$$=\lim_{\alpha\downarrow 0}\lim_{d_1\downarrow 0}\left(u_\alpha\left(x;\sqrt{d_1}\right), v_\alpha\left(x;\sqrt{d_1}\right)\right)=(u^*(x),\xi^*).$$

6. From bifurcation to singular perturbation. We have already seen in §4 that the structure of $\mathcal{C}_\alpha^1$ is approximated well by that of $\mathcal{C}_0^1$ for $d_1\geq\delta_1$. On the other hand, we have shown in §5 that (SP) has singularly perturbed solutions for sufficiently small d_1. In this section we will show that $\mathcal{C}_\alpha^1$ is connected to singularly perturbed solutions when α is small; i.e., the bifurcating branch from $\bar{U}$ continues to exist in $\mathfrak{S}$ until it arrives at the solution $(u_\alpha(x;\varepsilon), v_\alpha(x;\varepsilon))$ of Theorem 5.2.

First we rewrite Theorem 3.7 in the following form:

THEOREM 3.7′. *There exists a positive number* $\bar{d}_1$ *such that* $\mathrm{Sec}_{d_1}(\mathcal{C}_{0,+}^1)$ (*resp.* $\mathrm{Sec}_{d_1}(\mathcal{C}_{0,-}^1)$) *consists of a unique element for* $0<d_1<\bar{d}_1$.

Now we can prove the main theorem (see Fig. 11).

THEOREM 6.1. *Suppose that* (A-0)–(A-6) *hold. Then, there exists a positive constant* α^* *such that, for* $0<\alpha<\alpha^*$, $\mathcal{C}^1_\alpha$ *exists globally with respect to* d_1 *and coincides with the singularly perturbed solutions* $(u_\alpha(x;\sqrt{d_1}), v_\alpha(x;\sqrt{d_1}))$ *and* $(u_\alpha(1-x;\sqrt{d_1}), v_\alpha(1-x;\sqrt{d_1}))$ *of Theorem* 5.2 *for sufficiently small* d_1.

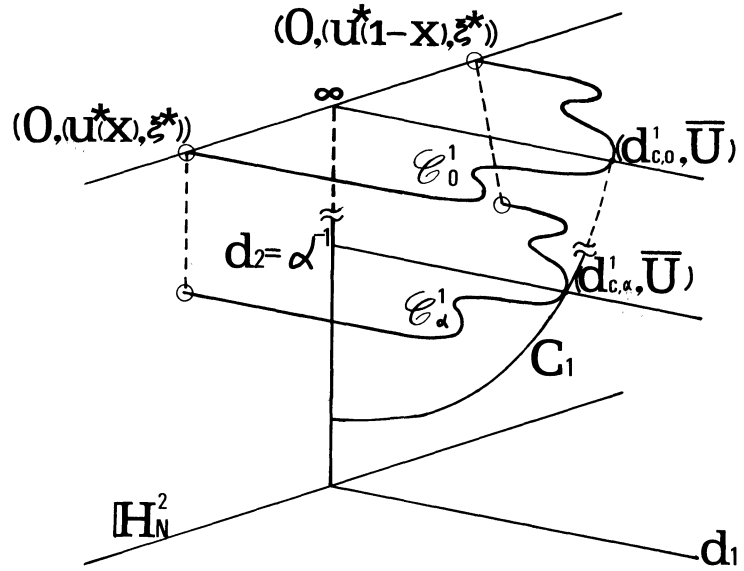

FIG. 11. *From bifurcation to singular perturbation. The branches diverge in* $\mathbb{H}^2_N$*-norm as* $d_1 \downarrow 0$ *(though remain finite in* L^∞*-norm); however, we draw them in a finite region.*

Proof. It is easy to verify that, for small α, $\mathcal{C}^1_\alpha$ is divided into two parts:

$$\mathcal{C}^1_\alpha = \mathcal{C}^1_{\alpha,+} \cup \mathcal{C}^1_{\alpha,-}, \qquad \mathcal{C}^1_{\alpha,+} \cap \mathcal{C}^1_{\alpha,-} \ni \{(d^1_{c,\alpha}, \bar{U})\},$$

where $\mathcal{C}^1_{\alpha,+}$ (resp. $\mathcal{C}^1_{\alpha,-}$) tends to $\mathcal{C}^1_{0,+}$ (resp. $\mathcal{C}^1_{0,-}$) as $\alpha \downarrow 0$ in the sense of Theorem 4.1. Moreover $\mathcal{C}^1_{\alpha,-}$ is the image of $\mathcal{C}^1_{\alpha,+}$ by reflection at $x=\frac{1}{2}$, i.e., any solution of $\mathcal{C}^1_{\alpha,-}$ can be obtained in the form $(d_1, u(1-x), v(1-x))$, where $(d_1, u(x), v(x))$ is a solution of $\mathcal{C}^1_{\alpha,+}$. Therefore it suffices to prove the theorem for the part $\mathcal{C}^1_{\alpha,+}$ of $\mathcal{C}^1_\alpha$.

First we take δ_1 in Theorem 4.1 as

$$\delta_1 < \min\{\varepsilon_0^2, \bar{d}_1\},$$

where ε_0 and $\bar{d}_1$ are the constants which appeared in Theorem 5.2 and Theorem 3.7′ respectively. Then, for some fixed d_1 with $\delta_1 \le d_1 < \min\{\varepsilon_0^2, \bar{d}_1\}$, we can take $\alpha^*(<\hat{\alpha}(\sqrt{d_1}))$ such that the distance between $\mathrm{Sec}_{d_1}(\mathcal{C}^1_{\alpha,+})$ and the singularly perturbed solutions $(u_\alpha(x;\sqrt{d_1}), v_\alpha(x;\sqrt{d_1}))$ in $(C^2(\bar{I}))^2$ is smaller than $\gamma_0 = \gamma_0(\sqrt{d_1})$ which appeared in Lemma 5.5 for $0<\alpha<\alpha^*$. This is possible because both solutions converge to the same solution $\mathrm{Sec}_{d_1}(\mathcal{C}^1_{0,+})$ as $\alpha \downarrow 0$ (see Theorem 4.1 and Lemma 5.3). The local uniqueness property of Lemma 5.5 implies that $\mathrm{Sec}_{d_1}(\mathcal{C}^1_{\alpha,+})$ must coincide with $(u_\alpha(x;\sqrt{d_1}), v_\alpha(x;\sqrt{d_1}))$. This completes the proof.

Remark 6.1. We can see from Theorem 6.1 that the asymptotic behavior of $\mathcal{C}^1_\alpha$ as $d_1 \downarrow 0$ is described by Theorem 5.2.

Remark 6.2. Using the transformation T^n of Lemma 2.1 and Theorem 6.1, we can obtain the corresponding theorem for $\mathcal{C}_\alpha^n$ for $n \geq 2$ (α^* is replaced by α^*/n^2).

7. Concluding remarks. In this paper, we are mainly concerned about the global existence of the bifurcating branch and its asymptotic behavior as $d_1 \downarrow 0$. Here we comment on several points of our results.

(i) *Stability.* As for stability , we know that the bifurcation direction determines stability near the simple critical bifurcation point (see, e.g., [5]) which corresponds to the point $(d_{c,\alpha}^n, \bar{U})$ of $\mathcal{C}_\alpha^n$. Then what about stability when the branch leaves the critical point? We especially want to know the *stability of singularly perturbed solutions.* This is a difficult problem; however, it seems to the author that Theorem 6.1 gives an insight for this problem. Let us consider the special case where the linearized operator of (SP) along $\mathcal{C}_\alpha^1$ is invertible except at $(d_{c,\alpha}^1, \bar{U})$, and never has pure imaginary eigenvalues. In this case Theorem 6.1 implies from the local stability result near $(d_{c,\alpha}^1, \bar{U})$ that singularly perturbed solutions $(u_\alpha(x;\varepsilon), v_\alpha(x;\varepsilon))$ are stable as stationary solutions of the evolution system.

$$\text{(P)} \qquad \begin{aligned} &\frac{\partial u}{\partial t} = d_1 u_{xx} + f(u,v), \\ &\qquad\qquad\qquad\qquad 0 < \alpha \ll 1. \\ &\frac{\partial v}{\partial t} = \frac{1}{\alpha} v_{xx} + g(u,v), \end{aligned}$$

The detailed spectral analysis for (SS) is a key to solving the stability problem for (P).

(ii) *No saturation case.* If the assumption (A-1) is not satisfied, the situation is completely different from ours. For example, in the case where the nonlinearity f is not folding over, such as in the Gierer–Meinhardt model with no saturation ($\kappa = 0$ in Example 2, see Fig. 12), the solution branch of the corresponding shadow system does not remain bounded in the L^∞-sense and the singularly perturbed solutions cannot be constructed by our method. Therefore the assumption that zero level curve of f is sigmoidal is indispensable for our study.

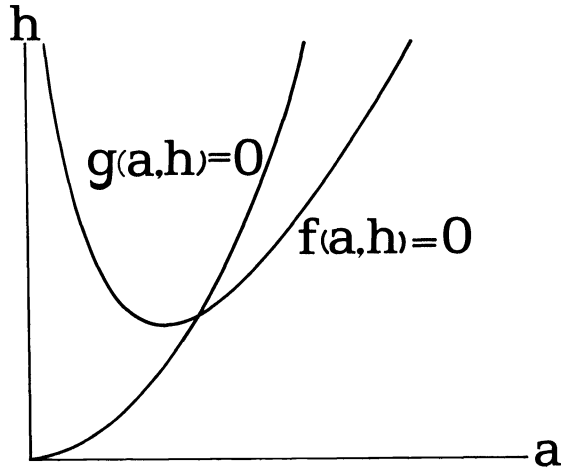

FIG. 12. *Functional forms of Gierer-Meinhardt model with no saturation.*

(iii) *The case where d_2 is not large.* In this case, the structure of the set of bifurcating solutions of (SP) is more complicated than in Theorem 6.1. One reason for such complexity is the appearance of double critical points (i.e., intersection of two curves of $\{C_n\}_{n=1}^{\infty}$ in §1), and the other is the occurance of secondary and tertiary bifurcations. Fujii, Mimura and Nishiura [8] have recently clarified some aspects of the mechanism of the global structure with the aid of a powerful numerical method (for numerical aspects, see also [9]). However, we have not yet obtained a total understanding of a global picture for (SP).

Appendix 1. The relation between the position of a transition layer and the asymptotic form of $\xi=\xi(E)$ as $E\uparrow E^{\infty}(\xi^*)$. We know from Theorem 3.3 that $\mathcal{S}_1$ goes into the point $(E^{\infty}(\xi^*),\xi^*)$ as $E\uparrow E^{\infty}(\xi^*)$. Let $\xi=\xi(E)$ be an arbitrary smooth path in T, which satisfies

$$\lim_{E\uparrow E^{\infty}(\xi^*)}\xi(E)=\xi^*.$$

Then $u(x;E,\xi(E))$ tends to the step function which has a discontinuity at $x=x_c$ as $E\uparrow E^{\infty}(\xi^*)$:

$$\lim_{E\uparrow E^{\infty}(\xi^*)}u(x;E,\xi(E))=\begin{cases}u_l(\xi^*), & 0\le x<x_c,\\ u_r(\xi^*), & x_c<x\le 1.\end{cases}$$

We study in the following how the position of x_c depends on the asymptotic form of $\xi(E)$ as $E\uparrow E^{\infty}(\xi^*)$.

First it follows from the asymptotic formulas (3.14) and (3.17) that x_c is determined by the following relation:

$$\text{(A.1.1)}\qquad \lim_{E\uparrow E^{\infty}(\xi^*)}\frac{\alpha_l^{-1/2}(\xi(E))\log(E_l^{\infty}(\xi(E))-E)}{\alpha_r^{-1/2}(\xi(E))\log(E_r^{\infty}(\xi(E))-E)}=\frac{x_c}{1-x_c}.$$

Expanding $E_l^{\infty}(\xi)$ and $E_r^{\infty}(\xi)$ at $\xi=\xi^*$, we have

$$\text{(A.1.2)}\quad E_l^{\infty}(\xi)=E^{\infty}+k_1^l(\xi-\xi^*)+\frac{k_2^l}{2}(\xi-\xi^*)^2+\cdots+\frac{k_n^l}{n!}(\xi-\xi^*)^n+o(|\xi-\xi^*|^n)$$

and

$$\text{(A.1.3)}\quad E_r^{\infty}(\xi)=E^{\infty}+k_1^r(\xi-\xi^*)+\frac{k_2^r}{2}(\xi-\xi^*)^2+\cdots+\frac{k_n^r}{n!}(\xi-\xi^*)^n+o(|\xi-\xi^*|^n),$$

where $E^{\infty}=E^{\infty}(\xi^*)$, $k_j^l=d^jE_l^{\infty}(\xi^*)/d\xi^j$ and $k_j^r=d^jE_r^{\infty}(\xi^*)/d\xi^j$ $(j=1,2,\ldots n)$. It follows from (A-2) that $k_1^l>0$ and $k_1^r<0$.

Suppose that $\xi(E)$ has an asymptotic form:

$$\text{(A.1.4)}\qquad \xi(E)=\xi^*+\xi_1(E-E^{\infty})+\xi_{1+\alpha}(E-E^{\infty})^{1+\alpha}(1+o(1)),$$
$$(\xi_{1+\alpha}\neq 0)\quad \text{as } E\uparrow E^{\infty},\qquad 0<\alpha\le 1.$$

Substituting (A.1.4) into (A.1.2) and (A.1.3), we obtain

$$\text{(A.1.5)}\quad E_l^{\infty}(\xi(E))-E=(k_1^l\xi_1-1)(E-E^{\infty})+C_{1+\alpha}^l(E-E^{\infty})^{1+\alpha}(1+o(1)),$$

$$\text{(A.1.6)}\quad E_r^{\infty}(\xi(E))-E=(k_1^r\xi_1-1)(E-E^{\infty})+C_{1+\alpha}^r(E-E^{\infty})^{1+\alpha}(1+o(1)),$$

where

$$C^l_{1+\alpha}\,(\text{resp. } C^r_{1+\alpha})=\begin{cases} k^l_1\xi_{1+\alpha}\,(\text{resp. } k^r_1\xi_{1+\alpha}), & 0<\alpha<1,\\ k^l_1\xi_2+k^l_2\xi_1^2/2\,(\text{resp. } k^r_1\xi_2+k^r_1\xi_1^2/2), & \alpha=1.\end{cases}$$

We note that $C^l_{1+\alpha}\neq 0$ for $0<\alpha<1$. Substituting (A.1.5) and (A.1.6) into (A.1.4), we can see that if $k^l_1\xi_1\neq 1$ and $k^r_1\xi_1\neq 1$, i.e., $\xi=\xi(E)$ is not tangent to the curves $E=E^\infty_l(\xi)$ and $E=E^\infty_r(\xi)$ at $\xi=\xi^*$, the limit of the left-hand side of (1) is equal to $\alpha_r^{1/2}/\alpha_l^{1/2}$; i.e., $u(x;E,\xi(E))$ tends to the step function $u_{\xi^*}(x)$ of Lemma 3.1 as $E\uparrow E^\infty(\xi^*)$. If the path $\xi=\xi(E)$ is tangent to the curve, for example, $E=E^\infty_l(\xi)$ of first order at $\xi=\xi^*$ (i.e., $k^l_1\xi_1=1$ and $C^l_{1+\alpha}\neq 0$ for $0<\alpha\leq 1$), the right-hand side of (A.1.5) starts from the second term. Therefore the limit of the left-hand side of (A.1.1) equal to $(1+\alpha)\alpha_r^{1/2}/\alpha_l^{1/2}$.

In general, if $\xi(E)$ has an asymptotic form

$$\text{(A.1.7)}\quad \xi(E)=\xi^*+\sum_{j=1}^{n}\xi_j(E-E^\infty)^j+\xi_{n+\alpha}(E-E^\infty)^{n+\alpha}(1+o(1)),\qquad 0<\alpha\leq 1$$

and if $\xi=\xi(E)$ is tangent to the curve $E=E^\infty_l(\xi)$ of nth order at $\xi=\xi^*$, we can see that the limit of (1) tends to $(n+\alpha)\alpha_r^{1/2}/\alpha_l^{1/2}$. This means that the more two curves $\xi=\xi(E)$ and $E=E^\infty_l(\xi)$ have a close contact, the more the region of x where $\lim_{E\uparrow E^\infty(\xi^*)}u(x;E,\xi(E))=u_l(\xi^*)$, extends. Let us denote the limit of the left-hand side of (A.1.1) by k, then the point of discontinuity x_c can be written as

$$\text{(A.1.8)}\qquad x_c=P(k),$$

where $P(k)=k/1+k$. We note that $P(k)$ is an strictly monotone increasing function for $k\geq 0$ satisfying $P(0)=0$ and $\lim_{k\uparrow\infty}P(k)=1$.

Thus we obtain

PROPOSITION A. *Let $\xi=\xi(E)$ be a path in T which satisfies*

$$\lim_{E\uparrow E^\infty(\xi^*)}\xi(E)=\xi^*.$$

Then, if the path $\xi=\xi(E)$ is not tangent to both curves $E=E^\infty_l(\xi)$ and $E=E^\infty_r(\xi)$, we have

$$\lim_{E\uparrow E^\infty(\xi^*)}u(x;E,\xi(E))=u_{\xi^*}(x),$$

where $u_{\xi^*}(x)$ is the step function which appears in Lemma 3.1. If the path $\xi=\xi(E)$ is tangent to the curve $E=E^\infty_l(\xi)$ (resp. $E=E^\infty_r(\xi)$) of nth order (and not of $(n+1)$th order) at $\xi=\xi^*$, and $\xi(E)$ has an asymptotic form such as (A.1.7), then we have

$$\lim_{E\uparrow E^\infty(\xi^*)}u(x;E,\xi(E))=\begin{cases}u_l(\xi^*), & 0\leq x<x_c,\\ u_r(\xi^*), & x_c<x\leq 1,\end{cases}$$

where $x_c=P((n+\alpha)\alpha_r^{1/2}/\alpha_l^{1/2})$ (resp. $P(\alpha_r^{1/2}/\{(n+\alpha)\alpha_l^{1/2}\})$). Especially if $\xi=\xi(E)$ is tangent to the curve $E=E^\infty_l(\xi)$ (resp. $E^\infty_r(\xi)$) of infinite order at $\xi=\xi^*$, we have

$$\lim_{E\uparrow E^\infty(\xi^*)}u(x;E,\xi(E))=\begin{cases}u_l(\xi^*), & 0\leq x<1\quad(\text{resp. } x=0),\\ u_r(\xi^*), & x=1\quad(\text{resp. } 0<x\leq 1).\end{cases}$$

Appendix 2. Invertibility of the linearized operator for small d_1. In this appendix we prove Lemma 5.4 in §5 through a sequence of lemmas. First we show a lemma which is a direct consequence of Lemma A.2.2 which will appear later.

LEMMA A.2.1. *$L_{\varepsilon,0}$ is invertible for $0<\varepsilon\le\hat{\varepsilon}$ for some positive constant $\hat{\varepsilon}$, where*

$$L_{\varepsilon,0}=\varepsilon^2\frac{d^2}{dx^2}+f_u(u_0(x;\varepsilon),\xi(\varepsilon))$$

and

$$(u_0(x;\varepsilon),\xi(\varepsilon))=\lim_{\alpha\downarrow 0}(u_\alpha(x;\varepsilon),v_\alpha(x;\varepsilon)).$$

COROLLARY A.2.2. *For any ε_1 with $0<\varepsilon_1<\hat{\varepsilon}$, there exists a positive constant $\alpha_1=\alpha_1(\varepsilon_1)$ such that $L_{\varepsilon,\alpha}$ is uniformly invertible for $\varepsilon_1\le\varepsilon\le\hat{\varepsilon}$ and $0\le\alpha\le\alpha_1$, where*

$$L_{\varepsilon,\alpha}=\varepsilon^2\frac{d^2}{dx^2}+f_u(u_\alpha(x;\varepsilon),v_\alpha(x;\varepsilon)).$$

Let us consider the following inhomogeneous linearized system of equations:

$$\text{(LP)}\qquad \mathcal{L}_{\varepsilon,\alpha}\begin{pmatrix}w\\ z\end{pmatrix}=\begin{pmatrix}L_{\varepsilon,\alpha} & f_v\\ g_u & \dfrac{1}{\alpha}\dfrac{d^2}{dx^2}+g_v\end{pmatrix}\begin{pmatrix}w\\ z\end{pmatrix}=\begin{pmatrix}F_1\\ F_2\end{pmatrix},$$

$$w_x(0)=w_x(1)=z_x(0)=z_x(1)=0,$$

where $(F_1,F_2)\in(C^0(\bar{I}))^2$, and $(w,z)\in(C^2(\bar{I}))^2$ are unknown functions. If we can solve (LP) uniquely for any (F_1,F_2), it follows from Banach's theorem that the linearized operator $\mathcal{L}_{\varepsilon,\alpha}$ is invertible. In what follows we solve the problem (LP) for $\varepsilon_1\le\varepsilon\le\hat{\varepsilon}$ and $0\le\alpha\le\alpha_1$ as defined in Corollary A.2.2. Operating $(L_{\varepsilon,\alpha})^{-1}$ to the first equation of (LP), we obtain

$$w+(L_{\varepsilon,\alpha})^{-1}(f_v z)=(L_{\varepsilon,\alpha})^{-1}F_1.$$

Substituting this into the second equation of (LP), we have

$$\text{(A.2.1)}\qquad \frac{1}{\alpha}z_{xx}+g_u\{(L_{\varepsilon,\alpha})^{-1}F_1-(L_{\varepsilon,\alpha})^{-1}(f_v z)\}+g_v z=F_2.$$

Since the operator d^2/dx^2 is not invertible under the zero flux boundary conditions, we decompose z as follows:

$$z=\eta+z^*,$$

when η is a constant function and $\int_I z^*\,dx=0$. It is convenient to introduce the projection P from $C^0(\bar{I})$ onto the range of the operator d^2/dx^2, i.e.,

$$Pu=u-\int_I u\,dx\quad\text{for } u\in C^0(\bar{I}).$$

Using this projection, we obtain an equivalent system of equations to (A.2.1):

$$\text{(A.2.2)}_{\text{a}}\qquad z^*_{xx}+\alpha P\left[g_u\{(L_{\varepsilon,\alpha})^{-1}F_1-(L_{\varepsilon,\alpha})^{-1}(f_v z)\}+g_v z-F_2\right]=0,$$

$$\text{(A.2.2)}_{\text{b}}\qquad \int_I\left[g_u\{(L_{\varepsilon,\alpha})^{-1}F_1-(L_{\varepsilon,\alpha})^{-1}(f_v z)\}+g_v z-F_2\right]dx=0.$$

The equation $(A.2.2)_a$ is uniquely solvable with respect to z^*. We denote its solution by $z^*(\varepsilon, \alpha, \eta; F_1, , F_2)$. Substituting this into $(A.2.2)_b$, we obtain the following equation:

(A.2.3)
$$G(\varepsilon,\alpha,\eta;F_1,F_2)=\int_I\left[g_u\left\{(L_{\varepsilon,\alpha})^{-1}F_1-(L_{\varepsilon,\alpha})^{-1}f_v(\eta+z^*)\right\}+g_v(\eta+z^*)-F_2\right]dx=0.$$

It is easy to see that G and $\partial G/\partial\eta$ are continuous functions for $\varepsilon_1\le\varepsilon\le\hat{\varepsilon}$ and $0\le\alpha\le\alpha_1$. Noting that $z^*(\varepsilon,0,\eta;F_1,F_2)\equiv 0$, we can see that if the coefficient of η with $\alpha=0$ does not vanish for some ε', i.e.,

$$\int_I\left\{-g_u(L_{\varepsilon,0})^{-1}f_v+g_v\right\}dx\neq 0, \tag{A.2.4}$$

then the implicit function theorem tells us that there exist positive constants δ and α_2 $(\le\alpha_1)$ such that (A.2.3) has a unique solution $\eta=\eta(\varepsilon,\alpha;F_1,F_2)$ for $\varepsilon'-\delta\le\varepsilon\le\varepsilon'+\delta$ and $0\le\alpha\le\alpha_2$. This implies the uniform invertibility of $\mathcal{L}_{\varepsilon,\alpha}$ on $[\varepsilon'-\delta,\varepsilon'+\delta]\times[0,\alpha_2]$. We can see from the above discussion that the problem (LP) for $\alpha=0$ is given by

$$\mathcal{L}_{\varepsilon,0}\begin{pmatrix}w\\ \eta\end{pmatrix}=\begin{pmatrix}L_{\varepsilon,0}w+f_v\eta\\ \int_I\{g_uw+g_v\eta\}dx\end{pmatrix}=\begin{pmatrix}F_1\\ \int_I F_2\,dx\end{pmatrix},$$

where η is a constant function.

We will show in the following that (A.2.4) does not vanish for all small ε, which leads to the conclusion of Lemma 5.4. In order to do that, it suffices to prove the next two lemmas. Here we introduce the orthonormal eigenfunctions in $L^2(I)$ and corresponding eigenvalues $\{\phi_\varepsilon^n(x),\sigma_\varepsilon^n\}_{n=0}^\infty$ of $L_{\varepsilon,0}\phi=\sigma\phi$, with $\phi_x(0)=\phi_x(1)=0$.

LEMMA A.2.2.

i) $0<\sigma_\varepsilon^0<K\exp(-C_1/\varepsilon)$ *as* $\varepsilon\downarrow 0$.

ii) $\sigma_\varepsilon^n\le -C_2$ *as* $\varepsilon\downarrow 0$ *for* $n\ge 1$,

where K, C_1 *and* C_2 *are positive constants independent of* ε.

LEMMA A.2.3.

$$\int_I(L_{\varepsilon,0})^{-1}f_v\,dx\to-\infty\quad\text{as }\varepsilon\downarrow 0.$$

Proof of Lemma A.2.2. For simplicity, we take the interval I in the space direction to be $(-1,1)$, and the point of discontinuity x^* (see Theorem 3.3) to be zero. Applying the change of variables $y=x/\varepsilon$ to the problem $L_{\varepsilon,0}\phi=\sigma\phi$, we obtain an eigenvalue problem on $(-1/\varepsilon,1/\varepsilon)$,

$$\frac{d^2}{dy^2}\psi+f_u(u_0(\varepsilon y;\varepsilon),\xi(\varepsilon))\psi=\sigma\psi \tag{EP}_\varepsilon$$

subject to zero flux boundary conditions. Letting $\varepsilon\downarrow 0$, we have an eigenvalue problem on $(-\infty,\infty)$,

$$\frac{d^2}{dy^2}\psi+f_u(u^\infty(y),\xi^*)\psi=\sigma\psi, \tag{EP}_0$$

where $u^\infty(y)$ is a strictly monotone increasing function which is a uniform limit of $u_0(\varepsilon y;\varepsilon)$ (we extend u_0 as a constant continuously to the outside of $(-1/\varepsilon,1/\varepsilon)$ and identify these two functions) and satisfies

$$\lim_{y\downarrow-\infty}u^\infty(y)=u_l(\xi^*),\qquad\lim_{y\uparrow+\infty}u^\infty(y)=u_r(\xi^*). \tag{A.2.5}$$

We can see from (A.2.5) that $-f_u(u^\infty(y),\xi^*)$ is a potential well. We denote the orthonormal system and the corresponding eigenvalues of $(\mathrm{EP})_\varepsilon$ by $\{\psi_\varepsilon^n(y),\ \sigma_\varepsilon^n\}_{n=0}^\infty$. Note that eigenvalues remain the same by the transformation $y=x/\varepsilon$ and $\psi_\varepsilon^n(y)=\sqrt{\varepsilon}\,\phi_\varepsilon^n(\varepsilon y)$. Since our problem is autonomous and $\xi(\varepsilon)$ is a constant function, $(\mathrm{EP})_\varepsilon$ with boundary conditions of Dirichlet type $\psi(1/\varepsilon)=\psi(-1/\varepsilon)=0$ always has simple zero eigenvalue as a principal eigenvalue, and the corresponding eigenfunction is given by $(d/dx)u_0(\varepsilon y;\varepsilon)$. After normalization we denote this eigenfunction by $\psi_{\varepsilon,D}^0(y)$. Therefore we can see from the shape of the potential $-f_u(u^\infty(y),\xi^*)$ that $(\mathrm{EP})_0$ has a simple zero eigenvalue (as a principal eigenvalue) and the corresponding eigenfunction $\psi_0^0(y)$ does not change sign on $(-\infty,\infty)$, decaying in exponential order as $|y|\to+\infty$. Moreover, both principal eigenfunctions $\psi_\varepsilon^0(y)$ and $\psi_{\varepsilon,D}^0(y)$ converge uniformly to $\psi_0^0(y)$ as $\varepsilon\downarrow 0$.

Now, by using the well-known property of the Wronskian, we have

$$W\left(\psi_{\varepsilon,D}^0,\psi_\varepsilon^0\right)\Big|_{-1/\varepsilon}^{1/\varepsilon}=\sigma_\varepsilon^0\int_{-1/\varepsilon}^{1/\varepsilon}\psi_{\varepsilon,D}^0(y)\,\psi_\varepsilon^0(y)\,dy, \tag{A.2.6}$$

where $W(\cdot,\cdot)$ denotes the Wronskian. Since the potential $-f_u(u^\infty(y),\xi^*)$ is of well type, we can see by using the comparison method that

$$\psi_\varepsilon^0\left(\pm\frac{1}{\varepsilon}\right)\left(\text{resp. }\left|\frac{d}{dy}\psi_{\varepsilon,D}^0\left(\pm\frac{1}{\varepsilon}\right)\right|\right)\le C_3\exp\left(-\frac{C_4}{\varepsilon}\right)\quad\text{as }\varepsilon\downarrow 0, \tag{A.2.7}$$

where C_3 and C_4 are positive constants independent of ε. Substituting (A.2.7) into (A.2.6), we obtain part i) of Lemma A.2.2. The second inequality of Lemma A.2.2 is a direct consequence of the fact that zero is the simple principal eigenvalue of $(\mathrm{EP})_0$.

Proof of Lemma A.2.3. Let us expand f_v by the orthonormal set $\{\phi_\varepsilon^n\}$:

$$f_v=\sum_{n=0}^{\infty}f_\varepsilon^n\phi_\varepsilon^n.$$

Then we obtain

$$(L_{\varepsilon,0})^{-1}f_v=\sum_{n=0}^{\infty}\frac{f_\varepsilon^n}{\sigma_\varepsilon^n}\phi_\varepsilon^n. \tag{A.2.8}$$

Integrating (A.2.8) on I, we have

$$\int_I(L_{\varepsilon,0})^{-1}f_v\,dx=\frac{f_\varepsilon^0}{\sigma_\varepsilon^0}\int_I\phi_\varepsilon^0\,dx+\int_I\sum_{n\ge 1}\frac{f_\varepsilon^n}{\sigma_\varepsilon^n}\phi_\varepsilon^n\,dx. \tag{A.2.9}$$

By using the Schwarz inequality and the Parseval relation, the second term of (A.2.9) is estimated as follows:

$$\int_I\sum_{n\ge 1}\frac{f_\varepsilon^n}{\sigma_\varepsilon^n}\phi_\varepsilon^n\,dx\le\frac{1}{|\sigma_\varepsilon^1|}\|f_v\|_{L^2(I)}. \tag{A.2.10}$$

It follows from Lemma A.2.2 that the right-hand side of (A.2.10) is bounded independently with respect to ε.

To determine the asymptotic behavior of the first term of (A.2.9), it is convenient to introduce the following notation:

$$\theta(\xi)=\{u\,|\,u\in(h_1(\xi),h_2(\xi)),f_u(u,\xi)\ge 0\},$$
$$\mathrm{OS}(\varepsilon,\xi)=\{y\,|\,f_u(u_0(\varepsilon y;\varepsilon),\xi(\varepsilon))\ge 0\}.$$

$OS(\varepsilon,\xi)$ denotes the interval where the eigenfunction ψ_ε^n behave in oscillatory manner. Since $\theta(\xi)$ is a compact set contained in $(h_1(\xi),h_2(\xi))$ (see (A-1)), we can see from (A-6) that there exists a positive constant κ such that

$$f_v(u,\xi)\leq -\kappa \quad \text{for } u\in\theta(\xi), \quad |\xi-\xi^*|\leq\frac{\delta}{2}.$$

Therefore we have

$$\begin{aligned} f_\varepsilon^0=\int_I f_v(u_0(x;\varepsilon),\xi(\varepsilon))\phi_\varepsilon^0(x)\,dx &= \varepsilon^{1/2}\int_{-1/\varepsilon}^{1/\varepsilon} f_v(u_0(\varepsilon y;\varepsilon),\xi(\varepsilon))\psi_\varepsilon^0(y)\,dy \\ &< -\kappa\varepsilon^{1/2}\int_{OS(\varepsilon,\xi)}\psi_\varepsilon^0(y)\,dy \\ &< -C_5\kappa\varepsilon^{1/2}, \end{aligned}$$

where C_5 is a positive constant independent of ε. Finally we have

$$\int_I \phi_\varepsilon^0\,dx=\varepsilon^{1/2}\int_{-1/\varepsilon}^{1/\varepsilon}\psi_\varepsilon^0(y)\,dy\geq C_6\varepsilon^{1/2},$$

where C_6 is a positive constant independent of ε. Combining these results and Lemma A.2.2, we have

$$\frac{f_\varepsilon^0}{\sigma_\varepsilon^0}\int_I\phi_\varepsilon^0\,dx\leq -\frac{C_5C_6\varepsilon}{\sigma_\varepsilon^0}\leq -\frac{C_5C_6\varepsilon}{K\exp(-C_1/\varepsilon)}\to -\infty \quad \text{as } \varepsilon\downarrow 0.$$

Thus we obtain the conclusion of Lemma A.2.3.

Finally we have to check the following

$$\left|\int_I g_u(L_{\varepsilon,0})^{-1}f_v\,dx\right|\to +\infty \quad \text{as } \varepsilon\downarrow 0. \tag{A.2.11}$$

We can see from (A.2.9) and the continuity of the integral with respect to ε that it suffices to show the following.

$$\int_{-\infty}^{\infty} g_u(u^\infty(y),\xi^*)\psi_0^0(y)\,dy\neq 0. \tag{A.2.12}$$

However (A.2.12) holds for almost all nonlinear g. In fact, if the left-hand side of (A.2.12) vanishes, we can perturb g slightly so that (A.2.12) holds for the new g. Therefore, generically speaking, the condition (A.2.12) always holds. Equation (A.2.11) implies that (A.2.4) does not vanish for all small ε, which completes the proof of Lemma 5.4.

Remark A.2.1. In many applications (see Examples 1 and 2 in §0), $|g_u|$ is bounded away from zero, i.e., $|g_u(u^\infty(y),\xi^*)|\geq\delta>0$ for some positive constant δ. In this case it is clear that (A.2.12) holds since $\psi_0^0(y)$ has a definite sign.

Acknowledgments. The author would like to express his sincere appreciation to Professor M. Yamaguti of Kyoto University for his invaluable suggestions and continuous encouragement. He is also indebted to Professors M. Mimura of Konan University and H. Fujii and Y. Hosono of Kyoto Sangyo University for their many stimulating discussions and useful comments. Many thanks should be extended to Professor H. Kodama of Kyoto University for his constructive comments.

Note added in proof. As for the open problems stated in the concluding remarks, the author has recently succeeded in showing the stability of singularly perturbed

solutions and the asymptotic behavior of the global branch as $d_1 \downarrow 0$ for the no saturation case. For the details, see the following paper and the forthcoming ones: Y. Nishiura, *Global structure of bifurcating solutions of some reaction-diffusion systems and their stability problems*, Proc. Fifth International Symposium on Computing Methods in Applied Sciences and Engineering, North-Holland, Amsterdam, 1982.

REFERENCES

[1] J. F. G. AUCHMITY AND G. NICOLIS, *Bifurcation analysis of nonlinear reaction diffusion equations—I. Evolution equations and the steady state solutions*, Bull. Math. Biol. 37 (1975), pp. 323–365.

[2] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[3] E. CONWAY, D. HOFF AND J. A. SMOLLER, *Large time behavior of solutions of systems of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 1–16.

[4] M. G. CRANDALL AND P. H. RABINOWITZ, *Bifurcation from simple eigenvalues*, J. Funct. Anal. 8 (1971), pp. 321–340.

[5] _____, *Bifurcation, perturbation of simple eigenvalues and linearized stability*, Arch. Rational Mech. Anal., 43 (1973), pp. 161–180.

[6] J. DIEUDONNÉ, *Calcul infinitesimal*, Hermann, Paris, 1968, Chaps. 3, 4.

[7] P. C. FIFE, *Boundary and interior transition layer phenomena for pairs of second-order differential equations*, J. Math. Anal. Appl., 54 (1976), pp. 497–521.

[8] H. FUJII, M. MIMURA AND Y. NISHIURA, *A picture of global bifurcation diagram in ecological interacting and diffusing systems*, Physica D (1982), to appear.

[9] H. FUJII AND M. YAMAGUTI, *Structure of singularities and its numerical realization in nonlinear elasticity*, J. Math. Kyoto Univ., 20 (1980), pp. 489–590.

[10] A. GIERER AND H. MEINHARDT, *A theory of biological pattern formation*, Kybernetik, 12 (1972), pp. 30–39.

[11] M. GOLUBITSKY AND V. GUILLEMIN, *Stable mappings and their singularities*, Graduate Texts in Mathematics 14, Springer-Verlag, Berlin, 1973.

[12] V. GUILLEMIN AND A. POLLACK, *Differential topology*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[13] J. P. KEENER, *Secondary bifurcation in nonlinear diffusion reaction equations*, Stud. Appl. Math., 55 (1976), pp. 187–211.

[14] _____, *Activators and inhibitors in pattern formation*, Stud. Appl. Math., 59 (1978), pp. 1–23.

[15] M. MIMURA AND J. D. MURRAY, *Spatial structures in a model substrate-inhibition reaction diffusion system*, Z. Naturforsch, 33C (1978), pp. 580–586.

[16] M. MIMURA AND Y. NISHIURA, *Spatial patterns for interaction-diffusion equations in biology*, Proc. International Symposium on Mathematical Topics in Biology, Kyoto, Japan, 1978, pp. 136–146.

[17] M. MIMURA, Y. NISHIURA AND M. YAMAGUTI, *Some diffusive prey and predator systems and their bifurcation problems*, Ann. New York Acad. Sci., 316 (1979), pp. 490–521.

[18] M. MIMURA AND Y. NISHIURA, *Spatial patterns for an interaction-diffusion equation in morphogenesis*, J. Math. Biol. 7 (1979), pp. 243–263.

[19] M. MIMURA, M. TABATA AND Y. HOSONO, *Multiple solutions of two-point boundary value problems of Neumann type with a small parameter*, SIAM J. Math. Anal., 11 (1980), pp. 613–631.

[20] L. NIRENBERG, *Topics in Nonlinear Functional Analysis*, Lecture Notes, Courant Institute of Mathematics Sciences, New York, 1974.

[21] Y. NISHIURA, *Global branching theorem for spatial patterns of reaction-diffusion systems*, Proc. Japan Acad. 55A, 6 (1979), pp. 201–204.

[22] P. H. RABINOWITZ, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[23] D. H. SATTINGER, *Topics in Stability and Bifurcation Theory*, Lecture Notes in Mathematics 309, Springer-Verlag, Berlin, 1973.

[24] _____, *Group Theoretic Methods in Bifurcation Theory*, Lecture Notes in Mathematics 762, Springer-Verlag, Berlin, 1979.

[25] F. F. SEELIG, *Chemical oscillations by substrate inhibition. A parametrically universal oscillator type in homogeneous catalysis by metal complex formation*, Z. Naturforsch, 31A (1976), pp. 731–738.

[26] A. M. TURING, *The chemical basis of morphogenesis*, Phil. Trans. Roy. Soc. London, B237, 32 (1952), pp. 37–72.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0005 $01.00/0

ON THE COMPARISON OF SOLUTIONS OF RELATED PROPERLY AND IMPROPERLY POSED CAUCHY PROBLEMS FOR FIRST ORDER OPERATOR EQUATIONS*

KAREN A. AMES†

Abstract. Solutions of Cauchy problems for certain classes of first order operator equations are compared with solutions of associated perturbed equations. We do not require either the original problem or the perturbed problem to be well posed in the sense of Hadamard. The logarithmic convexity method is used to derive Hölder stability inequalities relating solutions of the perturbed and unperturbed problems in a suitably chosen measure. Several special cases are treated in order to demonstrate how the Lagrange identity method can be employed in the comparison of solutions as well as to indicate how certain data assumptions and requirements on the solutions can be relaxed.

1. Introduction. The aim of this paper is to compare solutions of Cauchy problems for operator equations of first order with solutions of associated perturbed equations. Neither the original problem nor the perturbed problem is required to be well posed in the sense of Hadamard. The situation in which the original problem is improperly posed and the perturbed problem well posed is referred to as the quasireversibility method (see Lattès and Lions [9]). One ill-posed problem which has been extensively studied by this method is the initial-boundary value problem for the backward heat equation (Miller [11], Ewing [6], Colton and Wimp [5], Showalter [14], [15]). In several physically interesting problems, however, the perturbed problem is the one which models the system under consideration and this perturbed problem may itself be ill posed (e.g., see Coleman, Duffin and Mizel [4]). In such cases, both the perturbed and unperturbed problems are determined a priori and one has no freedom in choosing a comparison problem.

Much of the past work on the quasireversibility method has dealt with problems for which an exact formal representation of the solution of both the perturbed and unperturbed problems can be given. Other results depend on special properties of the operators (e.g., Miller [11]). Our goal here is to obtain more general results which are not so strongly dependent on the form of the operators in the equations and do not require the perturbed problem to be well posed.

To make these ideas more precise, we shall be interested in comparing the solution of an original Cauchy problem of the form

$$Pu_t+Mu=F(t,u), \quad t\in[0,T), \qquad u(0)=f_1, \tag{1.1}$$

with the solution of a perturbed problem of the form

$$Pw_t+Mw+\varepsilon N_1 w=F(t,w), \quad t\in[0,T), \qquad w(0)=f_2 \tag{1.2}$$

or

$$Pw_t+Mw+\varepsilon N_2 w_t=F(t,w), \quad t\in[0,T), \qquad w(0)=f_2. \tag{1.3}$$

Here ε is a small positive parameter lying in an interval $0\le\varepsilon\le\varepsilon_0$. The definitions and properties of the operators and spaces involved in (1.1)–(1.3) will be made precise in the next section. One wishes to determine a stabilizing constraint set such that if u

* Received by the editors August 4, 1980, and in revised form July 2, 1981. This work was partly supported by the U.S. Army Research Office.

† Department of Mathematics, Iowa State University, Ames, Iowa 50011.

and w are both elements of this set and if f_1 and f_2 are "close" in the appropriate sense, then the solutions u and w will be "close" over a determinable time interval. Of course, if the equations describe a real physical problem, the constraint set must be realizable.

It is important to note here that for given data, the solutions of these kinds of problems may fail to exist for some or perhaps all values of ε in the interval $0 \leq \varepsilon \leq \varepsilon_0$ which is under consideration. If we allow for small variations in the data over this range of ε values, then we can in part overcome this difficulty.

Logarithmic convexity arguments are used to compare the solution of problem (1.1) (assumed to exist) with the solutions of either (1.2) or (1.3). We will show that if u and w belong to the appropriate spaces of functions, then their difference in a suitably chosen measure is of order ε to some positive power which is a function of t for $0 \leq t < T$.

We emphasize that in order to establish our stability inequalities, the solutions of the problems under study are assumed to exist. It is not our purpose in this paper to discusss the question of existence. If we were assured of the existence of the solution w for a sequence of values ε_k tending to zero such that $0 < \varepsilon_k \leq \varepsilon_0$ and of the existence of the solution u (in the appropriate space), then our results indicate that w would converge to u in the chosen norm through this sequence of values as $\varepsilon_k \to 0$.

Sections 3 and 4 of this paper are devoted to the development of stability inequalities for the problems specified in §2. We then consider several special cases in §5 in order to indicate how the Lagrange identity method may be applied to certain equations.

In a forthcoming paper, we shall present some generalizations of results that are established in [1] for second order operator equations. This subsequent paper will treat Cauchy problems for an equation of the form $Pu_{tt} + Lu_t + Mu = F(t, u, u_t)$ and the comparison equation $Pw_{tt} + Lw_t + Mw + \varepsilon Nw = F(t, w, w_t)$.

2. Statement of problems. Let H be a real Hilbert space with inner product $(\cdot, \cdot)$ and norm $\|\cdot\| = (\cdot, \cdot)^{1/2}$. We define $X \subseteq H$ to be a dense linear subspace of H and denote by P and M two linear operators (bounded or unbounded) which map X into H. We consider the following (generally ill-posed) problem

$$Pu_t + Mu = F(t, u), \quad t \in [0, T), \qquad u(0) = f_1, \tag{2.1}$$

where $f_1 \in H$ and $T > 0$. For the development of our results, we make the following assumptions:

(i) the operators P and M as well as the space H are independent of t;

(ii) P is symmetric and there is a constant $\lambda > 0$ such that $\lambda^2(P\phi, \phi) \geq \|\phi\|^2$ for all $\phi \in X$;

(iii) M is symmetric;

(iv) $u \in C^1([0, T); X)$;[1]

(v) the nonlinear term $F(t, u)$ satisfies, for $u_1, u_2 \in C^1([0, T); X)$, the uniform Lipschitz condition

$$\|F(t, u_1) - F(t, u_2)\| \leq \kappa \|u_1 - u_2\|,$$

where κ is a nonnegative constant.

[1] Here $u \in C^1([0, T); X)$ means that $u : [0, T) \to H$ is differentiable in the strong sense and that for each $t \in [0, T)$, $u, u_t \in X$.

Our goal is to investigate the relationship between the solution of (2.1) and the solutions of each of the following two problems:

$$Pw_t + Mw + \varepsilon N_1 w = F(t,w), \quad t \in [0,T), \qquad w(0) = f_2, \tag{2.2}$$

$$Pw_t + Mw + \varepsilon N_2 w_t = F(t,w), \quad t \in [0,T), \qquad w(0) = f_2. \tag{2.3}$$

Here $f_2 \in H$ and ε is a small positive parameter. We assume that N_1 and N_2 are symmetric linear operators mapping X into H and that they are independent of t. In addition, we impose the restriction that N_2 is positive semi-definite. The solutions of both (2.2) and (2.3) are assumed to belong to the space $C^1([0,T); X)$.

For the purpose of comparing the solutions u and w, we define $v \equiv w - u$ so that $v \in C^1([0,T); X)$ satisfies one of the following problems, depending upon whether (2.2) or (2.3) is under consideration:

Problem A.

$$Pv_t + Mv = -\varepsilon N_1 w + F(t,w) - F(t,u), \quad t \in [0,T), \qquad v(0) = f.$$

Problem B.

$$Pv_t + Mv = -\varepsilon N_2 w_t + F(t,w) - F(t,u), \quad t \in [0,T), \qquad v(0) = f.$$

Here the Cauchy data $f = f_2 - f_1$ are assumed to be small in the sense that there exist nonnegative constants k_i, $i = 1, \cdots, 4$ such that $(Pf,f) \le k_1 \varepsilon^2$, $|(Mf,f)| \le k_2 \varepsilon^2$, $|(N_1 f,f)| \le k_3 \varepsilon$, and $(N_2 f,f) \le k_4 \varepsilon$.

In the following two sections, we will develop stability inequalities for Problems A and B using a logarithmic convexity argument which has been widely employed in the study of improperly posed problems (see Payne [12]). We propose to prove that if u and w belong to the appropriate spaces of functions, then their difference v in a suitably chosen norm is of order ε to some positive power which is a function of t for $0 \le t < T$.

3. Inequalities for Problem A. In this section we prove the inequality which, under certain stabilizing conditions, yields continuous dependence estimates for Problem A of §2. We now establish the following theorem:

THEOREM 1. *Let u be a solution of* (2.1) *such that* $\sup_{t \in [0,T)} \| N_1 u \| \le R_1$ *for a prescribed constant R_1 and let w be a solution of* (2.2). *Assume also that a solution v of Problem A satisfies $\int_0^T (Pv,v)\, d\eta \le R_2^2$ for a constant R_2 independent of ε. Then there exist computable constants C and R_3 independent of ε such that on any compact subinterval of* $[0,T)$

$$\int_0^t (Pv,v)\, d\eta \le C \varepsilon^{2[1-\delta(t)]} R_3^{2\delta(t)}, \tag{3.1}$$

where $0 \le \delta(t) < 1$.

Proof. Consider the functional

$$\Phi(t) = \int_0^t (Pv,v)\, d\eta + (T-t)(Pf,f) + Q^2, \tag{3.2}$$

where

$$Q^2 = \beta \varepsilon^2 \lambda^2 R_1^2 + \beta_1 \varepsilon |(N_1 f,f)| + \beta_2 (Pf,f) + \beta_3 |(Mf,f)| \tag{3.3}$$

and β, β_1, β_2 and β_3 are positive constants.

We show that as a function of t, Φ satisfies a second order differential inequality of the form

$$\Phi\Phi'' - (\Phi')^2 \ge -c_1 \Phi\Phi' - c_2 \Phi^2 \tag{3.4}$$

for computable, nonnegative constants c_1 and c_2. The solution of this differential inequality then leads to the desired bounds. We have

$$\frac{d\Phi}{dt}=(Pv,v)-(Pf,f)=2\int_0^t(Pv_\eta,v)\,d\eta. \tag{3.5}$$

Substitution of the differential equation for v into the above expression yields

$$\frac{d\Phi}{dt}=-2\int_0^t(Mv,v)\,d\eta-2\varepsilon\int_0^t(N_1w,v)\,d\eta+2\int_0^t(F(\eta,w)-F(\eta,u),v)\,d\eta.$$

A second differentiation followed by a reintroduction of the differential equation, substitution of $u+v$ for w and integration results in

$$\begin{aligned}\frac{d^2\Phi}{dt^2}=&4\int_0^t(Pv_\eta,v_\eta)\,d\eta+4\varepsilon\int_0^t(N_1u,v_\eta)\,d\eta-2\varepsilon(N_1u,v)\\&-4\int_0^t(F(\eta,w)-F(\eta,u),v_\eta)\,d\eta\\&+2(F(t,w)-F(t,u),v)-2(Mf,f)-2\varepsilon(N_1f,f).\end{aligned} \tag{3.6}$$

If we apply Schwarz's inequality to the terms $4\varepsilon\int_0^t(N_1u,v_\eta)\,d\eta$ and $-2\varepsilon(N_1u,v)$ and use the assumption that $\lambda^2(P\phi,\phi)\geq\|\phi\|^2$ for a positive constant λ, we find that

$$\begin{aligned}\frac{d^2\Phi}{dt^2}\geq&4\int_0^t(Pv_\eta,v_\eta)\,d\eta-4\varepsilon\lambda\left(\int_0^t\|N_1u\|^2d\eta\right)^{1/2}\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\\&-2\varepsilon\lambda\|N_1u\|(Pv,v)^{1/2}-4\int_0^t(F(\eta,w)-F(\eta,u),v_\eta)\,d\eta\\&+2(F(t,w)-F(t,u),v)-2(Mf,f)-2\varepsilon(N_1f,f).\end{aligned} \tag{3.7}$$

To bound the two terms involving the function $F(t,w)-F(t,u)$, we make use of Schwarz's inequality, the Lipschitz behavior of F and the hypothesis on the operator P to obtain the following inequalities:

$$\begin{aligned}2(F(t,w)-F(t,u),v)&\geq-2\|v\|\,\|F(t,w)-F(t,u)\|\\&\geq-2\kappa\|v\|^2\geq-2\kappa\lambda^2(Pv,v)\end{aligned} \tag{3.8}$$

and

$$\begin{aligned}-4\int_0^t(F(\eta,w)-F(\eta,u),v_\eta)\,d\eta&\geq-4\kappa\left(\int_0^t\|v_\eta\|^2d\eta\right)^{1/2}\left(\int_0^t\|v\|^2d\eta\right)^{1/2}\\&\geq-4\kappa\lambda\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\left(\int_0^t(Pv,v)\,d\eta\right)^{1/2}\end{aligned} \tag{3.9}$$

In view of inequalities (3.8) and (3.9), we can rewrite (3.7) as

$$\begin{aligned}\frac{d^2\Phi}{dt^2}\geq&4\int_0^t(Pv_\eta,v_\eta)\,d\eta-4\varepsilon\lambda\left(\int_0^t\|N_1u\|^2d\eta\right)^{1/2}\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\\&-2\varepsilon\lambda\|N_1u\|(Pv,v)^{1/2}-4\kappa\lambda\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\left(\int_0^t(Pv,v)\,d\eta\right)^{1/2}\\&-2\kappa\lambda^2(Pv,v)-2(Mf,f)-2\varepsilon(N_1f,f).\end{aligned}$$

If we restrict u to belong to that class of functions satisfying $\sup_{t\in[0,T)}\|N_1u\|\le R_1$ for a prescribed constant R_1, we are led to the following inequality:

$$
\begin{aligned}
\frac{d^2\Phi}{dt^2}\ge 4\int_0^t(Pv_\eta,v_\eta)\,d\eta-4\varepsilon\lambda R_1\sqrt{T}\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\\
-2\varepsilon\lambda R_1(Pv,v)^{1/2}-2\kappa\lambda^2(Pv,v)\\
-4\kappa\lambda\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\left(\int_0^t(Pv,v)\,d\eta\right)^{1/2}\\
-2(Mf,f)-2\varepsilon(N_1f,f).
\end{aligned}
\tag{3.10}
$$

The quantity $\Phi\Phi''-(\Phi')^2$ can now be formed from expressions (3.2), (3.5) and (3.10). We find that

(3.11)

$$
\begin{aligned}
\Phi\Phi''-(\Phi')^2\ge S^2+4Q_1^2\int_0^t(Pv_\eta,v_\eta)\,d\eta-4\varepsilon\lambda R_1\sqrt{T}\,\Phi\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\\
-2\varepsilon\lambda R_1\Phi(Pv,v)^{1/2}-4\kappa\lambda^2\Phi\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\left(\int_0^t(Pv,v)\,d\eta\right)^{1/2}\\
-2\kappa\lambda^2\Phi(Pv,v)-2\Phi(Mf,f)-2\varepsilon\Phi(N_1f,f),
\end{aligned}
$$

where we have set

$$
S^2=4\left\{\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)\left(\int_0^t(Pv,v)\,d\eta\right)-\left(\int_0^t(Pv_\eta,v)\,d\eta\right)^2\right\}
$$

and $Q_1^2=(T-t)(Pf,f)+Q^2$. We shall now indicate how the term $-4\kappa\lambda^2\Phi(\int_0^t(Pv_\eta,v_\eta)\,d\eta)^{1/2}(\int_0^t(Pv,v)\,d\eta)^{1/2}$ can be bounded below by an expression of the form $-\alpha_1\Phi S-\alpha_2\Phi\Phi'-\alpha_3\Phi(Pf,f)$. We have

$$
\begin{aligned}
D\equiv-4\kappa\lambda^2\Phi\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\left(\int_0^t(Pv,v)\,d\eta\right)^{1/2}\\
=-2\kappa\lambda^2\Phi\left\{S^2+(\Phi')^2\right\}^{1/2}\ge-2\kappa\lambda^2\Phi(|S|+|\Phi'|).
\end{aligned}
$$

We note that S is nonnegative as a result of Schwarz's inequality and that

$$
|\Phi'|\le(Pv,v)+(Pf,f)\le\Phi'+2(Pf,f).
$$

Hence,

$$
D\ge-2\kappa\lambda^2\Phi S-2\kappa\lambda^2\Phi\Phi'-4\kappa\lambda^2\Phi(Pf,f).
$$

After an application of the arithmetic-geometric mean inequality, we find that

$$
-2\varepsilon\lambda R_1\Phi(Pv,v)^{1/2}\ge-\Phi\left[\beta\varepsilon^2\lambda^2R_1^2+\frac{1}{\beta}(Pv,v)\right],
$$

where β is a positive constant. Noting that $(Pv,v)=\Phi'+(Pf,f)$, we can rewrite (3.11) as

$$
\begin{aligned}
\Phi\Phi''-(\Phi')^2\ge\{S^2-2\kappa\lambda^2\Phi S\}+\left\{4Q_1^2\int_0^t(Pv_\eta,v_\eta)\,d\eta-4\varepsilon\lambda R_1\sqrt{T}\,\Phi\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\right\}\\
-\left(4\kappa\lambda^2+\frac{1}{\beta}\right)\Phi\Phi'-\left[\left(6\kappa\lambda^2+\frac{1}{\beta}\right)(Pf,f)+2(Mf,f)\right]\Phi\\
-\beta\varepsilon^2\lambda^2R_1^2\Phi-2\varepsilon(N_1f,f)\Phi.
\end{aligned}
$$

If we complete the square on the appropriate terms in the above inequality and discard any nonnegative terms, we obtain

$$\Phi\Phi''-(\Phi')^2\geq-\kappa^2\lambda^4\Phi^2-\left(4\kappa\lambda^2+\frac{1}{\beta}\right)\Phi\Phi'-\frac{\varepsilon^2\lambda^2R_1^2T}{Q_1^2}\Phi^2$$
$$-\left[\beta\varepsilon^2\lambda^2R_1^2+2\varepsilon(N_1f,f)+\left(6\kappa\lambda^2+\frac{1}{\beta}\right)(Pf,f)+2(Mf,f)\right]\Phi.$$

Recalling the definitions of Q_1^2 and Q^2 and observing that $\Phi\geq Q_1^2\geq Q^2\geq\beta\varepsilon^2\lambda^2R_1^2$, we find that the last two terms to the right of the above inequality are bounded below by $-\alpha_1\Phi^2$ for a computable, nonnegative constant α_1. In fact, if we set $\beta_1=\beta_3=2$ and $\beta_2=6\kappa\lambda^2+\frac{1}{\beta}$ in (3.3), then $\alpha_1=\kappa^2\lambda^4+\frac{1}{\beta}+1$. Thus, there exist computable nonnegative constants c_1 and c_2 (e.g., $c_1=4\kappa\lambda^2+\frac{1}{\beta}$, $c_2=\alpha_1$) such that the given $\Phi(t)$ satisfies an inequality of the form (3.4).

We remark here that "computable" means that the constants are known functions of T and the parameters appearing in the hypotheses on the operators and the function F as well as any coefficients introduced by the application of the arithmetic-geometric mean inequality.

It can be shown (see Levine [10]) that if a solution $\Phi(t)$ of (3.4) vanishes at some value t_0 in the interval $[0,T)$, then it must vanish identically in $[0,T)$. We may thus assume without loss of generality that $\Phi(t)>0$ for all $t\in[0,T)$. It follows that the change of variable $\sigma=e^{-c_1t}$ transforms (3.4) into

$$\frac{d^2}{d\sigma^2}\left\{\ln\left[\phi\sigma^{-c_2/c_1^2}\right]\right\}\geq0. \tag{3.12}$$

Integration of (3.12) by means of Jensen's inequality yields (in terms of the variable t)

$$\Phi(t)\leq e^{-c_2t/c_1}[\Phi(0)]^{1-\delta(t)}\left[\Phi(T)e^{c_2T/c_1}\right]^{\delta(t)}, \tag{3.13}$$

where

$$0\leq\delta(t)=\frac{1-e^{-c_1t}}{1-e^{-c_1T}}<1.$$

Since $\Phi(0)=Q_1^2$ involves only terms of $O(\varepsilon^2)$, it follows that $\Phi(0)$ is $O(\varepsilon^2)$. As has been noted previously (see Pucci [13], John [8]), it is not sufficient to have $\Phi(0)$ small in order to obtain from (3.13) a stability inequality on compact subintervals of $[0,T)$. To ensure that $\Phi(t)$ will be small for $0\leq t<T$, we must restrict the class of admissible solutions $v(t)$. A suitable stabilizing class is clearly indicated by (3.13), i.e., the set

$$\mathfrak{M}=\left\{v\in C^1([0,T),X):\int_0^T(Pv,v)\,d\eta\leq R_2^2\right\},$$

where R_2 is an a priori constant independent of ε. It follows that we can then compute a constant R_3 such that

$$\Phi(T)e^{c_2T/c_1}\leq R_3^2,$$

and inequality (3.13) leads to the result

$$\Phi(t)\leq e^{-c_2t/c_1}Q_1^{2[1-\delta(t)]}R_3^{2\delta(t)}$$

from which the assertion of the theorem follows immediately.

4. Inequalities for Problem B. An analysis similar to that employed in the previous section enables us to establish an analogue of Theorem 1 for Problem B.

THEOREM 2. *If u is a solution of Problem* (2.1) *satisfying* $\sup_{t\in[0,T)}\|N_2u_t\|\leq R_0$ *and if v is a solution of Problem* B *which lies in the class of functions* $\tilde{\mathfrak{M}}=\{v\in C^1([0,T),X):\int_0^T[(Pv,v)+\varepsilon(N_2v,v)]\,d\eta\leq R_1^2\}$, *then the following stability inequalities for t in the range $0\leq t<T$ hold*:

$$\int_0^t(Pv,v)\,d\eta\leq A\varepsilon^{2[1-\delta(t)]}R_2^{2\delta(t)}, \tag{4.1}$$

$$\int_0^t(N_2v,v)\,d\eta\leq A\varepsilon^{1-2\delta(t)}R_2^{2\delta(t)}. \tag{4.2}$$

Here A and the R_i $(i=0,1,2)$ are constants independent of ε and $0\leq\delta(t)<1$.

The proof of this theorem rests on the application of logarithmic convexity arguments to the functional

$$\Phi(t)=\int_0^t\{(Pv,v)+\varepsilon(N_2v,v)\}\,d\eta+(T-t)\{(Pf,f)+\varepsilon(N_2f,f)\}+Q^2, \tag{4.3}$$

where

$$Q^2=\beta\varepsilon^2\lambda^2R_0^2+\beta_1\varepsilon(N_2f,f)+\beta_2(Pf,f)+\beta_3|(Mf,f)| \tag{4.4}$$

and $\beta,\beta_1,\cdots,\beta_3$ are appropriately chosen positive constants. Since this functional can be shown to satisfy a second order differential inequality of the form (3.4), the desired stability estimates (4.1) and (4.2) can be obtained under the hypotheses of the theorem. More precisely,

$$\begin{aligned}\frac{d\Phi}{dt}&=(Pv,v)+\varepsilon(N_2v,v)-\{(Pf,f)+\varepsilon(N_2f,f)\}\\&=2\int_0^t\{(Pv_\eta,v)+\varepsilon(N_2v_\eta,v)\}\,d\eta.\end{aligned} \tag{4.5}$$

Use of the differential equation for Problem B leads to the expression

$$\frac{d\Phi}{dt}=-2\int_0^t(Mv,v)\,d\eta-2\varepsilon\int_0^t(N_2u_\eta,v)\,d\eta+2\int_0^t(F(\eta,w)-F(\eta,u),v)\,d\eta.$$

We then find

$$\begin{aligned}\frac{d^2\Phi}{dt^2}=&4\int_0^t\{(Pv_\eta,v_\eta)+\varepsilon(N_2v_\eta,v_\eta)\}\,d\eta+4\varepsilon\int_0^t(N_2u_\eta,v_\eta)\,d\eta-2\varepsilon(N_2u_t,v)\\&+2(F(t,w)-F(t,u),v)-4\int_0^t(F(\eta,w)-F(\eta,u),v_\eta)\,d\eta-2(Mf,f).\end{aligned}$$

If we make use of assumptions (ii) and (v) as well as Schwarz's inequality and the restriction on the solution $u(t)$, we generate the following inequality for $d^2\Phi/dt^2$:

$$\begin{aligned}\frac{d^2\Phi}{dt^2}\geq&4\int_0^t\{(Pv_\eta,v_\eta)+\varepsilon(N_2v_\eta,v_\eta)\}\,d\eta-4\varepsilon\lambda R_0\sqrt{T}\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\\&-2\varepsilon\lambda R_0(Pv,v)^{1/2}-2\kappa\lambda^2(Pv,v)-4\kappa\lambda^2\left(\int_0^t(Pv,v)\,d\eta\right)^{1/2}\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}\\&-2(Mf,f).\end{aligned}$$

We now form

$$\Phi\Phi''-(\Phi')^2\geq S^2+4Q_1^2\int_0^t(Pv_\eta,v_\eta)\,d\eta-4\varepsilon\lambda R_0\sqrt{T}\,\Phi\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}$$

$$-2\varepsilon\lambda R_0\Phi(Pv,v)^{1/2}-2\kappa\lambda^2\Phi(Pv,v) \tag{4.6}$$

$$-4\kappa\lambda^2\Phi\left(\int_0^t(Pv,v)\,d\eta\right)^{1/2}\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}-2\Phi(Mf,f).$$

Here

$$S^2=4\left[\int_0^t\{(Pv_\eta,v_\eta)+\varepsilon(N_2v_\eta,v_\eta)\}\,d\eta\right]\left[\int_0^t\{(Pv,v)+\varepsilon(N_2v,v)\}\,d\eta\right]$$

$$-4\left[\int_0^t\{(Pv_\eta,v)+\varepsilon(N_2v_\eta,v)\}\,d\eta\right]^2$$

is nonnegative as a result of Schwarz's inequality and we have written $Q_1^2=(T-t)\{(Pf,f)+\varepsilon(N_2f,f)\}+Q^2$. Application of the arithmetic-geometric mean and Schwarz's inequality to appropriate terms in (4.6) yields

$$\Phi\Phi''-(\Phi')^2\geq S^2-2\kappa\lambda^2\Phi S+4Q_1^2\int_0^t(Pv_\eta,v_\eta)\,d\eta-4\varepsilon\lambda R_0\sqrt{T}\,\Phi\left(\int_0^t(Pv_\eta,v_\eta)\,d\eta\right)^{1/2}$$

$$-\left(4\kappa\lambda^2+\frac{1}{\beta}\right)\Phi\Phi'-\left[\left(6\kappa\lambda^2+\frac{1}{\beta}\right)\{(Pf,f)+\varepsilon(N_2f,f)\}\right.$$

$$\left.+2(Mf,f)+\beta\varepsilon^2\lambda^2R_0^2\right]\Phi.$$

Upon completion of squares, the previous expression leads to the inequality

$$\Phi\Phi''-(\Phi')^2\geq -c_1\Phi\Phi'-c_2\Phi^2,$$

where c_1 and c_2 are computable, nonnegative constants. Hence, it follows that for $0\leq t<T$

$$\Phi(t)\leq e^{-c_2t/c_1}[\Phi(0)]^{1-\delta(t)}[\Phi(T)e^{c_2T/c_1}]^{\delta(t)}, \tag{4.7}$$

where $\delta(t)=(1-e^{-c_1t})/(1-e^{-c_1T})$. In the event that $v\in\tilde{\mathfrak{N}}$, inequality (4.7) contains the information that

$$\Phi(t)\leq A\varepsilon^{2[1-\delta(t)]}R_2^{2\delta(t)}$$

for computable constants A and R_2 independent of ε. This completes the proof of the theorem.

5. Some special cases. In order to establish the theorems of §§3 and 4, we made use of the logarithmic convexity method. Although this method is more widely applicable than the so-called Lagrange identity method (see [12]), it is nevertheless the case that in certain special classes of linear problems the latter method requires less severe data assumptions. In this section we specialize the operators F, N_1 and N_2, sketch how these specializations lead to a less restrictive constraint class requirement and demonstrate how data assumptions can sometimes be relaxed if we employ the Lagrange identity method in the comparison of solutions.

Throughout this section we shall assume that F is independent of the solution u. The general linear problem of the form $F=\rho u+\mathcal{F}(t)$ where $\rho\in H$ is independent of t and $\mathcal{F}\in H$ for all $t\in[0,T)$ may be reduced to this case by merely redefining the operator M as $M=\tilde{M}-\rho I$.

In the following examples we assume that $P\equiv I$ and, unless otherwise indicated, retain hypotheses (i)–(iv) as well as the requirements on the operators N_1 and N_2.

5.1. Problem A with $N_1=M^2$ and $F\equiv 0$. The first problem with which we deal is

$$v_t+Mv=-\varepsilon M^2 w,\quad t\in[0,T)\qquad v(0)=f.$$

We introduce a class of functions $\mathfrak{W}=\{\phi\in C^1([0,T),X):\int_0^{t^*}\|\phi\|^2\,d\eta\le m^2\}$ for a prescribed constant m and some $t^*>T$. Our results in this case can be stated in the following corollary.

COROLLARY 1. *Let u and w be solutions of* (2.1) *and* (2.2), *respectively, with $P=I$, $F\equiv 0$, $N_1=M^2$. If $u\in\mathfrak{W}$ and if w lies in that class of functions for which $\int_0^T\|w\|^2\,d\eta\le n^2$ where n is a constant independent of ε, then the solution of* (5.1) *satisfies the following stability inequality for $0\le t<T$:*

$$\int_0^t\|v\|^2\,d\eta\le C\varepsilon^{2(1-t/T)}R_2^{2t/T}. \tag{5.2}$$

Here C and R_2 are constants independent of ε.

The proof of this corollary via the logarithmic convexity method is similar to that of Theorem 1. The functional $\Phi(t)=\int_0^t\|v\|^2\,d\eta+(T-t)\|f\|^2+Q^2$, where $Q^2=\beta_1\varepsilon^2R_1^2+\beta_2\varepsilon^2\|Mf_1\|^2+\beta_3\|f\|^2+\beta_4\|Mf\|^2$, can be shown to satisfy an inequality of the form (3.4) with $c_1=0$. To establish this result, we use a slightly different expression for $d^2\Phi/dt^2$ than previously cited. Instead of (3.6), we take

$$\begin{aligned}\frac{d^2\Phi}{dt^2}=4\int_0^t\|v_\eta\|^2\,d\eta+2\varepsilon\int_0^t\big(M^2u,v_\eta\big)\,d\eta-2\varepsilon\int_0^t\big(M^2u_\eta,v\big)\,d\eta\\ -2(Mf,f)-2\varepsilon\|Mf\|^2-2\varepsilon(Mf_1,Mf).\end{aligned} \tag{5.3}$$

The desired differential inequality is obtained by employing the Schwarz and arithmetic-geometric mean inequalities in (5.3) and using the resulting expression in our calculation of $\Phi\Phi''-(\Phi')^2$. The result is also contingent upon a restriction of the solution u to an appropriate class of functions which will ensure that terms of the form $\int_0^t\|M^2u_\eta\|^2\,d\eta$ and $\int_0^t\|M^2u\|^2\,d\eta$ are bounded for $t\in[0,T)$.

We show here that if $u\in\mathfrak{W}$ and if its initial data satisfy a particular boundedness condition, then the previously indicated integrals are bounded. To this end, we introduce a function $\gamma(t)\in C^6$ $(t\ge 0)$ defined as follows:

$$\begin{aligned}&\gamma(t)\equiv 1,\quad 0\le t\le t_0\le T,\qquad 0\le\gamma(t)\le 1,\quad t_0\le t\le t_1,\\ &\gamma(t)\equiv 0,\quad t\ge t_1.\end{aligned} \tag{5.4}$$

Then we have

$$\int_0^t\|M^2u_\eta\|^2\,d\eta\le\int_0^\infty\gamma(\eta)\|M^2u_\eta\|^2\,d\eta.$$

Noting that $\gamma^{(\alpha)}(t)=0$ (for $\alpha=1,\cdots,6$) at both limits of integration and recalling that M is a symmetric operator, we find that repeated substitution of the differential

equation $u_t+Mu=0$ (now assumed to hold for $0\le t<t_1$) and integration by parts yields

$$\begin{aligned}\int_0^\infty \gamma(\eta)\|M^2u_\eta\|^2\,d\eta &= -\int_0^\infty \gamma(\eta)\big(M^3u, M^2u_\eta\big)\,d\eta\\ &= \frac{1}{2}(M^3u, M^2u)|_{t=0}+\frac{1}{2}\int_0^\infty \gamma'(\eta)(M^3u, M^2u)\,d\eta\\ &= \hat{Q}+\frac{1}{4}\int_0^\infty \gamma''(\eta)\|M^2u\|^2\,d\eta\\ &\vdots\\ &= \hat{Q}+\frac{1}{64}\int_0^\infty \gamma^{(\mathrm{vi})}(\eta)\|u\|^2\,d\eta\end{aligned}$$

where $\hat{Q}=\frac{1}{2}(M^3f_1, M^2f_1)$. Since $\gamma^{(\mathrm{vi})}(t)$ is bounded for $t\ge 0$, it follows that if we require $u\in\mathfrak{W}$ (choosing $t^*=t_1$) and $\hat{Q}$ to be bounded, then we can compute a constant R_1 such that $\int_0^t\|M^2u_\eta\|^2\,d\eta\le R_1^2$. In a similar way, we can compute a bound for $\int_0^t\|M^2u\|^2\,d\eta$.

Remark 5.1. If M is a negative semi-definite operator, then the data term $\hat{Q}$ is nonpositive and can consequently be discarded from the bounding inequality.

In addition to the restrictions on the solutions u and w, we must also assume that $\|f\|$ and $\|Mf\|$ are $O(\varepsilon)$ and that $\|Mf_1\|\le\tilde{C}$ for a prescribed constant $\tilde{C}$ in order to obtain (5.2).

The result (5.2) can be deduced from a Lagrange identity analysis of problem (5.1) and, as we shall see, the requirements on the initial data can be somewhat relaxed. Let us assume that v^* is any solution of the adjoint equation of (5.1), that u^* satisfies the equation $u_t^*=Mu^*$ and that $w^*=u^*+v^*$. For such functions, we have the following identity:

$$0=\int_0^t\Big\{\big(v^*, v_\eta+Mv+\varepsilon M^2w\big)-\big(v, -v_\eta^*+Mv^*+\varepsilon M^2w^*\big)\Big\}\,d\eta,$$

which, upon integration, leads to the equation

$$[(v^*,v)]\Big|_0^t=\varepsilon\int_0^t\{(v, M^2u^*)-(v^*, M^2u)\}\,d\eta.$$

If we assume $0\le t\le 2t<T$ and choose $v^*(\eta)=v(2t-\eta)$ and $u^*(\eta)=u(2t-\eta)$, we find that

$$\|v\|^2\le(v(2t),v(0))+\varepsilon\int_0^{2t}|(v(\eta), M^2u(2t-\eta))|\,d\eta.$$

Application of Schwarz's inequality followed by an integration from 0 to t yields

$$\int_0^t\|v\|^2\,d\eta\le\int_0^t(v(2\eta),f)\,d\eta+\varepsilon\int_0^t\left\{\left(\int_0^{2\sigma}\|v\|^2\,d\eta\right)^{1/2}\left(\int_0^{2\sigma}\|M^2u\|^2\,d\eta\right)^{1/2}\right\}d\sigma.$$

After using Schwarz's inequality again, we can manipulate the previous inequality to obtain

$$\begin{aligned}\int_0^t\|v\|^2\,d\eta\le&\left(\frac{t}{2}\right)^{1/2}\|f\|\left(\int_0^{2t}\|v\|^2\,d\eta\right)^{1/2}\\ &+\varepsilon t\left(\int_0^{2t}\|v\|^2\,d\eta\right)^{1/2}\left(\int_0^{2t}\|M^2u\|^2\,d\eta\right)^{1/2}.\end{aligned}\tag{5.5}$$

Thus,[2] if $u \in \mathfrak{W}$ for $t^* > T$, $\int_0^T \|w\|^2 d\eta \leq n^2$ and the initial data are small in the sense that $\|f\| = O(\varepsilon)$, then there are nonnegative constants C and R such that

$$\int_0^t \|v\|^2 d\eta \leq C\varepsilon R, \qquad 0 \leq t \leq \frac{T}{2}. \tag{5.6}$$

Here R is such that $\int_0^T \|v\|^2 d\eta \leq R^2$. Hölder continuous dependence on the data for $t \in [0, T)$ can be generated using (5.5) and (5.6) (see Payne [12]). Hence, we see that our stability results for this particular case are actually obtainable under somewhat less restrictive conditions on the initial data than needed in the logarithmic convexity analysis.

One example of this case is the choice $M = \Delta$, the Laplace operator. It is this prototype that Lattès and Lions used to illustrate the method of quasireversibility. If we assume the equations for u and w hold in a bounded domain $D \subseteq \mathbb{R}^n$ with a sufficiently smooth boundary ∂D and if we take as our boundary conditions $u(x,t) = 0$ and $w(x,t) = \Delta w(x,t) = 0$ for $x \in \partial D$, then the resulting initial-boundary value problems can be readily compared using the previous arguments. The results of Corollary 1 thus validate the quasireversibility procedure in this particular case. We note that since Δ is a negative operator, the observation of Remark 5.1 is valid and allows us to establish the stability inequality under less restrictive data assumptions. In addition, the a priori requirement that $\int_0^T \|w\|^2 d\eta \leq n^2$ in Corollary 1 can be eliminated in this case. In fact, this restriction can be dropped whenever w satisfies an equation of the form $w_t = -[\varepsilon G(M) + M]w$ where $G(M)$ is a nonnegative operator and M is nonpositive. In these situations, it is possible to bound $\int_0^T \|w\|^2 d\eta$ in terms of $\int_0^T \|u\|^2 d\eta$.

Remark 5.2. The results of Corollary 1 can be easily generalized to include the case in which $N_1 = \sum_{j=1}^k c_j M^j$.

5.2. Problem B with $F = \mathfrak{F}(t)$. We now consider problem (2.1) in the case where F is independent of the solution u and compare its solution with that of problem (2.3). In our analysis, we set $F = \mathfrak{F}(t)$ where $\mathfrak{F}$ is a prescribed vector-valued function. Using the Lagrange identity method, we shall prove the following corollary. We remark here that a proof via the logarithmic convexity method is merely a specialization of the argument used in the proof of Theorem 2.

COROLLARY 2. *Let u and w be solutions of* (2.1) *and* (2.3), *respectively, with $P = I$ and $F = \mathfrak{F}(t)$. The initial data f corresponding to the difference problem is required to satisfy $\|f\| = O(\varepsilon)$ and $\|N_2 f\| = O(1)$. In addition, we impose the requirements that $\int_0^{t^*} \|N_2 \mathfrak{F}\|^2 d\eta \leq n^2$ for a constant n and $t^* > T$ and that the operators M and N_2 commute. If u belongs to that class of functions defined by the conditions that $\int_0^{t^*} \|N_2 u\|^2 d\eta \leq m_1^2$ and that the difference $v = w - u$ satisfies $\int_0^T \|v\|^2 d\eta \leq m_2^2$ for prescribed constants m_1 and m_2, then it follows that for t in the range $0 \leq t < T$*

$$\int_0^t \|v\|^2 d\eta = O(\varepsilon^{2(1-t/T)}) \quad \text{and} \quad \int_0^t (N_2 v, v)\, d\eta = O(\varepsilon^{1-2t/T}).$$

Proof. We define v^* and u^* to be solutions of the respective adjoint equations corresponding to the special case of Problems B and (2.1) under consideration and take $w^* = v^* + u^*$. Then, we have

$$\begin{aligned} 0 &= \int_0^t \left\{ (v^*, v_\eta + Mv + \varepsilon N_2 w_\eta) - (v, -v_\eta^* + Mv^* - \varepsilon N_2 w_\eta^*) \right\} d\eta \\ &= \left[(v^*, v) + \varepsilon (N_2 v, v^*) \right]\Big|_0^t + \varepsilon \int_0^t \left\{ (v^*, N_2 u_\eta) + (v, N_2 u_\eta^*) \right\} d\eta. \end{aligned}$$

[2] We must again require that u satisfies $u_t + Mu = 0$ for $0 \leq t < t_1$.

We again assume $0\leq t\leq 2t<T$; then the choices $v^*(\eta)=v(2t-\eta)$ and $u^*(\eta)=u(2t-\eta)$ allow us to write the previous expression as

$$\|v\|^2+\varepsilon(N_2v,v)\leq(v(2t),f)+\varepsilon(N_2f,v(2t))+\varepsilon\int_0^{2t}|(v(\eta),N_2u_\eta(2t-\eta))|\,d\eta.$$

Using Schwarz's inequality and integrating both sides from 0 to t, we obtain after some manipulation the following inequality

$$\begin{aligned}\int_0^t\{\|v\|^2+\varepsilon(N_2v,v)\}\,d\eta\leq\left(\frac{t}{2}\right)^{1/2}[\|f\|+\varepsilon\|N_2f\|]\left(\int_0^{2t}\|v\|^2d\eta\right)^{1/2}\\ +\varepsilon t\left(\int_0^{2t}\|N_2u_\eta\|^2d\eta\right)^{1/2}\left(\int_0^{2t}\|v\|^2d\eta\right)^{1/2}.\end{aligned}\tag{5.7}$$

If we introduce a function $\gamma(t)\in C^2$ $(t\geq 0)$ as defined in (5.4), we can show by substituting the differential equation for u and judiciously applying the arithmetic-geometric mean and Schwarz inequalities that

$$\begin{aligned}\int_0^{2t}\|N_2u_\eta\|^2d\eta\leq\int_0^\infty\gamma(\eta)\|N_2u_\eta\|^2d\eta\\ \leq B+\alpha_1\int_0^{t_1}\|N_2u\|^2d\eta+\alpha_2\int_0^{t_1}\|N_2\mathcal{F}\|^2d\eta,\end{aligned}\tag{5.8}$$

where $B=(N_2u,N_2Mu)|_{t=0}$ and α_1 and α_2 are nonnegative constants which are computable from the bounds satisfied by $\gamma'(t)$ and $\gamma''(t)$. We note that inequality (5.8) was obtained under the assumption that $N_2M=MN_2$. Hence, if we require B to be bounded, the hypotheses on u and $\mathcal{F}$ in the corollary are sufficient to guarantee that $\int_0^{2t}\|N_2u_\eta\|^2d\eta\leq R^2$ for a constant R. In view of this bound and the given restrictions on v and the initial data, the desired results can be deduced from (5.7).

Remark 5.3. If $N_2=M^{2j}$ for a positive integer j or if N_2 is a linear combination of such operators, then it is possible to establish a bound for $\int_0^{2t}\|N_2u_\eta\|^2d\eta$ by assuming that u satisfies the weaker condition $\int_0^{t^*}\|u\|^2d\eta\leq m^2$ for a prescribed constant m and some $t^*>T$. In order to do this, several additional conditions must be imposed on the function $\mathcal{F}(t)$. It should also be noted that if M is a negative semi-definite operator, then the data term B is nonpositive. Since it may then be dropped from (5.8), the boundedness condition on the initial data for u is no longer necessary.

One example to which these remarks are relevant is that of $M=\Delta=-N_2$. This case is of particular interest when Δ is the one-dimensional Laplacian and $\mathcal{F}$ is a scalar function of t alone because of its connection with models of such physical phenomena as clay consolidation (Taylor [16]) and nonsteady shearing flow in second-order fluids (Huilgol [7]).

Equations of the type $w_t+\Delta w-\varepsilon\Delta w_t=\mathcal{F}(t)$ also appear in the theory of seepage of liquids in fissured rocks backward in time [2] and in a theory of backward heat conduction in a material for which the conductive and thermodynamic temperature do not coincide [3]. In the first case, the reduced equation ($\varepsilon=0$) models seepage of liquids under elastic conditions while in the second, setting $\varepsilon=0$ leads to the classical heat conduction problem (backward in time).

6. Concluding remarks. In this paper we have assumed that solutions of the equations under consideration are $C^1([0,T);X)$. However, results similar to those obtained here can most certainly be established for weak solutions. For certain classes of equations, it is convenient to work with appropriately defined weak solutions since they may exist, whereas classical solutions may not exist. (See Payne [12] and the references cited therein.)

At this point, the reader may ask whether there are other perturbations of problem (2.1) which would also serve as good or perhaps better approximations. If the function F is independent of the solution u, we can prove the same type of results presented in this paper using a perturbed problem of the form

$$Pw_t + Mw - \varepsilon Pw_{tt} = \mathcal{F}(t), \quad t \in [0, T), \qquad w(0) = f_2, \qquad w_t(0) = g_2.$$

These stability estimates are obtainable from a Lagrange identity analysis similar to those exhibited in §5.

Whether or not one could prove similar results for equations with nonsymmetric, time dependent operators or for more general nonlinear problems is another question to be pursued. We propose to determine to what extent we could generalize the operators in the equations considered here and still guarantee Hölder stability. If we are also willing to work with weaker norms, then it is possible that the desired results can be obtained for a fairly broad class of operators.

Acknowledgment. The author wishes to thank Professor L. E. Payne of Cornell University for his encouragement, guidance and patience during the preparation of both this paper and the dissertation from which these results have been taken.

REFERENCES

[1] K. A. AMES, *Comparison results for related properly and improperly posed problems, with applications to mechanics*, Ph.D. thesis, Cornell University, Ithaca, NY, 1980.

[2] G. BAHRENBLATT, I. ZHELTOV AND I. KOCHIVA, *Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks*, J. Appl. Math. Mech., 24 (1960), pp. 1286–1303.

[3] P. J. CHEN AND M. E. GURTIN, *On a theory of heat conduction involving two temperatures*, ZAMP, 19 (1968), pp. 614–627.

[4] B. COLEMAN, R. J. DUFFIN AND V. MIZEL, *Instability, uniqueness and nonexistence theorems for the equation* $u_t = u_{xx} - u_{xtx}$ *on a strip*, Arch. Rational Mech. Anal., 19 (1965), pp. 100–116.

[5] D. COLTON AND J. WIMP, *The construction of solutions to the heat equation backward in time*, Math. Meth. Appl. Sci., 1 (1979), pp. 32–39.

[6] R. E. EWING, *The approximation of certain parabolic equations backward in time by Sobolev equations*, this Journal, 6 (1975), pp. 283–294.

[7] R. R. HUILGOL, *A second order fluid of the differential type*, Internat. J. Nonlinear Mech. 3 (1968), pp. 471–482.

[8] F. JOHN, *Continuous dependence on data for solutions of partial differential equations with a prescribed bound*, Comm. Pure Appl. Math., 13 (1960), pp. 551–585.

[9] R. LATTÈS AND J. L. LIONS, *The Method of Quasireversibility, Applications to Partial Differential Equations*, American Elsevier, New York, 1969.

[10] H. A. LEVINE, *Logarithmic convexity and the Cauchy problem for some abstract second-order differential inequalities*, J. Differential Equations, 8 (1970), pp. 34–55.

[11] K. MILLER, *Stabilized quasi-reversibility and other nearly-best-possible methods for non-well-posed problems*, Symposium on Non-well-posed problems and Logarithmic Convexity, Lecture Notes in Mathematics, 316, Springer-Verlag, New York, 1973, pp. 161–176.

[12] L. E. PAYNE, *Improperly Posed Problems in Partial Differential Equations*, CBMS Regional Conference Series in Applied Mathematics, 22, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[13] C. PUCCI, *Discussione del problema di Cauchy per le equazioni di tipo ellitico*, Ann. Mat. Pura Appl., 46 (1958), pp. 131–153.

[14] R. E. SHOWALTER, *The final value problem for evolution equations*, J. Math. Anal. Appl., 47 (1974), pp. 563–572.

[15] R. E. SHOWALTER, *Quasi-Reversibility of first and second order parabolic evolution equations*, Pitman Research Notes in Mathematics, 1, Pitman, London, 1975, pp. 76–84.

[16] D. W. TAYLOR, *Research on Consolidation of Clays*, MIT Press, Cambridge, MA., 1942.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0006 $01.00/0

THE DIFFERENTIABILITY WITH RESPECT TO A PARAMETER OF THE SOLUTION OF A LINEAR ABSTRACT CAUCHY PROBLEM*

DENNIS W. BREWER†

Abstract. The differentiability with repect to a parameter of the solution of a linear inhomogeneous abstract Cauchy problem is considered by employing the theory of strongly continuous semigroups. Criteria for differentiability are developed which allow the parameter to appear in unbounded terms of the generator. Sufficient conditions are given for the solvability of an inhomogeneous Cauchy problem with zero initial condition by the variation of constants formula. These conditions are used to show that the derivatives with respect to a parameter satisfy certain sensitivity equations in time. The results are applied to the differentiability of the solution of a linear delay differential equation with respect to its delays.

1. Introduction. The purpose of this paper is to consider the differentiability with respect to a parameter p of the solution of the linear abstract Cauchy problem

$$x'(t)=A(p)x(t)+u(t), \qquad x(0)=x_0. \tag{1}$$

Differentiability results of this type are required for the application of gradient methods to parameter identification problems for control systems governed by equations which can be formulated as abstract Cauchy problems. (See [2] for an example.) In [6] this problem is considered for solutions of (1) in which $A(p)$ is of the form $A+B(p)$, where $B(p)$ is a bounded linear operator. In the present paper we wish to prove differentiability under assumptions which allow the parameter to appear in unbounded terms. In §2 we consider the differentiability with respect to p of the solution of the linear homogeneous equation

$$x'(t)=A(p)x(t), \qquad x(0)=x_0. \tag{2}$$

In §3 we obtain similar results for the solution of the linear inhomogeneous equation

$$x'(t)=A(p)x(t)+u(t), \qquad x(0)=0. \tag{3}$$

In §4 we prove that under certain conditions these derivatives with respect to p satisfy inhomogeneous evolution equations called sensitivity equations. Section 5 concerns an application of these results to the sensitivity of the solution of a functional differential equation with respect to multiple delays.

2. The homogeneous equation. Let P be an open subset of a normed linear space $\mathcal{P}$ with norm $|\cdot|$, and let X be a Banach space with norm $\|\cdot\|$. For every $p\in P$, let $A(p)$ be a linear operator on $D(A(p))$ in X. Throughout this paper we assume

(H1) $A(p)$ generates a strongly continuous semigroup $S(t;p)$ on X.

(H2) $D(A(p))=D$ is independent of p.

(H3) $\|S(t;p)x\|\le Me^{\omega t}\|x\|$, $x\in X$, $t\ge 0$, $p\in P$, for some constants M and ω independent of p, x, t.

If $x_0\in D$, then the solution of (2) is given by $x(t)=S(t;p)x_0$. We will consider the slightly more general problem of differentiating $S(t;p)x_0$ with respect to p for any fixed $x_0\in X$. Differentiation will be in the sense of Fréchet.

* Received by the editors June 4, 1980, and in revised form June 2, 1981. This research was done while the author was on leave of absence at Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

† Department of Mathematics, University of Arkansas, Fayetteville, Arkansas 72701.

We will assume that $A(p)=A+B(p)$, where A and $B(p)$ both have domain D and A is independent of p. Fixing $p_0\in P$ and $T>0$, we assume $B(p)$ satisfies the following hypothesis:

(H4) For every $p\in P$, there is a constant $K>0$ such that

$$\int_0^T \| B(p)S(t;p_0)x\|\, dt\le K\|x\|, \qquad x\in D.$$

Note that (H4) allows $B(p)$ to be an unbounded operator (see §5). It does imply, however, that the linear mapping $x\to B(p)S(\cdot;p_0)x$ from D into $L^1(0,T;X)$ is bounded on D. Let $F(p)$ denote the bounded linear extension of this mapping to $\overline{D}=X$. Let $|||\cdot|||$ denote the norm in $L^1(0,T;X)$. Concerning $F(p)$ we assume the following:

(H5) There is a closed subspace Y of X such that
(i) $F(p)x_0\in L^1(0,T;Y)$ for every $p\in P$, and
(ii) for every $\varepsilon>0$, there exists $\delta>0$ such that

$$|||F(p_0+h)y-F(p_0)y|||\le\varepsilon\|y\|$$

whenever $y\in Y$ and $|h|\le\delta$.

The next theorem gives sufficient conditions for the differentiability of $S(t;p)x_0$ with respect to p.

THEOREM 1. *Suppose* (H1)–(H5) *hold. In addition, suppose*

(H6) $F(p)x_0$ *is Fréchet differentiable with respect to p at* p_0.

Then for every $t\in[0,T]$, $S(t;p)x_0$ *is Fréchet differentiable with respect to p at* p_0 *and*

$$D_pS(t;p_0)x_0=\int_0^t S(t-s;p_0)\big[D_pF(p_0)x_0\big](s)\,ds, \qquad 0\le t\le T.$$

Remark. Before beginning the proof of this theorem, some comments regarding hypothesis (H4) are in order. This hypothesis is clearly satisfied for bounded perturbations as in [6]. It is quite similar to one of the hypotheses employed by Hille and Phillips to obtain a perturbation series, namely

$$\text{(H4)}^* \qquad \begin{aligned} &B(p)S(t;p_0) \quad \text{bounded on } D \text{ for } t>0,\\ &\int_0^T \| B(p)S(t;p_0)\|_D\, dt<\infty. \end{aligned}$$

(see [8, p. 394] and following). Note that (H4)* implies (H4). In §5 we will discuss an important application to delay differential equations for which (H4) holds but (H4)* does not hold. Of course the perturbation series approach given in [8] yields derivatives of all orders for a semigroup with respect to a perturbation, whereas here we obtain only the first. Generally speaking, (H4)* seems best suited for applications to parabolic partial differential equations, whereas (H4) yields a perturbation theory applicable to delay equations and perhaps more general forms of functional differential equations as indicated in §5. Hypotheses (H4) may not be applicable to hyperbolic partial differential equations or neutral functional differential equations for which the solution operator has no smoothing properties in time.

This paper deals only with a constructive method for a linear abstract Cauchy problem. Differentiability properties for solutions of various types of nonlinear equations have been investigated using fixed point properties of uniform contractions, e.g., [7, §2.4].

The author is indebted to the referees for some of the remarks made here.

Proof. Choose $x\in D$, p_0, $p_0+h\in P$, and let $w(t)=S(t;p_0+h)x-S(t;p_0)x$, $t\geq 0$. Then it is easy to see that w is strongly differentiable and satisfies $w'(t)=A(p_0+h)w(t)+[B(p_0+h)-B(p_0)]S(t;p_0)x, t\geq 0$, $w(0)=0$. Note that for $\lambda>\omega$

$$[B(p_0+h)-B(p_0)]S(t;p_0)x$$
$$=A(p_0+h)(\lambda I-A(p_0))^{-1}S(t;p_0)(\lambda I-A(p_0))x-S(t;p_0)A(p_0)x.$$

Since $A(p_0+h)(\lambda I-A(p_0))^{-1}$ is bounded by the closed graph theorem, $[B(p_0+h)-B(p_0)]S(t;p_0)x$ is strongly continuous in t. Therefore, by standard arguments (e.g., [9, p. 488]), we have

$$S(t;p_0+h)x-S(t;p_0)x=\int_0^t S(t-s;p_0+h)(B(p_0+h)-B(p_0))S(s;p_0)x\,ds \tag{4}$$

whenever $p_0+h\in P$, $t\geq 0$ and $x\in D$. It now follows from the definition of F that if $x\in D$ then

$$S(t;p_0+h)x-S(t;p_0)x=\int_0^t S(t-s;p_0+h)[(F(p_0+h)x)(s)-(F(p_0)x)(s)]\,ds. \tag{5}$$

Since $F(p)$ is a bounded operator from X into $L^1(0,T;X)$, it is easy to see that both sides of (5) are bounded operators on X, and therefore (5) is true for all $x\in X$. According to (4) and (H3), if $C=Me^{\omega T}$, then

$$\|S(t;p_0+h)x-S(t;p_0)x\|\leq C|||F(p_0+h)x-F(p_0)x|||. \tag{6}$$

Let $\varepsilon>0$ be given; then by (6) and (H5) there exists $\delta>0$ such that if $|h|\leq\delta$, then

$$\|S(t;p_0+h)y-S(t;p_0)y\|\leq\varepsilon C\|y\|, \tag{7}$$

for all $y\in Y$.Therefore, again by (H5),

$$\int_0^t\|[S(t-s;p_0+h)-S(t-s;p_0)][(F(p_0+h)x_0)(s)-(F(p_0)x_0)(s)]\|\,ds$$
$$\leq\int_0^t\varepsilon C\|(F(p_0+h)x_0)(s)-(F(p_0)x_0)(s)\|\,ds \tag{8}$$
$$\leq\varepsilon C|||F(p_0+h)x_0-F(p_0)x_0|||,\qquad 0\leq t\leq T,\quad |h|\leq\delta.$$

According to (H6) there is a number $\eta\in(0,\delta)$ such that

$$|||F(p_0+h)x_0-F(p_0)x_0-D_pF(p_0)h|||\leq\varepsilon|h|,\qquad |h|\leq\eta,$$

where $D_pF(p_0)=D_p[F(p_0)x_0]$ is a bounded linear operator from the parameter space $\mathcal{P}$ into $L^1(0,T;X)$. Let M_1 denote the norm of this operator. Then if $|h|\leq\eta$ and $0\leq t\leq T$

the above inequalities yield

$$\begin{aligned}
&\left\| S(t;p_0+h)x_0 - S(t;p_0)x_0 - \int_0^t S(t-s;p_0)\left[D_pF(p_0)h\right](s)\,ds \right\| \\
&\quad \le \int_0^t \left\| \left[S(t-s;p_0+h)-S(t-s;p_0)\right]\left[(F(p_0+h)x_0)(s)-(F(p_0)x_0)(s)\right] \right\| ds \\
&\qquad + \int_0^t \left\| S(t-s;p_0)\left[(F(p_0+h)x_0)(s)-(F(p_0)x_0)(s)-\left(D_pF(p_0)h\right)(s)\right] \right\| ds \\
&\quad \le \varepsilon C \left| \left| F(p_0+h)x_0 - F(p_0)x_0 \right| \right| + C \left| \left| F(p_0+h)x_0 - F(p_0)x_0 - D_pF(p_0)h \right| \right| \\
&\quad \le \varepsilon C \left| \left| F(p_0+h)x_0 - F(p_0)x_0 - D_pF(p_0)h \right| \right| + \varepsilon C \left| \left| D_pF(p_0)h \right| \right| \\
&\qquad + C \left| \left| F(p_0+h)x_0 - F(p_0)x_0 - D_pF(p_0)h \right| \right| \\
&\quad \le \varepsilon^2 C|h| + \varepsilon C M_1 |h| + \varepsilon C|h| .
\end{aligned}$$

This inequality yields the desired result.

The conclusion of Theorem 1 becomes more transparent if we define a "derivative" of the unbounded operator $B(p)$ in the following way.

DEFINITION 1. Let $\tilde{B}(p):\mathcal{D}(\tilde{B}(p))\subset L^1(0,T;X)\to L^1(0,T;X)$ be defined by

$$\mathcal{D}(\tilde{B}(p)) = \{w\in L^1(0,T;D): B(p)(w(\cdot))\in L^1(0,T;X)\}$$

and $(\tilde{B}(p)w)(t)=B(p)(w(t))$, a.e. $t\in(0,T)$. We then define

$$B'(p_0):\mathcal{D}(B'(p_0))\subset L^1(0,T;X)\to \mathcal{B}(\mathcal{P};L^1(0,T;X))$$

by

$$\begin{aligned}
\mathcal{D}(B'(p_0)) = \{ & w\in L^1(0,T;D): w\in\mathcal{D}(\tilde{B}(p_0+h)) \text{ for } |h| \text{ sufficiently} \\
& \text{small and } \tilde{B}(p)w \text{ is differentiable with respect to } p \text{ at } p_0\}
\end{aligned}$$

and

$$B'(p_0)w = D_p\tilde{B}(p_0)w.$$

COROLLARY 2. *Suppose the hypotheses of Theorem* 1 *hold with* $x_0\in D$. *Then* $S(\cdot;p_0)x_0\in\mathcal{D}(B'(p_0))$ *and*

$$D_pS(t;p_0)x_0 = \int_0^t S(t-s;p_0)\left[B'(p_0)(S(\cdot;p_0)x_0)\right](s)\,ds, \qquad 0\le t\le T.$$

Proof. Since $x_0\in D$, $S(\cdot;p)x_0\in\mathcal{D}(\tilde{B}(p))$ for $p\in P$ by (H4). Furthermore, by definition,

$$\left[\tilde{B}(p)(S(\cdot;p_0)x_0)\right](t) = B(p)S(t;p_0)x_0 = (F(p)x_0)(t).$$

Therefore, by (H6), $S(\cdot;p_0)x_0\in\mathcal{D}(B'(p_0))$ and

$$B'(p_0)(S(\cdot;p_0)x_0) = D_pF(p_0)x_0.$$

The result now follows from Theorem 1.

3. The inhomogeneous equation. Suppose $u\in L^1(0;T;X)$; then the solution of the inhomogeneous equation (3), whenever it exists, is given by

$$x(t) = \int_0^t S(t-s;p)u(s)\,ds.$$

We will now consider differentiability with respect to p of expressions of this form.

THEOREM 3. *Suppose hypotheses* (H1)–(H5) *hold. Fix* $u\in L^1(0,T;X)$ *and define*

$$Q(t;p)=\int_0^t S(t-s;p)u(s)\,ds,\qquad 0\le t\le T,\quad p\in P.$$

Also define a mapping $G(p)$ *on* $L^1(0,T;D)$ *by*

$$[G(p)w](t)=\int_0^t B(p)S(t-s;p_0)w(s)\,ds,\qquad 0\le t\le T,\quad p\in P.$$

Then $G(p)$ *may be extended to a bounded linear mapping on* $L^1(0,T;X)$. *Suppose*

(H7) $G(p)u\in L^1(0,T;Y)$, $p\in P$ *and*

(H8) $G(p)u$ *is differentiable with respect to* p *at* p_0.

Then $Q(t;p)$ *is differentiable at* p_0 *and*

$$D_pQ(t;p_0)=\int_0^t S(t-s;p_0)\big[D_pG(p_0)u\big](s)\,ds.$$

Proof. First we show that $G(p)$ may be extended to a bounded linear operator on $L^1(0,T;X)$. If $p\in P$ and $w\in L^1(0,T;D)$, then

$$\begin{aligned}|||G(p)w|||&\le\int_0^T\int_0^t\|B(p)S(t-s;p_0)w(s)\|\,ds\,dt\\&=\int_0^T\int_s^T\|B(p)S(t-s;p_0)w(s)\|\,dt\,ds\\&=\int_0^T\int_0^{T-s}\|B(p)S(t;p_0)w(s)\|\,dt\,ds\\&\le\int_0^T\int_0^T\|B(p)S(t;p_0)w(s)\|\,dt\,ds\\&\le\int_0^T K\|w(s)\|\,ds=K|||w|||,\end{aligned}$$

by hypothesis (H4). Therefore $G(p)$ is bounded on $L^1(0,T;D)$ and so has a unique bounded extension to $L^1(0,T;X)$.

Suppose $w\in L^1(0,T;D)$. $p_0+h\in P$; then by (4),

$$\begin{aligned}&\int_0^t\big[S(t-s;p_0+h)-S(t-s;p_0)\big]w(s)\,ds\\&\quad=\int_0^t\int_0^{t-s}S(t-s-\tau;p_0+h)\big[B(p_0+h)-B(p_0)\big]S(\tau;p_0)w(s)\,d\tau\,ds\\&\quad=\int_0^t\int_0^{t-s}S(\tau;p_0+h)\big[B(p_0+h)-B(p_0)\big]S(t-s-\tau;p_0)w(s)\,d\tau\,ds\end{aligned}$$

by Fubini's theorem. Therefore

$$\begin{aligned}&\int_0^t S(t-s;p_0+h)w(s)\,ds-\int_0^t S(t-s;p_0)w(s)\,ds\\&\quad=\int_0^t S(t-s;p_0+h)\big[G(p_0+h)w\big](s)\,ds-\int_0^t S(t-s;p_0+h)\big[G(p_0)w\big]\,ds\end{aligned}\tag{9}$$

for $0\le t\le T$, $w\in L^1(0,T;D)$. It is easy to see that both sides of (9) are bounded linear operators from $L^1(0,T;X)$ into X. Therefore (9) must hold for all $w\in L^1(0,T;X)$. In

particular, for $w=u$ we obtain

$$Q(t;p_0+h)-Q(t;p_0)=\int_0^t S(t-s;p_0+h)\big[G(p_0+h)u-G(p_0)u\big](s)\,ds, \tag{10}$$
$$0\le t\le T,\quad p_0, p_0+h\in P.$$

Let $\varepsilon>0$ be given; then there is a $\delta>0$ such that

$$\big|\big|\big|G(p_0+h)u-G(p_0)u-\big[D_pG(p_0)u\big]\big|\big|\big|\le\varepsilon|h|$$

whenever $|h|\le\delta$. In particular, if M_2 is the norm of $D_pG(p_0)u$, then

$$|||G(p_0+h)-G(p_0)u|||\le|h|(\varepsilon+M_2),\qquad |h|\le\delta.$$

Therefore, by (10), (H5), (H7) and (7),

$$\begin{aligned}
&\big\|Q(t;p_0+h)-Q(t;p_0)-\int_0^t S(t-s;p_0)\big[(D_pG(p_0)u)h\big](s)\,ds\big\|\\
&\le\int_0^t\big\|\big[S(t-s;p_0+h)-S(t-s;p_0)\big]\big[(G(p_0+h)u)(s)-(G(p_0)u)(s)\big]\big\|\,ds\\
&\quad+\int_0^t\|S(t-s;p_0)\|\,\big\|(G(p_0+h)u)(s)-(G(p_0)u)(s)-\big[D_pG((p_0)u)h\big](s)\big\|\,ds\\
&\le\varepsilon C\int_0^t\|(G(p_0+h)u)(s)-(G(p_0)u)(s)\|\,ds\\
&\quad+C\big|\big|\big|G(p_0+h)u-G(p_0)u-\big[D_pG(p_0)u\big]h\big|\big|\big|\\
&\le\varepsilon C(\varepsilon+M_2)|h|+\varepsilon C|h|
\end{aligned}$$

for $|h|$ sufficiently small. This estimate yields the desired result.

4. Sensitivity equations. We now consider the question of whether for fixed $h\in\mathcal{P}$, the derivatives $[D_pS(t;p_0)x_0]h$ and $[D_pQ(t;p_0)]h$ are solutions of abstract evolution equations. According to Theorems 1 and 3, both are of the form $\int_0^t S(t-s;p_0)f(s)\,ds$, where f is $[D_pF(p_0)x_0]h$ and $[D_pG(p_0)u]h$, respectively. A function of the form $v(t)=\int_0^t S(t-s)f(s)\,ds$, where S is a C_0-semigroup generated by an operator A, is sometimes called a weak solution of the abstract Cauchy problem

$$v'(t)=Av(t)+f(t),\qquad v(0)=0. \tag{11}$$

Note that v is well defined on $[0,T]$ if $f\in L^1(0,T;X)$. Unfortunately, there exist $f\in C([0,T];X)$ for which v can be nowhere differentiable, e.g., $f(t)=T(t)x$, $x\notin D(A)$ works if A generates a (C_0) group. If $f\in C^1([0,T];X)$, then v is strongly differentiable and satisfies (11). (See [9, Thm. 1.19, p. 488] or [12, Lem. 6.1, p. 215].) If S is holomorphic [9, Thm. 1.27, p. 493], or if S satisfies $S(t)X\subset D(A)$, $t>0$ [5], [13], then this hypothesis on f may be weakened. In [1] it is shown that v satisfies (11) in a dual space setting for quite general f.

In the present context we wish to show that v satisfies (11) a.e. without assuming $f\in C^1$ or $S(t)X\subset D(A)$, $t>0$. Instead we will take advantage of the special structure of S and f already employed in §§2 and 3. In [10, Lemma 6.2, p. 136] it is shown that if $f\in C([0,T];X)$ and $v(t)\in D(A)$, then the right derivative of v exists at t and satisfies (11). The following theorem is an extension of this idea.

THEOREM 4. *Suppose* $f\in L^1(0,T;X)$ *and* A *generates a* C_0*-semigroup* $S(t)$ *on the Banach space* X. *Let* $v(t)=\int_0^t S(t-s)f(s)\,ds$, $0\le t\le T$. *Suppose* $v(t)\in D(A)$ *a.e. and* $Av\in L^1(0,T;X)$. *Then* v *is differentiable a.e. and satisfies*

$$v'(t)=Av(t)+f(t) \quad a.e.\ t\in(0,T), \qquad v(0)=0. \tag{12}$$

The proof of Theorem 4 requires several preliminary lemmas. The first is a vector-valued version of a result due to Titchmarsh. (See [4, Thm. 2.1.6, p. 92].)

LEMMA 5. *Suppose* $z\in L^1_{\mathrm{loc}}(\mathbb{R};X)$ *and*

$$\lim_{h\to 0^+}\frac{1}{h}\int_a^b \|z(x+h)-z(x)\|\,dx=0 \tag{13}$$

for every bounded interval (a,b). *Then there is a constant* c *such that* $z(x)=c$ *almost everywhere.*

Proof. Let $F_n(x)=n\int_0^{1/n} z(t+x)\,dt$, $n=1,2,3,\cdots$. Then F_n is continuous and (13) implies that F_n has a right derivative which is zero. (See [11, Thm. 3.2, p. 170] for a similar argument). It is well known that any continuous function with a continuous right derivative is differentiable. Therefore $F_n'=0$. By [11, Thm. 2.1, p. 166], each F_n is constant. Since

$$\int_a^b \|F_n(x)-z(x)\|\,dx\to 0 \quad \text{as } n\to\infty$$

for every bounded interval (a,b), it follows that z is constant a.e.

The proof of Theorem 4 depends on a somewhat stronger notion of differentiability than differentiability a.e.

DEFINITION 2. Suppose $f:\mathbb{R}\to X$. Then f has a *local derivative* g if

$$\lim_{h\to 0}\int_a^b \left\|\frac{1}{h}[f(x+h)-f(x)]-g(x)\right\|dx=0$$

for every bounded interval (a,b). The right local derivative of f is defined similarly with $h\to 0$ replaced by $h\to 0^+$.

LEMMA 6. *Suppose* $f\in L^1_{\mathrm{loc}}(\mathbb{R};X)$ *has a right local derivative* $g\in L^1_{\mathrm{loc}}(\mathbb{R};X)$. *Then* g *is the local derivative of* f.

Proof. Let $w(x)=\int_0^x g(s)\,ds$ for $x\in\mathbb{R}$. By [11, Thm. 3,3, p. 171], w is locally differentiable with local derivative g. It remains to show that $f-w$ is constant a.e. Let $z=f-w$. Then

$$\int_a^b \left\|\frac{1}{h}[z(x+h)-z(x)]\right\|dx$$
$$\le\int_a^b \left\|\frac{1}{h}[f(x+h)-f(x)]-g(x)\right\|dx+\int_a^b \left\|\frac{1}{h}[w(x+h)-w(x)]-g(x)\right\|dx.$$

Since this expression goes to zero as $h\to 0^+$ for every bounded interval (a,b), the result follows from Lemma 5.

Proof of Theorem 4. Note that v is continuous on [0,T] by the strong continuity of S and the dominated convergence theorem. Extend v to $\mathbb{R}$ by setting $v(t)=v(0)$ for $t<0$ and $v(t)=v(T)$ for $t>T$. We will show that v has a right local derivative g for which $g(t)=Av(t)+f(t)$ for a.e. $t\in(0,T)$. The conclusion of Theorem 4 then follows

from Lemma 6 and the fact that continuity and local differentiability imply differentiability a.e. (See [11, Thm. 3.4, p. 171].)

If $t, t+h \in [0,T]$ and $h>0$, then

$$\begin{aligned}
&\frac{1}{h}[v(t+h)-v(t)] \\
&\quad=\frac{1}{h}\int_t^{t+h} S(t+h-s)f(s)\,ds+\frac{1}{h}\int_0^t S(t+h-s)f(s)ds-\frac{1}{h}\int_0^t S(t-s)f(s)\,ds \\
&\quad=\frac{1}{h}\int_t^{t+h} S(t+h-s)f(s)\,ds+\frac{1}{h}[S(h)-I]v(t) \\
&\quad=\frac{1}{h}\int_t^{t+h} S(t+h-s)[f(s)-f(t)]ds+\frac{1}{h}\int_t^{t+h}[S(t+h-s)-I]f(t)\,ds \\
&\qquad+f(t)+\frac{1}{h}[S(h)-I]v(t) \\
&\quad=\frac{1}{h}\int_0^h S(h-s)[f(s+t)-f(t)]\,ds+\frac{1}{h}\int_0^h[S(s)-I]f(t)\,ds \\
&\qquad+f(t)+\frac{1}{h}[S(h)-I]v(t).
\end{aligned}$$

Choose a constant C such that $\|S(t)\| \le C$ for $0 \le t \le T$. Then

$$\begin{aligned}
&\left\|\frac{1}{h}[v(t+h)-v(t)]-Av(t)-f(t)\right\| \\
&\qquad\le\frac{1}{h}C\int_0^h\|f(s+t)-f(t)\|\,ds+\max_{0\le s\le h}\|[S(s)-I]f(t)\| \\
&\qquad\quad+\left\|\frac{1}{h}[S(h)-I]v(t)-Av(t)\right\| \quad \text{a.e. } t\in(0,T).
\end{aligned}$$

Therefore, if $h>0$,

$$\begin{aligned}
&\int_0^{T-h}\left\|\frac{1}{h}[v(t+h)-v(t)]-Av(t)-f(t)\right\|dt \\
&\qquad\le\frac{1}{h}C\int_0^{T-h}\int_0^h\|f(s+t)-f(t)\|ds\,dt+\int_0^T\max_{0\le s\le h}\|[S(s)-I]f(t)\|dt \\
&\qquad\quad+\int_0^T\left\|\frac{1}{h}[S(h)-I]v(t)-Av(t)\right\|dt.
\end{aligned} \tag{14}$$

The first term on the right is bounded by

$$C\int_0^{T-h}\int_0^1\|f(t+sh)-f(t)\|\,ds\,dt=C\int_0^1\int_0^{T-h}\|f(t+sh)-f(t)\|\,dt\,ds,$$

where we extend f as zero outside $(0,T)$ if necessary. Since the inner integral goes to zero uniformly in s as $h\to 0^+$ by standard arguments, this term goes to zero as $h\to 0^+$.

The second and third terms on the right of (14) go to zero as $h\to 0^+$ since their integrands go to zero pointwise a.e. and are dominated by $(C+1)\|f(t)\|$ and $(C+1)\|Av(t)\|$, respectively. Finally, by definition of v,

$$\begin{aligned}\int_{T-h}^{T}\left\|\frac{1}{h}[v(t+h)-v(t)]-Av(t)-f(t)\right\|dt&\\ \le\int_{T-h}^{T}\left\|\frac{1}{h}[v(T)-v(t)]\right\|dt+\int_{T-h}^{T}\|Av(t)+f(t)\|\,dt&\\ \le\max_{T-h\le t\le T}\|v(T)-v(t)\|+\int_{T-h}^{T}\|Av(t)+f(t)\|\,dt,&\end{aligned}$$

which goes to zero as $h\to 0^+$ since v is continuous on $\mathbb{R}$ (as extended above) and $Av+f$ is integrable on $(0,T)$. These estimates yield the desired conclusion.

The following lemma gives conditions under which v is a solution of (12) in the special case when the generator $A(p_0)$ satisfies the conditions of Theorem 1.

LEMMA 7. *Suppose* (H1)–(H5) *hold. Let* $R(\lambda)$ *denote the resolvent of* $A(p_0)$, *which exists as a bounded operator on* X *for* $\lambda>\omega$. *Let* $D_0=\{\lambda R(\lambda)y : y\in Y,\ \lambda>\omega\}$. *In the same spirit as* (H4) *we assume*:

(H9) *There is a constant* $K_0>0$ *such that*

$$\int_0^T\|A(p_0)S(t;p_0)x\|\,dt\le K_0\|x\|,\qquad x\in D_0.$$

Then $v(t)=\int_0^t S(t-s;p_0)f(s)\,ds$ *is a solution of*

$$v'(t)=A(p_0)v(t)+f(t),\qquad v(0)=0 \tag{15}$$

in the sense of Theorem 4, *for every* $f\in L^1(0,T;Y)$.

Note that

$$\left\|\int_0^T A(p_0)S(t;p_0)x\,dt\right\|=\|S(T;p_0)x-x\|\le K_0\|x\|\quad\text{for } x\in D.$$

Hypothesis (H9) demands that this inequality hold with the norms brought inside the integral for all $x\in D_0$. The application given in §5 has the property that (H9) does not hold for all $x\in D$.

Proof. According to Theorem 4, it is sufficient to show that $v(t)\in D$ a.e. and $A(p_0)v\in L^1(0,T;X)$. Fix $f\in L^1(0,T;Y)$ and let $f_\lambda(t)=\lambda R(\lambda)f(t)$, $\lambda>\omega$. Then $f_\lambda\in L^1(0,T;D_0)$. Define a mapping $[Jw](t)=\int_0^t A(p_0)S(t-s;p_0)w(s)\,ds$, $0\le t\le T$, $w\in L^1(0,T;D_0)$. Using (H9) one can show, just as in the proof of Theorem 3 that $J: L^1(0,T;D_0)\to L^1(0,T;X)$ is bounded on $L^1(0,T;D_0)$. Therefore J may be extended to a bounded linear operator on the smallest closed subspace of $L^1(0,T;X)$ which contains $L^1(0,T;D_0)$.

Let $v_\lambda(t)=\int_0^t S(t-s;p_0)f_\lambda(s)\,ds$. By the dominated convergence theorem,

$$|||f_\lambda-f|||\to 0\quad\text{as }\lambda\to\infty$$

and therefore

$$v_\lambda(t)\to v(t)\quad\text{in } X\quad\text{for } 0\le t\le T. \tag{16}$$

Since $S(t-s;p_0)f_\lambda(s)\in L^1(0,T;D)$ for each $t\in[0,T]$ and $A(p_0)$ is closed, we have

$$A(p_0)v_\lambda(t)=\int_0^t A(p_0)S(t-s;p_0)f_\lambda(s)\,ds=[J(f_\lambda)](t).$$

Since $f_\lambda \in L^1(0,T;D_0)$ and J is bounded on $L^1(0,T;\overline{D_0})$, we have

$$|||J(f_\lambda)-J(f)|||\to 0 \quad \text{as } \lambda\to\infty.$$

Therefore, there is a subsequence $\lambda_k\to\infty$ such that

$$A(p_0)v_{\lambda_k}(t)=\left[J(f_{\lambda_k})\right](t)\to J(f)(t) \quad \text{in } X \tag{17}$$

almost everywhere on $(0,T)$. Since $A(p_0)$ is closed, (16) and (17) together imply that $v(t)\in D$ a.e. $t\in(0,T)$ and $A(p_0)v=J(f)\in L^1(0,T;X)$. This completes the proof.

We are now in a position to prove the motivating result of this section.

THEOREM 8. *Suppose* (H1)–(H9) *hold. Fix* $h\in\mathcal{P}$ *and let* $v_1(t)=[D_pS(t;p_0)x_0]h$, $v_2(t)=[D_pQ(t;p_0)]h$, $f_1=[D_pF(p_0)x_0]h$ *and* $f_2=[D_pG(p_0)u]h$. *Then* v_1, v_2 *are differentiable almost everywhere on* $(0,T)$ *and satisfy*

$$\begin{aligned} &v_i'(t)=A(p_0)v_i(t)+f_i(t) \quad \text{a.e. } t\in(0,T),\\ &v_i(0)=0 \end{aligned}$$

for $i=1,2$.

Proof. By (H5) and (H7), f_1, $f_2\in L^1(0,T;Y)$. The result now follows immediately from Lemma 7.

5. An application. In this section we consider the differentiability with respect to $p=(p_1,p_2,\cdots,p_n)$ of the solution of the delay differential equation

$$\begin{aligned} &x'(t)=a_0x(t)+\sum_{k=1}^{n}a_kx(t-p_k)+u(t), \qquad t>0,\\ &x(0)=\eta, \qquad x_0=\varphi, \end{aligned} \tag{18}$$

where $p^*>\max\{p_k\}>0$ is fixed, $\eta\in\mathbb{R}$, $a_k\in\mathbb{R}$, $k=0,1,2,\cdots,n$, $\varphi\in L^1(-p^*,0)$ with norm denoted by $\|\varphi\|_1$, $x_t(s)=x(t+s)$, $t\geq 0$, $-p^*\leq s\leq 0$ and $u\in L^1_{\text{loc}}(0,\infty)$. By a solution $x(t;p)$ of (18) we mean a function $x:(-p^*,+\infty)\to\mathbb{R}$ such that x is absolutely continuous on compact subsets of $[0,\infty)$ and satisfies the first line of (18) a.e. on $(0,\infty)$ and such that $x(0)=\eta$ and $x(t)=\varphi(t)$ a.e. on $(-p^*,0)$. It can be shown that (18) has a unique solution $x(t;p)$ for every pair of initial data $(\eta,\varphi)\in\mathbb{R}\times L^1(-p^*,0)$. This problem can be placed in a semigroup setting in a standard way. Define a Banach space $X=\mathbb{R}\times L^1(-p^*,0)$ with norm $\|(\eta,\varphi)\|=|\eta|+\|\varphi\|_1$. Let $\mathcal{P}=\mathbb{R}^n$ and let P be the cross product of the interval $(-p^*,0)$ with itself n times. For $p=(p_1,p_2,\cdots,p_n)\in P$ define the operator $A(p)$ in X by

$$D(A(p))=\{(\eta,\varphi)\in X:\varphi \text{ is a.c., } \varphi'\in L^1(-p^*,0),\ \eta=\varphi(0)\},$$

$$A(p)(\eta,\varphi)=\left(a_0\varphi(0)+\sum_{k=1}^{n}a_k\varphi(-p_k),\varphi'\right).$$

Then it is well known that $A(p)$ generates a semigroup $S(t;p)$ satisfying

$$S(t;p)(\eta,\varphi)=(y(t),y_t), \qquad t\geq 0, \quad (\eta,\varphi)\in X, \tag{19}$$

where y is the solution of

$$\begin{aligned} &y'(t)=a_0y(t)+\sum_{k=1}^{n}a_ky(t-p_k), \qquad t>0,\\ &(y(0),y_0)=(\eta,\varphi). \end{aligned} \tag{20}$$

It may be shown by standard results that $A(p)$ and $S(t;p)$ satisfy (H1)–(H3). The fact that ω may be chosen independent of p for a fixed p^* is verified by the estimates made in [2]. Let $D=D(A(p))$ as above; then $A(p)=A+B(p)$, where

$$B(p)(\eta,\varphi)=\left(\sum_{k=1}^{n} a_k\varphi(-p_k),0\right), \qquad (\eta,\varphi)\in D,$$
$$A(\eta,\varphi)=(a_0\varphi(0),\varphi'), \qquad (\eta,\varphi)\in D.$$

We now consider whether (H4) holds. Fix any $T>0$. Note that by (H3) and (19) there is a constant $C>0$ such that $|y(t)|\le C\|(\eta,\varphi)\|$ for $0\le t\le T$, where y satisfies (20). Fix $p_0=(r_1,r_2,\cdots,r_n)\in P$ and let $y(t)=y(t;p_0)$ satisfy (20) with $p=p_0$. Let $A_M=\max|a_k|$. Then

$$\begin{aligned}\int_0^T \|B(p)S(t;p_0)(\eta,\varphi)\|\,dt &= \int_0^T \|B(p)(y(t),y_t)\|\,dt\\ &= \int_0^T \left|\sum_{k=1}^{n} a_k y(t-p_k)\right| dt\\ &\le \sum_{k=1}^{n} |a_k| \int_0^T |y(t-p_k)|\,dt\\ &\le A_m \sum_{k=1}^{n} \int_{-p_k}^{0} |\varphi(t)|\,dt + A_m \sum_{k=1}^{n} \int_0^T |y(t)|\,dt\\ &\le A_M n(\|\varphi\|_1 + CT\|(\eta,\varphi)\|) \le A_M n(CT+1)\|(\eta,\varphi)\|\end{aligned}$$

for $(\eta,\varphi)\in D$. This verifies (H4). Note that using the notation of Theorem 1 we have

$$F(p)(\eta,\varphi)=\left(\sum_{k=1}^{n} a_k y(t-p_k),0\right) \tag{21}$$

for $(\eta,\varphi)\in X$, where $y=y(t;p_0)$.

Note that the mapping

$$(\eta,\varphi)\to\left(\sum_{k=1}^{n} a_k y(t-p_k),0\right)$$

is not bounded as a mapping from X into X for $0<t<\max\{p_k\}$. Therefore (H4)* does *not* hold for this example.

We will now verify (H5). Let $Y=\mathbb{R}\times\{0\}$. Then $F(p)x\in L^1(0,T;Y)$ for $x\in X$ and $p\in P$. Furthermore, if $h=(h_1,h_2,\cdots,h_n)$, then

$$|||F(p_0+h)(\eta,0)-F(p_0)(\eta,0)|||=\int_0^T \left|\sum_{k=1}^{n} a_k(z(t-r_k-h_k)-z(t-r_k))\right| dt,$$

where z satisfies

$$\begin{aligned} &z'(t)=a_0z(t)+\sum_{k=1}^{n} a_k z(t-r_k), \qquad t>0,\\ &(z(0),z_0)=(\eta,0).\end{aligned} \tag{22}$$

Since $|z(t)| \leq C|\eta|$ for $0 \leq t \leq T$, (22) implies that there is a constant C_1 such that $|z'(t)| \leq C_1|\eta|$ for a.e. $t \in (0,T)$. Fix k and assume $h_k > 0$. Then

$$\begin{aligned}
&\int_0^T |z(t-r_k-h_k)-z(t-r_k)|\,dt \\
&\quad \leq \int_{-r_k}^{0} |z(t-h_k)-z(t)|\,dt + \int_0^{h_k} |z(t-h_k)-z(t)|\,dt + \int_{h_k}^{T} |z(t-h_k)-z(t)|\,dt \\
&\quad \leq \int_0^{h_k} |z(t)|\,dt + \int_0^T |z(t+h_k)-z(t)|\,dt \\
&\quad \leq h_k C|\eta| + h_k T C_1|\eta| \\
&\quad = |h_k|\,|\eta|(C+TC_1).
\end{aligned}$$

The case $h_k < 0$ is similar. Therefore,

$$|||F(p_0+h)(\eta,0)-F(p_0)(\eta,0)||| \leq C_2|h|\,\|(\eta,0)\|,$$

where C_2 is a constant and $|h|$ denotes $\Sigma_{k=1}^n |h_k|$. This inequality verifies (H5).

Note that (H7) is immediately verified by the definition of G and the subspace Y for any $u \in L^1(0,T;X)$.

We next consider (H9) which was needed in the proof of Lemma 7. Calculation shows that every element of D_0 is of the form $(\eta, \eta e^{\lambda s})$, where η is a constant depending on λ and $-p^* \leq s \leq 0$. If y satisfies (20) with $p = p_0$ and $(y(0), y_0) = (\eta, \eta e^{\lambda s})$, then

$$\begin{aligned}
&\int_0^T \|A(p)S(t;p_0)(\eta,\eta e^{\lambda s})\|\,dt \\
&\quad = \int_0^T \Big|a_0 y(t) + \sum_{k=1}^n a_k y(t-p_k)\Big|\,dt + \int_0^T \int_{-p^*}^0 |y'(t+s)|\,ds\,dt \\
&\quad \leq \int_0^T \Big|a_0 y(t) + \sum_{k=1}^n a_k y(t-p_k)\Big|\,dt + \int_0^{p^*} \int_{t-p^*}^0 |\eta\lambda e^{\lambda s}|\,ds\,dt \\
&\qquad + \int_0^{p^*} \int_0^t |y'(s)|\,ds\,dt + \int_{p^*}^T \int_{t-p}^t |y'(s)|\,ds\,dt.
\end{aligned}$$

The first, third and fourth terms are easily estimated using (18) and the techniques used to verify (H4). Direct integration shows that the second term is bounded by $p^*|\eta| \leq p^*\|(\eta, \eta e^{\lambda s})\|$. Therefore (H9) holds. Note that (H9) does *not* hold for all $x \in D$. This proof depends on the monotonicity of the functions which appear as second coordinates of elements of D_0.

Suppose $x_0 \in D$. Then it is not difficult to show, using [4, Thm. 2.17, p. 93], for example, that $F(p)x_0$ is differentiable with respect to p at $p = p_0$ in the sense of Definition 1 and that

$$\big[D_p F(p_0)x_0\big]h = \Big(-\sum_{k=1}^n a_k y'(t-r_k)h_k, 0\Big),$$

where y satisfies

$$\begin{aligned}
&y'(t) = a_0 y(t) + \sum_{k=1}^n a_k y(t-r_k), \qquad t > 0, \\
&(y(0), y_0) = x_0.
\end{aligned} \tag{23}$$

Therefore, by Theorem 1, $S(t;p)x_0$ is differentiable with respect to p at $p=p_0$ for $0\le t\le T$ and $x_0\in D$ and satisfies

$$\left[D_pS(t;p_0)x_0\right]h=\int_0^t S(t-s;p_0)\left(-\sum_{k=1}^n a_k y'(s-r_k)h_k,0\right)ds.$$

Furthermore, Theorem 8 shows in this context that $[D_pS(t;p_0)x_0]h=(v(t),v_t)$, where v satisfies

$$v'(t)=a_0v(t)+\sum_{k=1}^n a_kv(t-r_k)-\sum_{k=1}^n a_ky'(t-r_k)h_k,\qquad t>0,$$
$$(v(0),v_0)=(0,0).$$

Suppose $u\in L^1(0,T;D)$. Then since $B(p)$ is a difference of closed operators we have by definition that

$$G(p)u=B(p)\int_0^t S(t-s;p_0)u(s)\,ds.$$

Therefore

$$G(p)u=\left(\sum_{k=1}^n a_kx(t-p_k),0\right),\tag{24}$$

where x is the solution of

$$\begin{aligned}&x'(t)=a_0x(t)+\sum_{k=1}^n a_kx(t-r_k)+u(t),\qquad t>0,\\&(x(0),x_0)=(0,0).\end{aligned}\tag{25}$$

Since $G(p)$ is continuous on $L^1(0,T;X)$ and the solution of (25) depends continuously on u, we have that (24) holds for all $u\in L^1(0,T;X)$. Since solutions of (25) are absolutely continuous on $(-p^*,T)$, we have as above that $G(p)u$ is differentiable at $p=p_0$ in the sense of Fréchet and

$$\left[D_pG(p_0)u\right]h=\left(-\sum_{k=1}^n a_kx'(t-r_k)h_k,0\right),$$

where x satisfies (25).

Therefore, by Theorem 3, $Q(t;p)=\int_0^t S(t-s;p)u(s)ds$ is differentiable with respect to p at $p=p_0$ for every $0\le t\le T$, $u\in L^1(0,T;X)$. Furthermore,

$$D_pQ(t;p_0)h=\int_0^t S(t-s;p_0)\left(-\sum_{k=1}^n a_kx'(s-r_k)h_k,0\right)ds.$$

Theorem 8 shows that $D_pQ(t;p_0)h=(g(t),g_t)$, where g is the solution of

$$g'(t)=a_0g(t)+\sum_{k=1}^n a_kg(t-r_k)-\sum_{k=1}^n a_kx'(t-r_k)h_k,\qquad t>0,$$
$$(g(0),g_0)=(0,0).$$

This application may be generalized in various ways. For example, no essential difficulties arise if $\mathbb{R}$ is replaced by $\mathbb{R}^m$ or if $L^1(-p^*,0)$ is replaced by $L^p(-p^*,0)$, $p\ge1$. These methods are also applicable to more general forms of functional differential

equations. For example, in [3] a semigroup theory is developed for a class of linear retarded functional differential equations. Hypothesis (ii) of [3] yields hypothesis (H4) of this paper. This and other applications of the results presented here will be the subject of further investigation.

Acknowledgment. The author gratefully acknowledges the contribution to this work, through many helpful discussions, of Professor John A. Burns.

REFERENCES

[1] J. M. BALL, *Strongly continuous semigroups, weak solutions and the variation of constants formula*, Proc. Amer. Math. Soc. 63 (1977), pp. 370–373.

[2] J. A. BURNS AND E. M. CLIFF, *An abstract quasi-linearization algorithm for estimating parameters in hereditary systems*, IEEE Trans. Automat. Contr., 25 (1980), pp. 126–129.

[3] J. A. BURNS AND T. L. HERDMAN, *Adjoint semigroup theory for a class of functional differential equations*, this Journal, 7 (1976), pp. 729–745.

[4] P. L. BUTZER AND H. BERENS, *Semi-groups of Operators and Approximation*, Springer-Verlag, New York, 1967.

[5] M. G. CRANDALL AND A. PAZY, *On differentiability of weak solutions of a differential equation in Banach space*, J. Math. Mech., 18 (1969), pp. 1007–1016.

[6] J. S. GIBSON AND L. G. CLARK, *Sensitivity analysis for a class of evolution equations*, J. Math. Anal. Appl., 58 (1977), pp. 22–31.

[7] J. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[8] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, AMS Colloquium Publications, 31, American Mathematical Society, Providence, R.I., 1957.

[9] T. KATO, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 1976.

[10] S. G. KREIN, *Linear Differential Equations in Banach Space*, Nauka, Moscow, 1967; Eng. transl., Transl. Math. Monographs, vol. 29, American Mathematical Society, Providence, R.I., 1972.

[11] J. MIKUSINSKI, *The Bochner Integral*, Birkhauser, Basel and Stuttgart, 1978.

[12] R. S. PHILLIPS, *Perturbation theory for semi-groups of linear operators*, Trans. Amer. Math. Soc., 74 (1953), pp. 199–221.

[13] G. F. WEBB, *Regularity of solutions to an abstract inhomogeneous linear differential equation*, Proc. Amer. Math. Soc., 62 (1977), pp. 271–277.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0007 $01.00/0

THE TWO-DIMENSIONAL EIGENVALUE RANGE AND EXTREMAL EIGENVALUE PROBLEMS*

B. E. WILLNER† AND T. J. MAHAR‡

Abstract. If $a(x)$ and $b(x)$ are measurable functions on $[0, 1]$ satisfying $0<a(x)<b(x)$, let $C(a(x), b(x))$ denote the class of functions $\varphi(x)\in L_1[0, 1]$ such that $a(x)\le\varphi(x)\le b(x)$. Given such a $\varphi(x)$, let $\lambda_1(\varphi)$ and $\lambda_2(\varphi)$ denote the first two eigenvalues of the problem $y''+\lambda\varphi(x)y=0$, $y(0)=y(1)=0$. In this paper we characterize the set $(\lambda_1(\varphi),\lambda_2(\varphi))$ as φ varies over $C(a(x),b(x))$. Explicit analytic and numerical results are given when $a(x)$ and $b(x)$ are constant. The connection between this problem and the extremization of functions of eigenvalues is also discussed.

1. Introduction.

Consider the eigenvalue problem

$$y''+\lambda\varphi(x)y=0, \qquad y(0)=y(1)=0 \tag{1.1}$$

with $\varphi(x)>0$. The eigenvalues of this problem form an unbounded sequence $\lambda_j(\varphi)>0$. If $f(x_1,x_2)$ is a real, differentiable function for $x_1,x_2>0$, define the functional $F(\varphi)$ by

$$F(\varphi)=f(\lambda_1(\varphi),\lambda_2(\varphi)). \tag{1.2}$$

Given measurable functions $a(x)$ and $b(x)$ such that $0<a(x)<b(x)$ on $[0, 1]$, let

$$C(a(x),b(x))=\{\varphi(x)\in L_1[0,1]: a(x)\le\varphi(x)\le b(x)\}. \tag{1.3}$$

The extremal eigenvalue problem for $F(\varphi)$ can be formulated as follows: for what functions φ in $C(a(x),b(x))$ does $F(\varphi)$ achieve an extreme value? Willner and Mahar [1] and Gentry and Banks [2] derive general characterizations of those functions φ in $C(a(x),b(x))$ which extremize functions $f(\lambda_1,\cdots,\lambda_N)$. Willner and Mahar [3] and Keller [4] investigate the problem of extremizing the ratio of the first two eigenvalues when $a(x)$ and $b(x)$ are constant functions and give explicit characterizations of the extremizing functions $\varphi(x)$ as a function of the parameters a and b.

The extremal eigenvalue problem can be reformulated as a geometrical question. Given $a(x)$ and $b(x)$, characterize the set

$$\Lambda=\{(\lambda_1(\varphi),\lambda_2(\varphi)): \varphi\in C(a(x),b(x))\}. \tag{1.4}$$

As discussed in [1] and [2], Λ is compact. The extremal eigenvalue problem for arbitrary f is solved once Λ is known, in the sense that it is then reduced to the calculus problem of extremizing a function of two variables over a compact set.

In this paper we investigate the two-dimensional eigenvalue range Λ. After certain general results are derived for arbitrary $a(x)$ and $b(x)$, we specialize to the case with $a(x)$ and $b(x)$ constant to obtain additional results. The particular case $a(x)\equiv 1$ and $b(x)\equiv 4$ is then studied numerically. A proof that the set Λ corresponding to $C(1,4)$ has been determined is given under an assumption on the correctness of the numerical calculations. A characterization of the interior of Λ is also given. While the extremal eigenvalue problem for N eigenvalues, $N>2$, could be studied in a similar manner, we leave this problem to some future investigation.

The analysis is based on the following observation: if the function $f(x_1,x_2)$ does not achieve an extemum in the interior of Λ, the extremum must occur on the

* Received by the editors December 3, 1980, and in revised form June 10, 1981.

† IBM Thomas J. Watson Research Center, Yorktown Heights, New York 15098.

‡ The Technological Institute, Northwestern University, Evanston, Illinois 60201. This author's research was supported by the National Science Foundation under Grant MCS79-02541.

boundary. The condition that f have no interior extrema is

$$|\overline{\nabla f}|>0 \tag{1.5}$$

in the interior of Λ. As we are interested in determining the boundary of Λ, we will characterize those functions φ in $C(a(x),b(x))$ which extremize functions f satisfying (1.5).

2. Results for arbitrary bounding functions. For $\varphi(x)$ in $C(a(x),b(x))$, let λ_1 and λ_2 denote the first two eigenvalues of (1.1) with corresponding eigenfunctions $y_i(x)$ normalized by

$$\int_0^1 \varphi y_i^2=1, \qquad i=1,2.$$

Given f satisfying (1.5), define the function

$$g(\varphi,f,x)=-\lambda_1\frac{\partial f}{\partial x_1}y_1^2(x)-\lambda_2\frac{\partial f}{\partial x_2}y_2^2(x) \tag{2.1}$$

with the derivatives of f evaluated at (λ_1,λ_2). The results of [1], [2] and [4] include the following theorem.

THEOREM 1. *If $f(\lambda_1(\varphi),\lambda_2(\varphi))$ achieves its maximum over $C(a(x),b(x))$ at $\bar{\varphi}(x)$, then*:

i) $\bar{\varphi}(x)=b(x)$ *when* $g(\bar{\varphi},f,x)>0$,

ii) $\bar{\varphi}(x)=a(x)$ *when* $g(\bar{\varphi},f,x)<0$.

If f achieves its minimum over $C(a(x),b(x))$ at $\underline{\varphi}(x)$, then:

iii) $\underline{\varphi}(x)=a(x)$ *when* $g(\underline{\varphi},f,x)>0$;

iv) $\underline{\varphi}(x)=b(x)$ *when* $g(\underline{\varphi},f,x)<0$.

We will use Theorem 1 and Lemma 2.1, below, to characterize extremizing functions $\varphi(x)$.

LEMMA 2.1. *The function $g(\varphi,f,x)$ has either no zeros, one zero (simple or double), or two simple zeros for $x\in(0,1)$.*

Proof. If $\partial f/\partial x_1$ is zero at (λ_1,λ_2), then g has one double zero in $(0,1)$ since y_2 has one simple zero there . If $\partial f/\partial x_2$ is zero at (λ_1,λ_2), then g has no zeros on $(0,1)$. In the case that neither derivative of f vanishes, we rewrite (2.1) as

$$g(\varphi,f,x)=-\lambda_1\frac{\partial f}{\partial x_1}y_1^2\left[1+\frac{\lambda_2}{\lambda_1}\cdot\frac{\partial f/\partial x_2}{\partial f/\partial x_1}\cdot r^2\right], \tag{2.2}$$

where $r(x)\equiv y_2(x)/y_1(x)$. We take $y_1(x)$ to be positive on $(0,1)$. The function $y_2(x)$ has one simple zero at $\bar{x}\in(0,1)$; we take $y_2>0$ on $(0,\bar{x})$. An elementary argument shows $r(x)$ is decreasing on $(0,1)$. Since $r(\bar{x})=0$, r^2 is decreasing on $(0,x)$ and increasing on $(\bar{x},1)$. Now (2.2) and the monotonicity properties of r^2 show that g can have at most one simple zero in $(0,\bar{x})$ and in $(\bar{x},1)$.

Combining Theorem 1 and Lemma 2.1, we obtain:

THEOREM 2. *The function $\varphi(x)$ which extremizes $f(\lambda_1,\lambda_2)$ has the following forms*:

$$\varphi(x)\equiv a(x) \quad \text{or} \quad \varphi(x)\equiv b(x), \qquad 0\le x\le 1 \tag{2.3a}$$

if $g(\varphi,f,x)$ never vanishes or has a double zero in $(0,1)$;

$$\varphi(x)=\begin{cases} a(x) \quad (b(x)), & 0\le x<x_0,\\ b(x) \quad (a(x)), & x_0<x\le 1\end{cases} \tag{2.3b}$$

if g has a simple zero at $x_0 \in (0,1)$;

$$\varphi(x)=\begin{cases} a(x) \quad (b(x)), & 0\le x<x_0, \\ b(x) \quad (a(x)), & x_0<x<x_1, \\ a(x) \quad (b(x)), & x_1<x\le 1, \end{cases} \tag{2.3c}$$

if g has simple zeros at x_0, x_1 *in* $(0,1)$. *The sign of g at* $x=0$ *and the nature of the extremum determine whether* $\varphi(x)$ *starts as* $a(x)$ *or* $b(x)$.

We now use Theorem 2 to determine relationships between the eigenfunctions of (1.1) when φ extremizes some $f(\lambda_1,\lambda_2)$ over $C(a(x),b(x))$. There needn't be any special relationship in case (2.3a), as it is always possible to extremize some $f(\lambda_1,\lambda_2)$ by $\varphi(x)\equiv a(x)$ or $\varphi(x)\equiv b(x)$. This is called the no-jump case.

The possibilities in (2.3b) are called the one-jump (discontinuity) case. Since r^2 is decreasing on $(0,\bar{x})$ and increasing on $(\bar{x},1)$, it follows that

$$\left(\frac{y_2'(0)}{y_1'(0)}\right)^2 > \left(\frac{y_2(x_0)}{y_1(x_0)}\right)^2 \ge \left(\frac{y_2'(1)}{y_1'(1)}\right)^2 \tag{2.4}$$

if $x_0<\bar{x}$ ($\bar{x}$ is the zero of y_2 in $(0,1)$), while

$$\left(\frac{y_2'(0)}{y_1'(0)}\right)^2 \le \left(\frac{y_2(x_0)}{y_1(x_0)}\right)^2 < \left(\frac{y_2'(1)}{y_1'(1)}\right)^2 \tag{2.5}$$

if $\bar{x}<x_0$. Note that $\bar{x}=x_0$ cannot occur here. The possibilities in (2.3c) are called the two-jump case. We see from (2.2) and the definition of r that

$$\left(\frac{y_2(x_0)}{y_1(x_0)}\right)^2 = \left(\frac{y_2(x_1)}{y_1(x_1)}\right)^2 \tag{2.6}$$

and $x_0<\bar{x}<x_1$ in this case.

3. Constant bounding functions. For the remainder of the paper, we assume $a(x)\equiv a^2$ and $b(x)\equiv b^2$ for some constants a and b. The constants appear as squares for convenience; furthermore, we will denote the first two eigenvalues by λ_1^2 and λ_2^2.

We use Theorem 2 in this special case to further characterize functions φ which extremize some $f(\lambda_1,\lambda_2)$ over $C(a^2,b^2)$. In the no-jump case, the functions $\varphi(x)\equiv a^2$ and $\varphi(x)\equiv b^2$ extremize certain functions f; nothing more can be said here.

If $\varphi(x)\in C(a^2,b^2)$, then $\varphi(1-x)\in C(a^2,b^2)$ and $\lambda_j(\varphi(1-x))=\lambda_j(\varphi(x))$ for all j. Thus, we may assume in the one-jump case that $\varphi(x)$ starts as b^2 and then switches to a^2. The jump point of $\varphi(x)$ is x_0, the location of the simple zero of g in $(0,1)$. We now rule out the possibility $\bar{x}<x_0$, $\bar{x}$ being the zero of y_2 in $(0,1)$, so the relations in (2.4) always hold in the one-jump case. With $\varphi(x)$ as above (i.e., b^2 and then a^2), define

$$I_j(x)=\left(y_j'(x)\right)^2+\lambda_j^2\varphi(x)y_j^2(x), \qquad j=1,2. \tag{3.1}$$

If we assume $\bar{x}<x_0$, then the relations in (2.5) hold. As the I_j are piecewise constant, we have

$$\frac{I_2(x_0^+)}{I_1(x_0^+)}=\left(\frac{y_2'(1)}{y_1'(1)}\right)^2 > \left(\frac{y_2(x_0)}{y_1(x_0)}\right)^2. \tag{3.2}$$

However, using (2.4) and the identity

$$\lambda_2^2(a^2-b^2)y_2^2(x_0)=\left(\frac{y_2(x_0)}{y_1(x_0)}\right)^2\lambda_2^2(a^2-b^2)y_1^2(x_0),$$

we easily verify that we also have

$$\frac{I_2(x_0^+)}{I_1(x_0^+)}<\left(\frac{y_2(x_0)}{y_1(x_0)}\right)^2. \tag{3.3}$$

Since (3.2) and (3.3) are inconsistent, we must have $x_0<\bar{x}$ and the relations in (2.4). Using piecewise trigonometric representations of the $y_j(x)$, the relations in (2.4) become

$$\left(\frac{\sin\lambda_2 bx_0}{\sin\lambda_1 bx_0}\right)^2<\left(\frac{\lambda_2}{\lambda_1}\right)\leq\left(\frac{\sin\lambda_2 a(1-x_0)}{\sin\lambda_1 a(1-x_0)}\right)^2. \tag{3.4}$$

We now consider the two-jump case, so that

$$\varphi(x)=\begin{cases}c^2, & 0\leq x<x_0,\\ d^2, & x_0<x<x_1,\\ c^2, & x_1<x\leq 1,\end{cases} \tag{3.5}$$

where $c=a$ and $d=b$ or $c=b$ and $d=a$. Using piecewise trigonometric representations of the eigenfunctions $y_j(x)$, the extremization condition (2.6) can be rewritten as

$$\frac{1+\left(\frac{c}{d}\right)^2\cot^2\lambda_1 c(1-x_1)}{1+\left(\frac{c}{d}\right)^2\cot^2\lambda_2 c(1-x_1)}=\frac{1+\left(\frac{c}{d}\right)^2\cot^2\lambda_1 cx_0}{1+\left(\frac{c}{d}\right)^2\cot^2\lambda_2 cx_0}. \tag{3.6}$$

Thus, if $\varphi(x)$ given by (3.5) extremizes some f over $C(a^2,b^2)$, we have $x_0<\bar{x}<x_1$ and (3.6) (x_0 and x_1 are the zeros of g on (0.1)). It can be shown that $x_0=1-x_1$ when $c=a$ and $b=d$, so that the extremizing function $\varphi(x)$ is symmetric about $x=\frac{1}{2}$ in this case. As the proof is neither conceptually nor computationally enlightening, we omit the details. The proof consists of a lengthy sequence of elementary, albeit obscure, arguments. Details will be sent upon request.

We now summarize the results of this section. Assume the bounding functions $a(x)$ and $b(x)$ have the constant values a^2 and b^2. If $\varphi(x)\in C(a^2,b^2)$ extremizes some function $f(\lambda_1,\lambda_2)$ satisfying (1.5), one of the following possibilities must occur.

No-jump case.

$$\varphi(x)\equiv a^2 \quad \text{or} \quad \varphi(x)\equiv b^2. \tag{3.7}$$

One-jump case.

$$\varphi(x)=\begin{cases}b^2, & 0\leq x<x_0,\\ a^2, & x_0<x\leq 1,\end{cases} \tag{3.8}$$

$x_0<\bar{x}$, and

$$\frac{\sin\lambda_2 bx_0}{\sin\lambda_1 bx_0}<\frac{\lambda_2}{\lambda_1}\leq\frac{\sin\lambda_2 a(1-x_0)}{\sin\lambda_1 a(1-x_0)}. \tag{3.9}$$

Two-jump case.

$$\varphi(x)=\begin{cases} a^2, & 0\le x<x_0, \\ b^2, & x_0<x<x_1, \\ a^2, & x_1<x\le 1, \end{cases} \tag{3.10a}$$

$x_0=1-x_1$, so $\varphi(x)$ is symmetric about $x=\frac{1}{2}$;

$$\varphi(x)=\begin{cases} b^2, & 0\le x<x_0, \\ a^2, & x_0<x<x_1, \\ b^2, & x_1<x\le 1, \end{cases} \tag{3.10b}$$

$x_0<\bar{x}<x_1$. Further,

$$\frac{1+\left(\frac{b}{a}\right)^2\cot^2\lambda_1 b(1-x_1)}{1+\left(\frac{b}{a}\right)^2\cot^2\lambda_2 b(1-x_1)}=\frac{1+\left(\frac{b}{a}\right)^2\cot^2\lambda_1 bx_0}{1+\left(\frac{b}{a}\right)^2\cot^2\lambda_2 bx_0}. \tag{3.11}$$

The numerical results of the next section show that extremizing functions $\varphi(x)$ need not be symmetric about $x=\frac{1}{2}$ in this subcase.

4. Numerical results. In this section we present the results of numerical computations for the particular example $a^2=1$ and $b^2=4$. If there exists (x_0,x_1) such that a function $\varphi(x)$ can be described by (3.10a), we call it type 1; if by (3.10b), type 2. By allowing the jump points to coincide either with each other or the boundary points 0 and 1, the no-jump and one-jump cases can be thought of as degenerate two-jump cases.

Figure 1 is a graph of the points in the x_0,x_1-plane which are the jump points of some extremizing $\varphi(x)$. In principle, a separate graph should be provided for the type 1 and type 2 cases. However, since $\varphi(x)$ and $\varphi(1-x)$ have the same eigevalues, we use the region $x_1\ge 1-x_0$ to graph type 1 points and $x_1\le 1-x_0$ to graph type 2 points. Since $x_1\ge x_0$, no points appear for $x_1<x_0$. The coordinates of all labeled points are given in Table 1 to four decimal places.

TABLE 1

	x_0	x_1	First eigenvalue	Second eigenvalue	Type
A	1.	1.	9.8696	39.4784	1
A	0.	1.	9.8696	39.4784	2
B	0.	1.	2.4674	9.8696	1
B	0.	0.	2.4674	9.8696	2
C	0.8465	1.	9.1063	28.9270	1
C	0.	0.8465	9.1063	28.9270	2
D	0.1174	0.8826	9.2305	30.0674	2
P(*S*1)	0.1471	1.	9.1967	29.6929	2
P(*S*2)	0.1195	0.8805	9.1967	29.6929	2
*P*1	0.1910	0.4270	4.6804	25.4354	1
*P*2	0.2550	0.8760	6.9479	20.6787	2
*P*31	0.0098	0.1469	9.1993	29.7190	1
*P*32	0.1464	0.9692	9.2019	29.7442	2
*P*41	0.0182	0.1469	9.2007	29.7434	1
*P*42	0.1464	0.9552	9.1926	29.6642	2

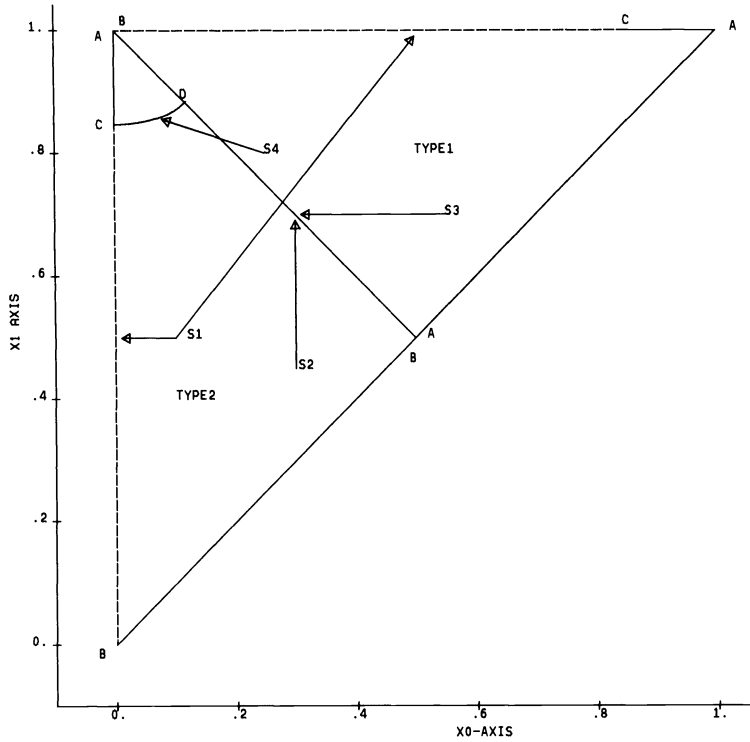

FIG. 1.

Type 1. In the nondegenerate two-jump case of type 1, we have $x_1 = 1 - x_0$. This line is labeled $S3$. The end points, (0, 1) and (.5, 5), are labeled B and A respectively. At B, $\varphi(x) \equiv b^2$, while $\varphi(x) \equiv a^2$ at A. The one-jump case is given by $x_0(1 - x_1) = 0$ and is labeled $S1$. The point (1, 1) is labeled A since $\varphi(x) \equiv a^2$ there. Any point on the line segment from (.5, .5) to (1, 1) is considered an "A" point and corresponds to the no-jump case. The one-jump case satisfies $x_0 < \bar{x}$ and (3.9). Only points between A and C along $S1$ satisfy these criteria. The dashed line from B to C indicates we already know these points do not correspond to possible extremizers.

Type 2. In this case, points where $x_0 = 1 - x_1$ are called B since $\varphi(x) \equiv b^2$. Similarly, (0, 1) is labeled A. All points on $\overline{BB}$ are in the no-jump case. The one-jump case is given by $\overline{BCA}$ as discussed previously, with no points on $\overline{CA}$ corresponding to extremizers.

We have numerically determined two curves in the (x_0, x_1)-plane which satisfy (3.11) in the two-jump case. These are the open line segment $\overline{AB}$ and the curve connecting C on the x_1-axis to D on the segment $\overline{AB}$. The point C divides the extremizing and nonextremizing one-jump cases. This second curve is labeled $S4$.

In Figs. 2–5 we present the points in the (λ_1, λ_2)-plane corresponding to φ's with jump points as in Fig. 1. Some points corresponding to other, nonextremizing, φ's are also shown. The coordinates of all labeled points are given in Table 1. Figure 2a shows the image of Fig. 1. The curves corresponding to $S1$–$S4$ are indicated, as well as the points A–D. Figure 2b is a pictographic representation of Fig 2a using only straight lines. The important features are highly exaggerated. A simple closed curve is formed by $S3$ from B to A, $S1$ from A to P, and $S2$ from P to B. We define R as the closed region bounded by this curve; subregions $R1$–$R4$ are defined as shown in Fig. 2c. Figures 3, 4 and 5 are blow-ups of the subrectangle from Fig. 2.

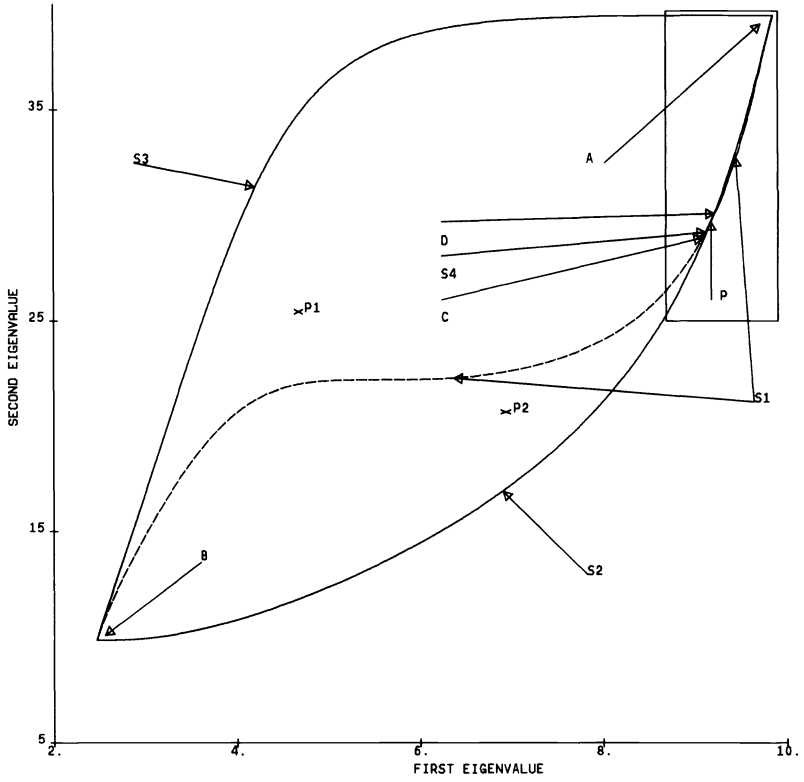

FIG. 2A.

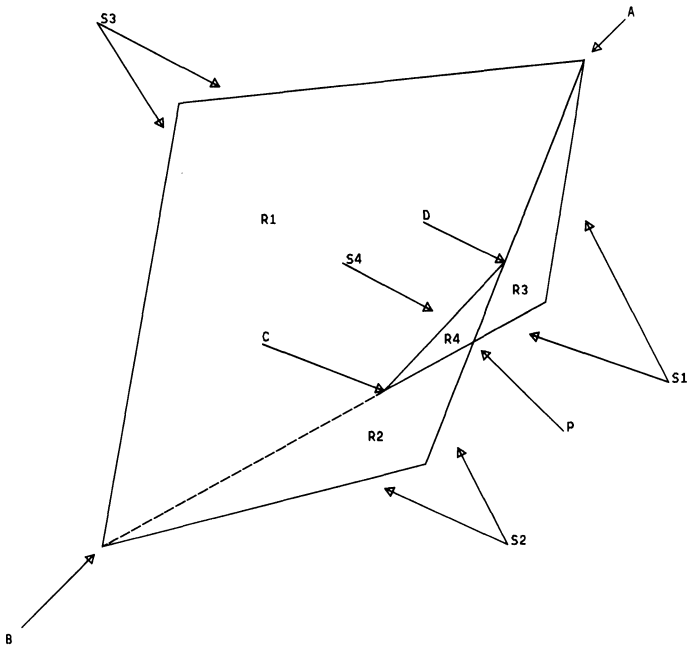

FIG. 2B. *Pictographic representation.*

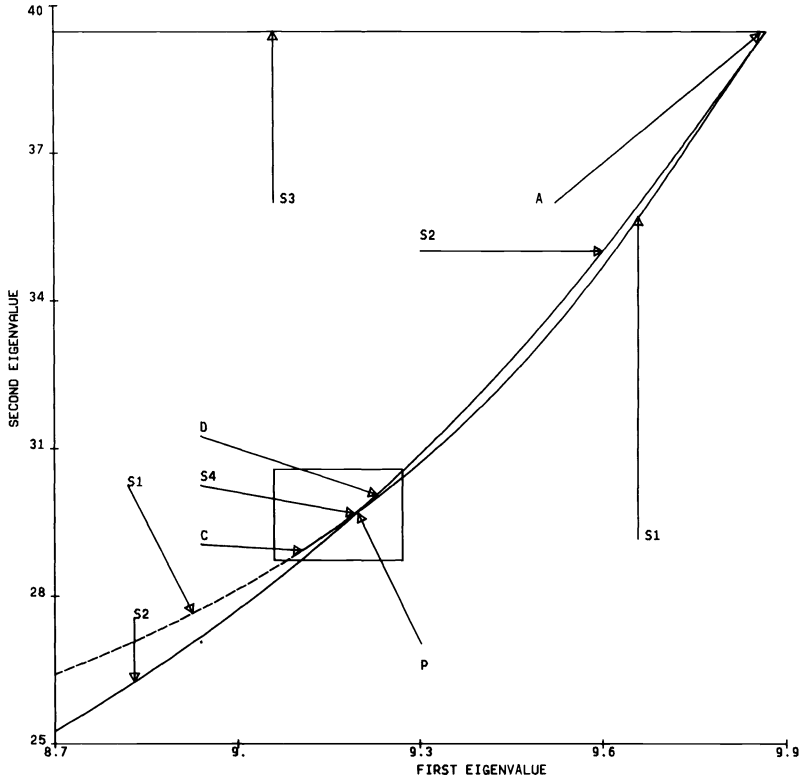
40
37
34
31
28
25
SECOND EIGENVALUE
8.7
9.
9.3
9.6
9.9
FIRST EIGENVALUE
S3
A
S2
S1
D
S4
C
S1
S2
P

FIG. 3.

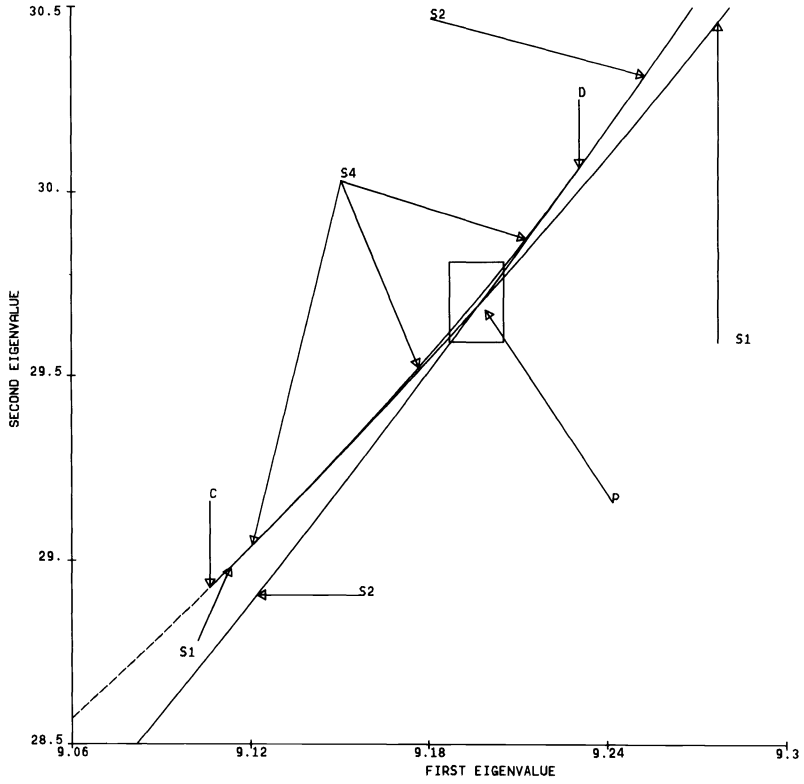
30.5
30.
29.5
29.
28.5
SECOND EIGENVALUE
9.06
9.12
9.18
9.24
9.3
FIRST EIGENVALUE
S2
D
S4
S1
C
P
S2
S1

FIG. 4.

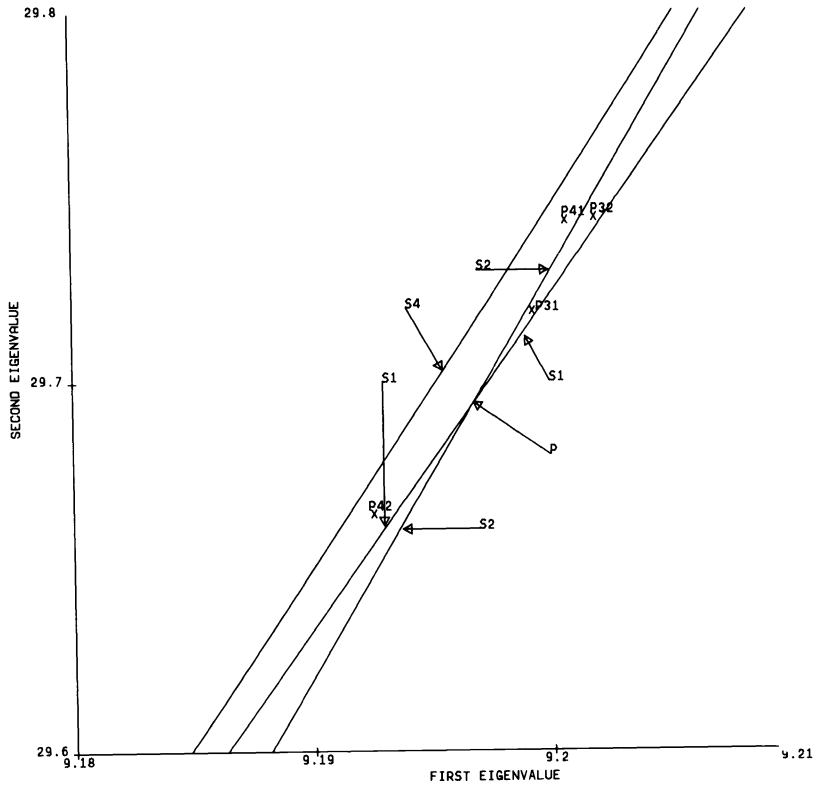

FIG. 5.

5. Λ for $C(1,4)$. We now prove that the region R in Fig. 2a is Λ, the two-dimensional eigenvalue range, for $C(1,4)$. Assuming that the numerical calculations are sufficiently accurate, Newton's method could be used to prove the existence of eigenvalues close to the computed ones. Numerical experience suggests that the computational accuracy can be substantially increased, so we assume the numerical results are correct.

THEOREM 3. *The region R is the two-dimensional eigenvalue range Λ for $C(1,4)$. That is, $(\lambda_1,\lambda_2)\in R$ if and only if there exists $\varphi\in C(1,4)$ such that* (1.1) *has λ_1 and λ_2 as its first two eigenvalues.*

Proof. First, assume (λ_1,λ_2) belongs to Λ but not R, and let $(\tilde{\lambda}_1,\tilde{\lambda}_2)$ be a point exterior to Λ and R such that it can be connected to (λ_1,λ_2) by a curve which does not intersect R. There exists $(\bar{\lambda}_1,\bar{\lambda}_2)$ on this curve which is a boundary point of Λ. Let $(\sigma_n,\tau_n)\in\Lambda$ converge to $(\bar{\lambda}_1,\bar{\lambda}_2)$ and set $f_n(\lambda_1,\lambda_2)=(\sigma_n-\lambda_1)^2+(\tau_n-\lambda_2)^2$. The points Q_n in Λ which minimize the f_n must lie on one of the S-labeled curves. However, since $(\bar{\lambda}_1,\bar{\lambda}_2)$ is outside R, there exists $\delta>0$ such that f_n evaluated at Q_n is no smaller than δ for all n. This is impossible since $(\sigma_n,\tau_n)\to(\bar{\lambda}_1,\bar{\lambda}_2)\in R$. Thus, if $(\lambda_1,\lambda_2)\in\Lambda$, then $(\lambda_1,\lambda_2)\in R$ also.

Now, let $\mathfrak{L}=(\lambda_1,\lambda_2)\in R$. If $\mathfrak{L}$ lies on some S curve, it is in Λ by definition of the S curves. If $\mathfrak{L}$ is not on any of these curves, it must be in one of the four subregions of R. As all cases are similar, assume $\mathfrak{L}\in R1$. Connect $\mathfrak{L}$ and $P1$ in $R1$ with an arc lying entirely in $R1$. If $\mathfrak{L}\notin\Lambda$, there exists $\bar{\mathfrak{L}}$ on this arc which is a boundary point of Λ. Let $\mathfrak{L}_n\notin\Lambda$ converge to $\bar{\mathfrak{L}}$, $\mathfrak{L}_n=(\sigma_n,\tau_n)$. Again consider minimizing the functions f_n. The

minimizing points Q_n all lie on S curves. Since $\bar{\mathfrak{Z}}$ is in the interior of $R1$, the $\mathfrak{Z}_n$ are also for n large. Thus, there exists $\delta>0$ such that the f_n evaluated at Q_n are no smaller than δ. But, $\mathfrak{Z}_n\to\bar{\mathfrak{Z}}\in\Lambda$, and so the f_n at Q_n must approach zero. Thus, $\mathfrak{Z}\in R$ implies $\mathfrak{Z}\in\Lambda$.

Theorem 3 shows that R and Λ are the same, so we have determined the two-dimensional eigenvalue range for $C(1,4)$.

THEOREM 4. *Every point in the interior of* $R1$ *corresponds to a* φ *of type* 1, *but not type* 2. *Every point in the interior of* $R2$ *corresponds to a* φ *of type* 2, *but not type* 1. *Points in the interior of* $R3$ *and* $R4$ *correspond to a* φ *of type* 1 *and a* φ *of type* 2.

Proof. We define $PC(1,4)$ as the class of functions $\varphi(x)$ described in (3.5), with $c=1$ and $d=2$ or $c=2$ and $d=1$, such that $0<x_0<x_1<1$. Elements of $PC(1,4)$ have two jumps. By utilizing piecewise trigonometric representations of the corresponding eigenfunctions $y_i(x)$ for (1.1), tedious algebraic and trigonometric manipulations can be used to show that

$$\left|\frac{\partial(\lambda_1,\lambda_2)}{\partial(x_0,x_1)}\right|\neq 0 \tag{5.1}$$

if the extremization condition (3.11) is not satisfied. Here, x_0 and x_1 are the jump points of φ and λ_1, λ_2 are the first two eigenvalues of the corresponding problem (1.1). Condition (5.1) guarantees that the implicit function theorem can be applied.

Let $\Omega 1$ be the eigenvalue image of type 1 elements of $PC(1,4)$, so $c=1$ and $d=2$. Assume (λ_1,λ_2) is not in $\Omega 1$, but is interior to $R1$. Connect (λ_1,λ_2) to $P1\in\Omega 1$ with an arc interior to $R1$. There exists a point $\mathfrak{Z}$ on this arc which is a boundary point of $\Omega 1$. Since $\mathfrak{Z}\in R1$, the implicit function theorem implies $\mathfrak{Z}\notin\Omega 1$. Let $\mathfrak{Z}_n\in\Omega 1$ converge to $\mathfrak{Z}$. The points $\mathfrak{Z}_n$ are generated by φ_n in $PC(1,4)$ of type 1 with jump points $(x_{0,n},x_{1,n})$. Choosing subsequences if necessary, let $x_{0,n}\to x_0$ and $x_{1,n}\to x_1$. Use (x_0,x_1) to construct $\varphi(x)$ in $PC(1,4)$ of type 1 such $\varphi_n\to\varphi$ in the L_1-norm. Now, $\mathfrak{Z}=(\lambda_1(\varphi),\lambda_2(\varphi))$ by continuity of the eigenvalues [5]. If $0<x_0<x_1<1$ and (3.11) is not satisfied, we have the contradiction $\mathfrak{Z}\in\Omega 1$. If (3.11) is satisfied, or φ is a degenerate two-jump function, $\mathfrak{Z}$ would be on an S curve, another contradiction. Thus, $\mathfrak{Z}$ is not a boundary point of $\Omega 1$, the final contradiction. Hence, (λ_1,λ_2) in the interior of $R1$ implies (λ_1,λ_2) in $\Omega 1$. By similar reasoning, every point in the interior of $R2$ is the image of a type 2φ, while points in the interior of $R3$ and $R4$ are images of a φ of type 1 and a φ of type 2.

We now show that no point in the interior of $R2$ is the image of a type 1 φ. Assume some type 1 φ has its eigenvalues in the interior of $R2$. As above, we then have that every element of the interior of $R2$ is the image of a type 1 φ. Let $\mathfrak{Z}=(\lambda_1,\lambda_2)$ be a point on $S2$ in the interior of the segment BP, and let $\mathfrak{Z}_n$ in $\Omega 1$ converge to $\mathfrak{Z}$. The points $\mathfrak{Z}_n$ are generated by φ_n in $PC(1,4)$ of type 1 with jump points $(x_{0,n},x_{1,n})$. Choosing subsequences, $x_{0,n}\to x_0$ and $x_{1,n}\to x_1$, we use (x_0,x_1) to construct φ in $PC(1,4)$ such that $\varphi_n\to\varphi$ in the L_1-norm. Thus, $\mathfrak{Z}=(\lambda_1(\varphi),\lambda_2(\varphi))$ by the continuity of the eigenvalues [5]. If $x_0=1-x_1$, $\mathfrak{Z}$ is on $S3$, a contradiction. If φ is a degenerate two-jump function, $\mathfrak{Z}$ lies on AB along $S1$, another contradiction. Thus, (5.1) holds and the implicit function theorem shows there is a neighborhood about $\mathfrak{Z}$ corresponding to a neighborhood about (x_0,x_1). This is false since no point of $\Omega 1$ can lie outside $\Lambda=R$. Similarly, no point of $R1$ corresponds to a φ of type 2.

6. Conclusions. This paper has characterized the set of points $(\lambda_1(\varphi),\lambda_2(\varphi))$ as $\varphi(x)$ varies over $C(a(x),b(x))$. It seems curious that this question hasn't been studied already, since it is of theoretical interest to know precisely what the various eigenvalue ranges are. Although a complete characterization has been given only for the case

$a(x)\equiv 1$ and $b(x)\equiv 4$, the general results suggest similar characterizations hold for arbitrary choices of constant bounding functions. The case of nonconstant bounding functions does not appear to be amenable to the same sort of detailed analysis.

The results presented above also provide a method, simple in principle and not difficult numerically, for solving extremal eigenvalue problems with constant bounding functions. Compute all pairs $(\lambda_1(\varphi),\lambda_2(\varphi))$ with φ one of the possible extremizers given in (3.7)–(3.11). A set of curves SR will be generated. The answer to the extremal eigenvalue problem will lie on one of these curves if (1.5) is satisfied. If (1.5) is violated, a piecewise constant φ with two jumps such that (3.11) is violated, or a one-jump φ such that $\bar{x}\leq x_0$, will provide the answer. In all cases, the extremizing φ will be piecewise constant, have at most two jumps, and assume only the smallest and largest allowed values.

REFERENCES

[1] B. E. WILLNER AND T. J. MAHAR, *Extrema of functions of eigenvalues*, J. Math. Anal. Appl., 72 (1979), pp. 730–739.

[2] R. D. GENTRY AND D. O. BANKS, *Bounds for functions of eigenvalues of vibrating systems*, J. Math. Anal Appl., 51 (1975), pp. 100–128.

[3] T. J. MAHAR AND B. E. WILLNER, *An extremal eigenvalue problem*, Comm. Pure Appl. Math., (1976), pp. 517–529.

[4] J. B. KELLER, *The minimum ratio of two eigenvalues*, SIAM J. Appl. Math., 31 (1976), pp. 485–491.

[5] T. J. MAHAR AND B. E. WILLNER, *Sturm-Liouville eigenvalue problems whose eigenfunctions have linearly dependent squares*, Comm. Pure Appl. Math., (1980), pp. 567–578.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0008 $01.00/0

SEPARATED FLOW PAST A PLATE WITH SPOILER*

ALAN R. ELCRAT†

Abstract. The Helmholtz–Kirchhoff flow past a flat plate with a spoiler attached is studied using conformal mapping of the hodograph image. The hodograph is a Riemann surface with polygonal boundary and a theorem of Gilbarg showing how to map such a domain into a half-plane is used. The result obtained actually solves an inverse problem and can be used to "shoot" for a solution of the direct problem. Some numerical results for this procedure are presented at the end of this paper.

Introduction. This work is concerned with a two-dimensional potential flow model for fluid flow past an obstacle with a flap or "spoiler" raised into the stream which causes the flow to separate from its edge. This separation is modeled with the classical Helmholz–Kirchhoff device of a free streamline extending from the points of separation to infinity. The particular configuration dealt with here is a flat plate at angle of attack with a flat plate spoiler attached. This is intended to be a simple model of the flow past an airfoil with a spoiler raised on its top surface.

The analysis given here is based on conformal mapping of the hodograph or velocity plane and in its spirit goes back to Kirchhoff's solution for a flat plate perpendicular to the free stream [1]. The hodograph, however, is not simply covered and must be thought of as a Riemann surface, so in detail this work is inspired by that of Gilbarg and Serrin [2]. In fact, the piecewise linear boundary suggests the Schwarz–Christoffel transformation, and a fundamental tool in what follows is a generalization of these conformal maps to polygonal regions on Riemann surfaces by Gilbarg [3]. In principle, a great number of flows past polygonal obstacles might be treated using this device, but there are serious technical difficulties in computing the parameters needed to determine the conformal mappings required. The problem dealt with here is just simple enough to yield to relatively straight forward analysis, but, we feel, just complicated enough to be an interesting addition to the large literature on this subject.

1. Preliminary results. First we state a specialization of Gilbarg's theorem tailored to our needs.

PROPOSITION. *Suppose that $\zeta=f(t)$ is an analytic function defined on the upper half-plane* $\operatorname{Im} t>0$ *and that $f'(t)\neq 0$ except for a finite number of points a_j. Suppose that f maps the upper half-plane onto a Riemann surface with a closed polygonal boundary having interior angles $\beta_k\pi$ at the vertices, which are images of $t=b_k$, b_k real. Then*

$$f'(t)=A\prod_j (t-a_j)^{\alpha_j}(t-\bar{a}_j)^{\alpha_j}\prod_k (t-b_k)^{\beta_k-1},$$

where α_j is the order of the zero of f' at the point a_j in the upper half-plane and A is a complex constant.

Gilbarg also gave a theorem which guarantees the existence of a conformal mapping of a Riemann surface satisfying certain hypotheses onto a half-plane. We need only the above result, however, which states the form such a mapping must have if there is one.

* Received by the editors December 11, 1980.
† Department of Mathematics, Wichita State University, Wichita, Kansas 67208.

We also need the following result, which gives a sufficient condition that an analytic function defined on the upper half-plane be a conformal map onto a Riemann surface.

LEMMA. *Suppose that f is analytic on the open upper half-plane and extends continuously to the closed upper half-plane (as a subset of the Riemann sphere.) Suppose that $\Gamma = f(\mathcal{R} \cup \{\infty\})$ is a closed curve made up of a finite number of Jordan arcs and that Γ bounds a region R and a subregion R_0 of R in such a way that the winding number $n(\Gamma, z) = 2$ for $z \in R_0$ and $n(\Gamma, z) = 1$ for $z \in R - \bar{R}_0$.*

Suppose also that $f'(t) \neq 0$ except for one point t_0 where $f''(t_0) \neq 0$. Then $\zeta_0 = f(t_0) \in R_0$, and we may think of the image of the upper half-plane as a two-sheeted Riemann surface with a branch point at ζ_0. Further, f is a conformal map onto this Riemann surface.

Since $n(\Gamma, \zeta)$ = the number of times the value ζ is taken on by f in the upper half-plane, we need only recall that f is two to one in a neighborhood of t_0 in order to see that $\zeta_0 \notin R - \bar{R}_0$. A slight extension of this argument shows that $\zeta_0 \in R_0$. In fact, if ζ_0 were on the bounday of R_0 (which is contained in Γ by hypothesis), then the preceding argument can be applied to f restricted to the upper half-plane minus a small neighborhood of the $\hat{t} \in R$ such that $f(\hat{t}) = \zeta_0$ to obtain a contradiction. The neighborhoods on each of the two sheets are ordinary ones except at the point ζ_0, which lies on both sheets, where the so-called "cyclic neighborhoods" [4] are used.

2. Statement of the physical problem. All flows considered here are two-dimensional, irrotational and incompressible.

We are interested in the flow, constant and horizontal at infinity, past an inclined flat plate with a spoiler, cf. Fig. 1, which turns the corner at the leading edge D, stagnates at F and separates at A and B, the free streamlines extending to infinity (M). Bernoulli's equation, $p + \frac{1}{2}|\text{velocity}| = \text{constant}$, implies that the velocity is infinite at the leading edge D. We use the following conventions:

$\zeta = u - iv$, u, v Cartesian velocity components,

$w = \phi + i\psi$, ϕ, ψ potential and stream functions.

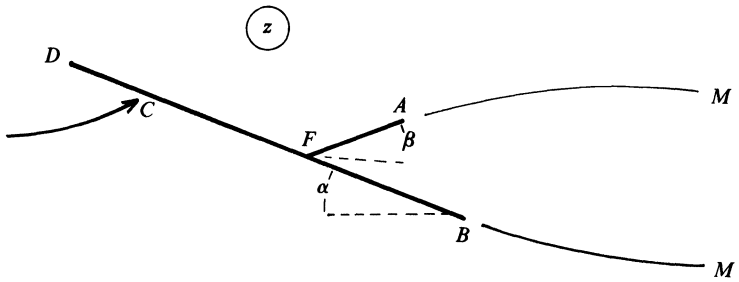

FIG. 1.

Then

$$\zeta = \frac{dw}{dz},$$

and ζ, w are analytic functions. The method of solution which we will use is then based on conformal mapping, and the first step is to use physical reasoning to discern what the hodograph image ζ should be. Here the straight-line boundary of the flow, the constant magnitude of the velocity on AMB and our assumptions of stagnation at F and infinite velocity at D indicate that the following image results, with the boundary

winding around the bounded region inside $ABFA$ twice. If we set $\eta=\ln(\zeta/v_\infty)$, we obtain a polygonal domain (to which the generalized Schwarz–Christoffel formula can be applied). On the other hand, if we assume that the stagnation streamline is given by $\psi=0$ and that $w=0$ at C, the velocity potential image is a plane slit along the positive axis with the origin correspondingto C and the top and bottom of the slit corresponding to $CDFAM$ and CBM, respectively. The technique which we use here is to construct a sequence of conformal maps which match the z- and ζ-planes, the appropriate matching of parameters determining a flow. The intermediate domain to which we map is the upper-half of the t-plane. The w-plane is mapped onto the upper-half of a τ-plane by $w=\sqrt{\tau}$, so that the first nontrivial step in this procedure may be taken to be the mapping of the η-plane image of the hodograph onto $\operatorname{Im} t>0$.

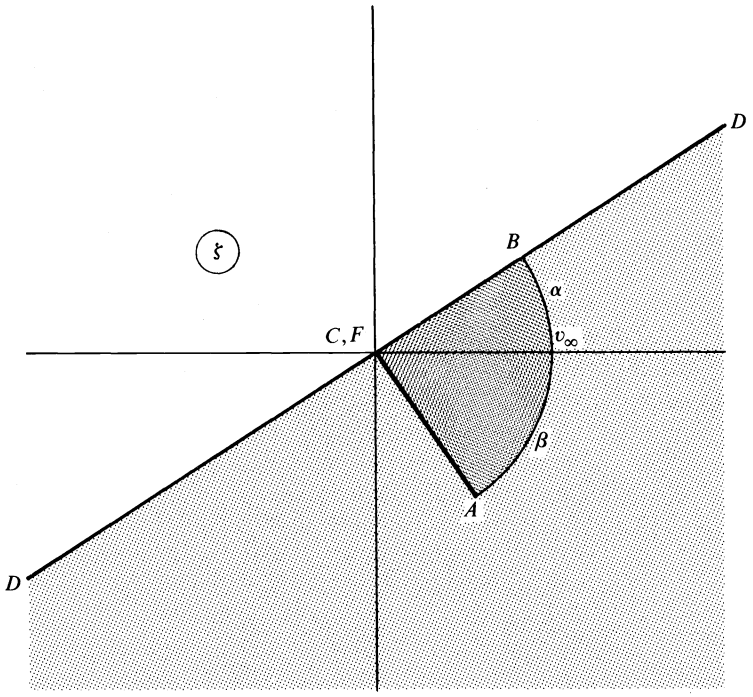

FIG. 2.

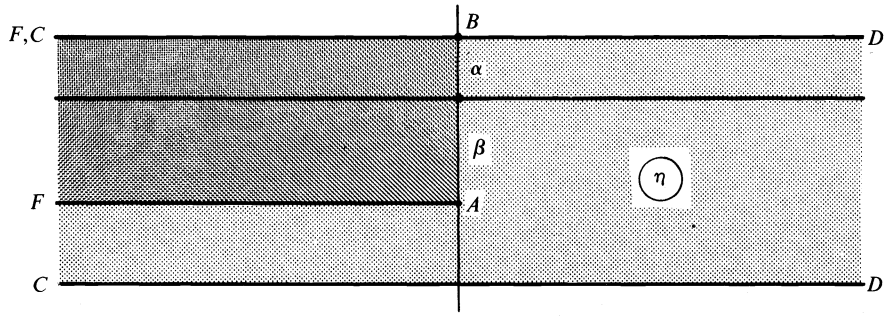

FIG. 3.

3. Mapping the hodograph. Gilbarg's theorem tells us that a mapping of $\operatorname{Im} t>0$ onto the domain of Fig. 3 which takes 0 to A, 1 to B and ∞ to F must satisfy

$$\frac{d\eta}{dt}=A\,\frac{(t-a)^2+b^2}{\sqrt{t}\sqrt{t-1}}\,\frac{1}{(t-c)}\,\frac{1}{(t-d)}$$

for some complex A, some $a+bi$, $b>0$ and some real c,d with $1<c<d$. Further, if we integrate the right-hand side from 0, we have

$$\eta=A\int_0^t \frac{(t-a)^2+b^2}{\sqrt{t}\sqrt{t-1}}\,\frac{dt}{(t-c)(t-d)}=\beta i \tag{1}$$

and

$$\alpha+\beta=-A\int_0^1 \frac{(t-a)^2+b^2}{\sqrt{t}\sqrt{1-t}}\,\frac{dt}{(t-c)(t-d)} \tag{2}$$

with A necessarily real and negative. The integral (1) is given by a Cauchy principal value when t passes through c and d, and the required changes in argument of η at these values imply [5, p. 26]

$$\frac{A}{d-c}\,\frac{(c-a)^2+b^2}{\sqrt{c(c-1)}}=-1 \tag{3}$$

and

$$\frac{A}{d-c}\,\frac{(d-a)^2+b^2}{\sqrt{d(d-1)}}=-1, \tag{4}$$

respectively. It is fortuitous that the integral in (1) can be done in closed form. The steps, which will not be given in detail here, are to expand the rational part of the integrand in partial fractions, substitute u such that $du=dt/\sqrt{t}$ and then use the appropriate trigonometric substitutions. The result can be written, using (3) and (4), as

$$I=\ln\left[\left(2t-1+2\sqrt{t(t-1)}\right)\left(\frac{t-c}{t-d}\right)^{1/A}\left(\frac{2\sqrt{t(t-1)}\,\sqrt{d(d-1)}+(2d-1)t-d}{2\sqrt{t(t-1)}\,\sqrt{c(c-1)}+(2c-1)t-c}\right)^{1/A}\right],$$

where I is the required antiderivative. Now, since $I(0)=\pi i$, we have

$$\eta=A[I(t)-\pi i]-\beta i$$

and (2) implies that $A=-\frac{\alpha+\beta}{\pi}$. It follows that

$$\frac{\zeta}{v_\infty}=e^{\alpha i}\left(2t-1+2\sqrt{t(t-1)}\right)^{-(\alpha+\beta)/\pi}\frac{(t-c)\left(2\sqrt{t(t-1)}\,\sqrt{d(d-1)}+(2d-1)t-d\right)}{(t-d)\left(2\sqrt{t(t-1)}\,\sqrt{c(c-1)}+(2c-1)t-c\right)}. \tag{5}$$

The parameters a and b appear only implicitly through (3), (4). We can check directly to see that this function maps the real axis onto the boundary of the required hodograph. Then the lemma of §1 implies that we do in fact have a Riemann surface of the required type with branch point at $\zeta(a+bi)$, which is the conformal image of the upper half-plane.

Finding this mapping now reduces to finding a solution c, d of (3), (4) for given $a+bi$, $b>0$. Since (4) can be written in the form

$$d=c+\frac{\alpha+\beta}{\pi}\frac{(c-a)^2+b^2}{\sqrt{c(c-1)}}, \tag{6}$$

we may think of this as a single equation for c,

$$G(c)=\frac{(c-a)^2+b^2}{\sqrt{c(c-1)}}-\frac{(d-a)^2+b^2}{\sqrt{d(d-1)}}=0, \tag{7}$$

with d given by (6). We may assume that $\alpha+\beta<\pi$ since this includes all feasible geometries. Then, by observing the asymptotic properties of G for c near 1 and ∞, we see that G is positive near 1 and ultimately negative. Therefore, the intermediate value theorem implies the existence of $c\in(1,\infty)$ such that $G(c)=0$ and (6) then gives the required pair (c,d). Since every solution (c,d) yields a conformal map of the required type and the mapping is uniquely determined by the conditions imposed by the action on three boundary points, we deduce that these equations have a unique solution $1<c<d<\infty$.

For the practical determination of the mapping, we need only solve (7). This fact together with the closed form expression for the integral in (1) are the features of the problem discussed in this paper that allow a relatively straightforward analysis.

It should be emphasized here that there are as many different mappings and as many corresponding ζ images as there are points $a+ib$, $b>0$.

4. An inverse solution of the physical problem. The solution of the flow problem, based on the above mapping formula (5), can be outlined as follows. The functions $\zeta(t)$ and $\tau=\sqrt{w}$ define mappings of the hodograph and velocity potential planes onto the upper halves of the t and τ planes, respectively. A correspondence between these is given by

$$t=\frac{p\tau+q}{\tau+s},$$

where p,q,s are real. If we observe that $\tau=0$ is mapped to $t=c$, and $\tau=\infty$ to $t=m$, where m is the unique value of $t\in(0,1)$ such that $\zeta(t)=v_\infty$, we get

$$t=\frac{m\tau+cs}{\tau+s},$$

that is,

$$\tau=\frac{-st+cs}{t-m}.$$

This implies that

$$\frac{d\tau}{dt}=\frac{s(m-c)}{(t-m)^2}.$$

Then

$$\frac{dz}{dt}=\frac{dz}{dw}\frac{dw}{d\tau}\frac{d\tau}{dt}=\frac{2}{\zeta}\frac{(c-m)s^2(t-c)}{(t-m)^3}$$

and, if we assume that $z=0$ when $t=1$,

$$z(t)=\frac{e^{-\alpha i}2(c-m)s^2}{v_\infty}\int_1^t \frac{g(t)}{(t-m)^3}dt, \tag{8}$$

where

$$g(t)=\frac{\left(2t-1+\sqrt{t(t-1)}\right)^{(\alpha+\beta)/\pi}(t-d)\left(2\sqrt{t(t-1)}\sqrt{c(c-1)}+(2c-1)t-c\right)}{\left(2\sqrt{t(t-1)}\sqrt{d(d-1)}+(2d-1)t-d\right)}.$$

Note that the value of s is not yet determined. We will think of s as a scale factor, and it will always be chosen so that the length $|BD|$ is normalized to one. The direct solution of the problem then reduces to finding (a,b) such $|DF|$ and $|FA|$ have specified lengths. We have no theorem to present for the direct problem. To attack the problem directly on the basis of $|DF|$ and $|FA|$ being functions of (a,b) seems hopeless from the theoretical point of view. An indirect proof based on the existence of appropriate conformal mappings may be possible, and we hope to return to this in a later work. On the other hand, we will adopt the opposite point of view here and study the inverse problem in which (a,b) is given and $|DF|$ and $|FA|$ are determined. From the theoretical point of view this problem is solved by (8). In the final section of the paper, we present computational results for the inverse problem which can be thought of as "shooting" computations for the direct problem. As a part of these, we calculate the total lift and drag predicted by this model using an idea of Gurevich ([6, pp. 81–84]). In fact, if $X+iY$ is the total force on our obstacle,

$$X+iY=\frac{-i\rho v_\infty^2}{2}\oint dz.$$

If we write this in dimensionless form

$$C_D+iC_L=\frac{X+iY}{\frac{1}{2}\rho v_\infty^2 L},$$

we obtain

$$(C_D+iC_L)(e^{-\alpha i})=-i\int_\gamma F(t)\,dt\Big/\int_1^d F(t)\,dt,$$

where γ is a small half circle in the upper half-plane centered at $t=m$ and $F(t)=g(t)/(t-m)^3$. The residue theorem then implies that

$$(C_D+iC_L)(e^{-\alpha i})=\pi g''(m)\Big/\int_1^d F(t)\,dt. \tag{9}$$

This is the formula used in our computations.

5. Some computational results. Our calculations of various flow quantities were done in the following steps.

1) Choose $a,b>0$.
2) Calculate c and then d using (6) and (7).
3) Calculate m, the solution of $\operatorname{Im}\zeta(t)=0$ for $t\in(0,1)$.
4) Calculate the length $|BD|$ using the integral over $[1,d]$ in (8).
5) Calculate the lengths $|DF|$ and $|FA|$ and normalize using the previous step.
6) Calculate $g''(m)$ and use (9) to obtain C_D and C_L.

The equations in 2) and 3) were solved using the subroutine ZEROIN provided with the book [7] of Forsythe, Malcolm and Moler. The integrals in 4) and 5) we computed using the adaptive quadrature subroutine QUANC8 given in [7]. All computations were done on the Wichita State University IBM 370-145 using double precision arithmetic. The infinite integrals which give $|DF|$ and $|FA|$ were transformed to integrals over finite intervals using a transformation $du=dt/(t-m)^{1+\delta}$ with the appropriate $\delta>0$ and then calculated using QUANC8. A listing of the author's program is available on request to anyone interested in further details of the computations.

In our presentation of the results in Tables 1 and 2, all values will be rounded to two decimal places. For brevity we have included only some representative cases, and only a, b among the various "nonphysical" parameters are given. We denote $|DF|$, $|FA|$ by LF, LS, respectively.

TABLE 1
$\alpha=15°, \beta=30°$

a	b	LF	LS	C_D	C_L
1.5	.5	.16	.19	.15	−.02
1.5	1.	.05	.09	.11	.20
1.0	.19	.41	.09	.09	.18
1.0	.17	.56	.12	.09	.11
1.0	.16	.67	.14	.09	.06
1.0	.15	.83	.17	.10	−.01

TABLE 2
$\alpha=10°, \beta=30°$

a	b	LF	LS	C_D	C_L
1.0	.09	.52	.06	.04	.13
1.0	.085	.62	.07	.04	.10
1.0	.08	.76	.08	.04	.05
1.0	.077	.87	.09	.04	.01
1.0	.075	.96	.10	.04	−.02
1.0	.0745	.98	.10	.04	−.02

It is also possible to obtain a value of LF greater than one for certain choices of the parameters a and b. Our tentative interpretation of this outcome is that the flow separates tangentially from the bottom of the plate before reaching the trailing edge. Since we are not interested in this situation, we have not done the more detailed calculations that would be necessary to justify this interpretation.

Acknowledgments. The author would like to gratefully acknowledge a number of stimulating conservations with William Wentz, director of the Walter M. Beech Memorial Wind Tunnel at Wichita State University, about the subject of this paper.

REFERENCES

[1] G. KIRCHHOFF, *Zür Theorie freier Flüssig-Keitsstrahlen*, J. Reine Angew. Math., 70 (1869), pp. 289–298.
[2] D. GILBARG AND J. SERRIN, *Free boundaries and jets in the theory of cavitation*, J. Math. and Physics, 29 (1950), pp. 1–12.
[3] D. GILBARG, *A generalization of the Schwarz-Christoffel transformation*, Proc. Nat. Acad. Sci., 35 (1949), pp. 609–611.
[4] H. COHEN, *Conformal Mapping on Riemann Surfaces*, McGraw-Hill, New York, 1967.
[5] P. HENRICI, *Applied and Computational Complex Analysis*, I, Wiley-Interscience, New York, 1974.
[6] M. I. GUREVICH, *The Theory of Jets in an Ideal Fluid*, Pergamon Press, New York, 1966.
[7] G. FORSYTHE, M. MALCOLM AND C. MOLER, *Computer Methods for Mathematical Computations*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0009 $01.00/0

POLYNOMIAL EXPANSIONS FOR SOLUTIONS OF $D_x^r u(x,t) = D_t u(x,t)$, $r = 2,3,4,\cdots$ *

HANS KEMNITZ†

Abstract. This paper is an extension of results by P. C. Rosenbloom and D. V. Widder [Trans. Amer. Math. Soc., 92 (1959), pp. 220–266] concerning the expansion of a solution $u(x,t)$ of the heat equation, $D_x^2 u(x,t) = D_t u(x,t)$, in a series of polynomial solutions.

It is found that a polynomial expansion

$$\sum_{n=0}^{\infty} a_n v_{r,n}(x,t)$$

converges in an infinite strip $|t|<\sigma$, where the polynomials

$$v_{r,n}(x,t) = n! \sum_{k+rl=n} \frac{x^k}{k!}\frac{t^l}{l!}$$

satisfy the partial differential equation $D_x^r u(x,t) = D_t u(x,t)$. Furthermore it is found that there exists a solution of $D_x^r u(x,t) = D_t u(x,t)$ which has a Maclaurin expansion in a strip $|t|<\sigma$ and which reduces to $f(x)$ for $t=0$ if and only if $f(x)$ is an entire function of special growth. All the proofs need only elementary calculus and make no use of fundamental solutions.

Introduction. In 1959 P. C. Rosenbloom and D. V. Widder [10] established the result that a polynomial series

$$\sum_{n=0}^{\infty} a_n v_n(x,t)$$

converges in a strip $|t|<\sigma$, where $v_n(x,t)$ are the heat polynomials

$$n! \sum_{k+2l=n} \frac{x^k}{k!}\frac{t^l}{l!}.$$

This theorem contains a formula for σ in terms of the coefficients a_n, analogous to Hadamard's formula for the radius of convergence for power series. The polynomials v_n satisfy the heat equation, $D_x^2 u(x,t) = D_t u(x,t)$, and those series expansions do the same in the whole strip of their convergence. D. V. Widder [12] showed in 1962 that there exists a solution of the heat equation which is equal to its Maclaurin series in the strip $|t|<\sigma$ and which reduces to $f(x)$ for $t=0$ if and only if $f(x)$ is an entire function of growth $(2, \frac{1}{4}\sigma)$.

In 1965 L. R. Bragg [2] found that a polynomial series in terms of

$$R_n^{\mu}(x,t) := \Gamma(\mu/2+n) \sum_{j=0}^{n} \frac{1}{\Gamma(\mu/2+j)} \binom{n}{j} x^{2j}(4t)^{n-j}$$

converges in a strip $|t|<\sigma$, where $R_n^{\mu}(x,t)$ are called radial heat polynomials and satisfying the radial heat equation

$$D_x^2 u(x,t) + \frac{\mu-1}{x} D_x u(x,t) = D_t u(x,t), \qquad \mu > 1.$$

The differential operator $D_x^2 + ((\mu-1)/x)D_x$ is the Laplacian in radial coordinates when $\mu = n$, a positive integer. D. T. Haimo [6] obtained the criterion that there exists a

* Received by the editors January 22, 1981, and in final form September 19, 1981.
† Wiesenstrasse 69, 7830 Emmendingen, West Germany.

solution of the radial heat equation which has a Maclaurin expansion in a strip $|t|<\sigma$ and which reduces to $f(x^2)$ for $t=0$ if and only if f is an even entire function of growth $(1,\frac{1}{4}\sigma)$.

In 1971 F. M. Cholewinski and D. T. Haimo [3] studied the Laguerre heat equation.

$$xD_x^2 u(x,t)+(\alpha+1-x)D_x u(x,t)=D_t u(x,t), \qquad 2\alpha>-1.$$

They found that a series expansion in terms of Laguerre heat polynomials

$$p_{n,\alpha}(x,t)=\sum_{k=0}^{n}\binom{n}{k}\frac{\Gamma(n+\alpha+1)}{\Gamma(n-k+\alpha+1)}(xe^{-t})^{n-k}(1-e^{-t})^k$$

converges in an unsymmetrical strip, i.e.,

$$\ln\frac{1}{1+4\sigma}<t<\ln\frac{1}{1-4\sigma}, \qquad 0<\sigma<\frac{1}{4},$$
$$\ln\frac{1}{1+4\sigma}<t<+\infty, \qquad \frac{1}{4}\leq\sigma.$$

D. T. Haimo [7] showed that there exists a solution of the Laguerre heat equation which has a Maclaurin expansion in such an unsymmetrical strip and which reduces to $f(x)$ for $t=0$ if and only if $f(x)$ is an entire function of growth $(1,\frac{1}{4}\sigma)$. It is remarkable that the determination of σ for both the Laguerre heat equation and the radial heat equation are completely equal because the polynomials are connected by the formula

$$p_{n,\alpha}(x,t)=R_n^{2\alpha+2}\left(x^{1/2}e^{-t/2},\frac{1-e^{-t}}{4}\right).$$

Other papers in which such aspects are transfered to partial differential equations of second order in one or several space variables include [4], [5], [8], [9], [11].

For a discussion of polynomial series it was essential to know the behavior of $v_n(x,t)$, $R_n^{\mu}(x,t)$, $p_{n,\alpha}(x,t)$ as $n\to\infty$. Therefore all proofs use nontrivial formulas of special functions. For example, Rosenbloom and Widder [10] employ the Poisson representation of the heat polynomials

$$v_n(x,t)=\int_{-\infty}^{+\infty}k(x-y,t)y^n\,dy,$$

where $k(x,t)$ is the fundamental solution of the heat equation. In the present note we give a method for estimating polynomials

$$v_{r,n}(x,t)=n!\sum_{k+rl=n}\frac{x^k}{k!}\frac{t^l}{l!}$$

by use of elementary calculus, in particular for the heat polynomials when $r=2$. Our estimates are not of the same quality as in [10], but they are sufficient to establish the principal result:

If

$$\limsup_{n\to\infty}\frac{r\cdot e}{n}|a_n|^{r/n}=\frac{1}{\sigma}<+\infty,$$

then the series

$$\sum_{n=0}^{\infty}\frac{a_n}{n!}v_{r,n}(x,t)$$

converges in an infinite strip $|t|<\sigma$ (Theorem 1.6). An application of this result leads to Cauchy problems for the partial differential equation (Corollary 2.5, Corollary 2.6)

$$D_x^r u(x,t)=D_t u(x,t),$$

because $v_{r,n}(x,t)$ satisfy the equation for fixed r. The necessary and sufficient condition for a solution of $D_x^r u(x,t)=D_t u(x,t)$ which is equal to its Maclaurin series in a strip $|t|<\sigma$ and which reduces to $f(x)$ for $t=0$ is that $f(x)$ have growth $(r/(r-1), ((r-1)/r)(1/r\sigma)^{1/(r-1)})$.

Now, we list the notation used in our study:

r	fixed integer, $r\geq 2$;
s	integer with $0\leq s\leq r-1$;
$\mathbb{K}$	field of real or complex numbers;
$\lvert\cdot\rvert$	Euclidean norm on $\mathbb{K}$;
$D_z u$	partial derivative of u with respect to $z\in\mathbb{K}$.

In [10] the heat polynomials $v_n(x,t)$ are defined by a generating function. In the same way we obtain for $x,t,z\in\mathbb{K}$

$$\exp(xz+tz^r):=\sum_{n=0}^{\infty} v_{r,n}(x,t)\frac{z^n}{n!}. \tag{0.1}$$

By use of Cauchy's rule for multiplying power series, we obtain the explicit expression

$$v_{r,n}(x,t)=n!\sum_{k=0}^{[n/r]}\frac{t^k}{k!}\frac{x^{n-rk}}{(n-rk)!}, \tag{0.2}$$

where $[n/r]$ means the largest integer $\leq n/r$. Setting $r=2$, we get back the heat polynomials, $v_{2,n}(x,t)=v_n(x,t)$. For further reference, we note the following evident properties of $v_{r,n}(x,t)$ for $n=mr+s$, $s\in\{0,\cdots,r-1\}$:

$$v_{r,n}(x,t)=n!\sum_{k=0}^{m}\frac{t^k}{k!}\frac{x^{r(m-k)+s}}{(r(m-k)+s)!}, \tag{0.3}$$

$$v_{r,s}(x,t)=x^s, \tag{0.4}$$

$$v_{r,n}(x,0)=x^n, \tag{0.5}$$

$$v_{r,n}(0,t)=\begin{cases}\dfrac{(rm)!}{m!}t^m, & s=0,\\ 0, & s\neq 0,\end{cases} \tag{0.6}$$

$$z^n v_{r,n}(x,t)=v_{r,n}(zx,z^r t), \tag{0.7}$$

$$|v_{r,n}(x,t)|\leq v_{r,n}(|x|,|t|), \tag{0.8}$$

$$D_x v_{r,n}(x,t)=\begin{cases}\dfrac{n!}{(n-1)!}v_{r,n-1}(x,t), & n\geq 1,\\ 0, & n=0,\end{cases} \tag{0.9}$$

$$D_t v_{r,n}(x,t)=\begin{cases}\dfrac{n!}{(n-r)!}v_{r,n-r}(x,t), & n\geq r,\\ 0, & n<r,\end{cases} \tag{0.10}$$

$$D_x^r v_{r,n}(x,t)=D_t v_{r,n}(x,t). \tag{0.11}$$

1. Strip of convergence. In this section we show that the polynomial series

$$\sum_{n=0}^{\infty} \frac{a_n}{n!} v_{r,n}(x,t) \tag{1.0}$$

converges in a strip $|t|<\sigma$, where σ is calculated with the aid of the coefficients $a_n \in \mathbb{K}$. For that, we need several simple preliminary results, based only on elementary calculus.

LEMMA 1.1. *For* $n := mr+s$, $0<\delta<+\infty$,

$$\frac{v_{r,n}(|x|,|t|)}{n!} \leq \frac{(\delta+|t|)^m}{m!} \frac{|x|^s}{s!} \exp \frac{|x|^r}{r\delta}.$$

Proof. By an induction argument for $\mu=0,1,2,\cdots$, it can be shown that, for fixed $r\geq 2$,

$$(\mu!)^r r^{\mu} \leq (r\mu)!.$$

Recalling that $n!s! \leq (n+s)!$, and using the above inequality with μ replaced by $(m-k)$, we obtain

$$\frac{1}{(r(m-k)+s)!} \leq \frac{1}{(r(m-k))!s!} \leq \frac{r^{k-m}}{s!(m-k)!(m-k)!}.$$

Now, from (0.3), noting that $|z|^n/n! \leq \exp|z|$, we have, for $\delta \neq 0$,

$$\begin{aligned}
\frac{v_{r,n}(|x|,|t|)}{n!} &= \sum_{k=0}^{m} \frac{|t|^k}{k!} \frac{|x|^s |x|^{r(m-k)}}{(r(m-k)+s)!} \\
&= \frac{\delta^m}{m!} \sum_{k=0}^{m} \frac{m!}{k!} \left(\frac{|t|}{\delta}\right)^k \frac{|x|^s}{(r(m-k)+s)!} \left(\frac{|x|^r}{\delta}\right)^{m-k} \\
&\leq \frac{\delta^m}{m!} \left(1+\frac{|t|}{\delta}\right)^m \frac{|x|^s}{s!} \exp \frac{|x|^r}{r\delta}
\end{aligned}$$

and the lemma is proved.An appeal to Stirling's formula yields the inequality

$$\frac{(\delta+|t|)^m}{m!} \leq c(mr+s)^{1/2} \left(\frac{(\delta+|t|)re}{rm+s}\right)^{(mr+s)/r},$$

where c is a constant which depends on δ. We thus obtain the following results:

COROLLARY 1.2. *For* $n=1,2,3,\cdots$, $0<\delta<+\infty$, $|x|<R$

$$\frac{v_{r,n}(|x|,|t|)}{n!} \leq M n^{1/2} \left(\frac{(\delta+|t|)re}{n}\right)^{n/r},$$

where M is a constant depending on δ, r, R.

LEMMA 1.3. *For* $n := mr+s$, $0<(x_0^r/t_0)<+\infty$

$$\frac{|v_{r,n}(x_0,t_0)|}{n!} \geq \frac{|x_0|^s}{s!} \frac{|t_0|^m}{m!}.$$

Proof. From (0.3) we obtain

$$\begin{aligned}
\frac{|v_{r,n}(x_0,t_0)|}{n!} &= \left| \sum_{k=0}^{m} \frac{t_0^k}{k!} \frac{x_0^s x_0^{r(m-k)}}{(r(m-k)+s)!} \right| \\
&= \frac{|t_0|^m}{m!} \frac{|x_0|^s}{s!} \sum_{k=0}^{m} \frac{m!s!}{k!(r(m-k)+s)!} \left(\frac{x_0^r}{t_0}\right)^{m-k}.
\end{aligned}$$

Hence, since the value of the latter sum is at least 1, the proof is complete.

COROLLARY 1.4. *For* $n=1,2,3,\cdots$, $0<(x_0^r/t_0)<+\infty$

$$\frac{|v_{r,n}(x_0,t_0)|}{n!}\geq Mn^{-1/2}\left(\frac{|t_0|re}{n}\right)^{n/r},$$

where M *is a constant depending on* x_0 *and* t_0.

Proof. By Stirling's formula we obtain for $n:=mr+s\geq 1$

$$\frac{|t_0|^m}{m!}\geq C(mr+s)^{-1/2}\left(\frac{|t_0|re}{mr+s}\right)^{(mr+s)/r}=Cn^{-1/2}\left(\frac{|t_0|re}{n}\right)^{n/r},$$

where C is a constant which depends on t_0. Hence the proposition follows easily from Lemma 1.3.

We are now in a position to show that if a polynomial series (1.0) converges at a point (x_0,t_0) with $0<(x_0^r/t_0)<+\infty$, then it converges in a strip $Z_\sigma:=\{(x,t)\in\mathbb{K}^2: |t|<\sigma\}$.

THEOREM 1.5. *If the series* (1.0) *converges at* (x_0,t_0), $0<(x_0^r/t_0)<+\infty$, *then*

i) *the series converges absolutely in the strip* $Z_{|t_0|}$;

ii) *the series converges uniformly in any compact region of* $Z_{|t_0|}$;

iii)

$$a_n=o\left(n^{1/2}\left(\frac{n}{re|t_0|}\right)^{n/r}\right),\qquad n\to+\infty.$$

Proof. Since the series (1.0) is assumed to converge at (x_0,t_0), the general term tends to zero as $n\to\infty$. It follows that by Corollary 1.4

$$\left|a_nMn^{-1/2}\left(\frac{|t_0|re}{n}\right)^{n/r}\right|\to 0,\qquad n\to+\infty,$$

and hence

$$|a_n|\leq cn^{1/2}\left(\frac{n}{|t_0|re}\right)^{n/r},\qquad n\geq 1$$

for some constant c which depends on x_0 and t_0. By use of Corollary 1.2 we have for $|x|<R$

$$\left|\frac{a_n}{n!}v_{r,n}(x,t)\right|\leq Cn\left(\frac{\delta+|t|}{|t_0|}\right)^{n/r},\qquad n\geq 1,$$

where C is a constant which depends on x_0, t_0, δ and R. But the series

$$\sum_{n=1}^{\infty} n\left(\frac{\delta+|t|}{|t_0|}\right)^{n/r}$$

converges for $|t|<|t_0|-\delta$. Since δ may be taken arbitrarily small, an application of Weierstrass M-test completes the proof.

There exist series (1.0) some of which converge for all (x,t), while others fail to converge when $t\neq 0$. We define a number σ as the radius of convergence, if the polynomial series (1.0) converges in a strip Z_σ. Because the limiting case $\sigma=0$ is not interesting in further studies, we omit it, but $\sigma=+\infty$ can be included.

THEOREM 1.6. *If*

$$(*) \qquad \limsup_{n\to+\infty} \frac{re}{n}|a_n|^{r/n}=\frac{1}{\sigma}<+\infty,$$

then the series

$$\sum_{n=0}^{\infty} \frac{a_n}{n!} v_{r,n}(x,t)$$

has the following properties:

i) *if it converges in the strip* Z_τ, *then* $\tau\le\sigma$;

ii) *it converges absolutely in the strip* Z_σ;

iii) *if it converges absolutely at some point* (x_0,t_0), $x_0\neq 0$, *then* $(x_0,t_0)\in\overline{Z}_\sigma$;

iv) *it converges uniformly in any compact region of* Z_σ.

Proof. If $0<\sigma_0<\sigma$, then the assumption $(*)$ implies that, for $n\ge n(\sigma_0)$,

$$|a_n|\le\left(\frac{n}{re\sigma_0}\right)^{n/r}.$$

Hence by Corollary 1.2 we have, for $|x|<R$,

$$\left|\frac{a_n}{n!} v_{r,n}(x,t)\right|\le M n^{1/2}\left(\frac{\delta+|t|}{\sigma_0}\right)^{n/r}, \qquad n\ge 1.$$

But the series

$$\sum_{n=1}^{\infty} n^{1/2}\left(\frac{\delta+|t|}{\sigma_0}\right)^{n/r}$$

converges for $|t|<\sigma_0-\delta$. Consequently, for δ arbitrarily near 0 and σ_0 arbitrarily near σ, we have ii) and iv). Now, suppose that the series (1.0) converges in the whole strip Z_τ where $\sigma<\tau$. Then in particular it would converge at (x_0,t_0), $0<(x_0^r/t_0)<+\infty$, $\sigma<|t_0|<\tau$. By Theorem 1.5, we obtain

$$\limsup_{n\to\infty} \frac{re}{n}|a_n|^{r/n}\le\frac{1}{|t_0|}.$$

The desired contradiction is evident and i) is proved. The property iii) follows in an obvious way from Theorem 2.4.

It should be noted that Theorem 1.6 does not preclude the possibility of convergence outside the strip Z_σ. This may happen, as in the example

$$(1.7) \qquad n:=mr+s, \qquad a_n=\begin{cases} 0, & s=0, \\ \dfrac{(rm+s)!}{((r-1)m)!}, & s\neq 0. \end{cases}$$

By Stirling's formula

$$\left(\frac{(rm+s)!}{((r-1)m)!}\right)^{r/rm+s}\sim\frac{rm+s}{e}\left(\frac{r}{r-1}\right)^{r-1},$$

so that

$$\limsup_{n\to\infty} \frac{re}{n}|a_n|^{r/n}=r\left(\frac{r}{r-1}\right)^{r-1}.$$

The strip of convergence is bounded, but by (0.6) the series (1.0) converges over the whole t-axis.

2. Polynomial expansions. In (0.11) we have noticed that the polynomials $v_{r,n}(x,t)$ satisfy the partial differential equation

$$D_x^r u(x,t) = D_t u(x,t). \tag{2.0}$$

We therefore would expect an analogous result for the polynomial expansion. The precise answer is described in Lemma 2.1. Furthermore, we will see that polynomial series are related to Maclaurin double series. This relationship allows polynomial series to be used in solving analytic Cauchy problems of (2.0) as we do in Corollary 2.5 and Corollary 2.6.

LEMMA 2.1. *If*

i) $$u_r(x,t) := \sum_{n=0}^{\infty} \frac{a_n}{n!} v_{r,n}(x,t),$$

ii) $$\limsup_{n\to\infty} \frac{re}{n} |a_n|^{r/n} = \frac{1}{\sigma} < +\infty,$$

then $u_r(x,t)$ satisfies equation (2.0) *and is analytic, as a function of two variables, in the strip Z_σ. The coefficients a_n have the determination*

iii) $$a_n = D_x^n u_r(0,0).$$

Proof. By Theorem 1.6 the series in i) converges uniformly in any compact region of Z_σ. Hence $u_r(x,t)$ is an analytic function in the whole strip Z_σ. Setting $t=0$

$$\sum_{n=0}^{\infty} \frac{a_n}{n!} v_{r,n}(x,0) = \sum_{n=0}^{\infty} \frac{a_n}{n!} x^n,$$

where we have used equation (0.5). Then iii) follows from Taylor's formula. For fixed integer p

$$\limsup_{n\to\infty} \frac{re}{n} |a_{n+p}|^{r/n} = \frac{1}{\sigma},$$

which follows easily from ii). By applying Theorem 1.6 the series

$$\sum_{n=0}^{\infty} \frac{a_{n+p}}{n!} v_{r,n}(x,t)$$

converges uniformly in any compact region of Z_σ. Consequently, appealing to equations (0.9) and (0.10), we note that the differential equation (2.0) holds in the strip Z_σ. In the following, we show that the expansion i) holds in some infinite strip $|t|<\sigma$ if and only if $u_r(x,t)$ is analytic, as a function of two variables, at some point of the x-axis and satisfies the differential equation (2.0).

THEOREM 2.2. *Under the conditons of Lemma* 2.1 $u_r(x,t)$ *has the Maclaurin expansion*

$$u_r(x,t) = \sum_{m,n=0}^{\infty} a_{m+rn} \frac{x^m}{m!} \frac{t^n}{n!}.$$

Proof. From the hypotheses, it follows that the series

$$\sum_{k=0}^{\infty} \frac{a_k}{k!} v_{r,k}(|x|, |t|)$$

converges absolutely in the strip Z_σ. Hence by (0.2)

$$\sum_{k=0}^{\infty} a_k \sum_{m+rn=k} \frac{x^m}{m!}\frac{t^n}{n!} = \sum_{m,n=0}^{\infty} a_{m+rn}\frac{x^m}{m!}\frac{t^n}{n!}.$$

and the result is established.

THEOREM 2.3. *If* $\rho>0$, $\sigma>0$, $|x|<\rho$, $|t|<\sigma$

i) $$D_x^r u_r(x,t)=D_t u_r(x,t), \qquad u_r(x,t)=\sum_{m,n=0}^{\infty} a_{m,n}\frac{x^m}{m!}\frac{t^n}{n!},$$

then $u_r(x,t)$ *can be extended to an analytic function* $U_r(x,t)$ *in the strip* Z_σ,

ii) $$U_r(x,t)=\sum_{k=0}^{\infty}\frac{a_{k,0}}{k!}v_{r,k}(x,t).$$

Proof. Since a power series may be differentiated term by term, we have for $|x|<\rho$, $|t|<\sigma$

$$D_t u_r(x,t)k=\sum_{m,n=0}^{\infty} a_{m,n+1}\frac{x^m}{m!}\frac{t^n}{n!},$$

$$D_x^r u_r(x,t)=\sum_{m,n=0}^{\infty} a_{m+r,n}\frac{x^m}{m!}\frac{t^n}{n!}.$$

Comparing the coefficients we see that $a_{m,n}=a_{m+rn,0}$. Since the series in ii) converges for $|x|<\rho$, $|t|<\sigma$ (and hence absolutely for $0<x_0<\rho$, $0<t_0<\sigma$) we have

$$u_r(x_0,t_0)=\sum_{m,n=0}^{\infty} a_{m+rn,0}\frac{x_0^m}{m!}\frac{t_0^n}{n!}=\sum_{k=0}^{\infty} a_{k,0}\sum_{m+rn=k}\frac{x_0^m}{m!}\frac{t_0^n}{n!}=\sum_{k=0}^{\infty}\frac{a_{k,0}}{k!}v_{r,k}(x_0,t_0).$$

But by Theorem 1.5 the latter series converges absolutely and uniformly in any compact region of the strip $Z_{|t_0|}$.

Since t_0 may be taken arbitrarily near σ, we obtain an analytic function $U_r(x,t)$ which is the continuation of $u_r(x,t)$. This completes the proof.

Now we have, from the expansions in Theorem 2.2 and Theorem 2.3ii, that the series

$$\sum_{k=0}^{\infty}\frac{a_k}{k!}v_{r,k}(x,t) \quad \text{converges in a strip } Z_\sigma$$

if and only if

$$\sum_{m,n=0}^{\infty} a_{m+rn}\frac{x^m}{m!}\frac{t^n}{n!} \quad \text{converges in a strip } Z_\sigma.$$

THEOREM 2.4. *If*

(*) $$\limsup_{n\to\infty}\frac{re}{n}|a_n|^{r/n}=\frac{1}{\sigma}<+\infty,$$

then the series

$$\sum_{m,n=0}^{\infty} a_{m+rn}\frac{x^m}{m!}\frac{t^n}{n!}$$

converges in the strip Z_σ *and, except perhaps when* $x=0$, *diverges for* $|t|>\sigma$.

Proof. Suppose the Maclaurin series converges at some point (x_0, t_0) with $x_0 \neq 0$, $|t_0| > \sigma$. Then, in particular, the series made up of the first r columns would converge absolutely so that

$$\frac{|x_0|^s}{s!} \sum_{n=0}^{\infty} \left| a_{s+rn} \frac{t_0^n}{n!} \right| < +\infty,$$

whence

$$\limsup_{n \to \infty} \frac{e}{n} |a_{s+rn}|^{1/n} \leq \frac{1}{|t_0|}.$$

Combining these r inequalities gives

$$\limsup_{n \to \infty} \frac{re}{n} |a_n|^{r/n} \leq \frac{1}{|t_0|} < \frac{1}{\sigma}.$$

But this contradicts the condition of convergence for the related polynomial series. An example of a Maclaurin double series, which converges in a bounded strip Z_σ but over the whole t-axis, is given by (1.7). Theorem 2.4 proves also Theorem 1.6iii. Under the conditions of Theorem 1.6, suppose the polynomial series also converges absolutely at some point (x_0, t_0), $x_0 \neq 0$, $|t_0| > \sigma$. Then the related Maclaurin series would converge absolutely at the same point, but this contradicts (*).

Now, we can apply our results to solve the analytic Cauchy problems of (2.0). First let us sum the Maclaurin series of Theorem 2.2 by rows. For that we need to define the growth of an entire function [1, p. 11]. We say that an entire function

$$f(x) = \sum_{n=0}^{\infty} \frac{a_n}{n!} x^n$$

has growth $\{\alpha, \beta\}$ if and only if

$$\limsup_{n \to \infty} \left(\frac{e}{n} \right)^{\alpha - 1} |a_n|^{\alpha/n} \leq \alpha\beta.$$

COROLLARY 2.5. *For $\sigma > 0$ and $f(x) := \sum_{m=0}^{\infty} (a_m/m!) x^m$:*

i) *If f has growth*

$$\left\{ \frac{r}{r-1}, \frac{r-1}{r} \left(\frac{1}{r\sigma} \right)^{1/(r-1)} \right\},$$

then

$$u_r(x,t) := \sum_{n=0}^{\infty} D_x^{rn} f(x) \frac{t^n}{n!}$$

is defined in Z_σ and satisfies $D_x^r u(x,t) = D_t u(x,t)$, $u(x,0) = f(x)$.

ii) *If*

$$u_r(x,t) = \sum_{m,n=0}^{\infty} a_{m+rn} \frac{x^m}{m!} \frac{t^n}{n!}$$

converges in Z_σ, then $u_r(x,0) = f(x)$ has growth

$$\left\{ \frac{r}{r-1}, \frac{r-1}{r} \left(\frac{1}{r\sigma} \right)^{1/(r-1)} \right\}.$$

Proof. The corollary follows immediately from the previous results and the equivalence of

$$\limsup_{n\to\infty}\frac{re}{n}|a_n|^{r/n}\leq\frac{1}{\sigma},$$

and

$$\limsup_{n\to\infty}\left(\frac{e}{n}\right)^{1/(r-1)}|a_n|^{r/(n(r-1))}\leq\left(\frac{1}{r\sigma}\right)^{1/(r-1)}.$$

Now we sum the Maclaurin double series by columns.

COROLLARY 2.6. *For* $s=0,\cdots,r-1$, $g_s(t):=\sum_{n=0}^{\infty}(a_{rn+s}/n!)t^n$:

i) *If* $g_s(t)$ *converges for* $|t|<\sigma_s$, $\sigma_s>0$, *then*

$$u_r(x,t):=\sum_{s=0}^{r-1}\sum_{m=0}^{\infty}D_t^m g_s(t)\frac{x^{rm+s}}{(rm+s)!}$$

is defined in Z_σ, $\sigma=\min\{\sigma_s : s=0,\cdots,r-1\}$, *and satisfies*

$$D_x^r u_r(x,t)=D_t u_r(x,t),\qquad D_x^s u_r(0,t)=g_s(t).$$

ii) *If*

$$u_r(x,t):=\sum_{m,n=0}^{\infty}a_{m+rn}\frac{x^m}{m!}\frac{t^n}{n!}$$

converges in Z_σ, $\sigma>0$, *then* $D_x^s u(0,t)=g_s(t)$ *converges for* $|t|<\sigma_s$, $\sigma_s\geq\sigma$.

Proof. The corollary is an immediate consequence of the previous results and the elementary relation

$$\limsup_{n\to\infty}\frac{re}{n}|a_n|^{r/n}\leq\frac{1}{\sigma}$$

if and only if for $\sigma_s\geq\sigma$, $s=0,\cdots,r-1$

$$\limsup_{m\to\infty}\frac{e}{m}|a_{rm+s}|^{1/m}\leq\frac{1}{\sigma_s}.$$

In [12] D. V. Widder pointed out that in general the classic Cauchy-Kowalewski solution of Corollary 2.6 ii) is known to be valid only in a sufficiently small neighborhood of the t-axis. For the heat equation he had obtained a solution "in the large" by using the special version ($r=2$) of Corollary 2.6. Now we have the analogous result for all partial differential equations (2.0).

REFERENCES

[1] R. P. BOAS, *Entire Functions*, Academic Press, New York, 1954.

[2] L. R. BRAGG, *The radial heat polynomials and related functions*, Trans. Amer. Math. Soc., 119 (1965), pp. 270–290.

[3] F. M. CHOLEWINSKI AND D. T. HAIMO, *Expansions in terms of Laguerre heat polynomials and of their Appell transforms*, J. Analyse Math., 24 (1971), pp. 285–322.

[4] D. T. HAIMO, *Expansions in terms of generalized heat polynomials and of their Appell transforms*, J. Math. Mech., 15 (1966), pp. 735–758.

[5] ———, *Series expansions of generalized temperature functions in n dimensions*, Canad. J. Math., 18 (1966), pp. 794–802.

[6] _____, *Series representations of generalized temperature functions*, SIAM J. Appl. Math., 15 (1967), pp. 359–367.
[7] _____, *Series expansions for dual Laguerre temperatures*, Canad. J. Math. 24 (1972), pp. 1145–1153.
[8] C. Y. Lo AND E. LANSING, *Polynomial expansions of solutions of* $u_{xx}+\varepsilon^2 u_{tt}=u_t$, J. Reine Angew. Math., 252 (1972), pp. 88–103.
[9] C. Y. Lo, *Series expansions of solutions of* $u_{xx}+u_{yy}+\varepsilon^2 u_{tt}=u_t$, this Journal, 3 (1972), pp. 461–473.
[10] P. C. ROSENBLOOM AND D. V. WIDDER, *Expansions in terms of heat polynomials and associated functions*, Trans. Amer. Math., Soc., 92 (1959), pp. 220–266.
[11] D. V. WIDDER, *Series expansions of solutions of the heat equation in n dimensions*, Ann. Mat. Pura Appl., (4) 55 (1961), pp. 389–409.
[12] _____, *Analytic solutions of the heat equation*, Duke Math. J., 29 (1962), pp. 497–503.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0010 $01.00/0

A SET OF HYPERGEOMETRIC ORTHOGONAL POLYNOMIALS*

RICHARD ASKEY† AND JAMES WILSON‡

Abstract. Polynomials orthogonal with respect to $|\Gamma(\alpha+ix)\Gamma(\gamma+ix)|^2$ on $(-\infty,\infty)$ are found.

1. Introduction. At the Helsinki Congress (1978), E. M. Nikishin [6] pointed out the importance of adding to the classical orthogonal polynomials other polynomials that are orthogonal to specific weight functions and finding explicit expressions for these polynomials and their three-term recurrence relation. One of us has given a large class of such polynomials and their hypergeometric representations [9], and jointly we have given the discrete case of even more general orthogonal polynomials which are basic hypergeometric series [1]. The particular weight function Nikishin proposed was $[\exp(2\pi\sqrt{x})-1]^{-1}$ on $[0,\infty)$. We will not solve this problem, but remark that some of the polynomials given by Wilson are orthogonal with respect to

$$w(x)=\frac{\exp\pi\sqrt{x}}{\exp\left(2\pi\sqrt{x}\right)-1}=\frac{1}{2\sinh\pi\sqrt{x}}$$

on $[0,\infty)$. This is the case $a=1$, $b=c=\frac{1}{2}$, $d=0$, after a change of variables. Wilson's polynomials will be given in the next section.

There is another interesting set of orthogonal polynomials which are hypergeometric functions that can be obtained from the polynomials in [9]. To find them, recall the Hahn polynomials

$$(1.1)\qquad Q_n(x;\alpha,\beta,N)={}_3F_2\left(\begin{matrix}-n,\ n+\alpha+\beta+1,\ -x\\ \alpha+1,\ -N\end{matrix};1\right),\qquad n=0,1,\cdots,N.$$

They are orthogonal on $x=0,1,\cdots,N$;

$$(1.2)\qquad \sum_{x=0}^{N} Q_n(x;\alpha,\beta,N)Q_m(x;\alpha,\beta,N)\binom{x+\alpha}{x}\binom{N-x+\beta}{N-x}=0,\qquad m\neq n.$$

A few years ago, W. Al-Salam asked one of us if all the orthogonality relations for these polynomials had been found. It is reasonably easy to argue that they have been. Here is the argument. If a set of polynomials $\{p_n(x)\}$ is orthogonal with respect to a positive measure, then

$$(1.3)\qquad xp_n(x)=A_n p_{n+1}(x)+B_n p_n(x)+C_n p_{n-1}(x),$$

with A_n, B_n, C_n real and $A_{n-1}C_n>0$, $n=1,2,\cdots$. For the Hahn polynomials

$$A_n=-\frac{(N-n)(n+\alpha+\beta+1)(n+\alpha+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)},\qquad C_n=-\frac{n(n+\beta)(n+N+\alpha+\beta+1)}{(2n+\alpha+\beta)(2n+\alpha+\beta+1)}.$$

This is an easy limiting case of the recurrence relation in [1], and can be found in many other sources. If N is not an integer, then $\lim_{n\to\infty}A_{n-1}C_n/n^2=-\frac{1}{16}$, so $A_{n-1}C_n>0$, $n=1,2,\cdots$, cannot hold. However Al-Salam knew that this was not the complete

*Received by the editors July 16, 1980, and in revised form June 8, 1981.

†Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706. The work of this author was supported in part by the National Science Foundation under grant MCS 78-07244A02.

‡Department of Mathematics, Iowa State University, Ames, Iowa 50011.

answer, since Bateman [2], Touchard [8] and Carlitz [3] had found other cases where there was orthogonality with respect to a positive measure. For

$$p_n(x)=i^n {}_3F_2\left(\begin{matrix}-n, n+1, \gamma+ix \\ 1, 2\gamma\end{matrix}; 1\right), \qquad 0<\gamma<1,$$

Carlitz [3] found an orthogonality relation equivalent to

$$\int_{-\infty}^{\infty} \frac{p_n(x)p_m(x)\,dx}{\sin^2\pi\gamma\cosh^2\pi x+\cos^2\pi\gamma\sinh^2\pi x}=0, \qquad m\neq n.$$

The case $\gamma=\frac{1}{2}$ was given earlier by Bateman [2] and Touchard [8].

It is not obvious that $p_n(x)$ is real, but it is. This fact will follow from more general results in the next section.

2. The orthogonal polynomials. To discover the more general set of orthogonal polynomials observe that the weight function for the Hahn polynomials is symmetric about the midpoint $x=N/2$ when $\alpha=\beta$. When this is the case the coefficient B_n in the recurrence relation (1.3) must be a constant, and so it can be removed by translation. After this translation the recurrence relation becomes

$$xp_n(x)=A_n p_{n+1}(x)+C_n p_{n-1}(x), \tag{2.1}$$

and $p_n(x)$ satisfies

$$p_n(-x)=(-1)^n p_n(x).$$

Set

$$q_n(x)=i^n p_n(ix),$$

so that (2.1) becomes

$$-xq_n(x)=A_n q_{n+1}(x)-C_n q_{n-1}(x). \tag{2.2}$$

If in (2.1) A_n and C_n are real and $A_{n-1}C_n<0$, then in (2.2) $A_{n-1}(-C_n)>0$. By a theorem usually attributed to Favard [5, Chapt. II, Thm. 1.5] (see Stone [7, Thm. 10.27] for an earlier treatment than Favard's, and even earlier ones exist) the polynomials $\{q_n(x)\}$ are orthogonal with respect to a positive measure. In the current case, set

$$P_n(x;\alpha,\gamma)=i^n {}_3F_2\left(\begin{matrix}-n, n+2\alpha+2\gamma-1, \gamma-ix \\ \alpha+\gamma, 2\gamma\end{matrix}; 1\right). \tag{2.3}$$

The recurrence relation (1.3) becomes (2.1) with

$$A_n=\frac{(n+2\gamma)(n+2\alpha+2\gamma-1)}{2(2n+2\alpha+2\gamma-1)}, \qquad C_n=\frac{n(n+2\alpha-1)}{2(2n+2\alpha+2\gamma-1)}. \tag{2.4}$$

Now $A_{n-1}C_n>0$ when $\alpha>0$, $\gamma>0$. The case $\alpha+\gamma=1$ was treated by Carlitz.

To find the weight function in the general case recall the quadratic transformation [4, 2.1.5(27)]

$${}_2F_1\left(\begin{matrix}2a, 2b \\ a+b+\frac{1}{2}\end{matrix}; t\right)={}_2F_1\left(\begin{matrix}a, b \\ a+b+\frac{1}{2}\end{matrix}; 4t(1-t)\right). \tag{2.5}$$

In general this only holds in the connected component of the set where $|t|<1$ and $|4t(1-t)|<1$ containing the point $t=0$. When $a=-n$, it holds for all t, since both sides are polynomials. Take $a=-n$ in (2.5), multiply by $t^{c-1}(1-t)^{d-c-1}$ and integrate over

[0, 1]. The result is

$$(2.6)\qquad {}_3F_2\left(\begin{matrix}-2n, 2b, c\\ b-n+\frac{1}{2}, d\end{matrix}; 1\right) = {}_4F_3\left(\begin{matrix}-n, b, c, d-c\\ b-n+\frac{1}{2}, d/2, (d+1)/2\end{matrix}; 1\right).$$

For $P_n(x;\alpha,\gamma)$ this gives

$$(2.7)\qquad P_{2n}(x;\alpha,\gamma) = (-1)^n {}_4F_3\left(\begin{matrix}-b, n+\alpha+\gamma-\frac{1}{2}, \gamma-ix, \gamma+ix\\ \alpha+\gamma, \gamma, \gamma+\frac{1}{2}\end{matrix}; 1\right).$$

Wilson's orthogonality in [9] is

$$(2.8)\qquad \int_0^\infty p_n(x^2;a,b,c,d)p_m(x^2;a,b,c,d) \cdot \left|\frac{\Gamma(a+ix)\Gamma(b+ix)\Gamma(c+ix)\Gamma(d+ix)}{\Gamma(2ix)}\right|^2 dx = 0, \qquad m\neq n,$$

when $a,b,c,d>0$. This also holds when one of the parameters is zero. The polynomial $p_n(x^2)$ is defined by

$$(2.9)\qquad p_n(x^2;a,b,c,d) = {}_4F_3\left(\begin{matrix}-n, n+a+b+c+d-1, a+ix, a-ix\\ a+b, a+c, a+d\end{matrix}; 1\right).$$

Take $c=0$, $d=\frac{1}{2}$, $a=\gamma$, $b=\alpha$. Then

$$(2.10)\qquad \int_{-\infty}^\infty p_n(x;\alpha,\gamma)p_m(x;\alpha,\gamma)|\Gamma(\alpha+ix)\Gamma(\gamma+ix)|^2 dx = 0, \qquad m\neq n,$$

when m and n are even. When one is even and the other is odd the integral vanishes by symmetry, since $P_n(-x)=(-1)^n P_n(x)$.

To complete the proof use the quadratic transformation [4, 2.11(19)].

$$ {}_2F_1\left(\begin{matrix}2a, 2b\\ a+b+\frac{1}{2}\end{matrix}; t\right) = (1-2t)\,{}_2F_1\left(\begin{matrix}a+\frac{1}{2}, b+\frac{1}{2}\\ a+b+\frac{1}{2}\end{matrix}; 4t(1-t)\right)$$

and integrate to obtain

$$(2.11)\qquad {}_3F_2\left(\begin{matrix}2a, 2b, c\\ a+b+\frac{1}{2}, d\end{matrix}; 1\right) = \frac{(d-2c)}{d}\,{}_4F_3\left(\begin{matrix}a+\frac{1}{2}, b+\frac{1}{2}, c, d-c\\ a+b+\frac{1}{2}, (d+1)/2, (d+2)/2\end{matrix}; 1\right)$$

when both series terminate. This is the same as

$$(2.12)\qquad P_{2n+1}(x;\alpha,\gamma) = \frac{(-1)^{n+1}x}{\gamma}\,{}_4F_3\left(\begin{matrix}-n, n+\alpha+\gamma+\frac{1}{2}, \gamma-ix, \gamma+ix\\ \alpha+\gamma, \gamma+\frac{1}{2}, \gamma+1\end{matrix}; 1\right)$$

This time take $a=\gamma$, $b=\alpha$, $c=\frac{1}{2}$, $d=1$. Wilson's orthogonality relation (2.8) gives (2.10) when m and n are odd.

The L^2-norm of the polynomials can be computed from Wilson's results [9], or by use of the recurrence relation (2.1), (2.4). It is

$$(2.13)\qquad \int_{-\infty}^\infty [P_n(x;\alpha,\gamma)]^2|\Gamma(\alpha+ix)\Gamma(\gamma+ix)|^2 dx = \frac{(1)_n(2\alpha)_n(\alpha+\gamma-\frac{1}{2})_n}{(2\gamma)_n(2\alpha+2\gamma-1)_n(\alpha+\gamma+\frac{1}{2})_n}A,$$

where

$$A = \int_{-\infty}^\infty |\Gamma(\alpha+ix)\Gamma(\gamma+ix)|^2 dx = \frac{\Gamma(\alpha)\Gamma(\alpha+\frac{1}{2})\Gamma(\gamma)\Gamma(\gamma+\frac{1}{2})\Gamma(\alpha+\gamma)\Gamma(\frac{1}{2})}{\Gamma(\alpha+\gamma+\frac{1}{2})}.$$

There is a second way to obtain this result from (2.8). It is not necessary to take $a,b,c,d>0$ in (2.8). They can be complex with positive real parts, provided they occur in complex conjugate pairs. Take $a=\gamma+iw$, $b=\gamma-iw$, $c=\alpha+iw$, $d=\alpha-iw$, $x=t+w$. To keep the polynomials real it suffices to consider

$$(a+b)_n(a+c)_n(a+d)_n\,{}_4F_3\left(\begin{matrix}-n, n+a+b+c+d-1, a+ix, a-ix\\ a+b, a+c, a+d\end{matrix};1\right).$$

Divide by $(2w)^n$ and let $w\to\infty$. The resulting polynomial is

$$(2\gamma)_n(\alpha+\gamma)_n i^n\,{}_3F_2\left(\begin{matrix}-n, n+2\alpha+2\gamma-1, \gamma-it\\ \alpha+\gamma, 2\gamma\end{matrix};1\right),$$

so the orthogonality relations (2.10) and (2.13) follow directly from Wilson's orthogonality without the use of quadratic transformations. We gave the first derivation since the method can be used in other settings. Also it seemed to explain the real reason for the existence of these orthogonal polynomials better than the second argument did.

There are special cases whose weight functions are interesting. When $\alpha=\frac{1}{2}$, $\gamma=1$, the weight function is a constant multiple of $x/\sinh 2\pi x$ on $(-\infty,\infty)$. Other choices of α and γ give $x\prod_{j=1}^{k}(x^2+j^2)\prod_{j=1}^{l}(x^2+(j-\frac{1}{2})^2)/\sinh 2\pi x$ and $x^2\prod_{j=1}^{k}(x^2+j^2)\prod_{j=1}^{l}(x^2+j^2)/\sinh^2\pi x$. There are similar extensions of the orthogonality relations of Bateman and Carlitz which can be given if they are needed. However the symmetry and simplicity of the weight function in (2.10) is so nice that it will probably be a more useful form than most of its special cases.

This set of polynomials is probably the only one of the ${}_3F_2$ Hahn polynomials which is orthogonal with respect to a positive absolutely continuous weight function, but we have not shown this.

3. Another set of orthogonal polynomials. The polynomials p_n defined in (2.9) are polynomials of degree n in x^2, and so of degree $2n$ in x. It is natural to ask if it is possible to explicitly find polynomials of degree $2n+1$, $n=0,1,\cdots$, so that the full set of polynomials are orthogonal on $(-\infty,\infty)$ with respect to the weight function in (2.8). The answer is yes. When one of the parameters vanishes (say $d=0$) the polynomials can be given as hypergeometric series:

$$p_{2n+1}(x;a,b,c)=x\,{}_4F_3\left(\begin{matrix}-n, n+a+b+c, a+ix, a-ix\\ a+b, a+c, a+1\end{matrix};1\right). \tag{3.1}$$

For

$$\int_{-\infty}^{\infty}p_{2n+1}(x;a,b,c)p_{2m+1}(x;a,b,c)\left|\frac{\Gamma(a+ix)\Gamma(b+ix)\Gamma(c+ix)\Gamma(ix)}{\Gamma(2ix)}\right|^2dx$$

$$=2\int_0^{\infty}p_n(x^2;a,b,c,1)p_m(x^2;a,b,c,1)\left|\frac{\Gamma(a+ix)\Gamma(b+ix)\Gamma(c+ix)\Gamma(1+ix)}{\Gamma(2ix)}\right|^2dx$$

and this vanishes when $m\neq n$ by (2.8).

In the general case the best representation we have found for the odd degree polynomials is as a sum of two hypergeometric series. They are given by

$$xq_{2n+1}(x;a,b,c,d)=\frac{p_{n+1}(x^2;a,b,c,d)}{p_{n+1}(0;a,b,c,d)}-\frac{p_n(x^2;a,b,c,d)}{p_n(0;a,b,c,d)}.$$

Clearly $q_{2n+1}(x)$ is a polynomial of degree $2n+1$ in x and only odd powers of x appear, so it is sufficient to show that

$$\int_0^\infty q_{2n+1}(x;a,b,c,d)x^{2k+1} \cdot \left| \frac{\Gamma(a+ix)\Gamma(b+ix)\Gamma(c+ix)\Gamma(d+ix)}{\Gamma(2ix)} \right|^2 dx=0, \qquad k=0,1,\cdots,n-1.$$

This is immediate from (2.8).

REFERENCES

[1] R. ASKEY AND J. WILSON, *A set of orthogonal polynomials that generalize the Racah coefficients or* $6-j$ *symbols*, this Journal, 10 (1979), pp. 1008–1016.

[2] H. BATEMAN, *Functions orthogonal in the Hermitian sense, a new application of basic numbers*, Proc. National Acad. Sci. 20 (1934), pp. 63–66.

[3] L. CARLITZ, *Bernoulli and Euler numbers and orthogonal polynomials*, Duke Math. J., 26 (1959), pp. 1–15.

[4] A. ERDÉLYI, *Higher Transcendental Functions*, vol. 1, McGraw-Hill, New York, 1953.

[5] G. FREUD, *Orthogonal Polynomials*, Pergamon Press, Oxford, 1971.

[6] E. M. NIKISHIN, *The Padé approximants*, Proc. International Congress of Mathematicians, Helsinki, 1978, O. Lehto, ed., vol. 2, Helsinki, 1980, pp. 623–630.

[7] M. H. STONE, *Linear Transformations in Hilbert Space and Their Applications to Analysis*, AMS Colloquium Publications, XV, American Mathematical Society, Providence, R. I., 1932.

[8] J. TOUCHARD, *Nombres exponentiels et nombres de Bernoulli*, Canadian J. Math. 8 (1956), pp. 305–320.

[9] J. WILSON, *Some hypergeometric orthogonal polynomials*, this Journal, 11 (1980), pp. 690–701.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0011 $01.00/0

ORTHOGONAL POLYNOMIALS, DUALITY AND ASSOCIATION SCHEMES*

DOUGLAS A. LEONARD†

Abstract. This paper reconstructs and characterizes the Askey–Wilson orthogonal polynomials as those having duals (in the sense of Delsarte) which are also orthogonal. It introduces the concepts of eigenvalues and Delsarte's duality to the study of orthogonal polynomials and provides those interested in P- and Q-polynomial association schemes with a closed form for their parameters.

In [6] D. Stanton discusses in depth the q-Krawtchouk polymomials and their relation to the classical root systems. In [1] R. Askey and J. Wilson exhibit a more general set of orthogonal polynomials which are basic hypergeometric extensions of the classical orthogonal polynomials.

P. Delsarte, in his work on association schemes [2], gives a notion of duality which in light of the discussion in [5] is applicable to the study of orthogonal polynomials. The duals of the Askey–Wilson polynomials are also orthogonal so it seems reasonable to ask which orthogonal polynomials have orthogonal duals. Under the assumption that there are sufficiently many polynomials to carry out all calculations, it is possible to reconstruct the Askey–Wilson polynomials as essentially the only such.

Despite the caveat against characterization theorems at the end of the Askey–Wilson paper, it is felt that this is a significant result in that it explains some of the importance of the polynomials and of Delsarte's duality. But more than that, it is central to the characterization of P- and Q-polynomial association schemes (see [5]) suggested by E. Bannai and begun by Y. Egawa [3]. For this reason the main theorem (found at the beginning of §2) is stated as a listing of eigenvalues and parameters for association schemes rather than as a characterization of Askey–Wilson polynomials as mentioned above.

1. Preliminaries. Except when otherwise noted, all unconstrained subscripts range over $0, 1, \cdots, N$, with $N=\infty$ allowed.

A sequence $(p_n(y))$ of polynomials is called *graded* if $\deg p_n = n$. A graded sequence $(p_n(y))$ of polynomials will be called *normalized* relative to some function $\mu(x)$ if $p_n(\mu(0))=1$. The function $\mu(x)$ will be called an *eigenfunction* if the $\mu_n=\mu(n)$ are all distinct. The μ_n will be called *eigenvalues*. (Given a sequence of eigenvalues, there are obviously many choices for an eigenfunction. This choice will not be important, as the function is only a notational convenience.) Note that if $(p_n(y))$ is a graded sequence of polynomials and $p_n(\mu(0))\neq 0$, then $(p_n(y)/p_n(\mu(0)))$ is a normalized sequence of polynomials relative to $\mu(x)$.

A normalized sequence $(p_n(y))$ of polynomials will be called *m-recurrent* relative to some eigenfunction $\mu(x)$ if it satisfies an m-term recurrence relation

$$\sum_{i=0}^{m-1} \lambda_{k,i} p_{k+1-i}(\mu(x)) = \mu(x) p_k(\mu(x)), \qquad 0 \leq k \leq N-1, \tag{1.1}$$

* Received by the editors February 9, 1981 and in final revised form July 27, 1981.
† Department of Mathematics, Auburn University, Auburn, Alabama 36849.

with $p_0(\mu(x))=1$, $p_{-1}(\mu(x))=0$ and $\lambda_{k,i}$ constants satisfying $\sum_{i=0}^{m-1}\lambda_{k,i}=\mu(0)$. In particular (see G. Szegö [7] or G. Freud [4]), note that orthogonal polynomials are 3-recurrent; that is, when normalized as described above they satisfy a 3-term recurrence relation

$$b_k p_{k+1}(\mu(x))+a_k p_k(\mu(x))+c_k p_{k-1}(\mu(x))=\mu(x)p_k(\mu(x)). \tag{1.2}$$

The coefficients b_k, a_k and c_k are the connection parameters familiar in the study of distance regular graphs (see [5]). This paper shall only treat 3-recurrence, since it is of importance in the study of orthogonal polynomials. However, it should be noted that the methods generalize easily to treat m-recurrence.

Two normalized sequences $(p_n(y))$ and $(p_n^*(y))$ of polynomials with respective eigenfunctions $\mu(x)$ and $\mu^*(x)$ are said to be *dual* if

$$p_k(\mu_i)=p_i^*(\mu_k^*). \tag{1.3}$$

In general $(p_n(y))$ may have many duals. However, once the two eigenfunctions $\mu(x)$ and $\mu^*(x)$ are chosen the dual sequence is unique, if its exists. This is the consequence of the following.

PROPOSITION 1. *Let $(p_n(y))$ and $(p_n^*(y))$ be dual sequences of normalized polynomials with eigenfunctions $\mu(x)$ and $\mu^*(x)$ respectively. Then there exists a sequence (M_j) with $M_0=1$ such that*

$$\begin{aligned} p_k(y)&=\sum_{j=0}^{k} M_j \prod_{m=0}^{j-1}(\mu_k^*-\mu_m^*)\prod_{m=0}^{j-1}(y-\mu_m),\\ p_i^*(y)&=\sum_{j=0}^{i} M_j \prod_{m=0}^{j-1}(\mu_i-\mu_m)\prod_{m=0}^{j-1}(y-\mu_m^*). \end{aligned} \tag{1.4}$$

$(p_k(y))$ and $(p_i^*(y))$ can be viewed as arising from the series

$$p(k,i)=\sum_{j=0}^{N} M_j \prod_{m=0}^{j-1}(\mu_k^*-\mu_m^*)\prod_{m=0}^{j-1}(\mu_i-\mu_m). \tag{1.5}$$

Proof. Given any graded sequence $(p_n(y))$ of polynomials and any sequence $(\alpha_m y+\beta_m)$ of linear polynomials, it is possible to write

$$p_n(y)=\sum_{j=0}^{n} A_{n,j}\prod_{m=0}^{j-1}(\alpha_m y+\beta_m) \tag{1.6}$$

for some constants $A_{n,j}$ as this can be viewed as a change of basis from $(p_n(y))$ to $(\prod_{m=0}^{j-1}(\alpha_m y+\beta_m))$. Therefore it is possible to write

$$\begin{aligned} p_k(y)&=\sum_{j=0}^{k} A_{k,j}\prod_{m=0}^{j-1}(y-\mu_m),\\ p_i^*(y)&=\sum_{j=0}^{i} A_{i,j}^*\prod_{m=0}^{j-1}(y-\mu_m^*). \end{aligned} \tag{1.7}$$

Since all the eigenvalues and all the dual eigenvalues are distinct,

$$\mu_{k,j}=\prod_{m=0}^{j-1}(\mu_k^*-\mu_m^*)\neq 0 \quad \text{for all } j\leq k,$$
$$\mu_{i,j}^*=\prod_{m=0}^{j-1}(\mu_i-\mu_m)\neq 0 \quad \text{for all } j\leq i. \tag{1.8}$$

So it is possible to rewrite (1.7) as

$$p_k(y)=\sum_{j=0}^{k} M_{k,j}\prod_{m=0}^{j-1}(\mu_k^*-\mu_m^*)\prod_{m=0}^{j-1}(y-\mu_m),$$
$$p_i^*(y)=\sum_{j=0}^{i} M_{i,j}^*\prod_{m=0}^{j-1}(\mu_i-\mu_m)\prod_{m=0}^{j-1}(y-\mu_m^*). \tag{1.9}$$

By duality (1.3),

$$0=\sum_{j=0}^{N}(M_{k,j}-M_{i,j}^*)\prod_{m=0}^{j-1}(\mu_k^*-\mu_m^*)\prod_{m=0}^{j-1}(\mu_i-\mu_m). \tag{1.10}$$

It is easy to show by induction that $M_{k,j}=M_{j,j}^*$ for all $k\geq j$ and that $M_{i,j}^*=M_{j,j}$ for all $i\geq j$. So let $M_j=M_{j,j}=M_{j,j}^*$. And since both sequences are normalized, $M_0=1$.

2. Main theorem. The following theorem is constructive. The form of the proof is discussed after the statement.

THEOREM. *Suppose* $(p_n(y))$ *is a normalized sequence of polynomials with eigenfunction* $\mu(x)$, *satisfying the 3-term recurrence relation*

$$b_k p_{k+1}(\mu(x))+a_k p_k(\mu(x))+c_k p_{k-1}(\mu(x))=\mu(x)p_k(\mu(x)) \tag{2.1}$$

for some constants b_k, a_k, c_k *with* $b_k+a_k+c_k=\mu(0)$. *Suppose* $(p_n^*(y))$ *is a normalized sequence of polynomials with eigenfunction* $\mu^*(x)$, *satisfying*

$$b_k^* p_{k+1}^*(\mu^*(x))+a_k^* p_k(\mu^*(x))+c_k^* p_{k-1}(\mu^*(x))=\mu^*(x)p_k(\mu^*(x)) \tag{2.2}$$

for some constants b_k^*, a_k^*, c_k^* *with* $b_k^*+a_k^*+c_k^*=\mu^*(0)$. *Suppose also that* $(p_n(y))$ *and* $(p_n^*(y))$ *are dual, so*

$$p_k(\mu_i)=p_i^*(\mu_k^*). \tag{2.3}$$

Then the eigenvalues and connection parameters (*and dually the dual eigenvalues and dual connection parameters*) *are given by*:

$$\mu_n=B_0+q^{-n}\frac{q^n-1}{q-1}\left(B_1+\frac{q^{n+1}-1}{q^2-1}B_2\right), \tag{2.4}$$

$$b_k=\frac{\left(B_1^*+B_2^*\dfrac{q^{k+1}-1}{q^2-1}\right)\left(q^{k+2}\dfrac{q^k-1}{q-1}\varepsilon_1-q^{k+2}\dfrac{q^{k-1}-1}{q-1}\varepsilon_0+d_k\right)}{\left(B_1^*+B_2^*\dfrac{q^{2k+2}-1}{q^2-1}\right)\left(B_1^*+B_2^*\dfrac{q^{2k+1}-1}{q^2-1}\right)},$$

where

$$d_k=\frac{(q^k-1)(q^{k-1}-1)}{(q-1)(q^2-1)}\left(B_1B_1^*(1-q^2)+B_1B_2^*+B_2B_1^*+B_2B_2^*\frac{q^{k+2}-1}{q^2-1}\right),$$

$$c_k=\frac{\dfrac{q^k-1}{q^2-1}\left(q^{k+1}\left(B_1^*+B_2^*\dfrac{q^{k+1}-1}{q^2-1}\right)\varepsilon_1-q^k(1+q)\left(B_1^*+B_2^*\dfrac{q^{k+2}-1}{q^2-1}\right)\varepsilon_0+e_k\right)}{\left(B_1^*+B_2^*\dfrac{q^{2k+1}-1}{q^2-1}\right)\left(B_1^*+B_2^*\dfrac{q^{2k}-1}{q^2-1}\right)},$$

where

$$e_k=q(q+1)\left(B_2\frac{q^{k-3}-1}{q^2-1}-B_1q^{k-3}\right)\left(B_1^*+B_2^*\frac{q^{k+1}-1}{q^2-1}\right)\left(B_1^*+B_2^*\frac{q^{k+2}-1}{q^2-1}\right),$$

and

$$a_k=B_0-(b_k+c_k),$$

where it is necessary to take limits in the cases $q=\pm 1$ (*which correspond to repeated roots in the quadratic equation arising from the* 3*-term recurrence in Proposition* 2).

The multiplicities m_i in each case are gotten by

$$m_i=\sum_{j=1}^{i}\frac{b_{j-1}^*}{c_j^*}, \tag{2.5}$$

and the polynomials $p_k(y)$ are given by

$$p_k(y)=\sum_{j=0}^{k}\prod_{m=0}^{j-1}\left(\frac{\mu_k^*-\mu_m^*}{\mu_j^*-\mu_m^*}\right)\prod_{m=0}^{j-1}\left(\frac{y-\mu_m}{b_m}\right). \tag{2.6}$$

The polynomials in (2.6) are those described by Askey–Wilson [1] though the eigenvalues differ by an affine transformation. This does not affect the polynomials but does allow the μ_m to conform with the eigenvalues in the association schemes (hence their names).

The proof is divided into three propositions. The first is used to determine the form of $\nu_n=\mu_n-\mu_{n-1}$ and does not depend on the dual recurrence relation (2.2). (Hence, it holds for any orthogonal polynomials having duals). The second proposition gives a relation between the eigenvalues and the dual eigenvalues. The third pieces together the results of the first two to determine ν_n under the assumptions of the theorem. The theorem follows by solving for the eigenvalues and parameters in each case.

PROPOSITION 2. *Let* $(p_n(y))$ *be a sequence of polynomials satisfying the* 3*-term recurrence relation* (2.1), *with eigenfunctions* $\mu(x)$ *and dual polynomials* $(p_n^*(y))$. *If* $N\geq 9$, *then* (ν_n) *satisfies a recurrence relation*

$$A\nu_n+B\nu_{n-1}+C\nu_{n-2}=0, \qquad 3\leq n\leq N-4 \tag{2.7}$$

for some constants A, B, C *with* $A\neq 0$.

Proof. $p_k(y)$ can be written as in (1.4). Equate coefficients of $\prod_{m=0}^{j-1}(y-\mu_m)$ in (2.1) to get the system of equations

$$b_kM_n\mu_{k+1,n}+a_kM_n\mu_{k,n}+c_kM_n\mu_{k-1,n}=M_{n-1}\mu_{k,n-1}+\mu_nM_n\mu_{k,n} \tag{2.8}$$

for $0\leq k\leq N$, $0\leq n\leq N$, where $\mu_{m,-1}=0$, $\mu_{m,0}=1$, $M_{-1}=0$ and $b_N=0$.

It is now only a matter of judiciously eliminating variables to produce the desired recurrence (2.7).

Let $n=k+1$ in (2.8) to get

$$\varepsilon_k=\frac{M_k}{M_{k+1}}=b_k\frac{\mu_{k+1,k+1}}{\mu_{k,k}}, \qquad 0\le k\le N-1. \tag{2.9}$$

Since $M_0=1$,

$$M_j^{-1}=\prod_{m=0}^{j-1}\varepsilon_m=\left(\prod_{m=0}^{j-1}b_m\right)\mu_{j,j}. \tag{2.10}$$

Substitute this into (2.8) to get

$$b_k\mu_{k+1,n}+a_k\mu_{k,n}+c_k\mu_{k-1,n}=\varepsilon_{n-1}\mu_{k,n-1}+\mu_n\mu_{k,n}, \qquad 0\le n\le N, \quad 0\le k\le N. \tag{2.11}$$

Eliminate a_k from (2.11) to get

(2.12)

$$\begin{aligned} &b_k\mu_{k+1,n-1}(\mu^*_{k+1}-\mu^*_k)+c_k\mu_{k-1,n-1}(\mu^*_{k-1}-\mu^*_k)\\ &\quad=\varepsilon_{n-1}\mu_{k,n-1}-\varepsilon_{n-2}(\mu^*_k-\mu^*_{n-1})\mu_{k,n-2}+\nu_n\mu_{k,n}, \qquad 1\le n\le N, \quad 0\le k\le N. \end{aligned}$$

Eliminate c_k from (2.12) to get

$$\begin{aligned} &b_k\mu_{k+1,n-2}(\mu^*_{k+1}-\mu^*_k)(\mu^*_{k+1}-\mu^*_{k-1})\\ &\quad=\varepsilon_{n-}\ \mu_{k,n-1}-\varepsilon_{n-2}(\mu^*_k+\mu^*_{k-1}-\mu^*_{n-1}-\mu^*_{n-2})\mu_{k,n-2}\\ &\quad\quad+\varepsilon_{n-3}(\mu^*_{k-1}-\mu^*_{n-2})(\mu^*_k-\mu^*_{n-2})\mu_{k,n-3}\\ &\quad\quad+\nu_n\mu_{k,n}-\nu_{n-1}(\mu^*_{k-1}-\mu^*_{n-2})\mu_{k,n-1}, \qquad 2\le n\le N, \quad 0\le k\le N. \end{aligned} \tag{2.13}$$

Eliminate b_k from (2.13) to get

$$\begin{aligned} &f_{k,n}+(\mu^*_{n-2}+\mu^*_{n-3}-\mu^*_{k-1}-\mu^*_{k+1})f_{k,n-1}\\ &\quad+(\mu^*_{k+1}-\mu^*_{n-3})(\mu^*_{k-1}-\mu^*_{n-3})f_{k,n-2}=0 \quad \text{for } 3\le n\le N,\ 0\le k\le N, \end{aligned} \tag{2.14}$$

where

$$f_{k,n}=\left[\nu_n(\mu^*_k-\mu^*_{n-1})(\mu^*_k-\mu^*_{n-2})+\varepsilon_{n-1}(\mu^*_k-\mu^*_{n-2})-\varepsilon_{n-2}(\mu^*_k-\mu^*_{n-1})\right]\mu_{k,n-2}. \tag{2.15}$$

Eliminate μ^*_{k+1} and μ^*_{k-1} from (2.14) to get

$$\begin{vmatrix} f_{k,n-2} & f_{k,n-1}+\mu^*_{n-3}f_{k,n-2} & f_{k,n}+(\mu^*_{n-2}+\mu^*_{n-3})f_{k,n-1}+\mu^{*\,2}_{n-3}f_{k,n-2}\\ f_{k,n-3} & f_{k,n-2}+\mu^*_{n-4}f_{k,n-3} & f_{k,n-1}+(\mu^*_{n-3}+\mu^*_{n-4})f_{k,n-2}+\mu^*_{n-4}f_{k,n-3}\\ f_{k,n-4} & f_{k,n-3}+\mu^*_{n-5}f_{k,n-4} & f_{k,n-2}+(\mu^*_{n-4}+\mu^*_{n-5})f_{k,n-3}+\mu^*_{n-5}f_{k,n-4} \end{vmatrix}=0 \tag{2.16}$$

for $5\le n\le N$, $0\le k\le N$.

Let

$$f_n(y)=\nu_n(y-\mu^*_{n-1})(y-\mu^*_{n-2})+\varepsilon_{n-1}(y-\mu^*_{n-2})-\varepsilon_{n-2}(y-\mu^*_{n-1}), \tag{2.17}$$

where $f_1(y)=\nu_1(y-\mu_0^*)+\varepsilon_0$. Then

$$(2.18)\qquad \begin{vmatrix} f_{n-2}(y) & f_{n-1}(y)(y-\mu_{n-4}^*) & f_n(y)(y-\mu_{n-3}^*)(y-\mu_{n-4}^*) \\ & \quad +\mu_{n-3}^* f_{n-2}(y) & \quad +(\mu_{n-2}^*+\mu_{n-3}^*)f_{n-1}(y)(y-\mu_{n-4}^*) \\ & & \quad +\mu_{n-3}^{*2} f_{n-2}(y) \\ f_{n-3}(y) & f_{n-2}(y)(y-\mu_{n-5}^*) & f_{n-1}(y)(y-\mu_{n-4}^*)(y-\mu_{n-5}^*) \\ & \quad +\mu_{n-4}^* f_{n-3}(y) & \quad +(\mu_{n-3}^*+\mu_{n-4}^*)f_{n-2}(y)(y-\mu_{n-5}^*) \\ & & \quad +\mu_{n-4}^{*2} f_{n-3}(y) \\ f_{n-4}(y) & f_{n-3}(y)(y-\mu_{n-6}^*) & f_{n-2}(y)(y-\mu_{n-5}^*)(y-\mu_{n-6}^*) \\ & \quad +\mu_{n-5}^* f_{n-4}(y) & \quad +(\mu_{n-4}^*+\mu_{n-5}^*)f_{n-3}(y)(y-\mu_{n-6}^*) \\ & & \quad +\mu_{n-5}^{*2} f_{n-4}(y) \end{vmatrix}.$$

is a polynomial of degree at most nine for $5\le n\le N$. But it has factors $y-\mu_k^*$ for all $k\ge n-4$ from (2.16) and $k=n-5$ by inspection. So if there are ten values for k, that is, if $N-n\ge 4$, this must be identically zero. Equate leading coefficients to get

$$(2.19)\qquad \begin{vmatrix} \nu_{n-2} & \nu_{n-1} & \nu_n \\ \nu_{n-3} & \nu_{n-2} & \nu_{n-1} \\ \nu_{n-4} & \nu_{n-3} & \nu_{n-2} \end{vmatrix}=0, \qquad 5\le n\le N-4.$$

Viewed as a system of dependencies among column vectors, this means that the column rank is at most two. So if the first two columns are independent, this yields

$$(2.20)\qquad \begin{vmatrix} \nu_3 & \nu_4 & \nu_n \\ \nu_2 & \nu_3 & \nu_{n-1} \\ \nu_1 & \nu_2 & \nu_{n-2} \end{vmatrix}=0, \qquad 5\le n\le N-4,$$

which is a nondegenerate 3-term recurrence relation. And if the first two columns are dependent, then

$$(2.21)\qquad \begin{vmatrix} \nu_2 & \nu_n \\ \nu_1 & \nu_{n-1} \end{vmatrix}=0, \qquad 5\le n\le N-4,$$

which is a degenerate 3-term recurrence relation. Either yields the result.

PROPOSITION 3. *Suppose* $(p_n(y))$ *and* $(p_n^*(y))$ *satisfy the hypotheses* (2.1), (2.2) *and* (2.3) *of the theorem. Then*

$$(2.22)\qquad (\mu_k^*-\mu_{k-1}^*)(\mu_{k+1}-\mu_{k-2})=(\mu_{k+1}^*-\mu_{k-2}^*)(\mu_k-\mu_{k-1}), \qquad 2\le k.$$

Proof. Let $n=k$ in (2.13) and divide by $\mu_{k,k-1}$ to get

$$(2.23)\qquad \varepsilon_k\frac{\mu_k^*-\mu_{k-1}^*}{\mu_{k+1}^*-\mu_{k-2}^*}=\varepsilon_{k-1}-\varepsilon_{k-3}\frac{\mu_{k-1}^*-\mu_{k-2}^*}{\mu_k^*-\mu_{k-3}^*}+\nu_k\nu_k^*-\nu_{k-1}\nu_{k-1}^*.$$

Note that $\varepsilon_k=\varepsilon_k^*$, and divide (2.23) by its dual equation to get

$$(2.24)\qquad \frac{\mu_k^*-\mu_{k-1}^*}{\mu_{k+1}^*-\mu_{k-2}^*}\frac{\mu_{k+1}-\mu_{k-2}}{\mu_k-\mu_{k-1}}=\frac{\varepsilon_{k-1}-\varepsilon_{k-2}+\varepsilon_{k-3}\dfrac{\mu_{k-1}^*-\mu_{k-2}^*}{\mu_k^*-\mu_{k-3}^*}+\nu_k\nu_k^*-\nu_{k-1}\nu_{k-1}^*}{\varepsilon_{k-1}-\varepsilon_{k-2}+\varepsilon_{k-3}\dfrac{\mu_{k-1}-\mu_{k-2}}{\mu_k-\mu_{k-3}}+\nu_k\nu_k^*-\nu_{k-1}\nu_{k-1}^*}.$$

When $k=2$, the right-hand side is 1. So induct on $k\ge 2$ to get the result.

PROPOSITION 4. *Suppose that* $(p_n(y))$ *and* $(p_n^*(y))$ *satisfy the hypotheses* (2.1), (2.2) *and* (2.3) *of the theorem. Then* (v_n) *and* (v_n^*) *must be one of the following*:

$$\begin{aligned} &1) \quad v_n = Fq^{-n}\left(1+Gq_n^2\right), \qquad v_n^* = F^*q^{-n}\left(1+G^*q_n^2\right) \text{ with } FF^*q \neq 0 \\ &\text{or} \\ &2) \quad v_n = Fq^{-n}(1+Gr_n), \qquad v_n^* = F^*q^{-n}(1+Gr_n) \text{ with } FF^*Gqr \neq 0, \end{aligned} \tag{2.25}$$

where $r_n = \sum_{l=0}^{n-1} r^l$.

Let x_1 and x_2 be the solutions to $Ax^2+Bx+C=0$ from Proposition 2, and let y_1 and y_2 be the solutions to the dual (where x_2 and/or y_2 are zero if this is degenerate (linear)). These can be chosen so that 1) if $x_i = y_j^{-1}$ for any (i,j), then $x_1 = y_1^{-1}$; 2) if $x_i \neq y_j^{-1}$ for any (i,j) but $x_i = y_j$ for some (i,j), then $x_i = y_1$. Then with $q = x_1$, $qr = x_2$, $q^* = y_1$ and $q^*r^* = y_2$ and $z_n = \sum_{l=0}^{n-1} z^l$, from Proposition 2,

$$\begin{aligned} v_n &= Dq^n(1+Er_n), \qquad Dq \neq 0, \quad E \neq r-1, \\ v_n^* &= D^*(q^*)^n(1+E^*r_n^*), \qquad D^*q^* \neq 0, \quad E^* \neq r^*-1. \end{aligned}$$

From Proposition 3, $v_n v_{n+1}^* - v_{n+1}v_n^* = K$, a constant, so

$$\begin{aligned} K = DD^*(qq^*)^n\big[&(q^*-q)+E(q^*r_n - qr_{n+1}) \\ &+E^*(q^*r_{n+1}^* - qr_n^*) + EE^*(q^*r_nr_{n+1}^* - qr_{n+1}r_n^*)\big] \end{aligned} \tag{2.26}$$

$$\begin{aligned} K(1-qq^*) = DD^*(qq^*)^{n+1}\big[&Er^n(q^*-qr) + E^*(r^*)^n(q^*r^*-q) \\ &+EE^*q^*\left(r^n(r^*)^{n+1} + r_n(r^*)^{n+1} + r^nr_{n+1}^*\right) \\ &-EE^*q\left(r^{n+1}(r^*)^n + r^{n+1}r_n^* + r_{n+1}(r^*)^n\right)\big], \end{aligned} \tag{2.27}$$

$$\begin{aligned} K(1-qrq^*r^*)(1-qrq^*)(1-qq^*) \\ = DD^*(qq^*)^{n+3}(r^*)^{n+1}(r^*-r)E^*(1-r+E)(q^*r^*-q), \end{aligned} \tag{2.28}$$

$$\begin{aligned} K(1-qrq^*r^*)(1-qq^*r^*)(1-qq^*) \\ = DD^*(qq^*)^{n+3}r^{n+1}(r-r^*)E(1-r^*+E^*)(qr-q^*), \end{aligned} \tag{2.29}$$

$$K(1-qrq^*r^*)(1-qrq^*)(1-qq^*r^*)(1-qq^*) = 0. \tag{2.30}$$

If $E = E^* = 0$, then from (2.26), $q^* = q$ or q^{-1}.

If $E^* = 0$, $E \neq 0$, then from (2.27) $q^* = qr$ or $(qr)^{-1}$, so by the choice of q, qr, q^*, q^*r^*, $q^* = q$ or q^{-1} and dually, if $E = 0$, $E^* \neq 0$.

So assume $EE^* \neq 0$. From (2.30) and the choice of q, qr, q^*, q^*r^*, $K(1-qq^*) = 0$. So from (2.28) and (2.29) either $r = 4$ or $q^*r^* = q$ and $qr = q^*$. If $r \neq r^*$, then by the choice of q, qr, q^*, q^*r^*, $q^* = q$ or q^{-1}. If $r = r^*$, then from (2.27),

$$0 = E(q^*-qr) + E^*(q^*r-q) + EE^*(q^*-q)(1+r)r_{n+1}, \tag{2.31}$$

so $0 = (q^*-q)(1+r)r_{n+1}$. So $q^* = q$ and $r = 1$ or $E = E^*$, or $r = -1$ and $q^* = -q$ or $E = E^*$.

Proof of theorem. Case 2) of Proposition 4 does not satisfy (2.14) with $n = 3$ unless it reduces to case 1), $\mu_n = \mu_0 + \sum_{j=1}^n v_j$ and dually for μ_n^*. These can be written in the form in the conclusion for some constants B_0, B_1, B_2, B_0^*, B_1^* and B_2^*. In terms of these and $q, \varepsilon_0, \varepsilon_1$, it is posssible to solve for b_k, c_k and a_k using (2.13) with $n = 2$, (2.12) with $n = 1$ and (2.11) with $n = 0$, respectively. (2.5) can be taken as the definition of the multiplicities, and (1.4) and (2.10) give the form of the polynomials.

REFERENCES

[1] R. ASKEY AND J. WILSON, *A set of orthogonal polynomials that generalize the Racah coefficients or 6-j symbols*, this Journal, 10 (1979), pp. 1008–1016.

[2] P. DELSARTE, *An algebraic approach to the association schemes of coding theory*, Thesis, Université Catholique de Louvain, June, 1973, Philips Res. Repts. Suppl. 1973, no. 10.

[3] Y. EGAWA, *Characterization of $H(n, q)$ by the parameters*, J. Combin. Theory. A, to appear.

[4] G. FREUD, *Orthogonal Polynomials*, Pergamon Press, Oxford, 1971.

[5] D. LEONARD, *Parameters of association schemes that are both P- and Q-polynomial*, J. Combin. Theory A, to appear.

[6] D. STANTON, *Some q-Krawtchouk polynomials on Chevalley groups*, Amer. J. Math, 102 (1980), pp. 625–662.

[7] G. SZEGŐ, *Orthogonal Polynomials*, AMS Colloquium Publications 23, American Mathematical Society, Providence, R.I., 1975.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0012 $01.00/0

SERIES OF ORTHOGONAL POLYNOMIALS AS HYPERFUNCTIONS*

AHMED I. ZAYED† AND GILBERT G. WALTER‡

Abstract. The convergence of series of orthogonal polynomials and of associated functions of the second kind in a general setting is studied. Series whose coefficients satisfy $a_n = O(1+\varepsilon_n)^n$ are shown to converge in the sense of hyperfunctions in the first case and to holomorphic functions in the second. The latter are shown to be the analytic representations of the former.

1. Introduction. The location of the regular and singular points of the series

$$\sum_{n=0}^{\infty} a_n p_n(x), \tag{1.1}$$

where $\{p_n(x)\}$ is an orthonormal sequence of polynomials, is important in a number of problems in physics. In particular, in the theory of potential scattering, the $p_n(x)$'s are Legendre polynomials, and the series converges to the "scattering amplitude" [4]. When the $p_n(x)$'s are Gegenbauer polynomials, the series is associated with the "generalized axially symmetric potential theory" [5].

For a very general class of orthogonal polynomials, namely, the Erdös class, the series converges to a function holomorphic inside an ellipse with foci at ± 1 whenever (see [2], [3])

$$\overline{\lim_{n\to\infty}} |a_n|^{1/n} = \rho < 1.$$

For such $\{a_n\}$, (1.1) will be regular at all points in $[-1, 1]$.

Unfortunately, in many of the cases of practical interest, the coefficients do not behave that nicely; often the best one can say is that $\rho \leq 1$. In such cases the series (1.1) at best converges on the real axis, and indeed may not converge at all in any classical sense. However, there is an associated series, the series of functions of the second kind,

$$\sum_{n=0}^{\infty} a_n q_n(z), \tag{1.2}$$

which can sometimes be used to obtain a holomorphic function related to (1.1) [9], [11].

Even though (1.1) does not converge in any of the classical ways such as pointwise or in the mean, there are more general concepts of convergence which could be used to give it a meaning in these cases. It has been shown in [9] that if $a_n = O(n^p)$ for some integer p, then the series (1.1) converges in the sense of generalized functions and the series (1.2) converges to a holomorphic function. Unfortunately, this is still insufficient for a series with, say, $a_n = n^{\log n}$, for which one must use a still more general concept of convergence. This was done for Gegenbauer polynomials in [12].

In this work, we shall investigate the following two related problems:

(i) To find a concept of convergence sufficiently broad to include all series (1.1) for which

$$\overline{\lim_{n\to\infty}} |a_n|^{1/n} \leq 1$$

and $\{p_n\}$ is a member of the Erdös class.

*Received by the editors December 2, 1980.

†California Polytechnic State University, San Luis Obispo, California 93407.
‡University of Wisconsin at Milwaukee, Milwaukee, Wisconsin 53201.

(ii) To find a relation between the holomorphic function given by (1.2) off the real axis and (1.1) on the real axis.

We shall see that an appropriate space in which to attack the first problem is a certain space of hyperfunctions. This will enable us to obtain results similar to those obtained in [7] for trigonometric series and in [12] for series of Gegenbauer polynomials, but with greater generality.

We first present a few of the basic properties of orthogonal polynomials and of hyperfunctions, and then show that, under the conditions stated in (i), the series (1.2) is the analytic representation of (1.1). A characterization of hyperfunctions will be given and then used to study the local behavior of their analytic representations.

2. Some basic properties. In this section, we present some of the known basic properties of both the orthogonal polynomials under consideration and the hyperfunctions. Most properties of the former can be found in the book by Geronimus [2] and those of the latter in the book by Köthe [8].

Let $\{p_n(x)\}$ be a sequence of orthonormal polynomials on the real line with respect to the probability measure arising from the monotone function $\alpha(x)$. That is,

$$\begin{aligned} &\int_{-\infty}^{\infty} p_n(x)p_m(x)\,d\alpha(x)=\delta_{nm}, \qquad n,m=0,1,2,\cdots, \\ &p_0(x)=1. \end{aligned} \tag{2.1}$$

DEFINITION 2.1. The sequence $\{p_n(x)\}$ is said to *belong to Erdös class E* if $d\alpha(x)$ has support in $[-1,1]$ and $\alpha'(x)$ exists and is positive almost everywhere in $[-1,1]$. All such polynomials satisfy a recurrence formula

$$\begin{aligned} &xp_n(x)=\frac{\gamma_n}{\gamma_{n+1}}p_{n+1}(x)+\alpha_n p_n(x)+\frac{\gamma_{n-1}}{\gamma_n}p_{n-1}(x), \qquad n=0,1,2,\cdots, \\ &p_0(x)=1, \qquad p_{-1}(x)=0, \end{aligned} \tag{2.2}$$

and have associated functions of the second kind

$$q_n(z)=\int_{-1}^{1}\frac{p_n(x)}{x-z}d\alpha(x), \qquad z\in\mathbf{C}-[-1,1], \quad n=0,1,2,\cdots. \tag{2.3}$$

The $\{q_n(z)\}$ satisfy the same recurrence relation (2.2) except that

$$q_0(x)=\int_{-1}^{1}\frac{1}{x-z}d\alpha(x), \qquad q_{-1}(x)=-1,$$

The $q_n(z)$ serve as the expansion coefficients of $1/(x-z)$ as well, and we have

$$\frac{1}{x-z}=\sum_{n=0}^{\infty}q_n(z)p_n(x). \tag{2.4}$$

This series converges for $\{p_n(x)\}$ in E whenever x is inside and z is outside an ellipse with foci at -1 and 1 [2].

We now turn to hyperfunctions, give their definition and outline some of their basic properties.

Let $I=[a,b]$ be a bounded interval and define the space $\mathcal{H}(I)$ as follows: The function ϕ belongs to $\mathcal{H}(I)$ if there exists a complex neighborhood $U\supset I$ such that ϕ is holomorphic in U. Two such functions are identified if they agree on some neighborhood of I. For each U as above, let $A(U)$ denote the space of functions holomorphic in U and continuous on $\overline{U}$ (the closure of U), with topology defined by the norm $\|\phi\|_U=\sup_{z\in U}|\phi(z)|$. There is a natural map from $A(U)$ into $\mathcal{H}(I)$.

The topology of $\mathcal{H}(I)$ is defined to be the finest locally convex topology on $\mathcal{H}(I)$ for which all these maps are continuous. Provided with this topology, the space $\mathcal{H}(I)$ becomes a nonmetrizable complete Montel space. Proofs and more properties may be found in Köthe [8] and Johnson [7].

The strong dual $\mathcal{H}'(I)$ of $\mathcal{H}(I)$ is our space of generalized functions, which is usually called the space of hyperfunctions on I. Since each Montel space is reflexive, and the strong dual of the strong dual is the original space and its dual is a Montel space, it follows immediately that $\mathcal{H}'(I)$ is also a Montel space.

The space $\mathcal{H}'(I)$ has a concrete representation which is given as follows: For $f\in H'(I)$, there exists a holomorphic function $\hat{f}(z)$ in $\mathbf{C}-I$ with $\hat{f}(\infty)=0$ defined by $\hat{f}(z)=\langle f, 1/(t-z)\rangle$, where $\mathbf{C}$ is the complex sphere, $t\in I$ and $z\in\mathbf{C}-I$. The function $\hat{f}(z)$ is called the indicatrix of f.

Conversely, each function $\hat{f}(z)$ that is holomorphic in $\mathbf{C}-I$ with $\hat{f}(\infty)=0$ gives rise to a generalized function f in $\mathcal{H}'(I)$ in the following manner: If $\phi\in\mathcal{H}(I)$, then f is given by

$$\langle f,\phi\rangle=\frac{1}{2\pi i}\int_\gamma \hat{f}(z)\phi(z)\,dz,$$

where γ is a path enclosing I lying in the region of analyticity of ϕ.

To guarantee the uniqueness of this representation, we identify any two holomorphic functions in $\mathbf{C}-I$, say f_1 and f_2, if their difference f_1-f_2 is holomorphic in I. Hence, we have a one-to-one correspondence between the space of generalized functions $\mathcal{H}'(I)$ and the space of equivalent classes of holomorphic functions on $\mathbf{C}-I$. In fact, these two spaces are isomorphic [7].

Examples.

1. If $a=b$, i.e., $I=\{a\}$, then the function $(-1/2\pi i)/(z-a)$ defines the Dirac δ-function with support on $\{a\}$.

2. The characteristic function $\chi_{[a,b]}$ of the interval $I=[a,b]$ is defined as a hyperfunction (also denoted by $\chi_{[a,b]}$) by $\log((a-z)/(b-z))\in A(\mathbf{C}-I)$, in which we consider $\log(-z)$ a single-valued function of $z\in\mathbf{C}-[0,\infty)$ which assumes real values for $z\in(-\infty,0)$. In this case, we have

$$\begin{aligned}\langle\chi_{[a,b]},\phi\rangle&=\frac{-1}{2\pi i}\int_\gamma \log\frac{b-z}{a-z}\phi(z)\,dz\\&=\frac{-1}{2\pi i}\int_\gamma\phi(z)\int_a^b\frac{dx}{x-z}dz\\&=\frac{-1}{2\pi i}\int_a^b dx\int_\gamma\frac{\phi(z)}{x-z}dz=\int_a^b\phi(x)\,dx.\end{aligned}$$

3. Let $f\in L^1[a,b]$. The associated hyperfunction (also called f) is given by $\langle f,\phi\rangle=\int_a^b f(x)\phi(x)\,dx$.

4. Let g be continuous and $\alpha(x)$ be of bounded variation on $[a,b]$. The hyperfunction $g\,d\alpha$ is given by $\langle g\,d\alpha,\phi\rangle=\int_a^b g\phi\,d\alpha$.

5. The indicatrix of $p_n\,d\alpha$ considered as a hyperfunction is $q_n(z)$ (see (2.3)).

Now we should be able to exhibit the relationship between the orthogonal polynomials and hyperfunctions.

Throughout the rest of this section, I will denote the interval $[-1,1]$ and the orthogonal polynomials will be in the class E.

THEOREM 2.1. *Let* $\{\gamma_n\}$ *be a sequence of complex numbers such that*

$$\overline{\lim}_{n\to\infty}|\gamma_n|^{1/n}=\frac{1}{R}<1.$$

Then $\phi(x)=\sum_{n=0}^{\infty}\gamma_n p_n(x)$ *belongs to* $\mathcal{H}(I)$.

Conversely, if $\phi(x)\in\mathcal{H}(I)$, *then* ϕ *has an expansion in terms of* $\{p_n(x)\}$ *with coefficients satisfying the above condition.*

Proof. See Geronimus [2, Thms. 9.8, 9.9], of which this is merely a restatement. In fact, the series converges to $\phi(x)$ in the sense of $\mathcal{H}(I)$ and represents a holomorphic function in an ellipse whose foci are at ± 1 and the sum of whose semi-axes is R. □

THEOREM 2.2. (i) *Let* $\{a_n\}$ *be a sequence of complex numbers such that* $\overline{\lim}_{n\to\infty}|a_n|^{1/n}\le 1$. *Then the series* $\sum_{n=0}^{\infty}a_n p_n d\alpha$ *converges in* $\mathcal{H}'(I)$ *to (say)* f. *Moreover,* $a_n=\langle f,p_n\rangle$.

(ii) *If* $f\in\mathcal{H}'(I)$, *then the expansion coefficients of* f *given by* $a_n=\langle f,p\rangle$ *satisfy* $\overline{\lim}_{n\to\infty}|a_n|^{1/n}\le 1$. *Moreover, the expansion of* f *converges to it in the sense of* $\mathcal{H}'(I)$.

Proof. (i) Since the space $\mathcal{H}'(I)$ is weakly sequentially complete, it suffices to show that the sequence $S_N=\sum_{n=0}^{N}a_n p_n d\alpha$ converges weakly, i.e., $\lim_{N\to\infty}\langle S_N,\phi\rangle$ exists for all $\phi\in\mathcal{H}(I)$. From Theorem 2.1, we have for any $\phi\in\mathcal{H}(I)$ the expansion

$$\phi(x)=\sum_{m=0}^{\infty}\gamma_m p_m(x)\quad\text{with}\quad\overline{\lim_{m\to\infty}}\,|\gamma_m|^{1/m}<1.$$

Hence,

$$\lim_{N\to\infty}\langle S_N,\phi\rangle=\lim_{N\to\infty}\sum_{n=0}^{N}a_n\sum_{m=0}^{\infty}\gamma_m\langle p_n d\alpha,p_m\rangle$$

$$=\lim_{N\to\infty}\sum_{n=0}^{N}a_n\gamma_n=\sum_{n=0}^{\infty}a_n\gamma_n.$$

The last series converges since $\overline{\lim}_{n\to\infty}|a_n\gamma_n|^{1/n}<1$. That $a_n=\langle f,p_n\rangle$ follows immediately from the orthonormality of $\{p_n\}$.

(ii) Since $p_n\in\mathcal{H}(I)$, $a_n=\langle f,p_n\rangle$ is well defined. The series $\sum_{n=0}^{\infty}a_n p_n d\alpha$ converges to f in $\mathcal{H}'(I)$ since $\langle f,\phi-\sum_{n=0}^{N}\gamma_n p_n\rangle\to 0$, and hence

$$\langle f,\phi\rangle=\sum_{n=0}^{\infty}a_n\gamma_n=\sum_{n=0}^{\infty}a_n\langle p_n d\alpha,\phi\rangle.$$

To show that $\overline{\lim}_{n\to\infty}|a_n|^{1/n}\le 1$, we assume the contrary and let $\overline{\lim}_{n\to\infty}|a_n|^{1/n}=\rho>1$ and $2\varepsilon=\rho-1>0$. Choose $\phi\in\mathcal{H}(I)$ such that

$$\phi(x)=\sum_{n=0}^{\infty}\gamma_n p_n(x),\qquad\overline{\lim_{n\to\infty}}\,|\gamma_n|^{1/n}=\frac{1}{\rho-\varepsilon}<1.$$

Then, the series $\langle f,\phi\rangle=\sum_{n=0}^{\infty}a_n\gamma_n$ diverges since

$$\overline{\lim_{n\to\infty}}\,|a_n\gamma_n|^{1/n}=\frac{\rho}{\rho-1}>1,$$

which gives the desired contradiction. □

3. Analytic representations of hyperfunctions. We have seen that there exists a one-to-one correspondence between $\mathcal{H}'(I)$, the space of hyperfunctions on I, and the space of equivalence classes of functions that are holomorphic in $\mathbf{C}-I$ and vanish at ∞.

In this section, we shall manifest the same relationship between the two spaces expressed in terms of the orthogonal polynomials and functions of the second kind.

In order to conform with previous convention, we introduce:

DEFINITION 3.1. Let $f \in \mathcal{H}'(I)$. Then the *analytic representation* ((the indicatrix)$/2\pi i$) $\hat{f}(z)$ of f is defined by

$$\hat{f}(z)=\frac{1}{2\pi i}\left\langle f(t), \frac{1}{t-z}\right\rangle, \qquad z\in \mathbf{C}-I.$$

The polynomials are still in the class E and the notation of §2 remains unchanged.

THEOREM 3.1. (i) $f\in\mathcal{H}'(I)$ *if and only if* $f=\Sigma_{n=0}^{\infty} a_n p_n d\alpha$ *with* $\overline{\lim}_{n\to\infty}|a_n|^{1/n}\le 1$.

(ii) *The analytic representation of* $f=\Sigma_{n=0}^{\infty} a_n p_n d\alpha \in \mathcal{H}'(I)$ *is given by*

$$\hat{f}(z)=\frac{1}{2\pi i}\sum_{n=0}^{\infty} a_n q_n(z),$$

which is holomorphic in $\mathbf{C}-I$, *and* $\hat{f}(\infty)=0$.

(iii) *Given a holomorphic function* $\hat{f}(z)$ *in* $\mathbf{C}-I$ *with* $\hat{f}(\infty)=0$, *then there is a series in the form* $(1/2\pi i)\Sigma_{n=0}^{\infty} a_n q_n(z)$ *that converges to it. Moreover, this series is the analytic representation of some* $f\in\mathcal{H}'(I)$ *and* $f=\Sigma_{n=0}^{\infty} a_n p_n d\alpha$.

Proof. (i) This is a restatement of Theorem 2.2.

(ii) By definition, we have

$$\begin{aligned}\hat{f}(z)&=\frac{1}{2\pi i}\left\langle f, \frac{1}{t-z}\right\rangle\\ &=\frac{1}{2\pi i}\sum_{n=0}^{\infty} a_n\left\langle p_n d\alpha, \frac{1}{t-z}\right\rangle\\ &=\frac{1}{2\pi i}\sum_{n=0}^{\infty} a_n\int_{-1}^{1}\frac{p_n(t)}{t-z}d\alpha\\ &=\sum_{n=0}^{\infty} a_n q_n(z).\end{aligned}$$

It is an easy matter to show that $\hat{f}(z)$ is holomorphic in $\mathbf{C}-I$ and $\hat{f}(\infty)=0$, since $\overline{\lim}_{n\to\infty}|a_n|^{1/n}\le 1$ and $q_n(z)=O[(|\xi|^{-1}+\varepsilon)^n]$ as $n\to\infty$, where $z=\frac{1}{2}(\xi+1/\xi)$ and ε is arbitrary (see [2]).

(iii) Let $\hat{f}(z)$ be holomorphic in $\mathbf{C}-I$ with $\hat{f}(\infty)=0$. Then

$$\hat{f}(z)=\frac{1}{2\pi i}\int_{\gamma}\frac{\hat{f}(t)}{t-z}dt,$$

where γ is any path, negatively oriented enclosing I. By (2.4) we have

$$\hat{f}(z)=\frac{1}{2\pi i}\int_{\gamma}\sum p_n(t)\hat{f}(t)q_n(z)\,dt=\sum_{n=0}^{\infty} a_n q_n(z),$$

where the series converges uniformly for z outside of any ellipse with foci at ± 1. Hence

$$\int_{\gamma}\hat{f}(z)\phi(z)\,dz=\sum_{n=0}^{\infty} a_n\int_{\gamma} q_n(z)\phi(z)\,dz, \tag{3.1}$$

where γ is now another ellipse with foci at ± 1 lying entirely inside the domain of analyticity of the function $\phi(z)$. But by Theorem 2.1, $\phi(z)$ has an expansion of the form

$$\phi(z)=\sum_{n=0}^{\infty}\gamma_n p_n(z), \qquad \overline{\lim_{n\to\infty}}\,|\gamma_n|^{1/n}<1, \tag{3.2}$$

and the series converges uniformly inside and on γ. Substituting this into (3.1) gives

$$\int_\gamma \hat{f}(z)\phi(z)\,dz=\frac{1}{2\pi i}\sum_{n=0}^{\infty}a_n\sum_{m=0}^{\infty}\gamma_m\int_\gamma q_n(z)p_m(z)\,dz. \tag{3.3}$$

Upon using the relation given by (2.3) we obtain

$$\begin{aligned}\int_\gamma \hat{f}(z)\phi(z)\,dz&=\frac{1}{2\pi i}\sum_{n=0}^{\infty}a_n\sum_{m=0}^{\infty}\gamma_m\int_\gamma p_m(z)\,dz\int_{-1}^{1}\frac{p_n(t)}{t-z}\,d\alpha(t)\\ &=\frac{1}{2\pi i}\sum_{n=0}^{\infty}a_n\sum_{m=0}^{\infty}\gamma_m\int_{-1}^{1}p_n(t)\int_\gamma\frac{p_m(z)}{t-z}\,dz\,d\alpha(t)\\ &=\sum_{n=0}^{\infty}a_n\sum_{m=0}^{\infty}\gamma_m\int_{-1}^{1}p_n(t)p_m(t)\,d\alpha(t)=\sum_{n=0}^{\infty}a_n\gamma_n.\end{aligned} \tag{3.4}$$

Since the last series converges for any sequence $\{\gamma_n\}$ satisfying (3.2), it follows that $\overline{\lim}_{n\to\infty}|a_n|^{1/n}\le 1$ by the argument in Theorem 2.1.

Hence, by (i), there is an $f\in\mathcal{H}'(I)$ given by $\Sigma a_n p_n d\alpha$, and by (ii) its analytic representation is $(1/2\pi i)\Sigma a_n q_n(z)$. □

4. The convergence theorem. In the last section, we established a one-to-one correspondence between the hyperfunctions on $[-1,1]$ and their analytic representations given by

$$\begin{aligned}f&=\sum_{n=0}^{\infty}a_n p_n d\alpha,\\ \hat{f}(z)&=\frac{1}{2\pi i}\sum_{n=0}^{\infty}a_n q_n(z), \qquad z\notin[-1,1].\end{aligned} \tag{4.1}$$

In this section, we examine the jump of $\hat{f}(z)$ across the cut $[-1,1]$ in the global and the local sense.

Globally, we shall show that $\lim_{y\to 0}\{\hat{f}(x+iy)-\hat{f}(x-iy)\}=f$. More precisely, if ϕ is holomorphic on $[-1,1]$, then there exists a $\rho>0$ such that ϕ is holomorphic on $[-1-\rho,1+\rho]$ and we have

$$\lim_{y\to 0}\int_{-1-\rho}^{1+\rho}\{\hat{f}(x+iy)-\hat{f}(x-iy)\}\phi(x)\,dx=\langle f,\phi\rangle \quad \text{for all } \phi\in\mathcal{H}$$

Locally, we shall show that $\lim_{y\to 0}\{\hat{f}(x+iy)-\hat{f}(x-iy)\}=f(x)$ pointwise for $x\in(a,b)$, where $f(x)$ is interpreted in a certain sense.

THEOREM 4.1. *Let f and $\hat{f}(z)$ be defined as in* (4.1). *Then*

$$\lim_{y\to 0}\int_{-1-\rho}^{1+\rho}\{\hat{f}(x+iy)-\hat{f}(x-iy)\}\phi(x)\,dx=\langle f,\phi\rangle \quad \textit{for all } \phi\in\mathcal{H},$$

where ρ is taken so that $[-1-\rho,1+\rho]$ is inside the region of analyticity of ϕ.

Proof.

$$\begin{aligned}\hat{f}(z)-\hat{f}(\bar{z}) &= \frac{1}{2\pi i}\sum_{k=0}^{\infty} a_k[q_k(z)-q_k(\bar{z})] \\ &= \frac{1}{2\pi i}\sum_{k=0}^{\infty} a_k\int_{-1}^{1}\frac{p_k(t)}{t-z}d\alpha(t)-\int_{-1}^{1}\frac{p_k(t)}{t-\bar{z}}d\alpha(t) \\ &= \frac{1}{\pi}\sum_{k=0}^{\infty} a_k\int_{-1}^{1}p_k(t)\frac{y}{(t-x)^2+y^2}d\alpha(t) \\ &= \sum_{k=0}^{\infty} a_k\int_{-1}^{1}p_k(t)P_y(t,x)\,d\alpha(t),\end{aligned}$$

where $P_y(t,x)=y/(\pi[(t-x)^2+y^2])$ is the Poisson kernel for the upper half plane. Since the series converges uniformly in the outside of any nondegenerate ellipse with foci at ± 1, it must also converge uniformly for all x for fixed $y>0$. Hence, we can write

$$\begin{aligned}&\int_{-1-\rho}^{1+\rho}\{\hat{f}(x+iy)-\hat{f}(x-iy)\}\phi(x)\,dx \\ &\qquad = \sum_{k=0}^{\infty} a_k\int_{-1-\rho}^{1+\rho}\int_{-1}^{1}p_k(t)P_y(t,x)d\alpha(t)\phi(x)\,dx \\ &\qquad = \sum_{k=0}^{\infty} a_k\int_{-1}^{1}\{\hat{\phi}(t+iy)-\hat{\phi}(t-iy)\}p_k(t)\,d\alpha(t) \\ &\qquad = \langle f,\hat{\phi}(\cdot+iy)-\hat{\phi}(\cdot-iy)\rangle,\end{aligned}$$

where

$$2\pi i\hat{\phi}(t\pm iy)=\int_{-1-\rho}^{1+\rho}\frac{\phi(x)}{x-t\pm iy}dx.$$

It remains to prove that $[\hat{\phi}(x+iy)-\hat{\phi}(x-iy)]\to\phi(x)$ in the sense of $\mathcal{H}(I)$, that is, uniformly in some open set containing $[-1,1]$. But this is a straightforward application of Cauchy's theorem. Indeed, let φ be holomorphic in $\bar{R}$ where

$$R=\{(x,y)\,|\,|x|<1+\rho,\ |y|<\varepsilon\}$$

and let U be an open set containing $[-1,1]$ whose closure lies in R. Then for $x\in U$

$$\begin{aligned}\phi(x) &= \frac{1}{2\pi i}\int_{\partial R}\frac{\phi(t)}{t-x}dt \\ &= \frac{1}{2\pi i}\left(\int_{-i\varepsilon-1-\rho}^{i\varepsilon-1-\rho}+\int_{i\varepsilon-1-\rho}^{i\varepsilon+1+\rho}+\int_{i\varepsilon+1+\rho}^{-i\varepsilon+1+\rho}+\int_{-i\varepsilon+1+\rho}^{-i\varepsilon-1-\rho}\right)\frac{\phi(t)}{t-x}dt \\ &= I_1+I_2+I_3+I_4.\end{aligned}$$

As $\varepsilon\to 0+$, the first and third integrals approach 0 while the others satisfy

$$\begin{aligned}I_2 &= \frac{1}{2\pi i}\int_{-1-\rho}^{1+\rho}\frac{\phi(t+i\varepsilon)-\phi(t)}{t+i\varepsilon-x}dt+\frac{1}{2\pi i}\int_{-1-\rho}^{1+\rho}\frac{\phi(t)}{t+i\varepsilon-x}dt \\ &= O(1)+\hat{\phi}(x+i\varepsilon)\end{aligned}$$

and

$$I_4=O(1)-\hat{\phi}(x-i\varepsilon).$$

□

5. Characterization theorem. Hyperfunctions have been characterized in many different ways. In [7] Johnson shows that f is a hyperfunction on the unit circle T if and only if there is a sequence $\{f_n\}_{n=0}^{\infty}$ of continuous functions on T such that

$$\lim_{n\to\infty}(n!\|f_n\|_\infty)^{1/n}=0, \qquad f=\sum_{n=0}^{\infty} f_n^{(n)}.$$

If we move to the real line, we find the following characterization given by Baernstein [1] for the analytic representation of hyperfunctions on any compact subset K of the real line:

$$\hat{f}(z)=\sum_{n=0}^{\infty}\int_K \frac{f_n(t)}{(t-z)^{n+1}}\,dt, \qquad z\notin K,$$

and $\|f_n\|_p^{1/n}\to 0$ as $n\to\infty$, $p\in[1,\infty)$. However, no explicit representation for the hyperfunctions themselves has been given. One should not expect the nice characterization given by Johnson to be carried over to K because of the behavior of the derivatives of $\{f_n\}$ at the boundary points.

On attempting to characterize the solutions of the generalized axially symmetric potential differential equation, i.e., GASP functions, we [12] gave a representation for hyperfunctions on $[-1,1]$ by using the Gegenbauer orthogonal polynomials.

Analogously, in this section we give a characterization for hyperfunctions on $[a,b]$ by using the orthonormal sequence $\{p_n\}$. It will be shown that in $[a,b]$ every hyperfunction f can be written in the form

$$f=\sum_{n=0}^{\infty} f_n^{(n)}+g_1+g_2,$$

where $\{f_n\}$ is a sequence of continuous functions on $[a,b]$ satisfying $\lim_{n\to\infty}(n!\|f_n\|_\infty)^{1/n}=0$, g_1 and g_2 are hyperfunctions concentrated on $[a]$ and $[b]$ respectively.

Throughout this section the norm $\|\cdot\|$ will denote the sup norm $\|\cdot\|_\infty$.

DEFINITION 5.1. Let $(p_k\,d\alpha)^{(-n)}$ denote the nth antiderivative of $p_k\,d\alpha$, i.e.,

$$(p_k\,d\alpha)^{(-n)}(x)=\frac{1}{\Gamma(n)}\int_a^x p_k(t)(x-t)^{n-1}\,d\alpha(t).$$

LEMMA 5.1. *Let $d\alpha$ have support in $[-1,1]$. Then there exist c and b such that for $n>1$ we have*

$$|(p_k\,d\alpha)^{(-n)}(x)|\le cb^n k^{1-n}, \qquad k\ge n-1.$$

Proof. Let $R_n(t)=(t^{n-1}/\Gamma(n))H(t)$, where $H(t)$ is the Heaviside function. Then $R_n(t)\in C^{n-2}[-2,2]$, and by Jackson's theorem [6] there is a polynomial $\pi(t)$ of degree $k-1$ such that

$$|R_n(t)-\pi(t)|\le C\lambda^{n-1}\frac{(n-1)^{n-2}}{(n-2)!}k^{1-n}L^{n-1}4^{n-1}\le ca^n k^{1-n},$$

$$k\ge n-1, \quad n>1, \quad -2\le t\le 2,$$

where c and a are independent of n and k. Hence,

$$\begin{aligned}(p_k\,d\alpha)^{(-n)}(x)&=\int_{-1}^{1}p_k(t)R_n(x-t)\,d\alpha(t)\\&=\int_{-1}^{1}p_k(t)\big(R_n(x-t)-\pi(x-t)\big)\,d\alpha, \qquad -1\le x\le 1.\end{aligned}$$

and

$$|(p_k d\alpha)^{(-n)}(x)| \le \int_{-1}^{1} |p_k(t)|\, d\alpha(t) \max_{-2\le t\le 2} |R_n(t)-\pi(t)|,$$

from which the conclusion follows. □

THEOREM 5.1. *Let f be a hyperfunction on $[a,b]$. Then there are a sequence of continuous functions $\{f_n\}$ on $[a,b]$ and sequences of complex numbers $\{c_n\}$, $\{d_n\}$ satisfying*

$$\lim_{n\to\infty} (n!|c_n|)^{1/n} = 0 = \lim_{n\to\infty} (n!|d_n|)^{1/n},$$

$$\lim_{n\to\infty} (n!\| f_n\|)^{1/n} = 0, \tag{5.1}$$

such that

$$f = \sum_{n=0}^{\infty} f_n^{(n)} + \sum_{n=0}^{\infty} c_n \delta^{(n)}(t-a) + \sum_{n=0}^{\infty} d_n \delta^{(n)}(t-b) \tag{5.2}$$

and conversely.

Proof. We know that f can be written in the form

$$f = \sum_{k=0}^{\infty} a_k p_k d\alpha \quad \text{with} \quad \overline{\lim_{k\to\infty}} |a_k|^{1/k} \le 1.$$

Johnson's decomposition theorem [7] and its modified version [12] show that we can find a sequence $\{a_{k,n}\}$ and a finite-valued function B such that

$$a_k = \sum_{n=0}^{k} a_{k,n}, \qquad |a_{k,n}| \le \frac{B(\varepsilon)\varepsilon^{n+c+2}k^n}{(n+c+2)!} \tag{5.3}$$

for $0\le n\le k$, $k=1,2,3,\cdots$, all $\varepsilon>0$, and c is a positive integer that depends on the orthonormal family $\{p_k\}$. Then, by Lemma 5.1 "extended to $[a,b]$", we have

$$|(p_k d\alpha)^{(-n)}| \le C b^n k^{1-n},$$

where b and C are independent of both n and k. Now we write

$$f_n(x) = \sum_{k=n-c-2}^{\infty} a_{k,n-c-2} (p_k d\alpha)^{(-n)}(x) \tag{5.5}$$

for $[a,b]$ and zero otherwise.

Except for a trivial modification when $n=c+2$, we have

$$\begin{aligned} |f_n(x)| &\le \sum_{k=n-c-2}^{\infty} |a_{k,n-c-2}|\, |(p_k d\alpha)^{(-n)}(x)| \\ &\le C \sum_{k=n-1}^{\infty} \frac{B(\varepsilon)\varepsilon^n k^{n-c-2}}{n!} k^{1-n} b^n + \sum_{k=n-c-2}^{n-2} \frac{B(\varepsilon)\varepsilon^n k^{n-c-2}}{n!} \frac{(b-a)^n}{(n-1)!} \\ &\le \frac{B(\varepsilon)\varepsilon^n}{n!} \left\{ C b^n \sum_{k=n-1}^{\infty} \frac{1}{k^{c+1}} + O(1) \right\}. \end{aligned}$$

Thus $f_n(x)$ is continuous in (a,b), right and left continuous at a,b respectively. Moreover, we have

$$(n!\| f_n\|)^{1/n} \le [2B(\varepsilon)]^{1/n} \varepsilon b \quad \text{for all } \varepsilon>0,$$

which gives (5.1).

For any holomorphic function ϕ on $[a,b]$, we have by integration by parts

$$\int_a^b p_k(x)\phi(x)\,d\alpha=(p_k\,d\alpha)^{(-1)}\phi\Big|_a^b-(p_k\,d\alpha)^{(-2)}\phi^{(1)}\Big|_a^b+(p_k\,d\alpha)^{(-3)}\phi^{(2)}\Big|_a^b$$
$$\cdots+(-1)^{n-1}(p_k\,d\alpha)^{(-n)}\phi^{(n-1)}\Big|_a^b+(-1)^n\int_a^b(p_k\,d\alpha)^{(-n)}\phi^{(n)}\,dx.$$

But for $k\geq n$, $(p_k\,d\alpha)^{(-n)}(b)=0$; hence

$$\int_a^b p_k(x)\phi(x)\,d\alpha=\sum_{j=k+1}^{n}(-1)^{j-1}(p_k\,d\alpha)^{(-j)}\phi^{(j-1)}\Big|_a^b$$
$$+(-1)^n\int_a^b(p_k\,d\alpha)^{(-n)}\phi^{(n)}\,dx,$$

from which one obtains

$$(f_n^{(n)},\phi)=\sum_{k=n-c-2}^{\infty}a_{k,n-c-2}\sum_{j=k+1}^{n}(-1)^{j-1}(p_k\,d\alpha)^{(-j)}\phi^{(j-1)}\Big|_a^b$$
$$+(-1)^n(f_n,\phi^{(n)}),$$

and consequently

$$\sum_{n=c+2}^{\infty}(f_n^{(n)},\phi)=\sum_{n=c+2}^{\infty}(-1)^n(f_n,\phi^{(n)})$$
$$+\sum_{n=c+2}^{\infty}\sum_{k=n-c-2}^{\infty}a_{k,n-c-2}\sum_{j=k+1}^{n}(-1)^j(p_k\,d\alpha)^{(-j)}\phi^{(j-1)}\Big|_a^b. \tag{5.6}$$

The proof will be completed if we can show that the second series on the right-hand side of (5.6) converges at the endpoints (say at b). But this is a straightforward calculation if we observe that $\phi^{(j)}(b)=O(j!/\delta^j)$ since ϕ is holomorphic in some δ neighborhood of b, and that

$$|(p_k\,d\alpha)^{(-j)}(x)|\leq cb^jk^{1-j}.$$

Finally, it is not hard to see that

$$d_{m-1}=(-1)^{m-1}\sum_{j=0}^{c+1}\sum_{k=m-c-2+j}^{m-1}a_{k,k-j}(p_k\,d\alpha)^{(-m)}(b). \qquad \square$$

This characterization theorem enables us to investigate the local behavior of the jump $\hat{f}(z)-\hat{f}(\bar{z})$ at a point $x_0\in(a,b)$ as it is related to the behavior of the hyperfunction f at x_0.

Since hyperfunctions are, in general, functionals, the phrase "behavior of f at x_0" may seem somewhat meaningless. Fortunately, there is an already available concept known as "value of a generalized function at a point" that we shall employ to give $f(x_0)$ a meaning. This concept has been extended to hyperfunctions on the unit circle [10] and will be adopted here.

DEFINITION 5.2. The hyperfunction f on $[a,b]$ is said to have a *value* γ at $x_0\in(a,b)$ if there exist a representation $\Sigma_{n=0}^{\infty}f_n^{(n)}$ of f satisfying (5.1), a sequence of polynomials $\{g_n\}$ of degree less than n and a sequence of complex numbers $\{\gamma_n\}$ such that, for each

$\varepsilon>0$, there exists a δ such that

$$\left|\frac{f_n(x)-g_n(x)}{(x-x_0)^n}-\frac{\gamma_n}{n!}\right|\leq\frac{\varepsilon^{n+1}}{n!}\quad\text{for } 0<|x-x_0|<\delta,$$

$n=1,2,3,\cdots$ and $g(x_0)=\gamma_0$. The value of f at x_0 is given by $\gamma=\Sigma_{n=0}^{\infty}\gamma_n$.

Simply, this definition says that each function f_n has a Peano derivative γ_n at x_0 and that the convergence to that value is faster for larger n.

For the proofs of the facts that the series $\Sigma\gamma_n$ converges and that the limit is independent of the representation $\Sigma f_n^{(n)}$ of f, see [10].

THEOREM 5.2. *Let f have a value γ at $x_0\in(a,b)$. Then*

$$\lim_{y\to 0}\{\hat{f}(x_0+iy)-\hat{f}(x_0-iy)\}=\gamma.$$

Proof. From (5.2), we get

$$\hat{f}(z)=\frac{1}{2\pi i}\left\langle f,\frac{1}{t-z}\right\rangle=\frac{1}{2\pi i}\sum_{n=0}^{\infty}\left\langle f_n^{(n)},\frac{1}{t-z}\right\rangle+C(z),$$

where

$$C(z)=\frac{1}{2\pi i}\left\{\sum_{n=0}^{\infty}(-1)^n\frac{c_n}{(a-z)^{n+1}}+\sum_{n=0}^{\infty}(-1)^n\frac{d_n}{(b-z)^{n+1}}\right\},$$

which is finite by Theorem 5.1.

Using the same arguments as in the proof of Theorem 5.1, in particular (5.6), we can show that

$$\hat{f}(z)=\frac{1}{2\pi i}\sum_{n=0}^{\infty}(-1)^n n!\int_a^b\frac{f_n(t)}{(t-z)^{n+1}}dt+U(z)+V(z)+C(z),$$

where $U(z)$ is holomorphic in the entire complex plane with the point a removed and $V(z)$ is holomorphic in the entire complex plane with the point b removed. In fact, it is not hard to see that

$$\lim_{z\to\infty}U(z)=0=\lim_{z\to\infty}V(z).$$

Now we have

$$\hat{f}(z)-\hat{f}(\bar{z})=\frac{1}{2\pi i}\sum_{n=0}^{\infty}(-1)^n n!\int_a^b f_n(t)\left\{\frac{1}{(t-z)^{n+1}}-\frac{1}{(t-\bar{z})^{n+1}}\right\}dt$$
$$+\{U(z)-U(\bar{z})\}+\{V(z)-V(\bar{z})\}+\{C(z)-C(\bar{z})\}.$$

Then, for $x_0\in(a,b)$ we have

$$\lim_{y\to 0}\{\hat{f}(x_0+iy)-\hat{f}(x_0-iy)\}=\lim_{y\to 0}\frac{1}{2\pi i}\sum_{n=0}^{\infty}\int_a^b f_n(t)P^{(n)}(t,x_0,y)\,dt,$$

where

$$P^{(n)}(t,x_0,y)=(-1)^n n!\left[\frac{1}{(t-x_0-iy)^{n+1}}-\frac{1}{(t-x_0+iy)^{n+1}}\right].$$

The remainder of the proof follows the same arguments as in the proof of the main theorem in [10]. □

Acknowledgment. The form of the proof of Lemma 5.1 is due to P. Nevai, whom the authors wish to thank.

REFERENCES

[1] A. BAERNSTEIN, *A representation theorem for functions holomorphic off the real axis*, Trans. Amer. Math. Soc., 165 (1972), pp. 159–165.
[2] JA. L. GERONIMUS, *Polynomials Orthogonal on a Circle and Interval*, Pergamon Press, New York, 1960.
[3] _____, *Orthogonal Polynomials*, AMS Transl. (2) 108, American Mathematical Society, Providence, RI, 1977.
[4] R. P. GILBERT AND S. Y. SHIEH, *A new method in the theory of potential scattering*, J. Math. Phys., 7 (1966), pp. 431–433.
[5] R. P. GILBERT, *Function Theoretic Methods in Partial Differential Equations*, Academic Press, New York, 1969.
[6] D. JACKSON, *Theory of Approximation*, AMS Colloquium Publications XI, American Mathematical Society, Providence, RI, 1930.
[7] G. JOHNSON, *Harmonic functions on the unit disk*, Illinois J. Math., 12 (1968), pp. 366–385.
[8] G. KÖTHE, *Topological Vector Spaces* I, Springer-Verlag, New York, 1969.
[9] G. WALTER AND P. NEVAI, *Series of orthogonal polynomials as boundary values*, this Journal, 12 (1981), pp. 502–513.
[10] G. WALTER, *Local boundary behavior of harmonic functions*, Illinois J. Math., 16 (1970), pp. 491–501.
[11] A. I. ZAYED, *On the singularities of Gegenbauer expansions*, Trans. Amer. Math. Soc., 262 (1981), pp. 487–503.
[12] _____, *Hyperfunctions as boundary values of generalized axially symmetric potentials*, Illinois J. Math., 25 (1981), pp. 306–317.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0013 $01.00/0

THE STIELTJES SUMMABILITY METHOD AND SUMMING STURM–LIOUVILLE EXPANSIONS*

LOUISE A. RAPHAEL†

Abstract. The Stieltjes summability method for divergent integrals is defined. The name is derived from its close relationship with the Stieltjes transform. After proving some basic properties, the Stieltjes summability method is compared with Cesàro and Abel summability methods. The main result proves the Stieltjes summability of eigenfunction expansions associated with a certain class of singular Sturm–Liouville systems, and develops, using the Stieltjes means, a stable method of summing such expansions under perturbations of the coefficients.

1. Introduction. One summability method which has been applied to

$$\sum_n^\infty a_n \quad \text{or} \quad \sum_n^\infty a_n u_n(x) \tag{1.1}$$

is given by

$$\sum_n^\infty \frac{a_n}{1+b\lambda_n} \quad \text{or} \quad \sum_n^\infty \frac{a_n u_n(x)}{1+b\lambda_n}, \tag{1.2}$$

where b is the summation parameter and λ_n an unbounded increasing sequence. The latter series has been used to sum eigenfunction expansions associated with regular Sturm–Liouville systems, where $u_n(x)$ is an eigenfunction and λ_n the corresponding eigenvalue. We cite in particular the works of Hille [7], Tikhonov [13] and Titchmarsh [14].

The first series in (1.1) has been studied by Hudak [8] under the name T-summation. Also comparisons of the first summation method (1.2) with Cesàro summability were studied by Bromwich [4] and Moore [10] under a different formulation.

One natural extension of (1.2) to integrals on the half line is

$$\int_0^\infty f(x)(1+bg(x))^{-w}dx \quad \text{and} \quad \int_0^\infty f(y)u_n(x,y)(1+bg(y))^{-w}dy,$$

where g is a positive function increasing to ∞. This leads to the formal definition of Stieltjes summability:

DEFINITION 1. Let $f:[0,\infty)\to C$ (the set of complex numbers) be a Lebesgue measurable, locally integrable function and g a positive Lebesgue measurable function monotonically increasing to ∞ on $[0,\infty)$. We say that $\int_0^\infty f(x)dx$ is *Stieltjes summable of order* $w>0$ *to* L ($S(g,w)$ *summable to* L) if $\int_0^\infty f(x)(1+bg(x))^{-w}dx$ exists for $b>0$ and $\lim_{b\to 0}\int_0^\infty f(x)(1+bg(x))^{-w}dx = L$ (finite) and write $\int_0^\infty f = L\ S(g,w)$.

Remark 1. In this paper $\int_0^\infty f(x)dx$ means $\lim_{a\to\infty}\int_0^a f(x)dx$, provided the limit exists, or equals plus or minus infinity.

DEFINITION 2. The *Stieltjes means* are defined by $\int_0^\infty f(x)(1+bg(x))^{-w}dx$.

Remark 2. If $g(x)=x$ and $w=1$, the Stieltjes mean is closely related to the Stieltjes transform (Widder [17]). The Stieltjes transform is a twice iterated Laplace transform.

---

* Received by the editors September 13, 1978. This research was supported in part by the National Science Foundation under MISRIP grant HES75-08424, and by the Resource Center for Science and Engineering, Atlanta University.

† Department of Mathematics, Clark College, Atlanta, Georgia 30314. Present address: Department of Mathematics, Howard University, Washington, DC 20059.

It was the transform used by Stieltjes (1894) in his infinite continued fractions research. These transforms have been used for the moment problem on the semi-infinite interval and in Padé approximants (Bender and Orszag [2]).

The following is a brief outline of the remainder of the paper. In §2 we prove the regularity of the Stieltjes summability method and prove that it has the basic properties common to the standard regular summability methods. In §3 we compare the strength of the Stieltjes, Cesàro and Abel summability methods. In §4 we obtain our main result on the Stieltjes summability of a class of singular Sturm–Liouville expansions, and develop, using the Stieltjes means, a stable method of summing such expansions under the L^2 perturbation of the coefficients.

2. Basic properties. We begin by proving that the $S(g,w)$ method is regular, that is, if $\int_0^\infty f$ exists then $\int_0^\infty f$ is Stieltjes summable to the same sum. In what follows, we assume f, g, b and w are as in Definition 1 (unless noted otherwise).

PROPOSITION 1. *If* $\int_0^\infty f = L$, *then* $\int_0^\infty f = L\ S(g,w)$.

Proof. The existence of $\int_0^\infty (1+bg)^{-w} f$ follows from an analogue of Abel's lemma [3, pp. 473–475]. To prove regularity, we consider

$$\int_0^\infty f(x)\,dx - \int_0^\infty f(x)[1+bg(x)]^{-w}dx = \lim_{A\to\infty}\int_0^A f(x)\left[1-(1+bg(x))^{-w}\right]dx.$$

Let $h(x)=1-(1+bg(x))^{-w}$; so h is monotone and $h<1$. We choose M so that $A>M$ implies $|\int_M^A f|<\varepsilon$. Then

$$\left|\int_0^A f(x)h(x)\,dx\right| \le \left|\int_0^M f(x)h(x)\,dx\right| + \varepsilon$$

and

$$\lim_{A\to\infty}\left|\int_0^A f(x)h(x)\,dx\right| \le \left|\int_0^M f(x)h(x)\,dx\right| + \varepsilon.$$

Finally,

$$\limsup_{b\to 0}\left|\int_0^\infty f(x)h(x)\,dx\right| \le \varepsilon + \limsup_{b\to 0}\left|\int_0^M f(x)h(x)\,dx\right| \le \varepsilon,$$

finishing the proof.

Example 1. The integral $\int_0^\infty \exp(-x)\,dx = 1$ and $\int_0^\infty \exp(-x)\,dx = 1\ S(x,1)$. It is of historical interest to note that Euler summed the divergent series $f(b)=\sum_{k=0}^\infty (-1)^k k!\, b^k$ by rewriting $f(b)$ as $\int_0^\infty \exp(-x)(1+bx)^{-1}\,dx$ [6, p. 26].

PROPOSITION 2. *If* $f\ge 0$, *then* $\int_0^\infty f = \lim_{b\to 0}\int_0^\infty f(x)(1+bg(x))^{-w}\,dx$.

Remark 2. As usual, we allow both sides to be infinite.

Proof. The proof follows from a simple application of the monotone convergence theorem.

Consequently, every function f for which $\int_0^\infty f$ does not exist and which is $S(g,w)$ summable is oscillatory on $[0,\infty)$. In the following corollary, we define $f^+ = \frac{1}{2}(|f|+f)$ and $f^- = \frac{1}{2}(|f|-f)$ as is customary.

COROLLARY.

(a) *If* $\int_0^\infty f^+ = \infty$, $\int_0^\infty f^- < \infty$, *then* $\lim_{b\to 0}\int_0^\infty f(x)(1+bg(x))^{-w}\,dx = +\infty$.

(b) *If* $\int_0^\infty f^+ < \infty$, $\int_0^\infty f^- = \infty$, *then* $\lim_{b\to 0}\int_0^\infty f(x)(1+bg(x))^{-w}\,dx = -\infty$.

Example 2. Some examples of $S(g,w)$ summability are

$$\int_0^\infty \cos(mx)\,dx = 0 \quad S(x,1), \qquad \int_0^\infty \sin(mx)\,dx = \frac{1}{m} \quad S(x,1).$$

So

$$\int_0^\infty \exp(imx)\,dx=\frac{i}{m} \quad S(x,1), \qquad \int_0^\infty x\sin x\,dx=0 \quad S(x^2,1) \qquad \text{(see [1, p. 232])}.$$

These answers are in agreement with results obtained by other standard summability methods such as Cesàro and Abel.

The next proposition states the strength of the $S(g,w)$ method for a given $g(x)$ increases as the order w increases.

PROPOSITION 3. *If $v>w>0$ and $\int_0^\infty f(x)\,dx=L\ S(g,w)$, then $\int_0^\infty f(x)\,dx=L\ S(g,v)$.*

Proof. The proof technique is identical to that of Proposition 1, where now $h(x)$ is defined to be equal to $[1-(1+bg(x))^{v-w}]$.

We now note that the Stieltjes summability method has the additive, homogeneity and translative properties. (Translative means that $\int_0^\infty f\ S(g,w)=\int_0^c f+\int_c^\infty f\ S(g,w)$ for finite $c>0$.)

Proposition 4, which follows next, gives a reason for choosing g to be a monotone function increasing to ∞ in the definition of Stieltjes summability. Propositions 5 and 6 define the range of possible Stieltjes sums when f is fixed and g any admissible function. Propositions 4, 5 and 6 below were proven by Professor Kenneth Davidson of Waterloo University. The propositions are proven for Stieltjes summability of order $w=1$, but they are true for order $w>1$ by Proposition 3.

PROPOSITION 4. *Let f be real valued on $[0,\infty)$. If $\int_0^\infty f^+=\infty$, $\int_0^\infty f^-=\infty$ and $L\in\mathbb{R}$, then there exists $g>0$ such that $\lim_{b\to 0}\int_0^\infty(1+bg)^{-1}f=L$.*

Proof. Let A_n be disjoint subsets such that $[0,\infty)=\bigcup_{n=1}^\infty A_n$ and such that $\int_{A_1}f=L$, $\int_{A_n}f=0$ for $n\geq 2$, $\int_{A_n}|f|=a_n<\infty$ for $n\geq 1$. Set $g=n^2a_n$ on A_n. Then

$$\int_0^\infty |f|(1+bg)^{-1}=\sum_{n=1}^\infty \frac{a_n}{1+n^2a_nb}<\frac{1}{b}\sum_{n=1}^\infty\frac{1}{n^2}<\infty.$$

So

$$\int_0^\infty f(1+bg)^{-1}=\frac{L}{1+ba_1}\to L \quad \text{as } b\to 0.$$

PROPOSITION 5. *Let $F(x)=\int_0^x f(t)\,dt$, where f is real valued. If $\lim_x\inf F\leq L\leq \lim_x\sup F$, then there exists a monotone $g>0$ such that $\int_0^\infty f=L\ S(g,1)$.*

Proof. Choose $x_0=0$, $x_n\uparrow\infty$ such that $F(x_n)\to L$ and $b_n=F(x_n)-F(x_{n-1})$ satisfies $\sum_{n=1}^\infty|b_n|<\infty$. Let $a_n=\int_{x_{n-1}}^{x_n}|f|$. Define g as follows:

$$g(x)=\begin{cases} a_1 & \text{on } [x_0,x_1),\\ \max\{n^2a_n, n+g[x_{n-2},x_{n-1})\} & \text{on } [x_{n-1},x_n).\end{cases}$$

Then

$$\int_0^\infty |f|\,(1+bg)^{-1}\leq\sum_{n=1}^\infty\frac{a_n}{1+n^2a_nb}<\frac{1}{b}\sum_{n=1}^\infty\frac{1}{n^2}<\infty,$$

and so

$$\int_0^\infty f(1+bg)^{-1}=\sum_{n=1}^\infty\frac{b_n}{1+g(x_n)b}\to\sum_{n=1}^\infty b_n$$

by the Lebesgue dominated convergence theorem as $b\to 0$. But $\sum_{n=1}^\infty b_n=L$, and the theorem is proved.

PROPOSITION 6. *If $g>0$ is monotone, f is real valued and $\int_0^\infty f = L\ S(g(x),1)$, then for $F(x)=\int_0^x f(t)\,dt$, $\liminf_x F \le \int_0^\infty f = L\ S(g,1) \le \limsup_x F$.*

Proof. If h is monotone decreasing and $\lim_{x\to\infty} h(x)=0$, we write $h(x)=\int_x^\infty dh(t)$ (Stieltjes integral). If $f\cdot h$ is absolutely integrable, then by Fubini's theorem

$$\int_0^\infty fh\,dx = \int_0^\infty f(x)\int_x^\infty dh(t)\,dx = \int_0^\infty \int_0^t f(x)\,dx\,dh(t)$$
$$= \int_0^\infty F(t)\,dh(t) = \int_0^N F(t)\,dh(t) + \int_N^\infty F(t)\,dh(t).$$

Now

$$h(N)\cdot \inf_{t\ge N} F(t) \le \int_N^\infty F(t)\,dh(t) \le h(N)\cdot \sup_{t\ge N} F(t)$$

and

$$(h(0)-h(N))\cdot \inf_{0\le t\le N} F(t) \le \int_0^N F(t)\,dh(t) \le (h(0)-h(N))\cdot \sup_{0\le t\le N} F(t).$$

With $h_b=(1+bg)^{-1}=h, h_b\to 1$ pointwise as $b\to 0$, and

$$h(N)\cdot \inf_{t\ge N} F(t) + [h(0)-h(N)]\cdot \inf_{t\le N} F(t)$$
$$\le \int_0^\infty fh_b\,dx \le h(N)\cdot \sup_{t\ge N} F(t) + [h(0)-h(N)]\cdot \sup_{t\le N} F(t).$$

Let $b\to 0$ to obtain $\inf_{t\ge N} F(t) \le L \le \sup_{t\ge N} F(t)$, whence the desired conclusion follows upon letting $N\to\infty$.

3. Comparison theorems. The question of under what conditions Cesàro summability of order one implies Stieltjes summability was essentially answered for Fourier series and integrals [4], [10]. First we call to mind one definition of the Cesàro summability method for integrals.

DEFINITION 3. *$\int_0^\infty f(x)\,dx$ is $(C,1)$-summable* whenever $\lim_{u\to\infty} \frac{1}{u}\int_0^u ds \int_0^s f(x)\,dx$ has a definite value.

The next three propositions are reformulations of theorems stated in [4], [10] where g is as in Definition 1.

PROPOSITION 7. *If $f(x)$ is uniformly continuous on $[0,\infty)$ and $\int_0^\infty f(x)\,dx = L\ (C,1)$, then $\int_0^\infty f(x)\,dx = L\ S(g,w)$, $w\ge 2$.*

PROPOSITION 8. *If $\int_0^\infty f(x)\,dx = L\ (C,1)$ and $\lim_{x\to\infty}(1+bg(x))^{-w}\int_0^x f(y)\,dy = 0$ for every $b>0$ and $w\ge 1$, then $\int_0^\infty f(x)\,dx = L\ S(g,w)$, $w\ge 1$.*

The next proposition gives a sufficient condition for integrals which are $S(g,w)$ summable to be convergent in the ordinary sense.

PROPOSITION 9. *If $\int_0^\infty f = L\ S(g,w)$, $w\ge 1$ and $|xf(x)|<M$, where M is fixed and positive, then $\int_0^\infty f(x)\,dx$ is convergent and its value is L.*

We now set the order w of the generalized Stieltjes method to 1, $g(x)$ to x and compare Stieltjes summability with two Abel summability methods. Propositions 11, 12 and 13 generalize the results of J. Hudak [8] to divergent integrals.

DEFINITION 4. The integral $\int_0^\infty f$ is summable by the Abel method $A(x)$ to the sum L, if the Laplace transform of f, $\int_0^\infty \exp(-xs)f(x)\,dx$, converges for $s>0$ and $\lim_{s\to 0}\int_0^\infty f(x)\exp(-xs)\,dx = L$. We write $\int_0^\infty f = L\ A(x)$ and say $\int_0^\infty f$ is *Abel summable to L.*

Similarly, $\int_0^\infty f$ is summable by the Abel method $A(\ln x)$ to the sum L, if the Mellin transform of f, $\int_0^\infty f(x)x^{-s}\,dx = \int_0^\infty f(x)\exp(-s\ln x)\,dx$, converges for $s>0$ and $\lim_{s\to 0}\int_0^\infty f(x)x^{-s}\,dx = L$. Here we write $\int_0^\infty f = L\,A(\ln x)$.

For divergent series, it is known that Cesàro summability implies Abel summability. This generalization for the integral analogue is false [6, p. 136]. However, if $\int_0^\infty f$ is $(C,1)$ summable and $\int_0^\infty f(x)\exp(-xs)\,dx$ converges for every $s>0$, then $\int_0^\infty f$ is Abel summable to the same sum [6]. Similarly, the next proposition states that Abel summability plus convergence of $\int_0^\infty f(x)(1+bx)^{-1}\,dx$ for $b>0$ implies $S(x,1)$ summability to the same sum. Consequently, there is no natural chain of summability strengths without extra conditions.

So that this paper will be self-contained we now state a test for uniform convergence of improper integrals which is analogous to Dirichlet's test for infinite series [15, p. 23].

PROPOSITION 10. *The integral $\int_0^\infty f(x,s)\phi(x,s)\,dx$ is uniformly convergent if $\phi(x,s)$ is such that $\frac{\partial\phi}{\partial x}$ is continuous, $\phi(x,s)$ tends to zero monotonically and uniformly with respect to s, and if $\left|\int_0^x f(t,s)\,dt\right|$ is less than a constant independent of x and s for all x.*

In what follows the Stieltjes mean $\int_0^\infty f(x)(1+bx)^{-1}\,dx$ will be expressed as a twice iterated Laplace transform

$$\int_0^\infty f(x)(1+bx)^{-1}dx = \frac{1}{b}\int_0^\infty \exp\left(-\frac{s}{b}\right)\int_0^\infty f(x)\exp(-xs)\,dx\,ds.$$

PROPOSITION 11. *If $\int_0^\infty f = L\,A(x)$ and $\int_0^\infty (1+bx)^{-1}f$ converges for $b>0$, then $\int_0^\infty f = L\,S(x,1)$.*

Proof. Let $F(s)=\int_0^\infty f(x)\exp(-xs)\,dx$. As $F(s)\to L$ when $s\to 0$, there exists a $\delta>0$ such that $|F(s)-L|<\varepsilon$ for $0\le s\le\delta$. Moreover, $\frac{1}{b}\int_0^\infty \exp(-s/b)\,ds\int_0^\infty f(x)\exp(-xs)\,dx$ is bounded by some positive number, as the Stieltjes transform $\int_0^\infty f(x)(1+bx)^{-1}\,dx$ is equal to this twice iterated Laplace transform and the Stieltjes transform converges for $b>0$. Thus

$$\begin{aligned}
\left|\int_0^\infty f(x)(1+bx)^{-1}dx - L\right| &= \left|\frac{1}{b}\int_0^\infty \exp\left(-\frac{s}{b}\right)\int_0^\infty f(x)\exp(-xs)\,dx\,ds - L\right| \\
&\le \frac{1}{b}\int_0^\delta \exp\left(-\frac{s}{b}\right)\left|\int_0^\infty f(x)\exp(-xs)\,dx - L\right|ds \\
&\quad + \left[\frac{1}{b}\int_0^\delta \exp\left(-\frac{s}{b}\right)ds - 1\right]L \\
&\quad + \frac{1}{b}\int_\delta^\infty \exp\left(-\frac{s}{b}\right)\left|\int_0^\infty f(x)\exp(-xs)\,dx\right|ds \\
&\le \varepsilon\left|1-\exp\left(-\frac{\delta}{b}\right)\right| + \exp\left(-\frac{\delta}{b}\right)L + M\exp\left(-\frac{\delta}{b}\right)\to 0
\end{aligned}$$

as $b\to 0$ and $\varepsilon\to 0$, where $\left|\int_0^\infty f(x)\exp(-xs)\,dx\right|<M$.

PROPOSITION 12. *If the integral $\int_0^\infty f(x)\exp(-s\ln x)\,dx$ converges for s, $0<s<1$, then*

$$\int_0^\infty f(x)\exp(-s\ln x)\,dx = \left[\frac{\sin \pi s}{\pi}\right]\int_0^\infty b^{s-1}\int_0^\infty f(x)\frac{1}{1+bx}\,dx\,db$$

for $0<s<1$.

Proof. The integral $\int_0^\infty f(x)(1+bx)^{-1}\,dx = \int_0^\infty f(x)x^{-s}x^s(1+bx)^{-1}\,dx$ converges uniformly for $0<s<1$, and $0<\varepsilon\le b\le r<\infty$. As

$$\int_0^\infty b^{s-1}(1+bx)^{-1}db=\pi x^{-s}(\sin\pi s)^{-1}, \qquad 0<s<1,$$

then

$$\begin{aligned}
&\pi^{-1}\sin\pi s\int_\varepsilon^r b^{s-1}\int_0^\infty f(x)(1+bx)^{-1}dx\,db\\
&\quad=(\pi^{-1}\sin\pi s)\int_0^\infty f(x)\int_\varepsilon^r b^{s-1}(1+bx)^{-1}db\,dx\\
&\quad\to\pi^{-1}\sin\pi s\int_0^\infty f(x)\int_0^\infty b^{s-1}(1+bx)^{-1}db\,dx\\
&\quad=\int_0^\infty f(x)x^{-s}dx=\int_0^\infty f(x)\exp(-s\ln x)\,dx
\end{aligned}$$

as $\varepsilon\to 0$ and $r\to\infty$.

PROPOSITION 13. *If* $G(s)=\int_0^\infty f(x)\exp(-s\ln x)\,dx$ *converges for* $0<s<1$ *and* $\int_0^\infty f = L\ S(x,l)$, *then* $\int_0^\infty f = L\ A(\ln x)$.

Proof. By the definition of $A(\ln x)$, it suffices to prove that $|G(s)-L|$ approaches 0 as $s\to 0$, $0<s<1$. Given any $\varepsilon>0$, choose $\delta>0$ such that $\left|\int_0^\infty f(x)(1+bx)^{-1}dx-L\right|<\varepsilon$ for $0<b\le\delta$. By Proposition 10 $\int_0^\infty f(x)(1+bx)^{-1}dx$ converges uniformly for $b>\delta$. By Proposition 12 $\int_0^\infty b^{s-1}\int_0^\infty f(x)(1+bx)^{-1}dx$ is bounded by some positive number M for $0<s<1$. So

$$\begin{aligned}
|G(s)-L|&\le\left|\pi^{-1}\sin\pi s\int_0^\infty b^{s-1}\int_0^\infty b^{s-1}\int_0^\infty f(x)(1+bx)^{-1}dx\,db-L\right|\\
&=\left|\left(\pi^{-1}\sin\pi s\int_0^\delta b^{s-1}\int_0^\infty f(x)(1+bx)^{-1}dx\,db-\pi^{-1}\sin\pi s\int_0^\delta b^{s-1}L\,db\right)\right.\\
&\quad\left.+\left(\pi^{-1}\sin\pi s\int_0^\delta b^{s-1}L\,db-L\right)+\pi^{-1}\sin\pi s\int_\delta^\infty b^{s-1}\int_0^\infty f(x)(1+bx)^{-1}dx\,db\right|\\
&\le\pi^{-1}\sin\pi s\int_0^\delta b^{s-1}\left|\int_0^\infty f(x)(1+bx)^{-1}dx-L\right|db\\
&\quad+\left|\left[\pi^{-1}\sin\pi s\int_0^\delta b^{s-1}db-1\right]L\right|+\pi^{-1}\sin\pi s\left|\int_\delta^\infty b^{s-1}\int_0^\infty f(x)(1+bx)^{-1}dx\,db\right|\\
&\le(\pi s)^{-1}(\sin\pi s)\varepsilon\cdot\delta^s+\left|(\pi s)^{-1}\sin\pi s\cdot\delta^s-1\right|\cdot|L|+M\pi^{-1}\sin\pi s\to 0
\end{aligned}$$

as $s\to 0$ and $\varepsilon\to 0$. Then $\int_0^\infty f(x)x^{-s}dx\to L$ as $s\to 0$. Thus, $\int_0^\infty f(x)\,dx$ is summable by the $A(\ln x)$ method to the sum L.

4. An application. Tikhonov used a variational principle [11] to construct a stable method for summing regular Sturm–Liouville eigenfunction expansions [13]. In this section we extend Tikhonov's results by developing a stable method for approximating real-valued $L^2[0,\infty)$ functions pointwise by regularizing their singular $[0,\infty)$ Sturm–Liouville eigenfunction expansions. The spectrum of the singular system, unlike that of the regular system, has a continuous component. The methods of proof are analogous

to those of Tikhonov. My special thanks go to Dr. Mark Kon of Boston University, who suggested some of the techniques which have been used.

Let S-L denote the following modified Sturm–Liouville singular system:

$$-f''(x)+q(x)f(x)=\lambda f(x), \qquad 0\leq x<\infty,$$
$$f(0)=0, \qquad f(\infty)<\infty$$

and let $q(x)$ be a continuous, uniformly bounded ($-M\leq q(x)\leq M$), real valued $L^1[0,\infty)$ function.

Example 3. In S-L, let $q(x)=0$ or $q(x)=[\cosh^2 x]^{-1}$. Both of these are sometimes interpreted as potential energy in the one-dimensional Schrödinger equation.

Three observations are made about S-L. First, on $L^2[0,\infty)$, the S-L operator is self-adjoint and hence its spectrum is real. Secondly, S-L falls into the limit point case at infinity of the Sturm–Liouville problem on $[0,\infty)$. Lastly, according to [9, pp. 206, 211] the spectral function $\rho(\lambda)$ is absolutely continuous for $\lambda>0$, that is $d\rho(\lambda)=\rho'(\lambda)d\lambda$, $\lambda>0$.

For q as in S-L, the negative part of the spectrum is discrete and bounded from below, and may be empty. [9, Thm. 3.1, p. 209]. An example of the latter possibility is given by $q(x)\geq 0$. We denote the associated orthonormal eigenfunctions by $\{u_n(x)\}_n$, where the index set may be infinite, finite or empty.

For $\lambda>0$ and q as in S-L, the spectrum of S-L is continuous in $(0,\infty)$ [9, Thm. 3.2, p. 211]. The associated (unnormalized) eigenfunctions, denoted by $\{u_\lambda(x)\}_{\lambda>0}$, are defined by $u_\lambda(0)=0$ and $u'_\lambda(0)=1$.

Now as the spectrum is continuous for $\lambda>0$ and the spectral function $\rho(\lambda)$ is absolutely continuous there, any real-valued $L^2[0,\infty)$ function f can be written as the sum of its discrete and continuous parts. That is,

$$f(x)\sim\sum_n F_n u_n(x)+\int_0^\infty F_\lambda u_\lambda(x)\,d\rho(\lambda),$$

where the expansion coefficients are

$$F_n\sim\int_0^\infty u_n(x)f(x)\,dx, \qquad F_\lambda\sim\int_0^\infty u_\lambda(x)f(x)\,dx$$

and $\sim$ denotes L^2 convergence as the upper limit of summation or integration becomes infinite.

Let $\{u_\nu(x)\}_{\nu\in S}$ be a complete system of eigenfunctions as described above, with S the spectrum of S-L. We fix $\bar{f}\in L^2[0,\infty)$, let $\gamma>0$, and let $f^*\in L^2[0,\infty)$ be such that $\|\bar{f}-f^*\|_2<\gamma$. That is,

$$f^*(x)\sim\sum_n F_n^* u_n(x)+\int_0^\infty F_\lambda^* u_\lambda(x)\,d\rho(\lambda), \tag{4.1}$$

where

$$\bar{f}(x)\sim\sum_n \bar{F}_n u_n(x)+\int_0^\infty \bar{F}_\lambda u_\lambda(x)\,d\rho(\lambda), \qquad F_\nu^*=\bar{F}_\nu+\Delta_\nu, \qquad \nu\in\text{spectrum}$$

and the errors $\{\Delta_\nu\}_\nu$ satisfy the inequality

$$\left(\sum_n \Delta_n^2+\int_0^\infty \Delta_\lambda^2\,d\rho(\lambda)\right)^{1/2}\leq\gamma. \tag{4.2}$$

In general, direct summation of (4.1) does not yield a stable method for obtaining a pointwise approximation to $\bar{f}$ even at continuity points of $\bar{f}$. This is because there can exist a continuity point x_0 of $\bar{f}$ and an $\varepsilon>0$ such that for any $\gamma>0$, there exists an f^* such that

$$\|\bar{f}-f^*\|_2\leq\gamma \quad \text{and} \quad |\bar{f}(x_0)-f^*(x_0)|>\varepsilon.$$

For this reason, the problem of summing Sturm–Liouville eigenfunction expansions belongs to the class of ill-posed problems [12].

By using variational methods, the problem of finding a stable method for approximating $\bar{f}$ pointwise is transformed into a problem of finding a function belonging to a special class of functions which minimizes a given functional. Following Tikhonov's method for solving ill-posed problems, consider the functional (in the variable function f)

$$v(\alpha,f,f^*)=\alpha\int_0^\infty\left([f'(x)]^2+q(x)f^2(x)\right)dx+\sum_n(F_n-F_n^*)^2+\int_0^\infty(F_\lambda-F_\lambda^*)^2d\rho(\lambda)$$

where $\alpha=k\gamma^2$, k is a positive constant, and γ is as in (4.2). The role of α as a function of γ is essential in this discussion.

We will use heuristic methods to minimize the above functional, and will show rigorously that the minimizing function satisfies the requirements for a stable approximation.

To find the minimizing function $f=f_\alpha^*\in L^2[0,\infty)$ such that $f_\alpha^*(0)=0$, $f_\alpha^*(\infty)<\infty$, we take the first variation, denoted by δ:

$$\delta(\alpha A)=2\alpha\int_0^\infty[-f''(x)+q(x)f(x)](\delta f)\,dx, \tag{4.3}$$

$$\begin{aligned}\delta(B)&=2\sum_n(F_n-F_n^*)(\delta F_n),\\ \delta(C)&=2\int_0^\infty(F_\lambda-F_\lambda^*)(\delta F_\lambda)\,d\rho(\lambda).\end{aligned} \tag{4.4}$$

So (4.3) and (4.4) imply that the minimizing function $f=f_\alpha^*$ satisfies the equation

$$\alpha[f''(x)-q(x)f(x)]-[f(x)-f^*(x)]=0 \tag{4.5}$$

or

$$-f''(x)+\left[q(x)+\frac{1}{\alpha}\right]f(x)=\frac{1}{\alpha}f^*(x) \tag{4.6}$$

with $f(0)=0$ and $f(\infty)<\infty$ [16].

To find the expansion coefficients of f_α^* with respect to the eigenfunctions $\{u_\nu\}_\nu$, we multiply (4.5) by $u_\nu(x)$, integrate, use the self-adjointness criterion and solve the resulting equation

$$\alpha[-\lambda_\nu F_{\alpha,\nu}^*]-F_{\alpha,\nu}^*+F_\nu^*=0.$$

We thus obtain

$$F_{\alpha,\nu}^*=\frac{F_\nu^*}{1+\alpha\lambda_\nu}, \tag{4.7}$$

so that

$$f_\alpha^*(x)\sim\sum_n\frac{F_n^*}{1+\alpha\lambda_n}u_n(x)+\int_0^\infty\frac{F_\lambda^*}{1+\alpha\lambda}u_\lambda(x)\,d\rho(\lambda) \tag{4.8}$$

satisfies (4.5) and $f_\alpha^*(0)=0$, $f_\alpha^*(\infty)<\infty$, $1+\alpha\lambda_n\neq 0$. Before showing that (4.8) represents a stable method of finding a pointwise approximation to $\bar{f}$, we introduce more notation.

Let L_α be the self adjoint operator of (4.6) defined by

$$L_\alpha(f(x))=-f''(x)+\left[q(x)+\frac{1}{\alpha}\right]f(x),$$

with q as in S-L, and in particular $-M\leq q(x)\leq M$ for some $M>0$. Let $G_{q+1/\alpha}(x,s)$ be the Green function of $L_\alpha(f(x))=0$.

If we define $\bar{f}_\alpha$ by

$$\bar{f}_\alpha(x)\sim\sum_n \frac{\bar{F}_n}{1+\alpha\lambda_n}u_n(x)+\int_0^\infty \frac{\bar{F}_\lambda}{1+\alpha\lambda}u_\lambda(x)\,d\rho(\lambda), \tag{4.9}$$

and $\Delta_\alpha(x)$ similarly, then

$$f_\alpha^*(x)=\bar{f}_\alpha(x)+\Delta_\alpha(x). \tag{4.10}$$

Not being able to calculate $G_{q+1/\alpha}(x,s)$ directly, we calculate the Green function for

$$-f''(x)+l_i^2 f(x)=0,\qquad i=1,2,\quad f(0)=0,\quad f(\infty)<\infty, \tag{4.10}'$$

where $l_1^2=-M+\frac{1}{\alpha}>0$ and $l_2^2=M+\frac{1}{\alpha}$, and α is sufficiently small. The Green function for (4.10)′ is

$$G_{l_i}^2(x,s)=\begin{cases}\dfrac{1}{l_i}e^{-l_i x}\sinh(l_i s) & \text{for } s<x,\\[2ex] \dfrac{1}{l_i}e^{-l_i s}\sinh(l_i x) & \text{for } s>x,\quad i=1,2.\end{cases}$$

LEMMA 1. *$G_{m+1/\alpha}(x,s)\leq G_{q+1/\alpha}(x,s)\leq G_{-M+1/\alpha}(x,s)$ when $-M+\frac{1}{\alpha}>0$.*

Let $M>0$ and $I=[x-\eta,x+\eta]\subset\mathbb{R}^+$. In Lemma 2 we choose α such that $0<\alpha<1$ and $\alpha M<1$.

LEMMA 2. *The following hold as α and M approach 0^+:*

(a) $$\int_{x-\eta}^{x+\eta} G_{q+1/\alpha}(x,s)\,ds=\alpha[1+O(\alpha)]\left[1+O\left(\exp\left(-\sqrt{\frac{1-\alpha M}{\alpha}}\,\eta\right)\right)\right],$$

(b) $$\int_0^\infty G_{q+1/\alpha}^2(x,s)\,ds=O(\alpha^{3/2}),$$

(c) $$\int_0^{x-\eta} G_{q+1/\alpha}^2(x,s)\,ds=\alpha^{3/2}O\left(\exp\left(-2\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right),$$

(d) $$\int_{x+\eta}^\infty G_{q+1/\alpha}^2(x,s)\,ds=\alpha^{3/2}O\left(\exp\left(-2\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right),$$

(e) $$\left[\int_0^{x-\eta} G_{q+1/\alpha}^2(x,s)\,ds+\int_{x+\eta}^\infty G_{q+1/\alpha}^2(x,s)\,ds\right]^{1/2}=\alpha^{3/4}O\left(\exp\left(-\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right)$$

if $1-\alpha M>0$.

Proof.

(a)

$$\int_{x-\eta}^{x+\eta} G_{-M+1/\alpha}(x,s)\,ds$$

$$=\frac{\alpha}{1-\alpha M}\left[1-\exp\left(-\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right.$$

$$\left.+\frac{\exp\left((-2x-\eta)\sqrt{\frac{1-\alpha M}{\alpha}}\right)-\exp\left((-2x+\eta)\sqrt{\frac{1-\alpha M}{\alpha}}\right)}{2}\right]$$

$$=\alpha[1+O(\alpha)]\left[1+O\left(\exp\left(-\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right)\right].$$

Replacing $\alpha/(1-\alpha M)$ with $\alpha/(1+\alpha M)$, we have

$$\int_{x-\eta}^{x+\eta} G_{M+1/\alpha}(x,s)\,ds=\alpha[1+O(\alpha)]\left[1+O\left(\exp\left(-\eta\sqrt{\frac{1+\alpha M}{\alpha}}\right)\right)\right].$$

And as $G_{q+1/\alpha}$ is sandwiched between $G_{M+1/\alpha}$ and $G_{-M+1/\alpha}$ we have (a).

(b)

$$\int_0^\infty G^2_{-M+1/\alpha}(x,s)\,ds$$

$$=\frac{1}{4}\left[\frac{\alpha}{1-\alpha M}\right]^{3/2}\left[1-\exp\left(-4x\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right]-\frac{\alpha}{1-\alpha M}x\exp\left(-2x\sqrt{\frac{1-\alpha M}{\alpha}}\right)$$

$$+\frac{1}{8}\left[\frac{\alpha}{1-\alpha M}\right]^{3/2}\left[1-2\exp\left(-2x\sqrt{\frac{1-\alpha M}{\alpha}}\right)+\exp\left(-4x\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right]$$

$$=O(\alpha^{3/2}).$$

Similarly $\int_0^\infty G^2_{M+1/\alpha}(x,s)\,ds=O(\alpha^{3/2})$ and thus (b) is established.

(c)

$$\int_0^{x-\eta} G^2_{-M+1/\alpha}(x,s)\,ds$$

$$=\frac{1}{8}\left[\frac{\alpha}{1-\alpha M}\right]^{3/2}\left[\exp\left(-2\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)-\exp\left((-4x+2\eta)\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right]$$

$$-\frac{1}{2}\left[\frac{\alpha}{1+\alpha M}\right]\exp\left(-2x\sqrt{\frac{1-\alpha M}{\alpha}}\right)(x-\eta)$$

$$=\alpha^{3/2}[1+O(\alpha^{3/2})]O\left(\exp\left(-2\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right)$$

$$=\alpha^{3/2}O\left(\exp\left(-2\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right).$$

Similarly

$$\int_0^{x-\eta} G^2_{M+1/\alpha}(x,s)\,ds = \alpha^{3/2} O\left(\exp\left(-2\eta\sqrt{\frac{1+\alpha M}{\alpha}}\right)\right)$$

and so (c) is proved.

(d)

$$\begin{aligned}
&\int_{x+\eta}^{\infty} G^2_{-M+1/\alpha}(x,s)\,ds \\
&\quad = \frac{1}{8}\left[\frac{\alpha}{1-\alpha M}\right]^{3/2}\left[\exp\left(-2\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right) - \exp\left((-4x-2\eta)\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right. \\
&\qquad\qquad \left. -2\exp\left(-2(x+\eta)\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right] \\
&\quad = \alpha^{3/2}\left[1+O(\alpha^{3/2})\right] O\left(\exp\left(-2\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right) \\
&\quad = \alpha^{3/2} O\left(\exp\left(-2\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right).
\end{aligned}$$

Thus (d) is proved.

(e) follows from adding the orders of (c) and (d), which is

$$\alpha^{3/2} O(\exp(-2\eta\sqrt{(1-\alpha M)/\alpha})),$$

and taking the square root, which is $\alpha^{3/4} O(\exp(-\eta\sqrt{(1-\alpha M)/\alpha}))$.

The main result of this section says that a sequence of Stieltjes means as in (4.8), $\{f^*_\alpha\}_{\alpha>0}$, formed from approximating an $L^2[0,\infty)$ function $\bar{f}$ by f^*, converges pointwise to $\bar{f}$ on the set of continuity points of $\bar{f}$, as long as α and $\|\bar{f}-f^*\|_2$ are scaled appropriately. That is, if for any $\gamma>0$ we choose an approximating $L^2[0,\infty)$ function f^* such that $\|\bar{f}-f^*\|^2_L \le \gamma$, then if f^*_α is related to f^* as above, and if $\alpha = k\gamma^2$ for some $k>0$, then

$$\lim_{\alpha\to 0} f^*_\alpha(x) = \lim_{\alpha\to 0} \sum_n \frac{F^*_n}{1+\alpha\lambda_n} u_n(x) + \lim_{\alpha\to 0}\int_0^\infty \frac{F^*_\lambda}{1+\alpha\lambda} u_\lambda(x)\,d\rho(\lambda) = \bar{f}(x)$$

at any given continuity point x of $\bar{f}$. We note that above, f^* depends implicitly on γ.

LEMMA 3. *If $f^*(x)\in L^2[0,\infty)$, and f^*_α is as in* (4.8), *then*

$$f^*_\alpha(x) = \frac{1}{\alpha}\int_0^\infty G_{q+1/\alpha}(x,s) f^*(s)\,ds, \qquad x>0 \tag{4.11}$$

under the above assumption about $q(x)$.

Proof. We first show that the right-hand side of (4.11) is continuous. Since $G(x,s)$ is continuous in x and s, it suffices to show that

$$\int_M^\infty G_{q+1/\alpha}(x,s) f^*(s)\,ds \to 0 \quad \text{as } M\to\infty,$$

uniformly in x on bounded x-intervals, which follows after an application of the Cauchy–Schwarz inequality, and the bound in Lemma 1. To show that $f^*_\alpha(x)$ is

continuous, we write

$$f_\alpha^*(x)=\int_{-b}^{\infty}\frac{F_\nu^* u_\nu(x)}{1+\alpha\nu}\,d\rho(\nu), \tag{4.12}$$

where now the integral includes the discrete sum over negative eigenvalues, which are obtained from jumps in the total spectral function $\rho(\nu)$ over negative ν. Here $-b$ is a finite lower bound for the spectrum of S-L, and $u_\nu(x)$ for $\nu<0$ is a solution of the S-L equation, satisfying $u_\nu(0)=0$ and $u_\nu'(0)=1$. The correct normalization for the eigenfunction is included in $\rho(\nu)$. Since $u_\nu(x)$ is continuous in ν and x, it suffices to show that

$$\int_M^{\infty}\frac{F_\nu^* u_\nu(x)}{1+\alpha\nu}\,d\rho(\nu)\to 0 \quad \text{as } M\to\infty,$$

uniformly in finite x-intervals. This follows by the Cauchy–Schwarz inequality and the fact that

$$\int_M^{\infty}\left(\frac{u_\nu(x)}{1+\alpha\nu}\right)^2 d\rho(\nu)\to 0 \quad \text{as } M\to\infty,$$

uniformly in x on finite intervals ([9, Corollary I, p. 116]).

Let the operator A_α^* be defined by setting $A_\alpha^* f^*$ equal to the right side of (4.11), and A_α by

$$A_\alpha : f^*\to\int_{-b}^{\infty}\frac{F_\nu^* u_\nu(x)}{1+\alpha\nu}\,d\rho(\nu).$$

Since the ranges of both A_α and A_α^* consist of continuous functions, it now suffices to show that A_α^* and A_α coincide as linear operators in the Hilbert space $L^2[0,\infty)$. By the Parseval equality, it follows that A_α is bounded. To show that A_α^* is bounded, we note that $G_{-M+1/\alpha}(x,s)\le e^{-l|x-s|}$, where $l=[-M+\frac{1}{\alpha}]^{1/2}$. By Lemma 1, it thus suffices to show that

$$A_\alpha^* : f^*(x)\to\int_0^{\infty} e^{-l|x-s|}f^*(s)\,ds$$

is bounded. The integral on the right is a convolution, and the boundness of A_α^* becomes clear upon application of the Fourier transform.

It now suffices to show that A_α and A_α^* agree on a dense subset of $L^2[0,\infty)$, say on $C_c^\infty[0,\infty)$. We then show that if $f^*\in C_c^\infty[0,\infty)$, then both $A_\alpha f^*$ and $A_\alpha^* f^*$ solve the differential equation

$$-y''+\left(q+\frac{1}{\alpha}\right)y=\frac{1}{\alpha}f^*, \qquad y(0)=0,\quad y(\infty)<\infty. \tag{4.13}$$

Since, for sufficiently small α, $-\frac{1}{\alpha}$ is not in the spectrum of $L=-d^2/dx^2+q$, the solution is thus clearly unique and given by $A_\alpha^* f$. A fairly standard argument involving differentiation under an integral shows also that $A_\alpha f^*$ solves (4.13), so that $A_\alpha f^*=A_\alpha^* f^*$, for $f^*\in C_c^\infty[0,\infty)$. The proof is then complete.

Also, the solutions for (4.10) can be represented as

$$\begin{aligned} f_\alpha^*(x)&=\frac{1}{\alpha}\int_0^{\infty}G_{q+1/\alpha}(x,s)f^*(s)\,ds,\\ \bar f_\alpha(x)&=\frac{1}{\alpha}\int_0^{\infty}G_{q+1/\alpha}(x,s)\bar f(s)\,ds, \qquad (4.14)\\ \Delta_\alpha(x)&=\frac{1}{\alpha}\int_0^{\infty}G_{q+1/\alpha}(x,s)\Delta(s)\,ds \end{aligned}$$

PROPOSITION 14. *Let $\bar{f}\in L^2[0,\infty)$ be real valued. Let $k>0$ and $\alpha=k\gamma^2$. If $\bar{f}$ is continuous at $x\in(0,\infty)$, then for any $\varepsilon>0$ there exists a $\gamma>0$ such that $|f_\alpha^*(x)-\bar{f}(x)|<\varepsilon$ for any $L^2[0,\infty)$ function f^* such that*

$$\|f^*-\bar{f}\|_{L^2}=\|\Delta\|_{L^2}=\left(\sum_n \Delta_n^2+\int_0^\infty \Delta_\lambda^2\, d\rho(\lambda)\right)^{1/2}\le\gamma.$$

Proof. To prove the proposition we use arguments analogous to Tikhonov's [13]. The idea of the proof is to find estimates for $|f_\alpha^*(x)-\bar{f}(x)|\le|\bar{f}_\alpha(x)-\bar{f}(x)|+|\Delta_\alpha(x)|$ and to show that these estimates converge to 0 as $\alpha\to0$ ($\gamma\to0$).

We first bound $|\Delta_\alpha(x)|$. By the Cauchy–Schwarz inequality and the proof of Lemma 2(b)

(4.15)
$$|\Delta_\alpha(x)|\le\frac{1}{\alpha}\|\Delta\|_{L^2}\left[\int_0^\infty G_{q+1/\alpha}^2(x,s)\,ds\right]^{1/2}=\frac{\gamma}{\alpha}O(\alpha^{3/4})=O(\alpha^{1/4})\to0\quad\text{as }\alpha\to0.$$

Let $I=[x-\eta,x+\eta]$, and $\eta=\eta(\alpha)$ such that as $\alpha\to0$, $\eta\to0$ and $\exp(-\eta\alpha^{-1/2})\alpha^{-1/4}\to0$. Let $w(\eta)=\sup_{x\in I}\bar{f}(x)-\inf_{x\in I}\bar{f}(x)$. As $\eta\to0$, $w(\eta)\to0$. In finding upper and lower bounds for $\bar{f}_\alpha$ we use an order argument.

Let $K=[0,x-\eta]\cup[x+\eta,\infty)$. From (4.14),

$$\begin{aligned}\bar{f}_\alpha(x)&=\frac{1}{\alpha}\int_{x-\eta}^{x+\eta}G_{q+1/\alpha}(x,s)\bar{f}(s)\,ds+\frac{1}{\alpha}\int_K G_{q+1/\alpha}(x,s)\bar{f}(s)\,ds\\&\le\frac{1}{\alpha}[\bar{f}(x)+w(\eta)][\alpha(1+O(\alpha))]\left[1+O\left(\exp\left(-\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right)\right]\\&\quad+\frac{1}{\alpha}\|\bar{f}\|_{L^2}\left[\int_K G_{q+1/\alpha}^2(x,s)\,ds\right]^{1/2}\\&=[\bar{f}(x)+w(\eta)](1+O(\alpha))\left[1+O\left(\exp\left(-\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right)\right]\\&\quad+\|\bar{f}\|_{L^2}\frac{1}{\alpha^{1/4}}O\left(\exp\left(-\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right)\end{aligned}$$

by Lemma 2(a), (e) and Cauchy–Schwarz. Similarly,

$$\begin{aligned}\bar{f}_\alpha(x)&\ge[\bar{f}(x)-w(\eta)](1+O(\alpha))\left[1+O\left(\exp\left(-\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right)\right]\\&\quad-\|\bar{f}\|_{L^2}\frac{1}{\alpha^{1/4}}O\left(\exp\left(-\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right).\end{aligned}$$

Thus,

(4.16)
$$\begin{aligned}|\bar{f}_\alpha(x)-\bar{f}(x)|&\le w(\eta)(1+O(\alpha))\left[1+O\left(\exp\left(-\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right)\right]\\&\quad+\|\bar{f}\|_{L^2}\frac{1}{\alpha^{1/4}}O\left(\exp\left(-\eta\sqrt{\frac{1-\alpha M}{\alpha}}\right)\right)\to0\end{aligned}$$

as $\alpha\to0$ for $w(\eta)\to0$ since $\alpha^{-1/4}O(\exp(-\eta\sqrt{(1-\alpha M)/\alpha}))\to0$ and $\|\bar{f}\|_{L^2}$ is bounded.

Finally combining (4.15) and (4.16), we see that

$$|f_\alpha^*(x)-\bar{f}(x)|\leq|\bar{f}_\alpha(x)-\bar{f}(x)|+|\Delta_\alpha(x)|\to 0$$

as $\alpha\to 0$ if $\|\Delta\|_{L^2}\leq\gamma$, $\alpha=k\gamma^2$, and the proof is complete.

COROLLARY. *Let $\bar{f}\in L^2[0,\infty)$ be real valued and f_n^* be a sequence of functions in $L^2[0,\infty)$ such that $\|f_n^*-\bar{f}\|_2=\gamma_n\to 0$. Let $\alpha_n=k\gamma_n^2$ for a fixed $k>0$. Then $f_{n,\alpha_n}^*\to_{n\to\infty}\bar{f}$ at any point of continuity of $\bar{f}$, where the Fourier coefficients of f_{n,α_n}^* are related to those of f_n^* by (4.7), with $\alpha=\alpha_n$.*

In particular, this corollary states that the singular S-L eigenfunction expansion of any $L^2[0,\infty)$ function f is Stieltjes summable to f on the set of continuity points of f (let $f_n^*=f_f$).

Acknowledgments. I would like to express my appreciation to a host of mathematicians who through mathematical conversations helped me to know the meaning of the word colleague: Dale Alspach, Ted Chihara, Ken Davidson, Harvey Diamond (especially in connection with §§2 and 3), Evans Harrell and Mark Kon; and most importantly to Ken Hoffman who extended an invitation to spend a year at MIT.

I am deeply grateful to the referee for his extraordinarily patient critical reading and helpful suggestions.

REFERENCES

[1] M. ABRAMOWITZ AND I. STEGUN, eds., *Handbook of Mathematical Functions (with Formulas, Graphs, and Mathematical Tables)* NBSAM55, National Bureau of Standards, Washington, DC, 1964.

[2] C. BENDER AND S. ORSZAG, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978.

[3] T. BROMWICH, *An Introduction to the Theory of Infinite Series*, 2nd ed., Cambridge University Press, London, 1926.

[4] _____, *On the limits of certain infinite series and integrals*, Math. Annalen, 65 (1908), pp. 350–369.

[5] H. DIAMOND, M. KON AND L. RAPHAEL, *Stable summation methods for a class of singular Sturm-Liouville expansions*, Proc. Amer. Math. Soc., 81 (1981), pp. 279–286.

[6] G. H. HARDY, *Divergent Series*, Oxford University Press, London, 1963.

[7] E. HILLE, *Lectures on Ordinary Differential Equations*, Addison-Wesley, Reading, Ma, 1969.

[8] JU. I. HADAK, *Two inclusion theorems for a method of generalized summation of $T(\{\lambda_k\})$ series*, Soviet Math. Dokl., 13 (1972), pp. 304–308.

[9] B. M. LEVITAN AND I. S. SARGSJAN, *Introduction to Spectral Theory: Self Adjoint Ordinary Differential Operators*, American Mathematical Society, Providence, RI, 1975.

[10] C. N. MOORE, *On the introduction of convergence factors into summable series and summable integrals*, AMS Trans., 8 (1907), pp. 299–330.

[11] A. N. TIKHONOV, *Regularization of incorrectly posed problems*, Soviet Math. Dokl., 4 (1963), pp. 1624–1627.

[12] A. N. TIKHONOV AND V. Y. ARSENIN, *Solutions of Ill Posed Problems*, V. H. Winston, Washington, 1977.

[13] A. N. TIKHONOV, *Stable methods for the summation of Fourier series*, Soviet Math. Dokl., 5 (1964), pp. 641–644.

[14] E. C. TITCHMARSH, *Eigenfunction Expansions*, Part I, Oxford University Press, London, 1962.

[15] _____, *The Theory of Functions*, 2nd ed., Oxford University Press, London, 1939.

[16] R. WEINSTOCK, *Calculus of Variations*, McGraw-Hill, New York, 1952.

[17] D. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1941.

SIAM J. MATH. ANAL.
Vol. 13, No. 4, July 1982

0036-1410/82/1304-0014 $01.00/0

INTEGRATION OF INTERVAL FUNCTIONS II. THE FINITE CASE*

L. B. RALL†

Abstract. Caprani, Madsen and Rall [Siam J. Math. Anal., 12 (1981), pp. 321–341] have shown previously that the use of interval values leads to a simple theory of integration in which all functions, interval and real, are integrable. Here, a simplified construction of the interval integral is given for the case that the integrand and interval of integration are finite; the interval integral is shown to be the intersection of the interval Darboux sums corresponding to the partitions of the interval of integration into subintervals of equal length. A rate of convergence of these interval Darboux sums to the interval integral is given for Lipschitz continuous integrands. An alternate approach to interval integration in the unbounded case via finite interval integrals is presented. The results give theoretical support to interval methods for the solution of integral equations and finding extreme values of functionals defined in terms of integrals.

1. Introduction. The construction of the interval integral, given in the general case in [1], can be simplified drastically in the case that the interval of integration is finite and the integrand is a bounded interval function. (Definitions of the necessary concepts will be given below.) In particular, the use of the extended real number system is not required, so all computations can be done by ordinary interval arithmetic [3], [4]. Furthermore, it is not necessary to consider all partitions of the interval of integration into subintervals as the partition into subintervals with equal lengths will be shown to suffice. This eliminates an inherently nonconstructive portion of the definition of the interval interval, the formation of the so-called interval Riemann sums.

In addition to the simplification of the construction of the interval interval in this case, rates of convergence of the Darboux sums based on the equipartition of the interval of integration to the interval integral will be derived for sufficiently smooth integrands. Another approach to improper interval integrals will also be given.

2. Interval functions. Following the definitions in [1], an *interval function* Y defined on an interval $X=[a,b]$ assigns the interval value

$$Y(x)=[\underline{y}(x),\bar{y}(x)] \tag{2.1}$$

to each real number $x\in X$, where $\underline{y},\bar{y}$ are real functions called respectively the *lower* and *upper boundary functions* (or *endpoint functions*) of Y.

The *vertical extent* of Y on X is defined to be the interval

$$\nabla Y(X)=\left[\inf_{x\in X}\{\underline{y}(x)\},\ \sup_{x\in X}\{\bar{y}(x)\}\right]. \tag{2.2}$$

In this paper only intervals of integration with finite *width* $w(X)=b-a$ and *bounded* interval functions such that $w(\nabla Y(X))<+\infty$ will be considered. This is the *finite case*.

The notation $Y=[\underline{y},\bar{y}]$ will also be used for interval functions. Real functions y may be identified with the interval functions $y=[y,y]$ with equal endpoint functions, which are called *degenerate* interval functions [1].

---

* Received by the editors February 9, 1981. This research was sponsored in part by the U.S. Army Research Office under contract DAAG29-80-C-0041, and by the Danish Natural Science Research Council under grant 511-15849.

† Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706.

3. Interval integrals. In general, the *interval integral* of an interval function Y over the interval $X=[a,b]$ is the interval

$$\int_X Y(x)\,dx=\int_a^b Y(x)\,dx=\left[\int_X \underline{y}(x)\,dx, \int_X \bar{y}(x)\,dx\right], \tag{3.1}$$

where $\int_X \underline{y}(x)\,dx$ denotes the lower Darboux integral of the lower endpoint function $\underline{y}$ over the interval X and $\overline{\int}_X \bar{y}(x)\,dx$ gives the upper Darboux integral of the upper endpoint function $\bar{y}$ over X [2]. As these Darboux integrals always exist in the extended real number system, it follows that all interval (and hence all real) functions are integrable in this sense. The definite and indefinite interval integrals have many properties similar to those of the Riemann integral [1].

The construction of the interval integral, carried out in [1] in the spirit of interval analysis, is done in three steps. The first step consists of partition of the interval X into subintervals $X_i=[x_{i-1},x_i]$, $i=1,2,\cdots,n$ by means of points

$$a=x_0\le x_1\le\cdots\le x_{i-1}\le x_i\le\cdots\le x_{n-1}\le x_n=b \tag{3.2}$$

to obtain the *partition*

$$\Delta_n=(X_1,X_2,\cdots,X_n) \tag{3.3}$$

of X and the corresponding *interval Darboux sum*

$$\sum_{\Delta_n} Y(X)=\sum_{i=1}^{n} w(X_i)\cdot\nabla Y(X_i). \tag{3.4}$$

Next, for each positive integer n, let $\mathfrak{D}_n$ denote the set of all partitions (3.3). The *interval Riemann sum of order n* is then defined to be

$$\sum_n Y(X)=\bigcap_{\Delta_n\in\mathfrak{D}_n}\sum_{\Delta_n} Y(X). \tag{3.5}$$

Finally, the interval integral of Y over X is given by

$$\int_a^b Y(x)\,dx=\bigcap_{n=1}^{\infty}\sum_n Y(X), \tag{3.6}$$

which is nonempty, as the interval Riemann sums form a decreasing sequence of nonempty closed sets [1], and agrees with (3.1). This construction will be simplified in the finite case.

4. The finite case. The interval integral (3.6) will be said to be *finitely defined* if the integrand Y is a bounded interval function and the interval of integration $X=[a,b]$ is finite. The *equipartition* $\bar{\Delta}_n$ of X is defined by the points

$$x_i=a+ih,\qquad h=\frac{b-a}{n},\qquad i=0,1,\cdots,n, \tag{4.1}$$

so that

$$w(X_i)=x_i-x_{i-1}=\frac{b-a}{n}=\frac{w(X)}{n}=h,\qquad i=1,2,\cdots,n. \tag{4.2}$$

The corresponding interval Darboux sum is

$$\sum_{\bar{\Delta}_n} Y(X)=\bar{\sum_n} Y(X)=\frac{w(X)}{n}\sum_{i=1}^{n}\nabla Y(X_i). \tag{4.3}$$

THEOREM 4.1. *In the finite case,*

$$\int_a^b Y(x)\,dx = \bigcap_{n=1}^{\infty} \overline{\sum_n} Y(X). \tag{4.4}$$

Thus, this construction requires only the formation of the single interval Darboux sum (4.3) for each positive integer n and skips the (nonconstructive) calculation of interval Riemann sums (3.5) entirely. Furthermore, (4.4) agrees with the definition of the interval integral given by R. E. Moore [2], [3], in the case that the endpoint functions $y, \bar{y}$ of Y are assumed to be continuous. Theorem 4.1 will be proved in §6 based on results on subintervals established in the next section.

5. Two lemmas on subintervals. The first lemma simplifies the proof of the mean interval-value theorem for interval integrals over a finite interval of integration.

LEMMA 5.1. *If* $Z_i=[c_i,d_i]\subset Z=[c,d]$ *are finite intervals,* $i=1,2,\cdots,n$, *and* $\alpha_i\geq 0$ *with* $\Sigma_{i=1}^n \alpha_i = 1$, *then*

$$\sum_{i=1}^{n} \alpha_i Z_i \subset Z. \tag{5.1}$$

Proof. This follows at once from the elementary inequalities

$$a\leq \alpha_1 a_1+\alpha_2 a_2+\cdots+\alpha_n a_n\leq \alpha_1 b_1+\alpha_2 b_2+\cdots+\alpha_n b_n\leq b \tag{5.2}$$

for convex combinations of real numbers. Q.E.D.

On the assumption that Theorem 4.1 holds this gives the mean interval-value theorem [1] for the interval integral (4.4) as

$$\frac{1}{n}\sum_{i=1}^{n} \nabla Y(x_i) = \overline{Y}_n \subset \nabla Y(X) \tag{5.3}$$

by Lemma 5.1 and from (4.4)

$$\int_a^b Y(x)\,dx = w(X)\cdot \bigcap_{n=1}^{\infty} \overline{Y}_n = w(X)\cdot \overline{Y}, \tag{5.4}$$

where $\overline{Y}\subset \nabla Y(X)$.

The *excess width* of an interval $Z=[c,d]$ over a subinterval $Z'=[c',d']\subset Z$ is defined to be

$$e(Z,Z')=\max\{c'-c,d-d'\}. \tag{5.5}$$

It is evident that

$$e(Z,Z')\leq w(Z)=d-c. \tag{5.6}$$

A *symmetric interval* is an interval S of the form $S=[-s,s]$, where $s\geq 0$.

LEMMA 5.2. *If* $Z'\subset Z$, *then for each symmetric interval* $S=[-s,s]$ *with* $s\geq e(Z,Z')$, *one has*

$$Z\subset Z'+S. \tag{5.7}$$

In particular,

$$Z\subset Z'+[-w(Z),w(Z)]. \tag{5.8}$$

Proof. The inclusion (5.7) follows from the definition (5.5) and the definition of interval addition [2], [3]; (5.8) follows immediately by (5.6). Q.E.D.

6. Proof of Theorem 4.1. It is to be shown that definitions (3.6) and (4.4) of the interval integral agree in the finite case. Set

$$I=\bigcap_{n=1}^{\infty}\overline{\sum_{n}}\,Y(X). \tag{6.1}$$

As the interval integral (3.6) is contained in each Darboux sum $\Sigma_{\Delta_n} Y(X)$, it follows that

$$\int_a^b Y(x)\,dx\subset\overline{\sum_{n}}\,Y(X),\qquad n=1,2,\cdots, \tag{6.2}$$

and thus

$$\int_a^b Y(x)\,dx\subset I=\bigcap_{n=1}^{\infty}\overline{\sum_{n}}\,Y(X). \tag{6.3}$$

Suppose that a partition point p, $x_{i-1}\le p\le x_i$ is introduced into an interval X_i. By Lemma 5.1, one has

$$w[x_{i-1},p]\cdot\nabla Y([x_{i-1},p])+w[p,x_i]\cdot\nabla Y([p,x_i])\subset w(X_i)\cdot\nabla Y(X_i). \tag{6.4}$$

Consider an arbitrary interval Darboux sum $\Sigma_{\Delta_m} Y(X)$ for some positive integer m. For each $n\ge m$, the partition points $x_1, x_2,\cdots,x_{m-1}$ of the interval Darboux sum are interior to at most $m-1$ subintervals of $\overline{\Sigma}_n Y(X)$, with total length not exceeding $((m-1)/n)w(X)$. After deletion of these subintervals from $\bar{\Delta}_n$, the remaining partition points of $\bar{\Delta}_n$ will belong to the subintervals of Δ_m. By (6.4) and Lemma 5.2,

$$\overline{\sum_{n}}\,Y(X)\subset\sum_{\Delta_m}Y(X)+\frac{(m-1)w(X)}{n}[-w(\nabla Y(X)),w(\nabla Y(X))]. \tag{6.5}$$

As (6.5) holds for each partition Δ_m and positive integer $n\ge m$, from (3.5),

$$\overline{\sum_{n}}\,Y(X)\subset\sum_{m}Y(X)+\frac{(m-1)w(X)}{n}w(\nabla Y(X))\cdot[-1,1]. \tag{6.6}$$

As $w(\nabla Y(X))<+\infty$, taking the intersectioin of both sides of (6.6) with respect to n gives

$$I\subset\sum_{m}Y(X)+[0,0]=\sum_{m}Y(X). \tag{6.7}$$

From (6.7) it follows that

$$I\subset\bigcap_{m=1}^{\infty}\sum_{m}Y(X)=\int_a^b Y(x)\,dx. \tag{6.8}$$

Comparison of (6.3) and (6.8) yields (4.4). Q.E.D.

This result can also be established using the relationships expressed in terms of elementary integrals of step functions as upper and lower limits of the interval Darboux sums [1] as in [2, pp. 54–56].

7. A rate of convergence for smooth integrands. As in ordinary interval analysis, an interval function $Y=[\underline{y},\bar{y}]$ is *continuous* if the real functions $\underline{y},\bar{y}$ are continuous. Similarly, Y is *Lipschitz continuous* if a Lipschitz constant $L>0$ exists for both $\underline{y}$ and $\bar{y}$, that is, $|\underline{y}(x)-\underline{y}(z)|\le L|x-z|$ and $|\bar{y}(x)-\bar{y}(z)|\le L|x-z|$ for $x,z\in X$.

Interval integrals of continuous interval functions can be expressed in terms of the Riemann (R) integrals of their endpoint functions [1]:

$$\int_a^b Y(x)\,dx=\left[(\mathrm{R})\int_a^b \underline{y}(x)\,dx,\ (\mathrm{R})\int_a^b \bar{y}(x)\,dx\right]. \tag{7.1}$$

One may also write

$$(\mathrm{R})\int_a^b y(x)\,dx=\sum_{i=1}^{n}(\mathrm{R})\int_{x_{i-1}}^{x_i} y(x)\,dx \tag{7.2}$$

for a Riemann integrable function y and a given partition Δ_n of $X=[a,b]$. If y is continuous, then on each subinterval X_i, $i=1,2,\cdots,n$,

$$(\mathrm{R})\int_{x_{i-1}}^{x_i} y(x)\,dx=y(\xi_i)(x_i-x_{i-1})=y(\xi_i)w(X_i),\qquad \xi_i\in X_i \tag{7.3}$$

[2, p. 209]. Furthermore,

$$\nabla y(X_i)=[c_i,d_i]=[y(\eta_i),y(\zeta_i)],\qquad \eta_i,\zeta_i\in X_i. \tag{7.4}$$

Thus, if y is Lipschitz continuous, then

$$d_i\cdot w(X_i)-(\mathrm{R})\int_a^b y(x)\,dx=[y(\zeta_i)-y(\xi_i)]\cdot w(X_i)\le L\cdot w(X_i)^2 \tag{7.5}$$

and

$$(\mathrm{R})\int_a^b y(x)\,dx-c_i\cdot w(X_i)=[y(\xi_i)-y(\eta_i)]\cdot w(X_i)\le L\cdot w(X_i)^2. \tag{7.6}$$

Applying (7.5) and (7.6) to $\bar{y}$ and y respectively for the equipartition with $w(X_i)=w(X)/n$ gives the following inequality for the excess width of $\overline{\sum}_n Y(X)$ over the interval integral (7.1) of Y.

THEOREM 7.1. *If Y is a Lipschitz continuous interval function, then*

$$e\left(\overline{\sum_n} Y(X),\ \int_a^b Y(x)\,dx\right)\le\frac{L\cdot w(X)}{n}. \tag{7.7}$$

The use of interval Darboux sums as approximations to the interval integral is an extension of the crude method of upper and lower Riemann sums [5] for the approximation of the integral of a Riemann integrable real function. The Darboux sums are generally easy to inclose and give rigorous upper and lower bounds for the value of the integral, but the rate of convergence as given by (7.7) is slow. Of course, the use of partitions other than the equipartition may be of benefit in some cases, but for smooth functions, the improvement may be marginal. For example, for

$$Y(x)=[0,3x^2],\qquad \int_0^1[0,3x^2]\,dx=[0,1], \tag{7.8}$$

the equipartition for $n=2$ gives

$$\overline{\sum_2} Y([0,1])=\tfrac{1}{2}\left[0,\frac{3}{4}\right]+\tfrac{1}{2}[0,3]=[0,\tfrac{15}{8}]=[0,1.875]. \tag{7.9}$$

The interval Riemann sum in this case corresponds to the use of the partition point $x_1=1/\sqrt{3}$ and has the value

$$\sum_2 Y([0,1])=\frac{[0,1]}{\sqrt{3}}+[0,3]\cdot\left(1-\frac{1}{\sqrt{3}}\right)=\left[0,3-\frac{2}{\sqrt{3}}\right]\subset[0,1.846]. \tag{7.10}$$

Although this is better than (7.9), extra labor was required to determine the optimal partition, and this additional effort increases rapidly with n.

8. Inner improper interval integrals. In [1] an interval integral was said to be unbounded if its value is an infinite interval. These unbounded interval integrals arise if the integrand or the interval of integration is unbounded. Relationships were developed in [1] between the value of the finite endpoint of a semi-infinite, or *improper* interval integral and the improper Riemann integral of the corresponding endpoint function of the integrand. Here, an approach to improper interval integration will be made via finitely defined interval integrals.

Case I. $Y(x)$ is an unbounded interval function on a finite interval of integration $X=[a,b]$; that is, $\nabla Y(X)=+\infty$, $w(X)<+\infty$. Here, the functions

$$Y_N(x)=Y(x)\cap[-N,N] \tag{8.1}$$

are defined for each positive integer N. The corresponding finitely defined interval integrals

$$I_N Y(X)=\int_a^b Y_N(x)\,dx, \qquad N=1,2,3,\cdots, \tag{8.2}$$

are finite and may be obtained from (4.4). For $M>N$,

$$I_N Y(X)\subset I_M Y(X), \tag{8.3}$$

because the interval integral is inclusion monotone, and $Y_N(X)\subset Y_M(X)$ for $M>N$ [1]. The *inner improper interval integral* in this case is defined to be

$$(\mathrm{I})\int_a^b Y(x)\,dx=\lim_{N\to\infty} I_N Y(X)\subset\int_a^b Y(x)\,dx, \tag{8.4}$$

the inclusion following again from $Y_N(X)\subset Y(X)$ and inclusion monotonicity of the interval integral. It follows that the inner improper interval integral exists (in the extended real number system) if the interval of integration is finite. The following examples are taken from [1].

(a) $Y(x)=x^{-1/3}$, a real function, $X=[0,1]$.

$$I_N Y([0,1])=\int_0^{N^{-3}} N\,dx+\int_{N^{-3}}^1 x^{-1/3}\,dx=\tfrac{1}{2}[3-N^{-2},3-N^{-2}]. \tag{8.5}$$

Thus,

$$(\mathrm{I})\int_0^1 x^{-1/3}\,dx=\left[\frac{3}{2},\frac{3}{2}\right]\subset\int_0^1 x^{-1/3}\,dx=\left[\frac{3}{2},\infty\right]. \tag{8.6}$$

(b) $Y(x)=x^{-1}$, $X=[0,1]$. Here,

$$I_n Y([0,1])=\int_0^{N^{-1}} N\,dx+\int_{N^{-1}}^1 x^{-1}\,dx=[1+\ln N,\ 1+\ln N] \tag{8.7}$$

and

$$(\mathrm{I})\int_0^1 x^{-1}\,dx=[\infty,\infty]=\int_0^1 x^{-1}\,dx, \tag{8.8}$$

an infinite integral. The standard definition of the improper Riemann (IR) integral of real functions over a finite interval ([2, p. 88]) gives the following result.

THEOREM 8.1. *If the endpoint functions $\underline{y},\bar{y}$ of Y have improper Riemann* (IR) *integrals over the finite interval $X=[a,b]$, then*

$$(8.9)\qquad (\mathrm{I})\int_a^b Y(x)\,dx=\left[(\mathrm{IR})\int_a^b \underline{y}(x)\,dx,\ (\mathrm{IR})\int_a^b \bar{y}(x)\,dx\right].$$

Of course, in case Y is bounded, or the real function y is bounded and Riemann (R) integrable, one may take

$$(8.10)\qquad (\mathrm{I})\int_a^b Y(x)\,dx=\int_a^b Y(x)\,dx,\qquad (\mathrm{IR})\int_a^b y(x)\,dx=(\mathrm{R})\int_a^b y(x)\,dx,$$

respectively.

Finitely defined interval integrals may also be used to construct an improper integral over infinite intervals of integration. For simplicity of notation, take $Y(x)=[0,0]$ outside X and the interval of integration to be the real line $R=[-\infty,\infty]$.

DEFINITION 8.1. *If*

$$(8.11)\qquad I_+Y=\lim_{N\to\infty}(\mathrm{I})\int_0^N Y(x)\,dx,\qquad I_-Y=\lim_{N\to-\infty}(\mathrm{I})\int_N^0 Y(x)\,dx$$

exist, then the *improper interval integral* of Y over $R=[-\infty,\infty]$ is defined to be

$$(8.12)\qquad (\mathrm{I})\int_{-\infty}^{\infty} Y(x)\,dx=I_+Y+I_-Y.$$

Justification. By use of the rules for extended interval arithmetic given in [1], the interval (8.12) is well defined if the limits (8.11) exist, as the formulas $[\infty-\infty,\cdot]=[-\infty,\cdot]$, $[\cdot,\infty-\infty]=[\cdot,\infty]$ resolve any "indeterminant forms" which may arise. The actual interval of integration may be indicated in (8.12) if different from R.

The following example is also taken from [1].

(c) $Y(x)=-e^{-x}$, $X=[0,\infty]$. Here,

$$(8.13)\qquad (\mathrm{I})\int_0^N(-e^{-x})\,dx=\int_0^N(-e^{-x})\,dx=[-1+e^{-N},-1+e^{-N}],$$

and, since $I_-Y=[0,0]$,

$$(8.14)\qquad (\mathrm{I})\int_0^{\infty}(-e^{-x})\,dx=I_+Y=[-1,-1],$$

a finite interval, while the value of the interval integral [1] is the infinite interval

$$(8.15)\qquad \int_0^{\infty}(-e^{-x})\,dx=[-\infty,-1].$$

Finally, the definition of the improper Riemann integral over an infinite interval of integration ([2, p. 94]) gives the following result.

THEOREM 8.2. *If the endpoint functions $\underline{y},\bar{y}$ of Y have improper Riemann integrals over $R=[-\infty,\infty]$, then*

$$(8.16)\qquad (\mathrm{I})\int_{-\infty}^{\infty} Y(x)\,dx=\left[(\mathrm{IR})\int_{-\infty}^{\infty}\underline{y}(x)\,dx,\ (\mathrm{IR})\int_{-\infty}^{\infty}\bar{y}(x)\,dx\right].$$

9. Acknowledgments. The author is grateful to Ole Caprani, Kaj Madsen and Prof. Dr. Wolfgang Walter for helpful discussions on the subject of this paper.

REFERENCES

[1] O. CAPRANI, K. MADSEN AND L. B. RALL, *Integration of interval functions*, this Journal, 12 (1981), pp. 321–341.
[2] E. J. MCSHANE, *Integration*, Princeton University Press, Princeton, NJ, 1944.
[3] R. E. MOORE, *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1966.
[4] _____, *Methods and Applications of Interval Analysis*, SIAM Studies in Applied Mathematics 2, Society for Industrial and Applied Mathematics, Philadelphia, 1979.
[5] L. B. RALL, *Numerical integration and the solution of integral equations by the use of Riemann sums*, SIAM Rev., 7 (1965), pp. 55–64.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0001 $01.00/0

ON THE EXISTENCE AND UNIQUENESS OF A LOCAL CLASSICAL SOLUTION OF AN INITIAL BOUNDARY-VALUE PROBLEM FOR INCOMPRESSIBLE NONHOMOGENEOUS VISCOUS FLUIDS*

BUI AN TON†

Abstract. The existence of a unique local solution $\{u, \nabla p, \rho, \chi\}$ in the space of Hölder continuous functions of the initial boundary-value problem

$$\rho(x,t)\left\{\frac{\partial u}{\partial t}+u\cdot\nabla u\right\}-\nabla\{\chi(x,t)\nabla u\}+\operatorname{grad} p=\rho f, \qquad \nabla\cdot u=0 \quad \text{on } G\times(0,T),$$
$$u(x,t)=0 \quad \text{on } \partial G\times(0,T), \qquad u(x,0)=0 \quad \text{on } G,$$

and of the initial-value problem

$$\frac{\partial\rho}{\partial t}+u\cdot\operatorname{grad}\rho=0, \qquad \rho(x,t)>0 \quad \text{on } G\times(0,T), \qquad \rho(x,0)=\rho_0(x) \quad \text{on } G$$

with

$$\frac{\partial\chi}{\partial t}+u\cdot\operatorname{grad}\chi=0, \qquad \chi(x,t)>0 \quad \text{on } G\times(0,T), \qquad \chi(x,0)=\chi_0(x,\rho_0(x))$$

is shown. G is a bounded open subset of R^3. The method of successive approximations and Lagrangian coordinates as developed by Solonnikov are used.

Introduction. The purpose of this paper is to show the existence of a unique local solution in the space of functions with Hölder continuous derivatives of an initial boundary-value problem for an incompressible *nonhomogeneous* viscous fluid.

Let u, ρ, χ be the velocity, the density and the viscosity of the fluid respectively. The motion of the fluid is described by the initial boundary-value problem

$$\begin{aligned}&\rho\left\{\frac{\partial u}{\partial t}+(u\cdot\nabla)u\right\}-\nabla\{\chi(c,t)\nabla u\}+\nabla p=\rho f \quad \text{on } Q_T=G\times(0,T),\\ &\nabla\cdot u=0 \quad \text{on } Q_T, \qquad u(x,t)=0 \quad \text{on } S_T=\partial G\times(0,T),\\ &u(x,0)=0 \quad \text{on } G,\end{aligned} \tag{0.1}$$

where G is a bounded open subset of R^3 with a smooth boundary ∂G.

The conservation of mass is expressed by the initial-value problem

$$\frac{\partial\rho}{\partial t}+u\cdot\nabla\rho=0, \qquad \rho(x,t)>0 \quad \text{on } Q_T, \qquad \rho(x,0)=\rho_0(x)>0 \quad \text{on } G. \tag{0.2}$$

We shall assume that the viscosity does not change along a fluid particle path, i.e.,

$$\frac{\partial\chi}{\partial t}+u\cdot\nabla\chi=0, \qquad \chi(x,t)>0 \quad \text{on } Q_T, \qquad \chi(x,0)=\chi_0(x,\rho_0(x)) \quad \text{on } G \tag{0.3}$$

where χ_0 is the initial viscosity.

The system (0.1)–(0.3) is the well-known Navier-Stokes equations when ρ and χ are both constants. For nonhomogeneous fluids with *constant viscosity*, the pioneering work is due to Kajikov [1]. He proved the existence of a weak global solution as well as

* Received by the editors January 14, 1981, and in final revised form July 15, 1981.
† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6K 2R7.

that of a strong solution of (0.1)–(0.2), the problem of the unicity of the strong solution remaining open. For a precise definition of weak and strong solutions of (0.1)–(0.2) we refer to [5].

In [5] Lions has shown the existence of a weak global solution of (0.1)–(0.2) using the Galerkin method and estimates of Kajikov's type when χ is *constant*. By a different approach using the method of successive approximations and the compensated compactness argument of Murat, the writer has proved in [9] the existence of a weak global solution of (0.1)–(0.2) when χ is a function of ρ. It is a special case of (0.1)–(0.3) and includes the problem of a mixture of two liquids with varying concentrations.

For nonhomogeneous fluids with *constant* viscosity, using L^p-estimates with $p>3$, Ladyženskaya and Solonnikov [4] have shown the existence of a *unique* strong local solution of (0.1)–(0.2). When G is a bounded open subset of R^2, Kajikov has proved in [2] the existence of a unique global solution of (0.1)–(0.2) by applying an inequality due to Rabinowitz.

In this paper we shall establish the existence of a unique local solution $\{u, \operatorname{grad} p, \rho, \chi\}$ of (0.1)–(0.3) in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})\times C^{\alpha,\alpha/2}(Q_{T^*})\times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$, where $Q_{T^*}=G\times(0,T^*)$ for some $0<T^*<T$. We shall use the method of successive approximations and Lagrangian coordinates. The approach taken in this paper has been used by the writer in [10] for the equations for the theory of shallow waters and by Tani [7] for the equations of the theory of compressible fluids. It seems new in the case of incompressible fluids, even for the Navier-Stokes equations, and allows us to prove the existence of Hölder continuous solutions without invoking the cumbersome machinery of the theory of hydrodynamic potentials. In the use of the Lagrangian coordinates we shall follow Solonnikov's approach.

In §1, the notation and the main results of the paper as well as a detailed outline of the proof of the main theorem are given. The transformation relating Lagrangian and Eulerian coordinates is studied in §2. The existence of a unique global solution of a simple linear initial boundary-value problem is proved in §3. We consider in §4, a linear initial boundary-value problem of the type (0.1) with ∇ replaced by $\nabla_w=A(w)\nabla$, where $A(w)$ is a matrix related to the transformation from Eulerian to Lagrangian coordinates. The proof of the main result is carried out in §5.

1. Let G be an open subset of R^3 with a smooth boundary ∂G of class $C^{2+\alpha}$. Set $D_j=\partial/\partial x_j$, $1\le j\le 3$ and denote by: $Q_T=G\times(0,T)$, $S_T=\partial G\times(0,T)$, $0<\alpha<1$,

$$H_x^{\alpha}(u;Q_T)=\sup_{\substack{(x,t)\neq(y,t)\\(x,t),(y,t)\text{ in }Q_T}}\{|u(x,t)-u(y,t)|\,|x-y|^{-\alpha}\},$$

$$H_t^{\alpha}(u;Q_T)=\sup_{\substack{t\neq s\\(x,t),(x,s)\text{ in }Q_T}}\{|u(x,t)-u(x,s)|\,|t-s|^{-\alpha}\}.$$

By $C^{\alpha,\alpha/2}(Q_T)$ we mean the space of functions $u(x,t)$ in $C(\bar{Q}_T)$ having finite norm

$$\|u\|_{C^{\alpha,\alpha/2}(Q_T)}=\|u\|_{C(Q_T)}+H_x^{\alpha}(u;Q_T)+H_t^{\alpha/2}(u;Q_T).$$

Similarly with u and ∇u in $C(\bar{Q}_T)$, we denote by

$$\|u\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}=\|u\|_{C(Q_T)}+\sum_{j=1}^{3}\|D_ju\|_{C^{\alpha,\alpha/2}(Q_T)}+H_t^{(1+\alpha)/2}(u;Q_T)$$

when the right-hand side is finite. Set

$$\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}=\|u\|_{C(Q_T)}+\|\nabla u\|_{C(Q_T)}+\|\nabla u\|_{C(Q_T)}+\left\|\frac{\partial u}{\partial t}\right\|_{C^{\alpha,\alpha/2}(Q_T)}$$
$$+\sum_{j=1}^{3}H_t^{(1+\alpha)/2}(D_j u;Q_T)+\sum_{j,k=1}^{3}\|D_jD_k u\|_{C^{\alpha,\alpha/2}(Q_T)}.$$

The main result of the paper is the following theorem.

THEOREM 1.1. *Let f be a vector function in $C^{\alpha,\alpha/2}(Q_T)$, ρ_0 be a scalar function in $C^{2+\alpha}(G)$ with*

$$0<m=\text{g.l.b.}_G\,\rho_0\le\rho_0(x)\le\text{l.u.b.}_G\,\rho_0=M.$$

Let χ_0 be a scalar function in $C^{2+\alpha}([m,M]\times G)$ with

$$0<c_1\le\text{g.l.b.}_{[m,M]\times G}\,\chi_0\le\chi_0(n,x)\le\text{l.u.b.}_{[m,M]\times G}\,\chi_0\le c_2.$$

Then there exist:

(i) *a nonempty interval $(0,T^*)$ with $T^*\le T$,*

(ii) *a unique $\{u,\operatorname{grad}p,\rho,\chi\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})\times C^{\alpha,\alpha/2}(Q_{T^*})\times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})\times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$,*

a solution of (0.1)–(0.3).

Remarks. 1) It will be clear from the proof of the theorem that if $\{f,\rho_0,\chi_0\}$ is in $C^{k+\alpha,(k+\alpha)/2}(Q_T)\times C^{2+k+\alpha}\times C^{2+k+\alpha}([m,N]\times G)$ for k a positive integer, then $\{u,\operatorname{grad}p,\rho,\chi\}$ will be in the appropriate space.

2) The general problem of nonhomogeneous viscous heat conducting fluids with nonzero diffusion coefficient is still open.

We shall now give a detailed outline of the proof of Theorem 1.1.

Step 1. It will be carried out in §2. The basic transformation

$$x=\xi+\int_0^t w(\xi,s)\,ds=X(\xi,t) \tag{1.1}$$

relating Eulerian coordinates $x=(x_1,x_2,x_3)$ to Lagrangian coordinates $\xi=(\xi_1,\xi_2,\xi_3)$ will be studied. Let

$$a^{jk}(\xi,t)=\delta_{jk}+\int_0^t\frac{\partial}{\partial\xi_j}w_k(\xi,s)\,ds,\qquad 1\le j,k\le 3.$$

Conditions on w and on T so that the matrix $U(\xi,t)=((a^{jk}(\xi,t)))$ has an inverse are given. Some simple estimates on U and on its inverse are established.

Step 2. In §3 we study the initial boundary-value problem

$$\rho_0\frac{\partial v}{\partial t}-\nabla\{\chi_0\nabla v\}+\nabla q=\rho_0 f,\qquad \nabla\cdot v=0\quad\text{on } Q_T,$$
$$v(\xi,t)=0\quad\text{on } S_T,\qquad v(\xi,0)=0\quad\text{on } G.$$

Step 3. Let $(T+T^\beta)\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\le\delta, 0<2\beta<1-\alpha$ with δ small. We consider in §4 the linear initial boundary value problem

$$\rho_0\frac{\partial v}{\partial t}-\nabla_w\{\chi_0\nabla_w v\}+\nabla_w q=\rho_0 f,\qquad \nabla_w\cdot v=g\quad\text{on } Q_T,$$
$$v(\xi,t)=0\quad\text{on } S_T,\qquad v(\xi,0)=0\quad\text{on } G.$$

∇_w is the operator

$$A(w)\nabla = U^{-1}\nabla = \sum_{j=1}^{3} a_{jk}(\xi,t)\frac{\partial}{\partial \xi_j}, \qquad 1\le k\le 3,$$

where a_{jk} are the entries of U^{-1}.

Step 4. We construct a sequence $\{v^n, \nabla q^n\}$, solutions of the initial boundary-value problems

$$\rho_0 \frac{\partial v^n}{\partial t} - \nabla_{n-1}\{\chi_0 \nabla_{n-1} v^n\} + \nabla_{n-1} q^n = \rho_0 f, \qquad \nabla_{n-1}\cdot v^n = 0 \quad \text{on } Q_T,$$
$$v^n(\xi,t) = 0 \quad \text{on } S_T, \qquad v^n(\xi,0) = 0 \quad \text{on } G \quad \text{for } n = 1,2,\cdots,$$

where $\nabla_n = A(v^n)\nabla$.

With the estimates obtained from Step 3, we show that there exist:

(i) a nonempty interval $(0, T^*)$ independent of n,

(ii) a constant K independent of n such that

$$\|v^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})} + \|\nabla q^n\|_{C^{\alpha,\alpha/2}(Q_{T^*})} \le K.$$

Let $n \to +\infty$ and we have $\{v, \nabla q\}$, a solution of

$$\rho_0 \frac{\partial v}{\partial t} - \nabla_v\{\chi_0 \nabla_v v\} + \nabla_v q = \rho_0 f, \qquad \nabla_v \cdot v = 0 \quad \text{on } Q_{T^*},$$
$$v(\xi,t) = 0 \quad \text{on } S_{T^*}, \qquad v(\xi,0) = 0 \quad \text{on } G.$$

Returning to the Eulerian coordinates via the transformation (1.1) we get the theorem.

2. Let w be a vector function in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ with $w = 0$ on S_T, and consider the one-parameter family of transformations

$$x = \xi + \int_0^t w(\xi,s)\,ds = X(\xi,t) \tag{2.1}$$

of G into G_t. Set

$$a^{jk}(\xi,t) = \delta_{jk} + \int_0^t \frac{\partial}{\partial \xi_j} w_k(\xi,s)\,ds, \qquad 1\le j,k\le 3, \tag{2.2}$$

where δ_{jk} is the Kronecker delta function.

The matrix $U(\xi,t) = ((a^{jk}(\xi,t)))$ is the Jacobian of the transformation X connecting Lagrangian coordinates ξ to Eulerian coordinates x. In this section we shall study U. It is known that without any further condition on w $\det(U) \ne 0$ only for small t. We express that restriction by assuming

$$(T + T^\beta)\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)} \le \delta < \frac{1}{8}, \qquad 0 < 2\beta < 1-\alpha. \tag{2.3}$$

PROPOSITION 2.1. *Suppose that* (2.3) *is verified. Then*

$$\frac{1}{2} \le \det(U(\xi,t)) \le \frac{3}{2} \quad \textit{for } 0\le t\le T.$$

Proof. We have

$$\delta_{jk} - t \max_{j,k} \left\| \frac{\partial}{\partial \xi_j} w_k \right\|_{C(Q_t)} \le a^{jk}(\xi,t) \le \delta_{jk} + t \max_{j,k} \left\| \frac{\partial}{\partial \xi_j} w_k \right\|_{C(Q_t)}.$$

With (2.3) we get

$$\frac{1}{2}\leq 1-3\delta-6\delta^2-6\delta^3\leq \det(U(\xi,t))\leq 1+3\delta+6\delta^2+6\delta^3\leq\frac{3}{2}.$$

LEMMA 2.1. *Let w be in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ and satisfy* (2.3). *Let $a^{jk}(\xi,t)$ be given by* (2.2) *and $U(\xi,t)=((a^{jk}(\xi,t)))$. Then*

1) $\|U-I\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\leq 2\delta,$

2) $\|A\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\leq C_1,$

3) $\|A-I\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\leq C_2\|U-I\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\leq 2C_2\delta.$

C_1 and C_2 are independent of δ and of w. $A(\xi,t)$ denotes $(U^*)^{-1}$.

Proof. This is Solonnikov [6, Lemma 4, p. 1332].

LEMMA 2.2. *Let v,w be two vector functions in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ satisfying* (2.3). *Let U_v, U_w be the Jacobian of the transformation* (2.1) *corresponding to v and to w respectively. Then for any $\varepsilon>0$ there exists $c(\varepsilon)>0$ such that*

$$\|U_v-U_w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\leq\varepsilon\|v-w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+c(\varepsilon)\int_0^t\|v-w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}ds.$$

Similarly

$$\|A_v-A_w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\leq\varepsilon\|v-w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+c_1(\varepsilon)\int_0^t\|v-w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}ds.$$

$c(\varepsilon)$ and $c_1(\varepsilon)$ are independent of $0<t<T$.

Proof. Cf. [6, Lemma 5, p. 1333].

LEMMA 2.3. *Suppose the hypotheses of Lemma* 2.2 *are satisfied. Then*

1) $$\left\|\frac{\partial}{\partial t}U_w\right\|_{C(Q_t)}+\left\|\frac{\partial}{\partial t}A_w\right\|_{C(Q_t)}\leq c\|\nabla w\|_{C(Q_t)},\qquad 0\leq t\leq T,$$

2) $$\left\|\frac{\partial}{\partial t}(A_v-A_w)\right\|_{C(Q_t)}\leq\varepsilon\|v-w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+c(\varepsilon)\int_0^t\|v-w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}ds$$

if $v(\xi,0)=w(\xi,0)$.

Proof. Cf. [6, Lemma 5, p. 1333].

LEMMA 2.4. *Let w be as in Lemma* 2.1. *Then*

$$\|\nabla w\|_{C(Q_T)}\leq\|\nabla w(\cdot,0)\|_{C(G)}+t^{(1+\alpha)/2}\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}\leq\delta+\|\nabla w(\cdot,0)\|_{C(G)}$$

for $0\leq t\leq T$.

Proof. Since:

$$\nabla w(\xi,t)=\nabla w(\xi,0)+\{\nabla w(\xi,t)-\nabla w(\xi,0)\},$$

we have

$$\|\nabla w\|_{C(Q_T)}\leq\|\nabla w(\cdot,0)\|_{C(G)}+t^{(1+\alpha)/2}\int_0^t\|\nabla w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_s)}ds$$

$$\leq\|\nabla w(\cdot,0)\|_{C(G)}+t^{(1+\alpha)/2}\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}.$$

LEMMA 2.5. *Let v, w be as in Lemma 2.2 and let f be a vector function in $C^{\alpha,\alpha/2}(Q_T)$. Set*

$$f^w(\xi,t)=f\Big(\xi+\int_0^t w(\xi,s)\,ds,t\Big).$$

Then

(i) $$\|f^w\|_{C^{\alpha,\alpha/2}(Q_t)}\le c_1\|f\|_{C^{\alpha,\alpha/2}(Q_t)},$$

(ii) $$\|f^v-f^w\|_{C^{\alpha,\alpha/2}(Q_t)}\le\varepsilon\|v-w\|_{C^{\alpha,\alpha/2}(Q_t)}+c(\varepsilon)\int_0^t\|v-w\|_{C^{\alpha,\alpha/2}(Q_s)}ds$$

for any $\varepsilon>0$ and $0\le t\le T$.

$c(\varepsilon)$ is independent of t.

Proof. This is [6, Lemma 7, p. 1351].

3. In this section we shall show the existence of a unique solution of the linear initial boundary-value problem

$$\begin{aligned}&\rho_0(\xi)\frac{\partial v}{\partial t}-\nabla\{\chi_0(\xi,\rho_0(\xi))\nabla v\}+\nabla q=\rho_0(\xi)f, \qquad \nabla\cdot v=g \quad \text{on } Q_T,\\ &v(\xi,t)=0 \quad \text{on } S_T, \qquad v(\xi,0)=0 \quad \text{on } G.\end{aligned}\tag{3.1}$$

It is Step 2 of the proof of Theorem 1.1 as outlined in §1.

The main result of the section is the following theorem.

THEOREM 3.1. *Let ρ_0,χ_0 be as in Theorem* 1.1, *let f be a vector function in $C^{\alpha,\alpha/2}(Q_T)$ and g be a scalar function in $C^{1+\alpha,(1+\alpha)/2}(Q_T)$. Suppose that*

(i) $\frac{\partial g}{\partial t}=\nabla\cdot b+h$ *in the generalized sense on Q_T with the vector function b in $C^{\alpha,\alpha/2}(Q_T)$ and the scalar function h in $C^{\alpha,\alpha/2}(Q_T)$,*

(ii) $g(\xi,0)=0$ *and* $\int_G g(\xi)\,d\xi=0$.

Then there exists a unique solution $\{v,\nabla q\}$ of (3.1) *in $C^{2+\alpha,(2+\alpha)/2}(Q_T)\times C^{\alpha,\alpha/2}(Q_T)$. Moreover,*

$$\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+\|\nabla q\|_{C^{\alpha,\alpha/2}(Q_T)}\le K\mathcal{R}(f,g,b,h;T).$$

K depends only on the bounds of ρ_0,χ_0. $\mathcal{R}(f,g,b,h;T)$ is the expression

$$\|f\|_{C^{\alpha,\alpha/2}(Q_T)}+\|g\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}+\|b\|_{C^{\alpha,\alpha/2}(Q_T)}+\|h\|_{C^{\alpha,\alpha/2}(Q_T)}.$$

First we have the proposition

PROPOSITION 3.1. *Let g be as in Theorem* 3.1. *Then there exists u in $C^{2+\alpha,(2+\alpha)/2}_{(Q_T)}$ such that*

$$\nabla\cdot u=g \quad \text{on } G, \qquad u\cdot n=0 \quad \text{on } \partial G;$$

n is the unit exterior normal vector to ∂G. Moreover,

$$\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\le K\mathcal{R}(0,g,b,h;T).$$

Proof. Consider the Neumann problem

$$\Delta\phi=g \quad \text{on } G, \qquad \frac{\partial\phi}{\partial n}=0 \quad \text{on } \partial G.$$

There exists a unique $u=\nabla\phi$, solution of the problem and

$$\|u(\cdot,t)\|_{C^{2+\alpha}(G)}\le K\|g(\cdot,t)\|_{C^{1+\alpha}(G)}.$$

K is independent of t. It is easy to see that

$$u=\nabla\phi=\nabla_\xi\left(\int_G N(\xi,y)g(y,t)\,dy\right),$$

where $N(\xi,y)$ is the Neumann function whose singularity is of the form $|\xi-y|^{-1}$. Since

$$\frac{\partial g}{\partial t}=\nabla\cdot b+h,$$

$$\frac{\partial u}{\partial t}=\nabla_\xi\left(-\int_G \nabla_y N(\xi,y)b(y,t)\,dy+\int_{\partial G} N(\xi,y)b\cdot n\,dy+\int_G N(\xi,y)h(y,t)\,dy\right).$$

It follows from the Hopf-Korn inequality

$$\left\|\frac{\partial u}{\partial t}\right\|_{C^{\alpha,\alpha/2}(Q_T)}\le K\left\{\|b\|_{C^{\alpha,\alpha/2}(Q_T)}+\|h\|_{C^{\alpha,\alpha/2}(Q_T)}\right\}.$$

Combining this with the above estimate we obtain the stated result.

PROPOSITION 3.2. *Let u be as in Proposition* 3.1. *Then there exists a vector function ω in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ such that*

$$\nabla\cdot\omega=0 \quad \text{on } Q_T, \qquad \omega=-u \quad \text{on } S_T.$$

Moreover,

$$\|\omega\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\le K\mathcal{R}(0,g,b,h;T).$$

Proof. First let us note that $\omega\cdot n=u\cdot n=0$ on ∂G. The compatibility condition is verified. The proposition is reduced to the standard construction of a solenoidal vector taking prescribed values on the boundary. Cf. [3]. It is known that such ω exists and

$$\|\omega\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\le K\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}.$$

Taking into account Proposition 3.1, we get the stated result. Set

$$F=\rho_0 f-\rho_0\frac{\partial}{\partial t}(u+\omega)+\nabla\{\chi_0\nabla(u+\omega)\}. \tag{3.2}$$

Applying Propositions 3.1–3.2 we have

$$\|F\|_{C^{\alpha,\alpha/2}(Q_T)}\le K\mathcal{R}(f,g,b,h;T). \tag{3.3}$$

With $v=w+u+\omega$, the initial boundary-value problem (3.1) may be rewritten as:

$$\begin{aligned}&\rho_0\frac{\partial w}{\partial t}-\nabla\{\chi_0(\xi,\rho_0(\xi))\nabla w\}+\nabla q=F, \qquad \nabla\cdot w=0 \quad \text{on } Q_T,\\ &w(\xi,t)=0 \quad \text{on } S_T, \qquad w(\xi,0)=-u(\xi,0)-\omega(\xi,0) \quad \text{on } G.\end{aligned} \tag{3.4}$$

First let us consider the special case when ρ_0,χ_0 are both constants.

LEMMA 3.1. *Let u,ω be as in Propositions* 3.1–3.2 *respectively. Suppose that ρ_0,χ_0 are two positive constants. Then there exists a unique solution $\{w,\nabla q\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_T)\times C^{\alpha,\alpha/2}(Q_T)$ of* (3.4). *Moreover,*

$$\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+\|\nabla q\|_{C^{\alpha,\alpha/2}(Q_T)}\le K\mathcal{R}(f,g,b,h;T).$$

Proof. The lemma follows from well-known results of the theory of Stokes equations and from (3.3).

We now establish an a priori estimate for the solutions of (3.4).

LEMMA 3.2. *Suppose all the hypotheses of Theorem* 3.1 *are satisfied and suppose that* $\{w, \nabla q\}$ *is a solution of* (3.4) *in the space* $C^{2+\alpha,(2+\alpha)/2}(Q_T)\times C^{\alpha,\alpha/2}(Q_T)$. *Then*

$$\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+\|\nabla q\|_{C^{\alpha,\alpha/2}(Q_T)}\leq K\mathfrak{R}(f,g,b,h;T).$$

Proof. 1) Let $\{\zeta_k\}$ be a finite partition of unity corresponding to an open covering $\{G_k\}$ of G with $\operatorname{supp}(\zeta_k)$ in G_k and $\operatorname{diam}(G_k)\leq\lambda$. Let ξ^k be a fixed but arbitrary point of $G_k\cap G$ and set

$$\rho_0^k=\rho_0(\xi^k),\quad \chi_0^k=\chi_0(\xi^k,\rho_0^k),\quad w_k=\zeta_k w,\quad F_k=\zeta_k F,\quad q_k=\zeta_k q.$$

We have from (3.4)

$$\begin{aligned}&\rho_0^k\frac{\partial}{\partial t}w_k-\nabla\{\chi_0^k\nabla w_k\}+\nabla q_k=\mathfrak{F}_k,\qquad \nabla\cdot w_k=\nabla\zeta_k\cdot w\quad\text{on } Q_T,\\ &w_k(\xi,t)=0\quad\text{on } S_T,\qquad w_k(\xi,0)=\zeta_k w(\xi,0)\quad\text{on } G,\end{aligned}\tag{3.5}$$

with

$$\mathfrak{F}_k=F_k+(\rho_0^k-\rho_0)\frac{\partial}{\partial t}w_k-q\nabla\zeta_k-\nabla\xi_k\chi_0\nabla w-\nabla\{\chi_0\nabla\zeta_k\cdot w\}+\nabla\{(\chi_0-\chi_0^k)\nabla w_k\}.$$

2) Since

$$\frac{\partial}{\partial t}(w\cdot\nabla\zeta_k)=\nabla\zeta_k\cdot\{F-\rho_0^{-1}\nabla g+\rho_0^{-1}\nabla(\chi_0\nabla w)\},$$

we may write it as $\nabla\cdot\tilde{b}+\tilde{h}$, with

$$\tilde{b}_j=-g\rho_0^{-1}D_j\zeta_k+\sum_{n=1}^{3}\chi_0\rho_0^{-1}D_jw_nD_n\zeta_k,\qquad 1\leq j\leq 3,$$

$$\tilde{h}=\rho_0^{-1}F\cdot\nabla\zeta_k+g\nabla\cdot\{\rho_0^{-1}\nabla\zeta_k\}-\sum_{j,n=1}^{3}D_j\{\chi_0\rho_0^{-1}D_j\zeta_k\}D_jw_n.$$

An elementary computation shows that

$$\begin{aligned}&\|\tilde{h}\|_{C^{\alpha,\alpha/2}(Q_T)}+\|\tilde{b}\|_{C^{\alpha,\alpha/2}(Q_T)}\\ &\qquad\leq K\lambda^{-1}\{\|F\|_{C^{\alpha,\alpha/2}(Q_T)}+\|g\|_{C^{\alpha,\alpha/2}(Q_T)}+\|w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\}.\end{aligned}\tag{3.6}$$

K is independent of λ, k.

Using an argument as in Proposition 3.1–3.2 and then by applying Lemma 3.1, we obtain by a standard argument (after taking into account (3.6))

$$\begin{aligned}&\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+\|\nabla q\|_{C^{\alpha,\alpha/2}(Q_T)}\\ &\qquad\leq K\{\mathfrak{R}(f,g,b,h;T)=\|q\|_{C^{\alpha,\alpha/2}(Q_T)}+\|w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\}.\end{aligned}\tag{3.7}$$

3) We now estimate q. It is clear that q is a solution of the Neumann problem

$$\begin{aligned}&\Delta q=\nabla\cdot\left\{F-\rho_0\frac{\partial w}{\partial t}+\nabla(\chi_0\nabla w)\right\}\quad\text{on } G,\\ &\frac{\partial q}{\partial n}=\left\{F-\rho_0\frac{\partial w}{\partial t}+\nabla(\chi_0\nabla w)\right\}\cdot n\quad\text{on } \partial G.\end{aligned}\tag{3.8}$$

The compatibility condition is satisfied and the solutions of (3.8) are defined up to an additive constant. We shall take the solution with zero constant. An integral representation of the solution is known:

$$q(\xi,t)=-\int_G \nabla_y N(\xi,y)\left\{F-\rho_0\frac{\partial w}{\partial t}+\nabla(\chi_0\nabla w)\right\}dy, \tag{3.9}$$

where $N(\xi,y)$ is the Neumann function whose singularity is of the form $|\xi-y|^{-1}$. An integration by parts yields

$$q(\xi,t)=-\int_G \nabla_y N(\xi,y)\left\{F-\rho_0\frac{\partial w}{\partial t}\right\}dy+\int_G \nabla_y^2 N(\xi,y)\chi_0\nabla w\,dy$$
$$+\int_{\partial G}\nabla_y N(\xi,y)\chi_0\nabla w\cdot n\,dy.$$

Applying the Calderon–Zygmund kernel theorem to the second integral and Young's theorem to the other two integrals, we get

$$\|q\|_{L^2(Q_T)}\le K\left\{\|F\|_{C^{\alpha,\alpha/2}(Q_T)}+\left\|\frac{\partial w}{\partial t}\right\|_{L^2(Q_T)}+\|w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\right\}$$
$$\le K\left\{\mathcal{R}(f,g,b,h;T)+\left\|\frac{\partial w}{\partial t}\right\|_{L^2(Q_T)}+\|w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\right\}.$$

4) Since Q_T is bounded, the following natural injection mappings are all compact:

$$C^{2+\alpha,(2+\alpha)/2}(Q_T)\subset C^{1+\alpha,(1+\alpha)/2}(Q_T)\subset C^{\alpha,\alpha/2}(Q_T)\subset L^2(Q_T).$$

Thus, it is known that for any $\varepsilon>0$ there exists $M(\varepsilon)>0$ such that

$$\|w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\le\varepsilon\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+M(\varepsilon)\|w\|_{L^2(Q_T)}, \tag{3.10}$$

and

$$\|q\|_{C^{\alpha,\alpha/2}(Q_T)}\le\varepsilon\|\nabla q\|_{C^{\alpha,\alpha/2}(Q_T)}+M(\varepsilon)\|q\|_{L^2(Q_T)}. \tag{3.11}$$

Similarly,

$$\left\|\frac{\partial w}{\partial t}\right\|_{L^2(Q_T)}\le\varepsilon\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+M(\varepsilon)\|w\|_{L^2(Q_T)}. \tag{3.12}$$

Taking into account (3.9)–(3.12) we obtain from (3.7)

$$\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+\|\nabla q\|_{C^{\alpha,\alpha/2}(Q_T)}\le K\left\{\mathcal{R}(f,g,b,h;T)+\|w\|_{L^2(Q_T)}\right\}. \tag{3.13}$$

It is rather trivial to check that

$$\|w\|_{L^2(Q_T)}\le K\mathcal{R}(f,g,b,h;T).$$

The lemma is proved.

Proof of Theorem 3.1. 1) Since $\rho_0,\chi_0(\xi,\rho_0(\xi))$ are two time-independent positive functions, the Galerkin method gives the existence of a unique weak solution w of (3.4) with

$$\operatorname*{ess\,sup}_{0\le t\le T}\|w(\cdot,t)\|_{L^2(G)}+\|\nabla w\|_{L^2(Q_T)}$$
$$\le K\left\{\|F\|_{L^2(Q_T)}+\|u\cdot(\cdot,0)\|_{L^2(G)}+\|\omega(\cdot,0)\|_{L^2(G)}\right\}.$$

Let j_ε be the Friedrichs mollifier with respect to t and denote by $w_\varepsilon = j_\varepsilon^* w$, $F_\varepsilon = j_\varepsilon^* F$. It follows from (3.4) that

$$\begin{aligned} &\rho_0 \frac{\partial w_\varepsilon}{\partial t} - \nabla\{\chi_0 \nabla w_\varepsilon\} + \nabla q_\varepsilon = F_\varepsilon, \qquad \nabla \cdot w_\varepsilon = 0 \quad \text{on } Q_T, \\ &w_\varepsilon(\xi,t) = 0 \quad \text{on } S_T, \qquad w_\varepsilon(\xi,0) = -u(\xi,0) - \omega(\xi,0) \quad \text{on } G. \end{aligned} \tag{3.14}$$

We may rewrite it as

$$-\nabla\{\chi_0 \nabla w_\varepsilon\} + \nabla q_\varepsilon = F_\varepsilon - \rho_0 \frac{\partial w_\varepsilon}{\partial t}, \quad \nabla \cdot w = 0 \quad \text{on } G, \qquad w_\varepsilon(\xi,t) = 0 \quad \text{on } \partial G \tag{3.15}$$

for almost all t on $(0,T)$.

The elliptic problem (3.15) satisfies the complementary condition of Agmon, Douglis and Nirenberg; cf. [8, p. 33]. Since $F_\varepsilon = \rho_0 \partial w_\varepsilon/\partial t$ is in $L^2(Q_T)$, w_ε is in $L^2(0,T; W^{2,2}(G))$. On the other hand,

$$\frac{\partial^k}{\partial t^k} F_\varepsilon, \frac{\partial^k}{\partial t^k} w_\varepsilon \quad \text{are in } L^2(Q_T) \text{ for all } k \geq 0.$$

Thus by the above arguments $(\partial^k/\partial t^k) w_\varepsilon$ is in $L^2(0,T; W^{2,2}(G))$ for all $k \geq 0$. Applying the Sobolev imbedding theorem we obtain w_ε and $\partial w_\varepsilon/\partial t$ in $C^{\alpha,\alpha/2}(Q_T)$. Hence $F_\varepsilon - \rho_0 \partial w_\varepsilon/\partial t$ is in $C^{\alpha,\alpha/2}(Q_T)$.

Using now Schauder's estimates for (3.15), we get $\{w_\varepsilon, \nabla q_\varepsilon\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_T) \times C^{\alpha,\alpha/2}(Q_T)$.

Applying Lemma 3.2 we have

$$\| w_\varepsilon \|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)} + \| \nabla q_\varepsilon \|_{C^{\alpha,\alpha/2}(Q_T)} \leq K \mathfrak{R}(F_\varepsilon, g, b, h; T). \tag{3.16}$$

Since

$$\| F_\varepsilon \|_{C^{\alpha,\alpha/2}(Q_T)} \leq c \| F \|_{C^{\alpha,\alpha/2}(Q_T)} \leq K_1 \mathfrak{R}(f, g, b, h; T),$$

we obtain

$$\| w_\varepsilon \|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)} + \| \nabla q \|_{C^{\alpha,\alpha/2}(Q_T)} \leq K \mathfrak{R}(f, g, b, h; T). \tag{3.17}$$

The different K's are all independent of ε.

Let $\varepsilon \to 0$ and we have by taking subsequences

$$\{w_\varepsilon, \nabla q_\varepsilon\} \to \{w, \nabla q\} \quad \text{in } C^{2+\gamma,(2+\gamma)/2}(Q_T) \times C^{\gamma,\gamma/2}(Q_T)$$

for $0 < \gamma < \alpha$. It is clear that $\{w, \nabla q\}$ is a solution of (3.4) and hence $\{v = w + u + \omega, \nabla q\}$ is a solution of (3.1).

In view of (3.17) we also have

$$\| w \|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)} + \| \nabla q \|_{C^{\alpha,\alpha/2}(Q_T)} \leq K \mathfrak{R}(f, g, b, h; T).$$

It is obvious that the solution is unique. The theorem is proved.

4. Let w be in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ with $w(\xi,0) = 0$ on G and

$$(T + T^\beta) \| w \|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)} \leq \delta < \frac{1}{8} \quad \text{for } 0 < 2\beta < 1 - \alpha.$$

Consider the initial boundary-value problem

$$\begin{aligned}&\rho_0\frac{\partial v}{\partial t}-\nabla_w(\chi_0\nabla_w v)+\nabla_w q=\rho_0 f, \qquad \nabla_w\cdot v=g \quad \text{on } Q_T,\\ &v(\xi,t)=0 \quad \text{on } S_T, \qquad v(\xi,0)=0 \quad \text{on } G,\end{aligned} \tag{4.1}$$

where $\nabla_w=A(w)\nabla$ with $A(w)$ as in Lemma 2.1.

In this section we shall show the existence of a unique solution of (4.1). It is Step 3 of the proof of Theorem 1.1 as outlined in §1. The main result of the section is the following theorem.

THEOREM 4.1. *Let f,ρ_0,χ_0 be as in Theorem* 1.1, *let g be as in Theorem* 3.1 *and let w be as above. Then there exists a unique solution $\{v,\nabla q\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_T)\times C^{\alpha,\alpha/2}(Q_T)$ of* (4.1). *Moreover,*

$$\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+\|\nabla q\|_{C^{\alpha,\alpha/2}(Q_T)}\le K\mathcal{R}(f,g,b,h;t).$$

K is independent of t with $0\le t\le T$; it depends on the bounds of ρ_0,χ_0. $\mathcal{R}$ is as in Theorem 3.1.

We shall use the method of successsive approximation. Consider the linear initial boundary-value problems

$$\begin{aligned}&\rho_0\frac{\partial v^n}{\partial t}-\nabla(\chi_0\nabla v^n)+\nabla q^n\\ &\quad=\rho_0 f+(I-A)\nabla q^{n-1}-(I-A)\nabla\{\chi_0\nabla v^{n-1}\}-A\nabla\{\chi_0(I-A)\nabla v^{n-1}\}\\ &\quad=f^n \quad \text{on } Q_T,\\ &\nabla\cdot v^n=g+(I-A)\nabla\cdot v^{n-1}=g^n \quad \text{on } Q_T, \qquad v^n(\xi,t)=0 \quad \text{on } S_T,\\ &v^n(\xi,0)=0 \quad \text{on } G, \quad n=1,2,\cdots.\end{aligned} \tag{4.2}$$

LEMMA 4.1. *Let $\{v^0,\nabla q^0\}$ be an element of $C^{2+\alpha,(2+\alpha)/2}(Q_T)\times C^{\alpha,\alpha/2}(Q_T)$ with $v^0(\xi,0)=0$ on G and $v^0(\xi,t)=0$ on S_T. Suppose all the hypotheses of Theorem* 4.1 *are satisfied. Then for each n there exists a unique $\{v^n,\nabla q^n\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_T)\times C^{\alpha,\alpha/2}(Q_T)$, a solution of* (4.2).

Proof. 1) We have f^1 in $C^{\alpha,\alpha/2}(Q_T)$. Since $A(\xi,0)=I, g^1(\xi,t)=g(\xi,t)+(I-A)\nabla\cdot v^0$ is zero at $t=0$. The compatibility condition

$$\int_G g^1(\xi,t)\,d\xi=0$$

is verified. Indeed, since $v^0(\xi,t)=0$ on S_T, with our hypothesis on g it suffices to check that

$$\int_G A\nabla_\xi\cdot v^0\,d\xi=0.$$

Let $u(x,t)=v^0(X^{-1}(\xi,t),t)$, where

$$x=\xi+\int_0^t w(\xi,s)\,ds=X(\xi,t).$$

Then $A\nabla_\xi\cdot v^0=\nabla_x\cdot u$ and $u(x,t)=0$ on ∂G_t. Therefore

$$\int_G A\nabla_\xi\cdot v^0\,d\xi=0.$$

We have $\partial g^1/\partial t=\partial g/\partial t+(\partial/\partial t)\{(I-A)\nabla\cdot v^0\}$ in the generalized sense on Q_T. We now show that

$$\frac{\partial g^1}{\partial t}=\nabla\cdot b^1+h^1,$$

with b^1, h^1 in $C^{\alpha,\alpha/2}(Q_T)$. Let us note that

$$A\nabla\cdot\phi=\nabla\cdot(A\phi)-\nabla A\cdot\phi,$$

where ∇A is a vector with components $\Sigma_{j=1}^3(\partial/\partial\xi_j)(a_{jk})$, $1\le k\le 3$. So

$$\begin{aligned}\frac{\partial}{\partial t}\{(I-A)\nabla\cdot v^0\}&=\nabla\cdot(I-A)\frac{\partial v^0}{\partial t}-\nabla\cdot\left(\frac{\partial A}{\partial t}v^0\right)+\frac{\partial}{\partial t}(\nabla A)\cdot v^0+\nabla A\cdot\frac{\partial v^0}{\partial t}\\&=\nabla\cdot\left\{(I-A)\frac{\partial v^0}{\partial t}-\frac{\partial A}{\partial t}v^0\right\}+\frac{\partial}{\partial t}(\nabla A)\cdot v^0+\nabla A\cdot\frac{\partial v^0}{\partial t}.\end{aligned}$$

Thus,

$$b^1=b+\left\{(I-A)\frac{\partial v^0}{\partial t}-\frac{\partial A}{\partial t}v^0\right\},\qquad h^1=h+\frac{\partial}{\partial t}(\nabla A)\cdot v^0+\nabla A\cdot\frac{\partial v^0}{\partial t}.$$

2) Applying Lemma 2.1, we have

$$\begin{aligned}\|b^1\|_{C^{\alpha,\alpha/2}(Q_t)}\le\|b\|_{C^{\alpha,\alpha/2}(Q_t)}+\delta\left\|\frac{\partial v^0}{\partial t}\right\|_{C^{\alpha,\alpha/2}(Q_t)}\\+K\left\|\frac{\partial A}{\partial t}\right\|_{C^{\alpha,\alpha/2}(Q_t)}\|v^0\|_{C(Q_t)}+K\|\nabla w\|_{C(Q_t)}\|v^0\|_{C^{\alpha,\alpha/2}(Q_t)}.\end{aligned}\tag{4.3}$$

Since $v^0(\xi,0)=0$, we may write

$$v^0(\xi,t)=\int_0^t\frac{\partial}{\partial s}v^0(\xi,s)\,ds.$$

Thus,

$$\|v^0\|_{C^{\alpha,\alpha/2}(Q_t)}\le\int_0^t\left\|\frac{\partial v^0}{\partial s}\right\|_{C^{\alpha,\alpha/2}(Q_s)}ds\le t\|v^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}.\tag{4.4}$$

Also, with $w(\xi,0)=0$ on G, by Lemma 2.4

$$\|\nabla w\|_{C(Q_t)}\le t^{(1+\alpha)/2}\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}\le\delta.\tag{4.5}$$

Noting that

$$\left\|\frac{\partial A}{\partial t}\right\|_{C^{\alpha,\alpha/2}(Q_t)}\le K\|w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)},\tag{4.6}$$

we have from (4.3) by taking into account (4.4)–(4.6)

$$\begin{aligned}\|b^1\|_{C^{\alpha,\alpha/2}(Q_t)}\le\|b\|_{C^{\alpha,\alpha/2}(Q_t)}+K\delta\Big\{\|v^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}\\+\int_0^t\|v^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}ds\Big\}.\end{aligned}\tag{4.7}$$

K is independent of t.

3) For h^1, we get

$$\begin{aligned}\|h^1\|_{C^{\alpha,\alpha/2}(Q_t)} \le{}& \|h\|_{C^{\alpha,\alpha/2}(Q_t)} + \|\nabla A\|_{C^{\alpha,\alpha/2}(Q_t)} \left\| \frac{\partial v^0}{\partial t} \right\|_{C(Q_t)} \\ &+ \|\nabla A\|_{C(Q_t)} \left\| \frac{\partial v^0}{\partial t} \right\|_{C^{\alpha,\alpha/2}(Q_t)} + \left\| \frac{\partial}{\partial t}(\nabla A) \right\|_{C^{\alpha,\alpha/2}(Q_t)} \|v^0\|_{C(Q_t)} \\ &+ \left\| \frac{\partial}{\partial t}(\nabla A) \right\|_{C(Q_t)} \|v^0\|_{C^{\alpha,\alpha/2}(Q_t)}.\end{aligned} \tag{4.8}$$

Since $\nabla^2 w(\xi,t) = \nabla^2 w(\xi,t) - \nabla^2 w(\xi,0) + \nabla^2 w(\xi,0)$ and $w(\xi,0)=0$ on G, it is easy to check that

$$\left\| \frac{\partial}{\partial t}(\nabla A) \right\|_{C(Q_t)} \le K \|\nabla^2 w\|_{C(Q_t)} \le K t^{\alpha/2} \|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} \le K\delta. \tag{4.9}$$

Also,

$$\left\| \frac{\partial}{\partial t}(\nabla A) \right\|_{C^{\alpha,\alpha/2}(Q_t)} \le K \|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}. \tag{4.10}$$

Moreover,

$$\|\nabla A\|_{C^{\alpha,\alpha/2}(Q_t)} + \|\nabla A\|_{C(Q_t)} \le Kt \|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} \le K\delta. \tag{4.11}$$

From (4.9)–(4.11) we obtain by using (4.8)

$$\|h^1\|_{C^{\alpha,\alpha/2}(Q_t)} \le \|h\|_{C^{\alpha,\alpha/2}(Q_t)} + K\delta \|v^0\|_{C(Q_t)} \|v^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}. \tag{4.12}$$

With (4.4) we get

$$\|h^1\|_{C^{\alpha,\alpha/2(Q_t)}} \le \|h\|_{C^{\alpha,\alpha/2}(Q_t)} + K\delta \|v^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}. \tag{4.13}$$

It is trivial to check that g^1 is in $C^{1+\alpha,(1+\alpha)/2}(Q_t)$.

Applying Theorem 3.1 we get the existence of a unique solution $\{v^1, \nabla q^1\}$ of (4.2). Repeating exactly the same arguments with $\{v^0, \nabla q^0\}$ replaced by $\{v^{n-1}, \nabla q^{n-1}\}$, we get the solution $\{v^n, \nabla q^n\}$. The lemma is proved.

Proof of Theorem 4.1. 1) Set $V^n = v^n - v^{n-1}$, $P^n = q^n - q^{n-1}$. From (4.2) we obtain

$$\begin{aligned}&\rho_0 \frac{\partial V^n}{\partial t} - \nabla\{\chi_0 \nabla V^n\} + \nabla P^n \\ &\quad = (I-A)\nabla P^{n-1} - (I-A)\nabla\{\chi_0 \nabla V^{n-1}\} - A\nabla\{\chi_0 (I-A)\nabla V^{n-1}\} \quad \text{on } Q_T, \\ &\nabla \cdot V^n = (I-A)\nabla \cdot V^{n-1} \quad \text{on } Q_T, \qquad V^n(\xi,0) = 0 \quad \text{on } G, \qquad V^n(\xi,t) = 0 \quad \text{on } S_T.\end{aligned} \tag{4.14}$$

From Theorem 3.1 and from (4.7), (4.13) we have

$$\begin{aligned}&\|V^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + \|\nabla P^n\|_{C^{\alpha,\alpha/2}(Q_t)} \\ &\quad \le K\delta \left\{ \|\nabla P^{n-1}\|_{C^{\alpha,\alpha/2}(Q_t)} + \|V^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + \int_0^t \|V^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)} ds \right\}.\end{aligned}$$

Take $K\delta \le \frac{1}{2}$ and summing from $n=1$ to N we get

$$A_n(t) \le 2\left\{ \|\nabla P^0\|_{C^{\alpha,\alpha/2}(Q_t)} + \|V^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + \int_0^t A_N(s)\, ds \right\},$$

where

$$A_N(t)=\sum_{n=1}^{N}\left\{\|V^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\|\nabla P^n\|_{C^{\alpha,\alpha/2}(Q_t)}\right\}.$$

It follows from the Gronwall lemma that $A_\infty(t)<\infty$ for all $0\le t\le T$. Thus,

$$\{v^n,\nabla q^n\}\to\{v,\nabla q\}\quad\text{in } C^{2+\alpha,(2+\alpha)/2}(Q_T)\times C^{\alpha,\alpha/2}(Q_T).$$

It is trivial to check that $\{v,\nabla q\}$ is a solution of (4.1).

2) We now establish an estimate for $\{v,\nabla q\}$. Applying Lemmas 2.2–2.4 we have

$$\begin{aligned}&\|f^n\|_{C^{\alpha,\alpha/2}(Q_t)}+\|g^n\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\\ &\le K\left\{\|f\|_{C^{\alpha,\alpha/2}(Q_t)}+\|g\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\right\}\\ &\quad+K\delta\left\{\|v^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\|\nabla q^{n-1}\|_{C^{\alpha,\alpha/2}(Q_t)}\right\}.\end{aligned}\tag{4.15}$$

For $\partial g^n/\partial t=\nabla\cdot b^n+h^n$, we obtain as in (4.7) and in (4.13) the estimate

$$\begin{aligned}&\|b^n\|_{C^{\alpha,\alpha/2}(Q_t)}+\|h^n\|_{C^{\alpha,\alpha/2}(Q_t)}\\ &\le K\left\{\|b\|_{C^{\alpha,\alpha/2}(Q_t)}+\|h\|_{C^{\alpha,\alpha/2}(Q_t)}\right\}\\ &\quad+K\delta\left\{\|v^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\int_0^t\|v^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}ds\right\}.\end{aligned}\tag{4.16}$$

Applying Theorem 3.1 we have

$$\begin{aligned}&\|v^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\|\nabla q^n\|_{C^{\alpha,\alpha/2}(Q_t)}\\ &\le K\mathcal{R}(f,g,b,h;t)+K\delta\left\{\|v^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\int_0^t\|v^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}ds\right.\\ &\qquad\left.+\|\nabla q^{n-1}\|_{C^{\alpha,\alpha/2}(Q_t)}\right\}.\end{aligned}\tag{4.17}$$

Let $n\to+\infty$ and take $K\delta\le\frac{1}{2}$. It follows from the first part that

$$\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\|\nabla q\|_{C^{\alpha,\alpha/2}(Q_t)}\le 2K\mathcal{R}(f,g,b,h;t)+\int_0^t\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}ds.$$

It follows from the Gronwall lemma that

$$\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\|\nabla w\|_{C^{\alpha,\alpha/2}(Q_t)}\le K_1\mathcal{R}(f,g,b,h;t).$$

K_1 is independent of t. It is now clear that the solution is unique.

5. In this section we shall carry out the proof of Theorem 1.1. Consider the linear initial boundary value problems

$$\begin{aligned}&\rho_0\frac{\partial v^n}{\partial t}-\nabla_{n-1}\{\chi_0\nabla_{n-1}v^n\}+\nabla_{n-1}q^n=\rho_0 f^{n-1},\qquad \nabla_{n-1}\cdot v^n=g\quad\text{on } Q_T,\\ &v^n(\xi,t)=0\quad\text{on } S_T,\qquad v^n(\xi,0)=0\quad\text{on } G,\qquad n=1,2,\cdots,\end{aligned}\tag{5.1}$$

with

$$\nabla_{n-1}=A(v^{n-1})\nabla_\xi\quad\text{and}\quad f^{n-1}(\xi,t)=f\left(\xi+\int_0^t v^{n-1}(\xi,s)\,ds,t\right).$$

LEMMA 5.1. *Let* f, ρ_0, χ_0 *be as in Theorem* 1.1. *Let* g *be as in Theorem* 3.1 *and* v^0 *be an element of* $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ *with* $v^0(\xi,0)=0$ *on* G, $v^0(\xi,t)=0$ *on* S_T. *Suppose that*

$$(T+T^\beta)\|v^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\leq\delta<\frac{1}{8}, \qquad 0<2\beta<1-\alpha.$$

Then there exist:

(i) *a nonempty interval* $(0,T^*)$ *independent of* n,

(ii) *a unique solution* $\{v^n, \nabla q^n\}$ *of* (5.1) *for each* n.

Moreover,

$$\|v^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})}+\|\nabla q^n\|_{C^{\alpha,\alpha/2}(Q_{T^*})}\leq K\mathcal{R}(f,g,b,h;T^*).$$

K *is independent of* n. *The expression* $\mathcal{R}$ *is as in Theorem* 3.1.

Proof. 1) Let v^0 be as in the lemma. Then it follows from Theorem 4.1 that there exists a unique $\{v^1, \nabla q^1\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_T)\times C^{\alpha,\alpha/2}(Q_T)$, a solution of (5.1). Moreover, by Lemma 2.5 we have:

$$\|v^1\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\|\nabla q^1\|_{C^{\alpha,\alpha/2}(Q_t)}\leq K\mathcal{R}(f,g,b,h;t).$$

K is independent of t. Thus, there exists a nonempty subinterval $(0,T^*)$ such that

$$\{T^*+(T^*)^\beta\}\|v^1\|_{C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})}\leq K\{T^*+(T^*)^\beta\}\mathcal{R}(f,g,b,h;T^*)\leq\delta<\frac{1}{8}.$$

2) Repeating the same argument with v^1 instead of v^0, we get

$$\|v^2\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\|\nabla q^2\|_{C^{\alpha,\alpha/2}(Q_t)}\leq K\mathcal{R}(f,g,b,h;t)$$

for $0\leq t\leq T^*$. Again,

$$\{T^*+(T^*)^\beta\}\|v^2\|_{C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})}\leq K\{T^*+(T^*)^\beta\}\mathcal{R}(f,g,b,h;T^*)\leq\delta<\frac{1}{8}$$

with the same T^* as before.

By induction we get the lemma.

THEOREM 5.1. *Let* f, ρ_0, χ_0 *be as in Theorem* 1.1. *Then there exist*

(i) *a nonempty interval* $(0,T^*)$,

(ii) *a unique* $\{v, \nabla q\}$ *in* $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})\times C^{\alpha,\alpha/2}(Q_{T^*})$, *a solution of the initial boundary-value problem*

$$\begin{aligned}&\rho_0\frac{\partial v}{\partial t}-\nabla_v\{\chi_0(\xi,\rho_0(\xi))\nabla_v v\}+\nabla_v q=\rho_0 f^v, \qquad \nabla_v\cdot v=0 \quad \text{on } Q_{T^*},\\ &v(\xi,t)=0 \quad \text{on } S_T, \qquad v(\xi,0)=0 \quad \text{on } G,\end{aligned} \tag{5.2}$$

with $\nabla_v=A(v)\nabla$, where $A(v)$ is as in Lemma 2.1 and

$$f^v=f\Big(\xi+\int_0^t v(\xi,s)\,ds,t\Big).$$

Proof. 1) Let $\{v^n, \nabla q^n\}$ be as in Lemma 5.1 with $g=0$. It follows from the lemma that

$$\|v^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\|\nabla q^n\|_{C^{\alpha,\alpha/2}(Q_t)}\leq K_1 \tag{5.3}$$

for $0\leq t\leq T^*$. The constant K_1 and the nonempty interval $(0,T^*)$ are independent of n.

Since Q_{T^*} is bounded we have by taking subsequences

$$\{v^n, \nabla q^n\}\to\{v, \nabla q\} \quad \text{in } C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})\times C^{\alpha,\alpha/2}(Q_{T^*})$$

for any $0<\gamma<\alpha$. In view of (5.3) it is clear that $\{v, \nabla q\}$ is also in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})\times C^{\alpha,\alpha/2}(Q_{T^*})$, and it is trivial to check that $\{v, \nabla q\}$ is a solution of (5.2).

2) We now show that the solution is unique. Suppose $\{v, \nabla q\}$ and $\{w, \nabla r\}$ are two solutions of (5.2) with all the properties stated in the theorem. Set $uv-w$ and $p=q-r$. Then we have

(5.4)

$$\rho_0 \frac{\partial u}{\partial t} - \nabla_v\{\chi_0 \nabla_v u\} + \nabla_v p$$
$$= \rho_0(f^v - f^w) + (A(w)-A(v))\nabla r$$
$$+ \nabla_v\{\chi_0(A(v)-A(w))\nabla w\} + (A(v)-A(w))\nabla\{\chi_0 A(w)\nabla w\} \quad \text{on } Q_{T^*},$$
$$\nabla_v \cdot u = (A(w)-A(v))\nabla\cdot w \quad \text{on } Q_{T^*}, \qquad u(\xi,t)=0 \quad \text{on } S_T, \qquad u(\xi,0)=0 \quad \text{on } G.$$

Applying Lemmas 2.2, 2.3 we obtain

(5.5)

$$\begin{aligned}\|(A(w)-A(v))\nabla u\|_{C^{\alpha,\alpha/2}(Q_t)} &+ \|\nabla_v\{\chi_0(A(v)-A(w))\nabla w\}\|_{C^{\alpha,\alpha/2}(Q_t)} \\ &+ \|(A(v)-A(w))\nabla\{\chi_0 A(w)\nabla w\}\|_{C^{\alpha,\alpha/2}(Q_t)} \\ &+ \|(A(w)-A(v))\nabla\cdot w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} \\ &\le L\left\{\varepsilon\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + c(\varepsilon)\int_0^t \|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\right\}.\end{aligned}$$

L is a constant independent of ε and of t.

As shown earlier in Lemma 4.1, the expression $g=(A(w)-A(v))\nabla\cdot w$ satisfies the compatibility conditions of Theorem 3.1. We now consider:

$$\frac{\partial g}{\partial t} = \frac{\partial}{\partial t}\{(A(w)-A(v))\nabla\cdot w\} = \nabla\cdot b + h$$

in the generalized sense on Q_{T^*}.

As in the proof of Lemma 4.1, we have

$$b = (A(w)-A(v))\frac{\partial w}{\partial t} + \frac{\partial}{\partial t}(A(w)-A(v))w,$$

$$h = -\nabla\{(A(w)-A(v))\}\cdot\frac{\partial w}{\partial t} - \nabla\frac{\partial}{\partial t}(A(w)-A(v))\cdot w.$$

Applying Lemmas 2.2–2.4 we obtain

$$\|b\|_{C^{\alpha,\alpha/2}(Q_t)} \le L\left\{\varepsilon\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + c(\varepsilon)\int_0^t \|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds + \|u\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\right\}.$$

L is independent of ε, t. Since $u(\xi,0)=0$ and

$$C^{2+\alpha,(2+\alpha)/2}(Q_t) \subset C^{1+\alpha,(1+\alpha)/2}(Q_t) \subset C(Q_t)$$

are all compact,

$$\|u\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} \le \varepsilon\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + c(\varepsilon)\int_0^t \|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds.$$

Thus

$$\|b\|_{C^{\alpha,\alpha/2}(Q_t)} \le L\left\{\varepsilon\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + c(\varepsilon)\int_0^t \|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\right\}. \tag{5.6}$$

Similarly,

$$\|h\|_{C^{\alpha,\alpha/2}(Q_t)} \leq L\left\{\varepsilon\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + c(\varepsilon)\int_0^t \|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\right\}. \tag{5.7}$$

From Lemma 2.5 we get

$$\|f^v - f^w\|_{C^{\alpha,\alpha/2}(Q_t)} \leq L\left\{\varepsilon\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + c(\varepsilon)\int_0^t \|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\right\}. \tag{5.8}$$

The different constants L are all independent of ε, t.

Applying Theorem 4.1 and taking into account (5.5)–(5.8) we obtain, by a trivial argument,

$$\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + \|\nabla p\|_{C^{\alpha,\alpha/2}(Q_t)} \leq K\int_0^t \|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds,$$

for $0 \leq t \leq T^*$. Hence $u = \nabla p = 0$. The theorem is proved.

Proof of Theorem 1.1. 1) Let u be in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})$ with $u=0$ on S_{T^*}, and consider the ordinary differential equation

$$\frac{d}{ds}Y(s;x,t) = u(Y(s;x,t),s), \qquad Y(s;x,t)\Big|_{s=t} = x \tag{5.9}$$

for every (x,t) in Q_{T^*}, $0 \leq s \leq t$.

Since u is in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})$, there exists a unique solution curve passing through (x,t). Set

$$\xi = Y(0;x,t).$$

The mapping $(x,t) \to (\xi,t)$ is a 1-1 mapping of Q_{T^*} onto Q_{T^*} and of S_{T^*} onto S_{T^*}. The inverse of that mapping, we denote by $x = X(\xi,t)$. Set

$$v(\xi,t) = u(X(\xi,t),t).$$

Equation (5.9) implies that

$$\frac{d}{ds}X(\xi,s) = v(\xi,s), \qquad X(\xi,0) = \xi. \tag{5.10}$$

Thus,

$$x = X(\xi,t) = \xi + \int_0^t v(\xi,s)\,ds.$$

An elementary computation shows that (0.2) may be rewritten as

$$\frac{\partial}{\partial t}\tilde{\rho}(\xi,t) = 0, \qquad \tilde{\rho}(\xi,0) = \rho_0(\xi) \quad \text{on } G. \tag{5.11}$$

Hence $\tilde{\rho}(\xi,t) = \rho_0(\xi)$.

Similarly, (0.3) becomes:

$$\frac{\partial}{\partial t}\mathcal{T}(\xi,t) = 0, \qquad \mathcal{T}(\xi,0) = \chi_0(\xi,\rho_0(\xi)). \tag{5.12}$$

Thus, $\mathcal{T}(\xi,t) = \chi_0(\xi,\rho_0(\xi))$.

Finally, (0.1) may be expressed as

$$\begin{aligned} &\rho_0\frac{\partial v}{\partial t} - \nabla_v\{\chi_0\nabla_v v\} + \nabla_v q = \rho_0 f^v, \qquad \nabla_v \cdot v = 0 \quad \text{on } Q_{T^*}, \\ &v(\xi,t) = 0 \quad \text{on } S_{T^*}, \qquad v(\xi,0) = 0 \quad \text{on } G. \end{aligned} \tag{5.13}$$

2) From Theorem 5.1 we have a unique solution $\{v, \nabla q\}$ of (5.13) in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*}) \times C^{\alpha,\alpha/2}(Q_{T^*})$.

Then with

$$u(x,t) = v(X^{-1}(x,t),t), \qquad p(x,t) = q(X^{-1}(x,t),t),$$
$$\rho(x,t) = \rho_0(X^{-1}(x,t),t), \qquad \chi(x,t) = \chi_0(X^{-1}(x,t)), \rho_0(X^{-1}(x,t)),$$

we have a unique solution $\{u, \nabla p, \rho, \chi\}$ of (0.1)–(0.3). Clearly $\{u, \nabla p, \rho, \chi\}$ is in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*}) \times C^{\alpha,\alpha/2}(Q_{T^*}) \times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*}) \times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$. The theorem is proved.

Acknowledgment. The writer is indebted to the referees for their comments.

REFERENCES

[1] V. A. KAJIKOV, *Resolution of boundary-value problems for nonhomogeneous viscous fluids*, Dokl. Akad. Nauk USSR, 216 (1974), pp. 1008–1010. (In Russian.)

[2] _____, *On the theory of classical plane solutions of nonhomogeneous viscous fluids*. Vsesoiuznyi Seminar Chil. Metodam Mekh. Zhid, 5th Katsivel, Ukraine 1975 (part 2), pp. 65–76. (In Russian.)

[3] O. A. LADYŽENSKAYA, *On the Mathematical Theory of Viscous Incompressible Fluid Flows*, Gordon-Breach, New York, 1960.

[4] O. A. LADYŽENSKAYA AND V. A. SOLONNIKOV, *Unique solvability of an initial and boundary-value problem for viscous incompressible nonhomogeneous fluids*, Zap. Nauk. Sem. Leningrad Otd. Mat., 52 (1975), pp. 52–109; Soviet Math, 10 (1978), pp. 697.

[5] J. L. LIONS, *On some problems connected with Navier-Stokes equations*, in Nonlinear Evolution Equations, M. Crandall, ed., Academic Press, New York, 1978, pp. 59–84.

[6] V. A. SOLONNIKOV, *Solvability of a problem on the motion of a viscous incompressible fluid bounded by a free surface*, Math USSR Izv. 11 (1977), pp. 1323–1357.

[7] A. TANI, *On the first initial boundary-value problem of compressible viscous fluid motion*, Publ. RIMS, Kyoto Univ., 13 (1977), pp. 193–253.

[8] R. TEMAN, *Navier-Stokes Equations*, North-Holland, Amsterdam–New York, 1979.

[9] B. A. TON, *On the initial boundary-value problem for nonhomogeneous incompressible heat conducting fluids*. Rocky Mtn. J. Math., 11 (1981), pp. 99–112.

[10] _____, *Existence and uniqueness of a classical solution of an initial boundary value problem of the theory of shallow waters*, this Journal, 12 (1981), pp. 229–241.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0002 $01.00/0

TRAVELING WAVE AND MULTIPLE TRAVELING WAVE SOLUTIONS OF PARABOLIC EQUATIONS*

PATRICK S. HAGAN†

Abstract. We consider scalar equations $u_t=f(u_{xx},u_x,u)$ with $\frac{\partial}{\partial\alpha}f(\alpha,\beta,\gamma)\geq 1$. We first determine the stability of the *monotonic* traveling wave solutions $u(t,x)=\phi(x-ct,c)$. We then study the continued existence and bifurcations of these solutions as the wavespeed c varies. We use these continuation results to explore the connection between the initial conditions $u(0,x)$ and the wavespeed(s) of the resulting solution $u(t,x)$.

1. Introduction.

We consider scalar equations

$$u_t=f(u_{xx},u_x,u), \quad \text{where } \frac{\partial}{\partial\alpha}f(\alpha,\beta,\gamma)\geq 1 \quad \text{for all } \alpha,\beta,\gamma. \tag{1.1}$$

An example of such an equation is $u_t=[q(u)u_x]_x+c(u)u_x+r(u)$ with $q(u)\geq 1$, which occurs when a quantity is governed by diffusion, a transport/convection term and a reaction term.

In [1] we considered the *nonmonotonic* traveling wave solutions $u=\phi(x-ct)$ of (1.1). We found that most of these solutions are unstable to *all* nonpositive and to *all* nonnegative initial perturbations. Now we consider the *monotonic* waves.

First we will obtain sharp stability results for these waves. These results, combined with those in [1], permit the stability of any traveling wave $u=\phi(x-ct)$ to be determined by inspection of its trajectory in the phase plane of

$$\phi'=v, \qquad f(v',v,\phi)+cv=0. \tag{1.2}$$

We will classify the monotonic solutions of (1.2) as $S\to S$, $N\to S$, $S\to N$ or $N\to N$ according as whether $\phi(-\infty),0$ and $\phi(+\infty),0$ are both saddle points, a node and a saddle point, a saddle point and a node or both nodes. For each of these classes, we will suppose that a monotonic solution of that type with a speed c_0 exists. Then, as c varies from c_0 we will determine the continued existence and bifurcations of monotonic solutions of the same class.

These continuation results will show the sharpness of the stability results. They will also be used to explore the relation between $u(0,x)$ and the qualitative behavior of $u(t,x)$. Specifically, for a large class of initial conditions, we will find that for any fixed η

$$\lim_{t\to\infty} u(t,\eta+ct)=\begin{cases}\phi_0 & \text{if } c<c_1,\\ \phi_1 & \text{if } c_1<c<c_2,\\ \vdots & \vdots \quad \vdots\\ \phi_{m-1} & \text{if } c_{m-1}<c<c_m,\\ \phi_m & \text{if } c_m<c,\end{cases} \tag{1.3}$$

* Received by the editors July 9, 1979 and in final revised form August 25, 1981. This research was supported by the Office of Naval Research, the Air Force Office of Scientific Research, the Army Research Office and the National Science Foundation.

† Current address: Exxon Corporate Research Laboratory, P.O. Box 45, Linden, New Jersey 07036.

where $m \geq 1$. Here $\phi_0, \cdots, \phi_m$ are distinct constant solutions of (1.1) and $c_1 < c_2 < \cdots < c_m$. We will determine m and the wavespeeds c_j from $u(0,x)$. In particular, if $m > 1$ we will call $u(t,x)$ a multiple, or stacked, traveling wave.

Finally, the existence and stability of any particular type of traveling wave solution will be shown to be a generic property of (1.1).

2. Mathematical preliminaries. We assume that f satisfies:

H1. For some $\Delta > 0$ and for all a, b, c, the derivatives $f_1(a,b,c)$, $f_2(a,b,c)$ and $f_3(a,b,c)$ exist and are Hölder continuous with exponent Δ in all arguments.

H2. $f_1(a,b,c) \geq 1$ for all a, b, c.

H3. At every u where $f(0,0,u) = 0$, $f_3(0,0,u) \neq 0$.

The last assumption guarantees that every singular point in the phase plane of (1.2) is an ordinary saddle point, node, spiral point or center. In the Appendix we will extend the stability results to the case of $f(0,0,u) \equiv 0$ for all u. This case often occurs in shock models [2].

To simplify later exposition, we now define B_x^2 as the set of all functions $\psi(t,x)$ which have ψ, ψ_t, ψ_x and ψ_{xx} existing, continuous and bounded for all t in $[0, T]$ and all x for every $T > 0$. For convenience, we also introduce the following theorem at this point.

THE MAXIMUM PRINCIPLE. *Assume that* H1 *and* H2 *are satisfied. Let* $u(t,x)$ *and* $v(t,x)$ *be any functions in* B_x^2 *which satisfy*

$$u_t - f(u_{xx}, u_x, u) \geq v_t - f(v_{xx}, v_x, v) \quad \text{for all } x, \quad \text{for all } t \geq 0. \tag{2.1}$$

1) *If* $u(0,x) \geq v(0,x)$ *for all* x, *then* $u(t,x) \geq v(t,x)$ *for all* x *and for all* $t \geq 0$.

2) *If* $c(t)$ *is continuously differentiable, if* $u(0,x) \geq v(0,x)$ *for all* $x \geq \bar{x} + c(0)$ *and if* $u(t, \bar{x} + c(t)) \geq v(t, \bar{x} + c(t))$ *for all* $t \geq 0$, *then* $u(t,x) \geq v(t,x)$ *for all* $x \geq \bar{x} + c(t)$, *all* $t \geq 0$.

The proof of this principle is standard (see, e.g., [3], [4]) and will be omitted.

The above theorem motivates the definitions of $\bar{u}(t,x)$ and $\underline{u}(t,x)$ as an *upper function* and a *lower function* of (1.1) if and only if $\bar{u}$ and $\underline{u}$ are in B_x^2 and satisfy

$$\bar{u}_t - f(\bar{u}_{xx}, \bar{u}_x, \bar{u}) \geq 0 \quad \text{for all } x, \quad \text{for all } t \geq 0, \tag{2.2}$$

$$\underline{u}_t - f(\underline{u}_{xx}, \underline{u}_x, \underline{u}) \leq 0 \quad \text{for all } x, \quad \text{for all } t \geq 0. \tag{2.3}$$

We will obtain most of our results by first constructing suitable upper and lower functions $\bar{u}(t,x)$ and $\underline{u}(t,x)$. Then we will use the maximum principle to conclude that $\bar{u}(t,x) \geq u(t,x) \geq \underline{u}(t,x)$ for all solutions $u(t,x)$ satisfying $\bar{u}(0,x) \geq u(0,x) \geq \underline{u}(0,x)$. This procedure has been widely used [5]–[10] for equations of the form $u_t = u_{xx} + h(u)$.

3. Some preliminary results. Here we introduce some well-known results about monotonic traveling wave solutions of (1.1) and then use these results to establish useful upper and lower functions.

Suppose that $u = \phi(x - ct)$ is a solution of (1.1). Then $\phi = \phi(x)$, $v = \phi'(x)$ satisfies (1.2). Consider the phase plane representation of the solutions of (1.2) at any fixed value of c. The singular points in this phase plane are the points $\phi_0, 0$ with $f(0,0,\phi_0) = 0$. We note that:

1) if $f_3(0,0,\phi_0) < 0$, then $\phi_0, 0$ is a saddle point at each c;

2) if $f_3(0,0,\phi_0) > 0$, then $\phi_0, 0$ is a node, spiral point or center depending on the value of c.

Now suppose that ϕ is monotonic and bounded. Let $\phi_- \equiv \phi(-\infty)$ and $\phi_+ \equiv \phi(+\infty)$. Then $\phi_-, 0$ is either a saddle point or a node, as is $\phi_+, 0$. Following [11, Chapt. 13], we define

$$k_{1,2}^-(c) \equiv \frac{-f_2^- - c \pm \sqrt{(f_2^- + c)^2 - 4f_1^- f_3^-}}{2f_1^-} \quad \text{with } k_1^-(c) \le k_2^-(c), \tag{3.1}$$

where $f_i^- \equiv f_i(0,0,\phi_-)$, $i = 1,2,3$. Then, as $x \to -\infty$ the asymptotic behavior of $\phi(x)$ is given by:

1) if $\phi_-, 0$ is a saddle point,

$$\phi(x) \sim \phi_- + a \exp[k_2^-(c)x] \quad \text{as } x \to -\infty; \tag{3.2}$$

2) if $\phi_-, 0$ is a node, then typically $\phi(x)$ decays at the *usual* rate

$$\phi(x) \sim \begin{cases} \phi_- + a \exp[k_1^-(c)x] & \text{as } x \to -\infty \quad \text{if } k_1^-(c) \ne k_2^-(c), \\ \phi_- + ax \exp[k_1^-(c)x] & \text{as } x \to -\infty \quad \text{if } k_1^-(c) = k_2^-(c), \end{cases} \tag{3.3}$$

but it may decay at the *accidental* rate

$$\phi(x) \sim \phi_- + a \exp[k_2^-(c)x] \quad \text{as } x \to -\infty. \tag{3.4}$$

In (3.2)–(3.4), a is some nonzero constant.

Similarly, we define

$$k_{1,2}^+(c) \equiv \frac{-f_2^+ - c \pm \sqrt{(f_2^+ + c)^2 - 4f_1^+ f_3^+}}{2f_1^+} \quad \text{with } k_1^+(c) \ge k_2^+(c), \tag{3.5}$$

where $f_i^+ \equiv f_i(0,0,\phi_+)$, $i = 1,2,3$. Then, as $x \to +\infty$ the asymptotic behavior of $\phi(x)$ is given by (3.2), (3.3) or (3.4) with ϕ_-, $k_1^-(c)$ and $k_2^-(c)$ replaced by ϕ_+, $k_1^+(c)$ and $k_2^+(c)$. Moreover, the asymptotic formulas for $\phi'(x)$ and $\phi''(x)$ as $x \to \pm\infty$ are obtained by differentiating the formulas for $\phi(x)$.

Finally, suppose that ϕ is nonconstant and monotonic. Then clearly, $\phi'(x) \ge 0$ for all x or $\phi'(x) \le 0$ for all x. Also, from the uniqueness of solutions of (1.2), it is easily shown that $\phi'(x) \ne 0$ for all x. Moreover, the asymptotic formulas for $\phi'(x)$ and $\phi''(x)$ show that $\phi''(x)/\phi'(x)$ goes to either $k_1^-(c)$ or $k_2^-(c)$ as $x \to -\infty$ and goes to $k_1^+(c)$ or $k_2^+(c)$ as $x \to +\infty$. Thus $|\phi''(x)/\phi'(x)|$ is bounded. Similarly, for any x_0, $|\phi(x) - \phi_-|/\phi'(x)$ is bounded for all $x \le x_0$ and $|\phi(x) - \phi_+|/\phi'(x)$ is bounded for all $x \ge x_0$.

We now establish the following lemmas with these facts.

LEMMA 1. *Assume that* H1, H2 *and* H3 *are satisfied. Suppose that* $u = \phi(x - ct)$ *is a bounded increasing in* x *solution of* (1.1). *Let* h_0 *be any constant and let* $\varepsilon_{1,2}^{\pm}$ *be any four positive constants small enough so that*

$$f(0,0,\phi_- + z) \ne 0 \quad \text{for all } z \ne 0 \text{ in } [-\varepsilon_1^-, \varepsilon_2^-], \tag{3.6}$$

$$f(0,0,\phi_+ + z) \ne 0 \quad \text{for all } z \ne 0 \text{ in } [-\varepsilon_1^+, \varepsilon_2^+]. \tag{3.7}$$

Then $\underline{u}$ *is a lower function and* $\bar{u}$ *is an upper function of* (1.1), *where*:

1) *if* $\phi_-, 0$ *and* $\phi_+, 0$ *are both nodes*,

$$\underline{u}(h_0,t,x) \equiv \phi(x - ct - h_0), \qquad \bar{u}(h_0,t,x) \equiv \phi(x - ct + h_0); \tag{3.8}$$

2) *if* $\phi_-,0$ *is a node and* $\phi_+,0$ *is a saddle,*

(3.9)

$$\underline{u}(\varepsilon_1^+,h_0,t,x)\equiv\phi(x-ct-h(\varepsilon_1^+,t))-q(\varepsilon_1^+,t)\{\phi(x-ct-h(\varepsilon_1^+,t))-\phi_-\}/(\phi_+-\phi_-),$$
$$\bar{u}(\varepsilon_2^+,h_0,t,x)\equiv\phi(x-ct+h(\varepsilon_2^+,t))+q(\varepsilon_2^+,t)\{\phi(x-ct+h(\varepsilon_2^+,t))-\phi_-\}/(\phi_+-\phi_-),$$

3) *if* $\phi_-,0$ *is a saddle and* $\phi_+,0$ *is a node,*

(3.10)

$$\underline{u}(\varepsilon_1^-,h_0,t,x)\equiv\phi(x-ct-h(\varepsilon_1^-,t))-q(\varepsilon_1^-,t)\{\phi_+-\phi(x-ct-h(\varepsilon_1^-,t))\}/(\phi_+-\phi_-),$$
$$\bar{u}(\varepsilon_2^-,h_0,t,x)\equiv\phi(x-ct+h(\varepsilon_2^-,t))+q(\varepsilon_2^-,t)\{\phi_+-\phi(x-ct+h(\varepsilon_2^-,t))\}/(\phi_+-\phi_-);$$

4) *if* $\phi_-,0$ *and* $\phi_+,0$ *are both saddles,*

(3.11)

$$\underline{u}(\varepsilon_1^-,\varepsilon_1^+,h_0,t,x)\equiv\phi(x-ct-h(\varepsilon_1,t))-q(\varepsilon_1,t)\left\{1+\frac{\varepsilon_1^+-\varepsilon_1^-}{\varepsilon_1^++\varepsilon_1^-}\tanh\kappa[x-ct-h(\varepsilon_1,t)]\right\},$$

$$\bar{u}(\varepsilon_2^-,\varepsilon_2^+,h_0,t,x)\equiv\phi(x-ct+h(\varepsilon_2,t))+q(\varepsilon_2,t)\left\{1+\frac{\varepsilon_2^+-\varepsilon_2^-}{\varepsilon_2^++\varepsilon_2^-}\tanh\kappa[x-ct+h(\varepsilon_2,t)]\right\}.$$

In the above, $h(\varepsilon,t)$ *and* $q(\varepsilon,t)$ *are*

$$h(\varepsilon,t)\equiv\varepsilon K(\varepsilon)[1-e^{-ts(\varepsilon)}]+h_0,\qquad q(\varepsilon,t)\equiv\varepsilon e^{-ts(\varepsilon)},\tag{3.12}$$

where $K(\varepsilon)$ *and* $s(\varepsilon)$ *are some positive constants that are bounded as* $\varepsilon\to 0$. *Also, in* (3.11) $\varepsilon_1\equiv\frac{1}{2}(\varepsilon_1^-+\varepsilon_1^+)$, $\varepsilon_2\equiv\frac{1}{2}(\varepsilon_2^-+\varepsilon_2^+)$ *and* $\kappa>0$ *is any sufficiently small constant.*

These upper and lower functions at $t=0$ are sketched in Figs. 1–4. As $t\to\infty$, $q(t)\to 0$ and $h(t)\to$ constant. So, the upper and lower functions in Figs. 2–4 eventually look like the ones in Fig. 1. We note that the functions in (3.11) were constructed previously in [5] for equations of the form $u_t=u_{xx}+h(u)$.

Proof. We prove only that $\bar{u}(\varepsilon_2^+,h_0,t,x)$ in (3.9) is an upper function. Very similar calculations show that the other functions in parts 2), 3), and 4) are upper and lower functions. The functions in part 1) are solutions of (1.1), and so they are also upper and lower functions.

We need to show that $\bar{u}(\varepsilon_2^+,h_0,t,x)$ satisfies (2.2). For simplicity, we define $\bar{q}(\varepsilon_2^+,t)\equiv q(\varepsilon_2^+,t)/(\phi_+-\phi_-)$. We have

$$\begin{aligned}&\bar{u}=\phi+\bar{q}(\phi-\phi_-),\qquad \bar{u}_t=(h_t-c)(1+\bar{q})\phi'+\bar{q}_t(\phi-\phi_-),\\ &\bar{u}_x=(1+\bar{q})\phi',\qquad \bar{u}_{xx}=(1+\bar{q})\phi''.\end{aligned}\tag{3.13}$$

We define

(3.14)

$$C_i=\sup_{\substack{0\le\alpha,\beta\le\bar{q}(\varepsilon_2^+,0)\\ \text{all } x}}|f_i([1+\alpha]\phi'',[1+\alpha]\phi',\phi_-+[1+\beta][\phi-\phi_-])|,\qquad i=1,2,3.$$

We can also define the positive constant $s=\frac{1}{2}\min_{-\varepsilon_1^+\le z\le\varepsilon_2^+}\{-f(0,0,\phi_++z)/z\}$ since $f_3(0,0,\phi_+)<0$, $f(0,0,\phi_+)=0$ and $f(0,0,\phi_++z)\neq 0$ for all $z\neq 0$, $-\varepsilon_1^+\le z\le\varepsilon_2^+$. Finally, since $\phi''(x)/\phi'(x)$ is bounded, we define $m=\sup|\phi''(x)/\phi'(x)|$.

Using (3.13), we find that

$$\begin{aligned}\bar{u}_t-f(\bar{u}_{xx},\bar{u}_x,\bar{u})\ge{}&h_t\phi'+\bar{q}_t(\phi-\phi_-)-c\bar{q}\phi'+f(\phi'',\phi',\phi)\\ &-f([1+\bar{q}]\phi'',[1+\bar{q}]\phi',\phi+\bar{q}[\phi-\phi_-]).\end{aligned}\tag{3.15}$$

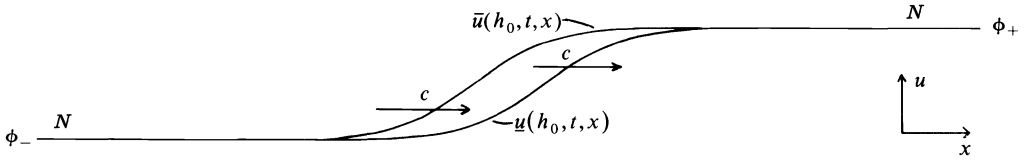

FIG. 1. *The upper and lower functions when* $\phi_-,0$ *and* $\phi_+,0$ *are both nodes* (N).

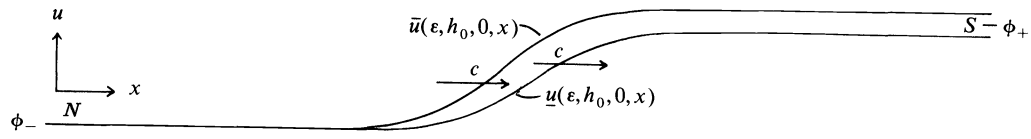

FIG. 2. *The upper and lower functions* $\bar{u}$ *and* $\underline{u}$ *at* $t=0$ *when* $\phi_-,0$ *is a node* (N) *and* $\phi_+,0$ *is a saddle point* (S).

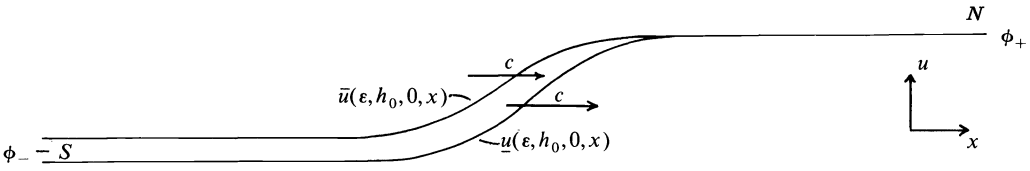

FIG. 3. *The upper and lower functions* $\bar{u}$ *and* $\underline{u}$ *at* $t=0$ *when* $\phi_-,0$ *is a saddle point* (S) *and* $\phi_+,0$ *is a node* (N).

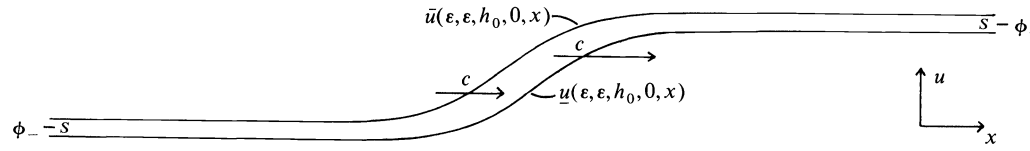

FIG. 4. *The upper and lower functions* $\bar{u}$ *and* $\underline{u}$ *at* $t=0$ *when* $\phi_-,0$ *and* $\phi_+,0$ *are both saddle points.*

For $x-ct-h\geq x_0$ and x_0 large enough, this gives

$$\begin{aligned}\bar{u}_t-f(\bar{u}_{xx},\bar{u}_x,\bar{u})\geq{}&h_t\phi'+\bar{q}_t(\phi-\phi_-)-|C_1\bar{q}\phi''|\\&-(C_2+c)\bar{q}\phi'+s\bar{q}(\phi-\phi_-).\end{aligned}\tag{3.16}$$

Since $\bar{q}_t\equiv -s\bar{q}$, we note that (2.2) is satisfied for all $x-ct-h\geq x_0$ if $h_t\geq(C_1m+C_2+c)\bar{q}$. Similarly, for $x-ct-h\leq x_0$, (3.15) yields

$$\bar{u}_t-f(\bar{u}_{xx},\bar{u}_x,\bar{u})\geq h_t\phi'-s\bar{q}(\phi-\phi_-)-\left[C_1|\phi''|+(C_2+c)\phi'+C_3(\phi-\phi_-)\right]\bar{q}.\tag{3.17}$$

Since $(\phi(x)-\phi_-)/\phi'(x)$ is bounded for all $x\leq x_0$, we can let

$$n\equiv\sup_{x\leq x_0}|\phi'(x)-\phi_-|/\phi'(x).$$

Thus, if $h_t\geq[C_1m+C_2+c+(C_3+s)n]\bar{q}$, then (2.2) is satisfied for all $x-ct-h\leq x_0$. Therefore, we select $K=[C_1m+C_2+c+(C_3+s)n]/s$ in (3.12). Then $\bar{u}$ satisfies (2.2) for all x, all $t\geq 0$ and is an upper function. □

The upper and lower functions in Lemma 1 will lead to sharp stability results for $\phi(x-ct)$ unless $\phi(x)$ decays at the accidental rate to a node as $x\to-\infty$ or as $x\to+\infty$. For these exceptional solutions, we now obtain sharper upper and lower functions.

LEMMA 2. *Assume that* H1, H2 *and* H3 *are satisfied. Suppose that* $u=\phi(x-ct)$ *is a bounded, increasing in x solution of* (1.1). *Let* $\phi_-\equiv\phi(-\infty)$, *let* $\phi_+\equiv\phi(+\infty)$ *and let the four constants* $k_{1,2}^{\pm}(c)$ *be defined by* (3.1) *and* (3.5). *Finally, define* $\psi_S(x)$ *and* $\psi_N(x)$ *to be any functions in* C^3 *satisfying*

$$
\begin{aligned}
&\psi_S(x)\equiv e^x, \quad \psi_N(x)\equiv e^x, \quad x\le 0,\\
&1\le\psi_S(x)\le 2, \quad 0\le\psi_N(x)\le 2, \quad 0\le x\le 1,\\
&0\le\psi_S'(x)\le 2, \quad 0\le|\psi_N'(x)|\le 2, \quad 0\le x\le 1,\\
&0\le|\psi_S''(x)|\le 4, \quad 0\le|\psi_N''(x)|\le 4, \quad 0\le x\le 1,\\
&\psi_S(x)\equiv 2, \quad \psi_N(x)\equiv 0, \quad 1\le x.
\end{aligned}
\tag{3.18}
$$

For any sufficiently small $\varepsilon>0$ *and* $\delta>0$, *the following* $\bar{u}$ *and* $\underline{u}$ *are upper and lower functions of* (1.1):

1) *if* $\phi_-,0$ *is a node, if* $k_1^-(c)\neq k_2^-(c)$ *and if* $\phi(x)$ *decays to* ϕ_- *at the accidental rate as* $x\to-\infty$, *then*

1a) *if* $\phi_+,0$ *is a saddle point,*

$$
\bar{u}(\varepsilon,\delta,h_0,t,x)\equiv\phi(x-ct+h(t))+\eta(t)\psi_S\big([k_1^-(c)+\varepsilon][x-ct+h(t)]\big),
\tag{3.19}
$$

1b) *if* $\phi_+,0$ *is a node,*

$$
\bar{u}(\varepsilon,\delta,h_0,t,x)\equiv\phi(x-ct+h(t))+\eta(t)\psi_N\big([k_1^-(c)+\varepsilon][x-ct+h(t)]\big);
\tag{3.20}
$$

2) *if* $\phi_+,0$ *is a node, if* $k_1^+(c)\neq k_2^+(c)$ *and if* $\phi(x)$ *decays to* ϕ_+ *at the accidental rate as* $x\to+\infty$, *then*

2a) *if* $\phi_-,0$ *is a saddle point,*

$$
\underline{u}(\varepsilon,\delta,h_0,t,x)\equiv\phi(x-ct-h(t))-\eta(t)\psi_S\big([k_1^+(c)-\varepsilon][x-ct-h(t)]\big),
\tag{3.21}
$$

2b) *if* $\phi_-,0$ *is a node,*

$$
\underline{u}(\varepsilon,\delta,h_0,t,x)\equiv\phi(x-ct-h(t))-\eta(t)\psi_N\big([k_1^+(c)-\varepsilon][x-ct-h(t)]\big).
\tag{3.22}
$$

In the above equations

$$
\eta(t)\equiv\delta\varepsilon^4\varepsilon^{2/\Delta}e^{-\varepsilon^3 t}, \qquad h(t)\equiv h_0+K\delta\varepsilon^{2/\Delta}\big(1-e^{-\varepsilon^3 t}\big),
\tag{3.23}
$$

where K *is a positive constant,* h_0 *is arbitrary and* Δ *is the Hölder exponent in* H1.

Proof. Lemma 2 follows from calculations like the ones which established Lemma 1. □

These new upper and lower functions are very similar to the ones in Lemma 1 for the corresponding cases, and they look like the functions sketched in Figs. 1–3. The key difference is that when $\phi(x)$ decays to a node at the accidental rate, the upper and lower functions in Lemma 1 also decay to the node at the accidental rate. However, the ones in Lemma 2 decay to the node at an exponential rate that is only ε larger than the usual decay rate.

4. Stability. We first show that a constant solution $u\equiv\phi_0$ is stable if $\phi_0,0$ is a saddle point and is unstable otherwise.

THEOREM 3 (stability of constant solutions). *Assume that* H1, H2 *and* H3 *are satisfied, and let* $u\equiv\phi_0$ *be a constant solution of* (1.1). *Then there is an* $\varepsilon_0>0$ *and* $k>0$

such that:

1) *Suppose that* $\phi_0, 0$ *is a saddle point. If* $u(t,x)$ *is any solution of* (1.1) *in* B_x^2 *that satisfies*

$$-\varepsilon e^{-kt} \le u(t,x) - \phi_0 \le \varepsilon e^{-kt} \quad \text{for all } x \tag{4.1}$$

at $t=0$ *for some* $0 \le \varepsilon \le \varepsilon_0$, *then* $u(t,x)$ *also satisfies* (4.1) *for all* $t \ge 0$.

2) *Suppose that* $\phi_0, 0$ *is a node, spiral point or center. If* $u(t,x)$ *is any solution of* (1.1) *in* B_x^2 *that satisfies either*

$$\phi_0 + \delta \operatorname{sech} kx \le u(0,x) \quad \textit{for all } x \quad \textit{or} \tag{4.2}$$

$$\phi_0 - \delta \operatorname{sech} kx \ge u(0,x) \quad \textit{for all } x \tag{4.3}$$

for any $\delta > 0$, *then* $u(t,x)$ *also satisfies*

$$\varepsilon_0 \le |u(t,x) - \phi_0| \quad \textit{for some } x \textit{ and some } t \ge 0. \tag{4.4}$$

Proof. If $\phi_0, 0$ is a saddle point, then $f_3(0,0,\phi_0) < 0$. Let $k \equiv -\frac{1}{2} f_3(0,0,\phi_0)$. Then there is an $\varepsilon_0 > 0$ such that $f_3(0,0,\phi_0 + h) < -k$ for all $|h| < \varepsilon_0$. Clearly, $\hat{u}(h,t) \equiv \phi_0 + he^{-kt}$ is an upper function for all $0 \le h \le \varepsilon_0$ and is a lower function for all $-\varepsilon_0 \le h \le 0$. So, suppose that $u(t,x)$ is any solution of (1.1) in B_x^2 which satisfies $|u(0,x) - \phi(x)| < \varepsilon$ for all x and some $0 < \varepsilon \le \varepsilon_0$. Then

$$\hat{u}(-\varepsilon, t) \le u(t,x) \le \hat{u}(\varepsilon, t) \quad \text{for all } x \tag{4.5}$$

is true at $t=0$. So, the maximum principle implies that (4.5) remains true for all $t \ge 0$. This establishes (4.1) and stability.

Now suppose that $\phi_0, 0$ is a node, spiral point or center. Define $\underline{u}(h,t,x) \equiv \phi_0 + he^{\mu t} \operatorname{sech} kx$, where $\mu \equiv \frac{1}{2} f_3(0,0,\phi_0) > 0$ and where $k > 0$ will be selected later. We find that

$$\begin{aligned} \underline{u}_t - f(\underline{u}_{xx}, \underline{u}_x, \underline{u}) = he^{\mu t} (\operatorname{sech} kx) \{ -\tfrac{1}{2} f_3 + f_2 k \tanh kx + f_1 k^2 (1 - 2\tanh^2 kx) \} \\ + O\big([he^{\mu t} \operatorname{sech} kx]^{1+\Delta} \big). \end{aligned} \tag{4.6}$$

Here $f_i \equiv f_i(0,0,\phi_0)$, $i = 1,2,3$, and Δ is the Hölder exponent in H1. Let $k > 0$ be chosen so small that the quantity in braces is smaller than $-\frac{1}{4} f_3$ for all x. Then there is an $\varepsilon_0 > 0$ such that $\underline{u}_t - f(\underline{u}_{xx}, \underline{u}_x, \underline{u}) \le 0$ for all x whenever $0 < he^{\mu t} \le \varepsilon_0$.

So, let $u(t,x)$ be any solution of (1.1) in B_x^2 which satisfies $\underline{u}(\delta, 0, x) \le u(0,x)$ for all x. The maximum principle implies that $\underline{u}(\delta, t, x) \le u(t,x)$ for all x and all $t > 0$ with $0 < \delta e^{\mu t} \le \varepsilon_0$. This establishes (4.4) when $u(0,x)$ satisfies (4.2). The case of $u(0,x)$ satisfying (4.3) is treated similarly. □

We now determine the stability of increasing in x solutions $\phi(x - ct)$. Decreasing solutions can be treated by first applying the transformation $u \to -u$ to (1.1) and to $u = \phi(x - ct)$.

THEOREM 4 (the stability of monotonic waves). *Assume that* H1, H2 *and* H3 *are satisfied. Suppose that* $u = \phi(x - ct)$ *is a bounded increasing solution of* (1.1). *Let* $\phi_- \equiv \phi(-\infty)$, *let* $\phi_+ \equiv \phi(+\infty)$ *and let the four constants* $k_{1,2}^{\pm}(c)$ *be defined by* (3.1) *and* (3.5). *For every* $\varepsilon > 0$, *there is a* $\delta(\varepsilon) > 0$, *such that if* $u(t,x)$ *is any solution of* (1.1) *in* B_x^2 *that satisfies*

$$\begin{aligned} -\delta(\varepsilon) w_1^-(x) \le u(0,x) - \phi(x) \le +\delta(\varepsilon) w_2^-(x) \quad \text{for all } x \le 0, \\ -\delta(\varepsilon) w_1^+(x) \le u(0,x) - \phi(x) \le +\delta(\varepsilon) w_2^+(x) \quad \text{for all } x \ge 0, \end{aligned} \tag{4.7}$$

then for all $t \geq 0$, $u(t,x)$ *satisfies*

$$\begin{aligned} &-\varepsilon w_1^-(x-ct) \leq u(t,x)-\phi(x-ct) \leq +\varepsilon w_2^-(x-ct) \quad \text{for all } x \leq ct, \\ &-\varepsilon w_1^+(x-ct) \leq u(t,x)-\phi(x-ct) \leq +\varepsilon w_2^+(x-ct) \quad \text{for all } x \geq ct. \end{aligned} \tag{4.8}$$

Here $w_1^-(x)$ *and* $w_2^-(x)$ *are defined by*

1) *if* $\phi_-, 0$ *is a saddle point then* $w_1^-(x) \equiv w_2^-(x) \equiv 1$;

2) *if* $\phi_-, 0$ *is a node and* $\phi(x)$ *decays at the usual rate as* $x \to -\infty$, *then* $w_1^-(x) \equiv w_2^-(x) \equiv e^{k_1^-(c)x}$ *if* $k_1^-(c) \neq k_2^-(c)$ *and* $w_1^-(x) \equiv w_2^-(x) \equiv (1-x)e^{k_1^-(c)x}$ *if* $k_1^-(c) = k_2^-(c)$;

3) *if* $\phi_-, 0$ *is a node and* $\phi(x)$ *decays at the accidental rate as* $x \to -\infty$, *then* $w_1^-(x) \equiv e^{k_2^-(c)x}$ *and* $w_2^-(x) \equiv e^{k_1^-(c)x+\rho x}$ *if* $k_1^-(c) \neq k_2^-(c)$ *and* $w_1^-(x) \equiv w_2^-(x) \equiv e^{k_1^-(c)x}$ *if* $k_1^-(c) = k_2^-(c)$. *Here* ρ *can be any sufficiently small positive constant.*

Similarly, $w_1^+(x)$ *and* $w_2^+(x)$ *are defined by*

1) *if* $\phi_+, 0$ *is a saddle point, then* $w_1^+(x) \equiv w_2^+(x) \equiv 1$;

2) *if* $\phi_+, 0$ *is a node and* $\phi(x)$ *decays at the usual rate as* $x \to +\infty$, *then* $w_1^+(x) \equiv w_2^+(x) \equiv e^{k_1^+(c)x}$ *if* $k_1^+(c) \neq k_2^+(c)$ *and* $w_1^+(x) \equiv w_2^+(x) \equiv (1+x)e^{k_1^+(c)x}$ *if* $k_1^+(c) = k_2^+(c)$;

3) *if* $\phi_+, 0$ *is a node and* $\phi(x)$ *decays at the accidental rate as* $x \to +\infty$, *then* $w_1^+(x) \equiv e^{k_1^+(c)x-\rho x}$ *and* $w_2^+(x) \equiv e^{k_2^+(c)x}$ *if* $k_1^+(c) \neq k_2^+(c)$ *and* $w_1^+(x) \equiv w_2^+(x) \equiv e^{k_1^+(c)x}$ *if* $k_1^+(c) = k_2^+(c)$. *Here* ρ *can be any sufficiently small positive constant.*

Crudely speaking, this theorem shows that bounded monotonic solutions $u = \phi(x-ct)$ are stable to all perturbed initial conditions $u(0,x)$ which are near to $\phi(x)$ for all x and also

1) decay asymptotically to ϕ_- at the usual rate as $x \to -\infty$ if $\phi_-, 0$ is a node and $\phi(x)$ goes to ϕ_- at the usual rate;

2) decay asymptotically to ϕ_- at a rate faster than the usual rate and no faster than the accidental rate as $x \to -\infty$ if $\phi_-, 0$ is a node and $\phi(x)$ goes to ϕ_- at the accidental rate;

3) satisfy the analogous restriction if $\phi_+, 0$ is a node.

Later we will find that Theorem 4 is sharp in the following sense: Suppose that $u(0,x)$ decays to a node at an exponential rate different than the one(s) allowed by the theorem. We will find that then $u(t,x)$ would evolve either into a wave with a speed *different* than the speed c of $\phi(x-ct)$ or into two or more stacked traveling waves. In either case, $|u(t,x)-\phi(x-ct)|$ would not remain uniformly small for all t.

Also, in §8 we will find that the stability results in Theorem 4 are generic.

Proof of Theorem 4. Suppose that $\phi(x)$ does not decay at the accidental rate to a node as $x \to -\infty$ or as $x \to +\infty$. Select constants $\varepsilon > 0$ and $h_0 > 0$ and let $\bar{u}(\varepsilon, h_0, t, x)$ and $\underline{u}(\varepsilon, h_0, t, x)$ be the appropriate upper and lower functions in Lemma 1. [If part 1) of the lemma applies, take $\bar{u}(\varepsilon, h_0, t, x)$ to be $\bar{u}(h_0, t, x)$; if part 4) applies take $\bar{u}(\varepsilon, h_0, t, x)$ to be $\bar{u}(\varepsilon, \varepsilon, h_0, t, x)$. Define $\underline{u}$ similarly.] For all $\varepsilon > 0$ small enough, $\bar{u}$ and $\underline{u}$ are an upper and a lower function of (1.1). Thus, let $u(t,x)$ be any solution in B_x^2 that satisfies

$$\underline{u}(\varepsilon, h_0, t, x) \leq u(t,x) \leq \bar{u}(\varepsilon, h_0, t, x) \quad \text{for all } x \tag{4.9}$$

at $t = 0$. The maximum principle shows that (4.9) is also satisfied for all $t \geq 0$. This immediately implies a type of stability since $\varepsilon > 0$ and $h_0 > 0$ can be taken as small as we wish. Inspection of the formulas in Lemma 1 for $\bar{u}$ and $\underline{u}$ and of the asymptotic formulas (3.1)–(3.5) for $\phi(x)$ shows that (4.8) implies the stability claimed by Theorem 4 for the nonaccidental cases.

Now suppose that $\phi(x)$ decays at the accidental rate to a node as $x \to -\infty$ or as $x \to +\infty$. For these cases the theorem is proven by repeating the above arguments with

$\bar{u}$ and $\underline{u}$ replaced wherever possible by the functions $\bar{u}(\rho,\varepsilon,h_0,t,x)$ and $\underline{u}(\rho,\varepsilon,h_0,t,x)$ in Lemma 2. □

In the Appendix we extend Theorems 3 and 4 to the case of $f(0,0,u)\equiv 0$ for all u and also point out other extensions.

Many authors [5]–[10], [12], [13] have extensively investigated monotonic waves in equations of the form $u_t=u_{xx}+h(u)$. For example, in [5] *asymptotic* stability results are obtained for solutions $u=\phi(x-ct)$ of this equation when $\phi(\pm\infty)$, 0 are both saddle points. Also, in [14] *asymptotic* stability results are obtained for solutions $u=\phi(x-ct)$ of $u_t=u_{xx}+h(u,u_x)$. However, the class of initial perturbations $u(0,x)-\phi(x)$ treated in [14] is *much* more restrictive than the class allowed by Theorem 4.

In each of the next three sections, we will assume that a monotonic $S\to S$, $N\to S$ or $N\to N$ wave exists. Then we will explore the implications of its existence.

5. $S\to S$ waves. We first show that $S\to S$ waves are unique, modulo translations in x.

THEOREM 5 (existence for $S\to S$ waves). *Assume that* H1, H2 *and* H3 *are satisfied. Suppose that $u=\phi(x-c_0t)$ is a bounded, increasing in x, solution of* (1.1). *Let $\phi_-\equiv\phi(-\infty)$, let $\phi_+\equiv\phi(+\infty)$ and assume that $\phi_-,0$ and $\phi_+,0$ are both saddle points of* (1.2). *If $u\equiv\bar{\phi}(x-\bar{c}t)$ is a monotonic solution of* (1.1) *with $\bar{\phi}(-\infty)=\phi_-$ and $\bar{\phi}(+\infty)=\phi_+$, then there is a constant h such that $\bar{\phi}(x-\bar{c}t)\equiv\phi(x-c_0t-h)$ for all x and all t. In particular, $\bar{c}\equiv c_0$.*

Proof. At any c, all solutions $u=\phi(x-ct)$ of (1.1) with $\phi(+\infty)=\phi_+$ are represented by a single trajectory in the phase plane of (1.2). Thus, at any c there is at most one monotonic solution $u=\phi(x-ct)$ with $\phi(+\infty)=\phi_+$ and $\phi(-\infty)=\phi_-$, modulo translations in x. By using phase planes, one can also show that there is at most one speed c_0 at which a monotonic solution $u=\phi(x-c_0t)$ of (1.1) with $\phi(+\infty)=\phi_+$ and $\phi(-\infty)=\phi_-$ exists. We shall not show this here because it is an immediate corollary of the next theorem. □

We now show that the existence of a monotonic $S\to S$ wave $\phi(x-c_0t)$ implies that for a large class of initial conditions $u(0,x)$, the solutions $u(t,x)$ must travel with speed c_0. First we introduce the definition that $u(t,x)$ travels with speed c_0 if and only if

$$\lim_{\substack{t\to\infty\\ \eta \text{ fixed}}} u(t,\eta+ct)=\begin{cases}\phi_0 & \text{for all } \eta \quad \text{if } c<c_0,\\ \phi_1 & \text{for all } \eta \quad \text{if } c>c_0\end{cases} \tag{5.1}$$

where ϕ_0 and ϕ_1 are distinct constant solutions of (1.1).

THEOREM 6 (wavespeed for $S\to S$ waves). *Assume that* H1, H2 *and* H3 *are satisfied. Suppose that $u=\phi(x-c_0t)$ is a bounded, increasing in x, solution of* (1.1). *Let $\phi_-\equiv\phi(-\infty)$, let $\phi_+\equiv\phi(+\infty)$ and assume that $\phi_-,0$ and $\phi_+,0$ are saddle points in the phase plane of* (1.2). *Let $\varepsilon_{1,2}^{\pm}$ be any four positive constants small enough to satisfy* (3.6) *and* (3.7). *If $u(t,x)$ is any solution of* (1.1) *in B_x^2 satisfying*

$$\begin{aligned}
&\phi_--\varepsilon_1^-\le u(0,x)\le\phi_-+\varepsilon_2^- \quad \text{for all } x\le -x_0,\\
&\phi_+-\varepsilon_1^+\le u(0,x)\le\phi_++\varepsilon_2^+ \quad \text{for all } x\ge +x_0,\\
&\phi_--\varepsilon_1^-\le u(0,x)\le\phi_++\varepsilon_2^+ \quad \text{for all } x
\end{aligned} \tag{5.2}$$

for any x_0, then $u(t,x)$ travels with speed c_0.

Proof. Select constants $\underline{\varepsilon}_{1,2}^{\pm}$ that satisfy $\underline{\varepsilon}_1^-<\varepsilon_1^-$, $\underline{\varepsilon}_2^->\varepsilon_2^-$, $\underline{\varepsilon}_1^+>\varepsilon_1^+$, $\underline{\varepsilon}_2^+>\varepsilon_2^+$, and that are small enough for $\underline{u}(\underline{\varepsilon}_1^-,\underline{\varepsilon}_1^+,h_0,t,x)$ and $\bar{u}(\underline{\varepsilon}_2^-,\underline{\varepsilon}_2^+,h_0,t,x)$ in part 4) of Lemma 1 to

be a lower and an upper function of (1.1). Now (5.2) implies that for some h_0 sufficiently large

$$\underline{u}(\varepsilon_1^-,\varepsilon_1^+,h_0,t,x)\le u(t,x)\le\bar{u}(\varepsilon_2^-,\varepsilon_2^+,h_0,t,x) \quad \text{for all } x \tag{5.3}$$

is true at $t=0$. Thus, the maximum principle implies that (5.3) is satisfied for all $t\ge 0$. Examining (5.3), (3.11) and (3.12), we note that for any ε, $q(\varepsilon,t)\to 0$ and $h(\varepsilon,t)\to$ constant as $t\to+\infty$. Hence, (5.3) becomes

$$\phi(x-c_0t-h(\varepsilon_1,+\infty))\le u(t,x)\le\phi(x-c_0t+h(\varepsilon_2,+\infty)) \tag{5.4}$$

asymptotically as $t\to+\infty$, and (5.1) is satisfied with $\phi_0=\phi_-$ and $\phi_1=\phi_+$. □

Clearly results analogous to Theorems 5 and 6 hold for decreasing in x, $S\to S$ waves $u=\phi(x-c_0t)$.

6. $N\to S$ waves. Now we assume that $u=\phi(x-c_0t,c_0)$ is a bounded, increasing in x, solution to (1.1). We also assume that $\phi_-\equiv\phi(-\infty,c_0)$, 0 is a node and that $\phi_+\equiv\phi(+\infty,c_0)$, 0 is a saddle point in the phase plane of (1.2) at $c=c_0$.

First we note that $\phi_+,0$ is a saddle point for every c and that $\phi_-,0$ is an unstable node at every $c\le c_{\max}$, where

$$c_{\max}=-2\sqrt{f_1(0,0,\phi_-)f_3(0,0,\phi_-)}-f_2(0,0,\phi_-). \tag{6.1}$$

Let $v_+(\phi,c)$ be the phase plane trajectory of the solutions of (1.2) that go to $\phi_+,0$ as $x\to+\infty$. For $c\le c_{\max}$, let $v_-(\phi,c)$ be the trajectory of the solutions which decay to $\phi_-,0$ at the *accidental* rate as $x\to-\infty$. Select ϕ_0 as any point with $\phi_-<\phi_0<\phi_+$, such that there are no singular points $\bar{\phi},0$ with $\phi_-<\bar{\phi}\le\phi_0$.

Suppose that $\phi(x,c_0)$ decays to ϕ_- at the *usual* rate as $x\to-\infty$. Then $0<v_+(\phi_0,c)<v_-(\phi_0,c)$. Now $v_+(\phi,c)$ and $v_-(\phi,c)$ are continuous in c when $v_+(\phi,c)\ne 0$ and $v_-(\phi,c)\ne 0$. So there is a $\delta>0$ such that $0<v_+(\phi,c)$ for all $\phi_0\le\phi<\phi_+$ and $0<v_+(\phi_0,c)<v_-(\phi_0,c)$ whenever $|c-c_0|\le\delta$, $c\le c_{\max}$. Since there is no singular point $\bar{\phi},0$ with $\phi_-<\bar{\phi}\le\phi_0$, the phase plane shows that for all $|c-c_0|\le\delta$, $c\le c_{\max}$, $v_+(\phi_-,c)=0$ and $0<v_+(\phi,c)$ for all $\phi_-<\phi<\phi_+$. Thus, for each c with $|c-c_0|\le\delta$, $c\le c_{\max}$, there is an increasing (in x) solution $u=\phi(x-ct,c)$ of (1.1) which has $\phi(+\infty,c)=\phi_+$ and which decays to ϕ_- at the *usual* rate as $x\to-\infty$. Moreover, since $\phi_+,0$ is a saddle point, at each c there is at most one monotonic solution $u=\phi(x-ct,c)$ with $\phi(-\infty,c)=\phi_-$ and $\phi(+\infty,c)=\phi_+$, modulo translations in x.

Now assume that $\phi(x,c_0)$ decays to ϕ_- at the *accidental* rate as $x\to-\infty$. A somewhat similar phase plane analysis shows that there is a $\delta>0$ such that for each c with $c_0-\delta\le c<c_0$, there is a monotonic solution $u=\phi(x-ct,c)$ with $\phi(+\infty,c)=\phi_+$ and which decays to ϕ_- at the *usual* rate as $x\to-\infty$. Moreover, it shows that no monotonic solutions $u=\phi(x-ct,c)$ exist with $\phi(-\infty,c)=\phi_-$ and $\phi(+\infty,c)=\phi_+$ for any $c>c_0$.

Define (c_1,c_2) as the largest open interval containing c_0 such that for each c in (c_1,c_2) there is a monotonic solution $u=\phi(x-ct,c)$ with $\phi(+\infty,c)=\phi_+$ and which decays at the usual rate as $x\to-\infty$. If $c_1>-\infty$, then at $c=c_1$ either:

1) there is a monotonic solution $u=\phi(x-c_1t,c_1)$ with $\phi(+\infty,c_1)=\phi_+$ and which decays at the *accidental* rate to ϕ_- as $x\to-\infty$; or

2) the trajectory $v_+(\phi,c_1)$ intersects, but does not cross, the $v=0$ axis at at least one point ϕ_0 in (ϕ_-,ϕ_+).

For if neither 1) or 2) occurred, then continuity arguments like those above would show that there is a monotonic solution $u=\phi(x-c_1t,c_1)$ with $\phi(+\infty,c_1)=\phi_+$ and which decays to ϕ_- at the usual rate as $x\to-\infty$.

The first possibility cannot occur, because it would imply that there are no monotonic solutions $u=\phi(x-ct,c)$ with $\phi(-\infty,c)=\phi_-$ and $\phi(+\infty,c)=\phi_+$ for $c>c_1$, contrary to assumption. Thus, the second possibility must occur. Let $\phi_0^{(i)}$, $i=1,\cdots,m$, be all the points between ϕ_- and ϕ_+ at which $v_+(\phi,c_1)=0$ and let

$$\phi_- \equiv \phi_0^{(0)} < \phi_0^{(1)} < \cdots < \phi_0^{(m)} < \phi_0^{(m+1)} \equiv \phi_+ . \tag{6.2}$$

For each $i=1,\cdots,m$, $\phi_0^{(i)},0$ is a singular point. In fact, it must be a saddle point since the trajectory $v_+(\phi,c_1)$ both enters and leaves $\phi_0^{(i)},0$. *Thus, as c decreases to c_1, the monotonic $N\to S$ wave $u=\phi(x-ct,c)$ bifurcates into $m+1$ monotonic waves $u=\phi^{(i)}(x-c_1t,c_1)$, $i=0,1,\cdots,m$, where $\phi^{(i)}(-\infty,c_1)=\phi_0^{(i)}$ and $\phi^{(i)}(+\infty,c_1)=\phi_0^{(i+1)}$.* Moreover, $\phi^{(0)}(x-c_1t,c_1)$ decays to $\phi_-=\phi_0^{(0)}$ at the usual rate as $x\to-\infty$.

Since $u=\phi^{(0)}(x-c_1t,c_1)$ is a monotonic $N\to S$ wave, we can apply preceding results. Thus, we learn that for each c in some interval (c_1',c_2'), there is a monotonic $N\to S$ wave $u=\phi^{(0)}(x-ct,c)$ with $\phi^{(0)}(+\infty,c)=\phi_0^{(1)}$ and which decays to $\phi_-\equiv\phi_0^{(0)}$ at the usual rate as $x\to-\infty$. *So as c decreases past c_1, we can describe the bifurcation as the monotonic $N\to S$ waves $\phi(x-ct,c)$ shedding $m\geq 1$ monotonic $S\to S$ waves $\phi^{(i)}(x-c_1t,c_1)$, $i=1,\cdots,m$ at $c=c_1$, leaving the monotonic $N\to S$ waves $\phi^{(0)}(x-ct,c)$.* Moreover, if there is no intermediate saddle point $\bar\phi,0$ with $\phi_-<\bar\phi<\phi_+$, then no bifurcation can occur. So c_1 must be $-\infty$ in this case.

The $c=c_2$ case is analyzed similarly. One finds that either $c_2=c_{\max}$, which is where the unstable node $\phi_-,0$ changes into an unstable spiral point, or that there is a monotonic solution $u=\phi(x-c_2t,c_2)$ with $\phi(+\infty,c_2)=\phi_+$ and which decays to ϕ_- at the *accidental* rate as $x\to-\infty$.

We now note that no monotonic solutions $u=\phi(x-ct,c)$ with $\phi(+\infty,c)=\phi_+$ and $\phi(-\infty,c)=\phi_-$ exist for any $c\leq c_1$ nor for any $c>c_2$. This follows from $v_+(\phi,c)$ being increasing in c at each ϕ with $v_+(\bar\phi,c)>0$ for all $\phi\leq\bar\phi<\phi_+$ and from $v_-(\phi,c)$ being decreasing in c at each ϕ with $v_-(\phi,c)>0$ for all $\phi_-<\bar\phi\leq\phi$.

The following theorem summarizes these results.

THEOREM 7 (existence for $N\to S$ waves). *Assume that* H1, H2 *and* H3 *are satisfied. Suppose that $u=\phi(x-c_0t,c_0)$ is a bounded, increasing in x, solution of* (1.1). *Let $\phi_-\equiv\phi(-\infty,c_0)$, let $\phi_+\equiv\phi(+\infty,c_0)$ and assume that $\phi_-,0$ is a node and $\phi_+,0$ is a saddle point of* (1.2) *at $c=c_0$. Then there is a $c_1\geq-\infty$ and a $c_2\leq c_{\max}$ such that for each c in (c_1,c_2), there is a unique (modulo translations in x) monotonic solution $u=\phi(x-ct,c)$ of* (1.1) *with $\phi(-\infty,c)=\phi_-$ and $\phi(+\infty,c)=\phi_+$. Also, for any $c\leq c_1$ and $c>c_2$, no monotonic solution $u=\phi(x-ct,c)$ exists with $\phi(-\infty,c)=\phi_-$ and $\phi(+\infty,c)=\phi_+$. Furthermore,*

1) *$\phi(x,c)$, $\phi_x(x,c)$ are both continuously differentiable in c;*

2) *for all $c_1<c<c_2$, $\phi(x,c)\to\phi_-$ at the usual rate as $x\to-\infty$;*

3) *if $c_2<c_{\max}$, then $\phi(x,c_2)\to\phi_-$ at the accidental rate as $x\to-\infty$.*

If $c_1>-\infty$, then there are $m\geq1$ saddle points $\phi_0^{(i)},0$ with $\phi_-\equiv\phi_0^{(0)}<\phi_0^{(1)}<\cdots<\phi_0^{(m)}<\phi_0^{(m+1)}\equiv\phi_+$. As c decreases to c_1, $\phi(x,c)$ bifurcates into $m+1$ monotonic solutions $u=\phi^{(i)}(x-c_1t,c_1)$ with $\phi^{(i)}(-\infty,c_1)=\phi_0^{(i)}$ and $\phi^{(i)}(+\infty,c_1)=\phi_0^{(i+1)}$, $i=0,1,\cdots,m$. In particular, if there is no saddle point $\bar\phi,0$ with $\phi_-<\bar\phi<\phi_+$, then $c_1=-\infty$. □

We now find how the speed of $u(t,x)$ depends on $u(0,x)$.

THEOREM 8 (wavespeed for $N\to S$ waves). *Assume that* H1, H2 *and* H3 *are satisfied, and suppose that $u=\phi(x-c_0t,c_0)$ is a bounded, increasing in x, solution of* (1.1). *Let $\phi_-\equiv\phi(-\infty,c_0)$, let $\phi_+\equiv\phi(+\infty,c_0)$ and assume that $\phi_-,0$ is a node and $\phi_+,0$ is a saddle point of* (1.2) *at $c=c_0$. Define $k_1^-(c)$ by* (3.1) *and also define c_1 and c_2 as in Theorem* 7. *Finally, let $\varepsilon_1^+>0$ and $\varepsilon_2^+>0$ be any constants small enough to satisfy* (3.7).

Suppose that $u(t,x)$ *is any solution of* (1.1) *in* B_x^2 *that satisfies*

$$\begin{aligned} &\phi_- \le u(0,x) \le \phi_+ + \varepsilon_2^+ \quad \text{for all } x, \\ &\phi_+ - \varepsilon_1^+ \le u(0,x) \le \phi_+ + \varepsilon_2^+ \quad \text{for all } x \ge x_0, \end{aligned} \tag{6.3}$$

for some x_0. *Then*

1) *if, for any* c *with* $c_1 < c < c_2$, *there is an* α *and a* β *such that*

$$0 < \alpha < |u(0,x) - \phi_-| e^{-k_1^-(c)x} \quad \text{for all } x \le x_0, \tag{6.4}$$

$$|u(0,x) - \phi_-| e^{-k_1^-(c)x} < \beta \quad \text{for all } x \le x_0, \tag{6.5}$$

then $u(t,x)$ *travels with speed* c;

2) *if there is a* $\beta > 0$ *such that*

$$|u(0,x) - \phi_-| e^{-k_1^-(c_2)x} < \beta \quad \text{for all } x \le x_0, \tag{6.6}$$

then $u(t,x)$ *travels with speed* c_2.

Roughly speaking, if $u(0,x)$ satisfies (6.3), then the speed of $u(t,x)$ is governed solely by the asymptotic decay rate of $u(0,x)$ as $x \to -\infty$: the *slower* the decay rate, the *faster* $u(t,x)$ travels to the *left*. Part 2) shows that all solutions decaying faster than $\exp\{+k_1^-(c_2)x\}$ as $x \to -\infty$ (including solutions with $u(0,x) \equiv \phi_-$ for all sufficiently small x) travel with speed c_2. So, part 2) is an extension of the results obtained in [9], [10], [12], [13].

One can also obtain one-sided results. Namely, if $u(0,x)$ satisfies (6.3) and (6.4), then $u(t,x)$ travels no faster than speed c. If $u(0,x)$ satisfies (6.3) and (6.5), then $u(t,x)$ travels no slower than speed c.

Proof of Part 1). Assume that $u(t,x)$ is in B_x^2 and satisfies (6.3)–(6.5). Select constants $\bar{\varepsilon}_1 > \varepsilon_1^+$, $\bar{\varepsilon}_2 > \varepsilon_2^+$ small enough so that $f(0,0,\phi_+ + z) \neq 0$ for all $z \neq 0$ in $[-\bar{\varepsilon}_1, +\bar{\varepsilon}_2]$. Now Theorem 7 shows that there is a monotonic solution $u = \phi(x - ct, c)$ with $\phi(-\infty, c) = \phi_-$ and $\phi(+\infty, c) = \phi_+$. So part 2) of Lemma 1 yields the upper and lower functions $\bar{u}(\bar{\varepsilon}_2, h_0, t, x)$ and $\underline{u}(\bar{\varepsilon}_1, h_0, t, x)$ for the wave $\phi(x - ct, c)$. Equations (6.3)–(6.5), (3.9) and (3.12) imply that, for some h_0 sufficiently large,

$$\underline{u}(\bar{\varepsilon}_1, h_0, t, x) \le u(t,x) \le \bar{u}(\bar{\varepsilon}_2, h_0, t, x) \quad \text{for all } x \tag{6.7}$$

is true at $t = 0$. The maximum principle now implies that (6.7) holds for all $t \ge 0$. Examining the definitions of $\underline{u}$ and $\bar{u}$ in (3.9), (3.12) exactly as we did for (5.3) in the $S \to S$ case, we find that (6.7) implies that $u(t,x)$ travels with speed c.

Part 2). Assume that $u(t,x)$ is in B_x^2 and satisfies (6.3) and (6.6). For each γ in $(c_1, c_2]$, define $\underline{u}(\varepsilon, h_0, \gamma, t, x)$ and $\bar{u}(\varepsilon, h_0, \gamma, t, x)$ as the upper and lower functions $\underline{u}(\varepsilon, h_0, t, x)$ and $\bar{u}(\varepsilon, h_0, t, x)$ defined in (3.9) for the monotonic wave $u = \phi(x - \gamma t, \gamma)$. Select $\bar{\varepsilon}_1 > \varepsilon_1^+$ and $\bar{\varepsilon}_2 > \varepsilon_2^+$ as before.

First, note that for each $c_1 < \gamma < c_2$, (6.3) and (6.6) imply that, for some constant $h_0(\gamma)$,

$$\phi_- \le u(t,x) \le \bar{u}(\bar{\varepsilon}_2, h_0(\gamma), \gamma, t, x) \quad \text{for all } x \tag{6.8}$$

is true at $t = 0$. The maximum principle then implies that (6.8) is true for all $t \ge 0$. This shows that for every $c < \gamma$ and every γ in (c_1, c_2),

$$\lim_{t \to \infty} u(t, \eta + ct) = \phi_- \quad \text{for all } \eta. \tag{6.9}$$

Thus, (6.9) is valid for all $c < c_2$.

We now must show that, for every $c>c_2$,

$$\lim_{t\to\infty} u(t,\eta+ct)=\phi_+ \quad \text{for all } \eta, \tag{6.10}$$

which will complete the proof. Choose any $c>c_2$, and let δ be any sufficiently small constant with $0<\delta<c-c_2$. A continuity argument similar to the one used for Theorem 7 shows that if $\delta>0$ is small enough, then there is a *nonmonotonic* solution $u=\phi(x-(c_2+\delta)t, c_2+\delta)$ which has $\phi(+\infty, c_2+\delta)=\phi_+$, has $\phi(\underline{x}(\delta), c_2+\delta)=\phi_-$ for some $\underline{x}(\delta)$ and has $\phi_x(x, c_2+\delta)>0$ for all $x\geq x(\delta)$. Furthermore, (1.1) is autonomous so we can take $\underline{x}(\delta)=0$.

We define

(6.11)

$$\underline{u}(\bar{\varepsilon}_1, x_0, c_2+\delta, t, x)=\phi(x-(c_2+\delta)t-h(t), c_2+\delta) \\ -q(t)\{\phi(x-(c_2+\delta)t-h(t), c_2+\delta)-\phi_-\}/(\phi_+-\phi_-),$$

where $h(t)\equiv x_0+\bar{\varepsilon}_1 K(1-e^{-st})$ and $q(t)\equiv\bar{\varepsilon}_1 e^{-st}$. A calculation very similar to the one used to prove Lemma 1 shows that there is a $K>0$ and $s>0$ such that $\underline{u}_t-f(\underline{u}_{xx}, \underline{u}_x, \underline{u})\leq 0$ for all $x\geq(c_2+\delta)t+h(t)$ and all $t\geq 0$.

Since $u(0,x)$ satisfies (6.3), clearly

$$\begin{aligned} &\phi_-\leq u(0,x) \quad \text{for all } x, \\ &\underline{u}(\bar{\varepsilon}_1, x_0, c_2+\delta, 0, x)\leq u(0,x) \quad \text{for all } x\geq x_0. \end{aligned} \tag{6.12}$$

The maximum principle shows that $u(t,x)\geq\phi_-$ for all x and for all $t\geq 0$. Since $\underline{u}(\bar{\varepsilon}_1, x_0, c_2+\delta, t, x)=\phi_-$ at $x=(c_2+\delta)t+h(t)$, we can again apply the maximum principle. This yields

$$u(t,x)\geq\underline{u}(\bar{\varepsilon}_1, x_0, c_2+\delta, t, x) \quad \text{for all } x\geq(c_2+\delta)t+h(t), \quad \text{all } t\geq 0. \tag{6.13}$$

Finally, we note that $\bar{u}(t,x)\equiv\phi_++\bar{\varepsilon}_2 e^{-\mu t}$ is an upper function for sufficiently small $\mu>0$. Since (6.3) shows that

$$u(t,x)\leq\phi_++\bar{\varepsilon}_2 e^{-\mu t} \quad \text{for all } x \tag{6.14}$$

at $t=0$, the maximum principle implies that (6.14) is true for all $t\geq 0$. Since $c>c_2+\delta$, together (6.14), (6.13) and (6.11) establish (6.10). □

Suppose that $u(0,x)$ satisfies (6.3) and that $u(0,x)\sim\phi_-+ae^{kx}$ as $x\to-\infty$. If $c_1=-\infty$, then Theorem 8 yields the speed c of $u(t,x)$ for every $k>0$. However, if $c_1>-\infty$, then Theorem 8 yields the speed only for $k>k_1^-(c_1)$. One naturally wonders how u evolves if $0<k<k_1^-(c_1)$.

To answer this assume that $u=\phi(x-ct, c)$, $c_1<c\leq c_2$, are a family of monotonic $N\to S$ solutions of (1.1). Let $\phi_-\equiv\phi(-\infty, c)$, let $\phi_+\equiv\phi(+\infty, c)$ and assume that $c_1>-\infty$. From Theorem 8 we know that $\phi(x,c)$ sheds at least one monotonic $S\to S$ wave at $c=c_1$. We consider only the typical case: we assume that at $c=c_1$, $u=\phi(x-ct, c)$ sheds only one monotonic $S\to S$ wave $u=\phi_2(x-c_1t, c_1)$, leaving the family of monotonic $N\to S$ waves $u=\phi_1(x-ct, c)$, $c_1'<c\leq c_2'$, where $c_1'<c_1<c_2'$. Here $\phi_1(-\infty, c)=\phi_-$, $\phi_1(+\infty, c)=\phi_2(-\infty, c_1)=\phi_0$ and $\phi_2(+\infty, c_1)=\phi_+$, where $\phi_0, 0$ is some saddle point with $\phi_-<\phi_0<\phi_+$.

Now let $\varepsilon_1^+>0$ and $\varepsilon_2^+>0$ be any constants small enough to satisfy (3.7). Suppose that $u(t,x)$ is any solution of (1.1) in B_x^2 that satisfies (6.3) for these ε_1^+, ε_2^+. Also suppose that for some $\alpha>0$, $\beta>0$,

$$\alpha\leq\left[u(0,x)-\phi_-\right]e^{-k_1^-(c_0)x}\leq\beta \quad \text{for all } x\leq x_0 \tag{6.15}$$

for some c_0 with $c_1' < c_0 < c_1$. We will show that $u(t,x)$ evolves into two stacked traveling waves: an upper wave with speed c_1 and a lower wave with speed c_0.

Select $\bar{\varepsilon}_1 > \varepsilon_1^+$ and $\bar{\varepsilon}_2 > \varepsilon_2^+$ small enough so that $f(0,0,\phi_+ + z) \neq 0$ for all $z \neq 0$ in $[-\bar{\varepsilon}_1, \bar{\varepsilon}_2]$. Lemma 1 shows that for some k sufficiently small

$$\bar{u}^+(\bar{\varepsilon}_2, h_0, t, x) \equiv \phi_2(\bar{z}, c_1) + \tfrac{1}{2}\bar{q}(\bar{\varepsilon}_2, t)\{1 + \tanh k\bar{z}\} \tag{6.16}$$

is an upper function, where $\bar{z} \equiv x - c_1 t + \bar{h}(\bar{\varepsilon}_2, t)$ and where $\bar{h}(\bar{\varepsilon}_2, t)$ and $\bar{q}(\bar{\varepsilon}_2, t)$ are given by (3.12). Further, let γ in (c_1, c_2) be chosen arbitrarily near c_1. Then Lemma 1 shows that

$$\underline{u}^+(\bar{\varepsilon}_1, h_0, \gamma, t, x) \equiv \phi(\underline{z}, \gamma) - \underline{q}(\bar{\varepsilon}_1, t)\{\phi(\underline{z}, \gamma) - \phi_-\}/(\phi_+ - \phi_-) \tag{6.17}$$

is a lower function, where $\underline{z} \equiv x - \gamma t - \underline{h}(\bar{\varepsilon}_1, t)$ and where $\underline{h}(\bar{\varepsilon}_1, t)$ and $\underline{q}(\bar{\varepsilon}_1, t)$ are given by (3.12). The upper and lower functions in (6.16) and (6.17) will enable us to show that the upper part of $u(t,x)$ travels with speed c_1.

We now construct upper and lower functions with speed c_0. For our lower function, we will use

$$\underline{u}^-(h_0, t, x) \equiv \phi_1(x - c_0 t - h_0, c_0). \tag{6.18}$$

To construct an upper function, consider the phase plane of (1.2) at c_0. Select $\bar{\phi} > \phi_-$ so that $\hat{\phi}, 0$ is *not* a singular point for all $\hat{\phi}$ in $(\phi_-, \bar{\phi}]$. Let $v(\phi, \alpha)$ be the trajectories of the solutions which decay to $\phi_-, 0$ as $x \to -\infty$ and which have $v(\hat{\phi}, \alpha) = \alpha$. In particular, let $v(\phi, \alpha_0)$ be the trajectory of $\phi_1(x - c_0 t, c_0)$. Consider $v(\phi, \alpha_0 + \eta)$ for small $\eta > 0$. For $\eta > 0$ small enough, the phase plane shows that $v(\phi, \alpha_0 + \eta) > 0$ for all $\phi_- < \phi \leq \phi_+ + \bar{\varepsilon}_2$. Let $u = \phi(x - c_0 t, c_0, \eta)$ be a solution of (1.1) represented by $v(\phi, \alpha_0 + \eta)$. Also let $\bar{x}(\eta)$ be the smallest x at which $\phi(\bar{x}(\eta), c_0, \eta) = \phi_+ + \bar{\varepsilon}_2$. Then for all $\eta > 0$ sufficiently small, $\phi(x, c_0, \eta)$ decays to ϕ_- at the usual rate and $\phi(x, c_0, \eta)$ is increasing in x for all $x \leq \bar{x}(\eta)$. Finally, note that $\bar{u}(\bar{\varepsilon}_2, t) \equiv \phi_+ + \bar{\varepsilon}_2 e^{-\mu t}$ is an upper function of (1.1) for all small enough $\mu > 0$. For our upper functions at $c = c_0$, we will use both

$$\begin{aligned} &\bar{u}^-(\eta, h_0, t, x) \equiv \phi(x - c_0 t + h_0, c_0, \eta) \quad \text{for all } x \leq \bar{x}(\eta) + c_0 t - h_0, \\ &\bar{u}(\bar{\varepsilon}_2, t) \equiv \phi_+ + \bar{\varepsilon}_2 e^{-\mu t} \quad \text{for all } x. \end{aligned} \tag{6.19}$$

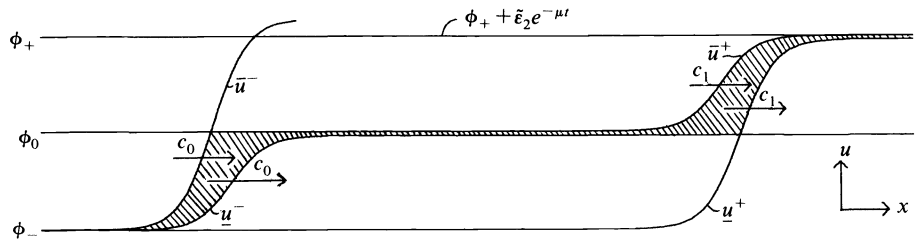

FIG. 5. *Relation* (6.20) *implies that* $u(t,x)$ *must remain in the shaded region for all* $t > 0$.

Since $u(t,x)$ satisfies (6.3) and (6.15), for each $\gamma > c_1$ we can choose h_0 so large that

$$\begin{aligned} &\underline{u}^+(\bar{\varepsilon}_1, h_0, \gamma, t, x) \leq u(t,x) \leq \bar{u}^+(\bar{\varepsilon}_2, h_0, t, x) \quad \text{for all } x, \\ &\underline{u}^-(h_0, t, x) \leq u(t,x) \leq \phi_+ + \bar{\varepsilon}_2 e^{-\mu t} \quad \text{for all } x, \\ &u(t,x) \leq \bar{u}^-(\eta, h_0, t, x) \quad \text{for all } x \leq \bar{x}(\eta) + c_0 t - h_0, \end{aligned} \tag{6.20}$$

is satisfied at $t = 0$. The maximum principle now implies that (6.20) is true for all $t \geq 0$. Relation (6.20) is illustrated in Fig. 5 for t large. Specifically, (6.20) implies that $u(t,x)$

must remain in the shaded region for all time. Since γ can be chosen arbitrarily near c_1, $u(t,x)$ clearly evolves into two stacked waves: an upper wave with speed c_1 and a lower wave with speed c_0.

Now, suppose that at some $c_1' > -\infty$ the secondary $N \to S$ waves $\phi_1(x-ct,c)$ shed another monotonic $S \to S$ wave $\phi_{12}(x-c_1't, c_1')$, leaving the family of $N \to S$ waves $\phi_{11}(x-ct,c)$, $c_1'' < c \le c_2''$, where $c_1'' < c_1' < c_2''$. Then if $u(0,x)$ satisfies (6.3) and (6.15) for some c_0 with $c_1'' < c_0 < c_1' < c_1$, similar arguments show that $u(t,x)$ evolves into three stacked waves: an upper wave with speed c_1, a middle wave with speed c_1' and a lower wave with speed c_0.

Similarly, if further bifurcations occur as c decreases, then some solutions will evolve into four or more stacked traveling waves. Note that only a finite number of bifurcations can occur as $c \to -\infty$, because there can only be a finite number of saddle points between $\phi_-, 0$ and $\phi_+, 0$.

The following theorem summarizes this discussion.

THEOREM 9 (stacked waves in the $N \to S$ case). *Assume that* H1, H2 *and* H3 *are satisfied. Suppose that* $u = \phi(x-\tilde{c}t, \tilde{c})$ *is a bounded, increasing in* x, *solution of* (1.1). *Let* $\phi_- \equiv \phi(-\infty, \tilde{c})$, *let* $\phi_+ \equiv \phi(+\infty, \tilde{c})$ *and assume that* $\phi_-, 0$ *is a node and* $\phi_+, 0$ *is a saddle point of* (1.2) *at* $c = \tilde{c}$. *Define* c_1, c_2 *and* $k_1^-(c)$ *as in Theorem* 7 *and assume that* $c_1 > -\infty$. *Let the* $n \ge 1$ *speeds at which the waves* $\phi(x-ct,c)$ *bifurcate be* $c_1 \equiv c_1^{(1)} > c_1^{(2)} > \cdots > c_1^{(n)}$. *Finally, assume that each bifurcation is into exactly one increasing in* x $S \to S$ *wave* $u \equiv \phi_{SS}^{(i)}(x-c_1^{(i)}t, c_1^{(i)})$ *and a family of increasing in* x $N \to S$ *waves* $u \equiv \phi_{NS}^{(i)}(x-ct,c)$, $c_1^{(i+1)} < c \le c_2^{(i+1)}$, *where* $c_1^{(i+1)} < c_1^{(i)} < c_2^{(i+1)}$ *and where* $c_1^{(n+1)} \equiv -\infty$. *Define* $\phi_0^{(0)} > \phi_0^{(1)} > \cdots > \phi_0^{(n)}$ *by* $\phi_0^{(0)} \equiv \phi_+$, $\phi_{NS}^{(i)}(-\infty, c) = \phi_-$, $\phi_{NS}^{(i)}(+\infty, c) = \phi_{SS}^{(i)}(-\infty, c_1^{(i)}) = \phi_0^{(i)}$, $\phi_{SS}^{(i)}(+\infty, c_1^{(i)}) \equiv \phi_0^{(i-1)}$, $i = 1, \cdots, n$.

Let $\varepsilon_1^+ > 0$ *and* $\varepsilon_2^+ > 0$ *be any constants small enough so that* (3.7) *is satisfied. Let* $u(t,x)$ *be any solution of* (1.1) *in* B_x^2 *and suppose that for some* $x_0 > 0$, $\alpha > 0$ *and* $\beta > 0$, $u(0,x)$ *satisfies* (6.3) *and* (6.15) *for some* c_0 *in* $(c_1^{(j+1)}, c_1^{(j)})$ *for some* $j \ge 1$. *Then* $u(t,x)$ *evolves into a stack of* $j+1$ *waves with speeds* $c_1^{(i)}$, $i = 1, \cdots, j$, *and* c_0. *Specifically, for every* η

$$\lim_{t \to +\infty} u(t, \eta + ct) = \begin{cases} \phi_+ \equiv \phi_0^{(0)} & \text{if } c > c_1 \equiv c_1^{(1)}, \\ \phi_\wedge^{(i)} & \text{if } c_1^{(i+1)} < c < c_1^{(i)}, \qquad i = 1, \cdots, j-1, \\ \phi_0^{(j)} & \text{if } c_0 < c < c_1^{(j)}, \\ \phi_- & \text{if } c < c_0. \end{cases}$$

We note that Theorem 9 resembles some of the results in [5], [8].

For any particular case, the contents of Theorems 8 and 9 can be summarized in a diagram like Fig. 6. These diagrams are constructed by

(i) drawing $k = k_2^-(c)$ with dashed lines, drawing $k = k_1^-(c)$ for $c < c_2$ with a solid curve and drawing $k = k_1^-(c)$ for $c_2 < c \le c_{\max}$ with dashed lines if $c_2 < c_{\max}$;

(ii) marking the curve $k = k_1^-(c)$ with x's for all $c < c_2$;

(iii) marking the point c_2, $k_2^-(c_2)$ with an x and drawing the solid vertical line $c = c_2$, $k > k_1^-(c_2)$;

(iv) drawing solid vertical lines $c = c_1^{(i)}$, $0 \le k < k_1^-(c_1^{(i)})$ and marking the points $k = 0$, $c = c_1^{(i)}$ with x's, where $c = c_1^{(i)}$ are the speeds (if any) at which the monotonic $N \to S$ waves shed a monotonic $S \to S$ wave.

Then the points c, k of the monotonic solutions $u = \phi(x-ct,c)$ have been marked with x's, where k is given by $\phi(x,c) \sim \phi_- + ae^{kx}$ as $x \to -\infty$. Moreover, note that each horizontal line $k = \bar{k}$ intersects the solid lines and curve at $m \ge 1$ speeds $c_1 > c_2 > \cdots > c_m$.

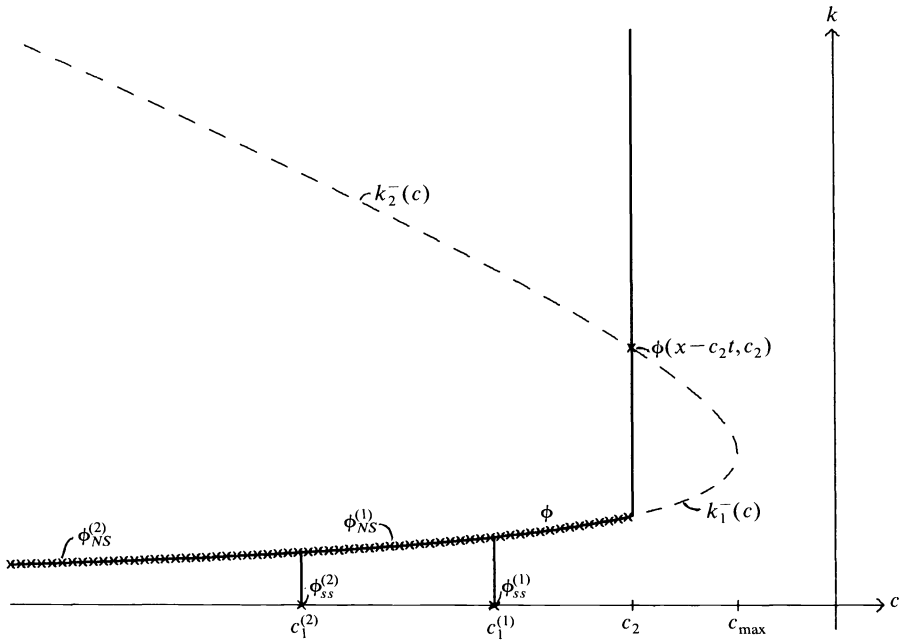

FIG. 6. *Summary of Theorems* 8 *and* 9 *for a monotonic* $N\to S$ *wave with* $c_2<c_{\max}$, *which sheds one* $S\to S$ *wave at* $c=c_1^{(1)}$ *and another at* $c=c_1^{(2)}$.

The upshot of Theorems 8 and 9 is that if $u(0,x)$ satisfies (6.3) and if $0<\alpha<[u(0,x)-\phi_-]\cdot\exp(-\bar{k}x)<\beta$ then $u(t,x)$ evolves into a stack of m waves with speeds c_i, $i=1,\cdots,m$.

The results for $S\to N$ waves are essentially the same as Theorems 7, 8 and 9. In fact, if $u=\phi(x-ct,c)$ is a $S\to N$ wave, then the transformation $x\to -x$, $c\to -c$ applied to (1.1), (1.2) and $u=\phi(x-ct,c)$ reduces the $S\to N$ wave to a $N\to S$ wave.

7. $N\to N$ waves. The following theorems are the analogues of Theorems 7, 8 and 9 for monotonic $N\to N$ waves. The proofs of these theorems are omitted since they involve no new ideas.

THEOREM 10 (existence for $N\to N$ waves). *Assume that* H1, H2 *and* H3 *are satisfied. Suppose that* $u=\phi(x-c_0t,c_0)$ *is a bounded, increasing in* x, *solution of* (1.1). *Let* $\phi_-\equiv\phi(-\infty,c_0)$, *let* $\phi_+\equiv\phi(+\infty,c_0)$ *and assume that* $\phi_-,0$ *and* $\phi_+,0$ *are both nodes of* (1.2) *at* $c=c_0$. *Finally, assume that* $c_{\min}<c_{\max}$, *where*

$$
\begin{aligned}
c_{\max}&\equiv -2\{f_1(0,0,\phi_-)f_3(0,0,\phi_-)\}^{1/2}-f_2(0,0,\phi_-),\\
c_{\min}&\equiv +2\{f_1(0,0,\phi_+)f_3(0,0,\phi_+)\}^{1/2}-f_2(0,0,\phi_+).
\end{aligned}
\tag{7.1}
$$

Then there is a c_1 *and a* c_2 *with* $c_{\min}\le c_1<c_2\le c_{\max}$ *such that for each* c *in* (c_1,c_2) *there is a family of increasing in* x $N\to N$ *solutions* $u=\phi(x-ct,c,\alpha)$, $0<\alpha\le 1$, *of* (1.1). *Moreover,*

1) $\phi(x,c,\alpha)$ *and* $\phi_x(x,c,\alpha)$ *are differentiable in* c *and* α;

2) $\phi(-\infty,c,\alpha)=\phi_-,\phi(+\infty,c,\alpha)=\phi_+$ *and the phase plane trajectories* $v(\phi,c,\alpha)$ *of* $\phi(x,c,\alpha)$ *are increasing in* α *at each* ϕ *in* (ϕ_-,ϕ_+);

3) $\phi(x,c,1)$ *decays at the accidental rate either as* $x\to-\infty$, *as* $x\to+\infty$, *or both. Further, there is at most one* $\bar{c}$ *in* (c_1,c_2) *such that* $\phi(x,\bar{c},1)$ *decays at the accidental rate both as* $x\to-\infty$ *and as* $x\to+\infty$;

4) *at each* c *in* (c_1, c_2) *the trajectory* $v(\phi, c, \alpha)$ *at* $\alpha = 0$ *intersects the* $v = 0$ *axis at* $n \geq 1$ *saddle points* $\phi_0^{(i)}$, 0, $i = 1, \cdots, n$. *Letting* $\phi_- \equiv \phi_0^{(0)} < \phi_0^{(1)} < \cdots < \phi_0^{(n+1)} \equiv \phi_+$, *the trajectory* $v(\phi, c, 0)$ *represents* $n+1$ *monotonic solutions* $u = \phi^{(i)}(x - ct, c)$ *with* $\phi^{(i)}(-\infty, c) = \phi_0^{(i)}$ *and* $\phi^{(i)}(+\infty, c) = \phi_0^{(i+1)}$, $i = 0, \cdots, n$. *Specifically,* $\phi^{(0)}(x - ct, c)$ *is a* $N \to S$ *wave and* $\phi^{(n)}(x - ct, c)$ *is a* $S \to N$ *wave*;

5) *if* $c_1 > c_{\min}$ *there is an increasing in* x $S \to N$ *wave* $u = \phi_{SN}(x - c_1 t, c_1)$ *which decays to* ϕ_+ *at the accidental rate as* $x \to +\infty$ *and also has* $\phi_- < \phi_{SN}(-\infty, c_1) < \phi_+$. *If* $c_2 < c_{\max}$ *there is an increasing in* x $N \to S$ *wave* $u = \phi_{NS}(x - c_2 t, c_2)$ *which decays to* ϕ_- *at the accidental rate as* $x \to -\infty$ *and has* $\phi_- < \phi_{NS}(+\infty, c_2) < \phi_+$;

6) *if* $u = \phi(x - ct, c)$ *is a monotonic solution of* (1.1) *with* $\phi(-\infty, c) = \phi_-$ *and* $\phi(+\infty, c) = \phi_+$, *then* $c_1 \leq c \leq c_2$. *Additionally, if* $c_1 < c < c_2$, *then there is an* α *in* $(0, 1]$ *and an* h *such that* $\phi(x - ct, c) \equiv \phi(x - ct + h, c, \alpha)$ *for all* x *and all* t.

THEOREM 11 (wavespeed for $N \to N$ waves). *Assume that* H1 H2 *and* H3 *are satisfied. Suppose that* $u = \phi(x - c_0 t, c_0)$ *is a bounded, increasing in* x, *solution of* (1.1). *Let* $\phi_- \equiv \phi(-\infty, c_0)$, *let* $\phi_+ \equiv \phi(+\infty, c_0)$ *and assume that both* $\phi_-, 0$ *and* $\phi_+, 0$ *are nodes of* (1.2) *at* $c = c_0$. *Let* $c_{\min}$, $c_{\max}$, c_1 *and* c_2 *be as in Theorem* 10 *and let* $k_{1,2}^{\pm}(c)$ *be defined by* (3.1) *and* (3.5). *Finally, assume that* $c_{\min} < c_{\max}$.

Suppose that $u(t, x)$ *is any solution of* (1.1) *in* B_x^2.

1) *If for some* c *in* (c_1, c_2),

$$\begin{aligned} &\phi_- < u(0, x) < \phi_+ \quad \text{for all } x, \\ &0 < \alpha < \left[u(0, x) - \phi_-\right] e^{-k_1^-(c)x} < \beta \quad \text{for all } x \leq 0, \\ &0 < \gamma < \left[\phi_+ - u(0, x)\right] e^{-k_1^+(c)x} < \delta \quad \text{for all } x \geq 0 \end{aligned} \tag{7.2}$$

for some $\alpha, \beta, \gamma, \delta$, *then* $u(t, x)$ *travels with speed* c.

2) *If for some* c *in* (c_1, c_2), $\phi(x, c, 1)$ *decays at the accidental rate as* $x \to -\infty$ *and at the usual rate as* $x \to +\infty$ *and if*

$$\begin{aligned} &\phi_- \leq u(0, x) < \phi_+ \quad \text{for all } x, \\ &\left[u(0, x) - \phi_-\right] e^{-k_1^-(c)x} < \alpha \quad \text{for all } x \leq 0, \\ &0 < \beta < \left[\phi_+ - u(0, x)\right] e^{-k_1^+(c)x} < \gamma \quad \text{for all } x \geq 0, \end{aligned} \tag{7.3}$$

for some α, β, γ, *then* $u(t, x)$ *travels with speed* c.

3) *If for some* c *in* (c_1, c_2), $\phi(x, c, 1)$ *decays at the accidental rate as* $x \to +\infty$ *and at the usual rate as* $x \to -\infty$, *and if*

$$\begin{aligned} &\phi_- < u(0, x) \leq \phi_+ \quad \text{for all } x, \\ &0 < \alpha < \left[u(0, x) - \phi_-\right] e^{-k_1^-(c)x} < \beta \quad \text{for all } x \leq 0, \\ &\left[\phi_+ - u(0, x)\right] e^{-k_1^+(c)x} < \gamma \quad \text{for all } x \geq 0, \end{aligned} \tag{7.4}$$

for some α, β, γ, *then* $u(t, x)$ *travels with speed* c.

4) *If for some* c *in* (c_1, c_2), $\phi(x, c, 1)$ *decays at the accidental rate both as* $x \to -\infty$ *and as* $x \to +\infty$ *and if*

$$\begin{aligned} &\phi_- \leq u(0, x) \leq \phi_+, \quad \text{for all } x, \\ &\left[u(0, x) - \phi_-\right] e^{-k_1^-(c)x} < \alpha \quad \text{for all } x \leq 0, \\ &\left[\phi_+ - u(0, x)\right] e^{-k_1^+(c)x} < \beta \quad \text{for all } x \geq 0 \end{aligned} \tag{7.5}$$

for some α, β, *then* $u(t, x)$ *travels with speed* c.

Roughly speaking, if $u(0,x)\sim\phi_-+a\exp(k^-x)$ as $x\to-\infty$, if $u(0,x)\sim\phi_++b\exp(-k^+x)$ as $x\to+\infty$ and if

$$k^-\geq k_1^-(c^*),\qquad k^+\geq -k_1^+(c^*)\quad\text{for some } c^* \text{ in } (c_1,c_2), \tag{7.6}$$

then Theorem 11 yields the speed c of $u(t,x)$. If (7.6) is *not* satisfied, then $u(t,x)$ evolves into a stack of two or more traveling waves. The following theorem gives precise results for the simplest case. The treatment of the other cases is similar.

THEOREM 12 (stacked waves in the $N\to N$ case). *Assume that* H1, H2 *and* H3 *are satisfied. Suppose $u=\phi(x-c_0t,c_0)$ is a bounded, increasing in x, solution of* (1.1). *Let $\phi_-\equiv\phi(-\infty,c_0)$, let $\phi_+\equiv\phi(+\infty,c_0)$ and assume that $\phi_-,0$ and $\phi_+,0$ are both nodes of* (1.2) *at $c=c_0$. Let $c_{\min}$, $c_{\max}$, c_1, c_2 and $k_{1,2}^{\pm}(c)$ be defined as in Theorem* 11 *and assume that $c_{\min}<c_{\max}$. Finally, let $\phi_{NS}(x-ct,c)$ and $\phi_{SN}(x-ct,c)$ be the $N\to S$ and $S\to N$ waves represented by the trajectory $v(\phi,c,0)$ in Theorem* 11. *Suppose that ϕ_{NS} does not shed any $S\to S$ waves for all $c\leq c_2$ and that ϕ_{SN} does not shed any for all $c\geq c_1$. Then $\phi_{NS}(-\infty,c)=\phi_-$, $\phi_{NS}(+\infty,c)=\phi_0$ for all $c\leq c_2$ and $\phi_{SN}(-\infty,c)=\phi_0$, $\phi_{SN}(+\infty,c)=\phi_+$ for all $c\geq c_1$ for some saddle point $\phi_0,0$ with $\phi_-<\phi_0<\phi_+$.*

Suppose that $u(t,x)$ is any solution of (1.1) *in B_x^2 satisfying*

$$\begin{aligned}
&\phi_-<u(0,x)<\phi_+ && \text{for all } x,\\
&\alpha<\left[u(0,x)-\phi_-\right]e^{-k^-x}<\beta && \text{for all } x\leq 0,\\
&\gamma<\left[\phi_+-u(0,x)\right]e^{k^+x}<\delta && \text{for all } x\geq 0
\end{aligned} \tag{7.7}$$

for some positive constants $\alpha,\beta,\gamma,\delta,k^-$ and k^+. If k^- and k^+ do not satisfy (7.6), *then $u(t,x)$ evolves into two stacked waves with speeds c_L and c_U. That is, for all η,*

$$\lim_{t\to\infty}u(t,\eta+ct)\equiv\begin{cases}\phi_- & \text{if } c<c_L,\\ \phi_0 & \text{if } c_L<c<c_U,\\ \phi_+ & \text{if } c_U<c,\end{cases} \tag{7.8}$$

where

1) *if $k^-=k_1^-(c_-)$ for some $c_-\leq c_2$, if $k^+=-k_1^+(c_+)$ for some $c_+\geq c_1$, and if $c_-<c_+$, then $c_L=c_-$ and $c_U=c_+$;*

2) *if $k^-=k_1^-(c_-)$ for some $c_-<c_1$ and if $k^+\geq -k_1^+(c_1)$, then $c_L=c_-$ and $c_U=c_1$; and*

3) *if $k^-\geq k_1^-(c_2)$ and if $k^+=-k_1^+(c_+)$ for some $c_+>c_2$, then $c_L=c_2$ and $c_U=c_+$.*

Theorems 11 and 12 are illustrated in Figs. 7 and 8 for the case where $c_1=c_{\min}$ and $c_2=c_{\max}$, where the underlying $N\to S$ wave ϕ_{NS} does not shed any $S\to S$ waves for $c\leq c_2$ and where the underlying $S\to N$ wave ϕ_{SN} does not shed any $S\to S$ waves for $c\geq c_1$. Additionally, for Fig. 7 we have assumed that for each c with $c_1\leq c\leq c_2$, the $N\to N$ wave $\phi(x-ct,c,1)$ decays at the usual rate as $x\to-\infty$ and at the accidental rate as $x\to+\infty$. For Fig. 8 we have assumed that there is a $\tilde{c}$ in (c_1,c_2) such that the $N\to N$ wave $\phi(x-\tilde{c}t,\tilde{c},1)$ decays at the accidental rate both as $x\to-\infty$ and as $x\to+\infty$.

In each figure we have plotted $k_i^-(c)$, $-k_j^+(c)$, $i=1,2$, $j=1,2$, with dashed lines. We have also marked with x's the points (k^-,k^+) for the monotonic waves $u=\phi(x-ct,c)$, where $\phi(x,c)\sim\phi_-+ae^{k^-x}$ as $x\to-\infty$ and $\phi(x,c)\sim\phi_+-be^{-k^+x}$ as $x\to+\infty$ determines k^- and k^+.

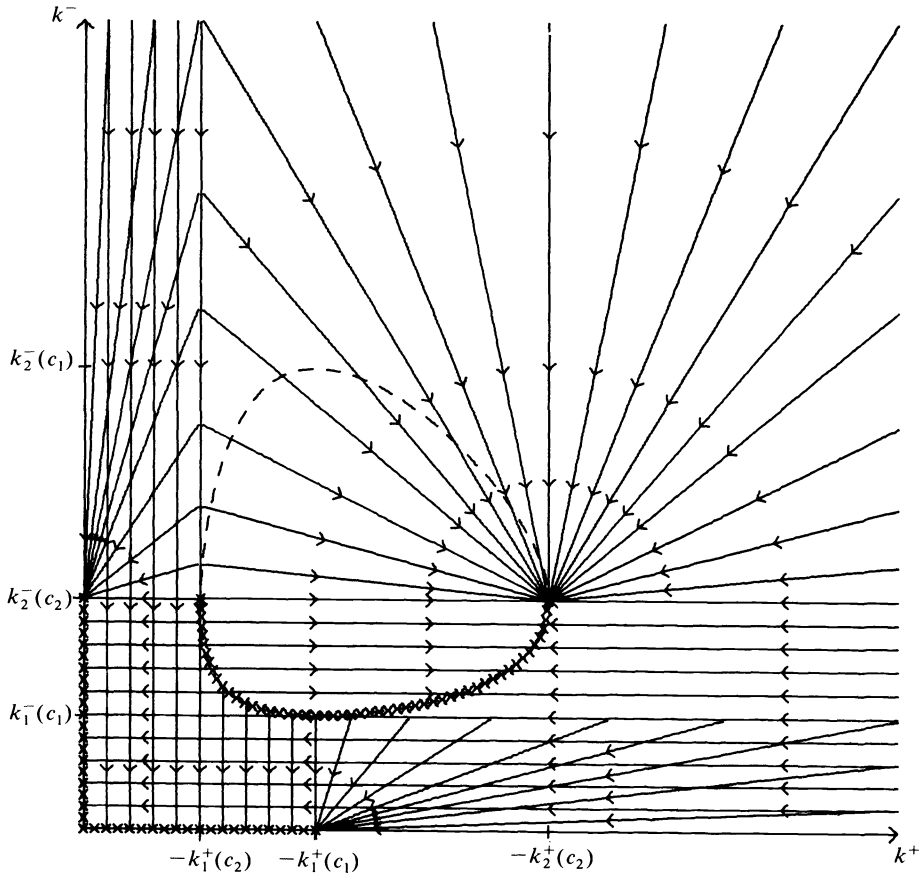

FIG. 7. *Summary of Theorems* 11 *and* 12 *when the* $N \to N$ *wave* $\phi(x-ct, c, 1)$ *decays at the usual rate as* $x \to -\infty$ *and at the accidental rate as* $x \to +\infty$ *for each* c *in* $[c_1, c_2]$.

Assume that $u(0,x)$ satisfies (7.7) for some $k^- > 0$, $k^+ > 0$. For the case treated by Fig. 7, Theorems 11 and 12 imply the following:

(A) Suppose that the point k^-, k^+ is marked with an x. Let $u = \phi(x-ct)$ be the wave represented by the x. Then $u(t,x)$ travels with speed c, which can be determined either from $k^- = k_{1,2}^-(c)$ or $k^+ = -k_{1,2}^+(c)$.

(B) Suppose a single trajectory crosses the point k^-, k^+ in Fig. 7. Follow the trajectory until it reaches a point $\bar{k}^-$, $\bar{k}^+$ marked with an x and let $u = \phi(x-\bar{c}t)$ be the wave represented by the x. Then $u(t,x)$ travels with speed $\bar{c}$.

(C) If neither A nor B occurs, then two trajectories cross the point k^-, k^+. Follow these trajectories until they reach points $\bar{k}^-$, $\bar{k}^+$ and $\bar{\bar{k}}^-$, $\bar{\bar{k}}^+$ marked with x's and let $\bar{\phi}(x-\bar{c}t)$ and $\bar{\bar{\phi}}(x-\bar{\bar{c}}t)$ be the waves represented by these x's. Then $u(t,x)$ evolves into two stacked traveling waves: one wave between $\bar{\phi}(-\infty)$ and $\bar{\phi}(+\infty)$ with speed $\bar{c}$ and one wave between $\bar{\bar{\phi}}(-\infty)$ and $\bar{\bar{\phi}}(+\infty)$ with speed $\bar{\bar{c}}$.

Fig. 8 summarizes Theorems 11 and 12 for its case similarly. For any other cases, the evolution of $u(t,x)$ into traveling and stacked traveling waves can be similarly found in terms of k^-, k^+ and summarized in diagrams like Figs. 7 and 8.

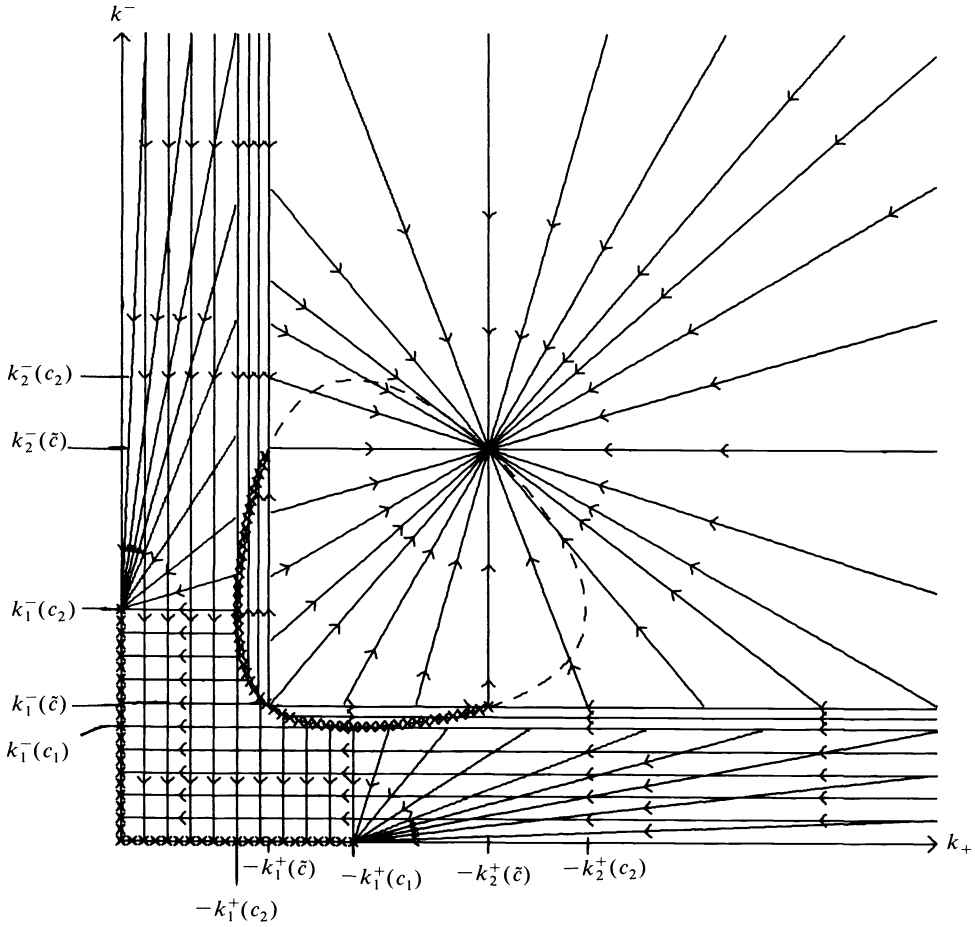

FIG. 8. *Summary of Theorems* 11 *and* 12 *when the* $N \to N$ *wave* $\phi(x-\tilde{c}t, \tilde{c}, 1)$ *decays at the accidental rate both as* $x \to -\infty$ *and as* $x \to +\infty$ *for some* $\tilde{c}$ *in* $[c_1, c_2]$.

8. Genericity. The following theorem shows that the existence of any particular type of traveling wave is generic.

THEOREM 13 (genericity). *Assume that* H1, H2 *and* H3 *are satisfied and assume that* $u=\phi(x-c_0t)$ *is a bounded, increasing in* x, *solution of* (1.1). *Let* $\phi_- \equiv \phi(-\infty)$, $\phi_+ \equiv \phi(+\infty)$ *and suppose that* $g(u_{xx}, u_x, u)$ *is any function that satisfies* H1. *Then there is an* $\varepsilon_0 > 0$ *such that, for each* ε *in* $[0, \varepsilon_0]$,

$$u_t = f(u_{xx}, u_x, u) + \varepsilon g(u_{xx}, u_x, u) \tag{8.1}$$

has an increasing in x *solution* $u=\phi(x-c(\varepsilon)t, \varepsilon)$. *Moreover,*

1) $\phi(x,\varepsilon)$, $c(\varepsilon)$, $\phi(-\infty,\varepsilon)$ *and* $\phi(+\infty,\varepsilon)$ *are continuously differentiable in* ε *with* $\phi(x,0)\equiv\phi(x)$, $c(0)=c_0$, $\phi(-\infty,\varepsilon)=\phi_-$ *and* $\phi(+\infty,0)=\phi_+$;

2) $\phi(-\infty,\varepsilon), 0$ *is a saddle point* [*node*] *of*

$$\phi' = v, \qquad f(v', v, \phi) + \varepsilon g(v', v, \phi) + c(\varepsilon)v = 0 \tag{8.2}$$

if and only if $\phi_-, 0$ *is a saddle point* [*node*] *of* (1.2) *at* $c=c_0$;

3) $\phi(+\infty,\varepsilon), 0$ *is a saddle point* [*node*] *of* (8.2) *if and only if* $\phi_+, 0$ *is a saddle point* [*node*] *of* (1.2) *at* $c=c_0$;

4) $\phi(x,\varepsilon)$ *decays at the usual rate* [*accidental rate*] *as* $x\to-\infty$ *if and only if* $\phi(x)$ *decays at the usual rate* [*accidental rate*] *as* $x\to-\infty$. *Similarly,* $\phi(x,\varepsilon)$ *decays at the usual rate* [*accidental rate*] *as* $x\to+\infty$ *if and only if* $\phi(x)$ *decays at the usual rate* [*accidental rate*] *as* $x\to+\infty$.

Now Theorem 4 yields exactly the same type of stability for the solution $u=\phi(x-c(\varepsilon)t,\varepsilon)$ of (8.1) as it does for the solution $u=\phi(x-c_0t)$ of (1.1). The only difference is that the constants $k^{\pm}_{1,2}(c_0)$ must be replaced by their analogues $k^{\pm}_{1,2}(c(\varepsilon),\varepsilon)$. Roughly speaking, Theorems 4 and 13 show that the existence and stability of any particular type of monotonic traveling wave are generic. Similarly, the results in Theorems 5–12 are generic also.

Appendix. Here we extend the stability results of Theorems 3 and 4 to the case of $f(0,0,u)\equiv 0$ for all u and point out some other extensions of these theorems.

THEOREM 14 (stability of constant solutions). *Assume that* H1 *and* H2 *are satisfied and that* $f(0,0,u)=0$ *for all* u. *Let* ϕ_0 *be any constant. Then* $u(t,x)\equiv\phi_0$ *is a constant solution of* (1.1). *Furthermore, if* $u(t,x)$ *is any solution of* (1.1) *in* B_x^2 *that satisfies*

$$\phi_0-\varepsilon\le u(t,x)\le\phi_0+\varepsilon \quad \text{for all } x \tag{A.1}$$

at $t=0$ *for some* ε, *then* (A.1) *is true for all* $t\ge 0$.

THEOREM 15 (stability of monotonic waves). *Assume that* H1 *and* H2 *are satisfied and that* $f(0,0,u)\equiv 0$ *for all* u. *Suppose that* $u=\phi(x-ct)$ *is a bounded, increasing in* x, *solution of* (1.1) *that satisfies*

$$\begin{aligned}\phi(x)&=\phi(-\infty)+ae^{k^-x}+o\left[e^{(k^-+\rho)x}\right] && \text{as } x\to-\infty,\\ \phi(x)&=\phi(+\infty)-be^{-k^+x}+o\left[e^{-(k^++\rho)x}\right] && \text{as } x\to+\infty,\end{aligned} \tag{A.2}$$

for some positive constants a,b,k^-,k^+ *and* ρ. *Also assume that the asymptotic behavior of* $\phi'(x)$ *and* $\phi''(x)$ *is correctly obtained by differentiating* (A.2). *Let* $\eta>0$ *be any constant. Then for every* $\varepsilon>0$ *there is a* $\delta(\varepsilon)>0$, *such that if* $u(t,x)$ *is any solution of* (1.1) *in* B_x^2 *that satisfies*

$$|u(0,x)-\phi(x)|\le\delta(\varepsilon)e^{-\eta|x|} \quad \text{for all } x, \tag{A.3}$$

then $|u(t,x)-\phi(x)|\le\varepsilon$ *for all* x *and all* $t\ge 0$.

Theorem 14 is proven by noting that for any $\varepsilon>0$, $\bar{u}\equiv\phi_0+\varepsilon$ and $\underline{u}\equiv\phi_0-\varepsilon$ are solutions of (1.1) and then using the maximum principle. Theorem 15 is proven by first showing that for all small enough $\varepsilon>0$, there is a $K>0$ such that

$$\begin{aligned}&u(\varepsilon,\delta,h_0,t,x)\equiv\phi(x-ct+h(t))+\eta(t)\operatorname{sech}\varepsilon[x-ct+h(t)], \quad \text{where}\\ &\eta(t)\equiv\delta\varepsilon^4\varepsilon^{2/\Delta}e^{-\varepsilon^3t}, \qquad h(t)\equiv h_0+K\delta\varepsilon^{1+2/\Delta}\left(1-e^{-\varepsilon^3t}\right)\end{aligned} \tag{A.4}$$

is an upper function for all small enough $\delta>0$, and is a lower function for all large enough $\delta<0$. The theorem then follows from the maximum principle.

When $f(0,0,u)\equiv 0$, (1.1) often has important unsteady solutions, whose stability can be determined via the maximum principle. For example, the stability of the N-wave solutions of Burger's equation can be found [4] by this approach.

Theorems 3 and 4 can be extended to many other classes of equations which possess maximum principles. For example,

$$u_t=f\left(u_{xx},u_x,u,\int_0^T\int_{-Y}^{Y}g[s,y,u(t-s,x-y)]\,dy\,ds\right) \tag{A.5}$$

has a maximum principle if $f_1(a,b,c,d)\geq 1$ and $f_4(a,b,c,d)\ g_3(\alpha,\beta,\gamma)\geq 0$ for all $a,b,c,d,\alpha,\beta,\gamma$. Similarly, the equation

$$u_t=f(u_{xx},u_{xy},u_{yy},u_x,u_y,u) \tag{A.6}$$

has a maximum principle if $f_1\cos^2\theta+2f_2\cos\theta\sin\theta+f_3\sin^2\theta\geq 1$ for all θ and all arguments of f. Also, the system

$$u_t^{(k)}=f^{(k)}\left(u_{xx}^{(k)},u_x^{(k)},\vec{u}\right),\qquad k=1,\cdots,m, \tag{A.7}$$

has a maximum principle if for each k, $f_1^{(k)}\geq 1$ and $\partial f^{(k)}/\partial u^{(l)}\geq 0$ for all $l\neq k$. These maximum principles can be used to find the stability of the monotonic solutions $u(t,x)=\phi(x-ct)$ of (A.5), of the monotonic plane waves $u(t,x,y)=\phi(k_1x+k_2y-ct)$ of (A.6) and of the monotonic solutions $\vec{u}(t,x)=\vec{\phi}(x-ct)$ of (A.7). These extensions are developed in [4].

REFERENCES

[1] P. S. HAGAN, *The instability of non-monotonic wave solutions of parabolic equations*, Stud. Appl. Math., 67 (1981), pp. 57–88.

[2] G. B. WHITMAN, *Linear and Nonlinear Waves*, John Wiley, New York, 1974.

[3] M. H. PROTTER AND H. F. WEINBERG, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[4] P. S. HAGAN, *Stability of waves in parabolic equations*, Thesis, California Institute of Technology, Pasadena, 1979.

[5] P. C. FIFE AND J. B. MCLEOD, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Rat. Mech. Anal. 65 (1977), pp. 335–361.

[6] P. C. FIFE, *Asymptotic states for equations of reaction and diffusion*, Bull. AMS, 84 (1978), pp. 693–776.

[7] ———, *Stationary patterns for reaction-diffusion equations*, Nonlinear Diffusion, Proc. NSF-CBMS Regional Conference on Nonlinear Diffusion, Research Notes in Math., Pitman, London, 14 (1977), pp. 81–121.

[8] ———, *Long time behavior of solutions of bistable nonlinear diffusion equations*, Arch. Rat. Mech. Anal., to appear.

[9] D. G. ARONSON AND H. F. WEINBERGER, *Multidimensional nonlinear diffusions arising in population genetics*, Adv. in Math., 30 (1978), pp. 33–58.

[10] ———, *Nonlinear diffusion in population genetics, combustion and nerve propagation*, in Partial Differential Equations and Related Topics, J. A. Goldstein, ed. Lecture Notes in Mathematics 446, Springer, New York, 1975, pp. 5–49.

[11] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[12] A. N. KOLMOGOROV, I. G. PETROVSKII AND N. S. PISKUNOV, *A study of the equation of diffusion with increase in the quantity of matter and its application to biological problem*, Bjul. Moskovsko Gos. Univ., 1:7 (1937), pp. 1–26. (In Russian)

[13] YA. I. KANEL', *On the stabilization of solutions of the equations of the theory of combustion with initial data of compact support*, Mat. Sb., 65 (1964), pp. 398–413. (In Russian)

[14] D. H. SATTINGER, *On the stability of waves of nonlinear parabolic systems*, Adv. in Math., 22 (1976), pp. 312–355.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0003 $01.00/0

GLOBAL EXISTENCE AND BOUNDEDNESS OF SOLUTIONS TO THE EXTENSIBLE BEAM EQUATION*

W. E. FITZGIBBON†

Abstract. The theory of semigroups of linear operators and the recently developed theory of cosine operators is used to provide a variation of parameters representation of solutions to the equation governing the transverse motion of an extensible beam and to insure L^2 boundedness of these solutions.

We use the theory of semigroups of linear operators and the recently developed theory of cosine operators to provide a variation of parameters representation of solutions to the equation governing the transverse motion of an extensible beam and to insure the L^2 boundedness of these solutions. Specifically, we consider the equation

$$\frac{\partial^2 u}{\partial t^2}+\frac{\alpha\partial^4 u}{\partial x^4}-\left(\beta+\kappa\int_0^1\left(\frac{\partial u(\xi,t)}{\partial\xi}\right)^2 d\xi\right)\frac{\partial^2 u}{\partial x^2}=0, \tag{1}$$

subject to the boundary conditions of the form

$$u(0,t)=u(1,t)=u_{xx}(0,t)=u_{xx}(1,t)=0. \tag{2}$$

Equation (1) was proposed by Wionowsky-Kreiger [30] for the transverse deflection of an extensible beam whose ends are held a fixed distance apart. The boundary conditions correspond to the ends of the beam being hinged. Boundary conditions appropriate for a beam with clamped ends would have the form

$$u(0,t)=u(1,t)=u_x(0,t)=u_x(1,t)=0. \tag{3}$$

Our methods apply equally well to clamped ends, but in the interest of brevity, we limit our discussion to boundary conditions of the form (2). The model differs from the one appearing in elementary treatments in that a nonlinearity has been introduced to represent the change in tension of the beam.

Nonlinear beam equations have been the subject of much recent activity. Ball [1] uses a Galerkin method to obtain weak solutions to (1) and obtains classical solutions by placing further restrictions on the regularity of initial data. In [5] Dickey demonstrates that the model allows a description of dynamic buckling of the beam. Ball [3] and Dickey [6] study the dynamic stability of equilibria of damped versions of (1). Additional treatments of related equations appear in [2], [18], [19], [28], [7], [8].

We treat (1) as an abstract second order differential equation. One way to view abstract second order equations is to rewrite them as first order systems. Our references for this approach shall be Goldstein [12], [13], [15], [16], Pazy [20] and Webb and Travis [23], [25]. An alternative to the first order system approach to abstract second order differential equations is provided by the developing theory of cosine operators. Here the reader is referred to Fattorini [9], [10], [11], Da Prato and Giusti, [4], Goldstein [14], [16], Travis and Webb [23], [24], [25], [26] and Webb [28].

---

* Received by the editors June 5, 1980, and in revised form May 11, 1981. Portions of this work were performed under the auspices of the U.S. Department of Energy, Argonne National Laboratory, Argonne, Illinois 60439. This work was supported by the Applied Mathematical Sciences Research Program (KC-0402) of the Office of Energy Research of the U.S. Department of Energy under contract W-31-109-ENG-38.

† Department of Mathematics, University of Houston, Houston, Texas 77004. This work was performed while the author was at the Department of Mathematics, University of California at San Diego, La Jolla, California 92093.

We now reformulate (1) as an abstract equation. We require that $\alpha, \kappa > 0$, the sign of β is unrestricted. As subsequent analysis shall reveal, we lose no generality by assuming $\alpha = 1$. In what follows X shall denote the Hilbert space $L^2[0,1]$. We define $A: X \to X$ by the equation

$$\begin{aligned} &Au = u'''', \\ (4) \qquad &D(A) = \{u \in X \mid u', u'', u''' \text{ are absolutely continuous,} \\ &\qquad u'''' \in X \text{ and } u(0) = u(1) = u''(0) = u''(1) = 0\}. \end{aligned}$$

It is known [8] that A so defined is a positive self-adjoint operator on X. The eigenvalues are of $\{\lambda_n = (n\pi)^4 \mid n \in Z^+\}$ and the corresponding eigenvectors are $\{z_n(s) = \sqrt{2} \sin n\pi s \mid n \in Z^+, s \in [0,1]\}$. Moreover, we have the following explicit spectral representation for A:

$$(5) \qquad Au = \sum_{n=1}^{\infty} (n\pi)^4 \langle u, z_n \rangle z_n.$$

Fractional powers of A are also positive self-adjoint operators and may be computed

$$(6) \qquad A^{\gamma} u = \sum_{n=1} \left((n\pi)^4\right)^{\gamma} \langle u, z_n \rangle z_n.$$

Specifically,

$$(7) \qquad A^{1/2} = \sum_{n=1}^{\infty} (n\pi)^2 \langle u, z_n \rangle z_n = -u''$$

and

$$(8) \qquad A^{1/4} u = \sum_{n=1}^{\infty} (n\pi) \langle u, z_n \rangle z_n.$$

We observe that

$$(9) \qquad \int_0^1 |u'|^2 = \sum_{n=1}^{\infty} (n\pi)^2 \langle u, z_n \rangle^2 = \langle A^{1/4} u, A^{1/4} u \rangle = \| A^{1/4} u \|^2.$$

We now define a nonlinear function $F: X \to X$ by

$$(10) \qquad Fu = \left(\beta + \kappa \langle A^{1/4} u, A^{1/4} u \rangle\right) A^{1/2} u.$$

In this setting (1) becomes

$$(11) \qquad \ddot{x}(t) + Ax(t) + Fx(t) = 0, \quad x(0) = \varphi, \quad \dot{x}(0) = \psi.$$

We remark that the boundary conditions (2) are accounted for in the specification of the domain of A.

We now convert (11) to a first order system. Toward this end we make $D(A)$ into a Banach space X_A by imposing the Euclidean graph norm

$$(12) \qquad \| x \|_A = \left(\| Ax \|^2 + \| x \|^2\right)^{1/2},$$

and we introduce the Banach space $\hat{X}$ by defining

$$(13) \qquad \hat{X} = X_A \times X$$

with

$$\| [\varphi, \psi] \|_{\hat{X}} = \left(\| \varphi \|_A^2 + \| \psi \|^2\right)^{1/2}.$$

We define an operator $\hat{A}: \hat{X} \to \hat{X}$ using the operator matrix

$$A = \begin{pmatrix} 0 & I \\ -A & 0 \end{pmatrix} \tag{14}$$

with $D(A) = D(A) \times D(A^{1/2})$. It is known [16] that $\hat{A}$ so defined is the infinitesimal generator of a strongly continuous group $\{\hat{T}(t) \mid -\infty < t < \infty\}$ of linear transformations on $\hat{X}$. The nonlinear operator F is used to define an operator $F: X \to X$ in the following fashion:

$$\hat{F}\hat{u} = \begin{pmatrix} 0 \\ Fu \end{pmatrix} \quad \text{for } \hat{u} = [u, v] \in \hat{X}. \tag{15}$$

It should be clear that there exists a positive nondecreasing function $L(\cdot)$, so that

$$\| \hat{F}\hat{u}_1 - \hat{F}\hat{u}_2 \|_{\hat{X}} \leq L(R) \| \hat{u}_1 - \hat{u}_2 \|_{\hat{X}} \quad \text{whenever } \hat{u}_1, \hat{u}_2 \in \hat{X} \text{ and } \sup \| \hat{u}_i \|_{\hat{X}} \leq R. \tag{16}$$

We seek solutions to the first order differential equation:

$$\hat{u}'(t) = \hat{A}\hat{u}(t) - \hat{F}\hat{u}(t), \qquad \hat{u}(0) = \hat{x}_0 = [\varphi, \psi] \in D(\hat{A}). \tag{17}$$

If π_1 and π_2 project $\hat{X}$ onto its first and second coordinates, respectively, we see that (17) has matrix representation

$$\frac{d}{dt} \begin{pmatrix} \pi_1 \hat{u}(t) \\ \pi_2 \hat{u}(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -A & 0 \end{pmatrix} \begin{pmatrix} \pi_1 \hat{u}(t) \\ \pi_2 \hat{u}(t) \end{pmatrix} - \begin{pmatrix} 0 \\ F(\pi_1 \hat{u}(t)) \end{pmatrix}. \tag{18}$$

It is clear that (11) is the second coordinate of (18).

We are now in a position to obtain the following global existence and boundedness theorem.

THEOREM 1. *Let $\hat{A}$ and $\hat{F}$ be defined via* (14) *and* (15) *respectively. If $\hat{x}_0 = [\varphi, \psi] \in D(\hat{A})$, then there exists a unique continuous function $\hat{u}: [0, \infty) \to \hat{X}$ which satisfies*

$$\hat{u}(t) = \hat{T}(t)\hat{x}_0 - \int_0^t \hat{T}(t-s)\hat{F}u(s)\, ds, \qquad \hat{u}(0) = \hat{x}_0. \tag{19}$$

For a.e. $t \in [0, \infty)$, $\hat{u}'(t)$ *exists and satisfies* (17); *consequently, there exists $x(\cdot): [0, \infty) \to X$ which satisfies* (11) *for* a.e. $t \in [0, \infty)$. *Moreover, there exists a constant K which depends on $\beta, \kappa, \|\psi\|$ and $\|A^{1/2}\varphi\|$ so that*

$$\sup_{t \in [0, \infty)} \| \hat{x}(t) \| \leq K \tag{20}$$

and

$$\sup_{t \in [0, \infty)} \| A^{1/2} x(t) \| \leq K. \tag{21}$$

Proof. The local Lipschitz continuity of $\hat{F}$ allows us to use the standard application of the Banach fixed point theorem to obtain local existence and uniqueness of solutions to (19). The usual arguments permit us to continue these solutions to a maximal interval of existence $[0, t_{\max})$. To see that the maximal interval of existence is infinite, we show that the solution of (19) can be continued beyond any finite interval of the form $[0, T)$. The bounds which allow this continuation are the desired global bounds (20) and (21). The function $\hat{u}(\cdot)$ is Lipschitz continuous on any compact subinterval.

Consequently, the reflexivity of $\hat{X}$ implies that $\hat{u}'(t)$ exists for a.e. t. We observe that

$$\frac{\hat{u}(t+h)-\hat{u}(t)}{h}=\frac{\hat{T}(t+h)\hat{x}_0-\hat{T}(t)\hat{x}_0}{h}-\frac{1}{h}\int_t^{t+h}\hat{T}(t+h-s)\hat{F}\hat{u}(s)\,ds$$

$$-\frac{\hat{T}(h)-I}{h}\int_0^t\hat{T}(t-s)\hat{F}u(s)\,ds.$$

It is clear that $\lim_{h\to 0}(\hat{T}(t+h)\hat{x}_0-\hat{T}(t)\hat{x}_0)/h=\hat{A}\hat{T}(t)\hat{x}_0$ and that

$$\lim_{h\to 0}\int_t^{t+h}\hat{T}(t+h-s)\hat{F}u(s)\,ds=\hat{F}\hat{u}(t)$$

for all t. The existence of $\hat{u}'(t)$ and the closedness of A imply that $\int_0^t\hat{T}(t-s)\hat{F}\hat{u}(s)\,ds\in D(\hat{A})$ and that

$$\lim_{h\to 0}\frac{\hat{T}(h)-I}{h}\left(\int_0^t\hat{T}(t-s)\hat{F}u(s)\,ds\right)=\hat{A}\int_0^t\hat{T}(t-s)\hat{F}\hat{u}(s)\,ds,$$

and we thereby obtain the a.e. existence of solutions to (17). Writing out the second component and using the definition of F (9) we have:

$$\ddot{x}(t)+Ax(t)+\left(\beta+\kappa\left\langle A^{1/4}x(t),A^{1/4}x(t)\right\rangle\right)A^{1/2}x(t)=0. \tag{22}$$

We take the inner product of (22) with $\dot{x}(t)$ to obtain

(23)

$$\begin{aligned}&\left\langle\ddot{x}(t),x(t)\right\rangle+\left\langle Ax(t),\dot{x}(t)\right\rangle+\left(\beta+\kappa\left\langle A^{1/4}x(t),A^{1/4}x(t)\right\rangle\right)\left\langle A^{1/2}x(t),\dot{x}(t)\right\rangle\\&\quad=\left\langle\ddot{x}(t),\dot{x}(t)\right\rangle+\left\langle A^{1/2}x(t),A^{1/2}\dot{x}(t)\right\rangle\\&\qquad+\beta\left\langle A^{1/4}x(t),A^{1/4}\dot{x}(t)\right\rangle+\kappa\left\langle A^{1/4}x(t),A^{1/4}x(t)\right\rangle\left\langle A^{1/4}x(t),A^{1/4}\dot{x}(t)\right\rangle\\&\quad=\frac{1}{2}\frac{d}{dt}\|\dot{x}(t)\|^2+\frac{1}{2}\frac{d}{dt}\|A^{1/2}x(t)\|^2+\frac{\beta}{2}\frac{d}{dt}\|A^{1/4}x(t)\|^2+\frac{\kappa}{4}\frac{d}{dt}\|A^{1/4}x(t)\|^4=0.\end{aligned}$$

Integrating (1.22) on $[0,t]$, we obtain

$$\begin{aligned}&\|\dot{x}(t)\|^2+\|A^{1/2}x(t)\|^2+\|A^{1/4}x(t)\|^2+\frac{\kappa}{2}\|A^{1/4}x(t)\|^4\\&\qquad\le\|\dot{x}(0)\|^2+\|A^{1/2}x(0)\|^2+\beta\|A^{1/4}x(0)\|^2+\frac{\kappa}{2}\|A^{1/4}x(0)\|^4.\end{aligned} \tag{24}$$

If $\beta<0$, the quantity $\beta\|A^{1/4}x(t)\|^2+\frac{\kappa}{2}\|A^{1/4}x(t)\|^4$ may be negative. However, we observe that the function $y=\beta x^2+\kappa/2x^4$ is bounded below and that the bound only depends on β and κ. Recalling that $x(0)=\varphi$ and $\dot{x}(0)=\psi$ and observing that $\|A^{1/4}\varphi\|$ can be bounded in terms of $\|A^{1/4}\varphi\|$, we have produced our desired bound for $\|\dot{x}(t)\|$ and $\|A^{1/2}x(t)\|$. We remark that this bound is independent of t. These bounds insure that $\|F\hat{u}(t)\|$ is bounded, and we can use the variation of parameters formula (19) to find $\lim_{t\to 1}\hat{u}(t)$. The local existence theory will carry the solution beyond T.

For notational convenience we specify $\mathcal{F}(\cdot):[0,\infty)\to X$,

$$\mathcal{F}(t)=\beta A^{1/2}x(t)+\kappa\left\langle A^{1/4}x(t),A^{1/4}x(t)\right\rangle A^{1/2}x(t). \tag{25}$$

The following lemma will be used to insure further regularity of $x(\cdot)$.

LEMMA 1. *If $\mathcal{F}(\cdot):[0,\infty)\to X$ is defined via* (25), *$\mathcal{F}(\cdot)$ is continuously differentiable.*

Proof. We observe that the continuity of $\hat{u}(\cdot):[0,\infty)\to\hat{X}$ insures that $\dot{x}(\cdot)$ is continuous. A glance at (18) reveals that $x(t)\in D(A)$ for a.e. $t\in[0,\infty)$. If we can show that $T>0$, $A^{1/2}\dot{x}(\cdot)\in L^\infty(0,T,X)$, we can use the closedness and maximality of $A^{1/2}$

together with the continuity of $\dot{x}(\cdot)$ to insure the continuity of $A^{1/2}\dot{x}(\cdot)$. Having obtained this continuity, the differentiation of $\mathcal{F}(\cdot)$ becomes a computational triviality. Thus our proof rests on establishing an L^∞ bound for $A^{1/2}\dot{x}(\infty)$. We take the L^2 inner product of (22) with $A^{1/2}\dot{x}(t)$ to obtain:

$$\begin{aligned}(26)\quad &\langle \ddot{x}(t), A^{1/2}x(t)\rangle + \langle Ax(t), A^{1/2}\dot{x}(t)\rangle + \beta\langle A^{1/2}x(t), A^{1/2}\dot{x}(t)\rangle \\ &\quad + \kappa\langle A^{1/4}x(t), A^{1/4}x(t), A^{1/4}(t)\rangle\langle A^{1/2}x(t), A^{1/2}\dot{x}(t)\rangle \\ &\quad = \frac{1}{2}\frac{d}{dt}\|A^{1/4}\dot{x}(t)\|^2 + \frac{1}{2}\frac{d}{dt}\|A^{3/4}x(t)\|^2 + \frac{\beta}{2}\frac{d}{dt}\|A^{1/2}x(t)\|^2 \\ &\quad + \frac{\kappa}{2}\langle A^{1/4}x(t), A^{1/4}x(t)\rangle\frac{d}{dt}\|A^{1/2}x(t)\|^2 = 0.\end{aligned}$$

We observe that $\|A^{1/2}x(t)\|^2$ is bounded, and we can produce a bound for $\|A^{3/4}x(t)\|^2$ by integrating (26). Taking the inner product of (22) with $A\dot{x}(t)$, we have

$$\begin{aligned}(27)\quad &\langle \ddot{x}(t), A\dot{x}(t)\rangle + \langle Ax(t), A\dot{x}(t)\rangle + \beta\langle A^{1/2}x(t), A\dot{x}(t)\rangle \\ &\quad + \kappa\langle A^{1/4}x(t), A^{1/4}x(t)\rangle\langle A^{1/2}x(t), A\dot{x}(t)\rangle \\ &\quad = \frac{1}{2}\frac{d}{dt}\|A^{1/2}\dot{x}(t)\|^2 + \frac{1}{2}\frac{d}{dt}\|Ax(t)\|^2 + \frac{\beta}{2}\frac{d}{dt}\|A^{3/4}x(t)\|^2 \\ &\quad + \frac{\kappa}{2}\langle A^{1/4}x(t), A^{1/4}x(t)\rangle\frac{d}{dt}\langle A^{3/4}x(t), A^{3/4}x(t)\rangle = 0.\end{aligned}$$

It is apparent that the boundedness of $\|A^{3/4}x(t)\|^2$ permits us to obtain the desired boundedness of $\|A^{1/2}\dot{x}(t)\|^2$ by integrating (27).

We now introduce the notion of cosine operators. Our treatment will be most cursory. For a thorough discussion of the subject, the interested reader is referred to Webb and Travis [24], [25], [26].

DEFINITION 1. A one-parameter family $\{C(t)|-\infty<t<\infty\}$ of bounded linear operators mapping a Banach space X into itself is called a *strongly continuous cosine family* if and only if:

(i) $C(s+t)+C(s-t)=2C(s)C(t)$ for all $s,t\in(-\infty,\infty)$;

(ii) $C(0)=I$;

(iii) $C(t)x$ is continuous in t for fixed $x\in X$.

The infinitesimal generator of a strongly continuous cosine family $\{C(t)|-\infty<T<\infty\}$ is the operator $A: X\to X$ defined by the equation

$$\begin{aligned}(28)\quad &Ax=\frac{d^2}{dt^2}C(0)x, \\ &D(A)=\{x\in X: C(t)x \text{ is twice continuously differentiable}\}.\end{aligned}$$

In [27] it is shown that the infinitesimal generators may be equivalently characterized as

$$(29)\quad Ax=\lim_{h\to 0}\frac{2(C(h)x-x)}{h^2}, \qquad D(A)=\left\{x\in X: \lim_{h\to 0}\frac{2(C(h)x-x)}{h^2} \text{ exists}\right\}.$$

Associated with every strongly continuous cosine family $\{C(t)|-\infty<t<\infty\}$, we have a strongly continuous sine family $\{S(t)|-\infty<t<\infty\}$, where $S(t)$ is defined by the equation:

$$(30)\quad S(t)x=\int_0^t C(s)x\,ds.$$

Webb and Travis [24] show that if $\mathcal{F}(\cdot):[0,\infty)\to X$ is a continuously differentiable function, then there exists a unique function $w:[0,\infty)\to X$ satisfying

$$w(t)=C(t)x+S(t)y+\int_0^t S(t-s)\mathcal{F}(s)\,ds.$$

Moreover, $w(\cdot)$ is twice continuously differentiable and satisfies the abstract inhomogeneous equation

$$\frac{d^2w(t)}{dt}=Aw(t)+\mathcal{F}(t),\quad w(0)=x,\quad w'(0)=y. \tag{31}$$

The operator A defined by (15) is self-adjoint. Therefore [16] it is known that $-A$ is the infinitesimal generator of a family of cosine operators. The following result is now immediate.

THEOREM 2. *Let $X=L^2[0,1]$ and suppose that $A:X\to X$ is defined via* (15). *The operator $-A$ is the infinitesimal generator of a strongly continuous cosine family $\{C(t)|-\infty<t<\infty\}$; let $\{S(t)|-\infty<t<\infty\}$ be the associated sine family. If $([\varphi,\psi]\in D(A)\times D/A^{1/2})$ and $\mathcal{F}(\cdot):[0,\infty)\to X$ is defined by* (1.24), *then the solution to* (1.10) *has variation of parameters representation*

$$x(t)=C(t)\varphi+S(t)\psi-\int_0^t S(t-s)\mathcal{F}(s)\,ds,\qquad t\geq 0.$$

Moreover, $x(\cdot):[0,\infty)\to X$ is twice continuously differentiable.

Subsequent to the submission of this manuscript, we learned of the existence of [20] by Holmes and Marsden. In the course of an extensive analysis of chaotic oscillations of a forced beam, they obtain global existence and boundedness for (1). We acknowledge the priority of their result and point out the difference in our approach. Holmes and Marsden obtain global existence and boundedness by utilizing energy functionals to extend smooth local semiflows on Banach spaces. We extend our local existence results via computations involving fractional powers of operators. Hopefully this approach will be useful for more complicated nonlinear problems. We also use cosine and sine operators to provide a variation of parameters representation for (1) and to insure additional regularity. To our knowledge this is the first application of cosine operators to the beam equation. We remark that the theory of cosine operators is incomplete. Hopefully, as the subject develops it will prove to be a useful tool for analyzing and computing solutions of beam equations.

REFERENCES

[1] J. BALL, *Initial-boundary value problems for an extensible beam*, J. Math. Anal. Appl., 42 (1973), pp. 61–90.

[2] _____, *On the asymptotic behavior of generalized processes with applications to nonlinear partial differential equations*, J. Diff. Eqs., 27 (1978), pp. 224–265.

[3] _____, *Stability theory for an extensible beam*, J. Diff. Eqs., 14 (1973), pp. 399–418.

[4] G. DA PRATO AND E. GIUSTI, *Una carotterizzazione dei funzioni coseno astrotte.* Boll. Unione Mat. Italiana, 22 (1967), pp. 357–362.

[5] R. DICKEY, *Dynamic stability of equilibrium states of an extensible beam*, Proc. Amer. Math. Soc., 41 (1973), pp. 94–102.

[6] _____, *Free vibrations and dynamic buckling of an extensible beam*, J. Math. Anal. Appl., 29 (1970), pp. 443–454.

[7] W. FITZGIBBON, *Representation and asymptotic behavior of strongly damped evolution equations* (to appear).

[8] _____, *Strongly damped quasilinear evolution equations* (to appear).

[9] H. FRATTORINI, *Ordinary differential equations in linear topological spaces.* I, J. Diff. Eqs., 5 (1968), pp. 72–105.
[10] _____, *Ordinary differential equations in linear topological spaces.* II, J. Diff. Eqs., 6 (1969), pp. 50–70.
[11] _____, *Uniformly bounded cosine functions in Hilbert space*, Indiana Univ. Math. J., 20 (1970), pp. 411–425.
[12] J. GOLDSTEIN, *A perturbation theorm for evolution equations and some applications*, Illinois J. Math. 18 (1974), pp. 196–207.
[13] _____, *On a connection between first order and second order differential equations in Banach spaces*, J. Math. Anal. Appl., 30 (1970), pp. 246–251.
[14] _____, *On the convergence and approximation of cosine functions*, Aequationes Math, 11 (1974), pp. 201–205.
[15] _____, *Semigroups and hyperbolic equations*, J. Functional Anal., 4 (1969), pp. 50–70.
[16] _____, *Semigroups of operators and abstract Cauchy problems*, Lecture Notes, Tulane Univ., 1970.
[17] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Springer Lecture Notes in Mathematics, Springer-Verlag, New York, 1982.
[18] J. EISELY, *Nonlinear vibration of beams and rectangular plates*, Z. Angew. Math. Phys., 15 (1964), pp. 167–175.
[19] P. HOLMES AND J. MARSDEN, *Bifurcations to divergence and flutter in flow induced oscillations: an infinite dimensional analysis*, Automatica, 14 (1978), pp. 419–448.
[20] _____, *A partial differential equation with infinitely many periodic orbits: chaotic oscillations of a forced beam* (to appear).
[21] N. C. HUANG AND W. NACHBAR, *Dynamic snap through imperfect viscoelastic shallow arches*, Trans. ASME. Ser. E. Appl. Mech., 35 (1968), pp. 289–296.
[22] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Lecture Note #10, University of Maryland, College Park, 1974.
[23] C. TRAVIS AND G. WEBB, *An abstract second order semilinear Volterra integrodifferential equation*, this Journal, 10 (1979), pp. 412–424.
[24] _____, *Compactness, regularity and uniform continuity properties of strongly continuous cosine families*, Houston J. Math., 3 (1977), pp. 555–567.
[25] _____, *Cosine families and abstract second order differential equations*, Acta. Math., 32 (1978), pp. 75–96.
[26] _____, *Perturbation of strongly continuous cosine family generators*, Colloquium Math. (to appear).
[27] _____, *Second order differential equations in Banach space*, to appear.
[28] G. WEBB, *Existence and asymptotic behavior for a strongly damped nonlinear wave equation*, to appear.
[29] _____, *A representation formula for strongly continuous cosine families* (to appear).
[30] S. WIONOWSKY-KREIGER, *The effect of an axial force on the vibration of hinged bars*, J. Applied Mech., 17 (1950), pp. 35–36.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0004 $01.00/0

ON THE STRUCTURE OF SOLUTIONS TO $\Delta^2 u=\lambda u$ WHICH SATISFY THE CLAMPED PLATE CONDITIONS ON A RIGHT ANGLE*

CHARLES V. COFFMAN†

Abstract. Let u be a nontrivial solution of $\Delta^2 u=\lambda u$ ($\lambda>0$) on the quarter circle $\{(x,y): 0<x,y,x^2+y^2<1\}$ and suppose that

$$u(x,0)=u_y(x,0)=0, \quad 0<x<1, \qquad u(0,y)=u_x(0,y)=0, \quad 0<y<1.$$

We show then that on any ray through the origin $u(x,y)$ either vanishes identically or oscillates infinitely often as $(x,y)\to(0,0)$.

1. Introduction. In the course of their numerical study of the eigenvalue problem

$$\Delta^2 u=\lambda u \quad \text{in } S, \qquad u=\frac{\partial u}{\partial n}=0 \quad \text{on } \partial S$$

on the unit square S, Bauer and Reiss [1] found that the first eigenfunction possesses nodal lines near the corners. The principal purpose of this note is to provide an analytical verification of the presence of such nodal lines.

We understand that it was A. Weinstein who first questioned whether the fundamental eigenfunction of the clamped plate might possess nodal lines. This issue was again raised by Szegö [11] (see also [5]) in connection with an isoperimetric problem. To the best of our knowledge, the only prior complete analytical demonstration of the existence of nodal lines for the fundamental eigenfunction is that given for the case of an annulus by Duffin and Shaffer [6], [7]. This work showed that, moreover, the fundamental eigenvalue need not be simple; the details can be found also in [3].

Our study will be strictly local in nature; that is to say, we confine our attention to a function $u=u(z)$ satisfying

$$\Delta^2 u=\lambda u \qquad (\lambda\geq 0) \tag{1.1}$$

inside the quarter-circle

$$\left\{z: 0<|z|<1, \ |\arg z|<\frac{\pi}{4}\right\} \tag{1.2}$$

and satisfying for $0<r<1$

$$u(re^{i\theta})=\frac{\partial u}{\partial\theta}(re^{i\theta})=0 \quad \text{when } |\theta|=\frac{\pi}{4}. \tag{1.3}$$

Our main result is the following.

THEOREM 1.1. *Let u be bounded and of class C^4 on the set* (1.2); *assume moreover that the second order partial derivatives of u are square integrable on* (1.2), *that* (1.1) *holds on the interior of* (1.2) *and* (1.3) *holds for* $0<r<1$. *Then for any fixed* θ_0, $|\theta_0|<\frac{\pi}{4}$, $u(re^{i\theta_0})$ *either oscillates as* $r\to 0$ *or vanishes identically.*

This result was proved for biharmonic functions (i.e., for $\lambda=0$) in [2], and we shall make use here of certain of the results from that paper. As in [2] our approach to the problem is based on the study of the transform

$$\hat{u}(p,\theta)=\int_0^1 r^{p-1}u(re^{i\theta})\,dr.$$

* Received by the editors December 15, 1980. This research was supported in part by the National Science Foundation under grant MCS 77-03643.

† Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

In the latter part of the paper, we construct certain special elementary solutions of (1.1), (1.3). We then discuss the expansion of the general solution in terms of these special solutions. This expansion generalizes the expansion for the biharmonic case which is discussed in [2] and which was introduced by Kondrat'ev [9].

There is a considerable literature concerning behavior of solutions to elliptic boundary value problems near an angular or conical point of the boundary. We mention in particular the extensive discussion by Grisvard in [8]. The behavior of eigenfunctions of the Laplacian at an angular point is discussed by Merigot in [10].

2. The transformed equation. Here we will summarize a number of results from [2]. We shall assume that the function u which is under consideration is bounded and of class C^4 on the set

$$Q_\rho=\left\{z:0<|z|<\rho,\ |\arg z|\leq\frac{\pi}{4}\right\}\qquad(\rho>1)$$

and that its partial derivatives of second order are square integrable on Q_ρ. We assume also that u satisfies the boundary conditions

$$u(re^{i\theta})=\frac{\partial u}{\partial\theta}(re^{i\theta})=0,\qquad 0<r<\rho,\quad|\theta|=\frac{\pi}{4}. \tag{2.1}$$

We define the transform

$$\hat{u}(p,\theta)=\int_0^1 r^{p-1}(re^{i\theta})\,dr,\qquad \mathrm{Re}\,p>0. \tag{2.2}$$

The justification of the steps which follow will be deferred to §7.

From the definition (2.2) one has

$$\int_0^1 r^{p+3}\Delta^2 u\,dr=\frac{d^4\hat{u}(p,\theta)}{d\theta^4}+2(p^2+2p+2)\frac{d^2\hat{u}(p,\theta)}{d\theta^2}+(p^4+4p^3+4p^2)\hat{u}(p,\theta)-f(p,\theta),$$

where $f(p,\theta)$ is a polynomial of degree three in p whose coefficients are continuous functions of θ (for the explicit expression see [2]).

If in addition to the conditions indicated above, u satisfies also the differential equation (1.1) on Q_ρ, then $u(p,\theta)$ satisfies

$$\frac{d^4\hat{u}}{d\theta^4}+2(p^2+2p+2)\frac{d^2\hat{u}}{d\theta^2}+(p^4+4p^3+4p^2)\hat{u}=f(p,\theta)+\lambda\hat{u}(p+4,\theta) \tag{2.3}$$

and the boundary conditions

$$\hat{u}(p,\theta)=\frac{\partial\hat{u}}{d\theta}(p,\theta)=0,\qquad|\theta|=\frac{\pi}{4}. \tag{2.4}$$

As was found in [2], the Green's function $K_p(\theta,\varphi)$ that corresponds to the differential expression on the left in (2.3) and the boundary conditions (2.4) is given by

$$\begin{aligned}K_p(\theta,\varphi)=(D(p))^{-1}\Big\{&\Big[w(p,-\varphi)w''\Big(p,\frac{\pi}{4}\Big)-w'(p,-\varphi)w'\Big(p,\frac{\pi}{4}\Big)\Big]w(p,\theta)\\&+\Big[w'(p,-\varphi)w\Big(p,\frac{\pi}{4}\Big)-w(p,-\varphi)w'\Big(p,\frac{\pi}{4}\Big)\Big]w'(p,\theta)\Big\},\\&-\frac{\pi}{4}<\theta<\varphi<\frac{\pi}{4},\end{aligned} \tag{2.5}$$

and

$$K_p(\varphi,\theta)=\overline{K_p(\theta,\varphi)}, \tag{2.6}$$

where for $p\neq 0,-1,-2$

$$w(p,\theta)=\frac{(p+2)\sin p\left(\theta+\frac{\pi}{4}\right)-p\sin\left[(p+2)\left(\theta+\frac{\pi}{4}\right)\right]}{4p(p+1)(p+2)} \tag{2.7}$$

and

$$D(p)=\frac{\left(p+1+\cos p\frac{\pi}{2}\right)}{p+1}\,\frac{\left(p+1-\cos p\frac{\pi}{2}\right)}{p(p+1)(p+2)}. \tag{2.8}$$

3. Analytic continuation of $\hat{u}(p,\theta)$. Throughout the remainder of the paper, whether explicitly stated or not, u will always be assumed to denote a solution of (1.1), (1.2) on Q_ρ ($\rho>1$) satisfying the conditions set down in the preceding section. The transform $\hat{u}(p,\theta)$ defined on the right half-plane by (2.2) admits an extension, also to be denoted by $\hat{u}(p,\theta)$, which is meromorphic in p on the entire plane. The result is as follows.

PROPOSITION 3.1. *Let u be a solution of* (1.1), (1.2) *on Q_ρ and satisfy the conditions set down in* §2. *Let $\hat{u}$ be defined by* (2.2). *For each fixed θ ($|\theta|<\frac{\pi}{4}$), $\hat{u}(p,\theta)$, is analytic in p on the right half-plane and admits a single-valued extension which is meromorphic in p on the entire plane.*

Proof. The analyticity in the right half-plane follows from the boundedness of u and the definition (2.2).

The transform is extended as a solution of the boundary value problem (2.3), (2.4), and for the purpose of making this extension, we use the representation

$$\hat{u}(p,\theta)=\int_{-\pi/4}^{\pi/4}K_p(\theta,\varphi)\left[f_p(p,\varphi)+\lambda\hat{u}(p+4,\varphi)\right]d\varphi. \tag{3.1}$$

Using (3.1), $\hat{u}(p,\theta)$ is extended inductively from the set S_n to the set S_{n+1}, where

$$S_n=\{p:\operatorname{Re}p>-4n\},\qquad n=0,1,2,\cdots.$$

The results of this procedure are made more explicit in the following assertion.

PROPOSITION 3.2. *Let u be as in Proposition* 3.1 *and let $\hat{u}$ denote the extension of its transform* (2.2) *to the entire plane. If for a given θ_0, p_0 is a pole of $\hat{u}(p,\theta_0)$, then* $\operatorname{Re}p_0<0$ *and*

$$D(p_0+4n)=0 \tag{3.2}$$

for some $n=0,1,2,\cdots$. Moreover, if $\hat{u}(p,\theta_0)$ has a pole at p_0 and (3.2) *holds, then $\hat{u}(p,\theta)$ cannot be regular at p_0+4n for all values of θ. Finally, if*

$$\operatorname{Re}p_0<0 \quad \textit{and} \quad D(p_0)=0,$$

then the residue at p_0 of $\hat{u}(p,\theta)$ has the form

$$A_p\left(\frac{\cos p\theta}{\cos p\frac{\pi}{4}}-\frac{\cos(p+2)\varphi}{\cos(p+2)\frac{\pi}{4}}\right) \tag{3.3}$$

or

$$A_p\left(\frac{\sin p\theta}{\sin p\frac{\pi}{4}}-\frac{\sin(p+2)\theta}{\sin(p+2)\frac{\pi}{4}}\right) \tag{3.4}$$

according as p_0 is a zero of the first or the second factor on the right in (2.8).

Proof. We note that in the inductive extension of $\hat{u}$ by means of (3.1) we will have $\hat{u}(p,\theta)$ regular at p_0 unless either p_0 is a root of $D(p)=0$ so that K_p is singular at p_0 or $\hat{u}(p+4,\varphi)$ is singular at $p=p_0$ for at least some values of φ; by induction it is clear that in the latter case p_0 must be a left-translate by a multiple of 4 of a root of $D(p)=0$. Note that none of these roots occur on the imaginary axis so that all poles of $\hat{u}(p,\theta)$ must lie in the open left half-plane.

Finally, suppose that p_0 is a root in the left half-plane of $D(p)=0$. Since all of the zeros of $D(p)$ are simple, it follows from (2.5), (2.6), (2.7) and (3.1) that $(p-p_0)\hat{u}(p,\theta)$ is bounded in the neighborhood of p_0. It readily follows that

$$v(\theta)=\lim_{p\to p_0}(p-p_0)\hat{u}(p,\theta)$$

exists and is a solution of

$$\frac{d^4v}{d\theta^4}+2(p_0^2+2p_0+2)\frac{d^2v}{d\theta^2}+(p_0^4+4p_0^3+4p_0^2)v=0, \tag{3.5}$$

$$v(\theta)=\frac{dv}{d\theta}(\theta)=0, \qquad |\theta|=\frac{\pi}{4}. \tag{3.6}$$

As a solution of (3.5), (3.6) must have one of the forms (3.3) or (3.4) as indicated; the final assertion of Proposition 3.2 is proved.

4. Green's function estimates. These estimates are embodied in the following lemmas.

LEMMA 4.1. *There exists a constant M such that for $|\varphi|, |\theta|<\frac{\pi}{4}$, $p\neq 0$ there holds*

$$|D(p)K_p(\theta,\varphi)|\leq M|p|^{-4}e^{|\operatorname{Im} p|(\pi-|\theta-\varphi|)}. \tag{4.1}$$

Proof. This can be read off directly from (2.5), (2.7) and (2.6).

LEMMA 4.2. *There exist constants c, M such that*

$$|D(p)|^{-1}\leq M|p|^4e^{-\pi|\operatorname{Im} p|}, \tag{4.2}$$

provided

$$|\operatorname{Im} p|>c \tag{4.3}$$

and

$$-|\operatorname{Im} p|\leq \operatorname{Re} p\leq\frac{1}{2}. \tag{4.4}$$

Proof. This estimate follows from (2.8) and the relations

$$\left|\cos p\frac{\pi}{2}\pm(p+1)\right|\geq\left|\cos p\frac{\pi}{2}\right|-|p|-1$$

and

$$\left|\cos p\frac{\pi}{2}\right|^2=\cosh^2\left(\frac{\pi}{2}|\operatorname{Im} p|\right)-\sin^2\left(\frac{\pi}{2}|\operatorname{Re} p|\right).$$

LEMMA 4.3. *There exists a constant M such that* (4.2) *holds provided*

$$\operatorname{Re} p = -(4n-1), \qquad n = 1, 2, \cdots. \tag{4.5}$$

Proof. When (4.5) holds we have

$$\left|\cos p\,\frac{\pi}{2} \pm (p+1)\right|^2 = |1+\operatorname{Re} p|^2 + \left|\sinh\left(\frac{\pi}{2}\operatorname{Im} p\right) \pm \operatorname{Im} p\right|^2;$$

the assertion therefore follows directly from (2.8).

5. Estimates for $\hat{u}(p,\theta)$. We shall now show that $\hat{u}(p,\theta)$ has at most polynomial growth as $p \to \infty$ in a suitably restricted way. It will then follow from Liouville's theorem that $\hat{u}(p,\theta)$ cannot be entire unless it vanishes identically in p.

LEMMA 5.1. *Let S be a set in the complex plane such that*

$$p \in S \quad \textit{implies} \quad p+4 \in S$$

and (4.2) *holds on*

$$S_1 = \left\{ p : p \in S,\ \operatorname{Re} p \le \frac{1}{2} \right\}$$

for some constant M for which (4.1) *also is valid. Let u be as in Proposition* 3.1, *and let λ satisfy*

$$0 < \lambda \pi M^2 < 1. \tag{5.1}$$

Then there exist constants k_1, k_2 depending only on u and M such that for $p \in S$,

$$|\hat{u}(p,\theta)| \le k_1 |p|^3 + k_2, \qquad |\theta| \le \frac{\pi}{4}. \tag{5.2}$$

Proof. First, the structure of the function $f(p,\theta)$, as described in §2, implies the existence of constants C_1, C_2 such that for all p and $|\theta| \le \frac{\pi}{4}$,

$$|f(p,\theta)| \le C_1 |p|^3 + C_2. \tag{5.3}$$

Secondly, since u is bounded on Q_ρ, we have

$$|\hat{u}(p,\theta)| \le C_3, \qquad \operatorname{Re} p > \frac{1}{2}, \quad |\theta| < \frac{\pi}{4} \tag{5.4}$$

for some constant C_3. In view of the hypothesis concerning the set S_1, we can combine (4.1) and (4.2) to obtain

$$|K_p(\theta,\varphi)| \le M^2, \qquad p \in S_1, \quad |\theta|, |\varphi| \le \frac{\pi}{4}. \tag{5.5}$$

From the representation (3.1) and (5.3), one, therefore, has

$$|\hat{u}(p,\theta)| \le \frac{\pi}{2} M^2 \left(C_1 |p|^3 + C_2\right) + \frac{\lambda\pi}{2} M^2 \sup_{|\varphi| \le \pi/4} |\hat{u}(p+4,\varphi)| \tag{5.6}$$

for $p \in S_1$.

Suppose now that $p \in S$ and $\frac{1}{2} > \operatorname{Re} p > -\frac{7}{2}$ so that $p \in S_1$ and $\operatorname{Re}(p+4) \ge \frac{1}{2}$; then (5.1), (5.6) and (5.4) imply

$$|\hat{u}(p,\theta)| \le \frac{\pi}{2} M^2 \left(C_1 |p|^3 + C_2\right) + C_3, \qquad \operatorname{Re} p > -\frac{7}{2}, \quad |\theta| \le \frac{\pi}{4}. \tag{5.7}$$

The proof will now be completed by induction as follows. Take

$$k_1 = \pi M^2 C_1, \qquad k_2 = \pi M^2 C_2 + C_3 \tag{5.8}$$

and note that with this choice of k_1, k_2, (5.7) implies (5.2) for $\mathrm{Re}\, p \geq -\frac{7}{2}$. Now let

$$\gamma \leq -\frac{7}{2},$$

and suppose that (5.2) holds for

$$p \in S, \qquad \mathrm{Re}\, p \geq \gamma. \tag{5.9}$$

For

$$p \in S, \qquad \gamma - 4 \leq \mathrm{Re}\, p \leq \gamma, \tag{5.10}$$

one then has from (5.6) and the assumed validity of (5.2) when (5.9) holds that

$$|\hat{u}(p,\theta)| \leq \frac{\pi}{2} M^2 \left(C_1 |p|^3 + C_2\right) + \frac{\lambda \pi}{2} M^2 \left(k_1 |p|^3 + k_2\right)$$

or taking into account (5.1)

$$|\hat{u}(p,\theta)| \leq \left(\frac{\pi}{2} M^2 C_1 + \frac{1}{2} k_1\right) |p|^3 + \frac{\pi}{2} M^2 C_2 + \frac{1}{2} k_2.$$

It follows from this last inequality and (5.8) that (5.2) extends to the set (5.10), and thus, the validity of (5.2) for all $p \in S$ follows by induction.

PROPOSITION 5.1. *Suppose that for a given θ_0, $\hat{u}(p,\theta_0)$ is entire. Then $u(re^{i\theta_0})$ vanishes identically in r.*

Proof. We define the set S as follows: $p \in S$ if $\mathrm{Re}\, p > \frac{1}{2}$ or if p satisfies either (4.3) and (4.4) or (4.5). It is clear that translation by 4 carries S into itself, and it follows from Lemmas 4.2 and 4.3 that with an appropriate choice of M the inequality (4.2) will hold for $p \in S$, $\mathrm{Re}\, p \leq \frac{1}{2}$. It follows from Lemma 5.1 that (5.2) holds on this set S provided λ satisfies (5.1). Since S contains squares with center at zero and with arbitrarily long sides, it follows from the Liouville estimates that $\hat{u}(p,\theta_0)$ must be a polynomial if it is entire. However, the transform (2.2) of a bounded function cannot be a polynomial other than the zero polynomial. This completes the proof in the case where (5.1) holds.

For the general case, i.e., λ positive but not satisfying (5.1), let $\varepsilon > 0$ be chosen so that

$$\varepsilon^4 \lambda \pi M^2 < 1.$$

We then write

$$\begin{aligned} \hat{u}(p,\theta) &= \int_0^{\varepsilon} r^{p-1} u(re^{i\theta})\, dr + \int_{\varepsilon}^{1} r^{p-1} u(re^{i\theta})\, dr \\ &= \varepsilon^p \int_0^1 r^{p-1} u(\varepsilon r e^{i\theta})\, dr + \int_{\varepsilon}^{1} r^{p-1} u(re^{i\theta})\, dr. \end{aligned}$$

Since the second integral on the right is an entire function of p for all θ, the function $u(\varepsilon r e^{i\theta_0})$ will have an entire transform if $u(re^{i\theta_0})$ does. The former function satisfies (1.1) with λ replaced by $\varepsilon^4 \lambda$, and thus, the preceding argument implies that

$$\int_0^1 r^{p-1} u(\varepsilon r e^{i\theta_0})\, dr \equiv 0$$

if $\hat{u}(p,\theta_0)$ is entire. By the uniqueness of the transform, $u(r,\theta_0)$ must therefore vanish identically for $0 < r < \varepsilon$, hence, by analyticity, for $0 < r < 1$. This completes the proof.

COROLLARY 5.1. *If for a given θ_0, $u(re^{i\theta_0})$ vanishes of infinite order as $r \to 0$, then $u(re^{i\theta_0})$ vanishes identically for $0 < r < 1$.*

COROLLARY 5.2. *Let the residues of* $\hat{u}(p,\theta)$ *at the roots in the left half-plane of* $D(p)=0$ *be given by* (3.3) *or* (3.4). *Then* $u(re^{i\theta})$ *is completely determined by the set of numbers*

$$\{A_p : D(p)=0,\ \operatorname{Re} p<0\}.$$

Proof. Corollary 5.1 is immediate since the vanishing of infinite order of $u(re^{i\theta_0})$ implies that the transform (2.2) is entire for $\theta=\theta_0$.

By Proposition 3.2, if all of the constants A_p are zero, then the transform must be entire; thus, Corollary 5.2 follows.

6. Oscillation of $u(re^{i\theta})$. Theorem 1.1 follows from Proposition 5.1 via a theorem of Doetsch, [4, p. 59], (see also Widder [12, p. 58]). Suppose that, for a given θ_0, $u(re^{i\theta_0})$ is nonnegative for small r. By Doetsch's theorem then, the integral

$$\int_0^1 r^{p-1}u(re^{i\theta_0})\,dr=\int_0^\infty e^{-pt}u(e^{-pt+i\theta_0})\,dt$$

if it fails to converge for all p must have a real pole on its abscissa of convergence. By inspection we see that $D(p)$, given by (2.8), has no real zeros; thus in view of Proposition 3.2 we conclude that the integral can have no real poles and hence must converge for all p. Finally, if the integral converges for all real p, then $\hat{u}(p,\theta_0)$ is entire and hence, by Proposition 5.1, $u(re^{i\theta_0})$ must vanish identically.

7. Justification of the transform method. The details of this justification are very similar to those for the biharmonic case as treated in [2]. Accordingly, we consider here only those aspects in which the present case differs from that treated in [2].

LEMMA 7.1. *There exist constants* $K_1, K_2, \cdots$, *such that if u satisfies* (1.1) *in a region* Ω *and*

$$|u(z)|\le M,\qquad z\in\Omega,$$

then for a partial differential operator $\partial_m=\partial^m/\partial x^k\partial y^{m-k}$ *of order m and* $z\in\Omega$,

$$|\partial_m u(z)|\le K_m M d_z^{-m},$$

where $d_z=\operatorname{dist}(z,\partial\Omega)$.

Proof. We prove the inequality for the case $m=1$, $\lambda=1$. Clearly there is no loss of generality in taking $\lambda=1$, (elementary consideration show that the constants K_n are independent of λ), and the general step in the inductive proof of the lemma is the same as that in [2].

Suppose that Ω is a disk of radius R centered at zero. The function u admits an expansion of the form

$$u(re^{i\theta})=\sum_{n=0}^{\infty}\{J_n(r)(A_n\cos n\theta+B_n\sin n\theta)+I_n(r)(C_n\cos n\theta+D_n\sin n\theta)\},$$

and

$$\frac{\partial u(0)}{\partial x}=\frac{1}{2}(A_1+C_1). \tag{7.1}$$

Since the terms $J_1(r)\cos\theta$ and $I_1(r)\cos\theta$ are orthogonal over the disk to the remaining terms in the series, we have

$$\pi\left(A_1^2\int_0^R J_1^2(r)r\,dr+2A_1C_1\int_0^R J_1(r)I_1(r)r\,dr+C_1^2\int_0^R I_1^2(r)r\,dr\right) \leq\int_0^{2\pi}\int_0^R|u(re^{i\theta})|^2 r\,dr\,d\theta\leq\pi R^2M^2. \tag{7.2}$$

Since $\int_0^R I_1^2(r)r\,dr\geq\int_0^R J_1(r)I(r)r\,dr$, $\int_0^R J_1^2(r)r\,dr$ and $R^{-4}\int_0^R I_1^2(r)r\,dr\to\frac{1}{16}$ as $R\to 0$, it follows from (7.2) that

$$(A_1+C_1)^2\leq K_1^2M^2R^{-2}$$

for some constant K_1. The desired inequality now follows from (7.1).

It follows from Lemma 7.1 that a partial derivative of arbitrary order of $u(re^{i\theta})$ grows no more rapidly than some negative power of r as $r\to 0$ and θ remains fixed. This suffices for the justification of the steps leading to (2.3).

Since our hypothesis in §2 includes square integrability of the second-order partial derivatives of u, it follows that the integral

$$\int_0^1 r^2\left|\frac{\partial u}{\partial\theta}\right|^2 dr$$

is bounded independently of θ on $|\theta|\leq\frac{\pi}{4}$. One can readily conclude from (2.1) that

$$\frac{\partial\hat{u}}{\partial\theta}(p,\theta)\to 0\quad\text{when }\theta\to\pm\frac{\pi}{4}$$

for $\mathrm{Re}\,p$ sufficiently large; the other boundary condition presents no problem.

8. Special solutions. In this section we shall construct certain special solutions of the problem (1.1), (1.2). These special solutions can be characterized by the following property: at all but one of the roots in the left half-plane of $D(p)=0$, the transform (2.2) of the solution u of (1.1), (1.2) is regular for all $\theta\in[-\frac{\pi}{4},\frac{\pi}{4}]$. By Corollary 5.2 of Proposition 5.1, such a solution is uniquely determined to within multiplication by a constant. Here we shall actually only consider those special solutions of this type whose transforms have a pole at a zero of the first factor on the right in (2.8). These are the solutions which are symmetric in θ. The solutions which are antisymmetric in θ are related in the same way to the zeros of the second factor on the right in (2.8); their construction is similar.

We put

$$h_p(r)=\sum_{k=0}^{\infty}\frac{1}{(2k)!\Gamma(p+2k+3)}\left(\frac{r}{2}\right)^{p+4k+2}=\frac{1}{2}\left(I_{p+2}(r)+J_{p+2}(r)\right) \tag{8.1}$$

and

$$g_p(r)=\sum_{k=0}^{\infty}\frac{1}{(2k+1)!\Gamma(p+2k+2)}\left(\frac{r}{2}\right)^{p+4k+2}=\frac{1}{2}\left(I_p(r)-J_p(r)\right). \tag{8.2}$$

Note that

$$h_p(r)\cos(p+2)\theta,\qquad g_p(r)\cos p\theta$$

are solutions of the partial differential equation (1.1) for $\lambda=1$.

We seek a function of the form

$$\Theta_p(r,\theta)=\sum_{n=0}^{\infty}\{A_n h_{p+4n}(r)\cos(p+4n+2)\theta+B_n g_{p+4n}(r)\cos(p+4n)\theta\} \tag{8.3}$$

with $\operatorname{Re} p>0$, which satisfies

$$\Theta_p(r,\theta)=\frac{\partial}{\partial\theta}\Theta_p(r,\theta)=0, \qquad |\theta|=\frac{\pi}{4}.$$

If we put $\theta=\frac{\pi}{4}$ in the expression on the right in (8.3) and then use (8.1) and (8.2), we get a power series in r whose coefficients must vanish if Θ_p is to vanish for $|\theta|=\frac{\pi}{4}$. We proceed similarly with the expression for $\frac{\partial}{\partial\theta}\Theta_p(r,\theta)$ that results from (8.3). We are thus led to the following system of equations:

$$\frac{A_n \sin p\frac{\pi}{4}}{\Gamma(p+4n+3)}-\frac{B_n\cos p\frac{\pi}{4}}{\Gamma(p+4n+2)}=\sin p\frac{\pi}{4}\Sigma_{1,n}+\cos p\frac{\pi}{4}\Sigma_{2,n}, \tag{8.4}$$

$$\frac{A_n(p+4n+2)\cos p\frac{\pi}{4}}{\Gamma(p+4n+3)}+\frac{B_n(p+4n)\sin p\frac{\pi}{4}}{\Gamma(p+4n+2)}=\cos p\frac{\pi}{4}\Sigma_{3,n}+\sin p\frac{\pi}{4}\Sigma_{4,n}, \tag{8.5}$$

where $\Sigma_{i,0}=0$, $i=1,2,3,4$, and for $n\geq 1$,

$$\Sigma_{1,n}=\sum_{k=1}^{n}\frac{(-1)^{k+1}A_{n-k}}{(2k)!\Gamma(p+4n-2k+3)}, \tag{8.6}$$

$$\Sigma_{2,n}=\sum_{k=1}^{n}\frac{(-1)^{k}B_{n-k}}{(2k+1)!\Gamma(p+4n-2k+2)}, \tag{8.7}$$

$$\Sigma_{3,n}=\sum_{k=1}^{n}\frac{(-1)^{k+1}(p+4n-4k+2)A_{n-k}}{(2k)!\Gamma(p+4n-2k+3)}, \tag{8.8}$$

$$\Sigma_{4,n}=\sum_{k=1}^{n}\frac{(-1)^{k+1}(p+4n-4k)B_{n-k}}{(2k+1)!\Gamma(p+4n-2k+2)}. \tag{8.9}$$

LEMMA 8.1. *Let* $\operatorname{Re} p>0$. *The infinite system of equations* (8.4), (8.5), $n=0,1,2,\cdots$, *admits a solution with* A_0, $B_0\neq 0$ *if and only if*

$$p+1+\cos p\frac{\pi}{2}=0. \tag{8.10}$$

Proof. The determinant of the coefficients of A_n and B_n on the left in (8.4), (8.5) is

$$\frac{p+4n+1+\cos p\frac{\pi}{2}}{\Gamma(p+4n+2)\Gamma(p+4n+3)}. \tag{8.11}$$

In order that (8.4), (8.5) admit a nontrivial solution when $n=0$, it, obviously, is necessary that (8.10) hold. Conversely, if (8.10) holds and (A_0,B_0) is a nontrivial solution of (8.4), (8.5), $n=0$, then since (8.11) does not vanish for $n\geq 1$, the infinite system can be solved recursively.

LEMMA 8.2. *Let* $\operatorname{Re} p>0$, *and let* (8.10) *hold. If* $\{A_n,B_n\}$, $n=0,1,\cdots$, *is a nontrivial solution of* (8.4), (8.5), *then there exists* $\mu\geq 1$ *such that*

$$\frac{n!|A_n|}{\mu^n|\Gamma(p+4n+3)|}, \frac{n!|B_n|}{\mu^n\Gamma(p+4n+2)|}=O(1) \tag{8.12}$$

as $n\to\infty$.

Proof. Suppose that the expressions on the left in (8.12) are bounded by M for all values of the index less than n. We first estimate $\Sigma_{1,n}$, defined in (8.6), as follows

$$
\begin{aligned}
|\Sigma_{1,n}| &\le M \sum_{k=1}^{n} \frac{\mu^{n-k}}{(2k)!(n-k)!|p+4n-2k+2|\cdots|p+4n-4k+3|} \\
&\le \frac{M\mu^{n-1}}{(n-1)!|p+2n+2|\,|p+3|} \sum_{k=1}^{n} \frac{(k-1)!}{(2k)!}\binom{n-1}{k-1}|p+n+3|^{-k+1} \\
&\le \frac{M\mu^{n-1}}{2(n-1)!|p+2n+2|\,|p+3|}\left(1+\frac{1}{|p+n+3|}\right)^{n-1} \\
&\le \frac{M\mu^{n-1}e}{2(n-1)!|p+2n+2|\,|p+3|}.
\end{aligned} \tag{8.13}
$$

Similarly,

$$
|\Sigma_{2,n}| \le \frac{M\mu^{n-1}e}{2(n-1)!|p+2n+1|\,|p+2|}, \tag{8.14}
$$

$$
|\Sigma_{3,n}| \le \frac{M\mu^{n-1}e}{2(n-1)!|p+3|}, \tag{8.15}
$$

$$
|\Sigma_{4,n}| \le \frac{M\mu^{n-1}e}{2(n-1)!|p+2|}. \tag{8.16}
$$

It follows from (8.10) that

$$
\left|\sin p\frac{\pi}{4}\right|, \left|\cos p\frac{\pi}{4}\right| \le \left(\frac{|p+2|}{2}\right)^{1/2}. \tag{8.17}
$$

We recall that the determinant of the coefficient matrix in (8.4), (8.5) is given by (8.11). If we now solve (8.4), (8.5) for A_n taking into account (8.10), and use the estimates (8.13)–(8.17), we get

$$
\begin{aligned}
|A_n| &\le \frac{|\Gamma(p+4n+3)|}{4n}\cdot\frac{M\mu^{n-1}e}{(n-1)!}\left\{\frac{|p+4n|}{2|p+2n+1|}+\frac{1}{2}\right\} \\
&\le \frac{M\mu^{n-1}e}{2n!}|\Gamma(p+4n+3)| \le \frac{M\mu^{n}}{n!}|\Gamma(p+4n+3)|,
\end{aligned}
$$

provided $\mu \ge \frac{1}{2}e$. The inductive estimate for B_n is similar.

THEOREM 8.1. *Let* $\mathrm{Re}\,p>0$, *and let* (8.10) *hold. Then there exists a nontrivial function* $\Theta_p(r,\theta)$ *of the form* (8.3) *defined for* $r>0$, $|\theta| \le \frac{\pi}{4}$ *and satisfying*

$$
\Delta^2\Theta_p = \Theta_p,
$$

and

$$
\Theta_p(r,\theta) = \frac{\partial}{\partial\theta}\Theta_p(r,\theta) = 0, \qquad r>0, \quad |\theta| = \frac{\pi}{4}.
$$

The series on the right in (8.3) *converges uniformly on* $\{(r,\theta): |r| \le R, |\theta| \le \frac{\pi}{4}\}$ *for any real* R. *For any* $\lambda>0$, *the transform*

$$
\int_0^1 r^{q-1}\Theta_p(\lambda r,\theta)\,d\theta
$$

has poles at $-p-2, -p-6, -p-10, \cdots$, *and not elsewhere. Finally,*

$$r^{-\mathrm{Re}\,p-2}\Theta_p(r,\theta)=O(1)$$

as $r\to 0$ *uniformly in* θ.

Proof. It is clear from inspection of (8.1) and (8.2) that

$$|\Gamma(p+4n+3)|r^{-\mathrm{Re}\,p-4n-2}|h_{p+4n}(r)|$$

and

$$|\Gamma(p+4n+2)|r^{-\mathrm{Re}\,p-4n-2}|g_{p+4n}(r)|$$

can be bounded independently of p and n on any interval of the form $(0,R)$. For a fixed p, the trigonometric terms which appear in (8.3) are bounded on $|\theta|\le\frac{\pi}{4}$ independently of n. In view of these uniformities, it follows from Lemma 8.2 that if $\{A_n, B_n\}$ satisfy (8.4), (8.5), then the series (8.3) converges uniformly for $|\theta|\le\frac{\pi}{4}$, $r<R$ for any R.

The remaining assertions of the theorem are proved in a straightforward way.

9. Expansion theorem. We begin by normalizing the functions that were constructed in the previous section. Let p satisfy (8.10) with

$$\mathrm{Re}\,p>0,$$

and take for a solution of (8.4), (8.5), $n=0$,

$$A_0=A_0(p)=-\frac{2^{p+2}\Gamma(p+3)}{\cos(p+2)\frac{\pi}{4}},\qquad B_0=B_0(p)=\frac{2^{p+2}\Gamma(p+2)}{\cos p\frac{\pi}{4}}.$$

Henceforth, $\Theta_p(r,\theta)$ will denote the function (8.3) that results from this choice of A_0, B_0. We then will have

$$\Theta_p(r,\theta)=r^{p+2}\left(\frac{\cos p\theta}{\cos p\frac{\pi}{4}}-\frac{\cos(p+2)\theta}{\cos(p+2)\frac{\pi}{4}}\right)+O(r^{\mathrm{Re}\,p+6})$$

as $r\to 0$ and the residue at $-(p-2)$ of the transform of $\Theta_p(r,\theta)$,

$$\int_0^1 r^{q-1}\Theta_p(r,\theta)\,dr,$$

will be

$$\left(\frac{\cos p\theta}{\cos p\frac{\pi}{2}}-\frac{\cos(p+2)\theta}{\cos(p+2)\frac{\pi}{4}}\right).$$

With the indicated choice of $A_0(p)$, $B_0(p)$, the constant M in the estimates (8.13)–(8.16) can be taken to be

$$2^{\mathrm{Re}\,p+3/2}|p|^{-1/2}$$

(since $|\cos p\,\frac{\pi}{4}|, |\cos(p+2)\frac{\pi}{4}|\ge 2^{-1/2}|p|^{1/2}$), giving

$$\frac{n!|A_n(p)|}{\mu^n|\Gamma(p+4n+3)|},\ \frac{n!|B_n(p)|}{\mu^n\Gamma(p+4n+2)}\le\frac{2^{\mathrm{Re}\,p+3/2}}{|p|^{1/2}}.$$

Next we note that

$$|p|^{-1/2}|\cos(p+4n+2)\theta|,\qquad |p|^{-1/2}|\cos(p+4n)\theta|$$

are bounded on $|\theta|\leq\frac{\pi}{4}$, independently of n and p so long as (8.12) holds. Taking into account the first assertion in the proof of Lemma 8.1, we have therefore

$$|\Theta_p(r,\theta)|\leq 2^{\operatorname{Re}p+5/2}r^{\operatorname{Re}p+2}K\sum_{n=0}^{\infty}\frac{\mu^n}{n!}r^{4n}\leq\sqrt{2}\,K\left(\frac{r}{2}\right)^{\operatorname{Re}p+2}e^{\mu r^4}, \tag{9.1}$$

where K is independent of p.

Suppose now that u satisfies the assumptions set down in §2 and that in addition u is an even function of θ. It is clear then from Proposition 3.2 that the poles of the transform $\hat{u}$ do not occur at the roots in the left-plane of the second factor of (2.8). We can estimate the residues of $\hat{u}$ at the zeros of the first factor of (2.8) by the methods of §5. Indeed, if p is a root in the right half-plane of (8.10) and the residue at $-(p-2)$ of $\hat{u}$ is

$$a_p\left(\frac{\cos p\theta}{\cos p\frac{\pi}{4}}-\frac{\cos(p+2)\theta}{\cos(p+2)\frac{\pi}{4}}\right),$$

then for λ sufficiently small, we will have the polynomial estimate

$$|a_p|\leq C|p|^{\kappa},$$

where C of course depends on u. This estimate, together with the uniformities for the functions Θ_p which were established above, enable us to prove the following.

THEOREM 9.1. *Let u be as in §2. Assume also that u is even in θ. Then the series*

$$\sum_p a_p\lambda^{-p/4-1/2}\Theta_p(\lambda^{1/4}r,\theta),$$

where the summation is over the roots in the right half-plane of (8.10), *converges uniformly to u on $|\theta|<\frac{\pi}{4}$, $0<r<\varepsilon$, for some $\varepsilon>0$.*

REFERENCES

[1] L. BAUER AND E. REISS, *Block five diagonal metrics and the fast numerical computation of the biharmonic equation*, Math. Comp., 26 (1972), pp. 311–326.

[2] C. V. COFFMAN AND R. J. DUFFIN, *On the structure of biharmonic functions satisfying the clamped plate conditions on a right angle*, Adv. in Appl. Math., 1 (1980), pp. 373–389.

[3] C. V. COFFMAN, R. J. DUFFIN AND D. H. SHAFFER, *The fundamental mode of vibration of a clamped annular plate is not of one sign*, Constructive Approaches to Mathematical Models, Academic Press, New York, 1979.

[4] G. DOETSCH, *Laplace-Transformation*, Dover, New York, 1943.

[5] R. J. DUFFIN, *Some problems of mathematics and science*, Bull. Amer. Math. Soc., 80 (1974), pp. 1053–1070.

[6] R. J. DUFFIN AND D. H. SHAFFER, *On the modes of vibration of a ring-shaped plate*, Bull. Amer. Math. Soc., 58 (1952), pp. 652.

[7] ———, *On the vibration of a ring-shaped plate*, Dept. of the Air Force Tech. Rep. AF8–TRIG, Carnegie Inst. of Tech., Pittsburgh, PA, 1952.

[8] P. GRISVARD, *Behavior of the solution of an elliptic boundary value problem in a polygonal or polyhedral domain*, Numerical Solution of Partial Differential Equations III, Proc. of Third Symposium (SYNSPADE) Univ. of Maryland, 1975, Academic Press, New York, 1976.

[9] V. A. KONDRAT'EV, *Boundary problems for elliptic equations with conical or angular points*, Trans. Moscow Math. Soc., 16 (1967), pp. 227–313.

[10] M. MERIGOT, *Régularité des functions propres du laplacian dans un cône*, C. R. Acad. Sci. Paris, Ser. A, 279 (1974), pp. 503–505.

[11] G. SZEGÖ, *On membranes and plates*, Proc. Nat. Acad. Sci. U.S.A., 36 (1950), pp. 210–216.

[12] D. V. WIDDER, *The Laplace Transform*, Princeton Univ. Press, Princeton, NJ, 1946.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0005 $01.00/0

SOLUTIONS FOR A FLUX-DEPENDENT DIFFUSION MODEL*

JONATHAN BELL†, CHRIS COSNER‡ AND WILLY BERTIGER§

Abstract. We study a one-dimensional continuous analogue of a system proposed by Mitchison to model vein formation in meristematic tissues of young leaves. The signal concentration satisfies a diffusion equation where the diffusion coefficient changes according to a differential equation which is flux dependent. We show that the system possesses a unique, global solution. We then examine the stability of the steady state solution which depends on a source strength parameter $\psi>0$. For ψ sufficiently small, the steady state is linearly and L^2 stable. But as ψ passes through a critical value, the stability changes and a Hopf bifurcation takes place.

1. Introduction. There is strong evidence suggesting that a flow pattern underlies vein development in plants. Mitchison [5] has derived a model from experimental evidence that a signal flows from a source in such a way that the capacity of a given pathway to transport this signal increases with the flux it carries. His model is spacially discrete, and it generates a well-defined pattern of strands. Thus, this flux-dependent facilitation seems to be a satisfactory hypothesis for many features of vein development.

Since the full model is very complicated to analyze, Mitchison proposed in an appendix to [5] to consider a continuous version of his model. Neglecting polar transport which he incorporates in the spacially discrete model for certain numerical experiments, the continuous model takes the form

$$\frac{\partial s}{\partial t}=\frac{\partial}{\partial x_1}\left(D_1\frac{\partial s}{\partial x_1}\right)+\frac{\partial}{\partial x_2}\left(D_2\frac{\partial s}{\partial x_2}\right),\qquad \frac{\partial D_i}{\partial t}=f\left(D_i,D_i\frac{\partial s}{\partial x_i}\right),\qquad i=1,2.$$

Here $s(t,x_1,x_2)$ is the signal concentration, and D_1,D_2 are diffusion coefficients for flux parallel to x_1 and x_2 axes, respectively. He then chose a specific form for f and showed that small disturbances from rest could grow.

An appealing aspect of the flux-dependent diffusion is that such a mechanism may lead to a very rich variety of pattern formation of importance in other contexts which can not be modelled effectively by classical reaction-diffusion systems. This new diffusion mechanism can be viewed as an alternative to the chemotaxis mechanism.

In this paper we analyze the spacially one-dimensional version of the above model which is the simplest continuous model which possesses the flux-dependent diffusive behavior that Mitchison has proposed. Hence, the system we consider is

$$\frac{\partial s}{\partial t}=\frac{\partial}{\partial x}\left\{d\frac{\partial s}{\partial x}\right\},\qquad 0<x<1,\quad t>0,$$

$$\frac{\partial d}{\partial t}=-d+g\left(d\frac{\partial s}{\partial x}\right),$$

$$s(x,0)\equiv 0,\quad d\frac{\partial s}{\partial x}\bigg|_{x=1}=\psi,\qquad s(x,0)=s_0(x),\quad d(x,0)=d_0(x)$$

* Received by the editors November 18, 1980, and in revised form May 28, 1981.

† Department of Mathematics, Texas A & M University, College Station, Texas 77843. Present address: State University of New York, Buffalo, New York 14260.

‡ Department of Mathematics, Texas A & M University, College Station, Texas 77843.

§ Chevron Oil Field Research Company, P.O. Box 446, La Habra, California 90631.

where g, s_0 and d_0 are sufficiently smooth and ψ is a positive constant. Thus the left boundary is a sink and the right boundary is a source of constant flux.

In the next section we specify conditions on g and show that the system has a solution and that it is unique. We obtain the solution by changing variables, constructing an iteration scheme and bounding the iterates and their derivatives to obtain convergence. If the initial data are positive, then the iterates are also and hence so is the solution. The approach used to obtain existence and uniqueness is similar to that used by Fateeva [1].

In the last section we study the steady state solution. In particular, there can be a $\psi_c > 0$ such that for $0 < \psi < \psi_c$ the steady state is linearly stable, while for $\psi > \psi_c$ the steady state is unstable. More complicated situations than this can also occur. We give a condition which guarantees that the steady state solution is asymptotically stable in the L^2-norm for $\psi > 0$ small. We then show that as ψ passes through ψ_c we obtain a Hopf-like bifurcation. That is, for $0 < \psi - \psi_c$ small, there bifurcates from the steady state a spacially dependent solution with a time oscillatory mode.

Since we consider only the one-dimensional model in this paper, our problem does not directly relate to the physical problem Mitchison addresses, but it does give indications of what to expect analytically from a more complete model. In particular, one can obtain oscillatory concentration changes from the diffusive mechanism alone. For the reason stated, this work has to be considered preliminary to the more physical model. The two-dimensional system seems to require methods of analysis different from those used here.

2. Solutions to the problem. The object of this section is to show the existence and uniqueness of solutions to the system

$$\begin{aligned} &s_t = (ds_x)_x, && (x,t) \in (0,1) \times (0,T], \\ &d_t = -d + g(ds_x), && (x,t) \in [0,1] \times (0,T] \end{aligned} \tag{2.1}$$

subject to the boundary and initial conditions

$$\begin{aligned} &s(0,t) = 0, \quad d(1,t)s_x(1,t) = \psi, \quad t \in [0,T], \\ &s(x,0) = s_0(x), \quad d(x,0) = d_0(x), \quad x \in [0,1]. \end{aligned} \tag{2.2}$$

We will assume that $s_0(0) = 0$ and $d_0(1)s_{0x}(1) = \psi$, that $d_0(x) \geq \delta_0 > 0$ and that d_0 and s_0 are smooth. We will also assume that the function g in (2.1) is smooth with all derivatives bounded on $\mathbb{R}$ and

$$0 < g_0 \leq g(\xi) \leq G_0 \tag{2.3}$$

for some constants g_0 and G_0.

To solve (2.1) we must introduce a new variable. Consider the system

$$\begin{aligned} &w_t = dw_{xx} + \left[\frac{g(w)}{d} - 1 \right] w \quad \text{on } (0,1) \times (0,T], \\ &d_t = -d + g(w) \quad \text{on } [0,1] \times (0,T] \end{aligned} \tag{2.4}$$

with

$$\begin{aligned} &w_x(0,t) = 0, \quad w(1,t) = \psi, \\ &w(x,0) = w_0(x) \equiv d_0(x)s_{0x}(x), \quad d(x,0) = d_0(x). \end{aligned} \tag{2.5}$$

We will assume that d_0 and s_0 are such that $w_{0x}(0) = 0$, $w_0(1) = \psi$, $w_{0xx}(1) = 0$ and that the additional compatibility conditions obtained by applying $\partial/\partial x$, $\partial^2/\partial x^2$ and $\partial/\partial t$ to

the first equation in (2.4) and substituting in the initial and boundary data at $t=0$, $x=0, 1$ are satisfied.

LEMMA 1. *Suppose that $w(x,t)$ and $d(x,t)$ are classical solutions of* (2.4), (2.5) *with $d(x,t)\geq k_0>0$. Let*

$$s(x,t)=\int_0^x \frac{w(\xi,t)}{d(\xi,t)}\,d\xi. \tag{2.6}$$

Then $s(x,t)$ and $d(x,t)$ satisfy (2.1), (2.2).

Remark. If s and d satisfy (2.1) and (2.2) and the compatibility conditions following (2.5) hold, then $w=ds_x$ also satisfies (2.4) and (2.5).

Proof. It is clear that $s(x,t)$ and $d(x,t)$ satisfy the initial and boundary conditions (2.2). Also, since (2.6) implies that $ds_x=w$, it follows that d satisfies the second equation in (2.1). Differentiating (2.6) with respect to t yields

$$s_t=\int_0^x \left(\frac{w(\xi,t)}{d(\xi,t)}\right)_t d\xi. \tag{2.7}$$

Solving for w_{xx} in (2.4) yields

$$w_{xx}=\frac{1}{d}w_t+\frac{1}{d}\left[\frac{-g(w)+d}{d}\right]w=\frac{1}{d}w_t-\frac{w}{d^2}d_t=\left(\frac{w}{d}\right)_t.$$

Integrating with respect to x and using the fact that $w_x(0,t)=0$, we have

$$w_x=\int_0^x w_{\xi\xi}(\xi,t)\,d\xi=\int_0^x \left(\frac{w(\xi,t)}{d(\xi,t)}\right)_t d\xi. \tag{2.8}$$

Combining (2.7) and (2.8) yields $s_t=w_x=(ds_x)_x$, the first equation in (2.1).

To solve (2.1) and (2.2), we solve (2.4), (2.5) by an iterative process and then obtain s via (2.6). The iteration is as follows: let $w^0(x,t)\equiv d_0(x)s_{0x}(x)$; let $d^0(x,t)\equiv d_0(x)$; then define d^N and w^N inductively to be the solutions of

$$w_t^N=d^{N-1}w_{xx}^N+\left[\frac{g(w^{N-1})}{d^{N-1}}-1\right]w^N, \qquad d_t^N=-d^N+g(w^{N-1}) \tag{2.9}$$

in $(0,1)\times(0,T]$ with the initial and boundary conditions

$$\begin{aligned} &w^N(0,t)=0, \qquad w^N(1,t)=\psi,\\ &w^N(x,0)=w_0(x)\equiv d_0(x)s_{0x}(x), \qquad d^N(x,0)=d_0(x). \end{aligned} \tag{2.10}$$

Observe that (2.9) consists of two independent linear equations. If d^{N-1} and w^{N-1} are Hölder continuous and $d^{N-1}(x,t)\geq\delta_1>0$, then (2.9) and (2.10) can be solved for w^N and d^N, and w^N and d^N will also be Hölder continuous. Also

$$\begin{aligned} d^N(x,t)&=e^{-t}d_0(x)+e^{-t}\int_0^t e^{\tau}g\big(w^{N-1}(x,\tau)\big)\,d\tau\\ &\geq e^{-t}\delta_0+e^{-t}\int_0^t g_0e^{\tau}\,d\tau\geq\min(g_0,\delta_0)>0. \end{aligned} \tag{2.11}$$

Thus, since d_0 and s_0 were assumed to be smooth and $d_0(x,t)\geq\delta_0$, it follows that the iteration is well defined. The compatibility conditions following (2.5) insure that the iterates w^N, and d^N also, are smooth enough on $[0,1]\times[0,T]$ for the computations needed to bound their derivatives uniformly in N. (See the discussion following [2, Thm. 2, §3].) Inequality (2.11) provides a uniform lower bound on the iterates d^N; a

similar computation, using the upper bound for $g(\xi)$ in (2.3) together with (2.11) yields, for all N,

$$|d^N(x,t)| \le D_0 \tag{2.12}$$

for some constant D_0 depending only on $g(\xi)$ and $d_0(x)$. Since d^N is bounded below and $|g|$ is bounded, it follows that for some constant H_0 independent of N,

$$\left| \frac{g(w^N)}{d^N} - 1 \right| \le H_0. \tag{2.13}$$

To see that $|w^N|$ is also uniformly bounded, we use a maximum principle argument. The following lemma is closely related to [2, Thm. 5, §1] and [6, Chap. 3, Thms. 5–7].

LEMMA 2. *Suppose that $u(x,t)$ is a classical solution of*

$$L[u] \equiv u_t - a(x,t)u_{xx} - b(x,t)u_x - c(x,t)u = f(x,t) \tag{2.14}$$

in $(0,1)\times(0,T]$, with

$$u_x(0,t) = 0, \quad u(1,t) = 0, \quad u(x,0) = 0. \tag{2.15}$$

Assume that the coefficients of L are bounded and continuous, with $c(x,t) \le K_0$ and $|f(x,t)| \le F_0$. Then there exists a constant $M_0 = M_0(K_0, F_0)$ so that

$$|u(x,t)| \le M_0 \quad \text{on } [0,1]\times[0,T]. \tag{2.16}$$

If $K_0 = -c_0 < 0$, then $M_0 = F_0/c_0$.

From Lemma 2 we obtain bounds for w^N.

LEMMA 3. *Suppose that $w^N(x,t)$ is a solution of* (2.9), (2.10) *and $w_0(x)$ is smooth. Then there exists a constant $W_0 = W_0(w_0, D_0)$ independent of N, so that*

$$|w^N(x,t)| \le W_0. \tag{2.17}$$

Proof. Let $u = w^N - w_0$. Then u satisfies the equation

$$u_t = d^{N-1}u_{xx} + \left[g(w^{N-1})/d^{N-1} - 1\right]u + d^{N-1}w_{0xx} + \left[g(w^{N-1})/d^{N-1} - 1\right]w_0.$$

It follows from (2.3) and (2.11) that for some H_0 independent of N,

$$|g(w^{N-1})/d^{N-1} - 1| \le H_0. \tag{2.18}$$

Similarly, $|d^{N-1}w_{0xx} + [g(w^{N-1})/d^{N-1} - 1]w_0|$ is bounded by a constant depending only on w_0 and D_0. Also $u_x(0,t) = 0$ and $u(x,0) = 0$, $u(1,t) = 0$; thus Lemma 2 applies to u and we have $|u| \le M_0(w_0, D_0)$. Finally, $|w^N| \le |u| + \sup|w_0|$, which yields (2.17).

The next step is to obtain estimates on the derivatives of w^N, d^N which are independent of N. The a priori bounds will allow us to prove the uniform convergence of the sequences $\{w^N\}$ and $\{d^N\}$ and via the Arzela–Ascoli theorem the convergence of subsequences of the derivatives of w^N and d^N occurring in (2.9). Taking the limit of the subsequences then yields a solution to (2.4).

LEMMA 4. *There exist constants D_1 and W_1 independent of N such that for all N, $|w_x^N| \le W_1$ and $|d_x^N| \le D_1$ in $[0,1]\times[0,T]$. Similarly, there exist constants D_2 and W_2 such that $|w_{xx}^N| \le W_2$ and $|d_{xx}^N| \le D_2$.*

Proof. To estimate w_x^N and d_x^N, we first obtain a uniform bound for w_x^N on the parabolic boundary of $(0,1)\times(0,T]$, then use a result from [3] to obtain a bound for w_x^N on the interior and finally bound d_x^N in terms of the bound on w_x^N. The second derivative estimates follow similarly.

Let $u=w^N-w_0$; then

$$u_x(0,t)=0, \quad u(x,0)=0 \quad u(1,t)=0.$$

As in Lemma 3, there exists a constant M_0 so that $|u|\leq M_0$. If we set $L^N[u]=-u_t+d^{N-1}u_{xx}$, then

$$L^N[u]=-\left[\frac{g(w^{N-1})}{d^{N-1}}-1\right]u-\left\{d^{N-1}w_{0xx}+\left[\frac{g(w^{N-1})}{d^{N-1}}-1\right]w_0\right\},$$

which implies that

$$|L^N[u]|\leq H_0M_0+E_0,$$

where E_0 is a constant depending on H_0, w_0 and D_0.

Let $v=e^{\alpha(x-1/2)}$. Then by (2.11), $L^N[v]=d^{N-1}\alpha^2 v\geq \min(g_0,\delta_0)\alpha^2$ in $[\frac{1}{2},1]\times[0,T]$ since $d^{N-1}\geq\min(g_0,\delta_0)$. Thus,

$$L^N[\pm u+v]\geq 0 \quad \text{in } \left[\tfrac{1}{2},1\right]\times[0,T] \quad \text{for } \alpha\geq\left(\frac{H_0M_0+E_0}{\min(g_0,\delta_0)}\right)^{1/2}$$

Furthermore, assume that α is large enough so that $e^{\alpha/2}\geq M_0+1$. Then $\pm u+v\leq e^{\alpha/2}$ on the parabolic boundary of $(\frac{1}{2},1)\times(0,T]$, so by the maximum principle, $\pm u+v\leq e^{\alpha/2}$ throughout $[\frac{1}{2},1]\times[0,T]$. Thus $\pm u_x+v_x\geq 0$ for $x=1$, or $\pm u_x+\alpha e^{\alpha/2}\geq 0$ at $x=1$. Therefore $|u_x|\leq\alpha e^{\alpha/2}$ at $x=1$, which implies that at $x=1$, $|w_x^N|\leq\alpha e^{\alpha/2}+|w_{0x}(1)|$. Note that from the initial and boundary conditions we can estimate $|w_x^N|$ at $x=0$ and at $t=0$ independent of N. Thus, on the parabolic boundary of $(0,1)\times(0,T]$, we have

$$|w_x^N|\leq\alpha e^{\alpha/2}+\sup_x|w_{0x}(x)|=B_0.$$

Let $v^N=w_x^N$. We have $v^N\leq V_0$ on the parabolic boundary of $(0,1]\times(0,T]$ independent of N. Differentiating (2.9) with respect to x yields

$$v_t^N=\left(d^{N-1}v_x^N\right)_x+\left(\left[\frac{g(w^{N-1})}{d^{N-1}}-1\right]w^N\right)_x. \tag{2.19}$$

Equation (2.19) is in divergence form, and the inhomogeneous term is of the form $\partial z/\partial x$, where $|z|\leq H_0W_0$. We thus may apply [3, Chap. III, Thm. 7.1] to conclude that for some V_2 independent of N,

$$\sup_{[0,1]\times[0,T]}|w_x^N|=\sup_{[0,1]\times[0,T]}|v|\leq V_2\equiv W_1. \tag{2.20}$$

Hence $|w_x^N|$ is bounded uniformly with respect to N. The uniform boundedness of d_x^N follows by differentiating the second equation of (2.9) with respect to x to obtain $d_{xt}^N=-d_x^N+g'(w^{N-1})w_x^{N-1}$. Since $|w_x^{N-1}|\leq W_1$, the argument leading to (2.12) may be repeated to obtain $|d_x^N|\leq D_1$.

To bound w_{xx}^N at $x=1$, note that since $w^N(1,t)=\psi$, $w_t^N(1,t)=0$; so at $x=1$, w_{xx}^N can be computed explicitly in terms of d^{N-1} and w^{N-1}.Since $|w^{N-1}|$ is bounded and d^{N-1} is bounded above and below uniformly, it follows that $|w_{xx}^N(1,t)|$ is uniformly bounded. At $x=0$ we have $w_x^N(0,t)=0$ so we can estimate $|w_{xx}^N(0,t)|$ via the same argument used to bound $|w_x^N(1,t)|$ earlier in the proof. To bound w_{xx}^N in the interior, let $\tilde{v}^N=w_{xx}^N$. We have

$$\tilde{v}_t^N=\left(d^{N-1}\tilde{v}^N\right)_{xx}+\left(\left[\frac{g(w^{N-1})}{d^{N-1}}-1\right]w^N\right)_{xx}=\left(d^{N-1}\tilde{v}_x^N+d_x^{N-1}\tilde{v}^N\right)_x+\tilde{h}_x, \tag{2.21}$$

where

$$\tilde{h}\equiv\left(\left[\frac{g(w^{N-1})}{d^{N-1}}-1\right]w^N\right)_x.$$

The second equation in (2.21) is in divergence form and d_x^{N-1} and $\tilde{h}$ are uniformly bounded, so again [3, Chap. III, Thm. 7.1] yields a uniform bound for $w_{xx}^N=\tilde{v}^N$ as in (2.20). A bound for d_{xx}^N is obtained from the bounds on w_{xx}^N and lower order terms just as the bound for d_x^N was obtained from that for w_x^N.

Remark. Since w^N, w_{xx}^N and d^N are uniformly bounded, it follows that w_t^N and d_t^N are also.

LEMMA 5. *The sequences* $\{w^N\}$ *and* $\{d^N\}$ *defined by* (2.9) *converge uniformly on* $[0,1]\times[0,T]$.

Proof. Let $v^N=w^N-w^{N-1}$, $c^N=d^N-d^{N-1}$ and $h^N=g(w^N)/d^N-1$. We then have

$$v_t^N=d^{N-1}v_{xx}^N+h^{N-1}v^N+c^{N-1}w_{xx}^{N-1}+(h^{N-1}-h^{N-2})w^{N-1} \tag{2.22}$$

on $(0,1)\times(0,T]$ with $v_x(0,t)=v(1,t)=v(x,0)\equiv 0$. But

$$h^{N-1}-h^{N-2}=\frac{\left[g(w^{N-1})-g(w^{N-2})\right]}{d^{N-1}}+g(w^{N-2})\left[\frac{1}{d^{N-1}}-\frac{1}{d^{N-2}}\right]$$

and $g(w^{N-1})-g(w^{N-2})=F_0^N v^{N-1}$ where

$$F_0^N(x,t)=\int_0^1 g'\left(\xi w^{N-1}+(1-\xi)w^{N-2}\right)d\xi, \tag{2.23}$$

so we can rewrite (2.22) as

$$v_t^N=d^{N-1}v_{xx}^N+h^{N-1}v^N+F_1^N v^{N-1}+F_2^N c^{N-1}, \tag{2.24}$$

where $F_1^N=F_0^N w^{N-1}/d^{N-1}$ and $F_2^N=w_{xx}^{N-1}-w^{N-1}g(w^{N-2})/d^{N-1}d^{N-2}$. Similarly, on $(0,1)\times(0,T]$ we have

$$c_t^N=-c^N+F_0^N v^{N-1} \tag{2.25}$$

with $c^N(x,0)=0$. It follows from the a priori estimates of Lemma 4 and from (2.13) that there exist constants k_0, k_1, k_2 and H_0 such that for all N,

$$|F_j^N|\le k_j,\quad j=0,1,2,\qquad |h^N|\le H_0. \tag{2.26}$$

Suppose that $\alpha>H_0$; then $-\alpha+h^N\le-\alpha+H_0<0$ for all N. Let $\tilde{v}^N=e^{-\alpha t}v^N$ and $\tilde{c}^N=e^{-\alpha t}c^N$. We have

$$\begin{aligned}\tilde{v}_t^N&=d^{N-1}\tilde{v}_{xx}^N+\left[-\alpha+h^{N-1}\right]\tilde{v}^N+F_1^N\tilde{v}^{N-1}+F_2^N\tilde{c}^{N-1},\\ \tilde{c}_t^N&=-(1+\alpha)\tilde{c}^N+F_0^N\tilde{v}^{N-1}.\end{aligned} \tag{2.27}$$

Let

$$p^N=\sup_{\substack{0\le x\le 1\\ 0\le t\le T}}|\tilde{v}^N|,\qquad q^N=\sup_{\substack{0\le x\le 1\\ 0\le t\le T}}|\tilde{c}^N|.$$

It follows from (2.27) and the fact that $\tilde{c}^N(x,0)=0$ that

$$\tilde{c}^N=e^{-(1+\alpha)t}\int_0^t e^{(1+\alpha)\tau}F_0^N(x,\tau)\tilde{v}^{N-1}(x,\tau)\,dt.$$

Hence, by (2.25) we have

$$q^N = \sup_{\substack{0\le x\le 1\\ 0\le t\le T}} |\tilde{c}^N| \le k_0 p^{N-1}. \tag{2.28}$$

Since $\tilde{v}^N$ satisfies homogeneous initial and boundary conditions, it follows from (2.26), (2.27) and Lemma 2 that

$$p^N = \sup|\tilde{v}^N| \le \frac{k_1 p^{N-1} + k_2 q^{N-1}}{\alpha - H_0}. \tag{2.29}$$

But (2.28) implies that $q^{N-1} \le k_0 p^{N-2}$, so (2.29) yields

$$p^N \le \frac{k_1 p^{N-1} + k_0 k_2 p^{N-2}}{\alpha - H_0}.$$

Choose α large enough such that $\alpha - H_0 \ge 3\max\{k_1, k_0 k_2\}$, then

$$p^N \le \frac{1}{3}\left(p^{N-1} + p^{N-2}\right). \tag{2.30}$$

Let $P = \max\{p^1, p^2\}$. Then by (2.30), $p^{2k+1} \le (\frac{2}{3})^k P$ and $p^{2k+2} \le (\frac{2}{3})^k P$. Thus, since $p^k \ge 0$, we may write $\Sigma_{N=1}^{\infty} p^N \le 2P \Sigma_{N=1}^{\infty} (\frac{2}{3})^N < \infty$ and conclude that the series $\Sigma_{N=1}^{\infty} p^N$ converges. By (2.28) the series $\Sigma_{N=1}^{\infty} q^N$ also converges. It follows that the series $\Sigma_{N=1}^{\infty} \tilde{v}^N$ and $\Sigma_{N=1}^{\infty} \tilde{c}^N$ converge uniformly. Since $v^N = e^{\alpha t}\tilde{v}^N$ and $c^N = e^{\alpha t}\tilde{c}^N$ and $e^{\alpha t}$ is bounded on $[0, T]$, the series $\Sigma_{N=1}^{\infty} v^N$ and $\Sigma_{N=1}^{\infty} c^N$ are also uniformly convergent on $[0, 1]\times[0, T]$. But $w^N = w_0 + \Sigma_{k=1}^{N} v^k$ and $d^N = d_0 + \Sigma_{k=1}^{N} c^k$, so the sequences $\{w^N\}$ and $\{d^N\}$ converge uniformly.

We can now prove the main result of this section.

THEOREM 1. *Suppose that $s_0(x)$ and $d_0(x)$ are smooth functions such that the compatibility conditions following* (2.5) *are satisfied. Then there exists a unique solution to* (2.4) *and* (2.5), *and hence to* (2.1) *and* (2.2), *in* $[0, 1]\times[0, T]$. *Since $T > 0$ was arbitrary, the solution, in fact, exists and is unique throughout* $[0, 1]\times[0, \infty)$.

Proof. By Lemma 1 and the remark following that lemma, it suffices to consider (2.4), (2.5). By (2.12) and Lemma 3, d^N and w^N are uniformly bounded. Lemma 5 and the remark following imply that w^N and d^N are uniformly Hölder continuous with respect to x and t.

Thus the coefficients of (2.9) are uniformly bounded and Hölder continuous. The Schauder theory of parabolic equations (see, for example, [3, Chap. IV]) then asserts that w_{xx}^N and w_t^N are uniformly Hölder continuous. Equation (2.9) implies that d_t^N is uniformly Hölder continuous. By the Arzela–Ascoli theorem, we can choose subsequences $\{w^{N_k}\}$ and $\{d^{N_k}\}$ of the iterates defined in (2.9) so that $\{w_{xx}^{N_k}\}$, $\{w_t^{N_k}\}$, and $\{d_t^{N_k}\}$ converge uniformly on $[0, 1]\times[0, T]$. By Lemma 5, the sequences $\{w^{N_k}\}$, $\{w^{N_k+1}\}$, $\{d^{N_k}\}$, $\{d^{N_k+1}\}$ converge uniformly on $[0, 1]\times[0, T]$, since the original sequences $\{w^N\}$ and $\{d^N\}$ converge. Passing to a limit as $k \to \infty$ in (2.9) yields a solution to (2.4), (2.5) in $[0, 1]\times[0, T]$.

To establish uniqueness we use an argument similar to that used in proving Lemma 6. Suppose that the pairs of functions (w_1, d_1) and (w_2, d_2) both satisfy (2.4) and (2.5). Let $u = w_1 - w_2$, $c = d_1 - d_2$. Then in $(0, 1)\times(0, T]$,

$$\begin{aligned} u_t &= w_{1t} - w_{2t} = d_1 w_{xx} - d_2 w_{2xx} + \left[\frac{g(w_1)}{d_1} - 1\right] w_1 - \left[\frac{g(w_2)}{d_2} - 1\right] w_2 \\ &= d_1 u_{xx} + c w_{2xx} + \left[\frac{g(w_1)}{d_1} - 1\right] u + \left[\frac{g(w_1)}{d_1} - \frac{g(w_2)}{d_2}\right] w_2. \end{aligned} \tag{2.31}$$

We have

$$\frac{g(w_1)}{d_1}-\frac{g(w_2)}{d_2}=\frac{g(w_1)-g(w_2)}{d_1}+g(w_2)\left(\frac{1}{d_1}-\frac{1}{d_2}\right)$$

and

$$g(w_1)-g(w_2)=(w_1-w_1)\int_0^1 g'(\theta w_1+(1-\theta)w_2)\,d\theta.$$

Let $f_1(x,t)=\int_0^1 g'(\theta w_1+(1-\theta)w_2)\,d\theta$; let $f_2(x,t)=-g(w_2)w_2/d_1d_2$. Then (2.41) may be written as

$$u_t=d_1u_{xx}+\left[\frac{g(w_1)}{d_1}-1\right]u+w_2f_1u+(w_{2xx}+f_2)c. \tag{2.32}$$

Also,

$$c_t=-c+f_1u. \tag{2.33}$$

Let $M_0=\max(\sup|w_1|,\sup|w_2|)$, $M_1=\sup|w_{2xx}|$, $F_1=\sup|f_1|$, $F_2=\sup|f_2|$. These constants will exist and be finite since (2.4) and (2.5) and the compatibility conditions on the data insure that w_1 and w_2 are smooth and bounded and d_1 and d_2 are bounded from above and below. Let $H_0=\sup|(g(w_1)/d_1-1)|$. Suppose that $\alpha>H_0$; let $\tilde{u}=e^{-\alpha t}u$ and $\tilde{c}=e^{-\alpha t}c$. Then

$$\begin{aligned}\tilde{u}_t&=d\tilde{u}_{xx}+\left(\left[\frac{g(w_1)}{d_1}-1\right]-\alpha\right)\tilde{u}+w_2f_1\tilde{u}+(w_{2xx}+f_2)\tilde{c},\\ \tilde{c}_t&=-(\alpha+1)\tilde{c}+f_1\tilde{u}.\end{aligned} \tag{2.34}$$

Also, $\tilde{u}(x,0)=\tilde{u}(1,t)=0$, $\tilde{u}_x(0,t)=0$ and $\tilde{c}(x,0)=0$. Let $U_0=\sup|\tilde{u}|$ and $C_0=\sup|\tilde{c}|$. By Lemma 2, (2.34) implies that

$$U_0=\sup|\tilde{u}|\le\frac{M_0F_1U_0+(M_1+F_2)C_0}{\alpha-H_0}. \tag{2.35}$$

Estimating $\tilde{c}$ from (2.44) by integrating, as in (2.11), yields

$$C_0=\sup|\tilde{c}|\le F_1U_0. \tag{2.36}$$

Combining (2.35) and (2.36) yields

$$0\le U_0\le\frac{U_0[M_0F_1+(M_1+F_2)F_1]}{\alpha-H_0}. \tag{2.37}$$

Choose $\alpha>H_0$ large enough so that $[M_0F_1+(M_1+F_2)F_1]/(\alpha-H_0)<1$; then (2.37) implies that $U_0=0$, which implies, via (2.36), that $C_0=0$. Hence $\tilde{u}=\tilde{c}\equiv0$, so $u=c\equiv0$, so $w_1\equiv w_2$ and $d_1\equiv d_2$, which proves the uniqueness assertion of the theorem.

3. Analysis of the steady state. In this section we examine the behavior of the steady state solution to system (2.4)–(2.5). It is easy to check that the unique steady state solution is given by

$$w\equiv\psi,\qquad d\equiv g(\psi). \tag{3.1}$$

Thus, the steady state solution to the original problem (2.1)–(2.2) is

$$s=\frac{\psi x}{g(\psi)},\qquad d=g(\psi). \tag{3.2}$$

We examine the linear stability of this steady state by linearizing (2.4) about the solution (3.1). Thus, consider

$$\overline{W}_t = g(\psi)W_{xx} + \frac{\psi g'(\psi)}{g(\psi)}\overline{W} - \frac{\psi}{g(\psi)}\overline{D}, \qquad \overline{D}_t = -\overline{D} + g'(\psi)\overline{W}. \tag{3.3}$$

Also, $\overline{W}(x,t)$ satisfies

$$\overline{W}_x(0,t) = 0, \qquad \overline{W}(1,t) = 0. \tag{3.4}$$

Here $g'(\psi)$ means $\frac{dg}{d\psi}$. Assuming $(\overline{W},\overline{D}) = e^{\lambda t}(W(x),D(x))$, then (3.3)–(3.4) reduce to

$$W_{xx} + \mu^2 W = 0, \qquad W_x(0) = 0 = W(1), \tag{3.5}$$

where

$$\mu^2 \equiv \frac{1}{g(\psi)}\left\{\frac{\psi g'(\psi)\lambda}{g(\psi)(1+\lambda)} - \lambda\right\}. \tag{3.6}$$

Therefore we have nontrivial solutions if and only if $\mu = \mu_n = (n+\frac{1}{2})\pi$, n an integer, and this corresponds to $\lambda = \lambda_n$, which from (3.6), satisfies

$$\lambda_n^2 - T_n(\psi)\lambda_n + U_n(\psi) = 0, \tag{3.7}$$

where

$$T_n(\psi) \equiv \psi g'(\psi)/g(\psi) - 1 - g(\psi)\pi^2\left(n+\frac{1}{2}\right)^2, \qquad U_n(\psi) \equiv g(\psi)\pi^2\left(n+\frac{1}{2}\right)^2. \tag{3.8}$$

Hence

$$2\lambda_n = 2\lambda_n^{\pm} = T_n(\psi) \pm \sqrt{T_n^2(\psi) - 4U_n(\psi)}\,. \tag{3.9}$$

Thus, (3.1) is linearly stable if and only if $\mathrm{Re}(\lambda_n) < 0$ for all n and unstable if there is an integer m such that $\mathrm{Re}(\lambda_m) > 0$. From (3.7) we see the following holds:

LEMMA 7. *$(w,d) = (\psi, g(\psi))$ is linearly stable if and only if $T_n(\psi) < 0$ for all n, if and only if $\psi g'(\psi) - g(\psi) < (\pi^2/4)g^2(\psi)$.*

Remark. Since g is bounded away from zero for positive arguments, we see that for ψ sufficiently small, the inequalities in Lemma 7 are guaranteed to hold.

It would be expected, at least for ψ small, that the steady state would not only be locally but also globally stable. This can be indicated formally by considering the problem (2.4)–(2.5) with g replaced by a piecewise constant approximation. We have been unable to obtain such a global stability result in the general case, but we can show asymptotic stability in an L^2 sense. We do this by considering the equivalent problem (2.1)–(2.2) with steady state solution (3.2).

LEMMA 8. *$(s,d) = (\psi x/g(\psi), g(\psi))$ is asymptotically stable in the L^2-norm provided*

$$\frac{1}{\delta_0}\left[G_1 D_0 + \frac{\psi}{g(\psi)}\right] + \frac{G_1\psi}{g'(\psi)} < 1. \tag{3.10}$$

Remark. The constants appearing in (3.10) are those defined in the previous section. Condition (3.10) is clearly satisfied if ψ is sufficiently small and δ_0 is sufficiently large. When $\psi = 0$, (3.10) is not needed.

Proof. Let

$$c(x,t) = d(x,t) - g(\psi), \qquad r(x,t) = s(x,t) - \frac{\psi x}{g(\psi)}, \tag{3.11}$$

where s and d are solutions to (2.1)–(2.2). Since $r(0,t)\equiv 0$, it follows that $r(x,t)=\int_0^x r_x(\xi,t)\,d\xi$. Also, $r_t=s_t=(ds_x)_x$ and $c_t=d_t=-c+g(ds_x)-g(\psi)$. Applying the mean value theorem, we have

$$c_t=-c+\bar{g}(x,t)[ds_x-\psi]$$

or

$$c_t=-c+\bar{g}(x,t)\left[dr_x-\frac{\psi c}{g(\psi)}\right], \tag{3.12}$$

where $\bar{g}(x,t)=g'(\xi)$ for some ξ depending on ds_x. Now compute

$$\frac{d}{dt}\int_0^1 r^2(\xi,t)\,d\xi=\int_0^1 r(ds_x)_x d\xi.$$

Integrating by parts and using the boundary data and (3.11), we obtain

$$\begin{aligned}\frac{d}{dt}\int_0^1 \frac{1}{2}r^2\,d\xi&=\psi r(1,t)-\int_0^1 ds_x r_x\,d\xi\\&=\psi\int_0^1 r_x(\xi,t)\,d\xi-\int_0^1 d\left(r_x+\frac{\psi}{g(\psi)}\right)r_x\,d\xi\\&=\int_0^1\left[-dr_x^2-\frac{\psi}{g(\psi)}cr_x\right]d\xi.\end{aligned}$$

Similarly, multiplying (3.12) by c and integrating, we obtain

$$\frac{d}{dt}\int_0^1\frac{1}{2}c^2(\xi,t)\,d\xi=\int_0^1\left\{\left[-1+\frac{\psi}{g(\psi)}\bar{g}\right]c^2+\bar{g}\,dcr_x\right\}d\xi.$$

Using $\bar{g}\le G_1=\sup|g'|$ and $\delta_0\le d\le D_0$ and Cauchy's inequality, we have

$$\begin{aligned}\frac{d}{dt}\int_0^1\frac{1}{2}(r^2+c^2)\,d\xi&\le\int_0^1\left\{-\delta_0 r_x^2+\left[G_1D_0+\frac{\psi}{g(\psi)}\right]cr_x+\left[-1+\frac{G_1\psi}{g(\psi)}\right]c^2\right\}d\xi\\&\le\int_0^1\left\{\left(\frac{1}{2\varepsilon}-\delta_0\right)r_x^2+\left(\frac{\varepsilon}{2}\left[G_1D_0+\frac{\psi}{g(\psi)}\right]+\frac{G_1\psi}{g(\psi)}-1\right)c^2\right\}d\xi.\end{aligned}$$

Since via the Cauchy–Schwarz inequality we have

$$2\int_0^1 r^2\,d\xi\le\int_0^1 r_x^2\,d\xi,$$

by choosing $\varepsilon>1/2\delta_0$ we may write

$$\frac{d}{dt}\int_0^1\frac{1}{2}(r^2+c^2)\,d\xi\le\int_0^1\left\{2\left(\frac{1}{2\varepsilon}-\delta_0\right)r^2+\left(\frac{\varepsilon}{2}\left[G_1D_0+\frac{\psi}{g(\psi)}\right]+\frac{G_1\psi}{g(\psi)}-1\right)c^2\right\}d\xi.$$

Hence, if we choose $1/2\delta_0<\varepsilon<2\delta_0$, then there is a K such that

$$\frac{d}{dt}\int_0^1\frac{1}{2}(r^2+c^2)\,d\xi\le -K\int_0^1(r^2+c^2)\,d\xi,$$

which implies asymptotic stability.

Within the class of g's we have been considering, it is not difficult to find examples where either the inequality in Lemma 7 holds for all $\psi>0$ or else there are ψ

such that the inequality is violated. For example, for $0<\beta$ small, let

$$g(\psi)=\frac{\beta+\psi^2}{1+k^2\psi^2};$$

then it is easy to verify that there are at least two $\bar{\psi}$ such that

$$\bar{\psi}g'(\bar{\psi})-g(\bar{\psi})=\frac{\bar{\psi}^2}{4}g^2(\bar{\psi}).$$

For very large or very small ψ, the inequality holds. Thus, when ψ increases through the smallest such $\bar{\psi}$, the steady state becomes unstable, but as ψ is increased further, the steady state eventually is "restabilized."

Consider a g with at least one such value, and let the smallest positive one be denoted ψ_c. Thus, we have the steady state solution linearly stable for $0<\psi<\psi_c$, while in some neighborhood of ψ_c with $\psi>\psi_c$ the steady state solution becomes unstable. We are interested in the nature of this change of stability as ψ passes through ψ_c. At $\psi=\psi_c$ (3.8) implies $T_0(\psi_c)=0$, $T_n(\psi_c)<0$ for $n>0$ and $U_n(\psi_c)>0$ for all n. Thus, if we define $\omega_0\equiv\pi\sqrt{g(\psi_c)}/2$, then $\lambda_0^{\pm}=\pm i\omega_0$. For $n>0$ we have $\mathrm{Re}(\lambda_n^{\pm})<0$. Also, for $\psi=\psi_c+\varepsilon$, $|\varepsilon|\ll 1$,

$$\lambda_0^{\pm}(\psi)=\frac{1}{2}T_0'(\psi_c)\varepsilon\pm i\omega_0\left(1+\frac{\varepsilon}{2}\frac{g'(\psi_c)}{g(\psi_c)}\right)+O(\varepsilon^2)\quad\text{as }\varepsilon\to 0, \tag{3.13}$$

so we suppose g is such that $T_0'(\psi_c)>0$. As we vary ψ through ψ_c, the $\lambda_0^{\pm}$ cross the imaginary axis with nonvanishing imaginary part and all other eigenvalues have negative real parts. It is in this respect that we mean the bifurcation is a Hopf-type bifurcation [4]. Linearized stability theory predicts the steady state loses its stability via an exponentially growing function of t of rate $O(T_0'(\psi_c)\varepsilon)$. But this behavior cannot represent the solution very long since the nonlinear terms become important.

To indicate the type of behavior we expect for $0<\psi-\psi_c\ll\varepsilon$, consider the solution to (3.3)–(3.4) for $\psi=\psi_c$. We can write it in the form

$$\begin{pmatrix}\bar{W}\\ \bar{D}\end{pmatrix}=A_0\begin{pmatrix}1\\ \dfrac{g'(\psi_c)(1-i\omega_0)}{1+\omega_0^2}\end{pmatrix}e^{i\omega_0 t}\cos\frac{\pi x}{2}+\text{c.c.}+\text{d.t.}, \tag{3.14}$$

where A_0 is an arbitrary constant, c.c. means complex conjugate and d.t. means (exponentially) decaying terms. For $\psi\equiv\psi_c=\varepsilon>0$ sufficiently small, the structure of the solution (w,d) of (2.4)–(2.5) should be similar to (3.14) but should grow at an exponential rate of $O(\varepsilon t)$. Our approach to analyzing the local bifurcation problem is to employ a formal two-timing perturbation approach. Details of the asymptotic solution will appear elsewhere. We will just indicate the approach here. Replace w by $w-\psi_c$ and d by $d-g(\psi_c)$ in (2.4)–(2.5) to obtain

$$\begin{aligned}w_t&=(g(\psi_c)+d)w_{xx}+(\psi_c+w)\left[\frac{g(\psi_c+w)}{g(\psi_c)+d}-1\right],\\ d_t&=-d-g(\psi_c)+g(\psi_c+w)\end{aligned} \tag{3.15}$$

with

$$w_x(0,t)=0,\qquad w(1,t)=\psi-\psi_c. \tag{3.16}$$

Define $\tau=\varepsilon t$ and let

(3.17)

$$w=\varepsilon w_1(x,t,\tau)+\varepsilon^2 w_2(x,t,\tau)+O(\varepsilon^3), \qquad d=\varepsilon d_1(x,t,\tau)+\varepsilon^2 d_2(x,t,\tau)+O(\varepsilon^3).$$

Upon substitution of (3.17) into (3.15)–(3.16) and equating powers of ε, we obtain a sequence of problems of the form

$$\begin{pmatrix} w_1 \\ d_1 \end{pmatrix}_t = \mathfrak{L}\begin{pmatrix} w_1 \\ d_1 \end{pmatrix}, \quad w_{1x}(0,t,\tau)=0, \quad w_1(1,t,\tau)=1, \tag{3.18}$$

$$\begin{pmatrix} w_2 \\ d_2 \end{pmatrix}_t = \mathfrak{L}\begin{pmatrix} w_2 \\ d_2 \end{pmatrix} + \mathfrak{N}\begin{pmatrix} w_1 \\ d_1 \end{pmatrix}, \quad w_{2x}(0,t,\tau)=0, \quad w_2(1,t,\tau)=0, \tag{3.19}$$

etc., where the operator $\mathfrak{L}$ is defined by the right-hand side of (3.3) and $\mathfrak{N}$ is a nonlinear term involving just (w_1, d_1). Solving (3.18) yields a solution of the form

$$w_1=\left(A_0(\tau)e^{i\omega_0 t}+1\right)\cos\frac{\pi x}{2}+\text{c.c.}+\text{d.t.},$$

$$d_1=\left(A_0(\tau)z_c e^{i\omega_0 t}+r_c\right)\cos\frac{\pi x}{2}+\text{c.c.}+\text{d.t.}$$

with $z_c\equiv g'(\psi_c)(1-i\omega_0)/(1+\omega_0^2)$ and $r_c\equiv\pi^2 g(\psi_c)g'(\psi_c)/4$. $A_0(\tau)$ as a function of the slow time variable is determined at the next stage in the perturbation procedure. Substitution of (3.20) into (3.19) yields the form for $\mathfrak{N}$ which must be orthogonal to solutions of (3.18), since the problem is self-adjoint. This gives a solvability condition for determining $A_0(\tau)$, and so the form of the lowest order perturbation is completely specified.

Therefore, as ψ passes through ψ_c, the signal concentration evolves to a time periodic oscillation of approximate period $2\pi/\omega_0$ but which has slow modulation and amplitude changes on the time scale $\tau=O(1)$.

REFERENCES

[1] G. M. Fateeva, *Boundary value problems for degenerate quasilinear parabolic equations*, Math. USSR Sbornik, 5 (4), (1968), pp. 509–532.

[2] A. M. Il'in, A. S. Kalashnikov and O. A. Oleinik, *Linear equations of the second order of parabolic type*, Russian Math. Surveys, 17 (3) (1962), pp. 1–143.

[3] O. A,. Ladyženskaja, V. A. Solonnikov and N. N. Ural'ceva, *Linear and quasilinear equations of parabolic type*, Transl. of Mathematical Monographs, Vol. 23, American Mathematical Society, Providence, RI, 1968.

[4] J. E. Marsden and M. McCracken, *The Hopf Bifurcation and Its Applications*, Applied Mathematical Sciences series, Vol. 19, Springer-Verlag, New York, 1976.

[5] G. J. Mitchison, *A model for vein formation in higher plants*, Proc. Roy. Soc. Lond. B, 207 (1980), pp. 79–109.

[6] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0006 $01.00/0

NONUNIFORM SEMI-INFINITE GROUNDED GRIDS*

A. H. ZEMANIAN†

Abstract. Semi-infinite resistive grounded grids are countably infinite electrical networks that arise from the discretization of the partial differential equation governing the minority-carrier density in a doped semiconductor. If the doping varies with depth from the surface of the semiconductor, the grid's resistances also vary with distance from the inputs to the grid. This nonuniformity prevents the use of the characteristic-resistance method for determining currents and voltages. A computational method for making such a determination is presented herein. It is based upon the theory of infinite continued fractions whose entries are positive operators on a Hilbert space. It is also know that the solution given by the method is precisely that solution for which the power dissipated in the network is finite. Finally, the method is extended to RLC networks, and this allows the computation of transient responses in semi-infinite grounded grids of positive-real impedances.

1. Introduction. The purpose of this work is to examine the behavior of a certain class of countably infinite electrical networks. Although individual networks in this class can be highly complex, the prototype of this class is comparatively simple and results from the discretization of a certain partial differential equation relating to semiconductor behavior. Therefore, to motivate our work, we first indicate why the prototype is of interest in the theory of semiconductors.

The partial differential equation that governs the minority-carrier density δ in a doped semiconductor is

$$\nabla^2\delta=\frac{\delta}{\tau D}, \tag{1.1}$$

where τ is the minority-carrier lifetime and D is the minority-carrier diffusion constant [9, p. 99]. Ordinarily, the doping concentration, and therefore τ as well, varies with distance from the surface through which the impurities are introduced. (There are of course lateral variations along the surface where the $p-n$ junctions appear, but these variations disappear just below that section.) Because of this, no closed-form solution for (1.1) exists, and computational techniques must be used to get an approximate determination of δ. However, the standard techniques, such as difference methods or finite-element methods, lead to excessively large computer times when the full thickness of the semiconductor wafer is modeled.

An alternative possibility is to assume that one surface of the wafer is at infinity and then make use of the theory of semi-infinite transmission lines. This approach was explored in [17] in the case where the doping does not vary with position. It led to the adaptation of the characteristic-resistance method to semi-infinite grounded grids, the kind of electrical network that arises from the discretization of (1.1). In fact, if the spatial variations for (1.1) are in only two dimensions, we get a square grid of resistances, all having the same value, with branches connecting the nodes of the grid to a common ground; the resistances of the latter branches represent the quantity τD. This is illustrated in Fig. 1, where a and the c_k denote conductances. For constant doping, the c_k are all the same; otherwise, they vary. The h_k are current sources representing the

* Received by the editors October 14, 1980, and in revised form July 30, 1981. This work was supported by the Air Force Office of Scientific Research under grant AFOSR 80-0205.

† Department of Electrical Engineering, State University of New York at Stony Brook, Stony Brook, New York 11794.

electrical excitation of the semiconductor at its surface. In the case of three spatial dimensions, we get the same configuration except that we now have a cubic grid.

This is the motivation for the problem attacked in this work. We wish to determine the currents and voltages in a semi-infinite grounded grid where the grid's resistances are allowed to vary with distance from its input section. We even allow resistances to vary in a certain restricted fashion for spatial displacements that remain equidistant from the input section. Our analysis is immediately extendable to far more complicated grids than the prototypical square or cubic grids mentioned above. Herein, we allow this generality.

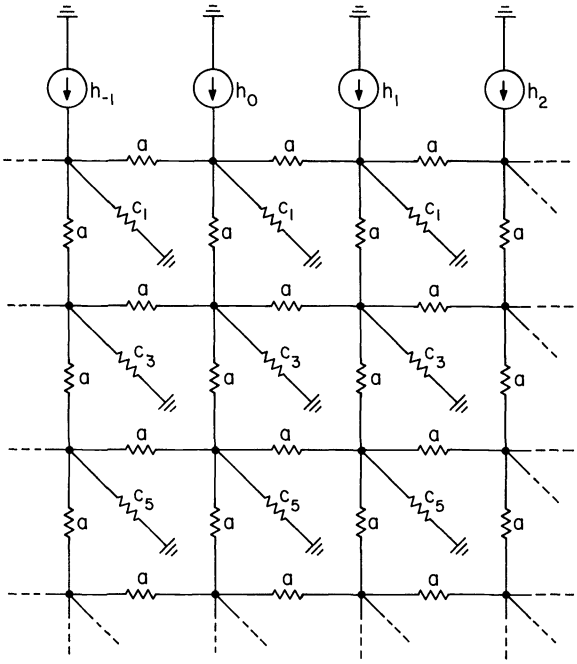

FIG 1.

The basis of our computational method is the theory of infinite continued fractions whose elements are positive operators on a Hilbert space. Those operators represent the admittances and impedances of ∞-ports consisting of sections of the grid lying parallel to the input section. The ∞-ports are connected together to make a semi-infinite ladder whose input impedance is the aforementioned continued fraction. Our analysis of the ladder network yields that unique set of voltages and currents for which the total power dissipated in the network is finite. By using the theory of Laurent operators, we also obtain a computational procedure for calculating the currents and voltages in the original grid.

All of this is extendable to grounded grids whose branches are positive-real impedances. We end this paper by indicating how the transient responses of such impedance grids can be computed. The solution we now obtain is characterized by a finite-power condition applied this time to points on the real positive axis of the complex-frequency domain.

Before proceeding, let us explain some of the notation we will be using. If H is a Hilbert space, $[H; H]$ denotes the Banach space of bounded linear operators that map H into H. By an "operator" we will always mean a member of $[H; H]$ for some H. The

symbol 1 is used to denote either the number one, or a function whose range is the singleton $\{1\}$, or the identity operator in $[H;H]$. Which meaning that symbol has in a particular case will either be stated or will be clear from the context in which it is used. If A is an operator, $W(A)$ denotes the numerical range of A:

$$W(A)=\{(Ax,x):x\in H,\ \|x\|=1\},$$

where (α,β) is the inner product of the elements α and β in H. The symbol (α,β) will also be used to denote an open interval between the real numbers α and β; once again, which meaning (α,β) has in particular cases will be either clear or specified. The symbols $[\alpha,\beta]$, $[\alpha,\beta)$, and $(\alpha,\beta]$ denote closed and semiclosed intervals with the endpoints α and β.

2. Semi-infinite grounded grids. The type of grounded grid we shall examine is indicated symbolically in Fig. 2. We have a sequence of infinite networks, which for the sake of illustration we indicate as being contained in a sequence of hypothetical boxes. We number these boxes by $k=1,3,5,\cdots$. We have shown only three nodes in each box, but it is understood that each box contains an infinity of them. The following is assumed.

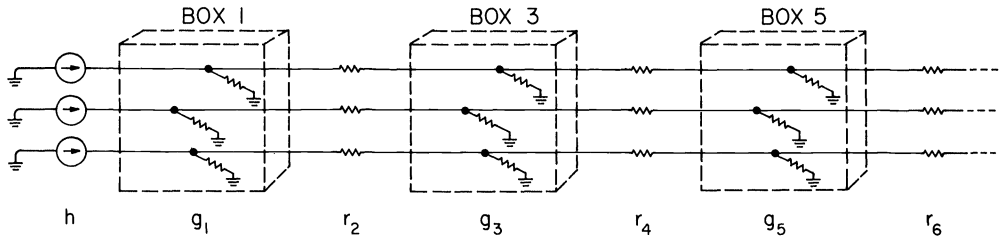

FIG 2.

Rule I. Every node is connected to a ground node through a positive conductance whose value c_k is the same for all the nodes in a particular box. The c_k can vary from box to box, that is, as k varies.

The nodes of a given box are connected together by conductances, which we have not shown in Fig. 1 so as not to clutter up the diagram. We assume that the graph of these interconnections within each box is isomorphic (in a graph-theoretical sense) to a uniform structure S_k, which we specify in Rule II. S_k need not be the same for every box. Let n be a positive integer (possibly greater than three) and let R^n denote real Euclidean n-space. The lattice points of R^n are the n-tuples $p=(p_1,\cdots,p_n)$, where each p_i is an integer.

Rule II. The nodes of each S_k occur at all the lattice points p of R^n; n is the same for every S_k. We number the nodes by their lattice numbers p. The degrees of the nodes of a particular S_k are finite and all the same, but those degrees can vary as k varies. Every branch of S_k is a positive conductance. Moreover, in a given S_k, if node p is connected to node q through a branch of conductance a, then every node j is connected to node $j+q-p$ through a branch with the same conductance a.

Thus, when $n=1$ and S_k is connected, S_k is simply a series of connections of conductances a that extends to infinity in both directions. When $n=2$, an infinity of possibilities arises. One of them is shown in Fig. 3, wherein a_1, a_2 and a_3 denote conductance values. Still more variety in possible configurations for the S_k arises as n increases beyond 2. Rule II implies that all the branches of S_k can be partitioned into a finite number of classes such that two branches are in the same class if and only if they are *parallel*, that is, if and only if the difference between the incident-node numbers of

one branch is equal to or the negative of the difference in the incident-node numbers of the other branch. We denote these classes of branches in S_k by $\Gamma_{k\mu}$, where $\mu=1,\cdots,j_k$. The single conductance value for all the branches in a given $\Gamma_{k\mu}$ is denoted by $a_{k\mu}$.

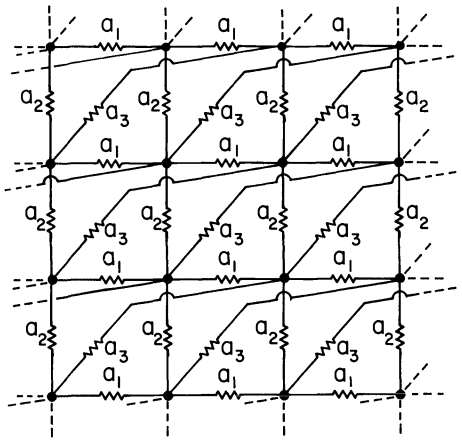

FIG 3.

Referring to Fig. 2, we impose still another condition.

Rule III. The nodes of box k are connected to the nodes of box $k+2$ in the following fashion. A node of box k is adjacent to a node of box $k+2$ if and only if the nodes have the same lattice number. Moreover, the branches connecting two consecutive boxes are all purely resistive and have the same resistance value, but that value is allowed to change as k changes. Furthermore, current generators are connected from ground to the nodes of the first box; these current generators are not in general the same.

Note that, under the three rules, Fig. 1 is a special case of Fig. 2.

3. Existence and uniqueness of solutions. We wish to examine the solutions of the countably infinite electrical networks satisfying the above three rules. By a *solution* we mean a set of branch currents and branch voltages that satisfy Kirchhoff's node and loop laws and Ohm's law. However, such networks have in general an infinity of solutions [14]. This is because power can be injected into the network from infinity. On the other hand, practical considerations (i.e., there is really no such thing as an infinite network—the idea is simply a mathematical convenience) dictate that the "natural" solutions are those that obtain their power only from the sources within the network. But, a particular infinite network may even have an infinity of natural solutions; see [15]. In this section, we shall impose conditions on our network that insure the existence of one and only one natural solution.

For a subsequent purpose, we shall allow our branch conductances to be operators on a certain Hilbert space. In particular, let H_r be any real Hilbert space. $l_2(H_r)$ will denote the real Hilbert space of vectors

$$x=[x_1,x_2,x_3,\cdots]^T,$$

where every element x_m is a member of H_r, the superscript T denotes matrix transpose, and

$$\|x\|=\left[\sum_{m=1}^{\infty}\|x_m\|^2\right]^{1/2}<\infty.$$

The inner product of two members a and b of $l_2(H_r)$ is

$$(x,y)=\sum_{m=1}^{\infty}(x_m,y_m).$$

Here, $\|x_m\|$ and (x_m,y_m) are of course the norm and inner product in H_r.

Another set of conditions we will employ are the following.

Conditions A. The currents and voltages of the network are members of H_r. Each branch is a parallel connection of a (possibly zero) current source $h\in H_r$ and a conductance g which is a positive invertible operator mapping H_r into H_r. There are no other current sources and no voltage sources. (Actually, voltage sources can be incorporated by making a Thevenin-to-Norton transformation.) The numerical ranges of all the conductance operators are uniformly contained in a fixed compact subinterval of the open half-axis $(0,\infty)$. The current sources (with any appropriate indexing) comprise a vector in $l_2(H_r)$.

In [17, Thm. 2.2] we proved the following theorem. It was established by modifying the circle of ideas concerning infinite electrical networks first introduced by Flanders [3].

THEOREM 3.1. *Let N be a connected infinite electrical network which is locally finite except possibly for one ground node; the ground node may be of infinite degree. Assume N satisfies Conditions* A. *Then, there exist a unique vector $v\in l_2(H_r)$ of branch voltages and a unique vector $i\in l_2(H_r)$ of branch currents such that Kirchhoff's node and loop laws and Ohm's law are satisfied.*

(When the ground node has infinite degree, it is not required to satisfy Kirchhoff's node law, that law being an assertion only about nodes of finite degree [12, p. 275].)

This theorem may be applied to any network satisfying Rules I through III, where, now, H_r is the real line, so long as the current-source values at the left-hand side of Fig. 2 are quadratically summable and all conductance values are contained in a compact subinterval of the open half-axis $(0,\infty)$.

4. ∞-ports and Laurent operators. The network in any box of Fig. 2 can be viewed as a grounded ∞-port, where the two terminals of each port are the ground node and one of the nodes within the box. In order to make use of Theorem 3.1., we shall restrict the voltage and current vectors of these ∞-ports to the Hilbert coordinate space $l_{2r}=l_2(R^1)$ but will alter the indexing of the components of any vector in l_{2r} to conform with Rule II. Let N^n denote the set of lattice points in R^n; that is, each member of N^n is an ordered n-tuple $p=(p_1,\cdots,p_n)$ whose entries are integers. A member of l_{2r} will now be an n-dimensional array $\{a_p : p\in N^n\}$ of real numbers a_p such that

$$\sum_{p\in N^n} a_p^2<\infty.$$

Thus, the inner product of two members $a=\{a_p\}$ and $b=\{b_p\}$ in l_{2r} is the n-tuple infinite series

$$(a,b)=\sum_{p\in N^n} a_p b_p.$$

A bounded linear mapping F of l_{2r} into l_{2r} has a matrix-like representation, but it should be borne in mind that its matrix $[F_{p,q}]$, where $p,q\in N^n$, is a $2n$-dimensional

array of real numbers. Thus, if $y=Fx$, where $y=(y_1,\cdots,y_n)\in l_{2r}$ and $x=(x_1,\cdots,x_n)\in l_{2r}$, then

$$y_p=\sum_{q\in N^n} F_{p,q}x_q.$$

Of course, not all $2n$-dimensional arrays of real numbers will represent bounded linear mappings of l_{2r} into l_{2r} [5, p. 126], but those we encounter below will do so.

Now consider the kth box of Fig. 2. As an ∞-port, it has a conductance operator whose matrix representation can be determined by making a nodal analysis. g_k has the structure of a Laurent matrix [1]; that is, upon letting $(g_k)_{p,q}$ denote the p,q entry of the matrix representation for g_k, we have for every $p,q,m\in N^n$

$$(g_k)_{p,q}=(g_k)_{p+m,q+m}. \tag{4.1}$$

This is an immediate consequence of Rules I and II.

Moreover, g_k truly is a bounded mapping of l_{2r} into l_{2r}. Indeed, for $x=\{x_q\}\in l_{2r}$, we may write the following, where every summation is understood to be over N^n.

$$\|g_kx\|^2=\sum_p\left|\sum_q(g_k)_{p,q}x_q\right|^2.$$

By virtue of Rule II, for each fixed p one finds only a finite number, say ν, of nonzero $(g_k)_{p,q}$ as q traverses N^n. Moreover, in view of (4.1), the same values appear whatever p is; the values merely shift their indices as p changes. Let M be a bound on those values. By applying Schwarz's inequality to the inner summation of the last expression and taking into account all the zero values of $(g_k)_{p,q}$, we get

$$\|g_kx\|^2\le M^2\nu^2\sum_m|x_m|^2=M^2\nu^2\|x\|^2.$$

This verifies our assertion.

A Laurent operator is a member of $[l_{2r};l_{2r}]$ that satisfies (4.1). We have proven that g_k is a Laurent operator.

Moreover, we can show that each g_k is positive and invertible by examining its numerical range. For any $x\in l_{2r}$,

$$(g_kx,x)=\sum_p\left[\sum_q(g_k)_{p,q}x_q\right]x_p. \tag{4.2}$$

By using the aforementioned properties of $(g_k)_{p,q}$, it is not difficult to see that the right-hand side converges absolutely and therefore can be rearranged. According to Rule I, the branch connecting node p to ground has conductance $c_k>0$. It therefore introduces the term $c_kx_p^2$ into the summation in (4.2). Now, consider any branch that is not incident to the ground node. Assume that it connects node p to node q and that its conductance is $a>0$. That branch introduces the following terms into the summation (4.2):

$$ax_p^2-2ax_px_q+ax_q^2=a(x_p-x_q)^2\ge0.$$

Now, we can partition all the branches that are not incident to ground (that is, all the branches in S_k) into a finite number of classes $\Gamma_{k\mu}$, where $\mu=1,2,\cdots,j_k$, as was explained in §2. The branches of any class all have the same conductance, say, $a_{k\mu}$. Thus, (4.2) can be rearranged into the following expression.

$$(g_k x, x) = c_k \sum_p x_p^2 + \sum_{\mu=1}^{j_k} a_{k\mu} \sum_{b \in \Gamma_{k\mu}} (x_{p_b} - x_{q_b})^2, \tag{4.3}$$

where p_b and q_b are the indices of the nodes incident to the branch b in class $\Gamma_{k\mu}$. By Rules I and IV, $c_k > 0$ and $a_{k\mu} > 0$. Hence,

$$(g_k x, x) \geq c_k \|x\|^2.$$

This proves that g_k is positive and invertible.

Actually, the branches connecting two consecutive boxes also comprise an ∞-port. We take the two nodes of each such branch as one of the ports of the ∞-port, and we number those ports in the same way as the nodes which they connect. According to Rule III, all those branches have the same positive resistance b_k (k is now even). Therefore, the ∞-port has the resistance operator $r_k = b_k 1$, where 1 is the identity operator on l_{2r}; that is, the element $(r_k)_{p,q}$, where $p, q \in N^n$, of r_k's matrix representation is

$$(r_k)_{p,q} = \begin{cases} b_k & \text{for } p = q, \\ 0 & \text{for } p \neq q. \end{cases}$$

Thus, each r_k is a positive invertible Laurent operator too.

5. A ladder network of operators. Because of the grounded nature of the g_k ∞-ports (k odd) and the disconnected form of the r_k ∞-ports (k even), we can connect them into the infinite ladder network of Fig. 4 without violating the port conditions. We shall analyze the network of Fig. 2 by a two-step procedure consisting of an analysis of Fig. 4, in which the individual ∞-port currents and ∞-port voltages are vectors in l_{2r}, followed by a determination of the interior branch currents and voltages of each ∞-port to get the branch currents and voltages of Fig. 2. To do so, we shall impose two further assumptions on Fig. 2.

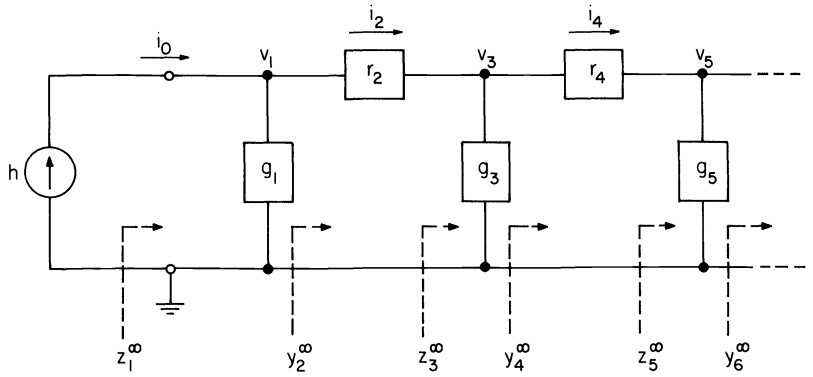

FIG 4.

Rule IV. (i) The vector h of current-source values h_p, $p \in N^n$, at the input of Fig. 2 is a member of l_{2r}.

(ii) There exist two real numbers α and γ with $0 < \alpha < \gamma < \infty$ such that the conductances-to-ground c_k satisfy $\alpha \leq c_k \leq \gamma$ for all odd k, the conductances $a_{k\mu}$ of the branches inside each box that are not incident to the ground node satisfy

$$\sum_{\mu=1}^{j_k} a_{k\mu} \leq \gamma$$

for all odd k, and the resistances b_k of the branches between the boxes satisfy $\alpha \le b_k \le \gamma$ for all even k.

We now show that Rule IV(ii) insures that the numerical ranges of all the operators g_k and r_k are uniformly bounded according to

$$W(g_k) \subset [\alpha, \beta], \quad k \text{ odd}, \qquad W(r_k) \subset [\alpha, \beta], \quad k \text{ even}, \tag{5.1}$$

where α is the constant in Rule IV and $0 < \alpha < \beta < \infty$. Since in this case we will also have $W(r_k^{-1}) \subset [\beta^{-1}, \alpha^{-1}]$, the assertion in Conditions A concerning the numerical ranges will be satisfied.

For the g_k we can argue from (4.3) as follows. Since

$$(x_{p_b} - x_{q_b})^2 \le 2x_{p_b}^2 + 2x_{q_b}^2,$$

we have

$$(g_k x, x) \le c_k \|x\|^2 + \sum_{\mu=1}^{j_k} a_{k\mu} \sum_{b \in \Gamma_{k\mu}} \left(2x_{p_b}^2 + 2x_{q_b}^2\right). \tag{5.2}$$

But, by Rule II, all the node voltages are traversed by x_{p_b} and by x_{q_b} as b traverses $\Gamma_{k\mu}$. Therefore, the right-hand side of (5.2) is equal to

$$c_k \|x\|^2 + 4\|x\|^2 \sum_{\mu=1}^{j_k} a_{k\mu}.$$

So, our assertion for the g_k follows when we set $\beta = 5\gamma$ and then invoke Rule IV(ii).

Since g_k is a strictly positive operator, this result on its numerical range also implies ([4, p. 62], [6, p. 145]) that for all k

$$\|g_k\| \le \beta, \qquad \|g_k^{-1}\| \le \alpha^{-1}. \tag{5.3}$$

The same conclusions for the r_k follow immediately from Rule IV(ii) since $r_k = b_k 1$, where now 1 denotes the identity operator on l_{2r}.

Now, refer to Fig. 4 again. The next thing we want to show is that the driving-point impedances z_k^∞, where k is odd, and the driving-point admittances y_k^∞, where k is even, exist and are positive invertible Laurent operators on l_{2r}. For $n > k$, we let z_k^n and y_k^n be the corresponding driving-point impedances and admittances when the ladder network is terminated at its nth element. To treat the driving-point impedances and admittances simultaneously, we introduce the immittance notation:

$$f_j = \begin{cases} g_j, & j = 1,3,5,\cdots, \\ r_j, & j = 2,4,6,\cdots \end{cases}$$

and

$$f_k^n = \begin{cases} z_k^n, & k = 1,3,5,\cdots, \quad n > k, \\ y_k^n, & k = 2,4,6,\cdots, \quad n > k. \end{cases} \tag{5.4}$$

The driving-point immittances of a finite (or infinite) ladder network are given by a finite (or infinite) continued fraction. That is, for $n < \infty$,

$$f_k^n = \frac{1}{f_k} + \frac{1}{f_{k+1}} + \cdots + \frac{1}{f_n}. \tag{5.5}$$

The inverse of a positive, invertible, Laurent operator and the sum of two such operators are again positive, invertible and Laurent. Therefore, every f_k^n also has these properties.

Now, Laurent operators commute [1]. This fact coupled with the fact that the numerical ranges of all the g_j and r_j are all contained in the interval $[\alpha, \beta]$, where $\alpha > 0$, allows us to invoke a theorem of Fair [2] to conclude that, as $n \to \infty$, (5.5) converges in the uniform operator topology. Its limit f_k^∞ is the driving-point impedance z_k^∞ or the driving-point admittance y_k^∞ depending on whether k is odd or even.

For $n > k+2$

$$f_k^n = \cfrac{1}{f_k + \cfrac{1}{f_{k+1} + f_{k+2}^n}}.$$

Since $W(f_k)$ and $W(f_{k+1})$ are both contained in $[\alpha, \beta]$, where $0 < \alpha < \beta < \infty$, and since f_{k+2}^n is a positive operator, we may invoke the spectral mapping theorem to write the following set inclusions, where the right-hand sides denote closed or half-closed intervals.

$$W(f_{k+1} + f_{k+2}^n) \subset [\alpha, \infty),$$

$$W\left(\frac{1}{f_{k+1} + f_{k+2}^n}\right) \subset \left(0, \frac{1}{\alpha}\right],$$

$$W\left(f_k + \frac{1}{f_{k+1} + f_{k+2}^n}\right) \subset \left[\alpha, \beta + \frac{1}{\alpha}\right],$$

$$W(f_k^n) \subset \left[\frac{1}{\beta + \frac{1}{\alpha}}, \frac{1}{\alpha}\right].$$

Since, for every $x \in l_{2r}$, $(f_k^n x, x) \to (f_k^\infty x, x)$ as $n \to \infty$, we can conclude that

$$W(f_k^\infty) \subset \left[\frac{1}{\beta + \frac{1}{\alpha}}, \frac{1}{\alpha}\right], \qquad 0 < \alpha < \beta < \infty. \tag{5.6}$$

In fact, we have established most of:

THEOREM 5.1. *Assume Rules* I *through* IV *for the grounded grid of Fig.* 2. *Then, the driving-point impedances and admittances* f_k^∞ *of the corresponding ladder network of operators shown in Fig.* 4 *exist as the limits under the uniform operator topology of the infinite continued fractions*

$$f_k^\infty = \cfrac{1}{f_k + \cfrac{1}{f_{k+1} + \cfrac{1}{f_{k+2} + \cdots}}}. \tag{5.7}$$

The f_k^∞ *are all positive, invertible, Laurent operators whose numerical ranges are uniformly bounded according to* (5.6).

Proof. There is only one thing left to prove, namely that each f_k^∞ is a Laurent operator. An operator in $[l_{2r}; l_{2r}]$ is Laurent if and only if it commutes with the shifting operator s_q, for every $q \in N^n$ [1, Thm. 2]. s_q is defined as follows: Let $x \in l_{2r}$ and for $p \in N^n$ let x_p be the pth element of x. Then, by definition, $s_q x = y$, where $y_p = x_{p-q}$.

Since each f_k^n is a Laurent operator, we have under the uniform operator topology that as $n \to \infty$

$$s_q f_k^\infty \leftarrow s_q f_k^n = f_k^n s_q \to f_k^\infty s_q.$$

This completes the proof.

6. The solution of the ladder network. Assume that the current source h of Fig. 4 is a member of l_{2r}. We now apply Kirchhoff's laws and Ohm's law to determine the l_{2r}-valued currents and voltages in Fig. 4.

For $k=0,2,4,\cdots$, Kirchhoff's node law applied to the nodes of Fig. 4 yields

$$i_{k+2} = i_k - g_{k+1} v_{k+1}.$$

By Ohm's law, $v_{k+1} = z_{k+1}^\infty i_k$. Therefore,

$$i_{k+2} = \theta_k i_k, \tag{6.1}$$

where

$$\theta_k = 1 - g_{k+1} z_{k+1}^\infty. \tag{6.2}$$

Here, 1 denotes the identity operator on l_{2r}.

For $k=1,3,5,\cdots$, Kirchhoff's loop law applied to the meshes of Fig. 4 and Ohm's law yield

$$v_{k+2} = v_k - r_{k+1} i_{k+1}, \qquad i_{k+1} = y_{k+1}^\infty v_k$$

and thus

$$v_{k+2} = \theta_k v_k, \tag{6.3}$$

where

$$\theta_k = 1 - r_{k+1} y_{k+1}^\infty. \tag{6.4}$$

Given $h \in l_{2r}$ in Fig. 4, these equations allow us to determine every voltage and every current in that ladder network recursively. In particular, for $k=2,4,6,\cdots$

$$i_k = \theta_{k-2}\theta_{k-4}\cdots\theta_0 h \tag{6.5}$$

and, for $k=3,5,7,\cdots$

$$v_k = \theta_{k-2}\theta_{k-4}\cdots\theta_1 v_1, \qquad v_1 = z_1^\infty h. \tag{6.6}$$

Our next objective is to establish several properties of the operator θ_k. By Theorem 5.1, z_{k+1}^∞ (k even) and y_{k+1}^∞ (k odd) are Laurent operators. So too are g_{k+1} (k even) and r_{k+1} (k odd). Furthermore, the composition and sum of two Laurent operators are also Laurent operators. Hence, θ_k is a Laurent operator for every k.

LEMMA 6.1. *If $A, B \in [H;H]$, where H is a Hilbert space, and if A is positive and commutes with B, then their numerical ranges satisfy $W(AB) \subset W(A)W(B)$.*

Proof. The square root $A^{1/2}$ commutes with every operator that commutes with A. Therefore,

$$(ABx,x) = (BA^{1/2}x, A^{1/2}x) \in W(B)\|A^{1/2}x\|^2 = W(B)(Ax,x).$$

Our lemma now follows immediately.

We now examine some numerical ranges. Let $k=0,2,4,\cdots$. According to (5.6) and (5.1),

$$W(y_{k+2}^\infty) \subset \left[\frac{\alpha}{1+\alpha\beta}, \frac{1}{\alpha}\right], \qquad W(g_{k+1}) \subset [\alpha,\beta].$$

Therefore,

$$W(g_{k+1}^{-1}) \subset \left[\frac{1}{\beta}, \frac{1}{\alpha}\right].$$

Since we are dealing with positive Laurent operators and Laurent operators commute, we can invoke Lemma 6.1.

$$W(y_{k+2}^{\infty} g_{k+1}^{-1}) \subset \left[\frac{\alpha}{\beta+\alpha\beta^2}, \frac{1}{\alpha^2}\right].$$

It can be seen from Fig. 4 that

$$z_{k+1}^{\infty} = (g_{k+1} + y_{k+2}^{\infty})^{-1} = g_{k+1}^{-1}(1 + y_{k+2}^{\infty} g_{k+1}^{-1})^{-1}. \tag{6.7}$$

Therefore,

$$g_{k+1} z_{k+1}^{\infty} = (1 + y_{k+2}^{\infty} g_{k+1}^{-1})^{-1}.$$

So,

$$W(g_{k+1} z_{k+1}^{\infty}) \subset \left[\left(1+\frac{1}{\alpha^2}\right)^{-1}, \left(1+\frac{\alpha}{\beta+\alpha\beta^2}\right)^{-1}\right].$$

The last closed interval is contained in the open interval $(0, 1)$. So, in view of (6.2),

$$\begin{aligned} W(\theta_k) &\subset \left[1-\left(1+\frac{\alpha}{\beta+\alpha\beta^2}\right)^{-1}, 1-\left(1+\frac{1}{\alpha^2}\right)^{-1}\right] \\ &= \left[\frac{1}{1+\alpha^{-1}\beta+\beta^2}, \frac{1}{1+\alpha^2}\right]. \end{aligned} \tag{6.8}$$

This shows that θ_k is a positive, invertible, strictly contractive operator with

$$\|\theta_k\| \leq \frac{1}{1+\alpha^2}. \tag{6.9}$$

For $k = 1, 3, 5, \cdots$, (6.7) is replaced by

$$y_{k+1}^{\infty} = (r_{k+1} + z_{k+2}^{\infty})^{-1}.$$

We can now apply the same argument to (6.4) to obtain (6.8) and (6.9) once again. This establishes:

THEOREM 6.1. *Under Rules* I *through* IV *and for every* $k = 0, 1, 2, \cdots$, θ_k *is a positive, invertible, Laurent operator in* $[l_{2r}; l_{2r}]$ *satisfying* (6.9).

Note that, by Rule IV(ii), α is independent of k.

We can now show that the solution given by (6.5) and (6.6) is precisely the one dictated by Theorem 3.1. Indeed, let the H_r of that theorem be l_{2r}. Since there is only one current source, the vector of current sources is a member of $l_2(l_{2r})$. We have already noted in §5 that $W(g_k) \subset [\alpha, \beta]$ and $W(r_k^{-1}) \subset [\beta^{-1}, \alpha^{-1}]$. By the analysis in the second and third paragraphs of this section, the solution given by (6.5) and (6.6) satisfies Kirchhoff's laws and Ohm's law. The rest of the hypothesis of Theorem 3.1 is clearly satisfied except perhaps for the requirements that the vector of all branch voltages and the vector of all branch currents be members of $l_2(l_{2r})$; this we now verify.

Summing over all odd k and using (6.6), we get for the vertical branches of Fig. 4

$$\sum \|v_k\|^2 = \|v_1\|^2 + \|\theta_1 v_1\|^2 + \|\theta_3\theta_1 v_1\|^2 + \|\theta_5\theta_3\theta_1 v_1\|^2 + \cdots.$$

By (6.9), $\|\theta_k\| \le K = (1+\alpha^2)^{-1} < 1$ for every odd k. Therefore,

$$\sum \|v_k\|^2 \le (1+K^2+K^4+K^6+\cdots)\|v_1\|^2$$
$$= (1-K^2)^{-1}\|v_1\|^2 < \infty.$$

For the horizontal branches of Fig. 4, we have $v_k = r_k i_k$, where now k is even. Summing over all such k and using (5.1) and (6.5), we get

$$\sum \|v_k\|^2 = \sum \|r_k i_k\|^2 \le \beta^2 \sum \|i_k\|^2$$
$$= (\|h\|^2 + \|\theta_0 h\|^2 + \|\theta_2\theta_0 h\|^2 + \|\theta_4\theta_2\theta_0 h\|^2 + \cdots)\beta^2$$
$$\le (1+K^2+K^4+K^6+\cdots)\beta^2\|h\|^2 = (1-K^2)^{-1}\beta^2\|h\|^2 < \infty.$$

So truly, the vector of all branch voltages is a member of $l_2(l_{2r})$.

Quite the same argument shows that the vector of all branch currents is also a member of $l_2(l_{2r})$. Thus, we have:

THEOREM 6.2. *Under Rules* I *through* IV, *the solution for the network of Fig.* 4 *given by* (6.5) *and* (6.6) *is the unique* (*finite power*) *solution dictated by Theorem* 3.1.

As was promised at the beginning of §5 we now have a two-step procedure for determining the solution for the grid of Fig. 2. We first determine the solution for the operator network of Fig. 4 and then determine the interior branch currents of each ∞-port to get the currents in the branches of Fig. 2. However, there is one more thing we should verify; namely, the solution for the grid of Fig. 2 given by this two-step procedure is the same as the solution specified by Theorem 3.1 when that theorem is applied directly to the grid of Fig. 2 with H_r now being the real line. This can be established in a completely straigtforward way. The details of the argument are spelled out in [18, §5].

7. A computational procedure. So far, we have established the existence and uniqueness of the solution in l_{2r} (i.e., the finite-power solution) for the network of Fig. 2. However, the question remains how one might compute the numerical values of the voltages and currents in the network, given the current sources h_j of Fig. 2. For this purpose, we use (6.6) to compute all the node voltages, from which all the currents in the grid can be determined. The first step is to determine $v_1 \in l_{2r}$, and this is facilitated by the isomorphism between l_2 and the corresponding space of Fourier series. Let's quickly survey that isomorphism and its effect on Laurent operators [1].

Let S denote the unit circle and S^n the Cartesian product of n replicates of S. $L_2(S^n)$ is as usual the Hilbert space of (equivalence classes of) real or complex, square integrable functions f on S^n with the norm

$$\|f\| = \left[\frac{1}{(2\pi)^n}\int_{S^n} |f(\omega)|^2 d\omega\right]^{1/2}, \qquad \omega = (\omega_1,\cdots,\omega_n), \quad 0 \le \omega_j < 2\pi.$$

On the other hand, l_2 denotes the complexification of l_{2r} [11, p. 137]. Thus, l_{2r} is a subset of l_2. Let $\mathfrak{F}$ denote the transformation that assigns to each $x = \{x_p : p \in N^n\} \in l_2$ the function

$$\tilde{x}(\omega) = \sum_{p\in N^n} x_p e^{i(p,\omega)},$$

where $(p,\omega) = p_1\omega_1 + \cdots + p_n\omega_n$. A standard result of Fourier series theory is that $\mathfrak{F}$ is a topological linear isomorphism from l_2 onto $L_2(S^n)$ such that $\|x\| = \|\tilde{x}\|$.

Let $z \in [l_{2r}; l_{2r}]$ and let its natural extension onto l_2 also be denoted by z [11, p. 138]. Also, let $\hat{z}$ be the mapping of $L_2(S^n)$ into $L_2(S^n)$ induced by $\mathcal{F}$; that is, $\hat{z} = \mathcal{F} z \mathcal{F}^{-1}$. Then, $\|z\| = \|\hat{z}\|$. It is a fact that z is a Laurent operator (that is, $z \in [l_{2r}; l_{2r}]$ and satisfies (4.1)) if and only if $\hat{z}$ is a multiplication. More specifically,

$$(\hat{z}\tilde{x})(\omega) = \tilde{z}(\omega)\tilde{x}(\omega), \tag{7.1}$$

where

$$\tilde{z}(\omega) = \sum_{q \in N^n} z_{0,q} e^{-i(q,\omega)}. \tag{7.2}$$

Here, the subscript 0 denotes the origin in N^n and $z_{0,q}$ is the $0,q$ entry of the matrix representation $[z_{p,q}]$ of z.

The mapping $z \mapsto \hat{z}$ of $[l_2; l_2]$ into $[L_2(S^n); L_2(S^n)]$ is linear, continuous and norm preserving; that is, $\|z\| = \|\hat{z}\|$. Also,

$$\|z\| = \operatorname{ess\,sup} |\tilde{z}(\omega)|. \tag{7.3}$$

Moreover, if z is positive and invertible, then $\tilde{z}(\omega)$ is real-valued, $\operatorname{ess\,inf} \tilde{z}(\omega) > 0$, and z^{-1} corresponds to multiplication by $[\tilde{z}(\omega)]^{-1}$; also, the numerical range of $\hat{z}$ is the closed interval between the essential supremum and the essential infimum of $\tilde{z}(\omega)$.

These results imply that z_1^∞, which exists as a Laurent operator according to Theorem 5.1, corresponds to multiplication by the function

$$\tilde{z}_1^\infty(\omega) = \frac{1}{\tilde{g}_1(\omega) + \frac{1}{b_2 + \frac{1}{\tilde{g}_3(\omega) + \frac{1}{b_4 + \cdots}}}}, \tag{7.4}$$

where for k odd $\tilde{g}_k(\omega)\cdot$ is the multiplication corresponding to g_k and for k even $b_k \cdot$ is the multiplication corresponding to $r_k = b_k 1$. By virtue of Rule II, each $\tilde{g}_k(\omega)$ is a finite Fourier series and hence a continuous function. Also, the range of $\tilde{g}_k(\omega)$ is contained in $[\alpha, \beta]$ where $\alpha > 0$.

The function in $L_2(S^n)$ corresponding to $v_1 = z_1^\infty h$ for a given $h = \{h_p : p \in N^n\} \in l_{2r}$ is

$$\tilde{v}_1(\omega) = \tilde{z}_1^\infty(\omega)\tilde{h}(\omega), \tag{7.5}$$

where

$$\tilde{h}(\omega) = \sum_{p \in N^n} h_p e^{i(p,\omega)}. \tag{7.6}$$

Thus, the node voltages $(v_1)_p$ for the nodes on the first box of Fig. 2 are

$$(v_1)_p = \frac{1}{(2\pi)^n} \int_{S^n} \tilde{v}_1(\omega) e^{-i(p,\omega)} d\omega. \tag{7.7}$$

The next step is the computation of the functions $\tilde{\theta}_k(\omega)$ for k odd by means of the following analogue to (6.4):

$$\tilde{\theta}_k(\omega) = 1 - \tilde{r}_{k+1}(\omega)\tilde{y}_{k+1}^\infty(\omega). \tag{7.8}$$

Here, $\tilde{r}_{k+1}(\omega) = b_{k+1}$, and

$$\tilde{y}_{k+1}^\infty(\omega) = \frac{1}{b_{k+1} + \frac{1}{\tilde{g}_{k+2}(\omega) + \frac{1}{b_{k+3} + \frac{1}{\tilde{g}_{k+4}(\omega) + \cdots}}}}. \tag{7.9}$$

The analogue to (6.6) then yields

$$\tilde{v}_k(\omega) = \tilde{\theta}_{k-2}(\omega)\tilde{\theta}_{k-4}(\omega) \cdots \tilde{\theta}_1(\omega)\tilde{v}_1(\omega). \tag{7.10}$$

Finally, the node voltages in the kth box of Fig. 2 are given by

$$(v_k)_p=\frac{1}{(2\pi)^n}\int_{S^n}\tilde{v}_k(\omega)e^{-i(p,\omega)}d\omega. \tag{7.11}$$

In practical computations we must either determine the continued fraction (7.4) in closed form, usually an unlikely prospect, or truncate it by an open-circuit or short-circuit operator admittance and estimate the resulting error, or, in the case where the grid approaches a uniform grid as $k\to\infty$, perhaps truncate it with the characteristic operator admittance of the uniform grid [17].

Let's consider how that truncation error might be estimated when the grid is terminated by an open-circuit or short-circuit operator after the nth section. In fact, let us approximate $\tilde{z}_1^{\infty}(\omega)$ by

$$\tilde{z}_1^n(\omega)=\frac{1}{\tilde{g}_1(\omega)+\dfrac{1}{b_2+\cdots+\dfrac{1}{f_n(\omega)}}}, \tag{7.12}$$

where $f_n(\omega)=b_n$ for n even and $f_n(\omega)=\tilde{g}_n(\omega)$ for n odd. Now, a property of convergent infinite continued fractions with positive terms is that its limit lies between any two consecutive truncations. Thus, $\tilde{z}_1^{\infty}(\omega)$ lies between $\tilde{z}_1^{n-1}(\omega)$ and $\tilde{z}_1^n(\omega)$. Hence,

$$|\tilde{z}_1^{\infty}(\omega)-\tilde{z}_1^n(\omega)|\le|\tilde{z}_1^{n-1}(\omega)-\tilde{z}_1^n(\omega)|. \tag{7.13}$$

Also, by virtue of Rule II, every $\tilde{g}_k(\omega)$ and therefore every $\tilde{z}_1^n(\omega)$ is a continuous function.

This allows us to bound the error generated by truncating (7.4) as follows. Let $v_1^a\in l_{2r}$ be the approximation of $v_1\in l_{2r}$ resulting from the replacement of $\tilde{z}_1^{\infty}(\omega)$ by $\tilde{z}_1^n(\omega)$. Then, letting $(x)_p$ denote the pth component of the vector $x\in l_{2r}$ and using (7.12), we may write

$$\begin{aligned}|(v_1)_p-(v_1^a)_p|&\le\frac{1}{(2\pi)^n}\int_{S^n}|\tilde{z}_1^{\infty}(\omega)-\tilde{z}_1^n(\omega)|\,|\tilde{h}(\omega)|\,d\omega\\ &\le\sup_{\omega}|\tilde{z}_1^{n-1}(\omega)-\tilde{z}_1^n(\omega)|\sup_{\omega}|\tilde{h}(\omega)|\frac{1}{(2\pi)^n}\int_{S^n}d\omega\\ &\le\sup_{\omega}|\tilde{z}_1^{n-1}(\omega)-\tilde{z}_1^n(\omega)|\sum|h_p|.\end{aligned} \tag{7.14}$$

Here, $\Sigma|h_p|$ denotes the sum over all the nonzero $|h_p|$, these usually being finite in number in practical cases. Because of (7.3) and the fact that (5.7) converges in the uniform operator topology, given the h_p with $\Sigma|h_p|$ convergent, we can make the right-hand side of (7.14) as small as we wish by choosing n large enough. That is, (7.14) can be used to control the error generated by truncating (7.4). However, this is a conservative approach; the bound (7.14) will be in general much larger than the actual error.

Bounds on the error generated in the computation of the $(v_k)_p$ by the continued-fraction expressions for the $\tilde{y}_{k+1}^{\infty}(\omega)$ can be estimated in exactly the same way, but now an error appears for each factor $\tilde{\theta}_k(\omega)$ as well as for $\tilde{v}_1(\omega)=\tilde{z}_1^{\infty}(\omega)\tilde{h}(\omega)$ in (7.10). Finally, when our nonuniform grid approaches a uniform one as $k\to\infty$, we will generate less error by terminating in the characteristic operator immittance of the uniform grid [17], and so our aforementioned bounds on the error will still be valid. Of course, other errors are generated by the numerical integrations of (7.7) and (7.11); these can be estimated by standard methods.

8. An example. We illustrate our computational procedure with an example. We assign values to the parameters in the grid of Fig. 1 as follows:

$$h_j = \begin{cases} 1 & \text{for } j=0, \\ 0 & \text{for } j\neq 0, \end{cases}$$

$$a=1,$$

$$c_{2m+1}=1+e^{-m} \quad \text{for } m=0,1,2,\cdots.$$

Consequently, the various functions of ω generated by the isomorphism $\mathcal{F}$ are

$$\tilde{h}(\omega)=1,$$

$$\tilde{r}_{2m}(\omega)=1 \quad \text{for } m=1,2,3,\cdots,$$

$$\tilde{g}_{2m+1}(\omega)=3+e^{-m}-2\cos\omega \quad \text{for } m=0,1,2,\cdots.$$

To compute approximately the driving-point impedance $\tilde{z}_1^{\infty}(\omega)$, we use the fact that our grid approaches a uniform grid as $m\to\infty$. So, we may replace the ladder network beyond node $2M+1$, where M is chosen sufficiently large, by its characteristic impedance $\tilde{z}_0(\omega)$. The latter can be determined by the method given in [17]; it is

$$\tilde{z}_0(\omega)=\tfrac{1}{2}\left\{-3+2\cos\omega+\left[(3-2\cos\omega)^2+4(3-2\cos\omega)\right]^{1/2}\right\}.$$

Then, for sufficiently large M, we have, to a high order of accuracy,

$$\tilde{z}_1^{\infty}(\omega)\approx\cfrac{1}{\tilde{g}_1(\omega)+\cfrac{1}{1+\cfrac{1}{\tilde{g}_3(\omega)+\cfrac{1}{1+\cdots+\cfrac{1}{\tilde{g}_{2M-1}(\omega)+\cfrac{1}{1+\tilde{z}_0(\omega)}}}}}}.$$

Similarly, for k odd and $k\ll 2M+1$,

$$\tilde{y}_{k+1}^{\infty}\approx\cfrac{1}{1+\cfrac{1}{\tilde{g}_{k+2}(\omega)+\cfrac{1}{1+\cfrac{1}{\tilde{g}_{k+4}(\omega)+\cdots+\cfrac{1}{\tilde{g}_{2M-1}(\omega)+\cfrac{1}{1+\tilde{z}_0(\omega)}}}}}}.$$

We now use (7.8), (7.10) and (7.11) to compute the node voltages for the nonuniform grid in the vicinity of the single current source $h_0=1$.

For the sake of illustration, we have chosen $M=12$ and have computed the node voltages for the first five rows of nodes (that is, for the first five boxes) and for the first eight columns of nodes on either side of the 0th column where $h_0=1$ appears. The results are displayed in Table 1. Since the node values have even symmetry around the 0th column, we have indicated their values only to the right of the 0th column. Computer execution time for these results was 38.5 seconds on the UNIVAC 1100 computer.

TABLE 1

		Column							
		0	1	2	3	4	5	6	7	8
Row	1	.28076	.07717	.02217	.00661	.00203	.00064	.00020	.00007	.00002
	2	.07199	.03413	.01308	.00465	.00161	.00055	.00019	.00006	.00002
	3	.02069	.01303	.00621	.00261	.00102	.00038	.00014	.00005	.00002
	4	.00643	.00474	.00265	.00127	.00056	.00023	.00009	.00004	.00001
	5	.00211	.00170	.00107	.00057	.00028	.00012	.00005	.00002	.00001

9. Nonuniform grids of positive-real impedances. We now turn from purely resistive networks to ones that may contain inductors, capacitors, transformers and so on. By using the results of [13], we can extend our analysis to grids of the form of Fig. 2

where now each branch is a positive-real driving-point impedance with no coupling between branches. We shall in fact examine the transient behavior of such networks by using the standard result that once the behavior of the network in the complex-frequency domain is determined its time-domain behavior can be examined through the inverse Laplace transform. The arguments needed to do this are quite similar to those given in [17, §8]. Therefore, we shall at this point merely define the needed concepts and then indicate how the arguments in [17, §8] have to be modified for our present considerations.

Here are some concepts from [13]. C_+ denotes the open right half of the complex plane C

$$C_+ = \{s \in C : \operatorname{Re} s > 0\}.$$

For $s \in C_+$, Ω_s is the closed cone

$$\Omega_s = \{z \in C : |\arg z| \leq |\arg s|\},$$

where it is understood that the origin is a member of Ω_s. As before, l_2 denotes the complexification of l_{2r}. By an "operator", we henceforth mean a continuous linear mapping of l_2 into l_2.

P is the set of all analytic operator-valued functions F on C_+ such that, for every $s \in C_+$, the numerical range $W[F(s)]$ of $F(s)$ is contained in Ω_s. Thus, if $F \in P$, $F(\sigma)$ is a positive operator for each $\sigma > 0$. P_i is the set of all $F \in P$ such that, for every fixed $s \in C_+$, $W[F(s)]$ is bounded away from the origin, that is, there exists a $\delta > 0$ depending in general on s such that $\operatorname{Re}(F(s)x, x) \geq \delta \|x\|^2$ for all $x \in l_2$. Thus, for each $s \in C_+$ and $F \in P_i$, $F(s)$ is an invertible operator. It was shown in [13] that if $F \in P_i$ and $G \in P$, then $F+G \in P_i$; also, if $F \in P_i$ and if F^{-1} denotes the function $s \mapsto [F(s)]^{-1}$, then $F^{-1} \in P_i$.

Next, let $F_1, F_2, F_3, \cdots \in P_i$ and let $Z_n(s)$ be the operator-valued finite continued fraction

$$Z_n(s) = \frac{1}{F_1(s)} + \frac{1}{F_2(s)} + \cdots + \frac{1}{F_n(s)}. \tag{9.1}$$

Then, by what has just been stated, $Z_n \in P_i$. The following theorem is a somewhat simplified version of [13, Thm. 1 and Corollary 1a].

THEOREM 9.1. *Assume the following three conditions*:

(i) $F_k \in P_i$ *for every* $i = 1, 2, 3, \cdots$.

(ii) *Given any compact set* $\Xi \subset C_+$, *there exists a constant* $\delta > 0$, *depending upon* Ξ, *such that* $\inf \operatorname{Re} W[F_k(s)] > \delta$ *for* $k = 1, 2$ *and* $s \in \Xi$.

(iii) *For each* $\sigma > 0$ *and all* $k = 1, 2, 3, \cdots$, *the operators* $F_k(\sigma)$ *commute with each other and* $W[F_k(\sigma)] > \delta_k(\sigma)$, *where the* $\delta_k(\sigma)$ *are positive numbers satisfying* $\sum_{k=1}^{\infty} \delta_k(\sigma) = \infty$.

Then, for every $s \in C_+$, *the sequence* $\{Z_n(s)\}_{n=1}^{\infty}$ *converges in the uniform operator topology, and the convergence is uniform with respect to s in any compact subset of* C_+. *Moreover, the limit function* $Z = \lim_{n \to \infty} Z_n$ *is a member of* P_i.

Next, we turn to the time domain and define the space $L_2(R, l_{2r})$. Consider the mappings of the real line R into l_{2r} which are square integrable on R under the l_{2r}-norm, and let two such functions be in the same equivalence class if they differ on

no more than a set of Lebesgue measure zero. $L_2(R, l_{2r})$ is the real Hilbert space of all such equivalence classes supplied with the inner product

$$(a,b)=\int_{-\infty}^{\infty} \sum_{p\in N^n} a_p(t)b_p(t)\,dt, \qquad t\in R, \quad a,b\in L_2(R,l_{2r}).$$

(See [11, Appendix G].)

In [17], we proved:

LEMMA 9.1. *Assume that the vector h is a member of $L_2(R,l_{2r})$ and that the support of h is bounded on the left. Let H be the Laplace transform of h, that is, the vector of Laplace transforms H_p of the components $h_p, p\in N^n$. Then, for each $s\in C_+$, $H(s)$ exists and is a member of l_2. Also, for each $\sigma>0$, $H(\sigma)$ is a member of l_{2r}.*

These are all the results we need to generalize our discussion to grids of impedances. To proceed, replace Rules I through IV with the following, where $s\in C_+$.

Rule I′. Same as Rule I except that the conductances c_k are replaced by scalar positive-real admittances $C_k(s)$.

Rule II′. Same as Rule II except that the conductance $a_{k\mu}$ (k odd) for every branch class $\Gamma_{k\mu}$ is replaced by a scalar positive-real admittance $A_{k\mu}(s)$.

Rule III′. Same as Rule III except that every resistance b_k (k even) connecting box $k-1$ to box $k+1$ is replaced by a scalar positive-real impedance $B_k(s)$.

Rule IV′. (i) In the time-domain, the vector $h=\{h_p : p\in N^n\}$ of current sources at the input of Fig. 2 is a member of $L_2(R,l_{2r})$, and the support of h is bounded on the left.

(ii) There exist two continuous functions $\alpha(\sigma)$ and $\gamma(\sigma)$ on the half line $R_+=\{\sigma\in R:\sigma>0\}$ with $0<\sigma(\sigma)<\gamma(\sigma)$ for every $\sigma\in R_+$ such that, for $k=1,3,5,\cdots$, we have $\alpha(\sigma)\le C_k(\sigma)\le\gamma(\sigma)$ and

$$\sum_{\mu=1}^{j_k} A_{k\mu}(\sigma)\le\gamma(\sigma),$$

and, for every $k=2,4,6,\cdots$, we have $\alpha(\sigma)\le B_k(\sigma)\le\gamma(\sigma)$.

We now decompose the grid of Fig. 2 into ∞-ports as before to get the ladder network of Fig. 4. In the frequency domain, the g_k for k odd and the r_k for k even are replaced by $[l_2;l_2]$-valued functions $Y_k(s)$ and $Z_k(s)=B_k(s)1$ respectively, where $s\in C_+$. Upon modifying the arguments that established (4.3) with manipulations suitable for a complex Hilbert space, we obtain for $x=\{x_p : p\in N^n\}\in l_2$

$$(Y_k(s)x,x)=C_k(s)\sum_p |x_p|^2+\sum_{\mu=1}^{j_k} A_{k\mu}(s)\sum_{k\in\Gamma_{k\mu}} |x_{p_b}-x_{q_b}|^2. \tag{9.2}$$

Using this equation and Rules I′ and IV′, we can argue in virtually the same way as in [17, §8] and thus establish that every Y_k and Z_k is a member of P_i, that $Y_k(\sigma)$ and $Z_k(\sigma)$ are Laurent operators for each $\sigma\in R_+$, that the hypothesis of Theorem 9.1 is satisfied when $F_k(s)=Y_k(s)$ for k even and $F_k(s)=Z_k(s)$ for k odd, and that the driving-point impedance $Z_1^\infty=\lim_{n\to\infty} Z_1^n$ is also a member of P_i.

Proceeding still further as in [17, §8], we conclude that the Y_k, Z_k, Z_1^∞ and θ^k, where

$$\theta^k(s)=1-Y_{k+1}(s)Z_{k+1}^\infty(s), \qquad k=0,2,4,\cdots,$$
$$\theta^k(s)=1-Z_{k+1}(s)Y_{k+1}^\infty(s), \qquad k=1,3,5,\cdots,$$

are all Laplace transforms of $[l_{2r}; l_{2r}]$-valued right-sided distributions whose supports are bounded on the left at the origin. Next, we generalize Ohm's law by means of distributional convolution [11, §5.2]:

$$v = r * i, \qquad i = g * v.$$

Finally, we conclude with:

THEOREM 9.2. *For the network of Fig. 2, assume that Rules* I′ *through* IV′ *hold. Then, there exists one and only one set of right-sided Laplace-transformable distributions for the branch voltages* v_m *in the network of Fig. 2 such that Kirchhoff's node and loop laws and (the generalized) Ohm's law are satisfied in the time domain and such that, for at least one* $\sigma > 0$ *and for* V_m *denoting the Laplace transform of* v_m, *we have*

$$\sum_m [V_m(\sigma)]^2 < \infty. \tag{9.3}$$

In this case, (9.3) *holds for all* $\sigma > 0$.

The branch voltages and branch currents for Fig. 2 can be determined from the components of the l_{2r}-valued distributions i_k and v_k, which are given in turn by (6.5) and (6.6) appropriately rewritten as distributional convolutions. Alternatively, we can work in the frequency domain, in which case (6.5) and (6.6) should be rewritten as multiplications of Laplace transforms.

10. The computation of transient responses. By using the method described in §7, we can compute the voltage $V(\sigma)$ at any node and the current $I(\sigma)$ in any branch at a finite set of points $\sigma = \sigma_1, \sigma_2, \cdots, \sigma_q$ on the real axis in C_+. Then using these values of $V(\sigma)$ or $I(\sigma)$, we can apply Papoulis' method [7], [8] to compute the corresponding transient response $v(t)$ or $i(t)$. This requires however that $V(s)$ and $I(s)$ tend to zero fast enough as $s \to \infty$ in C_+ to ensure that $v(t)$ or $i(t)$ be a sufficiently smooth function to allow the convergence of Papoulis' method.

For example, assume that as $s \to \infty$ in C_+ every $C_k(s)$ acts capacitively, that is, it is asymptotic to a constant times s, and every $A_{k\mu}(s)$ and $B_k(s)$ acts resistively, that is, it is asymptotic to a constant. Assume in addition that the Laplace transform $H_p(s)$ of every current generator is of order $O(|s|^{-j})$, where j is an integer greater than one. Also, assume that only a finite number of the $H_p(s)$ are not identically equal to zero. Then, it can be shown that every current flowing between box $k-1$ and box k has continuous derivatives up to the $(j+\frac{k}{2}-2)$th derivative and that every node voltage in box k has continuous derivatives up to the $(j+\frac{k}{2}-\frac{3}{2})$th derivative [19, §X]. Thus, these transients are quite smooth, and we may apply Papoulis' method.

REFERENCES

[1] A. BROWN AND P. R. HALMOS, *Algebraic properties of Toeplitz operators*, J. für Mathematik, 213 (1963), pp. 89–102.

[2] W. FAIR, *Noncommutative continued fractions*, this Journal, 2 (1971), pp. 226–232.

[3] H. FLANDERS, *Infinite networks: I—Resistive networks*, IEEE. Trans. Circuit Theory, CT-18 (1971), pp. 326–331.

[4] P. R. HALMOS, *A Hilbert Space Problem Book*, D. Van Nostrand, Princeton, NJ, 1967.

[5] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, Macmillan, New York, 1964.

[6] L. A. LIUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Frederick Ungar, New York, 1961.

[7] A. PAPOULIS, *A new method of inversion of the Laplace transform*, Quarterly Appl. Math., 14 (1956), pp. 405–414.

[8] _____, *A different approach to the analysis of tracer data*, SIAM J. Control, 11 (1973), pp. 466–474.
[9] S. M. SZE, *Physics of Semiconductor Devices*, Wiley-Interscience, New York, 1969.
[10] A. H. ZEMANIAN, *Generalized Integral Transformations*, Wiley-Interscience, New York, 1968.
[11] _____, *Realizability Theory for Continuous Linear Systems*, Academic Press, New York, 1972.
[12] _____, *Countably infinite networks that need not be locally finite*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 274–277.
[13] _____, *Continued fractions of operator-valued analytic functions*, J. Approximation Theory, 11 (1974), pp. 319–326.
[14] _____, *Infinite electrical networks*, Proc. IEEE, 64 (1976), pp. 6–17.
[15] _____, *The connections at infinity of a countable resistive network*, Circuit Theory Appl., 3 (1975), pp. 333–337.
[16] _____, *The limb analysis of countably infinite electrical networks*, J. Combin. Theory, Ser. B, 24 (1978), pp. 76–93.
[17] _____, *The characteristic-resistance method for grounded semi-infinite grids*, this Journal, 12 (1981), pp. 115–138.
[18] _____, *The characteristic-resistance method for grounded semi-infinite grids*, College of Engineering Tech. Rep. 330, State University of New York at Stony Brook, August, 1979.
[19] _____, *Nonuniform semi-infinite grounded grids*, College of Engineering Tech. Rep. 345, State University of New York at Stony Brook, September, 1980.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0007 $01.00/0

COMPACT PERTURBATIONS OF m-ACCRETIVE OPERATORS IN GENERAL BANACH SPACES*

S. GUTMAN†

Abstract. The evolution problem $\frac{du}{dt}+Au(t)\ni Bu(t)$, $u(0)=x$, where A is a nonlinear m-accretive operator and B is a nonlinear completely continuous operator $B: C([0,T];X)\to L^p(0,T;X)$, is studied and various results on a local and global existence of solutions are established. A generalization of the Weyl–Riesz criterion for compactness of sets in the spaces $L^p(\mathbb{R},dt,X)$, $1\le p<\infty$, X is a Banach space is proved.

1. Introduction.

In this paper we study the initial value problem

$$u+Au\ni Bu, \qquad u(0)=x, \tag{1}$$

where A is a given m-accretive (possibly multivalued) operator in a real Banach space X with norm $\|\cdot\|$ and B is a continuous operator from $C([0,T];X)$ into $L^p(0,T;X)$, $1\le p\le\infty$, $0<T\le\infty$. The case in which B has some Lipschitz-like properties was studied by Crandall and Nohel [4]. The case in which the operator A generates a compact semigroup was treated by Pazy [10] in the linear case and more recently by Vrabie [13] in the nonlinear case. For a t-dependent continuous m-accretive operator $A=A(t)$ and a completely continuous operator $B(t,u):[0,T]\times X\to X$, the existence of a solution of (1) was proved in [11] and [12]. We are interested in the situation where A is a general m-accretive operator and B is a completely continuous operator (i.e., B is continuous and compact).

To formulate our main results, we recall some basic definitions and refer the reader to [1] for the relevant background.

DEFINITION. A strong solution of (1) on $[0,T]$ is a function u which is continuous on $[0,T]$, absolutely continuous on $]0,T[$, differentiable almost everywhere on $[0,T]$ and satisfies $u(0)=x$ and $u'(t)+Au(t)\ni Bu(t)$ almost everywhere on $[0,T]$.

Note that in general Banach space absolute continuity does not imply differentiability almost everywhere. Nevertheless, if X is reflexive or, more generally, if X satisfies the Radon–Nykodým property, absolute continuity implies differentiability almost everywhere. Also absolute continuity of u and differentiability almost everywhere are equivalent to $u\in W^{1,1}(0,T;X)\cap C([0,T];X)$, see [6].

Let $F: X\to 2^{X^*}$ be the duality map, that is,

$$F(x)=\{x^*\in X^*:(x,x^*)=\|x\|^2=\|x^*\|^2\}.$$

For $(x,y)\in X\times X$ define

$$\langle y,x\rangle_S=\sup\{(y,x^*):x^*\in F(x)\}$$

and consider the initial value problem

$$u'+Au\ni f, \qquad u(0)=x, \tag{2}$$

where A is an m-accretive operator, $A: X\to 2^X$, $f\in L^1(0,T;X)$ and $x\in\overline{D(A)}$.

* Received by the editors March 2, 1981, and in revised form August 25, 1981. This paper contains a part of the author's Ph.D. thesis written under the direction of Professor A. Pazy at the Hebrew University of Jerusalem.

† Institute of Mathematics, The Hebrew University of Jerusalem, Jerusalem, Israel.

DEFINITION. The function $u(t):[0,T]\to\overline{D(A)}$ is called an *integral solution* of the intitial value problem (2) if $u(t)$ is continuous on $[0,T]$, $u(0)=x$ and the inequalities

$$\frac{1}{2}\|u(t)-x\|^2\leq\frac{1}{2}\|u(s)-x\|^2+\int_s^t\langle f(\tau)+y,u(\tau)-x\rangle_S\,d\tau$$

hold for each $[x,y]\in A$ and $0\leq s\leq t\leq T$.

It is known [2] that the initial value problem (2) has a unique integral solution for every $x\in\overline{D(A)}$ and $f\in L^1(0,T;X)$ and that this integral solution depends continuously on the data x and f. One can thus define the operator H that maps (f,x) into the integral solution u of (2), $H:L^1(0,T;X)\times\overline{D(A)}\to C([0,T];\overline{D(A)})$, and by the previous remarks H is continuous. For $B:C([0,T];X)\to L^1(0,T;X)$, one also defines the operator $G=H\cdot B$, $G:C([0,T];X)\to C([0,T];X)$, and G is obviously continuous.

For every $a\in[0,T]$ the operator $i_a:C([0,a];X)\to C([0,T];X)$ is defined by

$$i_a(u)(t)=\begin{cases}u(t), & 0\leq t\leq a,\\ u(a), & a\leq t\leq T\end{cases}$$

and the operator $j_a:C([0,T];X)\to C([0,a];X)$ by $j_a(v)(t)=v(t)$ for $0\leq t\leq a$. Thus, the operator G_a defined by $G_a=j_aHBi_a$ satisfies $G_a:C([0,a];X)\to C([0,a];\overline{D(A)})$ and is continuous.

DEFINITION. The function $u(t):[0,a]\to\overline{D(A)}$, $0\leq a\leq T$ is called an *integral solution* of the initial value problem (1) on $[0,a]$ if $u(t)$ is continuous on $[0,a]$, $u(0)=x$ and u is a fixed point of the operator G_a.

The function $u(t):[0,b[\to\overline{D(A)}$ is called an integral solution of (1) on $[0,b[$, $0<b\leq\infty$, if $u(t)$ is an integral solution of (1) on every closed interval $[0,a]$, $0\leq a<b$.

Remark. The fact that the function $u(t)$ is an integral solution of the initial value problem (1) on the interval $[0,b]$ does not imply, in general, that $u(t)$ is an integral solution of (1) on any smaller interval $[0,a]$, $0\leq a<b$. To avoid this difficulty we introduce the following class of operators (see Proposition 2.4 in this connection).

DEFINITION. We say that the operator $B:C([0,T];X)\to L^p(0,T;X)$ is *causal* if for each $a\in[0,T]$ and $v_1,v_2\in C([0,T];X)$ such that $v_1=v_2$ on $[0,a]$ we have $(Bi_av_1)(t)=(Bi_av_2)(t)$ almost everywhere for $t\in[0,a]$.

We can now state our main results.

THEOREM 1.1. (Local existence). *Let* $0<T<\infty$, $1\leq p\leq\infty$ *and let* $B:C([0,T];X)\to L^p(0,T;X)$ *be completely continuous. If* $x\in\overline{D(A)}$, *then there exists an* $a\in]0,T[$ *and an integral solution* u *of* (1) *on* $[0,a]$.

THEOREM 1.2. (Global existence). *Let* $0<T<\infty$, $1\leq p\leq\infty$ *and let* $B:C([0,T];X)\to L^p(0,T;X)$ *be completely continuous and causal. If* $x\in\overline{D(A)}$, *then there exists an integral solution* u *of* (1) *defined on a maximal interval of existence* $[0,T]$ *or* $[0,T_{\max}[$, $T_{\max}\leq T$. *In the second case, the solution* u *is unbounded on* $[0,T_{\max}[$.

The proofs of Theorems 1.1 and 1.2 will be given in §§3 and 4, where we also discuss further global existence results and regularity properties of the integal solutions. The next section, §2, is devoted to some preliminaries related to the properties of integral solutions u of (2). In §5 we deal with the continuous dependence of the solutions upon data. Finally, in the Appendix we give a generalization to vector valued functions of the Weyl–Riesz criterion of compactness of sets in the spaces $L^p(\mathbb{R})$. The results of the Appendix are independent of the other results of this paper.

2. Preliminaries. The notion of integral solution was introduced by P. Benilan [2] who also studied their main properties, (see also [1], [5], [8]). We start by recalling some of the properties of integral solutions that will be needed below.

Let A be an m-accretive operator $A: X \to 2^X$. We define for each $\lambda > 0$ $J_\lambda = (I+\lambda A)^{-1}$, $A_\lambda = \frac{1}{\lambda}(I - J_\lambda)$ and $|Ax| = \lim_{\lambda \downarrow 0} \| A_\lambda x \|$ for $x \in X$. Set $\hat{D}(A) = \{x \in X: |Ax| < \infty\}$. $\hat{D}(A)$ is called the generalized domain of A. Identify X with its natural embedding in X^{**} and define operator $\hat{A}$ by

$$\hat{A}x = \{z \in X^{**} : \forall [\xi, \eta] \in A, \exists w \in F(\xi - x), (\eta - z, w) \geq 0\}.$$

The following proposition is a standard result for accretive operators (see, e.g., [2, pp. 0.7–0.8]):

PROPOSITION 2.1. *Let A be m-accretive. Then*

1) $D(A) \subset \hat{D}(A) \subset \overline{D(A)}$.

2) $\| A_\lambda x \| \leq |Ax|$ *for* $x \in X$ *and* $\lambda > 0$.

3) $|Ax| = \inf\{\| y \| : y \in \hat{A}x\}$ *for* $x \in \hat{D}(A)$.

4) *If* $x_n \to x_0$ *as* $n \to \infty$, *then* $|Ax_0| \leq \liminf |Ax_n|$; *if* $|Ax_n|$ *is bounded, then* $x_0 \in \hat{D}(A)$.

Our next results state the existence of integral solutions of (2) and give some of their properties. They are proved in P. Benilan [2]; see also V. Barbu [1, Chap. 3].

PROPOSITION 2.2. *Let A be m-accretive, $x \in \overline{D(A)}$ and $f, g \in L^1(0, T; X)$. Then*

1) *The initial value problem* (2) *has a unique integral solution $u(t)$ on $[0, T]$. Hence the operator H given by $Hf = u$ is well defined and single-valued.*

2) *Let $x, y \in \overline{D(A)}$. Then*

$$\| Hf(t) - Hg(t) \| \leq \| Hf(s) - Hg(s) \| + \int_s^t \| f - g \| \, d\tau$$

for $0 \leq s \leq t \leq T$.

3)

$$\| u(t+h) - u(t) \| \leq \| u(s+h) - u(s) \| + \int_s^t \| f(\tau + h) - f(\tau) \| \, d\tau$$

for $0 \leq s \leq t \leq T$.

4) *Let $z \in \hat{D}(A)$. Then (see* [2, *Remark* 1.3])

$$\| u(h) - z \| \leq \| z - x \| + \int_0^h \| f(s) \| \, ds + |Az| \cdot h.$$

5) *If, moreover, f is of bounded variation on* $[0, T]$, *then*

$$\| u(t+h) - u(t) \| \leq \| u(h) - x \| + \int_t^{t+h} \operatorname*{var}_{[0,\tau]} f \, d\tau.$$

Note that part 5) follows from the fact that the function $t \to \| u(t+h) - u(t) \| - \int_t^{t+h} \operatorname{var}_{[0,\tau]} f \, d\tau$ is nonincreasing (see [2, Prop 1.6]).

PROPOSITION 2.3. *Let $x \in \hat{D}(A)$ and let $f \in L^1(0, T; X)$ be of bounded variation on* $[0, T]$. *Then*

1) $\| u(t+h) - u(t) \| \leq \| u(h) - x \| + \int_t^{t+h} \operatorname{var}_{[0,\tau]} f \, d\tau$;

2) $u(t)$ *is Lipschitz on* $[0, T]$ *and*

$$\| u \|_{\mathrm{Lip}} \leq |f(0+) - Ax| + \operatorname*{var}_{[0,T]} f$$

$$\left(\| u \|_{\mathrm{Lip}} = \sup\{(\| u(t+h) - u(t) \|)/h : h \neq 0, t, t+h \in [0, T]\} \right);$$

3) $\operatorname{var}_{[0,t]} u \leq (|f(0+) - Ax| + \operatorname{var}_{[0,t]} f) \cdot t$;

4) *if X is reflexive, or, more generally, X satisfies the Radon–Nikodým property and the set $A \subset X \times X$ is closed, then the integral solution u is also a strong solution of the initial value problem* (2).

The following proposition is a statement about "heredity" properties of the integral solutions of (1).

PROPOSITION 2.4. *Let B be a causal continuous operator, $B: C([0,T];X)\to L^p(0,T;X)$, $1\le p\le\infty$.*

1) *If u is an integral solution of* (1) *on the interval $[0,b]$, $0<b\le T$, then the function u is an integral solution of* (1) *on every smaller interval $[0,a]$, where $0<a<b$.*

2) *If u is an integral solution of* (1) *on every interval $[0,a]$, where $0<b\le T$ and $0<a<b$ and the limit $\lim u(t)$, $t\nearrow b$ exists, then the function u is an integral solution of* (1) *on the interval $[0,b]$.*

Proof. The results follow immediately from the definitions and the causality of the operator H (statement 2) of Proposition 2.2).

Since the main purpose of this paper is the study of compact perturbations of a given accretive operator, it seems natural to say also some words about compact sets in the space $L^p(0,T;X)$. The complete answer to this problem will be given in the Appendix. Here we give only a few simple sufficient conditions for such compactness. Denote by $|f|_p$ the norm of the function $f\in L^p(0,T;X)$.

PROPOSITION 2.5. *Let $p\in[1,\infty[$. Let $F\subset L^p(0,T;X)$ be such that*

1) $\sup\{|f|_p: f\in F\}=M<\infty$;

2) $\lim_{h\to 0}\int_0^T \|f(t+h)-f(t)\|^p\,dt=0$ *uniformly for* $f\in F$;[1]

3) *there exists a compact set $Q\subset X$ such that $f(t)\in Q$ almost everywhere on $[0,T]$.*

Then F is relatively compact in $L^p(0,T;X)$.

The assertion of Proposition 2.5 is a special case of Theorem A.1.

3. Local existence.

Proof of Theorem 1.1. Since the Lebesgue measure of the interval $[0,T]$ is finite, we have the inclusion $L^s(0,T;X)\subset L^r(0,T;X)$ for $r<s$ $(r,s\ge 1)$ and the strong topology in L^r is weaker than the strong topology in L^s (see, e.g., [3, Part IV, §6.5]). Hence we can consider the given completely continuous operator $B: C([0,T];X)\to L^p(0,T;X)$ as a completely continuous operator into $L^1(0,T;X)$. Let $z\in D(A)$ and take $R=4\max\{\|z-x\|,|Az|\}$. Denote $M_R=\{v\in C([0,T];X): \|v(t)-x\|\le R\ \forall t\in[0,T], v(0)=x\}$. M_R is a bounded, closed subset in the space $C([0,T];X)$, and therefore the set $B(M_R)\subset L^1(0,T;X)$ is relatively compact. By Lemma A.2 there exists $b>0$, $b\le T$ such that $\int_0^b \|f(t)\|\,dt\le\frac{R}{4}$ for all $f\in B(M_R)$. Consider the operator

$$G_b=j_bHBi_b,\qquad G_b: C([0,b];X)\to C([0,b];X).$$

(See the Introduction for the definitions.) Let $M_R^b=j_bM_R\subset C([0,b];X)$. By the inequality 4) of Proposition 2.2 we find that

$$\|(G_bv)(t)-z\|\le\|x-z\|+\int_0^b\|(Bi_bv)(s)\|\,ds+b|Az|\le\frac{3R}{4}$$

for every $t\in[0,b]$ and $v\in M_R^b$. Hence $\|(G_bv)(t)-x\|\le R$, $0\le t\le b$ and $G_bM_R^b\subset M_R^b$. Since the operator G_b is completely continuous and M_R^b is convex, bounded and closed, we can apply the Schauder fixed point theorem. Let u^* be a fixed point of G_b in M_R^b. By definition this is an integral solution of the initial value problem (1) on the interval $[0,b]$ and the proof is complete.

---

[1] Here and in the sequel we will assume that f is defined as 0 outside $[0,T]$.

COROLLARY 3.1. *Let* $0<T<\infty$ *and let* B *be a completely continuous operator*

$$B: C([0,T];X)\to C([0,T];X).$$

Then there exist $b>0$ *and an integral solution of* (1) *on the interval* $[0,b]$.

Proof. Since $C([0,T];X)\subset L^{\infty}(0,T;X)$, this corollary is a consequence of Theorem 1.1.

In the previous theorem, if $x\in\hat{D}(A)$ and B satisfies some further conditions, one gets integral solutions which are Lipschitz continuous. Let f be a function $f:[0,T]\to X$, $A\subset[0,T]$ and var(f,A) be a variation of the function f on the set A, i.e.,

$$\sup\left\{\sum_{k=1}^{n}\|f(t_{k+1})-f(t_k)\|: t_1\le t_2\le\cdots\le t_{n+1},\ t_1\cdots t_{n+1}\in A\right\}.$$

Define ess-var$_{[0,T]}f=\inf\{\text{var}(f,A): A$ is measurable, $A\subset[0,T],\mu([0,T]\setminus A)=0\}$. If f is a measurable function and ess-sup$\{\|f(t)\|,\ t\in\text{d}[0,T]\}<\infty$ denote by $[f]$ an equivalence class of this function in $L^{\infty}(0,T;X)$. Define an essential variation of $[f]$ by ess-var$_{[0,T]}[f]=$ess-var$_{[0,T]}\{g: g\in[f]\}$. This is a correct definition. Note also that in the case of continuous function f ess-var$_{[0,T]}[f]=\text{var}_{[0,T]}f$.

Denote by $V(0,T;X)\subset L^{\infty}(0,T;X)$ the set of all classes of measurable functions of essentially bounded variation with the norm $\|f\|_V=|f|_{\infty}+$ess-var$_{[0,T]}f$. Recall also that by a bounded operator we mean an operator that maps bounded sets into bounded sets.

THEOREM 3.2. *Let* $x\in\hat{D}(A)$ *and let* B *be a completely continuous operator from* $C([0,T];X)$ *into* $L^p(0,T;X)$. *Moreover, if* B *is a bounded operator from* $V(0,T;X)$ *into itself, then there exist* $b\in]0,T]$ *and a Lipschitz continuous integral solution* $v(t)$ *of* (1) *on* $[0,b]$.

Proof. As noted in the proof of Theorem 1.1, the operator B is a completely continuous operator into $L^1(0,T;X)$. Since B is bounded in $V(0,T;X)$, there exists a nondecreasing function $\varphi:[0,\infty[\to[0,\infty[$ such that $\|Bu\|_V\le\varphi(\|u\|_V)$ for all functions $u\in V(0,T;X)\cap C([0,T];X)$. Note that $|Bu|_{\infty}\le\varphi(\|u\|_V)$ and ess-var$_{[0,T]}(Bu)\le\varphi(\|u\|_V)$, and in particular $B(i_bu)$ is of bounded variation. Choose $R>\|x\|$ and set $\alpha=|Ax|+2\varphi(R)+1, b=\min\{R/3\alpha,T\}$. Consider the set

$$M=\{u\in V(0,T;X)\cap C([0,T];X): \|u\|_{\text{Lip}}\le\alpha,\ u(0)=x\}.$$

For all $u\in M$ we have $\text{var}_{[0,b]}u\le R/3$, $\|u-x\|_{\infty}\le R/3$ and $\|u\|_{\infty}\le 2R/3$.

By inequality 2) of Proposition 2.3 (see also [2, pp. 0.13, 1.12]), we find that

$$\|G_bu\|_{\text{Lip}}\le|Ax|+\|Bi_bu(0+)\|+\operatorname*{var}_{[0,T]}(Bi_bu)\le|Ax|+\varphi(R)+\varphi(R)<\alpha.$$

Hence $G_bM\subset M$. The set M is a convex closed and bounded subset of the space $C([0,T];X)$. Since the operator B is completely continuous from $C([0,T];X)$ into $L^p(0,T;X)$, by Schauder's fixed point theorem the operator G_b has a fixed point in M and this is Lipschitz integral solution of (1) on $[0,b]$.

Remark 3.1. Note that from the fact that u is Lipschitz and B is a bounded operator into $V(0,T;X)$ it follows that $B(u)$ is of bounded variation. Therefore, if X is reflexive, the conditions of Theorem 3.1 are satisfied and then the problem (1) has a strong solution by 4) of Proposition 2.3.

Remark 3.2. It follows from the proof of Theorem 3.1 that it is sufficient to check the boundedness of the operator B only on the set of all X-valued Lipschitz continuous functions in the space $V(0,T;X)\cap C([0,T];X)$.

COROLLARY 3.3. *Let B be a completely continuous operator from $C([0,T];X)$ into $C([0,T];X)$. Under the conditions of Theorem* 3.1, *the initial value problem* (1) *has a local Lipschitz integral solution.*

4. Global existence.

Proof of Theorem 1.2. As in the proof of Theorem 1.1, we restrict ourselves to the case $p=1$. Let $a\in]0,T[$ be such that there exists an integral solution u of (1) on the closed interval $[0,a]$. (This is guaranteed by Theorem 1.1.) We will extend this solution as an integral solution of (1) to a closed interval $[0,b]$, where $b>a$.

Let $R>0$ be such that $\|u-x\|_\infty\le R/2$. Define $\tilde{u}=i_a u$ and let $M_R=\{v\in C([0,T];X): v(t)=\tilde{u}(t)$ for $t\in[0,a]$ and $\|v(t)-u(a)\|\le R$ for $a\le t\le T\}$. Since M_R is bounded, the set $B(M_R)$ is relatively compact. By (iii) of Theorem A.1, there exists $0<\delta<a$ such that $\int_0^T\|f(\tau+h)-f(\tau)\|\,d\tau\le R/2$ for every $h\in[0,\delta]$ and $f\in B(M_R)$. Let $b=a+\delta\le T$ and consider the operator G_b and the set $M_R^b=j_b(M_R)$. Let $v\in M_R^b$ and $0<h\le\delta$, and from inequality (3) of Proposition 2.2, we find that

$$\|(G_bv)(a+h)-(G_bv)(a)\|\le\|(G_bv)(h)-(G_bv)(0)\|+\int_0^a\|(Bi_bv)(\tau+h)-(Bi_bv)(\tau)\|\,d\tau.$$

Since $h\le a$ and the operators B and H are causal, we have $(G_bv)(t)=u(t)$ for $t\le a$. Hence $(G_bv)(0)=u(0)=x$ and $(G_bv)(a)=u(a)$. Therefore $\|(G_bv)(a+h)-u(a)\|\le\|u(h)-u(0)\|+R/2\le R$ and $G_bM_R^b\subset M_R^b$.

By Schauder's fixed point theorem, G_b has a fixed point $u^*\subset M_R^b\subset C([0,b];X)$, i.e., an integral solution of (1) on $[0,b]$ and obviously, $u^*(t)=u(t)$ for $t\le a$. Thus every integral solution of (1) can be defined on a maximal interval of existence $[0,T_{\max}[$ or on the whole interval $[0,T]$. (We have used here Proposition 2.4.) Suppose now that an integral solution u is defined on $[0,T_{\max}[$ and is bounded, $\|u\|_\infty\le R$. By definition u is an integral solution on every interval $[0,a]$ for $0<a<T_{\max}$. Since the set $M=\{i_aj_au: a<T_{\max}\}$ is bounded in $C([0,T];X)$, the set $B(M)$ is relatively compact in $L^1(0,T;X)$, and by (3) of the Theorem A. for each $\varepsilon>0$ there exists $\delta>0$ such that, for $0<h\le\delta$, $\int_0^T\|f(\tau+h)-f(\tau)\|\,d\tau\le\varepsilon$, for all $f\in M$. Let $\{t_n\}_1^\infty$ be a sequence $t_n\to T_{\max}$, $t_n<T_{\max}$. Suppose that for $n>N$ $|t_n-T_{\max}|\le\frac{\delta}{2}$. If $\|u(h)-u(0)\|\le\varepsilon$ for $0\le h\le\delta$ and t_n, $t_m\le a<T_{\max}$, $n,m\ge N$, then

$$\begin{aligned}\|u(t_n)-u(t_m)\|&=\|(G_aj_au)(t_n)-(G_aj_au)(t_m)\|\\&\le\|u(|t_m-t_n|)-u(0)\|+\int_0^T\|(Bi_aj_au)(\tau+|t_n-t_m|)-(Bi_aj_au)(\tau)\|\,d\tau\\&\le 2\varepsilon.\end{aligned}$$

Therefore the sequence $\{u(t_n)\}_1^\infty$ is a Cauchy sequence and the limit of $u(t)$ when $t\to T_{\max}$ exists, contrary to the maximality of the interval $[0,T_{\max}[$ (using Proposition 2.4).

THEOREM 4.1. *If B is a causal operator and the conditions of Theorem* 3.2 *are satisfied, then every Lipschitz continuous integral solution of* (1) *on* $[0,a]$, $a\in[0,T[$ *can be extended to a larger interval* $[0,b]$, $b\in[0,T]$ *as a Lipschitz continuous integral solution.*

Proof. The process of the extension of solutions that was described in the proof of Theorem 1.2 can be applied to Lipschitz continuous integral solutions as it was done in the proof of Theorem 3.2.

THEOREM 4.2. *Let B be a causal completely continuous operator $B: C([0,T];X)\to L^p(0,T;X)$, $1\le p\le\infty$, $x\in\overline{D(A)}$. Suppose there exist two locally bounded functions $\varphi,\psi:[0,\infty)\to[0,\infty)$, $\lim_{s\to 0}\varphi(s)<1$, such that*[2] $|\chi([t,t+h])\cdot Bu|_p\le\varphi(h)\|\chi([0,t])\cdot u\|_\infty+\psi(h)$ *for every $u\in C([0,T];X)$ and $0\le t\le t+h\le T$.*

Then every integral solution v is defined on the whole interval $[0,T]$.

Proof. Let v be an integral solution of (1) that is defined on the maximal interval of existence $[0,T_{\max}[$ (see Theorem 1.2 in this connection).

Let $z\in\hat{D}(A)$. By inequality 4) of Proposition 2.2

$$\|v(t)-z\|\le T\cdot|Az|+\int_0^t\|(Bi_aj_av)(\tau)\|\,d\tau+\|x-z\|$$

for $t\le a<T_{\max}$. Let $\frac{1}{q}=1-\frac{1}{p}$ and $0<h\le 1$ such that $\varphi(s)\le c<1$ for each s, $0\le s\le h$. Then

$$\begin{aligned}\|v(t)\|&\le c_1+T\cdot|Az|+\int_0^{t-h}\|(Bi_aj_av)(\tau)\|\,d\tau+\int_{t-h}^t\|(Bi_aj_av)(\tau)|\,d\tau\\&\le c_1+\int_0^{T_{\max}-h}\|(Bi_aj_av)(\tau)\|\,d\tau+h^{1/q}|\chi([t-h,t])\cdot(Bi_aj_av)|_p\\&\le c_2+\varphi(h)\cdot\|j_{t-h}v\|_\infty+\psi(h)\le c_2+c\|j_{T_{\max}-h}v\|_\infty+\psi(h),\end{aligned}$$

where $c<1$. Hence $\|j_{T_{\max}-h}v\|_\infty\le\text{const}$.

Since $\psi(h)$ is a locally bounded function, we find that $\|v\|_\infty\le\text{const}$. Therefore by Theorem 1.2 the maximal interval of existence cannot be $[0,T_{\max}[$, and thus v is defined on the whole $[0,T]$.

COROLLARY 4.3. *Let B be a causal completely continuous operator from $C([0,\infty);X)$ into $L^p(0,T;X)$, $1\le p\le\infty$ and let the conditions of Theorem 4.2 be satisfied for every $T>0$. Then the initial value problem* (1) *has an integral solution $v(t)$ that is defined on $[0,\infty)$.*

5. Dependence on data. We consider a family of initial value problems

$$(1_n)\qquad u'+Au\ni B_nu,\qquad u(0)=x_n,$$

where A is a given m-accretive operator and B_n is a completely continuous operator

$$B_n: C([0,T];X)\to L^p(0,T;X),\qquad 1\le p\le\infty,$$

for $T<\infty$ and $x_n\in\overline{D(A)}$.

THEOREM 5.1. *Suppose that $|B_nv-Bv|_p\to 0$, $n\to\infty$ uniformly on every bounded set in $C([0,T];X)$, $x_n\to x$, and all the initial value problems (1_n) have integral solutions u_n on the same interval $[0,T]$ such that $\|u_n\|_\infty\le C$.*

Then (1) *has an integral solution u on $[0,T]$, and there exists a subsequence of $\{u_n\}_1^\infty$ that converges to u uniformly on $[0,T]$.*

Proof. Since $B_n\to B$, the operator B is completely continuous. Hence, if $M=\{u\in C([0,T];x):\|u\|_\infty\le C\}$, then $B(M)$ is a relatively compact set. As in the proof of Theorem 1.1, the theorem is reduced to the case $p=1$.

Without loss of generality we can suppose that $Bu_n\to f$ in $L^1(0,T;X)$ for some function f. Since $B_n\to B$ uniformly on M, we find that $B_nu_n\to f$ in $L^1(0,T;X)$. Let H_ng

[2] $\chi(I)$ is a characteristic function of the set I.

be a unique integral solution of the initial value problem $u'+Au\ni g$, $u(0)=x_n$, $g\in L^1(0,T;X)$. By Proposition 2.1,

$$\|H_n g-Hf\|_\infty \le \|H_n g(0)-Hf(0)\|+\int_0^T \|g-f\|\,d\tau,$$

where Hf is an integral solution of $u'+Au\ni f$, $u(0)=x$. Let $u=Hf$; then

$$\begin{aligned}\|u_n-u\|_\infty &= \|u_n-Hf\|_\infty = \|H_nB_nu_n-Hf\|_\infty \\ &\le \|u_n(0)-Hf(0)\|+\int_0^T \|B_nu_n(\tau)-f(\tau)\|\,d\tau.\end{aligned}$$

Hence $u_n\to u$ in $C([0,T];X)$. To complete the proof we have to prove that u is an integral solution of (1). Indeed, $u_n\to u$; therefore $Bu_n\to Bu$ in $L^1(0,T;X)$ and $HBu_n\to HB_u$ in $C([0,T];X)$. Since $B_n\to B$ and $\|HBu_n-H_nBu_n\|_\infty\le\|x-x_n\|$, we find that $H_nB_nu_n\to HBu$, or $u_n\to HBu$. But $u_n\to u$ and, therefore $HBu=u$; i.e., u is an integral solution of (1) in $[0,T]$.

6. An example. In this section we give a simple example to which the preceding theory applies. Let $H\subset R^N$ be a bounded domain with smooth boundary. Let $X=L^1(\Omega)$ with norm $\|\cdot\|$, and let A be m-accretive in X. An interesting example of an m-accretive operator in this space is the operator A defined as follows:

$$D(A)=\{u\in L^1(\Omega):\varphi(u)\in W_0^{1,1}(\Omega),\ \Delta\varphi(u)\in L^1(\Omega)\},$$

$Au=-\Delta\varphi(u)$ (in the sense of distributions), where φ is a real function on $\mathbb{R}$, $\varphi(0)=0$, $\varphi\in C(\mathbb{R})\cap C^1(\mathbb{R}\setminus\{0\})$ and $\varphi'(t)\ge 0$ for $t\neq 0$. (see, e.g., [9]).

THEOREM 6.1. *Let A be an m-accretive operator in $L^1(\Omega)$. Let $T>0$, $t\in[0,T]$ and $g(x,y,\tau,s):\bar\Omega\times\bar\Omega\times[0,T]\times\mathbb{R}\times\mathbb{R}$ be continuous of all its variables satisfying*

$$|g(x,y,\tau,s)|\le M(1+|s|^\beta),\qquad s\in\mathbb{R}$$

for some $M>0$ and $\beta\in[0,1[$. Then the initial value problem

$$\frac{\partial u}{\partial t}+Au\ni\int_0^t\int_\Omega g(x,y,\tau,u(x,\tau))\,dx\,d\tau,\qquad u(x,0)=u_0(x)\in\overline{D(A)}$$

has an integral solution defined on a maximal interval of existence $[0,T]$ or $[0,T_{\max}[$, $T_{\max}\le T$. In the second case the solution is unbounded as $t\to T_{\max}$.

For the proof of Theorem 6.1 we will need some preliminaries. Define

$$g_n(x,y,\tau,s)=\begin{cases} g(x,y,\tau,s) & \text{for } |s|\le n,\\ g(x,y,\tau,n\cdot s/|s|) & \text{for } |s|>n,\end{cases}$$

$n=1,2,3\cdots$, and

$$[B_nu](y,t)=\int_0^t\int_\Omega g_n(x,y,\tau,u(x,\tau))\,dx\,d\tau$$

for $u\in C([0,T]\times L^1(\Omega))$.

LEMMA 6.1. *The operator B_n is a completely continuous operator from $C([0,T],L^1(\Omega))$ into itself.*

Proof. Since g_n is bounded and uniformly continuous in $\bar\Omega^2\times[0,T]\times\mathbb{R}$, the operator B_n is a bounded operator in $C([0,T];L^1(\Omega))$. Moreover, B_n can be considered as a bounded operator from $C([0,T];L^1(\Omega))$ into $C([0,T]\times\bar\Omega)$. The image of every bounded set of $C([0,T];L^1(\Omega))$ is a bounded set of equicontinuous functions in $C([0,T]\times\bar\Omega)$. By Ascoli's theorem we get that B_n is a compact operator into $C([0,T]\times\bar\Omega)$. Since

$C([0,T]\times\bar{\Omega})$ is isometric to $C([0,T];C(\bar{\Omega}))$ and $C([0,T];C(\bar{\Omega}))\subset C([0,T];L^1(\Omega))$, we have that B_n is a compact operator from $C([0,T];L^1(\Omega))$ into itself.

To show that B_n is a continuous operator note that every function $u\in C([0,T];L^1(\Omega))$ can be considered as a function in the space $L^1([0,T]\times\Omega)$. For fixed n and y, $g_n(x,y,\tau,u(x,\tau)):[0,T]\times\Omega\to\mathbb{R}$ is a bounded measurable function. Let $\varphi_k\to\varphi$, $k\to\infty$, in $C([0,T];L^1(\Omega))$; then $\varphi_k\to\varphi$ in $L^1([0,T]\times\Omega)$, and therefore, there exists a subsequence $\{\varphi_i\}_1^\infty$ of $\{\varphi_k\}_1^\infty$ that converges to φ almost everywhere in $[0,T]\times\Omega$. Therefore, $g_n(x,y,\tau,\varphi_i(x,\tau))\to g_n(x,y,\tau,\varphi(x,\tau))$ as $i\to\infty$ almost everywhere in $[0,T]\times\Omega$. By Lebesgue dominated convergence theorem, $[B_n\varphi_i](y,t)\to[B_n\varphi](y,t)$ as $i\to\infty$ for every $(y,t)\in[0,T]\times\bar{\Omega}$. Since $\{[B_n\varphi_i]\}_{i=1}^\infty$ is an equicontinuous family of functions, it follows that $[B_n\varphi_i]\to[B_n\varphi]$, $i\to\infty$, uniformly in $[0,T]\times\bar{\Omega}$ and, therefore, $B_n\varphi_i\to B_n\varphi$, $i\to\infty$, in $C([0,T];L^1(\Omega))$. But this is sufficient for the continuity of B_n.

Proof of Theorem 6.1. It is enough to prove (see Theorem 1.2) that the operator

$$[BU](y,t)=\int_0^t\int_\Omega g(x,y,\tau,u(x,\tau))\,dx\,d\tau$$

is a causal and completely continuous from $C([0,T];L^1(\Omega))$ into $C([0,T];L^1(\Omega))$. From the estimate $|g(x,y,t,s)|\le M(1+|s|^\beta)$, we get that B is a bounded operator in this space. The causality of B is obvious. To show that B is a completely continuous operator, we prove that the operators B_n converge to B uniformly on bounded subsets of $C([0,T];L^1(\Omega))$. Suppose that $\varphi\in C([0,T];L^1(\Omega))$ and $\sup_t\|\varphi(x,t)\|\le C<\infty$. ($t\in[0,T]$, $x\in\Omega$.) Let $\varepsilon>0$, $\tau\in[0,T]$, $\varepsilon\le C$. If $N\ge(C/\varepsilon)^{1/(1-\beta)}$ and $\Omega(N,\tau)=\{x\in\Omega:|\varphi(x,\tau)|>N\}$, then $m(\Omega(N,\tau))\le\varepsilon$. Therefore, for each $y\in\Omega$, $t\in[0,T]$, $n\ge N$, we have

$$\begin{aligned}
|[B\varphi](y,t)-[B_n\varphi](y,t)|&\le\int_0^t\int_\Omega|g-g_n|\,dx\,d\tau\\
&=\int_0^t\int_{\Omega\setminus\Omega(N,\tau)}|g-g_n|\,dx\,d\tau+\int_0^t\int_{\Omega(N,\tau)}|g-g_n|\,dx\,d\tau\\
&\le M\int_0^t\int_{\Omega(N,\tau)}\left(2+n^\beta+|\varphi(x,\tau)|^\beta\right)dx\,d\tau\\
&\le 2MT\cdot\varepsilon+2M\int_0^t\int_{\Omega(N,\tau)}|\varphi(x,\tau)|^\beta\,dx\,d\tau\\
&\le 2MT\varepsilon+2Mn^{\beta-1}\cdot\int_0^t\|\varphi(x,\tau)\|\,d\tau\\
&\le 2MT\varepsilon+2Mn^{\beta-1}\cdot T\cdot C\le 2MT(\varepsilon+N^{\beta-1}\cdot C)\le 4MT\varepsilon.
\end{aligned}$$

Thus B is a completely continuous operator, and the proof is complete.

Remark. To obtain local existence of solutions, we do not need the causality of the operator B (see Theorem 1.1). Therefore, we can guarantee the local existence of solutions for more general operators than those treated in Theorem 6.1.

For example, we can consider initial value problems of the type

$$\frac{\partial u}{\partial t}+Au\ni\int_0^T\int_\Omega g(t,x,y,\tau,u(x,\tau))\,dx\,d\tau,\qquad u(x,0)=u_0(x)\in\overline{D(A)}.$$

The operator in the right-hand side of the equation is completely continuous under assumptions on g, which are similar to those considered in Theorem 6.1.

Remark. The previous arguments can be extended to the more general evolution problem

$$\frac{du}{dt}+A(t)u\ni B(u), \qquad u(0)=u_0,$$

where $A(t)$ is an m-accretive operator for almost all $t\in[0,T]$, $\overline{D(A(t))}=D$ is constant almost everywhere and B is a completely continuous operator. (We suppose that $A(t)$ has a standard dependence upon t (see [8].)) An example of such extension can be found in [4, pp. 323–324].

Appendix. A criterion for compactness of subset of $L^p(\mathbb{R},dt,X)$. The following theorem is the generalization of the criterion of Weyl–Riesz for the compactness of sets in the spaces $L^p(T,\mu)$. (see, e.g., [7, Part 4, 20.1]). We treat only the simplest case, where T is the group of reals, $\mu=dt$ is the Lebesgue measure and X is a Banach space. Modifications for the more general case, where T is a group and μ is the Haar's measure are straightforward.

Let $|f|_p$ be the norm in the space $L^p(\mathbb{R},dt,X)$, let $\|\cdot\|$ be the norm in the space X, and let $\chi(A)$ be the characteristic function of the set $A\subset\mathbb{R}$.

THEOREM A.1. *Let $1\le p<\infty$ and $F\subset L^p(\mathbb{R},dt,X)$. The set F is relatively compact in $L^p(\mathbb{R},dt,X)$ if and only if:*

(i) $\sup\{|f|_p : f\in F\}=M<\infty$.

(ii) $\lim_{N\to\infty}\int_{\|t\|\ge N}\|f(t)\|^p\,dt=0$ *uniformly for* $f\in F$.

(iii) $\lim_{h\to 0}\int_{\mathbb{R}}\|f(t+h)-f(t)^p\,dt=0$ *uniformly for* $f\in F$.

(iv) *For every $\varepsilon>0$ there exists a compact set $Q_\varepsilon\subset X$ such that for every $f\in F$ there exists a set $A_{f,\varepsilon}$ with $\mu(A_{f,\varepsilon})\le\varepsilon$ and $f(t)\in Q_\varepsilon$ for every $t\in\mathbb{R}\setminus A_{f,\varepsilon}$.*

The proof of the theorem will be given after some technical preliminaries.

DEFINITION. Let $r(t)\in C_0^\infty(\mathbb{R})$ with support in $[-1,1]$ and $0\le r\le 1$, $\int_{\mathbb{R}} r\,dt=1$. For $\varepsilon>0$ set $r_\varepsilon(t)=\frac{1}{\varepsilon}r(\frac{t}{\varepsilon})$ and for $f\in L^p(\mathbb{R},dt;X)$ let the ε-convolution of f be

$$f_\varepsilon(t)=\int_{\mathbb{R}} r_\varepsilon(t-\tau)f(\tau)\,d\tau=\int_{\mathbb{R}} r_\varepsilon(\tau)f(t-\tau)\,d\tau.$$

Set also $f^h(t)=f(t+h)$ and $\frac{1}{p}+\frac{1}{q}=1$.

Remark. It follows from the definitions, Hölder's inequality and Fubini's theorem that

1) $\|f_\varepsilon(t)\|_\infty\le|r_\varepsilon|_q\cdot|f|_p$.

2) $|f_\varepsilon-f|_p\le|r_\varepsilon|_q\cdot(2\varepsilon)^{1/p}\cdot\sup\{|f^h-f|_p : h\in[-\varepsilon,\varepsilon]\}\le 2\sup\{|f^h-f|_p : h\in[-\varepsilon,\varepsilon]\}$;

3) $\|f_\varepsilon(t)-g_\varepsilon(t)\|\le|r_\varepsilon|_q\cdot|f-g|_p$ for $f,g\in L^p(\mathbb{R},dt,X)$.

LEMMA A.1. *Let either one of the following two conditions hold:*

1) *A set $F\subset L^p(\mathbb{R},dt,X)$ is relatively compact.*

2) *A set $F\subset L^p(\mathbb{R},dt,X)$ satisfies the conditions* (i)–(iii) *of Theorem* A.1.

Then

1) $\|f_\varepsilon\|\le C(\varepsilon)$ *for all* $f\in F$.

2) $|f_\varepsilon-f|_p\to 0$ *uniformly in* F *as* $\varepsilon\to 0$.

3) $\|f_\varepsilon(t)-g_\varepsilon(t)\|\le C(\varepsilon)\cdot|f-g|_p$ *for all* $f,g\in F$.

Proof. Condition 2) is a consequence of condition 1). Indeed, (i)–(iii) are satisfied for a single function $f\in C(]-\infty,\infty[;X)$ with compact support. Therefore, the same holds for any finite set of such functions and, hence, for the precompact set F (since every function in L^p can be approximated by continuous functions with compact support). Let F satisfy (i)–(iii), then the results of the lemma follow easily from the previous remark with $C(\varepsilon)=M\cdot|r_\varepsilon|_q$.

LEMMA A.2. *Let the conditions of Lemma* A.1 *be satisfied. Then, for every* $\varepsilon>0$ *there exists a* $\delta>0$ *such that for every* $A\subset R$ *with* $\mu(A)\le\delta$ *we have* $|f\cdot\chi(A)|_p\le\varepsilon$ *for all* $f\in F$.

Proof. Let $\varepsilon>0$. By statement 2) of Lemma A.1, we can choose $\sigma>0$ such that $|f_\sigma-f|_p\le\frac{\varepsilon}{2}$ for all $f\in F$. By assertion (1) of Lemma A.1, there exists a $\delta>0$ such that $|f_\sigma\cdot\chi(A)|_p\le\frac{\varepsilon}{2}$, where $\mu(A)\le\delta$. Therefore $|f\cdot\chi(A)|_p\le\varepsilon$ for all $f\in F$ and the lemma is proved.

LEMMA A.3. *Let* $F\subset L^p(\mathbb{R},dt,X)$ *satisfy the conditions* (i)–(iv) *of Theorem* A.1. *Then for every* $t\in\mathbb{R}$ *and* $\varepsilon>0$, *the set* $B_\varepsilon(t)=\{f_\varepsilon(t):f\in F\}$ *is relatively compact in* X.

Proof. Given $\alpha>0$ we can choose by Lemmas A.1 and A.2 a $\delta>0$ such that $|f\cdot\chi(\mathbb{R}\setminus A)-f|_p\le\alpha/c(\varepsilon)$ for every measurable set A with $\mu(A)\le\delta$ and all $f\in F$ (here $c(\varepsilon)=|r_\varepsilon|_q\cdot M$). By (iv) we can find a compact set Q_δ and a corresponding set $A_{f,\delta}$ with $\mu(A_{f,\delta})\le\delta$. Let $\tilde f=f\cdot\chi(\mathbb{R}\setminus A_{f,\delta})$ and $Q=\overline{\operatorname{conv}}(Q_\delta\cup\{0\})$. The set Q is convex and compact and $\tilde f(t)\in Q$ for all $t\in\mathbb{R}$ and $f\in F$. Since $\tilde f_\varepsilon(t)\in\overline{\operatorname{conv}}\{\tilde f(\tau):\tau\in\mathbb{R}\}\subset Q$, the set $\{f_\varepsilon(t)\}_{f\in F}$ is relatively compact for every $t\in\mathbb{R}$. By (3) of Lemma A.1 $\|f_\varepsilon(t)-\tilde f_\varepsilon(t)\|\le c(\varepsilon)\cdot|f-\tilde f|_p\le\alpha$. Therefore, there exists a finite set $K\subset X$, $K=\{k_i\}_1^n$ such that for every $f_\varepsilon(t)$ there exists k_i with $\|f_\varepsilon(t)-k_i\|\le2\alpha$. Thus, the set $B_\varepsilon(t)=\{f_\varepsilon(t)\}$ is relatively compact.

Proof of Theorem A.1. We first prove the necessity of the conditions. Assume that $F\subset L^p(\mathbb{R},dt,X)$ is relatively compact. Note that the properties (i)–(iii) hold for a single function in L^p (since e.g. each such function can be approximated by a continuous function with compact support). Therefore, (i)–(iii) hold for every finite subset of L^p and, hence, for every relatively compact subset of L^p. To prove (iv) we recall that a set $E(\varepsilon)$ is said to be ε-dense in F if for every $f\in F$ there is an $e\in E(\varepsilon)$ such that $\|e-f\|<\varepsilon$. Let $0<\varepsilon<1$ and let $E(\varepsilon)$ be a finite ε-dense set in F consisting of continuous functions with compact support.

Set $E_n=E(\varepsilon^{n+1/p}/2^{n/p})$. For every $f\in F$ there is a sequence $\{g_n\}_{n=1}^\infty$ $g_n\in E_n$ such that $|g_n-f|_p<\varepsilon^{n+1/p}/2^{n/p}$. If $A_{f,\varepsilon}^n=\{t\in\mathbb{R}:\|g_n(t)-f(t)\|>\varepsilon^n\}$, then $\mu(A_{f,\varepsilon}^n)\le\varepsilon/2^n$. Let $A_{f,\varepsilon}=\bigcup_{n=1}^\infty A_{f,\varepsilon}^n$. Then $\mu(A_{f,\varepsilon})\le\varepsilon$, and for every $t\in\mathbb{R}\setminus A_{f,\varepsilon}$, we have $\|g_n(t)-f(t)\|\le\varepsilon^n$. Set

$$B_{f,\varepsilon}=\{x\in X:x=f(t)\text{ for }t\in\mathbb{R}\setminus A_{f,\varepsilon}\}$$

and $B_\varepsilon=\bigcup_{f\in F}B_{f,\varepsilon}$. We next prove that B_ε is totally bounded. Let $\delta>0$ be given and fix a natural $m>0$ such that $\varepsilon^m<\frac{\delta}{2}$ (this can be done since $\varepsilon<1$). Since the members of E_m are continuous with compact support, the set

$$E_m(\mathbb{R})=\{x\in X:x=h(t),h\in E_m,t\in\mathbb{R}\}$$

is compact in X. Let $K\subset E_m(\mathbb{R})$ be a $\delta/2$-dense finite subset of $E_m(\mathbb{R})$, then K is a δ-dense finite subset of B_ε. Indeed, if $x\in B_\varepsilon$ there is $f\in F$ and a $t\in\mathbb{R}\setminus A_{f,\varepsilon}$ such that $x=f(t)$. Since $\|g_m(t)-f(t)\|<\delta/2$, we have a $k\in K$ such that

$$\|x-k\|\le\|f(t)-k\|-\|f(t)-g_m(t)\|+\|g_m(t)-k\|\le\frac{\delta}{2}+\frac{\delta}{2}=\delta.$$

Finally, choosing $Q_\varepsilon=\overline{B_\varepsilon}$ we have property (iv).

We turn now to the proof of the sufficiency of the conditions. Suppose (i)–(iv) are satisfied. Let $\varepsilon>0$, $\varepsilon<1$ be given. By (ii) there exists $T>0$ such that $|f\cdot\chi(\mathbb{R}\setminus[-T,T])|_p<\varepsilon$ for all $f\in F$. Define $f^*(t)=f(t)\cdot\chi(I)$, where $I=[-T,T]$ then $|f-f^*|_p<\varepsilon$ and the set

$$F^*=\{f^*:f^*=f\cdot\chi(I),f\in F\}$$

also satisfies the properties (i)–(iv). Consider the set $F_\varepsilon^*=\{f_\varepsilon^*:f\in F\}$ (the set of ε-convolutions of the functions $f\in F$).

It follows from Lemmas A.1 and A.2 that it is an equicontinuous uniformly bounded family of functions (with $\operatorname{supp} f_\varepsilon^* \subset [-T-1; T+1]$). By Lemma A.3 the set $\{f_\varepsilon^*(t)\}$ is relatively compact in $C([-T-1, T+1]; X)$.

It follows from (iii) and the statement 2) of Lemma A.1 that $|f^* - f_\varepsilon^*|_p \to 0$ when $\varepsilon \to 0$ uniformly for $f \in F$. Hence, F^* is relatively compact and so is F.

Acknowledgments. I am indebted to Professor A. Pazy for numerous suggestions and permanent attention to this work and to the referee for several comments that improved the presentation of the paper.

REFERENCES

[1] V. BARBU, *Nonlinear semigroups and differential equations in Banach spaces*, Editure Academiei, Noordhoff, Groningen, 1976.

[2] P. BENILAN, *Equations d'evolution dans un espace de Banach quelconque et applications*, Thesis, Univ. Paris XI, Orsay, 1972.

[3] N. BOURBAKI, *Integration*, Chap. 1–4, Hermann, Paris, 1952.

[4] M. G. CRANDALL AND J. NOHEL, *An abstract functional differential equation and a related Volterra equation*, Israel J. Math., 29 (1978), pp. 313–328.

[5] M. CRANDALL AND A. PAZY, *Nonlinear evolution equations in Banach spaces*, Israel J. Math., 11 (1972), pp. 57–93.

[6] J. DIESTEL AND J. J. UHL, *Vector measures*, Mathematics Surveys 15, American Mathematical Society, Providence, RI, 1977.

[7] R. E. EDWARDS, *Functional Analysis*, Holt, Rinehart, Winston, New York, 1965.

[8] L. EVANS, *Nonlinear evolution equations in an arbitrary Banach space*, Israel J. Math., 26 (1977), pp. 1–42.

[9] ———, *Application of nonlinear semigroup theory to certain partial differential equations*, Proc. Symp. Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, New York, 1978, pp. 163–188.

[10] A. PAZY, *A class of semilinear equations of evolutions*, Israel J. Math., 20 (1975), pp. 23–26.

[11] E. SCHECHTER, *Evolution generated by continuous dissipative plus compact operators*, Bull. London Math. Soc., 13 (1981), pp. 303–308.

[12] D. VOLKMANN, *Ein Existenzsatz für gewöhnliche Differentialgleichungen in Banachräumen.*, Proc. Amer. Math. Soc., 80 (1980), pp. 297–300.

[13] J. VRABIE, *The nonlinear version of Pazy's local existence theorem*, Israel J. Math., 32 (1979), pp. 221–235.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0008 $01.00/0

A VERSION OF THE CHAIN RULE AND INTEGRODIFFERENTIAL EQUATIONS IN HILBERT SPACES*

HANS ENGLER†

Abstract. For Hilbert-space valued functions u, subdifferentials $\partial\varphi$ and certain scalar kernels a, an estimate on $\int_0^t\langle \dot{u}, d/ds\, a * \partial\varphi(u)\rangle$ is derived and then applied to certain integrodifferential equations.

0. Introduction. Consider the integrodifferential equation

$$\text{(I)} \qquad \dot{u}(t)+Au(t)+\int_0^t a(t-s)Bu(s)\,ds \ni f(t), \qquad 0\le t\le T$$

in a Hilbert space H with the scalar product $\langle\cdot,\cdot\rangle$, A and B being maximal monotone operators. To show the existence of functions $u:[0,T]\mapsto H$ that satisfy (I) and an initial condition $u(0)=u_0$, a crucial step usually consists in suitable estimates for certain approximating solutions u_n. A way to do this has been used in [1], [5]. It consists in multiplying (I) (or its approximating analogue) with Bu and integrating. If B is a subdifferential and the form $u\mapsto\langle Bu, Au\rangle$ is bounded from below, frequency conditions on the kernel a lead to the desired estimates.

In this note we want to propose a different approach: Differentiate (I) once (in fact, take difference quotients), multiply with $\dot{u}$ and integrate from 0 to t. After utilizing the monotonicity of A, one is left with the term

$$\int_0^t\left\langle \dot{u}(s), \frac{d}{ds}\int_0^s a(s-\tau)Bu(\tau)\,d\tau\right\rangle ds;$$

cf. [7], where the author handles it by assuming that Bu is dominated in a certain way by Au (either linearly in the norm sense or by a sign condition on $\langle Bu, Au\rangle$).

In §1 of this note, we show that the above term can be estimated directly for positive nonincreasing convex kernels a and subdifferentials B. In §2 we use this estimate to establish the existence of solutions of (I). In §3 some examples are given, A being a differential operator (of order $1, 2, 2m$) and B a polynomial function.

1. An estimate related to the chain rule. Let $\varphi: H\to]-\infty,\infty]$ be convex, lower semicontinuous (lsc) and nontrivial and $\partial\varphi$ be its subdifferential. Let $u\in W^{1,2}([0,T],H)$ and $v\in L^2(0,T;H)$ such that $v(t)\in\partial\varphi(u(t))$ a.e.. Then the chain rule holds:

$$\text{(1)} \qquad \varphi(u(t))-\varphi(u(0))=\int_0^t\langle \dot{u}(s), v(s)\rangle\,ds=\int_0^t\left\langle \dot{u}(s), \frac{d}{ds}(1*v)(s)\right\rangle ds$$

for all $0\le t\le T$ ([3]). Here, $a*v(s)=\int_0^s a(s-\tau)v(\tau)\,d\tau$, of course, if $a\in L^1(0,T;\mathbb{R})$. More generally, we have:

LEMMA 1. *Let φ, u, v be as above, $a\in C^2([0,T],\mathbb{R})$, $a^{(i)}\cdot(-1)^i\ge 0$ on $[0,T]$ $(i=0,1,2)$. Then*

$$\text{(2)} \qquad \int_0^t\left\langle \dot{u}(s), \frac{d}{ds}(a*v)(s)\right\rangle ds\ge a(t)\varphi(u(t))-a(0)\varphi(u(0))-\int_0^t\dot{a}(s)\varphi(u(s))\,ds.$$

---

* Received by the editors June 4, 1981.

† Institut für Angewandte Mathematik, Im Neuenheimer Feld 293, D-6900 Heidelberg, West Germany. Current address: Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706.

Proof. $[0,T]\ni s\mapsto a*v(s)$ is in $W^{1,\infty}([0,T],H)$ and (cf. [3])

$$\begin{aligned}\int_0^t\left\langle \dot u(s),\frac{d}{ds}(a*v)(s)\right\rangle ds&=\int_0^t\langle \dot u(s),a(0)v(s)\rangle\,ds+\int_0^t\langle \dot u(s),\dot a*v(s)\rangle\,ds\\&=a(0)(\varphi(u(t))-\varphi(u(0)))\\&\quad+\int_0^t\left\langle\int_\tau^t\dot a(s-\tau)\dot u(s)\,ds,v(\tau)\right\rangle d\tau.\end{aligned}\tag{3}$$

Choose $h>0$ so small that $1+h\cdot\dot a(0)\geq 0$. Then

$$\begin{aligned}&\int_0^t\left\langle\int_\tau^t\dot a(s-\tau)\dot u(s)\,ds,v(\tau)\right\rangle d\tau\\&\qquad\geq\frac{1}{h}\cdot\int_0^t\left(\varphi(u(\tau))-\varphi\left(u(\tau)-h\cdot\int_\tau^t\dot a(s-\tau)\dot u(s)\,ds\right)\right)d\tau,\end{aligned}$$

since $v(\tau)\in\partial\varphi(u(\tau))$ a.e.. Now for a.e. $\tau\in[0,T]$,

$$\int_\tau^t\dot a(s-\tau)\dot u(s)\,ds=-\dot a(0)u(\tau)+\dot a(t-\tau)u(t)-\int_\tau^t\ddot a(s-\tau)u(s)\,ds,$$

and since $1+h\cdot\dot a(0)\geq 0$, $\dot a(\cdot)\leq 0\leq\ddot a(\cdot)$, and

$$1+h\cdot\dot a(0)-h\dot a(t-\tau)+h\cdot\int_\tau^t\ddot a(s-\tau)\,ds=1,$$

we can use the convexity of φ to deduce that

$$\begin{aligned}&\int_0^t\left\langle\int_\tau^t\dot a(s-\tau)u(s)\,ds,v(\tau)\right\rangle d\tau\\&\quad\geq\frac{1}{h}\cdot\left(\int_0^t\varphi(u(\tau))\,d\tau-\left((1+h\dot a(0))\int_0^t\varphi(u(\tau))\,d\tau-h\cdot\int_0^t\dot a(t-\tau)\varphi(u(t))\,d\tau\right.\right.\\&\qquad\left.\left.+h\cdot\int_0^t\int_\tau^t\ddot a(s-\tau)\varphi(u(s))\,ds\,d\tau\right)\right)\\&\quad=-\dot a(0)\int_0^t\varphi(u(\tau))\,d\tau+(a(t)-a(0))\varphi(u(t))-\int_0^t(\dot a(s)-\dot a(0))\varphi(u(s))\,ds.\end{aligned}$$

Inserting this in (3) finally implies (2).

Remark. If $v_0\in\partial\varphi(u(0))$, then one deduces from (2)

$$\begin{aligned}\int_0^t\left\langle\dot u(s),\frac{d}{ds}(a*v)(s)\right\rangle ds&\geq a(t)\cdot(\varphi(u(t))-\varphi(u(0)))\\&\quad+\int_0^t\langle\dot u(s),(a(s)-a(t))v_0\rangle\,ds\end{aligned}\tag{4}$$

for $0\leq t\leq T$ by taking $\tilde\varphi(u):=\varphi(u)-\langle v_0,u-u(0)\rangle-\varphi(u(0))$ in (2). If $\dot u\in L^\infty(0,T;H)$, the right-hand side in (4) makes sense even if $a\in L^1(0,T;\mathbb{R})$ is only positive and nondecreasing.

2. Existence for integrodifferential equations. Let $A,B:H\to 2^H$ be maximal monotone operators, $B=\partial\varphi$ be the subdifferential of a convex lsc nontrivial functional $\varphi:H\to]-\infty,\infty]$. We shall use the following hypotheses which relate A and B:

(H0) For all positive C,

$$\{u\in H\mid\varphi(u)+|w|+|u|^2\leq C,w\in Au\}$$

is precompact in H.

(H1) There exists some continuous function $K^*: \mathbb{R}^+ \to \mathbb{R}^+$ such that for all $u \in D(A)$ $Bu \neq \emptyset$ and for $w \in Bu$,

$$|w| \leq K^*\big(|A^0 u| + |u|\big),$$

$A^0 u$ denoting the element with minimal norm in Au.

(H2) There exists some continuous function $K^*: \mathbb{R}^+ \to \mathbb{R}^+$ such that for all $u \in D(A)$ and $w \in Bu$,

$$|w| \leq K^*\big(|u| + \varphi(u)\big) \cdot \big(1 + |A^0 u|\big).$$

A preliminary result on local strong solutions of

$$(\mathrm{I}_0) \qquad \begin{aligned} &\dot{u}(t) + v(t) + a * w(t) \ni f(t) \quad \text{a.e. on } [0, T_0], \\ &\dot{u}, v, w \in L^1(0, T_0; H), \quad v \in Au, \quad w \in Bu \quad \text{a.e., } u(0) = u_0 \end{aligned}$$

is contained in:

LEMMA 2. *Let* (H0) *and* (H1) *hold for* A *and* B. *Let* $a \in BV([0, T], \mathbb{R})$, $f \in BV([0, T], H)$, $u_0 \in D(A)$. *Then there exists some* $T_0 > 0$ *and a solution* $u \in W^{1,\infty}([0, T_0], H)$ *of* (I_0); v *and* w *are essentially bounded on* $[0, T_0]$.

Proof. For $\lambda > 0$ let $B_\lambda := (1/\lambda)(\mathrm{Id} - (\mathrm{Id} + \lambda B)^{-1})$ be the Yosida approximation of B. B_λ is Lipschitz continuous with Lipschitz constant $1/\lambda$ [3]. Then the approximating equations

$$(5) \qquad \begin{aligned} &\dot{u}_\lambda(t) + v_\lambda(t) + a * B_\lambda u_\lambda(t) \ni f(t), \qquad 0 \leq t \leq T, \\ &v_\lambda(t) \in A u_\lambda(t) \quad \text{a.e.}, \qquad u_\lambda(0) = u_0 \end{aligned}$$

have unique strong solutions $u_\lambda \in W^{1,\infty}([0, T], H)$; $v_\lambda \in L^\infty(0, T; H)$ (cf. [4], [5]). We want to estimate $|v_\lambda(\cdot)| + |u_\lambda(\cdot)|$ independently of λ on some small interval $[0, T_0]$. Put $g_\lambda := f - a * B_\lambda u_\lambda$. Since $g_\lambda \in BV([0, T], H)$,

$$\begin{aligned} &|\dot{u}_\lambda(t)| \leq |\big(g_\lambda(0+0) - Au_0\big)^0| + \mathrm{Var}(g_\lambda; 0, t), \\ &|v_\lambda(t)| \leq |\big(g_\lambda(0+0) - Au_0\big)^0| + \mathrm{Var}(g_\lambda; 0, t) + \| g_\lambda \|_{L^\infty(0,t)} \end{aligned}$$

for a.e. $t \in [0, T]$ ([3]). Now

$$\begin{aligned} \mathrm{Var}(g_\lambda; 0, t) &\leq \mathrm{Var}(f; 0, t) + \big(|a(0)| + \mathrm{Var}(a; 0, t)\big) \cdot \int_0^t |B_\lambda u_\lambda(s)|\, ds \\ &\leq C_1 + C_2 \cdot \int_0^t K^*\big(|A^0 u_\lambda(s)| + |u_\lambda(s)|\big)\, ds \end{aligned}$$

and

$$|\big(g_\lambda(0+0) - Au_0\big)^0| = |\big(f(0+0) - Au_0\big)^0| = C_3.$$

Thus, $|v_\lambda(\cdot)| + |u_\lambda(\cdot)|$ satisfies an integral inequality

$$(6) \qquad |v_\lambda(t)| + |u_\lambda(t)| \leq C_4 + C_5 \cdot \int_0^t K^*\big(|v_\lambda(s)| + |u_\lambda(s)|\big)\, ds,$$

C_4, C_5 only depending on u_0 and f. There exists a maximal solution of (6) which is bounded by some constant C_6 on some small interval $[0, T_0]$. Hence, $v_\lambda(\cdot)$, $B_\lambda u_\lambda(\cdot)$, $\varphi(u_\lambda(\cdot))$ and $\dot{u}_\lambda(\cdot)$ are also bounded on $[0, T_0]$. Thus $\{u_\lambda \mid \lambda > 0\}$ is relatively compact in $C([0, T_0], H)$ by Arzela's theorem, and $\{v_\lambda \mid \lambda > 0\}$, $\{B_\lambda u_\lambda \mid \lambda > 0\}$ are relatively weak* compact in $L^\infty(0, T_0; H)$. Taking strongly resp. weak* convergent subsequences as $\lambda \downarrow 0$ ($u_\lambda \to u$, $v_\lambda \to v$, $B_\lambda u_\lambda \to w$) and using the maximal monotonicity of A and B to see that

$v \in Au$ and $w \in Bu$ a.e. on $[0, T_0]$, we conclude that there exists a local strong soluton of (I_0).

If $a \in BV([0,T],\mathbb{R})$ and $f \in BV([0,T],H)$, then $u(T_0) \in D(A)$ and $t \mapsto \int_0^{T_0} a(t-s)w(s)\,ds$ is in $BV([T_0,T],H)$. Thus, we can repeat this existence argument on some interval $[T_0, T_1]$, etc., and find (by Zorn's lemma) that there exists a maximal solution $u \in W^{1,\infty}_{\mathrm{loc}}([0,T^*[,H)$ on some maximal interval $[0,T^*[$. We also note that a bound on u and on $v \in Au$ on some interval $[0,\bar{t}[$ implies bounds for $w \in Bu$ (by (H1)) and $\dot{u}$ (from the equation) on the same interval. Then $\lim_{t \uparrow \bar{t}} u(t) \in D(A)$ exist, and the solution u can be continued on some larger interval, unless $\bar{t} = T$.

How to find such uniform bounds (by Lemma 1) is the object of

THEOREM 3. *Let* (H0), (H1), (H2) *hold for A and B. Let* $a \in C^2([0,T],\mathbb{R})$, $(-1)^i \cdot a^{(i)} \geq 0$ *for* $0 \leq i \leq 2$, $f \in BV([0,T],H)$ *and* $u_0 \in D(A)$. *Then there exists a strong solution* $u \in W^{1,\infty}([0,T],H)$ *of* (I_0), *and* v $(\in Au)$ *and* w $(\in Bu)$ *both are in* $L^\infty(0,T;H)$.

Proof. Without loss of generality, let $a(T) > 0$ (else we solve (I_0) step by step, or $a \equiv 0$).

Let u be a maximal solution on some interval $[0,T^*[$ (which exists by Lemma 2 and the following remarks). We show that $\lim_{t \uparrow T^*} u(t) \in D(A)$ exists (and thus $T^* = T$). For any measurable function $g: [0,T^*[\to H$, $0 < h < T^*$, $0 \leq t < T^* - h$, let

$$d_h g(t) := \frac{g(t+h) - g(t)}{h}$$

denote the forward difference quotient. We apply d_h to all the terms in (I_0), multiply by $d_h u(t)$ and use the monotonicity of A to get

$$\frac{1}{2}\frac{d}{dt}|d_h u(t)|^2 + \langle d_h u(t), d_h(a * w)(t)\rangle \leq \langle d_h u(t), d_h f(t)\rangle. \tag{7}$$

Integrating between 0 and t and estimating the right-hand side gives

$$\frac{1}{2}|d_h u(t)|^2 + \int_0^t \langle d_h u(s), d_h(a * w)(s)\rangle\, ds \leq \frac{1}{2}|d_h u(0)|^2 + |d_h u|_{L^\infty(0,t;H)} \cdot \mathrm{Var}(f;0,T).$$

As $h \downarrow 0$ we use Lebesgue's theorem and find that

$$\begin{aligned} &\frac{1}{2}|\dot{u}(t)|^2 + \int_0^t \left\langle \dot{u}(s), \frac{d}{ds}(a * w)(s)\right\rangle ds \\ &\qquad \leq \frac{1}{2}\left|\left(f(0+0) - Au_0\right)^0\right|^2 + |\dot{u}|_{L^\infty(0,t;H)} \cdot \mathrm{Var}(f;0,T). \end{aligned} \tag{8}$$

for almost all $t \in [0,T^*[$. Using the remark after Lemma 1 above, we estimate further

$$\begin{aligned} &\frac{1}{2}|\dot{u}(t)|^2 + a(t)\varphi(u(t)) \\ &\quad \leq a(t)\varphi(u_0) + \frac{1}{2}\left|\left(f(0+0) - Au_0\right)^0\right|^2 + |\dot{u}|_{L^\infty(0,t;H)} \cdot \mathrm{Var}(f;0,T;H) \\ &\qquad + \int_0^t \left\langle \dot{u}(s), (a(t) - a(s))B^0 u_0\right\rangle ds \\ &\quad \leq C_1(a,f,u_0) + |\dot{u}|_{L^\infty(0,t;H)} \cdot C_2(a,f,u_0) \end{aligned} \tag{9}$$

with suitable constants C_1, C_2. Hence $|\dot{u}(\cdot)|+\varphi(u(\cdot))$ is essentially bounded on $[0,T^*[$ by some constant $C_3(a,f,u_0)$. (H2) now implies that v ($\in Au$ a.e.) satisfies the integral inequality

$$\begin{aligned}|v(t)| &\le a * |w|(t)+|\dot{u}(t)|+|f(t)| \\ &\le a * \big(K^*(\varphi(u(\cdot))+|u(\cdot)|)\cdot(1+|v(\cdot)|)\big)(t)+|\dot{u}(t)|+|f(t)| \\ &\le C_4 \cdot a * |v(\cdot)|(t)+C_5\end{aligned}$$

with suitable constants C_4, C_5. Gronwall's lemma then shows that $v(\cdot)$ is essentially bounded on $[0,T^*[$. Hence, $\lim_{t\uparrow T^*} u(t)\in D(A)$ exists, and since $[0,T^*[$ was assumed to be maximal, $T^*=T$ follows.

Remarks.

(i) The proof generalizes directly to equations of the form

$$(\mathrm{I}_1) \qquad \dot{u}(t)+Au(t)+\sum_{i=1}^{n} a_i * B_i u(t) \ni f(t),$$

each $a_i\in C^2([0,T],\mathbb{R})$ and $B_i=\partial\varphi_i$ fulfilling the assumptions of Theorem 3.

(ii) No additional difficulty arises if A is replaced by $A+\lambda\cdot\mathrm{Id}$, $\lambda\in\mathbb{R}$ some constant: One observes that this gives rise to an additional term $\lambda\cdot|d_h u(t)|^2$ on the left-hand side in (7). This in turn causes an additional term $|\lambda|\cdot\int_0^t|\dot{u}(s)|^2\,ds$ on the right-hand side of (9), and the essential boundedness of $|\dot{u}(\cdot)|+\varphi(u(\cdot))$ still follows from Gronwall's lemma.

(iii) In the proof above we only used the remark after Lemma 1, which neither contains $\dot{a}$ nor $a(0)$. Then an approximation argument generalizes Theorem 3 to kernels a satisfying

$$(10) \qquad a\in L^1(0,T;\mathbb{R}), \qquad a \text{ positive, nonincreasing, convex.}$$

If (10) holds only on some small interval $[0,T_0]$ (and $a\in BV([T_0,T],\mathbb{R})$), one shows the existence of a solution on the interval $[0,T_0]$ and proceeds by translation and induction. This is possible since even then

$$t\mapsto\int_0^{\bar{t}} a(t-s)v(s)\,ds$$

is in $BV([\bar{t},T],H)$ if $v\in L^\infty(0,\bar{t};H)$ and $0<\bar{t}<T$.

(iv) If φ has bad growth properties, (H2) is rather weak. It can be compensated by a coercivity condition for A, namely:

(H3) There exists a lsc seminorm $\|\cdot\|_1$ on H and a continuous function $K^*:\mathbb{R}^+\to\mathbb{R}^+$ such that for all $v_i\in Au_i$ $(i=1,2)$,

$$\langle v_1-v_2, u_1-u_2\rangle \ge \|u_1-u_2\|_1^2$$

and for all $w\in Bu_1$,

$$|w|\le K^*(\|u_1\|_1+|u_1|)\cdot\big(1+|A^0u_1|\big).$$

If (H3) is satisfied instead of (H2), we get an additional term $\|d_h u(t)\|_1^2$ on the left-hand side of (7) in the proof of Theorem 3. This in turn gives an extra term $\int_0^t\|\dot{u}(s)\|_1^2\,ds$ on the left-hand side of (9). Since $\|u_0\|_1<\infty$, $\|u(\cdot)\|_1$ is a priori bounded on $[0,T^*[$; using (H3) instead of (H2) to estimate v ($\in Au$) in the sequel, we complete the proof as above.

(v) It should finally be noted that the technique of the proof of Theorem 3 can also be used to give a priori estimates for solutions of equations of the type (cf. [2], [8], [9])

$$(I_2) \qquad \dot{u}(t)+a*Bu(t)\ni f(t),$$

$$(I_3) \qquad \varepsilon u(t)+\delta Au(t)+a*Bu(t)\ni f(t), \qquad \varepsilon,\delta\geq 0.$$

3. Examples. Throughout this section $a\in C^2([0,T],\mathbb{R})$ fulfils $a^{(i)}\cdot(-1)^i\geq 0$ $(0\leq i\leq 2)$.

Example 1. Let $H:=L^2(0,1;\mathbb{R})$, $Au=u'$ for all $u\in D(A):=\{u\in H^{1,2}([0,1],\mathbb{R})\mid u(0)=0\}$. Then A is maximal monotone in H.

Let $p\geq 1$ and $Bu(x):=|u(x)|^p\cdot\operatorname{sign}u(x)$, if $u\in D(B)=L^{2p}(0,1;\mathbb{R})$. Then $B=\partial\varphi$ is the subdifferential of the convex, lsc, nontrivial functional φ,

$$\varphi(u)=\frac{1}{p+1}\cdot\int_0^1|u(x)|^{p+1}dx.$$

For all $u\in D(A)$

$$|Bu|_H^2=\int_0^1|u(x)|^{2p}dx\leq|u|_{L^\infty}^{2p}\leq|Au|_H^{2p},$$

and $\{u\in H\mid\varphi(u)+|Au|_H+|u|_H^2\leq C\}$ is relatively compact in H for all $C>0$ by suitable imbedding theorems. Finally, by some well-known Sobolev inequality (cf. [6]), there exists $C_0>0$ such that

$$|Bu|_H^2=\int_0^1|u|^{2p}\leq C_0\cdot\left(\int_0^1|u'|^2\right)\cdot\left(\int_0^1|u|^{p+1}\right)^{(2p-2)/(p+1)}$$

for all $u\in D(A)$. Thus A and B satisfy (H0)–(H2). It follows from Theorem 3 that for all $u_0\in H^{1,2}([0,1],\mathbb{R})$ such that $u_0(0)=0$, for all $f\in BV([0,T],H)$ and for all a as above a solution $u\in W^{1,\infty}([0,T],L^2(0,1;\mathbb{R}))\cap L^\infty(0,T;H^{1,2}([0,1],\mathbb{R}))$ exists of

$$\partial_t u(x,t)+\partial_x u(x,t)+\int_0^t a(t-s)|u(x,s)|^p\operatorname{sign}u(x,s)\,ds=f(x,t) \qquad \text{a.e. in } [0,1]\times[0,T],$$

$$u(\cdot,0)\equiv u_0 \quad \text{a.e. in } [0,1],$$

$$u(0,\cdot)\equiv 0 \quad \text{a.e. in } [0,T].$$

Example 2. Let $\Omega\subset\mathbb{R}^n$ be open, bounded, $\partial\Omega$ C^2-smooth. Let $G:\mathbb{R}^n\to\mathbb{R}$ be convex and continuously differentiable, $G(0)=0$, and put $g:=\nabla_x G$. Suppose there exist $c_0>0$, $q\geq 2$, such that for all $\xi\in\mathbb{R}^n$

$$(11) \qquad G(\xi)\geq c_0\cdot|\xi|^q.$$

Let $H=L^2(\Omega,\mathbb{R})$ and define for $u,v\in H$,

$$v\in Au:\Leftrightarrow u\in W_0^{1,1}(\Omega;\mathbb{R}),\quad g(\nabla_x u)\in L^1(\Omega,\mathbb{R}^n),\quad v=-\operatorname{div}g(\nabla_x u)\quad \text{in } \mathcal{D}'(\Omega,\mathbb{R}).$$

Then $A:H\to 2^H$ is maximal monotone (since it is a subdifferential, cf. [10]).

For $v\in Au$ we want to estimate $|v|_H$ from below by some norm of u in the special case $n>2q$. To this end we employ a technique by L. Veron ([10]): For $v\in Au$ and $p\geq 2$,

$$\int_\Omega v\cdot u\cdot|u|^{p-2}\geq c_1\cdot\|u^{(p+q-2)/q}\|_{W^{1,q}}\geq c_2\cdot\|u\|_{L^{\tilde{p}}}^{p+q-2},$$

$$(12) \qquad \tilde{p}=\frac{n(p+q-2)}{n-q}$$

with positive constants c_1, c_2, if additionally $u, v \in L^p(\Omega, \mathbb{R})$. For $p=2$ (12) implies

$$\int_\Omega v \cdot u \geq c_1 \cdot \|u\|^q_{W^{1,q}} \geq c_3 \cdot |u|_H \cdot \|u\|^{q-1}_{W^{1,q}},$$

hence

$$|v|_H \geq c_3 \cdot \|u\|^{q-1}_{W^{1,q}}. \tag{13}$$

For $p=1+n(q-1)/(n-2q)$, hence $\tilde{p}=2\cdot(p-1)$, (12) implies

$$\int_\Omega v \cdot u \cdot |u|^{p-2} \geq c_2 \cdot \|u\|^{p+q-2}_{L^{\tilde{p}}} = c_2 \cdot \|u\|^{q-1}_{L^{\tilde{p}}} \cdot |u^{p-1}|_H,$$

hence,

$$|v|_H \geq c_2 \cdot \|u\|^{q-1}_{L^{\tilde{p}}}, \qquad \tilde{p} = \frac{2n(q-1)}{n-2q} \tag{14}$$

if still $u, v \in L^p(\Omega, \mathbb{R})$. An approximation argument then implies (14) for all $u, v \in H$, since A restricted to $L^p(\Omega, \mathbb{R})$ is m-accretive for all $p \geq 2$ ([10]).

Next let $h: \mathbb{R} \to \mathbb{R}$ be continuous and nondecreasing, $h(0)=0$, $\bar{h}(r) := \int_0^r h(s)\,ds$ be its primitive. Define $Bu(x) := h(u(x))$ for all u such that $Bu \in H$. Then $Bu = \partial\varphi(u)$, $\varphi(u) = \int_\Omega \bar{h}(u(x))\,dx$ for all $u \in D(\varphi) = \{u \in H \mid \bar{h}(u) \in L^1(\Omega, \mathbb{R})\}$ ([3]), and φ is convex, lsc and nontrivial.

We now assume that there exist constants c^*, $\varepsilon, \rho, r > 0$, such that

$$\begin{aligned} &|h(t)| \leq c^* \cdot (|t|^\rho + 1), \\ &\bar{h}(t) \geq \varepsilon |t|^r - c^* \quad \text{for all } t \in \mathbb{R}. \end{aligned} \tag{15}$$

Concerning ρ and r, we assume further that

$$2 \cdot \rho \leq \tilde{p}, \quad q-1 \leq \rho \leq q-1+\frac{r \cdot q}{n}, \quad r \geq 2, \tag{16a,b,c}$$

$\tilde{p}$ as in (14). Then there exist constants δ and c_4 such that $\varphi(u) \geq \delta \cdot \|u\|^r_{L^r} - c_4$ for all $u \in D(\varphi)$, and

$$\{u \in H \mid |u|^2_H + \varphi(u) + |v| \leq C,\ v \in Au\}$$

is relatively compact in H for all $C \geq 0$ by (13) and the compactness of the imbedding $W^{1,q}(\Omega, \mathbb{R}) \to H$. Thus, (H0) holds. (16a) further implies that

$$|Bu|_H \leq c_5 \cdot (\|u\|^\rho_{L^{\tilde{p}}} + 1) \leq c_6 \cdot (|v|^{n/(n-2q)}_H + 1), \qquad v \in Au;$$

hence (H1) holds. Finally the interpolation inequality

$$\|u\|_{L^{2\rho}} \leq c_6 \cdot \|u\|^{1-a}_{L^r} \cdot \|u\|^a_{L^{\tilde{p}}},$$

$a=(q-1)/\rho$, ρ satisfying (16b) (cf. [6]) implies that

$$\begin{aligned} |Bu|_H &\leq c_7 \cdot (\|u\|^\rho_{L^{2\rho}} + 1) \\ &\leq c_8 \cdot (\|u\|_{L^r} + 1)^{\rho+1-q} \cdot (\|u\|^{q-1}_{L^{\tilde{p}}} + 1) \\ &\leq K^*(|u|_H + \varphi(u)) \cdot (1 + |v|_H) \quad \text{if } v \in Au; \end{aligned}$$

hence (H2) holds. (In the "limit case", $r=\rho+1$, (16b) holds if and only if $\rho \leq q-1+q^2/(n-q)$; cf. [5] for $q=2$.)

It follows that for all such g and h and for all a as above, for all $u_0 \in D(A)$ and for all $f \in BV([0,T], L^2(\Omega,\mathbb{R}))$ some $u \in W^{1,\infty}([0,T], L^2(\Omega,\mathbb{R}))$ and $v \in L^\infty(0,T; L^2(\Omega,\mathbb{R}))$ exist such that

$$\begin{aligned}
&\partial_t u(x,t) + v(x,t) + a * h(u(x,\cdot))(t) = f(x,t),\\
&v(x,t) = -\operatorname{div}_x g(\nabla_x u(x,t)) \quad \text{a.e. in } \Omega\times[0,T],\\
&u(\cdot,0) = u_0,\\
&u(\cdot,t) \in W_0^{1,1}(\Omega,\mathbb{R}) \quad \text{for a.e. } t\in[0,T].
\end{aligned}$$

All these considerations hold if $n > 2q$. If $n \le 2q$, (14) will hold with any $\tilde{p} < \infty$ ($\tilde{p} = \infty$, if $n < q$), which will give the condition $\rho < q - 1 + \frac{r}{2}$ in (16b).

Example 3. Let $\Omega \subset \mathbb{R}^n$ be open and bounded, $m \in \mathbb{N}$, $\partial\Omega$ be C^{2m}-smooth. Let

$$Au := \sum_{0\le|\alpha|,|\beta|\le m} \partial^\alpha\big(a^{\alpha\beta}(\cdot)\partial^\beta u(\cdot)\big)$$

if $u \in D(A) := (H^{2m,2} \cap H_0^{m,2})(\Omega,\mathbb{R})$ be a strongly elliptic operator with C^m-coefficients $a^{\alpha\beta}$ (cf. [6]). We shall assume that Gårding's inequality holds in the version

$$\langle Au, u\rangle_{L^2} \ge c\|u\|^2_{H^{m,2}}, \qquad c>0, \tag{17}$$

which is certainly true if a^{00} is big enough. This is no loss of generality by remark (iii) after Theorem 3. The well-known solution theory for linear elliptic equations in $L^2(\Omega,\mathbb{R}) =: H$ then implies that A is maximal monotone in H and that

$$|Au|_H \ge \varepsilon\|u\|_{H^{2m,2}} \quad \text{for all } u \in D(A) \tag{18}$$

with some positive ε.

As in Example 2, let $h: \mathbb{R} \to \mathbb{R}$ be continuous and nondecreasing, $h(0) = 0$, $\bar{h}$ be its primitive and define

$$B(u)(x) := h(u(x)), \qquad \varphi(u) := \int_\Omega \bar{h}(u(x))\,dx$$

for all u such that $h(u) \in H$ resp. $\bar{h}(u) \in L^1(\Omega,\mathbb{R})$. Then φ is convex, lsc, nontrivial, and $B = \partial\varphi$. Hence (18) together with well-known imbedding theorems imply that (H0) holds in this situation.

(H1) means here that

$$|B(u)|^2 = \int_\Omega |h(u)|^2 \le K^*(|Au| + |u|) \le K^{**}(\|u\|_{H^{2m,2}})$$

should hold for some function K^{**}; this is certainly true if

$$n < m_1 \tag{19a}$$

or

$$n = m_1, \quad |h(r)| \le C\cdot\big(|r|^{\tilde{p}} + 1\big) \quad \text{for all } r, \quad C, \tilde{p} > 0, \tag{19b}$$

or

$$n > m_1, \quad |h(r)| \le C\cdot\big(|r|^{\tilde{p}} + 1\big) \quad \text{for all } r, \quad C>0, \quad \tilde{p} = 1 + \frac{4}{n - m_1}, \tag{19c}$$

$m_1 = 4 \cdot m$ [6]. We sharpen these conditions by assuming that (19) holds with $m_1 = 2 \cdot m$ (thus $\tilde{p} = (n+2m)/(n-2m)$ in (19c)). Then for all $u \in D(A)$,

$$|Bu|_H^2 \leq K^*(\|u\|_{L^\infty}), \quad \text{if } n < 2m,$$

$$|Bu|_H^2 \leq c_0 \cdot \left(\int_\Omega |u|^{2\tilde{p}} + 1 \right) \leq c_1 \cdot \left(\|u\|_{H^{2m,2}}^2 \cdot \|u\|_{H^{m,2}}^{2\tilde{p}-2} + 1 \right)$$

$$\leq c_2 \cdot \left(1 + \|u\|_{H^{m,2}}^{2\tilde{p}-2}\right)\left(1 + |Au|_H^2\right), \quad \text{if } n \geq 2m,$$

$\tilde{p}$ as in (19), c_0, c_1, c_2 suitable constants, by some Sobolev inequality ([6]). These estimates together with (17) imply that condition (H3) (Remark (iv) after Theorem 3) holds, the seminorm $\|\cdot\|_1$ being the $H^{m,2}$-norm in this case. Remark (iv) then implies that for all such h, a and A and for all $f \in BV([0,T], L^2(\Omega, \mathbb{R}))$ and $u_0 \in H^{2m,2} \cap H_0^{m,2}(\Omega, \mathbb{R})$ some $u \in W^{1,\infty}([0,T], L^2(\Omega, \mathbb{R})) \cap L^\infty(0,T; H^{2m,2}(\Omega, \mathbb{R}))$ exists such that

$$\partial_t u(x,t) + \sum_{0 \leq |\alpha|, |\beta| \leq m} \partial^\alpha \left(a^{\alpha\beta}(x) \partial^\beta u(x,t) \right) + \int_0^t a(t-s) h(u(x,s))\, ds = f(x,t) \quad \text{a.e. in } \Omega \times [0,T],$$

$$u(\cdot, t) \in H_0^{m,2}(\Omega, \mathbb{R}) \quad \text{for a.e. } t,$$

$$u(\cdot, 0) \equiv u_0 \quad \text{a.e. in } \Omega.$$

In this case more regularity of the solution can be shown (depending on the data) by employing the theory of analytic semigroups in L^p-spaces [6] (resp. of C^α-semigroups [11]).

4. Concluding remarks. The main features in the above considerations are the use of Lemma 1 together with suitable interpolation inequalities in order to utilize the information that is given in the estimate (2). Since the class of kernels (positive, nonincreasing, convex) used there is rather small (compared to other techniques), one could look for some generalization of Lemma 1 by discarding some of the assumptions for a. That no version of Lemma 1 will hold if a is just a positive C^2-function is indicated by the following simple

Counterexample. Take $\mathbb{R}$ as the Hilbert space H, $a(t) = e^t$, $\varphi(u) = e^{-u}$ (thus $\partial\varphi(u) = -e^{-u}$) and $u(t) = t$ ($t \geq 0$). Then

$$\int_0^t \dot{u}(s) \cdot \frac{d}{ds} (a * \partial\varphi(u)(s))\, ds = -\sinh t. \tag{20}$$

On the other hand, let $b(s) = c \cdot e^s$ be a linear combination of $a, \dot{a}, \ddot{a}$. Then

$$a(t)\varphi(u(t)) - a(0)\varphi(u(0)) + \int_0^t b(s)\varphi(u(s))\, ds = c \cdot t,$$

$$\varphi(u(t)) - \varphi(u(0)) + \int_0^t b(s)\varphi(u(s))\, ds = e^{-t} - 1 + c \cdot t,$$

and both terms cannot be lower bounds for the expression in (20) for all positive t. By rescaling the equation, one also gets counterexamples for any finite t-interval.

Acknowledgment. Many thanks are due to Professor John A. Nohel for stimulating discussions on this note.

REFERENCES

[1] V. BARBU, *Nonlinear Volterra integro-differential equations in Hilbert spaces*, Conf. del Sem. di Mat. Bari, 143 (1976).

[2] _____, *Existence for nonlinear Volterra equations in Hilbert spaces*, this Journal, 10 (1979), pp. 552–569.

[3] H. BRÉZIS, *Opérateurs maximaux monotones*, North-Holland, Amsterdam, 1973.

[4] M. G. CRANDALL AND J. A. NOHEL, *An abstract functional differential equation and a related nonlinear Volterra equation*, Israel J. Math., 29 (1978), pp. 313–328.

[5] M. G. CRANDALL, S.-O. LONDEN AND J. A. NOHEL, *An abstract nonlinear Volterra integrodifferential equation*, JMAA, 64 (1978), pp. 701–735.

[6] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[7] G. GRIPENBERG, *On some integral and integrodifferential equations in a Hilbert space*, Ann. di Mat. Pura et Appl., 118 (1978), pp. 181–194.

[8] S.-O. LONDEN, *An existence result on a Volterra equation in a Banach space*, Trans. AMS, 235 (1978), pp. 285–304.

[9] S.-O. LONDEN AND O. J. STAFFANS, *A note on Volterra equations in a Hilbert space*, Proc. AMS, 70 (1978), pp. 57–62.

[10] L. VERON, *Coércivité et propriétés régularisantes des semi-groupes non linéaires dans les espaces de Banach*, Université de Tours, 1977.

[11] W. v. WAHL, *Neue Resolventenabschätzungen für elliptische Differentialoperatoren und semilineare parabolische Gleichungen*, Abh. Math. Sem. Hamburg, 46 (1977), pp. 179–204.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0009 $01.00/0

PERIODIC KOLMOGOROV SYSTEMS*

J. M. CUSHING†

Abstract. The existence of nontrivial periodic solutions of periodic Kolmogorov systems of ordinary differential equations is considered. Under very general conditions, a global continuum of solutions is shown to bifurcate from a noncritical periodic solution of a reduced system using the average inherent per unit growth rate of one component of the system as a bifurcation parameter. The positivity and stability of the bifurcating branch solutions are studied. The stability is shown to depend on the stability of the solution of the reduced system as well as the direction of bifurcation. These results extend and generalize earlier work on periodic Volterra–Lotka systems. Applications to mathematical ecology are given.

Key words. Kolmogorov systems, periodic solutions, bifurcation, stability

1. Introduction.

Differential equations of the form

$$x_i' = x_i h_i(x_1, \cdots, x_n), \qquad 1 \le i \le n, \quad ' = \frac{d}{dt}$$

are sometimes referred to as Kolmogorov equations. They arise in applications in which the per unit of change x_i'/x_i of dependent variables $x_i = x_i(t)$ are prescribed functions $h_i(x_1, \cdots, x_n)$ of these variables at any given time. Many models in population dynamics and mathematical ecology are of this form, of which the well-known logistic, Volterra predator-prey and Volterra–Lotka competition equations serve as perhaps the simplest and most famous examples. In such applications one is usually interested only in nonnegative solutions $x_i \ge 0$ and in positive solutions $x_i > 0$, if they exist. As indicated above the equations are autonomous and the algebraic zeros of the h_i play of course an important role as equilibria. Our concern in this paper is with Kolmogorov equations of the above type under the assumption that the h_i are no longer independent of time t, but explicitly dependent periodically on t. We will be interested in the existence of nonnegative and positive periodic solutions and in their stability or instability.

Several specific periodic Kolmogorov systems have been studied with regard to certain applications in mathematical ecology. Volterra predator-prey equations with periodic coefficients were studied by the author [2], and more general periodic predator-prey equations have recently been studied by Bardi [1]. Periodic Volterra–Lotka competition equations have been studied by de Mottoni and Schiaffino [7] and the author [3]. Volterra systems of arbitrary order with periodic coefficients were also considered by the author [4]. These papers were all motivated by the obvious fact that cyclic fluctuations in biological and environmental parameters can play an important role in the dynamics of population growth and that such periodic fluctuations might well be modelled by placing periodic coefficients in the classical equations of mathematical ecology.

The mathematical approach taken in this paper to the question of the existence of periodic solutions of periodic Kolmogorov systems is a generalization of that taken in

* Received by the editors November 4, 1980, and in revised form September 29, 1981. This material is based upon work supported by the National Science Foundation under grant MCS-7901307-01.

† Department of Mathematics and Program on Applied Mathematics, University of Arizona, Tucson, Arizona 85721.

previous papers by the author [2], [3], [4] and is based upon very general, abstract bifurcation theorems of Rabinowitz [8]. It will be shown (Theorem 1) for systems of size $n \geq 2$ that under rather general conditions there exists a critical value of the average inherent growth rate of one species, say x_1, at which there bifurcates a "global" branch of solutions which are positive solutions at least locally near the bifurcation point. This branch of solutions (to which, incidentally, there corresponds a branch of locally nonpositive solutions) bifurcates from a positive periodic solution of the "reduced system" of size $n-1$ obtained by eliminating x_1 from the system (i.e., by setting $x_1 \equiv 0$ in the system). Simple examples are given to show that this branch may not stay in the positive cone globally, however. Nonetheless, positive periodic solutions of periodic Kolmogorov systems of size $n \geq 2$ can be built up by repeated bifurcations of this type by applying this result repeatedly, starting from positive periodic solutions of scalar ($n=1$) periodic equations. Scalar equations are dealt with in Theorems 2 through 6. The stability of both the positive branch solutions and the periodic solutions of the reduced system are studied in §3, at least locally near the bifurcation point. As is typically the case in such bifurcation phenomena, it is found that the stability of the positive branch solutions depends on the stability properties of the solution of the reduced system and on the direction of bifurcation. In §4 the special case of planar ($n=2$) periodic Kolmogorov systems is discussed and examples of general periodic predator-prey and two-species competition models are given.

2. Periodic solutions.

(a). We begin by setting up some notation. If $v=\text{col}(v_i)$, $w=\text{col}(w_i)$ are n-vectors, then define $v \cdot w := \sum v_i w_i$ and $v \hat{} w := \text{col}(v_i w_i) = w \hat{} v$. For any $n \times n$ matrix $M=(m_{ij})$, define the $n \times n$ matrix $v \circ M := (v_i m_{ij})$. Note that $v \hat{} (Mw) = (v \circ M)w$.

Denote the Banach space of real valued, continuous p-periodic functions of a real variable t under the supremum norm $|\cdot|_0$ by B_p and denote the k-fold cross product of B_p with itself by B_p^k. Let $\text{av}(x) := p^{-1}\int_0^p x(t)\,dt$ denote the average of $x \in B_p$. Euclidean k-space will be denoted by R^k. A continuum (in a Banach space) is a closed connected set. The boundary of a set Ω will be denoted by $\partial\Omega$.

Consider the following general periodic system of n equations

$$(1)\quad \begin{aligned} &\text{(a)} \quad x' = x[a(t)+f(t,x,y)], \\ &\text{(b)} \quad y' = y \hat{} [b(t)+g(t,x,y)], \end{aligned}$$

where $x=x(t)$ is a scalar valued function of t and $y=y(t)$ is an $(n-1)$-vector valued function of t. For $n \geq 2$ let Ω^n denote an open set in $R \times R^{n-1}$ which contains the origin $(x,y)=(0,0)$. The functions f and g are assumed to satisfy the following hypotheses:

H1. $b \in B_p^{n-1}$, $n \geq 2$, and $f: R \times \Omega^n \to R$, $g: R \times \Omega^n \to R^{n-1}$ are continuous functions which are p-periodic in t and continuously differentiable in x and y with $f(t,0,0) \equiv g(t,0,0) \equiv 0$. Also, $a \in B_p$.

The system

$$(2)\quad y' = y \hat{} [b(t)+g(t,0,y)]$$

of $n-1$ equations will be referred to as the *reduced system*. By a solution of (1) or (2) we always mean continuously differentiable functions. A positive solution $(x(t),y(t))$ of (1) [or $y(t)$ of (2)] is a solution for which $x(t)>0$, $y_i(t)>0$ for all t and i [or $y_i(t)>0$

for all t and i]. Two assumptions concerning the reduced system will be made:

H2. The reduced system (2) of $n-1$ equations, $n\geq 2$, has a positive p-periodic solution $0<y=y_0(t)\in B_p^{n-1}$ such that $(0,y_0(t))\in\Omega^n$ for all t.

H3. The solution $y=y_0(t)$ of (2) in H2 is noncritical, i.e., all Floquet exponents of the linearization of (2) at $y_0(t)$ have nonzero real parts.

More will be said about H2 and H3 below. Note that H2 implies the existence of a nontrivial p-periodic solution of (1) on the boundary of the positive cone in $B_p\times B_p^{n-1}$, namely $(x,y)=(0,y_0)$. We are interested, however, in solutions lying in the interior of this cone.

It is easy to show that if $0\leq(x,y)\in B_p\times B_p^{n-1}$ is a solution of (1) ($0\leq(x,y)$ means that $x\geq 0$ and $y_i\geq 0$ for all i) then

$$\begin{aligned}&\operatorname{av}(a(t)+f(t,x(t),y(t)))=0 \quad \text{if } x>0,\\ &\operatorname{av}(b_i(t)+g_i(t,x(t),y(t)))=0 \quad \text{if } y_i>0.\end{aligned} \tag{3}$$

To see this, given $x>0$, divide (1a) by x and integrate over one period. A similar procedure is applied to (1b) given $y_i>0$.

Note that (1) has the *positivity property*: if a solution of (1) satisfies $x(t_0)>0$, $y(t_0)>0$ for some finite t_0, then $x(t)>0$, $y(t)>0$ for all $t\in(-\infty,+\infty)$. Also, if $x(t_0)=0$ or $y_i(t_0)=0$ for some $1\leq i\leq n$ and finite t_0, then $x(t)=0$ or $y_i(t)=0$ for all $t\in(-\infty,+\infty)$. These facts follow from

$$x(t)=x(t_0)\exp\left[\int_{t_0}^{t}(a(s)+f(s,x(s),y(s)))\,ds\right],$$

$$y_i(t)=y_i(t_0)\exp\left[\int_{t_0}^{t}(b_i(s)+g_i(s,x(s),y(s)))\,ds\right]$$

for $1\leq i\leq n$ and all t.

Define $\Omega_p^n(y_0):=\{(x,y)\in B_p\times B_p^{n-1}:(x(t),y(t)+y_0(t))\in\Omega^n$ for all $t\}$ for $n\geq 2$. This set is an open set in $B_p\times B_p^{n-1}$ which contains $(0,y_0)$. The following is our main theorem concerning the existence of p-periodic solutions of (1).

THEOREM 1. *Assume* H1, H2 *and* H3 *hold. Let* $a_0(t)\in B_p$ *be a given function with* $\operatorname{av}(a_0)=0$. *There exists a continuum* $C\subset\Omega_p^n(y_0)\times R$ *with the following properties*:

(i) $(x,y,\mu)\in C$ *implies that* $(x,y)\in B_p\times B_p^{n-1}$ *solves system* (1) *with* $a(t)=\mu+a_0(t)$;

(ii) $(0,y_0,\mu_0)\in C$, *where*

$$\mu_0:=-\operatorname{av}(f(t,0,y_0(t))); \tag{4}$$

(iii) *either* C *is unbounded or* $\partial(\Omega_p^n(y_0)\times R)\cap C\neq\varnothing$;

(iv) *in a suitably small open neighborhood of* $(0,y_0,\mu_0)$, $C=K^+\cup K^-$, *where* K^+ *and* K^- *are continua for which* $K^+\cap K^-=\{(0,y_0,\mu_0)\}$. *The solutions from* $K^+-\{(0,y_0,\mu_0)\}$ *are positive while those from* $K^--\{(0,y_0,\mu_0)\}$ *satisfy* $x(t)<0$, $0<y(t)<y_0(t)$ *for all* t.

In the case when $\Omega^n=R^n$ (i.e., when f and g are globally defined) and hence $\Omega_p^n(y_0)=B_p\times B_p^{n-1}$, the second alternative in (iii) is to be ruled out with the result that (iii) states simply that C is unbounded. This is the case for Volterra type systems in which the functions f and g are linear in x,y (see [2], [3], [4]). This theorem generalizes [3, Thm. 1(a)].

Proof. Define $z=y-y_0$ and let $a(t)=\mu+a_0(t)$ in (1). This results in the system

$$(5)\quad \begin{aligned} &\text{(a)} && x'=[\mu+a_0(t)+f(t,0,y_0(t))]x+r_1(x,z),\\ &\text{(b)} && z'=[y_0(t)\hat{}g_x(t,0,y_0(t))]x+[b(t)+g(t,0,y_0(t))]\hat{}z\\ &&&\quad +y_0(t)\hat{}[g_y(t,0,y_0)z]+r_2(x,z), \end{aligned}$$

where

$$\begin{aligned} r_1(x,z) &:= x[f(t,x,z+y_0)-f(t,0,y_0)],\\ r_2(x,z) &:= z\hat{}[g_x(t,0,y_0)x+g_y(t,0,y_0)z+r_3(t,x,z)]+y_0(t)\hat{}r_3(t,x,z),\\ r_3(t,x,z) &:= g(t,x,z+y_0)-g(t,0,y_0)-g_x(t,0,y_0)x-g_y(t,0,y_0)z. \end{aligned}$$

The operators $r_1:\Omega_p^n(y_0)\to B_p$ and $r_2:\Omega_p^n(y_0)\to B_p^{n-1}$ are continuous and higher order than linear near $(0,0)\in B_p\times B_p^{n-1}$, i.e., $|r_1(x,z)|_0$ and $|r_2(x,z)|_{0,n-1}=O(|x|_0^2+|z|_{0,n-1}^2)$, where $|z|_{0,n-1}=\Sigma_1^{n-1}|z_i|_0$.

We wish to formulate (5) as an operator equation to which certain global bifurcation theorems apply. To do this we first choose a real constant c such that

$$c\neq \operatorname{av}(f(t,0,y_0(t)))$$

and rewrite (5a) as

$$(5a')\qquad x'=[-c+a_0(t)+f(t,0,y_0(t))]x+[c+\mu]x+r_1(x,z).$$

We are now interested in solving (5a')–(5b) for positive solutions in $B_p\times B_p^{n-1}$.

Let $G_1(t,s)$ and $G_2(t,s)$ be the Green's functions for the linear equations

$$\begin{aligned} x' &= [-c+a_0(t)+f(t,0,y_0(t))]x,\\ z' &= [b(t)+g(t,0,y_0(t))]\hat{}z+y_0(t)\hat{}[g_y(t,0,y_0(t))z] \end{aligned}$$

respectively. $G_1(t,s)$ exists by the choice of c which makes the p-periodic coefficient of the first (scalar) equation have nonzero average. $G_2(t,s)$ exists by H3 since the second equation is the linearization of (2) at $y=y_0$. These Green's functions define compact linear operators $L_1:B_p\to B_p$ and $L_2:B_p^{n-1}\to B_p^{n-1}$ by means of the integrals

$$\begin{aligned} L_1\xi &:= \int_0^p G_1(t,s)\xi(s)\,ds, \qquad \xi\in B_p,\\ L_2\eta &:= \int_0^p G_2(t,s)\eta(s)\,ds, \qquad \eta\in B_p^{n-1}. \end{aligned}$$

The system (5a')–(5b) is equivalent to the pair of operator equations

$$(6)\qquad x=\lambda L_1x+H_1(x,z), \qquad z=L_2[y_0\hat{}g_x(t,0,y_0)]x+H_2(x,z)$$

for $(x,z)\in B_p\times B_p^{n-1}$, where $\lambda=c+\mu$ and $H_1:=L_1r_1:\Omega_p^n(y_0)\to B_p$ and $H_2:=L_2r_2:\Omega_p^n(y_0)\to B_p^{n-1}$ are completely continuous operators of higher order than linear near $(x,z)=(0,0)$. Equations (6) are in turn equivalent to the equations

$$(7)\qquad x=\lambda L_1x+H_1(x,z), \qquad z=\lambda L_2[y_0\hat{}g_x(t,0,y_0)]L_1x+H_3(x,z),$$

where $H_3(x,z):=L_2[y_0\hat{}g_x(t,0,y_0)]H_1(x,z)+H_2(x,z)$ is higher order than linear near $(x,z)=(0,0)$. Finally, (7) can be written in the concise form

$$(8)\qquad w=\lambda Lw+H(w),$$

where $w=(x,z)\in B_p\times B_p^{n-1}$, $L: B_p\times B_p^{n-1}\to B_p\times B_p^{n-1}$ is defined by

$$Lw := \left(L_1 x, L_2[y_0, g_x(t,0,y_0)]L_1 z\right)$$

and $H:\Omega_p^n(y_0)\to B_p\times B_p^{n-1}$ is defined by $H(x,z) := (H_1(x,z), H_3(x,z))$. The operator L is linear and compact and H is completely continuous and of order higher than linear near $w=0$.

Of course, $w=0$ is a trivial solution of (8) (which corresponds to the solution $x=0$, $y=y_0$ of (1)). To find nonzero solution $w\neq 0$ of (8), we can use directly the global bifurcation theorems of Rabinowitz [8], using λ (or what amounts to the same thing μ) as a bifurcation parameter (c being held fixed). Specifically, we will use [8, Thm. 1.25, Cor. 1.12]. To do this it remains to be shown that L has a characteristic value $\lambda=\lambda_0$ of odd multiplicity. We will in fact show that L has one and only one characteristic value and it is real and simple.

The linear equation $w=\lambda Lw$ is equivalent to the linearized, homogeneous p-periodic system obtained by setting $r_1\equiv 0$, $r_2\equiv 0$ and $\mu=\lambda-c$ in (5), namely

$$\begin{aligned} &\text{(a)} \quad x' = [\mu + a_0(t) + f(t,0,y_0(t))]x, \\ &\text{(b)} \quad z' = [y_0(t), g_x(t,0,y_0(t))]x + [b(t)+g(t,0,y_0(t))], z \\ &\qquad\qquad + y_0(t), [g_y(t,0,y_0(t))z]. \end{aligned} \tag{9}$$

This system can be solved for $(x,z)\in B_p\times B_p^{n-1}$ by first solving the simpler scalar equation (9a) for p-periodic $x\in B_p$ and then solving (9b) by means of the formula $z=L_2[y_0, g_x(t,0,y_0)]x$. Clearly, $x\equiv 0$ implies $z\equiv 0$, so (9) has a nontrivial solution if and only if $x\not\equiv 0$, which occurs if and only if $\mu=\mu_0$ as defined by (4). This shows that L has a unique characteristic value $\lambda=\lambda_0 := c+\mu_0$ which by definition of c is nonzero.

Finally we argue that $\lambda=\lambda_0$ is simple. Let $0\neq w_0=(x_0,z_0)\in B_p\times B_p^{n-1}$ be a characteristic solution: $w_0=\lambda_0 Lw_0$. Suppose $w\in B_p\times B_p^{n-1}$ satisfies $(I-\lambda_0 L)^2 w=0$. We wish to show that w is a multiple of w_0. Let $w^* := (I-\lambda_0 L)w$. Then $(I-\lambda_0 L)w^*=0$, which implies that $w^*=mw_0$ for some real $m\in R$. Thus $w^*=m\lambda_0 Lw_0$ and, hence $w=\lambda_0 L(w+mw_0)$, which implies that $w=(x,z)\in B_p\times B_p^{n-1}$ solves the nonhomogeneous linear system

$$\begin{aligned} x' &= [\mu_0 + a_0(t) + f(t,0,y_0)]x + \lambda_0 m x_0, \\ z' &= [b(t)+g(t,0,y_0)], z + y_0(t), [g_y(t,0,y_0)z] + \lambda_0 m z_0. \end{aligned}$$

The Fredholm alternative implies that $\lambda_0 m x_0$ must be orthogonal to the adjoint solution $1/x_0$. This means, because $\lambda_0\neq 0$, that $m=0$. Hence, $w^*=0$ or, in other words, $w=\lambda_0 Lw$, which implies the desired result that w is a multiple of w_0.

The results (i)–(iii) of Theorem 1 now follow from [8, Cor. 1.12] applied to the operator equation (8) on the bounded set $\Omega_p^n(y_0)\cap S(\rho)$, where $S(\rho)$ is the open sphere of arbitrary radius $\rho>0$, center $(0,0)$ in $B_p\times B_p^{n-1}$. Part (iv) follows from [8, Thm. 1.25] (also see §3 below). □

While Theorem 1 guarantees the existence of a global branch of p-periodic solutions of (1) in the sense that the branch is either unbounded or reaches the boundary of the domain of f and g, it asserts the positivity of solutions only on a subcontinuum K^+ of the branch lying in a neighborhood of the "bifurcation point" $(0,y_0,\mu_0)$. While it is shown in [8] that $C=C^+\cup C^-$, $C^+\cap C^- = \{(0,y_0,\mu_0)\}$, where $C^\pm$ are continua which are extensions of K^+ and K^- and that C^+ and C^- satisfy (iii), it does not follow that C^+ necessarily stays in the positive cone. The question which naturally arises then

concerns the nature of this maximal subcontinuum C^+ containing K^+: does C^+ lie in the positive cone or does it "leave" this cone and hence contain nonpositive solutions of (1)? By "leave" the cone we mean here that the intersection of C^+ with the boundary of the positive cone contains a point other than the bifurcation point. In the latter case we can, by the positivity property of solutions of (1), distinguish three possibilities: if C^+ leaves the positive cone then there exists a $(x,y)\in C^+$ for which either

$$\begin{array}{lll} \text{(a)} & x\equiv 0, & y\equiv 0, \quad \text{or} \\ \text{(b)} & x\equiv 0, & 0\leq y\not\equiv y_0 \text{ and } \not\equiv 0 \quad \text{or} \\ \text{(c)} & x>0, & y_i\equiv 0 \quad \text{for some } 1\leq i\leq n. \end{array} \tag{10}$$

If we define K_m^+ to be the maximal subcontinuum of C which consists of nonnegative solutions of (1) and contains K^+, then the following is immediate: *under the hypotheses of Theorem* 1, *either* K_m^+ *satisfies* (iii) *or it leaves the positive cone in the sense that* K_m^+ *contains a solution of* (1) *of one of the three types in* (10).

That K_m^+ can in fact leave the positive cone is easily established by examples. The classical Volterra predator-prey equations and Volterra–Lotka competition equations with constant coefficients and $n=2$ demonstrate this possibility as do the periodic coefficient versions of these equations studied in [2], [3]. In these examples, case (10c) always occurs, i.e., K_m^+ leaves the positive cone through a solution $(x,0,\mu^*)\in B_p\times B_p\times R$, $x>0$. We will give below a simple autonomous example to illustrate not only that K_m^+ leaves the positive cone, but that any of the three possibilities in (10) can, in fact, occur. Furthermore, this example will show that the stability of the positive solutions on K_m^+ can be lost before it leaves the cone. The stability of the positive periodic solutions on K_m^+ will be studied in §3 below.

Consider the following autonomous system of two ($n=2$) equations with constant coefficients:

$$\begin{gathered} x'=x(-\gamma+y), \qquad y'=y[(y-\beta)(\alpha-y)-x], \\ \gamma>0, \quad \beta>0, \quad \alpha<\beta. \end{gathered} \tag{11}$$

This system serves as a model for predator-prey interaction in which the per capita growth rate of the predator x is an increasing, linear function of prey population size y and that of the prey is such that the prey zero isocline has a "hump" (see Rosenzweig [9]). The "hump" lies in the positive quadrant if and only if $(\alpha+\beta)/2>0$. Equations (11) have the form (1) with $n=2$ and

$$a(t)\equiv -\gamma, \quad f\equiv y, \quad b(t)\equiv -\alpha\beta, \quad g\equiv(\alpha+\beta)y-y^2-x.$$

In Theorem 1, $a_0(t)\equiv 0$ and $\mu=-\gamma$ with

$$y_0(t)\equiv\beta \quad \text{and} \quad \mu_0=-\beta.$$

The continuum C of periodic solutions (equilibria in this autonomous example) is explicitly given by

$$x(t)\equiv(\gamma-\beta)(\alpha-\gamma), \quad y(t)\equiv\gamma, \quad \mu=-\gamma. \tag{12}$$

Thus, there are three cases.

1) When $\alpha=0$, the subcontinuum K_m^+ of C given by (12) for $0<\gamma<\beta$ connects the bifurcation point $(x,y,\mu)=(0,\beta,-\beta)$ with $(x,y,\mu)=(0,0,0)$ and hence case (10a) occurs.

2) When $\alpha>0$, K_m^+ is given by (12) with $\alpha<\gamma<\beta$; it leaves the positive quadrant at $(x,y,\mu)=(0,\alpha,-\alpha)$ and (10b) occurs.

3) When $\alpha<0$, K_m^+ is given by (12) with $0<\gamma<\beta$. In this case, however, it leaves the positive quadrant at $(x,y,\mu)=(-\alpha\beta,0,0)$ so that the case (10c) is seen to occur.

One can also show by standard linearization and eigenvalue techniques that when $(\alpha+\beta)/2>0$ the stability of the positive equilibrium (12) is lost as γ decreases through $(\alpha+\beta)/2$ (i.e., μ increases through $-(\alpha+\beta)/2$) and that a classical Hopf bifurcation occurs.

It is possible, of course, to set down simple conditions which rule out one or more of the possibilities in (10). For example, *if* $\operatorname{av}(b)\neq 0$, *then* (10a) *cannot occur*. This is because if $(x_n,y_n,\mu_n)\in K_m^+$ are such that $(x_n,y_n,\mu_n)\to(0,0,\mu^*)$, $\mu^*\in R$, then by (3)

$$0=\operatorname{av}\big(b(t)+g(t,x_n(t),y_n(t))\big),$$

which in the limit as $n\to+\infty$ implies the contradiction that $\operatorname{av}(b)=0$. Also, *if the reduced system* (2) *has a unique positive, p-periodic solution, then* (10b) *cannot occur*. This is because $(x_n,y_n,\mu_n)\in K_m^+$ such that $(x_n,y_n,\mu_n)\to(0,y,\mu^*)$, $y\neq y_0$, $\mu^*\in R$, implies that $y\in B_p$ solves the reduced system, a contradiction.

A simple example of the case when the subcontinuum K_m^+ of positive solutions does not leave the positive cone is given by

$$x'=x(\mu-y),\quad \mu\in R,\qquad y'=y(1+x-y)$$

for which $y_0(t)\equiv 1$, $\mu_0=1$ and $K_m^+=\{(\mu-1,\mu,\mu):\mu>\mu_0=1\}$.

Not only can C^+ leave the positive cone, but it can "leave and re-enter" the positive cone and it can even do this infinitely often. An example (again autonomous) is given by the system

$$x'=x(\mu-y),\quad \mu\in R,\qquad y'=y(x+\sin y)$$

for which $y_0(t)\equiv\pi$, $\mu_0=\pi$ and $C^+=\{(-\sin\mu,\mu,\mu):\mu>\pi\}$ which yields positive solutions for $\mu\in((2n-1)\pi,2n\pi)$ for $n=1,2,3,\cdots$. In this example $K_m^+=\{(-\sin\mu,\mu,\mu):\pi<\mu<2\pi\}$ leaves the cone at a solution of the form (10b), namely $(0,2\pi,2\pi)$. Note that in this case $C^-=\{(-\sin\mu,\mu,\mu):\mu<\pi\}$ contains no positive solutions. If, on the other hand, we take $y_0(t)\equiv 2\pi$, $\mu_0=2\pi$, then $C^+=\{(-\sin\mu,\mu,\mu):\mu<2\pi\}$ and $K^+=\{(-\sin\mu,\mu,\mu):\pi<\mu<2\pi\}$. Now $C^-=\{(-\sin\mu,\mu,\mu):\mu>2\pi\}$ contains positive solutions for $\mu\in((2n-1)\pi,2n\pi)$, $n=2,3,\cdots$.

(b) Hypothesis H2 assumes that the reduced system (2) associated with (1) has a positive p-periodic solution. This hypothesis itself can be fulfilled by use of Theorem 1 with n replaced by $n-1$ provided $n\geq 3$ and a reduced system of this reduced system has a positive p-periodic solution satisfying H3, etc. In this way positive p-periodic solutions of a general p-periodic Kolmogorov system

$$x_i'=x_ih_i(t,x_1,\cdots,x_{n-1},x_n) \tag{13}$$

can be built up from a positive p-periodic solution of a single scalar, $(n-1)$-fold reduced equation associated with one of the equations, say

$$x_n'=x_nh_n(t,0,\cdots,0,x_n), \tag{14}$$

by repeated bifurcations as given by Theorem 1. For this reason it is important to consider the scalar case $n=1$, as we do below, in order to start this algorithmic procedure. The main difficulty in carrying out this procedure is with the assumption H3 that at each step the periodic solution obtained is noncritical. While one can at least say that this hypothesis is "generic", no simple general criteria exist for testing this hypothesis when $n\geq 3$. In the case of linear functions f and g this requirement can be met by an assumption of strong diagonal dominance [4]. In §3 a criterion valid at least

locally near the point of bifurcation will be obtained ($\mu_1 \neq 0$). For $n=2$ the reduced equation is a scalar equation and H3 reduces to the nonvanishing of the average of the coefficient. This case will be considered in §4. We now turn to the scalar case $n=1$.

Consider the scalar periodic equation

$$y' = y(b(t) + g(t,y)) \tag{15}$$

under the assumption

H4. $g: R \times (\alpha, \beta) \to R$, $-\infty \leq \alpha < 0 < \beta \leq +\infty$, is continuous in $(t,y) \in R \times (\alpha, \beta)$, is continuously differentiable in $y \in (\alpha, \beta)$, is p-periodic in t and satisfies $g(t,0) \equiv 0$.

Define $\Omega_p := \{y \in B_p : y(t) \in (\alpha, \beta) \text{ for all } t\}$, an open set in B_p which contains $y(t) \equiv 0$.

THEOREM 2. *Let $b_0(t) \in B_p$, $\operatorname{av}(b_0) = 0$, be given and assume that* H4 *holds. Then there exists a continuum $C \subset \Omega_p \times R$ with the following properties*:

(i) *$(y, \mu) \in C$ implies y solves* (15) *with $b(t) = b_0(t) + \mu$;*

(ii) $(y, \mu) = (0,0) \in C$;

(iii) *either C is unbounded or $C \cap \partial(R \times \Omega_p) \neq \varnothing$;*

(iv) *in a suitable small open neighborhood of* $(0,0)$, $C = K^+ \cup K^-$, $K^+ \cap K^- = \{(0,0)\}$, *where $K^{\pm}$ are continua. The solutions from $K^+ - \{(0,0)\}$ are positive while those from $K^- - \{(0,0)\}$ are negative.*

Proof. This theorem is proved in virtually the same way as Theorem 1 (but with $n=1$, of course) except that $y_0(t)$, the p-periodic solution of the reduced system (2), is replaced by the identically zero solution of (15) and $\Omega_p^n(y_0)$ is replaced by Ω_p. Otherwise, the details are the same. □

If g is globally defined (i.e., $(\alpha, \beta) = R$ so that $\Omega_p = B_p$), then the second alternative in (iii) is to be eliminated with the result that C is unbounded.

As in the case $n \geq 2$, $C = C^+ \cup C^-$, $C^+ \cap C^- = \{(0,0)\}$, where the continua C^+ and C^- contain K^+ and K^- respectively and we again ask whether C^+ can leave the positive cone. By the positivity property of solutions of (15), C^+ can leave the positive cone only if $(0, \mu) \in C^+$ for some $\mu \in R$, $\mu \neq 0$. But this is impossible, because $(0, \mu) \in C^+$ implies that $(0, \mu)$ is a bifurcation point and, hence, $\mu = 0$ (since $w = \lambda Lw$ in the proof of Theorem 1 and hence of Theorem 2 has only one characteristic value $\lambda_0 = c$, hence $\mu_0 = 0$). Thus, *under the hypotheses of Theorem* 2, $C = C^+ \cup C^-$, $C^+ \cap C^- = \{(0,0)\}$, *where $C^{\pm}$ are continua for which $C^+ - \{(0,0)\}$ and $C^- - \{(0,0)\}$ contain positive and negative solutions of* (15), *respectively. Both $C^{\pm}$ satisfy the alternative* (iii) *of Theorem* 2. *Thus, if $\beta = +\infty$, then C^+ is unbounded.* (See [8, Thm. 1.27].)

Define the *spectrum* $\Sigma^+ \subset R$ associated with C^+ to be the range of the projection mapping $C^+ \to R$ defined by $(y, \mu) \to \mu$. Thus, for $\mu \in \Sigma^+$ there exists at least one positive p-periodic solution of (15) with $a(t) = a_0(t) + \mu$. Define $S^+ \subset B_p$ to be the range of the projection $C^+ \to B_p$ defined by $(y, \mu) \to y$. Both Σ^+ and S^+ are continua which contain $\mu = 0$ and $y \equiv 0$ respectively. Σ^+ is of course an interval which is possibly open, closed or half open or closed and is possibly infinite. Let R^+ denote the positive reals.

Theorems 3–6 give results concerning S^+ and the spectrum Σ^+ of (15).

THEOREM 3. *Assume that the hypotheses of Theorem* 2 *hold. Suppose that C^+ is unbounded (which happens if $\beta = +\infty$). Then either S^+ is unbounded (which happens if $\beta = +\infty$) or else S^+ is bounded in which case the spectrum Σ^+ is unbounded, $\beta < +\infty$ and there exist sequences $(y_j, \mu_j) \in C^+$ and $t_j \in [0,p]$ such that $t_j \to t_0 \in [0,p]$, $|y_j|_0 \to \beta$ and $|\mu_j| + |g(t_j, |y_j|_0)| \to +\infty$.*

Let Γ be the set of all convergent sequences $(t_j,\xi_j)\in[0,p]\times(0,\beta)$, $0<\beta\le+\infty$, for which $\xi_j\to\beta$ and $\lim g(t_j,\xi_j)$ exists (but is not necessarily finite). Define $g_{\inf}$, $g_{\sup}\in[-\infty,+\infty]$ by

$$g_{\inf}:=\inf_\Gamma\{\lim g(t_j,\xi_j)\}\le\sup_\Gamma\{\lim g(t_j,\xi_j)\}:=g_{\sup}.$$

THEOREM 4. *Assume the hypotheses of Theorem* 2. *Then* Σ^+ *is an interval whose closure contains* $(-g_{\inf}-\min b_0(t),0]$ *if* $g_{\inf}\ge-\min b_0(t)\ge0$ *or* $[0,-g_{\sup}-\max b_0(t))$ *if* $g_{\sup}\le-\max b_0(t)\le0$.

Proofs. In everything to follow, if $(y_j,\mu_j)\in C^+$, then t_j denotes a real number which satisfies $t_j\in[0,p]$, $y_j'(t_j)=0$, $y_j(t_j)=|y_j|_0$. Since $y_j'/y_j=b_0(t)+\mu_j+g(t,y_j(t))$ for all t, it follows that

$$\mu_j=-b_0(t_j)-g(t_j,|y_j|_0). \tag{16}$$

Without loss of generality, it is assumed that $t_j\to t_0\in[0,p]$. Suppose that C^+ is unbounded, which of course implies that either S^+ or Σ^+ is unbounded.

Proof of Theorem 3. If S^+ is bounded, then Σ^+ is unbounded which means that there exists a sequence $(y_j,\mu_j)\in C^+$ for which the μ_j are unbounded. Choosing a subsequence if necessary, assume that $|\mu_j|\to+\infty$. Since $|y_j|_0$ is a bounded sequence, we can assume (by choosing another subsequence if necessary) that $|y_j|_0\to\sigma$, where $\sigma\ge0$ is finite and $\sigma\le\beta$. The conclusions of Theorem 3 follow by taking limits in (16) and noting that if $\sigma\ne\beta$ (which happens if $\beta=+\infty$), then a contradiction results because the limit on the right-hand side would be finite by the continuity of g while $|\mu_j|\to+\infty$. If $\beta<+\infty$ the only way out of the contradiction is the conclusion that S^+ is unbounded or that S^+ is bounded, but $\sigma=\beta<+\infty$.

Proof of Theorem 4. Using Theorem 2(iii) (for C^+) and Theorem 3, choose a sequence $(y_j,\mu_j)\in C^+$ for which $\xi_j:=|y_j|_0\to\beta\le+\infty$. Extracting a subsequence if necessary assume $\mu_j\to\mu^*\in[-\infty,+\infty]$. From (16)

$$\mu^*=-b_0(t_0)-\lim g(t_j,\xi_j),$$

and hence Σ^+ is an interval whose closure contains both 0 (by Theorem 2(ii)) and μ^*. Since

$$-\max b_0(t)-g_{\sup}\le\mu^*\le-\min b_0(t)-g_{\inf},$$

the result follows. □

Example 1 (A generalized logistic equation). Consider (15) when g satisfies H4 with $\beta=+\infty$. Suppose that $g(t,y)\le0$ for all $(t,y)\in[0,p]\times[0,+\infty)$ and that $\lim g(t_j,\xi_j)=-\infty$ for any convergent sequence $(t_j,\xi_j)\in[0,p]\times[0,+\infty)$ for which $\xi_j\to+\infty$. Then from the above results we conclude that (15) has at least one positive p-periodic solution whenever $\mu=\operatorname{av}(b(t))>0$ (see Theorem 4 with $g_{\sup}=-\infty$) which is unbounded in norm as $\operatorname{av}(b(t))\to+\infty$ (Theorem 3). The periodic logistic, in which $g(t,y)=-c(t)y$, $c\in B_p$, $c(t)>0$, is a simple example.

THEOREM 5. *Assume the hypotheses of Theorem* 2 *hold. If* $g(t,y)$ *is bounded for* $(t,y)\in[0,p]\times[0,\infty)$, *then* Σ^+ *is bounded.*

Proof. $(y,\mu)\in C^+$ and (16) imply that $|\mu|\le|b_0|_0+\max|g(t,y)|$. □

EXAMPLE 2. Let $g(t,y)=-c(t)y/(1+y)$, $c\in B_p$, $c(t)>0$. Then $g_{\inf}=-\max c(t)$, $g_{\sup}=-\min c(t)$ and (15) has by Theorem 4 at least one positive p-period solution for $b(t)=b_0(t)+\mu$, $\operatorname{av}(b_0)=0$, $\mu\in[0,\min c(t)-\max b_0(t)]$ provided $\min c(t)>\max b_0(t)$. By Theorem 5 the spectrum Σ^+ is bounded.

THEOREM 6. *Assume the hypotheses of Theorem* 2 *hold. Suppose that* $\beta<+\infty$ *and that the sequence* $g(t_j,\xi_j)$ *is unbounded for all convergent sequences* $(t_j,\xi_j)\in[0,p]\times(0,\beta)$ *for which* $\xi_j\to\beta$. *Then the spectrum* Σ^+ *is unbounded and* S^+ *is bounded.*

Proof. Since $y\in S^+$ implies $|y|_0<\beta$, S^+ is bounded. Suppose that Σ^+ is bounded. Then C^+ is bounded, and by Theorem 2(iii) applied to C^+, it follows that $C^+\cap\partial(\Omega_p\times R)\neq\varnothing$. Thus, there exists a sequence $(y_j,\mu_j)\in C^+$ for which $\xi_j=|y_j|_0\to\beta$. From (16) and the stated hypothesis on g, it follows that μ_j is unbounded, a contradiction to the boundedness of Σ^+. □

Example 3. Let $g(t,y)=-c(t)\tan y$, $0<c(t)\in B_p$. Then $\alpha=-\pi$, $\beta=\pi$ and Theorem 6 implies that Σ^+ is unbounded and S^+ is bounded. Since $g_{\sup}=-\infty$ Theorem 4 implies that there exists at least one positive p-periodic solution of (15) when $\mu=\operatorname{av}(b(t))>0$. Theorem 6 shows that for a sequence $\mu_j=\operatorname{av}(b(t))\to+\infty$ there are solutions $y_j\in B_p$, $y_j(t)>0$, of (15) for which $|y_j|_0\to\pi$.

3. Local analysis and stability. In this section the (local asymptotic) stability of the p-periodic solutions of (1) lying in a neighborhood of the bifurcation point $(x,y,\mu)=(0,y_0,\mu_0)$ will be studied as it depends on the bifurcation parameter $\mu=\operatorname{av}(a)$. Both the solutions $(0,y_0,\mu)$, $\mu\in R$, and those lying on C in a neighborhood of the bifurcation point $(0,y_0,\mu_0)$ will be considered.

THEOREM 7. *Assume* H1, H2, H3 *hold and that* $n\geq 2$, *but with the added assumption that* f *and* g *are twice continuously differentiable in* x *and* y.

(i) *If one of the Floquet exponents of the linearization of the reduced system* (2) *at* $y_0(t)$ *has positive real part, then* $(x,y)=(0,y_0)\in B_p\times B_p^{n-1}$ *is an unstable solution of* (1) (*for any* $a\in B_p$).

(ii) *If all Floquet exponents of the linearization of* (2) *have negative real parts, then* $(x,y)=(0,y_0)\in B_p\times B_p^{n-1}$ *is* (*locally uniformly asymptotically*) *stable as a solution of* (1) *when* $\mu=\operatorname{av}(a)<\mu_0$ *and is unstable when* $\mu>\mu_0$ (*where* μ_0 *is given by* (4)).

Proof. The linearization of (1) at $(x,y)=(0,y_0)$ yields the linear system (9), a block triangular system whose Floquet exponents are those of the reduced system (2) plus

$$\operatorname{av}s\big(\mu+a_0(t)+f(t,0,y_0(t))\big)=\mu-\mu_0.$$

The theorem follows from standard linearization theorems for periodic systems of ordinary differential equations. □

In order to study the stability of the solutions of (1) lying on the bifurcating branch C near the point of bifurcation, Lyapunov–Schmidt small parameter expansions of the solutions and their Floquet exponents will be made. Thus,

$$\begin{aligned} x(t)&=x_1(t)\varepsilon+x_2(t)\varepsilon^2+x_3(t,\varepsilon)\varepsilon^2,\\ y(t)&=y_0(t)+y_1(t)\varepsilon+y_2(t)\varepsilon^2+y_3(t,\varepsilon)\varepsilon^2,\\ \mu&=\mu_0+\mu_1\varepsilon+\mu_2(\varepsilon)\varepsilon, \end{aligned}\tag{17}$$

where $|x_3(t,\varepsilon)|_0=O(|\varepsilon|)$, $|y_3(t,\varepsilon)|_{0,n-1}=O(|\varepsilon|)$ and $|\mu_2(\varepsilon)|=O(|\varepsilon|)$. To rigorously establish that the solutions on C near the bifurcation point $(0,y_0,\mu_0)$ corresponding to $\varepsilon=0$ have the form (17) is a routine application of classical Lyapunov–Schmidt techniques and will hence not be given here. To do this requires two continuous derivative for f and g. (See [5], [6] for abstract theorems particularly suited to the operator formulation of (1) given in the proof of Theorem 1 which establish the validity of (17).) The plan here, of course, is to determine the lower order coefficients x_1, y_1 and μ_1 in the

expansions (17) and then in an expansion for the Floquet exponents of the solution (17). This is done by substituting (17) into (1) and equating coefficients of like powers of ε.

The lowest order terms in ε in (1) yield the reduced equation (2) which is satisfied by $y_0(t)$ by definition. The first order terms in ε yield the linear system (1) with $\mu=\mu_0$:

$$(18)\quad \begin{aligned} &\text{(a)} && x_1' = x_1[\mu_0 + a_0(t) + f(t,0,y_0(t))], \\ &\text{(b)} && y_1' = [y_0(t)\hat{}\,g_x(t,0,y_0(t))]x_1 + y_0(t)\hat{}\,g_y(t,0,y_0(t))y_1 \\ & && \quad + [b(t)+g(t,0,y_0(t))]\hat{}\,y_1 \end{aligned}$$

for $(x_1,y_1)\in B_p\times B_p^{n-1}$. Thus

$$(19)\quad \begin{aligned} x_1(t) &= \exp\left(\int_0^t [\mu_0 + a_0(s) + f(s,0,y_0(s))]\,ds\right) > 0, \\ y_1(t) &= \int_0^p G_2(t,s)[y_0(s)\hat{}\,g_x(s,0,y_0(s))]x_1(s)\,ds. \end{aligned}$$

Finally, the second order terms in ε from (1a) only yield the scalar equation

$$\begin{aligned} x_2' = {} & x_2[\mu_0 + a_0(t) + f(t,0,y_0(t))] \\ & + x_1(t)[\mu_1 + f_x(t,0,y_0(t))x_1(t) + f_y(t,0,y_0(t))\cdot y_1(t)] \end{aligned}$$

for $x_2\in B_p$. This equation is a nonhomogeneous version of (18a) and consequently the nonhomogeneous term must be orthogonal to the adjoint solution $1/x_1(t)$. This orthogonality condition yields a formula for μ_1:

$$(20)\quad \mu_1 = -\mathrm{av}[f_x(t,0,y_0(t))x_1(t) + f_y(t,0,y_0(t))\cdot y_1(t)].$$

The sign of μ_1 determines the local "direction" of bifurcation of the branches K^+ and K^- (given by $\varepsilon>0$ and $\varepsilon<0$, respectively) in Theorem 1. This is provided $\mu_1\neq 0$, of course.

Let $N_n(\rho)$ denote the open ball in $B_p\times B_p^{n-1}\times R$ of radius $\rho>0$ and center $(0,y_0,\mu_0)$.

THEOREM 8. *In addition to the hypotheses of Theorem* 7, *assume* $\mu_1\neq 0$. *There exists a* $\rho>0$ *such that the following statements hold*:

(i) $(x,y,\mu)\in C^{\pm}\cap N_n(\rho)$ *implies* $\operatorname{sign}(\mu-\mu_0)=\pm\operatorname{sign}\mu_1$.

(ii) *If at least one Floquet exponent of the linearization of the reduced system* (2) *at* $y_0(t)$ *has a positive real part* (*hence*, $y_0(t)$ *is an unstable solution of* (2)), *then for* $(x,y,\mu)\in C^{\pm}\cap N_n(\rho)-\{(0,y_0,\mu_0)\}$ *the solution* $(x,y)\in B_p\times B_p^{n-1}$ *of* (1) *with* $a(t)=a_0(t)+\mu$ *is unstable*.

(iii) *On the other hand, if all Floquet exponents of the linearization of* (2) *at* $y_0(t)$ *have negative real parts* (*hence*, $y_0(t)$ *is a stable solution of* (2)), *then for* $(x,y,\mu)\in C^+\cap N(\rho)-\{(0,y_0,\mu_0)\}$ *the positive solution* $(x,y)\in B_p\times B_p^{n-1}$ *of* (1) *with* $a(t)=a_0(t)+\mu$ *is*:

(*locally uniformly asymptotically*) *stable if* $\mu_1>0$,

unstable if $\mu_1<0$.

(*For the nonpositive solutions from* $(x,y,\mu)\in C^-\cap N(\rho)-\{(0,y_0,\mu_0)\}$, *the inequalities are reversed*.)

Proof.

(i) This follows immediately from (17) and $x_1(t)>0$.

(ii) The linearization of (1) at the branch solution (17) yields

$$(21)\qquad \begin{aligned} z_1' &= [\mu + a_0(t) + f(t,x,y) + xf_x(t,x,y)]z_1 + [xf_y(t,x,y)]z_2, \\ z_2' &= [y \hat{} g_x(t,x,y)]z_1 + [b(t) + g(t,x,y)] \hat{} z_2 + y \hat{} [g_y(t,x,y)z_2], \end{aligned}$$

where $z_1(t)$ is a scalar valued function and $z_2(t)$ is an $(n-1)$-vector valued function. The stability properties of the branch solutions (17) are determined by the Floquet exponents of (21) which, because the coefficients of (21) depend on ε through (17), are functions of ε. For $\varepsilon = 0$, (21) reduces to (9), which has by assumption a Floquet multiplier with positive real part. Thus, (21) has such a Floquet exponent for sufficiently small ε.

(iii) In this case (9) has $n-1$ Floquet exponents with negative real parts. The remaining exponent is $\operatorname{av}(\mu_0 + a_0(t) + f(t,0,y_0(t))) = 0$. Our problem is to determine where in the complex plane this latter Floquet exponent is located for small $|\varepsilon| > 0$.

Now e is a Floquet exponent of (21) if and only if the linear homogeneous system for $(z,w) \in B_p \times B_p^{n-1}$,

$$(22)\qquad \begin{aligned} &\text{(a)} \quad z' = [\mu + a_0(t) + f(t,x,y) + xf_x(t,x,y) - e]z + [xf_y(t,x,y)]w, \\ &\text{(b)} \quad w' = [y \hat{} g_x(t,x,y)]z + [(b(t) + g(t,x,y)) \circ I + y \circ g_y(t,x,y) - eI]w, \end{aligned}$$

has a nontrivial p-periodic solution. Here I is the $(n-1)\times(n-1)$ identity matrix. (See the Appendix.)

The coefficients of (21) are at least twice continuously differentiable in ε and consequently so is $e = e(\varepsilon)$. We write

$$(23)\qquad e = e(\varepsilon) = e_1\varepsilon + e_2(\varepsilon)\varepsilon, \qquad |e_2(\varepsilon)| = O(|\varepsilon|)$$

for that exponent which vanishes at $\varepsilon = 0$ and look for a nontrivial p-periodic solution of (22) of the form

$$(24)\qquad \begin{aligned} z &= z_1(t) + z_2(t)\varepsilon + z_3(t,\varepsilon)\varepsilon \in B_p, \qquad |z_3|_0 = O(|\varepsilon|), \\ w &= w_1(t) + w_2(t)\varepsilon + w_3(t,\varepsilon)\varepsilon \in B_p^{n-1}, \qquad |w_3|_{0,n-1} = O(|\varepsilon|). \end{aligned}$$

It is a straightforward application of Lyapunov–Schmidt methods to show that (22) has a nontrivial solution of the form (24) for e given by (23) and, hence, these details will not be given. Instead, only enough coefficients in these expansions will be calculated in order to find a formula for the crucial coefficient e_1.

Substitute (23) and (24) into (22). The lowest order terms yield a linear system identical to (18) and hence $z_1 = x_1$ and $w_1 = y_1$ as given by (19). The coefficients of the first order ε terms in only equation (22a) yield the scalar, linear nonhomogeneous equation

$$\begin{aligned} z_2' = {} & [\mu_0 + a_0(t) + f(t,0,y_0(t))]z_2 + [x_1(t)f_y(t,0,y_0(t))] \cdot y_1 \\ & + [\mu_1 + x_1(t)f_x(t,0,y_0(t)) + y_1(t) \cdot f_y(t,0,y_0(t)) + x_1(t)f_x(t,0,y_0(t)) - e]x_1(t) \end{aligned}$$

for $z_2 \in B_p$ whose nonhomogeneous term must be orthogonal to the adjoint solution $1/x_1(t)$. Making use of the definition (20) of μ_1, one finds that this orthogonality condition reduces to

$$\operatorname{av}[y_1(t) \cdot f_y(t,0,y_0(t)) + x_1(t)f_x(t,0,y_0(t)) - e_1] = 0$$

or $e_1 = -\mu_1$. Thus for small $|\varepsilon|$ we have from (i) that

$$\operatorname{sign} e(\varepsilon) = \operatorname{sign} e_1 = -\operatorname{sign} \mu_1 = \mp \operatorname{sign}(\mu - \mu_0).$$

Since the branch solution is stable for $e(\varepsilon)<0$ and unstable for $e(\varepsilon)>0$, the result follows. □

Note that when $\mu_1 \neq 0$ and the Floquet exponents of the linearization of the reduced system (2) have negative real parts, then Theorems 7(ii) and 8(iii) show that a typical "exchange of stability" from the "trivial" solution $(0, y_0) \in B_p \times B_p^{n-1}$ to the branch solutions as the bifurcation parameter μ increases through the critical value μ_0.

It also follows from Theorem 8 and its proof that the solutions on the bifurcating branch C are noncritical, at least in a neighborhood of the bifurcation point $(0, y_0, \mu_0)$, provided $\mu_1 \neq 0$. As discussed in the previous section, this fact allows the repeated application of Theorem 1 in order to build up periodic solutions of general periodic Kolmogorov systems (13).

The scalar case (15) is easier to consider. The Lyapunov–Schmidt expansions

$$\begin{aligned} &y(t) = y_1(t)\varepsilon + y_2(t,\varepsilon)\varepsilon, \qquad |y_2|_0 = O(|\varepsilon|), \\ &\mu = \mu_1\varepsilon + \mu_2(\varepsilon)\varepsilon, \qquad |\mu_2| = O(|\varepsilon|) \end{aligned} \tag{25}$$

substituted into (15) with $b(t) = b_0(t) + \mu$, $b_0 \in B_p$ and $\operatorname{av}(b_0) = 0$, yield

$$y_1(t) = \exp\left[\int_0^t b_0(s)\,ds\right], \qquad \mu_1 = -\operatorname{av}\left[g_y(t,0)y_1(t)\right]. \tag{26}$$

A linearization of (15) at (25) yields

$$z' = \left[b(t) + g(t,y) + yg_y(t,y)\right]z,$$

whose Floquet exponent is

$$\operatorname{av}\left[b(t) + g(t,y) + yg_y(t,y)\right] = -\mu_1\varepsilon + O\left(|\varepsilon|^2\right).$$

Let $N(\rho)$ denote the open ball in B_p of radius ρ centered at 0.

THEOREM 9. *Assume that* H4 *holds and that* $\operatorname{av}[g_y(t,0)\exp(\int_0^t b_0(s)\,ds)] \neq 0$.

(i) *The trivial solution* $y \equiv 0$ *of the scalar equation* (15) *is* (*locally uniformly asymptotically*) *stable if* $\mu = \operatorname{av}(b) < 0$ *and unstable if* $\mu = \operatorname{av}(b) > 0$.

(ii) *There is a* $\rho > 0$ *such that if* $(y,\mu) \in C^+ \cap N(\rho) - \{(0,0)\}$ *the positive solution* $y \in B_p$ *of* (15) *with* $b(t) = b_0(t) + \mu$ *is*:

(*locally uniformly asymptotically*) *stable if* $\operatorname{av}\left[g_y(t,0)\exp\left(\int_0^t b_0(s)\,ds\right)\right] < 0$,

unstable if $\operatorname{av}\left[g_y(t,0)\exp\left(\int_0^t b_0(s)\,ds\right)\right] > 0$.

(*For the negative solutions from* $(y,\mu) \in C^- \cap N(\rho) - \{(0,0)\}$, *the inequalities are reversed.*)

The hypotheses of Theorems 8 and 9 require that $\mu_1 \neq 0$. If μ_1 should equal zero, then coefficients of higher order terms in the ε series expansions of the solution, the bifurcation parameter μ and the Floquet exponent e would have to be computed in order to determine the direction of bifurcation and the stability of the bifurcating solution branch. While this is in principle a straightforward repeated application of the Fredholm alternative, it quickly becomes tedious in a general setting. Although for special cases (such as the case when the lowest order terms in f and g are of order ≥ 2)

one can without too much difficulty write formulas for these coefficients, we will refrain from doing so here.

4. Systems in the plane and applications to theoretical ecology. In the subject of theoretical ecology, systems in the plane are of fundamental importance. It is, thus, of interest to consider the results of the previous sections for the special case $n=2$ and to apply them in particular to the basic types of systems which model the ecologically fundamental predator-prey and two species competition interactions.

(a) *Planar systems.* Consider the pair of scalar equations

$$x'=x[a(t)+f(t,x,y)], \qquad y'=y[b(t)+g(t,x,y)] \tag{27}$$

and the related reduced equation

$$y'=y[b(t)+g(t,0,y)]. \tag{28}$$

Assume that $y=y_0(t)\in B_p$ is a positive nontrivial solution of (28). Theorems 2–6 and 9 apply to the question of the existence of $y_0(t)$. In order to apply Theorems 1, 7 and 8 with $n=2$ to (27), the hypothesis H3 of noncriticality must be met by $y_0(t)$. Since the reduced equation (28) is scalar, $y_0(t)$ is noncritical if and only if

$$\operatorname{av}\big[y_0(t)g_y(t,0,y_0(t))\big]\neq 0. \tag{29}$$

The following theorem summarizes the application of Theorems 1, 7 and 8 to the planar system (27).

THEOREM 10. *Assume that* H1 *holds with* $n=2$. *Assume further that the scalar equation* (28) *has a positive solution* $0<y_0(t)\in B_p$ *for which* (29) *holds.*

(i) *Then the conclusions of Theorem* 1 *hold* (*with* $n=1$) *for* (27).

(ii) *Assume that f and g are twice continuously differentiable in x and y. If*

$$\operatorname{av}\big[y_0(t)g_y(t,0,y_0(t))\big]>0, \tag{30}$$

then $(0,y_0)\in B_p\times B_p$ *is an unstable solution of* (27). *On the other hand, if*

$$\operatorname{av}\big[y_0(t)g_y(t,0,y_0(t))\big]<0, \tag{31}$$

then $(0,y_0)\in B_p\times B_p$ *is an unstable solution of* (27) *for* $\mu=\operatorname{av}(a)>\mu_0:=-\operatorname{av}(f(t,0,y_0(t)))$ *and is a* (*locally uniformly asymptotically*) *stable solution for* $\mu<\mu_0$.

(iii) *Suppose that* $\mu_1\neq 0$ (*where* μ_1 *is given by* (20)). *Then there exists a* $\rho>0$ *such that* $(x,y,\mu)\in C^{\pm}\cap N_2(\rho)$ *implies* $\operatorname{sign}(\mu-\mu_0)=\pm\operatorname{sign}\mu_1$; (30) *implies that solutions* $(x,y)\in B_p\times B_p$ *of* (27) *corresponding to* $(x,y,\mu)\in C^{\pm}\cap N_2(\rho)-\{(0,y_0,\mu_0)\}$ *are unstable*; *and* (31) *implies that positive solutions* $(x,y)\in B_p\times B_p$ *of* (27) *corresponding to* $(x,y,\mu)\in C^{+}\cap N_2(\rho)-\{(0,y_0,\mu_0)\}$ *are unstable for* $\mu_1<0$ *and* (*locally uniformly asymptotically*) *stable for* $\mu_1>0$. (*For nonpositive solutions from* $C^{-}\cap N_2(\rho)-\{(0,y_0,\mu_0)\}$ *the inequalities are reversed.*)

(b) *Applications.* We consider (27) to be a model of the growth of two interacting species whose population densities are given in some units as functions of time $x=x(t)$, $y=y(t)$. The p-periodic coefficients $a(t)$ and $b(t)$ represent the inherent growth rates of each species x and y respectively in the absence of the other and in the absence of any self-inhibitory effects on its own per unit growth rate. The functions f and g, which are assumed globally defined and twice continuously differentiable in x and y, describe what effects that interspecies and intraspecies interactions have on per unit growth rates.

Assume that at least one of the species, say y, has a positive average inherent growth rate: $\operatorname{av}(b(t))>0$, so that it will, in the absence of inter- and intraspecies effects

on growth rates ($g\equiv 0$), exhibit unlimited growth. Furthermore, assume that at all density levels an increase in population density y never results in an increase in per unit growth rate of y (in the absence of x):

$$g_y(t,0,y)\leq 0 \quad (\not\equiv 0). \tag{32}$$

Finally, assume that the reduced equation (28) has a positive p-periodic solution $y_0(t)\in B_p$ (see Theorems 2–6 and 9). It follows that (31) holds and, hence, $y_0(t)$ is stable, that is species y has a stable periodic "carrying capacity" in the absence of species x.

An example is the periodic version of the famous logistic equation in which $g(t,0,y)=-y/K(t)$, $0<K(t)\in B_p$, which because it is integrable in closed form is easily seen to have a unique, positive globally stable p-periodic solution. Another example is the generalized periodic logistic considered in Example 1 of §2.

Of interest are the questions of the existence and stability of positive solutions of the system (27). From Theorem 10 we find that a global branch C^+ of solutions $(x,y)\in B_p\times B_p$ for which $a(t)=a_0(t)+\mu$, $a_0\in B_p$ and $\operatorname{av}(a_0)=0$ bifurcates from the solution $(0,y_0)\in B_p\times B_p$ at $\mu=\mu_0:=-\operatorname{av}(f(t,0,y_0(t)))$ and that $(0,y_0)$ suffers a loss of stability as $\mu=\operatorname{av}(a)$ increases through μ_0. The branch consists, at least locally near the bifurcation point, of positive solutions which are stable if $\mu_1>0$ (see (20)), in which case the bifurcation is "to the right" (i.e., $\mu>\mu_0$ on the branch) and unstable if $\mu_1<0$, in which case the bifurcation is "to the left". From the discussion in §2, we conclude that the global branch C^+ can leave the positive cone but only at solutions of the form $(x,y,\mu)=(0,y,\mu)$, $0<y\neq y_0$, or $(x,0,\mu)$, $x>0$, and that not all positive solutions from the branch C^+ need be stable. (Note: all of the above holds if (32) is replaced by the weaker assumption (31).)

Predator-prey systems. Let x denote a predator species which preys on species y. The restrictions

$$\begin{gathered} f_x(t,x,y)\leq 0, \quad f_y(t,x,y)\geq 0 \ (\not\equiv 0), \quad g_x(t,x,y)\leq 0 \ (\not\equiv 0) \\ \text{for } x>0, \quad y>0 \quad \text{and all } t\in[0,p] \end{gathered} \tag{33}$$

reflect the situation that an increase in predator density decreases the per unit growth rate of both the predator itself and the prey while an increase in prey density increases the predator's per unit growth rate at each instant of time.

Since the inequality (31) implies that the Green's function $G_2(t,s)$ is strictly positive (see the proof of Theorem 1 in §2), it follows from (19) and (33) that $y_1(t)<0$ for all t. Since $x_1(t)>0$ (see (19)) the formula (20) shows, together with (33), that $\mu_1>0$. (Note that (33) is sufficient, but not necessary for $\mu_1>0$.)

Thus for such predator-prey systems there is an exchange of stability from $(0,y_0(t))$ (*whose stability implies predator extinction*) *to the positive, p-periodic branch solutions* (*whose stability imply the stable coexistence of the prey and predator*) *as* $\mu=\operatorname{av}(a)$ *increases through* μ_0.

If we also assume that

$$\operatorname{av}[f(t,0,y_0(t))]>0,$$

then $\mu_0<0$. Thus, near bifurcation, that is, for $\mu<\mu_0$ close to μ_0, it follows that $\mu=\operatorname{av}(a)<0$ or, in other words, that the predator has a negative average inherent growth rate and would go to extinction in the absence of the prey.

Note that since $y_1(t)<0$ the prey density satisfies $y(t)<y_0(t)$ for all t (see (17)). Thus, near the bifurcation point the prey and the predator coexist, but do so in such a

way that the prey's density is at all times less than what it would be in the absence of the predator.

A classical example would be the periodic version of the famous Volterra–Lotka system in which

$$f(t,x,y)=-c_1(t)x+c_2(t)y, \qquad g(t,x,y)=-c_3(t)x-c_4(t)y,$$
$$0\leq c_i(t)\in B_p, \quad c_2(t)\not\equiv 0, \quad c_3(t)\not\equiv 0$$

to which all of the above applies. For a more detailed analysis of this specific example see [2]. Also see Bardi [1].

Two-species competition. Assume $\mathrm{av}(b)>0$. The restrictions

$$\begin{aligned} &f_x(t,x,y)\leq 0, \quad f_y(t,x,y)\leq 0\ (\not\equiv 0), \quad g_x(t,x,y)\leq 0\ (\not\equiv 0), \\ &g_y(t,x,y)\leq 0 \quad \text{for all } x>0, y>0 \text{ and } t\in[0,p] \end{aligned} \tag{34}$$

describe a case in which increases in either species density results in a decrease in both species' per unit growth rates at all times. If

$$\mathrm{av}[f(t,0,y_0(t))]<0,$$

then $\mu_0>0$, which implies that near bifurcation $\mu=\mathrm{av}(a)>0$ and species x (and hence both species) have unlimited growth in the absence of both inter- and intraspecific competition.

The direction of bifurcation is in this case indeterminate since the sign of μ_1 can be either $+$ or $-$ under assumptions (34). Note that again $y_1(t)<0$. In fact, the quantity μ_1 is precisely that which distinguishes between the case of stable coexistence and of competitive exclusion. It is the natural generalization to the periodic case of the determinant of the community matrix which accomplishes this same task in the autonomous case of the famous Volterra–Lotka competition model (in which f and g are linear in x and y) upon which the idea of the principle of competitive exclusion is theoretically based. This determinant, in fact, identically equals μ_1 when the above analysis and formulas are applied to this special autonomous case. For a more complete discussion of the periodic Volterra–Lotka competition equations see [3]. Also see [7].

Thus under the general assumptions (34) *two competing species can coexist if the bifurcation in Theorem* 10 *is to the "right" but do not coexist and suffer competitive exclusion if the bifurcation is to the "left".*

Appendix. Let $A(t)$ be a continuous, p-periodic $n\times n$ matrix valued function and consider the two systems

$$(*) \qquad x'(t)=A(t)x(t),$$
$$(**) \qquad y'(t)=[A(t)-\lambda I]y(t),$$

where I is the $n\times n$ identity matrix.

PROPOSITION. λ *is a Floquet exponent of* (*) *if and only if* (**) *has a nontrivial p-periodic solution.*

Proof. Let $X(t)$, $Y(t)$ be fundamental matrices of (*) and (**), respectively, which satisfy $X(0)=Y(0)=I$. Define $Z(t):=X(t)\exp(-\lambda t)$. A straightforward calculation shows that $Z(t)$ satisfies (**) and $Z(0)=I$. Thus,

$$(\dagger) \qquad Y(t)=X(t)\exp(-\lambda t).$$

First suppose that λ is a Floquet exponent of (*); that is, $\exp(\lambda p)$ is an eigenvalue of $X(p)$. Let $v\neq 0$ be a vector such that $X(p)v=\exp(\lambda p)v$ and define $y(t):=Y(t)v$.

By definition, $y(t)$ is a nontrivial soluton of (**). Moreover, it is p-periodic because

$$y(p)=Y(p)v=\exp(-\lambda p)X(p)v=\exp(-\lambda p)\exp(\lambda p)v=v=y(0)$$

by (†).

Conversely, suppose that (**) has a nontrivial p-periodic solution $y(t)$. Then $y(t)=Y(t)v$ for some $v\neq 0$. From (†)

$$X(p)v=\exp(\lambda p)Y(p)v=\exp(\lambda p)y(p)=\exp(\lambda p)y(0)=\exp(\lambda p)v,$$

so that λ is a Floquet exponent.

REFERENCES

[1] MARTINO BARDI, *Predator-prey models in periodically fluctuating environments*, J. Math. Biology, 12 (1981), pp. 127–140.

[2] J. M. CUSHING, *Periodic time-dependent predator-prey systems*, SIAM J. Appl. Math., 32 (1977), pp. 82–95.

[3] _____, *Two species competition in a periodic environment*, J. Math. Biology, 10 (1980), pp. 384–400.

[4] _____, *Stable limit cycles of time dependent multispecies interactions*, Math. Biosci., 31 (1976), pp. 259–273.

[5] _____, *Nontrivial periodic solutions of some Volterra integral equations*, Proc. Symposium on Volterra Equations, Lecture Notes in Mathematics, 737, Springer-Verlag, New York, 1979, pp. 50–66.

[6] _____, *Nontrivial periodic solutions of integrodifferential equations*, J. Integral Equations, 1 (1979), pp. 165–181.

[7] P. DE MOTTONI AND A. SCHIAFFINO, *Competition systems with periodic coefficients: a geometric approach*, J. Math. Biology, 11 (1981), pp. 319–335.

[8] P. H. RABINOWITZ, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[9] M. L. ROSENZWEIG, *Why the prey curve has a hump*, Amer. Nat., 103 (1969), pp. 81–87.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0010 $01.00/0

ASYMPTOTIC STABILITY AND THE METHOD OF ITERATED AVERAGES FOR ALMOST PERIODIC SYSTEMS*

STEPHEN C. PERSEK†

Abstract. New stability methods are developed for a class of nonperiodic systems that include a large general class of almost periodic systems. Estimates of global domains of stability and rates of decay on these domains are obtained. Examples of chaotic motion and strange attractors are treated.

1. Introduction. On the interval $0 \leq t < \infty$ consider the initial value problem

$$\frac{dw}{dt} = \varepsilon E(w,x,z,t,\varepsilon), \qquad w|_{t=0} = w_0 + \varepsilon w_1(\varepsilon), \tag{1a, 2a}$$

$$\frac{dx}{dt} = \varepsilon F(w,x,z,t,\varepsilon), \qquad x|_{t=0} = x_0 + \varepsilon x_1(\varepsilon), \tag{1a, 2b}$$

$$\frac{dz}{dt} = Az + \varepsilon H(w,x,z,t,\varepsilon), \qquad z|_{t=0} = z_0 + \varepsilon z_1(\varepsilon), \tag{1c, 2c}$$

where $\varepsilon > 0$ is a small scalar and A in $R^{n\times n}$ is a constant matrix and where w in R^k, x in R^l and z in R^n are column vectors. We assume that each eigenvalue of the matrix A has a negative real part and that the vector functions E and F enjoy some degree of periodicity in t to be stipulated later.

Now assume that the averages of E and F with respect to t exist for $z=0$ and $\varepsilon=0$, and denote these averages by $\bar{E}(w,x)$ and $\bar{F}(w,x)$, respectively. By replacing εE, εF and $Az+\varepsilon H$ in (la, b, c) with $\varepsilon\bar{E}(w,x)$, $\varepsilon\bar{F}(w,x)$ and Az, respectively, Bogoliubov and Mitropolskii [1] obtain an approximating system and show that if the appproximating system has an asymptotically stable rest point, then system (la, b, c) is asymptotically stable in the vicinity of the rest point. And for periodic systems, W. Loud and P. Sethna [14] extend this to a global result. Also in this regard, see A. Lazer [13] on the computation of periodic solutions. Finally, for almost periodic systems, J. Hale [3, pp. 113–169] treats local asymptotic stability of integral manifolds, and D. Gilsinn [2] then obtains the corresponding global result.

However, these averaging methods fail to provide significant results when $\bar{E}(w,x) \equiv 0$. For $\bar{E}(w,x) \equiv 0$, we single out three cases of interest. The first case involves (la, b, c) as a periodic system with a periodic solution. The stability of such a periodic solution can then be explored by the stability methods of S. Persek [7], [8]–although in some situations, it may be enough to apply Lyapunov's direct method (see T. Yoshizawa [11]) or Hopf bifurcation theory (see J. Marsden and M. McCracken [4]). The second case involves (la, b, c) as a periodic system with chaotic nonperiodic solutions. The stability of such nonperiodic solutions can be treated by the approach given in S. Persek [9]–although Lyapunov's methods may be sufficient in special cases. We note that chaotic nonperiodic solutions that are stable are called strange attractors (see C. Marzec and E. Spiegel [5], F. Moon [6] and Y. Ueda [12]).

The final case we mention (to be the subject of this paper) involves (la, b, c) as a system with only limited periodicity; i.e., E and F are nonperiodic finite sums of

* Received by the editors March 10, 1981 and in revised form September 29, 1981.

† 160 Banbury Road, Mineola, New York, 11501; CBA-MGT, St. John's University, Jamaica, New York 11439.

periodic vector functions. To treat the stability of solutions to such systems, we develop iterative methods that apply when $\bar{E}(w,x)\equiv 0$. And examples will be provided illustrating chaotic stable nonperiodic solutions to such systems—solutions similar in behavior to the strange attractors mentioned earlier.

2. Preliminaries. Let D_w, D_x, D_z be convex bounded open sets in R^k, R^l, R^n, respectively; let S_w, S_x, S_z be open sets having their closures contained in D_w, D_x, D_z, respectively; and let $\mathring{S}_w$, $\mathring{S}_x$, $\mathring{S}_z$ (with $z=0$ in $\mathring{S}_z$) be respective subsets of S_w, S_x, S_z. Choosing $\varepsilon_D>0$, define the set D consisting of points of the form (w,x,z,t,ε) by:

$$D=D_w\times D_x\times D_z\times[0,\infty)\times[0,\varepsilon_D].$$

In this paper the pointwise norm of any vector $\xi(t)$ with components $\xi_i(t)$ will be given by $|\xi(t)|=\Sigma_i|\xi_i(t)|$, and the pointwise norm of any matrix $\Lambda(t)$ with scalar entries $\Lambda_{ij}(t)$ will be given by $\|\Lambda(t)\|=\max_j\Sigma_i|\Lambda_{ij}(t)|$. We now require:

Hypothesis H1. (*Quasi-periodicity*). The vectors E and F in (1a), (b) can be represented as a finite sum of vector functions

$$E(w,x,z,t,\varepsilon)=\sum_{i=1}^{\nu}E^{(i)}(w,x,z,t,\varepsilon),$$

$$F(w,x,z,t,\varepsilon)=\sum_{i=1}^{\nu}F^{(i)}(w,x,z,t,\varepsilon),$$

where for $z=0$ and $\varepsilon=0$, each vector function pair $(E^{(i)},F^{(i)})$ is periodic in t with a period P_i independent of w and x. The periods P_i may be rationally independent of one another.

Hypothesis H2. (*Smoothness, boundedness*). H and each $E^{(i)}$ and $F^{(i)}$ are assumed bounded on the set D, with smooth and bounded derivatives on D up to at least third order. Furthermore, the real part of each eigenvalue of the matrix A is assumed to be negative. And finally, the norm of $(w_1(\varepsilon),x_1(\varepsilon),z_1(\varepsilon))$ in (2a, b, c) is uniformly bounded by a fixed constant for $0\le\varepsilon\le\varepsilon_D$.

Now if we let $V(t,s)=e^{A(t-s)}$, then by Hypothesis H2, constants K_A and $\delta>0$ exist such that

$$\|V(t,s)\|\le K_Ae^{-\delta(t-s)}$$

for $0\le s\le t<\infty$. Moreover, with $\gamma\ge 0$ any number, we define the vector averages $\bar{E}$ and $\bar{F}$ by

$$\bar{E}(w,x)=\lim_{t\to\infty}\frac{1}{t-\gamma}\int_\gamma^t E(w,x,z,s,\varepsilon)\big|_{z=0,\varepsilon=0}\,ds,$$

$$\bar{F}(w,x)=\lim_{t\to\infty}\frac{1}{t-\gamma}\int_\gamma^t F(w,x,z,s,\varepsilon)\big|_{z=0,\varepsilon=0}\,ds.$$

And since the point $(w,x,z,t,\varepsilon)\big|_{z=0,\varepsilon=0}$ will be abbreviated from now on as (w,x,t), $\bar{E}$ and $\bar{F}$ can be written in the form

$$\bar{E}(w,x)=\lim_{t\to\infty}\frac{1}{t-\gamma}\int_\gamma^t E(w,x,s)\,ds,$$

$$\bar{F}(w,x)=\lim_{t\to\infty}\frac{1}{t-\gamma}\int_\gamma^t F(w,x,s)\,ds.$$

Hypothesis H3. (*The iterated or nonlinear average* $\bar{E}^*$). Assume that $\bar{E}(w,x)\equiv 0$. With $\partial E/\partial(w,x)$ and $\partial E/\partial z$ Jacobian matrices, we further assume that the limit

$$\bar{E}^*(w,x) = -\lim_{t\to\infty}\frac{1}{t-\gamma}\int_\gamma^t\left\{\int_\gamma^s \frac{\partial E(w,x,\tau)}{\partial(w,x)}\,d\tau\right\}\begin{pmatrix} E(w,x,s)\\ F(w,x,s)\end{pmatrix} ds$$
$$+\lim_{t\to\infty}\frac{1}{t-\gamma}\int_\gamma^t\left(\frac{\partial E}{\partial z}\right)(w,x,s)\left\{\int_\gamma^s V(s,\tau)H(w,x,\tau)\,d\tau\right\} ds$$
$$+\lim_{t\to\infty}\frac{1}{t-\gamma}\int_\gamma^t\left(\frac{\partial E}{\partial \varepsilon}\right)(w,x,s)\,ds$$

exists independent of $\gamma\geq 0$, as a bounded $C^2(w,x)$ function with bounded derivatives on $D_w\times D_x$. We also assume that each separate term used to evaluate $\bar{E}^*$ approaches its particular limit uniformly in (w,x,γ) at the rate $O(1/(t-\gamma))$.

We now introduce the iterated-average system corresponding to (1, a, b, c), (2a, b, c):

(3a, 4a) $$\frac{d\rho}{dt}=\varepsilon^2\bar{E}^*(\rho,\psi), \qquad \rho|_{t=0}=w_0,$$

(3b, 4b) $$\frac{d\psi}{dt}=\varepsilon\bar{F}(\rho,\psi), \qquad \psi|_{t=0}=x_0,$$

(3c, 4c) $$\frac{d\zeta}{dt}=A\zeta, \qquad \zeta|_{t=0}=z_0,$$

and will require that (3a, b, c) has a stable rest point.

Hypothesis H4. (*Existence of a stable rest point*). We assume that (3a, b, c) has a rest point $(\rho^{(0)},\psi^{(0)},0)$ that is an interior point of the set $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$. We further assume that for $0<\varepsilon\leq\varepsilon_D$ and for all initial values in $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$, each solution $(\rho(t,\varepsilon),\psi(t,\varepsilon),\zeta(t,\varepsilon))$ of (3a, b, c), (4a, b, c) remains in $S_w\times S_x\times S_z$ for $0\leq t<\infty$ and approaches $(\rho^{(0)},\psi^{(0)},0)$ as $t\to\infty$.

We now consider the variational system

$$\frac{dU}{dt}=\begin{pmatrix} \varepsilon^2\left(\frac{\partial\bar{E}^*}{\partial w}\right)(\rho(t,\varepsilon),\psi(t,\varepsilon)) & \varepsilon^2\left(\frac{\partial\bar{E}^*}{\partial x}\right)(\rho(t,\varepsilon),\psi(t,\varepsilon))\\ \varepsilon\left(\frac{\partial\bar{F}}{\partial w}\right)(\rho(t,\varepsilon),\psi(t,\varepsilon)) & \varepsilon\left(\frac{\partial\bar{F}}{\partial x}\right)(\rho(t,\varepsilon),\psi(t,\varepsilon))\end{pmatrix}U,$$

with $U|_{t=s}=I_{k+l}$ (the identity matrix in $R^{(k+l)\times(k+l)}$), and write its solution $U(t,s)$ in the block matrix form

$$U(t,s)=\begin{pmatrix} U_{11}(t,s) & U_{12}(t,s)\\ U_{21}(t,s) & U_{22}(t,s)\end{pmatrix}.$$

Hypothesis H5. (*Regularity of* $\partial\bar{E}^*/\partial(w,x)$). We assume that the limit taken in Hypothesis H3 and the use of the operator $\partial/\partial(w,x)$ are interchangeable operations, and that each separate term used to evaluate $\partial\bar{E}^*/\partial(w,x)$ as a limit, approaches its particular value at the rate $O(1/(t-\gamma))$, uniformly for (w,x,γ) in $D_w\times D_x\times[0,\infty)$.

Hypothesis H6. (*Exponential asymptotic stability*). Positive constants K_E, K_F, λ_E, and λ_F are assumed to exist, independent of s and ε and independent of all $\rho(t,\varepsilon)$ and

$\psi(t,\varepsilon)$ described in Hypothesis H4, such that for $0\le s\le t<\infty$, $0<\varepsilon\le\varepsilon_D$:

$$\varepsilon\|U_{11}(t,s)\|+\|U_{12}(t,s)\|+\varepsilon\|U_{21}(t,s)\|\le\varepsilon K_E e^{-\varepsilon^2\lambda_E(t-s)},$$

$$\|U_{22}(t,s)\|\le\varepsilon K_E e^{-\varepsilon^2\lambda_E(t-s)}+K_F e^{-\varepsilon\lambda_F(t-s)}.$$

It should be noted that the validity of the last hypothesis can be checked easily in many cases—for example, if either

$$\frac{\partial\bar{E}^*}{\partial x}\equiv 0 \quad \text{or} \quad \frac{\partial\bar{F}}{\partial w}\equiv 0.$$

In fact, Hypothesis H6 usually holds in the vicinity of stable rest points of (3a, b, c).

3. Previous and new results. The following theorem has been proved in S. Persek and F. Hoppensteadt [10] and will be needed in the proof of our stability result:

THEOREM 1. *Let Hypotheses* H1–H4 *and* H6 *hold for the initial value problem* (1a, b, c), (2a, b, c) *with* $(w(t,\varepsilon),x(t,\varepsilon),z(t,\varepsilon))$ *a solution for* $0<\varepsilon\le\varepsilon_D$, *and let* $(\rho(t,\varepsilon),\psi(t,\varepsilon),\zeta(t,\varepsilon))$ *be a solution to the initial value problem* (3a, b, c), (4a, b, c) *for* $0<\varepsilon\le\varepsilon_D$. *Then constants* $K^*,\varepsilon^*>0$ *exist (with values depending on the sets* D, $S_w\times S_x\times S_z$, $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$ *and on the bounds in* H1–H4, H6) *such that for* $0<\varepsilon\le\varepsilon^*$, *the solutions* $(w(t,\varepsilon),x(t,\varepsilon),z(t,\varepsilon))$ *and* $(\rho(t,\varepsilon),\psi(t,\varepsilon),\zeta(t,\varepsilon))$ *exist on* $0\le t<\infty$, *and*

$$\sup_{0\le t<\infty}\{|w(t,\varepsilon)-\rho(t,\varepsilon)|+|x(t,\varepsilon)-\psi(t,\varepsilon)|+|z(t,\varepsilon)-\zeta(t,\varepsilon)|\}\le K^*\varepsilon$$

independent of the initial values (w_0,x_0,z_0) *in* $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$ *chosen.*

We will follow with the main result of this paper, which basically states that all solutions to system (1a, b, c) are exponentially asymptotically stable if their initial values lie in the region of attraction of an exponentially asymptotically stable rest point of the corresponding iterated-average system.

THEOREM 2. *Let Hypotheses* H1–H6 *hold for the initial value problem* (1a, b, c), (2a, b, c), *with* $(w^*(t,\varepsilon),x^*(t,\varepsilon),z^*(t,\varepsilon))$ *a solution on* $0\le t<\infty$, $0<\varepsilon\le\varepsilon^*$, *where*

$$(w^*,x^*,z^*)\big|_{t=0}=(w_0+\varepsilon w_1(\varepsilon),x_0+\varepsilon x_1(\varepsilon),z_0+\varepsilon z_1(\varepsilon)).$$

Let $(w^\#(t,\varepsilon),x^\#(t,\varepsilon),z^\#(t,\varepsilon))$ *satisfy* (1a, b, c) *on* $0\le t<\infty$, $0<\varepsilon\le\varepsilon^*$ *with*

$$(w^\#,x^\#,z^\#)\big|_{t=0}=(w_0^\#+\varepsilon w_1^\#(\varepsilon),x_0^\#+\varepsilon x_1^\#(\varepsilon),z_0^\#+\varepsilon z_1^\#(\varepsilon)),$$

where $(w_0^\#,x_0^\#,z_0^\#)$, *as well as* (w_0,x_0,z_0), *lie in* $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$. *For fixed* $N_2>0$, *assume that* $|(w_1(\varepsilon),x_1(\varepsilon),z_1(\varepsilon))|\le N_2$ *and* $|(w_1^\#(\varepsilon),x_1^\#(\varepsilon),z_1^\#(\varepsilon))|\le N_2$ *on* $0<\varepsilon\le\varepsilon^*$. *Then constants* $\mathring{K},\mathring{\varepsilon}>0$ *exist depending on* $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$ *and* N_2 *but not on the choice of* $(w^*,x^*,z^*)\big|_{t=0}$ *or* $(w^\#,x^\#,z^\#)\big|_{t=0}$ *such that for all* $0\le t<\infty$, $0<\varepsilon\le\mathring{\varepsilon}$:

$$|w^\#(t,\varepsilon)-w^*(t,\varepsilon)|+|x^\#(t,\varepsilon)-x^*(t,\varepsilon)|+|z^\#(t,\varepsilon)-z^*(t,\varepsilon)|\le\mathring{K}e^{-\varepsilon^2\lambda_E t/2}.$$

4. Applications. We now consider illustrations.

Example 1. With a,b,α and β constants and with w_1, w_2 and w_3 scalars, consider the almost periodic system

$$\frac{dw_1}{dt}=\varepsilon w_3^2(2-w_1^2-w_2^2)\sin t,$$

$$\frac{dw_2}{dt}=\varepsilon w_3^2(w_1^2-w_2^2)\sin\pi t,$$

$$\frac{dw_3}{dt}=\varepsilon\alpha w_1^2\cos t+2\varepsilon\beta w_1w_2\cos\pi t+\varepsilon aw_3\cos\Omega t+\varepsilon bw_3^2(4w_1-w_3)\sin\Omega t,$$

where Ω is constant and $\neq 0, 1$ or π. With (ρ_1, ρ_2, ρ_3) corresponding to (w_1, w_2, w_3), we obtain the iterated-average system

$$\frac{d\rho_1}{dt} = \varepsilon^2 \alpha \rho_1^2 \rho_3 (2 - \rho_1^2 - \rho_2^2),$$

$$\frac{d\rho_2}{dt} = \frac{2}{\pi} \varepsilon^2 \beta \rho_1 \rho_2 \rho_3 (\rho_1^2 - \rho_2^2),$$

$$\frac{d\rho_3}{dt} = \varepsilon^2 \rho_1 \rho_3^2 \left\{ \frac{2}{\Omega} ab - 2\alpha + \left(\alpha - \frac{\beta}{\pi}\right)\rho_1^2 + \left(\alpha + \frac{\beta}{\pi}\right)\rho_2^2 \right\} - \frac{\varepsilon^2}{\Omega} ab \rho_3^3.$$

Provided $\alpha > 0$, $\beta > 0$ and $ab/\Omega > 0$, the (ρ_1, ρ_2, ρ_3) system has stable rest points located at $(1, 1, 2)$ and $(1, -1, 2)$ and $(-1, 1, -2)$ and $(-1, -1, -2)$. With Re the real part of a number, set

$$0 < \lambda_E < \min\left(\frac{4ab}{\Omega}, 2\alpha + \frac{4\beta}{\pi} - \operatorname{Re}\sqrt{4\alpha^2 - \frac{48\alpha\beta}{\pi} + \frac{16\beta^2}{\pi^2}} \right)$$

and consider a bounded domain of stability for each of the rest points. Then by Theorem 2, constants $\mathring{K}$ and $\mathring{\varepsilon} > 0$ exist such that if $w^{\#}(t, \varepsilon)$ and $w^*(t, \varepsilon)$ are any solutions of the (w_1, w_2, w_3) system starting in the same domain of stability at $t = 0$, then

$$|w^{\#}(t, \varepsilon) - w^*(t, \varepsilon)| \leq \mathring{K} e^{-\varepsilon^2 \lambda_E t/2}$$

for $0 \leq t < \infty$ and $0 < \varepsilon \leq \mathring{\varepsilon}$.

We also consider the (ρ_1, ρ_2, ρ_3) system rest points located at $(\sqrt{2}, 0, 2\sqrt{2}[1 - \beta\Omega/(\pi ab)])$ and at $(-\sqrt{2}, 0, -2\sqrt{2}[1 - \beta\Omega/(\pi ab)])$ which are stable provided that $\alpha(1 - \beta\Omega/(\pi ab)) > 0$, $\alpha\beta < 0$ and $ab/\Omega > 0$. Choosing a value for $\tilde{\lambda}_E$ by setting

$$0 < \tilde{\lambda}_E < 8\left(1 - \frac{\beta\Omega}{\pi ab}\right) \min\left(2\alpha, -\frac{2\beta}{\pi}, \frac{ab}{\Omega} - \frac{\beta}{\pi}\right),$$

we consider a bounded domain of stability for each of the two rest points. Then by Theorem 2, constants $\tilde{K}$ and $\tilde{\varepsilon} > 0$ exist such that, for any solutions $w^{\#}(t, \varepsilon)$ and $w^*(t, \varepsilon)$ of the (w_1, w_2, w_3) system both starting in the same domain of stability at $t = 0$, it follows that

$$|w^{\#}(t, \varepsilon) - w^*(t, \varepsilon)| \leq \tilde{K} e^{-\varepsilon^2 \tilde{\lambda}_E t/2}$$

for $0 \leq t < \infty$ and $0 < \varepsilon \leq \tilde{\varepsilon}$.

Example 2. With w_1, w_2, x and z as scalars, consider the system

$$\frac{dw_1}{dt} = \varepsilon(3w_1 + w_2)\cos \pi t + \varepsilon z w_1 \cos t,$$

$$\frac{dw_2}{dt} = \varepsilon(w_1 - w_2)\sin \pi t - \varepsilon z w_2 \sin t,$$

$$\frac{dx}{dt} = \varepsilon(w_1 w_2 - x^2)\cos^2 t,$$

$$\frac{dz}{dt} = -z + \varepsilon(w_1^2 + w_2^2)\sin t.$$

With $(\rho_1,\rho_2,\psi,\zeta)$ corresponding to (w_1,w_2,x,z), we have the iterated average system:

$$\frac{d\rho_1}{dt}=-\frac{\varepsilon^2}{2\pi}(\rho_1-\rho_2)-\frac{\varepsilon^2}{4}\rho_1(\rho_1^2+\rho_2^2),$$
$$\frac{d\rho_2}{dt}=\frac{\varepsilon^2}{2\pi}(3\rho_1+\rho_2)-\frac{\varepsilon^2}{4}\rho_2(\rho_1^2+\rho_2^2),$$
$$\frac{d\psi}{dt}=\frac{\varepsilon}{2}(\rho_1\rho_2-\psi^2),\quad \frac{d\zeta}{dt}=-\zeta.$$

Then the $(\rho_1,\rho_2,\psi,\zeta)$ system has two stable rest points located at $(\sqrt{2/(5\pi)},\sqrt{18/(5\pi)},\sqrt{6/(5\pi)},0)$ and at $(-\sqrt{2/(5\pi)},-\sqrt{18/(5\pi)},\sqrt{6/(5\pi)},0)$. Moreover, a bounded domain of stability exists for each of these rest points such that Hypothesis H6 is valid for all $(\rho_1,\rho_2,\psi,\zeta)$ solutions starting in that domain. Hence, by Theorem 2 constants $\mathring{K}$ and $\mathring{\varepsilon}>0$ exist such that any two solutions $(w_1^{\#},w_2^{\#},x^{\#},z^{\#})$ and $(w_1^{*},w_2^{*},x^{*},z^{*})$ to the (w_1,w_2,x,z) system, which start in the same domain of stability at $t=0$, satisfy

$$|w_1^{\#}(t,\varepsilon)-w_1^{*}(t,\varepsilon)|+|w_2^{\#}(t,\varepsilon)-w_2^{*}(t,\varepsilon)|$$
$$+|x^{\#}(t,\varepsilon)-x^{*}(t,\varepsilon)|+|z^{\#}(t,\varepsilon)-z^{*}(t,\varepsilon)|\le\mathring{K}e^{-\varepsilon^2\lambda_E t/2}$$

for $0\le t<\infty$ and $0<\varepsilon\le\mathring{\varepsilon}$, where λ_E is from Hypothesis H6.

Example 3. Consider nonautonomous systems of the form

$$\frac{dw_1}{dt}=\varepsilon\sum_{\beta}f_\beta(w_1,w_2)\sin\Omega_\beta t+\varepsilon\sum_{\beta}g_\beta(w_1,w_2)\cos\Omega_\beta t,$$
$$\frac{dw_2}{dt}=\varepsilon\sum_{\beta}h_\beta(w_1,w_2)\sin\Omega_\beta t+\varepsilon\sum_{\beta}r_\beta(w_1,w_2)\cos\Omega_\beta t,$$

where the sums consist of finitely many terms and where the constants Ω_β can be rationally independent from one another. Replacing the arguments (w_1,w_2) in f_β, g_β, h_β and r_β with (ρ_1,ρ_2), we obtain the iterated-average system

$$\frac{d\rho_1}{dt}=\frac{\varepsilon^2}{2}\sum_{\beta}\frac{1}{\Omega_\beta}\left\{g_\beta\frac{\partial f_\beta}{\partial\rho_1}-f_\beta\frac{\partial g_\beta}{\partial\rho_1}+r_\beta\frac{\partial f_\beta}{\partial\rho_2}-h_\beta\frac{\partial g_\beta}{\partial\rho_2}\right\},$$
$$\frac{d\rho_2}{dt}=\frac{\varepsilon^2}{2}\sum_{\beta}\frac{1}{\Omega_\beta}\left\{g_\beta\frac{\partial h_\beta}{\partial\rho_1}-f_\beta\frac{\partial r_\beta}{\partial\rho_1}+r_\beta\frac{\partial h_\beta}{\partial\rho_2}-h_\beta\frac{\partial r_\beta}{\partial\rho_2}\right\}.$$

As an illustration consider

$$\frac{dw_1}{dt}=\varepsilon w_2\sin\Omega t+\varepsilon w_2\cos\Omega t+\varepsilon(w_2^2-2w_1w_2)\sin t+\varepsilon(2w_1+w_2)\cos t,$$
$$\frac{dw_2}{dt}=\varepsilon(w_1+2w_2)\sin\Omega t+\varepsilon(4w_1+2w_2)\cos\Omega t$$
$$+\varepsilon(w_1^2-2w_1w_2)\sin t+\varepsilon(w_1+2w_2)\cos t,$$

where Ω is irrational. The corresponding iterated-average system is given by

$$\frac{d\rho_1}{dt}=\frac{3\varepsilon^2}{2\Omega}\rho_1(1-\Omega\rho_1),\qquad \frac{d\rho_2}{dt}=\frac{3\varepsilon^2}{2\Omega}(2\rho_1-\rho_2-\Omega\rho_2^2).$$

For $\Omega>0$, the (ρ_1,ρ_2) system has a stable rest point at $(\frac{1}{\Omega},\frac{1}{\Omega})$. Consider a bounded domain of stability for the rest point and choose a positive constant $\lambda_E<3/(2\Omega)$. Then

by Theorem 2, constants $\mathring{K}$ and $\mathring{\varepsilon}>0$ exist such that if $w^{\#}(t,\varepsilon)$ and $w^{*}(t,\varepsilon)$ are any solutions of the (w_1,w_2) system starting in the domain of stability at $t=0$, then

$$|w^{\#}(t,\varepsilon)-w^{*}(t,\varepsilon)|\leq \mathring{K}e^{-\varepsilon^2\lambda_E t/2}$$

for $0\leq t<\infty$ and $0<\varepsilon\leq\mathring{\varepsilon}$. Note that the constants λ_E, $\mathring{K}$ and $\mathring{\varepsilon}$ may depend on the particular (irrational) value of Ω that has been chosen.

5. Proof of Theorem 2. With $(\rho^{(0)},\psi^{(0)},0)$ the rest point of Hypothesis H4, we will first show that any two solutions to system (1a, b, c) that approach the vicinity of the rest point thereafter exponentially decay into one-another. Theorem 1 will then be applied to show that any two solutions of system (1a, b, c) starting in the region of stability of the iterated-average system (3a, b, c) approach the vicinity of the rest point, and while doing so, their difference is majorized by an exponentially decaying function.

Because Theorem 1 will be used here, we note that it was proved in [10] under a less restrictive version of our present Hypothesis H3, and therefore only for the initial value problem on the interval $0\leq t<\infty$. Using the present version of Hypothesis H3, Theorem 1 and its proof are easily generalized to apply to the initial value problem but now on an arbitrary interval $t_1(\varepsilon)\leq t<\infty$, where $t_1(\varepsilon)\geq 0$. The constants K^{*}, $\varepsilon^{*}>0$ derived in Theorem 1 are possibly altered in the process but to values independent of the choice of $t_1(\varepsilon)$ (as well as the choice of initial values $(w,x,z)\big|_{t=t_1(\varepsilon)}$). We will employ this generalization.

Now let (w^{*},x^{*},z^{*}) and $(w^{*}+W,x^{*}+X,z^{*}+Z)$ be solutions to system (1a, b, c) on $t_1(\varepsilon)\leq t<\infty$ for $t_1(\varepsilon)\geq 0$ arbitrary, i.e., suppose

$$\frac{d}{dt}\begin{pmatrix} w^{*}\\ x^{*}\\ z^{*}\end{pmatrix}=\begin{pmatrix} 0\\ 0\\ Az^{*}\end{pmatrix}+\varepsilon\begin{pmatrix} E(w^{*},x^{*},z^{*},t,\varepsilon)\\ F(w^{*},x^{*},z^{*},t,\varepsilon)\\ H(w^{*},x^{*},z^{*},t,\varepsilon)\end{pmatrix}, \tag{5}$$

$$\frac{d}{dt}\begin{pmatrix} w^{*}+W\\ x^{*}+X\\ z^{*}+Z\end{pmatrix}=\begin{pmatrix} 0\\ 0\\ A(z^{*}+Z)\end{pmatrix}+\varepsilon\begin{pmatrix} E(w^{*}+W,x^{*}+X,z^{*}+Z,t,\varepsilon)\\ F(w^{*}+W,x^{*}+X,z^{*}+Z,t,\varepsilon)\\ H(w^{*}+W,x^{*}+X,z^{*}+Z,t,\varepsilon)\end{pmatrix}, \tag{6}$$

and for arbitrary constant $K_1>0$ assume the initial values at $t=t_1(\varepsilon)$ satisfy the restrictions

$$|w^{*}(t_1(\varepsilon),\varepsilon)-\rho^{(0)}|+|x^{*}(t_1(\varepsilon),\varepsilon)-\psi^{(0)}|+|z^{*}(t_1(\varepsilon),\varepsilon)|\leq K_1\varepsilon, \tag{7}$$

$$|W(t_1(\varepsilon),\varepsilon)|+|X(t_1(\varepsilon),\varepsilon)|+|Z(t_1(\varepsilon),\varepsilon)|<\varepsilon. \tag{8}$$

Then by our earlier discussion on the generalization of Theorem 1, constants K^{*} and $\varepsilon^{*}>0$ exist (depending on K_1 but not on the choice of $t_1(\varepsilon)$ or $(w^{*},x^{*},z^{*})\big|_{t=t_1(\varepsilon)}$ or $(W,X,Z)\big|_{t=t_1(\varepsilon)}$) such that on $0<\varepsilon\leq\varepsilon^{*}$:

$$\sup_{t_1(\varepsilon)\leq t<\infty}\{|w^{*}(t,\varepsilon)-\rho^{(0)}|+|x^{*}(t,\varepsilon)-\psi^{(0)}|+|z^{*}(t,\varepsilon)|\}\leq K^{*}\varepsilon,$$

$$\sup_{t_1(\varepsilon)\leq t<\infty}\{|w^{*}(t,\varepsilon)+W(t,\varepsilon)-\rho^{(0)}|+|x^{*}(t,\varepsilon)+X(t,\varepsilon)-\psi^{(0)}|+|z^{*}(t,\varepsilon)+Z(t,\varepsilon)|\}\leq K^{*}\varepsilon.$$

Expansion of equation (6). From (5) and (6), we obtain

(9)
$$\frac{d}{dt}\begin{pmatrix} W \\ X \\ Z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ AZ \end{pmatrix} + \varepsilon \begin{pmatrix} E(w^*+W, x^*+X, z^*+Z, t, \varepsilon) \\ F(w^*+W, x^*+X, z^*+Z, t, \varepsilon) \\ H(w^*+W, x^*+X, z^*+Z, t, \varepsilon) \end{pmatrix} - \varepsilon \begin{pmatrix} E(w^*, x^*, z^*, t, \varepsilon) \\ F(w^*, x^*, z^*, t, \varepsilon) \\ H(w^*, x^*, z^*, t, \varepsilon) \end{pmatrix},$$

and by Taylor's theorem, with Jacobian matrix notation,

(10)
$$\frac{d}{dt}\begin{pmatrix} W \\ X \\ Z \end{pmatrix} = L(w^*, x^*, z^*, t, \varepsilon)\begin{pmatrix} W \\ X \\ Z \end{pmatrix} + \varepsilon \begin{pmatrix} C_{1E}(W, X, Z, t, \varepsilon) \\ C_{1F}(W, X, Z, t, \varepsilon) \\ C_{1H}(W, X, Z, t, \varepsilon) \end{pmatrix},$$

where a constant N_1 exists depending on K_1, K^* and ε^* (but not on ε, $t_1(\varepsilon)$ or $(w^*, x^*, z^*)|_{t=t_1(\varepsilon)}$) such that for $t_1(\varepsilon) \le t < \infty$, $0 < \varepsilon \le \varepsilon^*$:

$$|C_{1E}(W, X, Z, t, \varepsilon)| + |C_{1F}(W, X, Z, t, \varepsilon)| + |C_{1H}(W, X, Z, t, \varepsilon)| \le N_1\left(|W|^2 + |X|^2 + |Z|^2\right)$$

and where the linear matrix operator $L(w, x, z, t, \varepsilon)$ is given by:

$$L(w, x, z, t, \varepsilon) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & A \end{pmatrix} + \varepsilon \frac{\partial}{\partial(w, x, z)} \begin{pmatrix} E(w, x, z, t, \varepsilon) \\ F(w, x, z, t, \varepsilon) \\ H(w, x, z, t, \varepsilon) \end{pmatrix},$$

with $\partial/\partial(w, x, z)$ the Jacobian operator.

Local asymptotic stability. Let $\Phi(t, s)$ be the fundamental matrix solution to the linear system

$$\frac{d\Phi}{dt} = L(w^*, x^*, z^*, t, \varepsilon)\Phi, \qquad \Phi|_{t=s} = I_{k+l+n}$$

with I_{k+l+n} the identity matrix on $R^{(k+l+n)\times(k+l+n)}$. As shown in §6, constants N_0 and $\varepsilon_0 > 0$ exist ($\varepsilon_0 \le \varepsilon^*$) independent of $t_1(\varepsilon)$ and any $(w^*, x^*, z^*)|_{t=t_1(\varepsilon)}$ satisfying (7) such that for $0 < \varepsilon \le \varepsilon_0$:

$$\|\Phi(t, s)\| \le N_0 e^{-3\varepsilon^2 \lambda_E (t-s)/4}, \qquad t_1(\varepsilon) \le s \le t < \infty.$$

Using the fact that $\Phi(t, t_1(\varepsilon))\Phi^{-1}(s, t_1(\varepsilon)) = \Phi(t, s)$, (10) becomes

(11)
$$\begin{pmatrix} W(t, \varepsilon) \\ X(t, \varepsilon) \\ Z(t, \varepsilon) \end{pmatrix} = \Phi(t, t_1(\varepsilon)) \begin{pmatrix} W(t_1(\varepsilon), \varepsilon) \\ X(t_1(\varepsilon), \varepsilon) \\ Z(t_1(\varepsilon), \varepsilon) \end{pmatrix} + \varepsilon \int_{t_1(\varepsilon)}^{t} \Phi(t, s) \begin{pmatrix} C_{1E}(W(s, \varepsilon), X(s, \varepsilon), Z(s, \varepsilon), s, \varepsilon) \\ C_{1F}(W(s, \varepsilon), X(s, \varepsilon), Z(s, \varepsilon), s, \varepsilon) \\ C_{1H}(W(s, \varepsilon), X(s, \varepsilon), Z(s, \varepsilon), s, \varepsilon) \end{pmatrix} ds.$$

Now for some $\beta > 0$, suppose that

$$|W(t_1(\varepsilon), \varepsilon)| + |X(t_1(\varepsilon), \varepsilon)| + |Z(t_1(\varepsilon), \varepsilon)| < \varepsilon \min\left(1, \beta, \frac{\beta}{N_0}\right),$$

and that $|W(s,\varepsilon)|+|X(s,\varepsilon)|+|Z(s,\varepsilon)|<\beta\varepsilon$ on $t_1(\varepsilon)\le s\le t$, $0<\varepsilon\le\varepsilon_0$. Then from (11) we have

$$\begin{aligned}|W(t,\varepsilon)|+|X(t,\varepsilon)|+|Z(t,\varepsilon)|&\\ \le N_0 e^{-3\varepsilon^2\lambda_E(t-t_1(\varepsilon))/4}&\{|W(t_1(\varepsilon),\varepsilon)|+|X(t_1(\varepsilon),\varepsilon)|+|Z(t_1(\varepsilon),\varepsilon)|\}\\ +\varepsilon\int_{t_1(\varepsilon)}^{t} N_0 e^{-3\varepsilon^2\lambda_E(t-s)/4}&\varepsilon\beta N_1\{|W(s,\varepsilon)|+|X(s,\varepsilon)|+|Z(s,\varepsilon)|\}\,ds\end{aligned}$$

and consequently

$$\begin{aligned}(12)\quad e^{3\varepsilon^2\lambda_E t/4}&\{|W(t,\varepsilon)|+|X(t,\varepsilon)|+|Z(t,\varepsilon)|\}\\ \le N_0 e^{3\varepsilon^2\lambda_E t_1(\varepsilon)/4}&\{|W(t_1(\varepsilon),\varepsilon)|+|X(t_1(\varepsilon),\varepsilon)|+|Z(t_1(\varepsilon),\varepsilon)|\}\\ +\varepsilon^2\beta N_0 N_1\int_{t_1(\varepsilon)}^{t} e^{3\varepsilon^2\lambda_E s/4}&\{|W(s,\varepsilon)|+|X(s,\varepsilon)|+|Z(s,\varepsilon)|\}\,ds.\end{aligned}$$

By a Gronwall inequality applied to (12), we then have for $0<\varepsilon\le\varepsilon_0$:

$$\begin{aligned}|W(t,\varepsilon)|+|X(t,\varepsilon)|+|Z(t,\varepsilon)|&\\ \le N_0\{|W(t_1(\varepsilon),\varepsilon)|+|X(t_1(\varepsilon),\varepsilon)|+|Z(t_1(\varepsilon),\varepsilon)|\}&e^{-\varepsilon^2(3\lambda_E/4-\beta N_0 N_1)(t-t_1(\varepsilon))}\\ <\beta\varepsilon e^{-\varepsilon^2(3\lambda_E/4-\beta N_0 N_1)(t-t_1(\varepsilon))}&.\end{aligned}$$

Choosing $\beta=\lambda_E/(4N_0N_1)$, we have for $0<\varepsilon\le\varepsilon_0$:

$$|W(t,\varepsilon)|+|X(t,\varepsilon)|+|Z(t,\varepsilon)|<\beta\varepsilon e^{-\varepsilon^2\lambda_E(t-t_1(\varepsilon))/2}\le\beta\varepsilon.$$

So for $\beta=\lambda_E/(4N_0N_1)$ and $0<\varepsilon\le\varepsilon_0$, the largest interval on which $|W(s,\varepsilon)|+|X(s,\varepsilon)|+|Z(s,\varepsilon)|<\beta\varepsilon$ is not $t_1(\varepsilon)\le s<t$ but rather $t_1(\varepsilon)\le s<\infty$. We have also shown that for

$$|W(t_1(\varepsilon),\varepsilon)|+|X(t_1(\varepsilon),\varepsilon)|+|Z(t_1(\varepsilon),\varepsilon)|<\varepsilon\min\left(1,\beta,\frac{\beta}{N_0}\right)$$

and $t_1(\varepsilon)\le t<\infty$, $0<\varepsilon\le\varepsilon_0$, $\beta=\lambda_E/(4N_0N_1)$, we have

$$|W(t,\varepsilon)|+|X(t,\varepsilon)|+|Z(t,\varepsilon)|<\beta\varepsilon e^{-\varepsilon^2\lambda_E(t-t_1(\varepsilon))/2}.$$

In summary, for arbitrary K_1 and $t_1(\varepsilon)\ge 0$ let (w^*,x^*,z^*) be any solution of system (1a, b, c) with

$$|(w^*,x^*,z^*)-(\rho^{(0)},\psi^{(0)},0)|_{t=t_1(\varepsilon)}<K_1\varepsilon.$$

Then constants $\alpha(K_1)$, $\beta(K_1)$ and $\varepsilon(K_1)>0$ exist independent of the choice of $t_1(\varepsilon)$ and $(w^*,x^*,z^*)|_{t=t_1(\varepsilon)}$ such that if (w^+,x^+,z^+) is any solution of (1a, b, c) satisfying

$$|(w^+,x^+,z^+)-(w^*,x^*,z^*)|_{t=t_1(\varepsilon)}<\alpha(K_1)\varepsilon$$

then

$$\begin{aligned}(13)\qquad |w^+(t,\varepsilon)-w^*(t,\varepsilon)|+|x^+(t,\varepsilon)-x^*(t,\varepsilon)|&\\ +|z^+(t,\varepsilon)-z^*(t,\varepsilon)|<\varepsilon\beta(K_1)e^{-\varepsilon^2\lambda_E(t-t_1(\varepsilon))/2}&\end{aligned}$$

for $t_1(\varepsilon)\le t<\infty$, $0<\varepsilon\le\varepsilon(K_1)$.

Global asymptotic stability. Let $N_2>0$ be fixed. Let $(w^*(t,\varepsilon),x^*(t,\varepsilon),z^*(t,\varepsilon))$ be a solution of system (1a, b, c), (2a, b, c), for any (w_0,x_0,z_0) in $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$ and for $|w_1(\varepsilon)|+|x_1(\varepsilon)|+|z_1(\varepsilon)|\le N_2$ on $0<\varepsilon\le\varepsilon_D$. Let $(w^\#(t,\varepsilon),x^\#(t,\varepsilon),z^\#(t,\varepsilon))$ be another solution of (1a, b, c) but satisfying

$$(w^\#,x^\#,z^\#)\big|_{t=0}=\big(w_0^\#,x_0^\#,z_0^\#\big)+\varepsilon\big(w_1^\#(\varepsilon),x_1^\#(\varepsilon),z_1^\#(\varepsilon)\big)$$

for any $(w_0^\#,x_0^\#,z_0^\#)$ in $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$ and $|w_1^\#(\varepsilon)|+|x_1^\#(\varepsilon)|+|z_1^\#(\varepsilon)|\le N_2$ on $0<\varepsilon\le\varepsilon_D$. Let $(\rho^*\psi^*,\zeta^*)$ and $(\rho^\#,\psi^\#,\zeta^\#)$ be the iterated-average solutions corresponding to (w^*,x^*,z^*) and $(w^\#,x^\#,z^\#)$, respectively. Then by Theorem 1, constants K_2 and $\varepsilon_2>0$ depending on $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$ and N_2 (but not on the specific choices of $(w^*,x^*,z^*)\big|_{t=0}$ or $(w^\#,x^\#,z^\#)\big|_{t=0}$) such that for $0<\varepsilon\le\varepsilon_2$:

$$\sup_{0\le t<\infty}\big\{|w^*(t,\varepsilon)-\rho^*(t,\varepsilon)|+|x^*(t,\varepsilon)-\psi^*(t,\varepsilon)|+|z^*(t,\varepsilon)-\zeta^*(t,\varepsilon)|\big\}\le K_2\varepsilon, \tag{14}$$

$$\sup_{0\le t<\infty}\big\{|w^\#(t,\varepsilon)-\rho^\#(t,\varepsilon)|+|x^\#(t,\varepsilon)-\psi^\#(t,\varepsilon)|+|z^\#(t,\varepsilon)-\zeta^\#(t,\varepsilon)|\big\}\le K_2\varepsilon. \tag{15}$$

Now by Hypotheses H4 and H6, a constant N_3 exists independent of the choices of (w_0,x_0,z_0) and $(w_0^\#,x_0^\#,z_0^\#)$ in $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$ such that for $0\le t<\infty$, $0<\varepsilon\le\varepsilon_D$:

$$|\rho^*(t,\varepsilon)-\rho^{(0)}|+|\psi^*(t,\varepsilon)-\psi^{(0)}|+|\zeta^*(t,\varepsilon)|\le N_3e^{-\varepsilon^2\lambda_E t/2}, \tag{16}$$

$$|\rho^\#(t,\varepsilon)-\rho^{(0)}|+|\psi^\#(t,\varepsilon)-\psi^{(0)}|+|\zeta^\#(t,\varepsilon)|\le N_3e^{-\varepsilon^2\lambda_E t/2}. \tag{17}$$

Now by (14), (w^*,x^*,z^*) stays close to (ρ^*,ψ^*,ζ^*) on $0\le t<\infty$, and by (16), (ρ^*,ψ^*,ζ^*) decays exponentially to $(\rho^{(0)},\psi^{(0)},0)$ on $0\le t<\infty$. So picking the largest interval $0\le t\le t_2(\varepsilon)$ on which $K_2\varepsilon\le N_3\exp(-\varepsilon^2\lambda_E t/2)$, (14) and (16) show that $|(w^*,x^*,z^*)-(\rho^{(0)},\psi^{(0)},0)|$ is majorized by a decaying exponential on this interval. Similar results hold for $|(w^\#,x^\#,z^\#)-(\rho^{(0)},\psi^{(0)},0)|$ from (15) and (17). In fact, choosing $t_2(\varepsilon)=-(2/(\varepsilon^2\lambda_E))\log(\varepsilon K_2/N_3)$, then from (14)–(17), we have for $0\le t\le t_2(\varepsilon)$, $0<\varepsilon\le\min(\varepsilon_2,N_3/K_2)$:

$$|w^*(t,\varepsilon)-\rho^{(0)}|+|x^*(t,\varepsilon)-\psi^{(0)}|+|z^*(t,\varepsilon)|\le 2N_3e^{-\varepsilon^2\lambda_E t/2}, \tag{18}$$

$$|w^\#(t,\varepsilon)-\rho^{(0)}|+|x^\#(t,\varepsilon)-\psi^{(0)}|+|z^\#(t,\varepsilon)|\le 2N_3e^{-\varepsilon^2\lambda_E t/2}, \tag{19}$$

and therefore also

$$\begin{aligned}|w^\#(t,\varepsilon)-w^*(t,\varepsilon)|+|x^\#(t,\varepsilon)-x^*(t,\varepsilon)|\\ +|z^\#(t,\varepsilon)-z^*(t,\varepsilon)|\le 4N_3e^{-\varepsilon^2\lambda_E t/2}.\end{aligned} \tag{20}$$

Employing inequalities (14)–(17) again and letting $t=t_2(\varepsilon)$, we obtain

$$\big|(w^*,x^*,z^*)-\big(\rho^{(0)},\psi^{(0)},0\big)\big|_{t=t_2(\varepsilon)}\le 2K_2\varepsilon, \tag{21}$$

$$\big|(w^\#,x^\#,z^\#)-\big(\rho^{(0)},\psi^{(0)},0\big)\big|_{t=t_2(\varepsilon)}\le 2K_2\varepsilon, \tag{22}$$

$$\big|(w^\#,x^\#,z^\#)-(w^*,x^*,z^*)\big|_{t=t_2(\varepsilon)}\le 4K_2\varepsilon \tag{23}$$

for $0<\varepsilon\le\min(\varepsilon_2,N_3/K_2)$. Now choosing $K_1\equiv 3K_2$ and $t_1(\varepsilon)\equiv t_2(\varepsilon)$ to be the values of K_1 and $t_2(\varepsilon)$ in inequality (13), we use (21) and (23) and employ our local stability

result (13) repeatedly (but not more times than $1+(4K_2/\alpha(K_1))$). It follows then for all $t_2(\varepsilon)\le t<\infty$, $0<\varepsilon\le\min(\varepsilon_2, N_3/K_2)$ that

$$\begin{aligned}(24)\qquad &|w^{\#}(t,\varepsilon)-w^*(t,\varepsilon)|+|x^{\#}(t,\varepsilon)-x^*(t,\varepsilon)|+|z^{\#}(t,\varepsilon)-z^*(t,\varepsilon)|\\ &\le\varepsilon\beta(K_1)\left\{1+\frac{4K_2}{\alpha(K_1)}\right\}e^{-\varepsilon^2\lambda_E(t-t_2(\varepsilon))/2}\\ &=N_3\beta(K_1)\left\{\frac{1}{K_2}+\frac{4}{\alpha(K_1)}\right\}e^{-\varepsilon^2\lambda_E t/2},\end{aligned}$$

where the constants $K_1\equiv 3K_2$, $\alpha(K_1)$, $\beta(K_1)$ and $\min(\varepsilon_2, N_3/K_2)$ are independent of the choices of (w_0,x_0,z_0) and $(w_0^{\#},x_0^{\#},z_0^{\#})$ in $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$. See the Appendix at the end of §5 for complete details.

Finally, defining

$$\mathring{\varepsilon}=\min\left(\varepsilon_2,\frac{N_3}{K_2}\right)$$

and

$$\mathring{K}=\max\left(4N_3, N_3\beta(K_1)\left\{\frac{1}{K_2}+\frac{4}{\alpha(K_1)}\right\}\right),$$

where $K_1\equiv 3K_2$; then from inequalities (20) and (24), we obtain

$$|w^{\#}(t,\varepsilon)-w^*(t,\varepsilon)|+|x^{\#}(t,\varepsilon)-x^*(t,\varepsilon)|+|z^{\#}(t,\varepsilon)-z^*(t,\varepsilon)|\le\mathring{K}e^{-\varepsilon^2\lambda_E t/2}$$

for all $0\le t<\infty$, $0<\varepsilon\le\mathring{\varepsilon}$, where the constants $\mathring{K}$ and $\mathring{\varepsilon}$ depend on the set $\mathring{S}_w\times\mathring{S}_x\times\mathring{S}_z$ and on the constant N_2 but not on the particular choices of (w_0,x_0,z_0), $(w_0^{\#},x_0^{\#},z_0^{\#})$, $(w_1(\varepsilon),x_1(\varepsilon),z_1(\varepsilon))$ and $(w_1^{\#}(\varepsilon),x_1^{\#}(\varepsilon),z_1^{\#}(\varepsilon))$. Consequently, the proof of Theorem 2 is complete.

Appendix to §5. With $m\ge 1$ any integer and j any integer with $0\le j\le m$, we define the new initial values

$$(w^{(j)},x^{(j)},z^{(j)})\big|_{t=t_2(\varepsilon)}=\left(\frac{j}{m}\right)\times(w^{\#},x^{\#},z^{\#})\big|_{t=t_2(\varepsilon)}+\left(1-\frac{j}{m}\right)\times(w^*,x^*,z^*)\big|_{t=t_2(\varepsilon)}$$

and let $(w^{(j)}(t,\varepsilon),x^{(j)}(t,\varepsilon),z^{(j)}(t,\varepsilon))$ be the solutions to (1a, b, c) corresponding to these initial values. Now by (21) and (22):

$$\left|(w^{(j)},x^{(j)},z^{(j)})-(\rho^{(0)},\psi^{(0)},0)\right|_{t=t_2(\varepsilon)}\le 2K_2\varepsilon,$$

and using (23) we have:

$$\left|(w^{(j+1)},x^{(j+1)},z^{(j+1)})-(w^{(j)},x^{(j)},z^{(j)})\right|_{t=t_2(\varepsilon)}\le 4K_2\varepsilon/m.$$

Let $K_1\equiv 3K_2$ and $t_1(\varepsilon)\equiv t_2(\varepsilon)$ in (13) and in the two inequalities immediately preceeding (13), and with $\alpha(K_1)\equiv\alpha(3K_2)$, we now choose m sufficiently large that $4K_2\varepsilon/m<\alpha(K_1)\varepsilon$. Since the two inequalities preceeding (13) now hold where $(w^{(j+1)},x^{(j+1)},z^{(j+1)})$ and $(w^{(j)},x^{(j)},z^{(j)})$ play the respective roles of (w^+,x^+,z^+) and (w^*,x^*,z^*), then (13) itself holds in the form

$$\begin{aligned}&|w^{(j+1)}(t,\varepsilon)-w^{(j)}(t,\varepsilon)|+|x^{(j+1)}(t,\varepsilon)-x^{(j)}(t,\varepsilon)|\\ &\qquad+|z^{(j+1)}(t,\varepsilon)-z^{(j)}(t,\varepsilon)|\le\varepsilon\beta(K_1)e^{-\varepsilon^2\lambda_E(t-t_2(\varepsilon))/2}\end{aligned}$$

for $t_2(\varepsilon)\le t<\infty$, $0<\varepsilon\le\varepsilon(K_1)$ and $0\le j\le m-1$, where $\beta(K_1)\equiv\beta(3K_2)$ and $\varepsilon(K_1)\equiv\varepsilon(3K_2)$. Hence on $t_2(\varepsilon)\le t<\infty$:

$$|w^{\#}(t,\varepsilon)-w^*(t,\varepsilon)|+|x^{\#}(t,\varepsilon)-x^*(t,\varepsilon)| + |z^{\#}(t,\varepsilon)-z^*(t,\varepsilon)|<\varepsilon m\beta(K_1)e^{-\varepsilon^2\lambda_E(t-t_2(\varepsilon))/2}$$

since $(w^{(m)},x^{(m)},z^{(m)})\equiv(w^{\#},x^{\#},z^{\#})$ and $(w^{(0)},x^{(0)},z^{(0)})\equiv(w^*,x^*,z^*)$. Inequality (24) then follows since m can be chosen to satisfy

$$\frac{4K_2}{\alpha(K_1)}<m\le 1+\frac{4K_2}{\alpha(K_1)}.$$

6. Stability of the linear system $\Phi'=L(w^*,x^*,z^*,t,\varepsilon)\Phi$. With (w^I,x^I,z^I) a vector, consider the general initial value problem

$$\frac{d}{dt}\begin{pmatrix} w^* \\ x^* \\ z^* \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \\ Az^* \end{pmatrix}+\varepsilon\begin{pmatrix} E(w^*,x^*,z^*,t,\varepsilon) \\ F(w^*,x^*,z^*,t,\varepsilon) \\ H(w^*,x^*,z^*,t,\varepsilon) \end{pmatrix}, \tag{5}$$

$$\begin{aligned}\frac{d}{dt}\begin{pmatrix} w^I \\ x^I \\ z^I \end{pmatrix}&=L(w^*,x^*,z^*,t,\varepsilon)\begin{pmatrix} w^I \\ x^I \\ z^I \end{pmatrix}\\ &=\begin{pmatrix} 0 \\ 0 \\ Az^I \end{pmatrix}+\varepsilon\frac{\partial}{\partial(w,x,z)}\begin{pmatrix} E(w,x,z,t,\varepsilon) \\ F(w,x,z,t,\varepsilon) \\ H(w,x,z,t,\varepsilon) \end{pmatrix}\Bigg|_{(w,x,z)=(w^*,x^*,z^*)}\times\begin{pmatrix} w^I \\ x^I \\ z^I \end{pmatrix}\end{aligned} \tag{25}$$

subject to the initial value restrictions

$$|w^*(s(\varepsilon),\varepsilon)-\rho^{(0)}|+|x^*(s(\varepsilon),\varepsilon)-\psi^{(0)}|+|z^*(s(\varepsilon),\varepsilon)|\le K^*\varepsilon \tag{26}$$

and $|w_0^I|+|x_0^I|+|z_0^I|<2$, where $(w^I,x^I,z^I)\big|_{t=s(\varepsilon)}=(w_0^I,x_0^I,z_0^I)$ and $s(\varepsilon)\ge 0$ is arbitrary. Then (5) and (25) are amenable to averaging, and employing Hypothesis H3 and (3a, b, c), we have

$$\frac{d}{dt}\begin{pmatrix} \rho^* \\ \psi^* \\ \zeta^* \end{pmatrix}=\begin{pmatrix} \varepsilon^2\bar{E}^*(\rho^*,\psi^*) \\ \varepsilon\bar{F}(\rho^*,\psi^*) \\ A\zeta^* \end{pmatrix}, \tag{27}$$

$$\begin{pmatrix} \rho^* \\ \psi^* \\ \zeta^* \end{pmatrix}_{t=s(\varepsilon)}=\begin{pmatrix} \rho^{(0)} \\ \psi^{(0)} \\ 0 \end{pmatrix},$$

$$\frac{d}{dt}\begin{pmatrix}\rho^I\\ \psi^I\\ \zeta^I\end{pmatrix}=\begin{pmatrix}\varepsilon^2\left(\frac{\partial \bar{E}^*}{\partial w}\right)(\rho^*,\psi^*)\rho^I+\varepsilon^2\left(\frac{\partial \bar{E}^*}{\partial x}\right)(\rho^*,\psi^*)\psi^I\\ \varepsilon\left(\frac{\partial \bar{F}}{\partial w}\right)(\rho^*,\psi^*)\rho^I+\varepsilon\left(\frac{\partial \bar{F}}{\partial x}\right)(\rho^*,\psi^*)\psi^I\\ A\zeta^I\end{pmatrix}, \tag{28}$$

$$\begin{pmatrix}\rho^I\\ \psi^I\\ \zeta^I\end{pmatrix}_{t=s(\varepsilon)}=\begin{pmatrix}w_0^I\\ x_0^I\\ z_0^I\end{pmatrix}.$$

Since by Hypothesis H4, $(\rho^{(0)},\psi^{(0)},0)$ is a rest point of (27), then on $s(\varepsilon)\le t<\infty$, $(\rho^*(t,\varepsilon),\psi^*(t,\psi),\zeta^*(t,\varepsilon))\equiv(\rho^{(0)},\psi^{(0)},0)$. And by Hypothesis H2 and the variational equation following Hypothesis H4, the solution to (28) on $s(\varepsilon)\le t<\infty$ is

$$\begin{pmatrix}\rho^I\\ \psi^I\end{pmatrix}=U(t,s(\varepsilon))\begin{pmatrix}w_0^I\\ x_0^I\end{pmatrix},$$

and

$$\zeta^I=V(t,s(\varepsilon))z_0^I.$$

Now by the generalization of Theorem 1 found in the beginning of §5, constants a_3 and $\varepsilon_3>0$ exist independent of the choice of $s(\varepsilon)$ such that

$$\begin{aligned}\sup_{s(\varepsilon)\le t<\infty}\ &\{|w^*(t,\varepsilon)-\rho^*(t,\varepsilon)|+|x^*(t,\varepsilon)-\psi^*(t,\varepsilon)|\\ &+|z^*(t,\varepsilon)-\zeta^*(t,\varepsilon)|+|w^I(t,\varepsilon)-\rho^I(t,\varepsilon)|\\ &+|x^I(t,\varepsilon)-\psi^I(t,\varepsilon)|+|z^I(t,\varepsilon)-\zeta^I(t,\varepsilon)|\}\le a_3\varepsilon\end{aligned} \tag{29}$$

for $0<\varepsilon\le\varepsilon_3$ independent of the initial value $(w^*,x^*,z^*,w^I,x^I,z^I)|_{t=s(\varepsilon)}$ provided that (26) holds and $|w_0^I|+|x_0^I|+|z_0^I|<2$.

Next, referring again to the beginning of §5, for arbitrary $t_1(\varepsilon)\ge 0$ and the constant $K_1>0$ with

$$|w^*(t_1(\varepsilon),\varepsilon)-\rho^{(0)}|+|x^*(t_1(\varepsilon),\varepsilon)-\psi^{(0)}|+|z^*(t_1(\varepsilon),\varepsilon)|\le K_1\varepsilon,$$

there exist constants K^*, $\varepsilon^*>0$ depending on K_1 but not on the choice of $t_1(\varepsilon)$ such that

$$\sup_{t_1(\varepsilon)\le t<\infty}\{|w^*(t,\varepsilon)-\rho^{(0)}|+|x^*(t,\varepsilon)-\psi^{(0)}|+|z^*(t,\varepsilon)|\}\le K^*\varepsilon$$

for $0<\varepsilon\le\varepsilon^*$. And for $s(\varepsilon)\ge t_1(\varepsilon)$, choosing $t=s(\varepsilon)$ in this result shows that (26) holds, thereby implying (29). So let $\Phi(t,s(\varepsilon))$ and $\Omega(t,s(\varepsilon))$ be the fundamental matrix solutions to (25) and (28) respectively, on $t_1(\varepsilon)\le s(\varepsilon)\le t<\infty$. Then by Hypothesis H6, a constant N_{12} exists such that

$$\|\Omega(t,s(\varepsilon))\|\le N_{12}e^{-\varepsilon^2\lambda_E(t-s(\varepsilon))}$$

for $t_1(\varepsilon)\le s(\varepsilon)\le t<\infty$, $0<\varepsilon\le\min(\varepsilon_D,\delta/\lambda_F,\lambda_F/\lambda_E)$. Letting $M=(4/\lambda_E)\log 2N_{12}$ and $\varepsilon_0=\min(\varepsilon^*,\varepsilon_D,\delta/\lambda_F,\lambda_F/\lambda_E,1/(16a_3N_{12}^3),\varepsilon_3)$, then by (29):

$$\left\|\Phi\left(s(\varepsilon)+\frac{M}{\varepsilon^2},s(\varepsilon)\right)\right\|\le\left\|\Omega\left(s(\varepsilon)+\frac{M}{\varepsilon^2},s(\varepsilon)\right)\right\|+a_3\varepsilon$$
$$\le N_{12}e^{-\lambda_E M}+a_3\varepsilon\le e^{-3\lambda_E M/4}$$

for all $s(\varepsilon)\ge t_1(\varepsilon)$ and $0<\varepsilon\le\varepsilon_0$. Hence, with $N_0=N_{12}+\frac{1}{2}$:

$$\|\Phi(t,\tau)\|\le N_0e^{-3\varepsilon^2\lambda_E(t-\tau)/4}$$

for $t_1(\varepsilon)\le\tau\le t<\infty$, $0<\varepsilon\le\varepsilon_0$, where N_0 depends on K_1 but not on $t_1(\varepsilon)$ or the choice of $(w^*,x^*,z^*)\big|_{t=t_1(\varepsilon)}$ satisfying (7).

REFERENCES

[1] N. Bogoliubov and Yu. Mitropolskii, *Asymptotic Methods in the Theory of Non-Linear Oscillations*, Gordon and Breach, New York, 1961.

[2] D. Gilsinn, *The method of averaging and domains of stability for integral manifolds*, SIAM J. Appl. Math., 29 (1975), pp. 628–660.

[3] J. Hale, *Oscillations in Nonlinear Systems*, McGraw-Hill, New York, 1963.

[4] J. Marsden and M. McCracken, *The Hopf Bifurcation and Its Applications*, Applied Mathematical Sciences, vol. 19, Springer-Verlag, New York, 1976.

[5] C. Marzec and E. Spiegel, *Ordinary differential equations with strange attractors*, SIAM J. Appl. Math., 38 (1980), pp. 403–421.

[6] F. Moon, *Experimental models for strange attractor vibrations in elastic systems*, New Approaches to Nonlinear Problems in Dynamics, P. Holmes, ed., Society for Industrial and Applied Mathematics, Philadelphia, 1980, pp. 487–495.

[7] S. Persek, *Asymptotic stability on slow time scales for periodic systems*, SIAM J. Appl. Math., 41 (1981), pp. 16–28.

[8] _____, *Conditional stability for periodic solutions of periodic systems containing a small parameter*, J. Differential Equations, 41 (1981), pp. 163–185.

[9] _____, *Stability methods for non-periodic solutions of periodic systems*, in preparation.

[10] S. Persek and F. Hoppensteadt, *Iterated averaging methods for systems of ordinary differential equations with a small parameter*, Comm. Pure Appl. Math., 31 (1978), pp. 133–156.

[11] T. Yoshizawa, *Stability Theory and the Existence of Periodic Solutions and Almost Periodic Solutions*, Applied Mathematical Sciences, vol. 14, Springer-Verlag, New York, 1975.

[12] Y. Ueda, *Steady motions exhibited by Duffing's equation: a picture book of regular and chaotic motions*, New Approaches to Nonlinear Problems in Dynamics, P. Holmes, ed., Society for Industrial and Applied Mathematics, Philadelphia, 1980, pp. 311–322.

[13] A. Lazer, *On the computation of periodic solutions of weakly nonlinear differential equations*, SIAM J. Appl. Math., 15 (1967), pp. 1158–1170.

[14] W. Loud and P. Sethna, *Some explicit estimates for domains of attraction*, J. Differential Equations, 2 (1966), pp. 158–172.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0011 $01.00/0

MULTIPARAMETER VARIATIONAL PRINCIPLES*

PAUL BINDING†

Abstract. Let T and V be self-adjoint operators on a separable Hilbert space H, where V is bounded and T has compact resolvent. Then a spectral theory, including eigenvector completeness, may be given variationally for the eigenvalue problem

$$Tx=\lambda Vx, \qquad 0\neq x\in H$$

provided either $T\gg 0$ (i.e., $(x,Tx)\geq\alpha\|x\|^2$ for some $\alpha>0$) or $V\gg 0$. These conditions are known as left and right definiteness (LD and RD) respectively.

Generalisations of LD and RD appropriate for the multiparameter problem

$$T_m x_m=\sum_{n=1}^{k}\lambda_n V_{mn}x_m, \qquad 0\neq x_m\in H_m, \quad m=1,\cdots,k,$$

have already been established in the literature. New variational principles are given here in terms of $\mathbb{R}^k$-valued functions on $\times_{m=1}^k H_m$ and $\otimes_{m=1}^k H_m$. In particular, when LD is assumed, a variational spectral theory is given, including eigenvector completeness. Such a theory is shown to be impossible under RD, although a finite dimensional version is established for a condition including both LD and RD.

1. The problem. Throughout, we shall consider the abstract multiparameter eigenvalue problem

$$\begin{aligned} &W_m(\boldsymbol{\lambda})x_m=0\neq x_m\in H_m,\\ &W_m(\boldsymbol{\lambda})=T_m-\sum_{n=1}^{k}\lambda_n V_{mn}, \qquad m=1,\cdots,k, \end{aligned} \tag{1.1}$$

where T_m and V_{mn} are self-adjoint operators on separable Hilbert spaces H_m. We shall also assume that the V_{mn} are bounded and that the T_m are bounded below with compact resolvents. Thus, for some real α, $T_m-\alpha I_m$ has a positive compact inverse, I_m being the identity on H_m.

The above assumptions have been used frequently in discussions of (1.1)—many ordinary differential and difference equation problems are covered. Applications occur in separation of variables for partial differential equations (p.d.e.) giving rise to systems of ordinary differential equations (o.d.e.) defining various "special functions" [1], [18], multipoint conditions for o.d.e. [2], [15], and quadratic eigenvalue problems [20], [21]. Difference equation approximations are treated by Atkinson in [3] and more general matrix problems in [4].

Two elementary examples where separation of variables leads to problems involving the definiteness conditions used here are given by vibrating annular and elliptic membranes. For the former, polar coordinates lead to a system of o.d.e. satisfying LD and RD, while for the latter, elliptic coordinates lead to a system satisfying LD but not RD. The variational principles of §§6 and 7 apply respectively, Bessel and Mathieu functions being the relevant special functions.

Consider the one-parameter right definite (RD) case where $k=1$ and $V_{11}\gg 0$, i.e.,

$$(x,V_{11}x)\geq\nu\|x\|^2 \quad \text{for all } x\in H_1 \tag{1.2}$$

---

*Received by the editors February 4, 1981, and in revised form June 4, 1981. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

† Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4.

for some $\nu>0$. Then we may define a new inner product by

$$[x,y]_0=(x,V_{11}y) \tag{1.3}$$

and write H_0 for the linear space H_1 with the $[\cdot,\cdot]_0$ inner product. It is easily shown that H_0 is homeomorphic to H_1 and that

$$\Gamma_1=V_{11}^{-1}T_1 \tag{1.4}$$

is self-adjoint on H_0 with compact resolvent. Thus the eigentuples for Γ_1 and hence for (1.1) may be generated variationally from the Rayleigh quotient

$$\gamma_1(x)=\frac{[x,\Gamma_1 x]_0}{[x,x]_0}, \qquad 0\neq x\in\mathcal{D}(\Gamma_1). \tag{1.5}$$

Since this coincides with the "generalised Rayleigh quotient" for (1.1), i.e.,

$$\gamma_1(x)=\frac{(x,T_1x)}{(x,V_{11}x)}, \qquad 0\neq x\in\mathcal{D}(T_1), \tag{1.6}$$

we see that the eigentuples of (1.1) may be determined by constrained extrema of a simple real-valued functional on H_1, viz. γ_1, defined in terms of the original operators T_1 and V_{11}. Moreover, a complete orthonormal basis of eigenvectors for H_0 is characterized directly.

Turning to the one-parameter left definite (LD) case, we assume $\alpha>0$ so $T_1\gg 0$ and we define the (incomplete) inner product

$$[x,y]_1=(x,T_1y) \tag{1.7}$$

on $\mathcal{D}(T_1)^2$. Let $\mathcal{D}_1$ denote the completion of $\mathcal{D}(T_1)$ with respect to $[\cdot,\cdot]_1$—except in the finite dimensional case, we no longer have a homeomorph of H_1. The analogue of (1.4) is

$$B_1=T_1^{-1}V_{11}, \tag{1.8}$$

which is now compact and self-adjoint on $\mathcal{D}_1$. Again we can generate eigentuples from constrained extrema of the Rayleigh quotient

$$\beta_1(x)=\frac{[x,B_1x]_1}{[x,x]_1}, \qquad 0\neq x\in\mathcal{D}_1, \tag{1.9}$$

and a complete orthonormal basis of eigenvectors for (1.1) is constructed, in this case for the orthocomplement of $\operatorname{Ker}V_{11}$ in $\mathcal{D}_1$. Again β_1 has a simple representation in terms of the original operators, viz.,

$$\beta_1(x)=\frac{(x,V_{11}x)}{(x,T_1x)}, \qquad 0\neq x\in\mathcal{D}(T_1). \tag{1.10}$$

The aim of this article is to establish multiparameter analogues of the above procedures for various definiteness conditions, in terms of constrained extrema of functionals of Rayleigh quotient type defined on $\times_{m=1}^{k}H_m$ and $\otimes_{m=1}^{k}H_m$. Very briefly, our main conclusions are that both approaches have analogues if both LD and RD hold, that the second has an analogue if LD holds, that there is no analogue for RD in the infinite dimensional case and that both approaches have finite dimensional analogues under a condition weaker than LD and RD. For a more detailed discussion we require some notation and definitions.

2. Basic notation. A typical element of $H=\times_{m=1}^{k} H_m$ is denoted by $x=(x_1,\cdots,x_k)$ with $x_m\in H_m$. We write

$$U_m=\{u_m\in\mathfrak{D}(T_m)\subseteq H_m:\|u_m\|=1\},$$

and we reserve the notation u_m for a general element of U_m. Then $U=\times_{m=1}^{k} U_m$ consists of all $u=(u_1,\cdots,u_k)$ for which u_m is a unit element of $\mathfrak{D}(T_m)$.

The operators T_m, V_{mn} and $W_m(\boldsymbol{\lambda})$ induce quadratic forms on U as follows:

$$t_m(u)=(u_m,T_m u_m),\qquad \mathbf{t}(u)=(t_1(u),\cdots,t_k(u)) \tag{2.1}$$

(throughout, the use of boldface denotes a vector, considered as a column if matrix notation is also employed);

$$v_{mn}(u)=(u_m,V_{mn}u_m),\qquad V(u)=[v_{mn}(u)] \tag{2.2}$$

(thus $V(u)$ is a $k\times k$ matrix);

$$w_m(\boldsymbol{\lambda},u)=(u_m,W_m(\boldsymbol{\lambda})u_m),\qquad \mathbf{w}(\boldsymbol{\lambda},u)=(w_1(\boldsymbol{\lambda},u),\cdots,w_k(\boldsymbol{\lambda},u)). \tag{2.3}$$

We then construct certain determinants as follows, via a fixed vector $\boldsymbol{\omega}\in\mathbb{R}^K$ and a real ω_0. Let

$$\delta(u)=\begin{vmatrix} \omega_0 & \boldsymbol{\omega}^T \\ -\mathbf{t}(u) & V(u)\end{vmatrix}, \tag{2.4}$$

and let $\delta_n(u)$ be the cofactor of ω_n in this $(k+1)\times(k+1)$ determinant, for $n=0,\cdots,k$. We also set

$$\boldsymbol{\delta}(u)=(\delta_1(u),\cdots,\delta_k(u)). \tag{2.5}$$

We note that

$$\delta_0(u)=\det V(u), \tag{2.6}$$

and we write $\delta_{0mn}(u)$ for the (m,n) cofactor in this $k\times k$ determinant. Observe that

$$\delta_n(u)=\sum_{m=1}^{k} t_m(u)\delta_{0mn}(u),\qquad n=1,\cdots,k. \tag{2.7}$$

Finally, we construct a "comparison" cone C as follows. Let $\mathbb{R}_+^k$ denote the positive orthant, with the origin, i.e.,

$$\mathbb{R}_+^k=\{\boldsymbol{\lambda}\in\mathbb{R}^k:\lambda_m>0 \text{ for } m=1,\cdots,k\}\cup\{\mathbf{0}\}, \tag{2.8}$$

and let $\overline{\mathbb{R}}^k$ be the nonnegative orthant. Let $\overline{U}_m$ be the closure of U_m, i.e., the unit sphere of H_m, and let $\overline{U}=\times_{m=1}^{k}\overline{U}_m$. It is obvious that δ_0 may be defined on $\overline{U}$ by (2.6) and similarly for the cofactors δ_{0mn}. We define

$$C=\{\boldsymbol{\lambda}\in\mathbb{R}^k:V(u)\boldsymbol{\lambda}\in\overline{\mathbb{R}}_+^k \text{ for some } u\in\overline{U}\}. \tag{2.9}$$

3. Definiteness conditions. We start with three conditions associated with left definiteness:

Condition LD_+. For some $\varepsilon>0$, $\delta_{0mn}(u)\geq\varepsilon$ for $1\leq m,\ n\leq k$ and for all $u\in U$.

Condition LD_δ. For some $\varepsilon>0$ and $\boldsymbol{\omega}\in\mathbb{R}^k$, $\Sigma_{n=1}^{k}\omega_n\delta_{0mn}(u)\geq\varepsilon$ for $m=1,\cdots,k$ and for all $u\in U$.

Condition LD_t. For some $\alpha>0$, $t_m(u)\geq\alpha$ for $m=1,\cdots,k$ and for all $u\in U$.

We shall define LD as the combination of LD_+ with LD_t. This is formally stronger than the conditions in the literature, which in particular assume LD_δ rather

than LD_+—see [17] for example. Nevertheless we lose no real generality since our conditions can still be arranged via a nonsingular affine eigenvalue transformation [8, Cor. 2.8]. Note that if LD holds then (2.7) gives

$$\delta_n(u) \geq k\alpha\varepsilon \quad \text{for } n=1,\cdots,k \text{ and } u\in U,$$

so we may construct the ratios

$$\beta_n(u)=\frac{\delta_0(u)}{\delta_n(u)}, \qquad \boldsymbol{\beta}(u)=(\beta_1(u),\cdots,\beta_k(u)) \tag{3.1}$$

by analogy with (1.10).

We turn now to right definiteness (RD), which we shall define as the combination of LD_t with the following:

Condition RD_δ. For some $\zeta>0$, $|\delta_0(u)|>\zeta$ for all $u\in U$.

The right definiteness conditions in the literature do not explicitly assume LD_t, but the latter may be arranged via an eigenvalue translation, assuming RD_δ [8, Lemma 2.1].

Under RD_δ, one can define the function $\boldsymbol{\gamma}$ on U by

$$\boldsymbol{\gamma}(u)=V(u)^{-1}\mathbf{t}(u), \qquad u\in U \tag{3.2}$$

(see (2.1), (2.2)). Two identities follow directly. First,

$$\boldsymbol{\omega}(\boldsymbol{\gamma}(u),u)=\mathbf{0} \tag{3.3}$$

comes from (2.3), while (2.5), (2.6) and an easy calculation yield

$$\boldsymbol{\gamma}(u)=\frac{\boldsymbol{\delta}(u)}{\delta_0(u)}. \tag{3.4}$$

We shall use two further definiteness conditions. One, implied by either LD or RD, is used by Atkinson [4, Chapt. 6] for the finite dimensional case:

Condition A. For some $\eta>0$, $|\delta(u)|\geq\eta$ for all $u\in U$.

This will be discussed in §9. The other condition is "properness", viz.

Condition P. LD and RD both hold.

This was introduced in [12] as the combination of LD_δ and RD_δ, but as we have noted above, LD_+ and LD_t are automatic modulo an affine transformation. According to [12] a cone is *proper* if it is closed, convex and contains no line. The connection between this and Condition P is as follows.

THEOREM 3.1. (i) *If C is proper, then LD_δ and RD_δ both hold.*

(ii) *If* P *holds, then C is proper and is contained in $\mathbb{R}^k_+$ (see* (2.8)).

Proof. All the contentions except the last follow from [12, Thm. 6.1]. For the last we assume P and consider $\boldsymbol{\lambda}\in C$, so from (2.7) there is $\boldsymbol{\mu}\in\overline{\mathbb{R}}^k_+$ such that

$$\lambda_m=\sum_{n=1}^{k}\frac{\delta_{0nm}(u)\mu_n}{\delta_0(u)}, \qquad m=1,\cdots,k.$$

It follows that $\lambda_m\geq 0$, so $\boldsymbol{\lambda}\in\overline{\mathbb{R}}^k_+$. Moreover, if $\lambda_m=0$, then $\boldsymbol{\mu}=\mathbf{0}$ so $\boldsymbol{\lambda}=\mathbf{0}$ and so we have $\boldsymbol{\lambda}\in\mathbb{R}^k_+$. Q.E.D.

4. Discussion. We shall now summarise our work and compare it with the relevant literature. Variational principles are standard for $k=1$ but comparatively underdeveloped for $k>1$—in fact, there seems to be no such treatment giving eigenvector completeness in infinite dimensions. We distinguish here between multiparameter variational principles, which characterize the eigentuples of (1.1) directly, and the use of one-parameter variational arguments as stepping stones to the analysis of (1.1) by other

techniques. There are several instances of the latter, e.g., in [10], [17], [20], [21] and [22], but the only article specifically dealing with the former seems to be [9], which treats RD_δ in finite dimensions.

We observe that γ, defined in (3.2), satisfies

$$w_1(\gamma(u),u)=0 \tag{4.1}$$

(see (2.3), (3.3)). When $k=1$, $\gamma(u)$ is simply the generalised Rayleigh quotient appropriate for RD_δ (see (1.6)). Equations of the same type as (4.1) have been used for variational principles in several contexts. For example, Duffin [14] has analysed a quadratic matrix eigenvalue problem via an analogue of (4.1). We remark that this can be recast as a right definite two-parameter problem [20] and that there are many extensions of [14] to more general nonlinear eigenvalue problems. Hadeler [16], apparently motivated by such ideas, defined a "generalised range" $\mathfrak{R}$ for one linear equation in several parameters. In fact, $\mathfrak{R}$ turns out to be the set of all $\gamma(u)$ satisfying (4.1) as $u=u_1$ ranges over U_1. Hadeler gave some variational connections between $\mathfrak{R}$ and a "spectrum" for his equation but did not pursue this topic. Binding [6] has generalised such definitions to nonlinear equations in several parameters and has given minimax-like relations for the eigentuples in terms of $\gamma(U)$. In the linear case, RD_δ is assumed and it is shown, for example, that there exists $\boldsymbol{\lambda}=\boldsymbol{\lambda}^0$ satisfying (1.1) and such that

$$\gamma(U)\subseteq\boldsymbol{\lambda}^0+C \tag{4.2}$$

(here C comes from (2.9)). Two disadvantages of this analysis are (i) there is no way to prove eigenvector completeness and (ii) C generates a partial order on $\mathbb{R}^k$ if and only if C is convex, which is true (assuming RD_δ) if and only if LD_δ holds, by Theorem 3.1. In particular, if LD_δ fails, then it is impossible to describe the minimax-like relations of [6] by collections of linear functionals on $\gamma(U)$.

Another approach to the right definite problem has been via the tensor product $H^\otimes=\otimes_{m=1}^k H_m$. The definition of the requisite linear operators Δ_n on $H^\otimes$ is given in §5, but for the present discussion it is enough to note that

$$\left(u^\otimes,\Delta_n u^\otimes\right)=\delta_n(u),\qquad u\in U,\quad n=0,\cdots,k. \tag{4.3}$$

Here $u^\otimes=u_1\otimes\cdots\otimes u_k$ is a "decomposable" element of $H^\otimes$. In finite dimensions, Binding and Browne [9] have shown that the eigentuples of (1.1) may be characterized by successive minimaximisations of k linear functionals over the set

$$\left\{\left((h,\Delta_1 h),\cdots,(h,\Delta_k h)\right)/(h,\Delta_0 h):0\neq h\in H^\otimes\right\}. \tag{4.4}$$

It turns out that the corresponding constrained maximisers may be taken decomposable, and thus (4.3) shows that (4.4) is simply the set

$$\{\boldsymbol{\delta}(u)/\delta_0(u):u\in U\},$$

which, by virtue of (3.4), is just $\gamma(U)$ again. Although this approach gives a direct construction of a complete orthonormal basis of eigenvectors for $H^\otimes$—with a new inner product involving Δ_0—there are still drawbacks. In particular, (iii) the k linear functionals may vary with the eigentuple being characterized, and (iv) the analysis of [9] is finite dimensional in nature. We note that Binding and Browne [10, §5] have generalised some of Hadeler's inclusions [16] by giving infinite dimensional descriptions of various images of γ, but there is no eigentuple characterization, and (i)–(iii) are not tackled.

In §6 we shall give an infinite dimensional version of the tensor product approach under Condition P. In fact, the k linear functionals may be taken as the k coordinates and are thus eigentuple independent. Further, Theorem 3.1 shows that (4.2) implies

$$\gamma_n(u) \leq \lambda_n^0, \qquad u \in U, \quad n = 1, \cdots, k.$$

Similarly, the other minimax-like relations of [6] can all be generated by the k coordinates, and we have overcome all the difficulties (i)–(iv). In §7 we extend the analysis to cover LD, using β (3.1). In this case new inner products are used involving the Δ_n and completion is involved. Also, the eigenvectors are complete not in $H^{\otimes}$ but in the orthocomplement of $\operatorname{Ker}\Delta_0$—cf. the approach via (1.8) in §1.

In §8 we show that our analysis cannot be extended to RD as well in the infinite dimensional case. We give an explanation for this in terms of the geometry of C. Roughly, if C is convex, then a nonsingular eigenvalue transformation exists under which $C \subseteq \mathbb{R}_+^k$ (Theorem 3.1), hence ensuring that the operators Δ_n have enough compactness properties for a variational characterization. Lack of convexity of C, already pinpointed in (ii) as one source of difficulty, is then seen in a different light. Indeed, we include an example satisfying RD but not LD_δ (so C is not convex) and for which the relevant operators on $H^{\otimes}$ have spectra which cannot be characterized by minimax principles. In the final section, we discuss Condition A. In view of the results of §8 and the fact that RD implies A, we restrict ourselves to finite dimensions. Atkinson has given a spectral theory for this case in [4, Chapt. 6], and we show that such a theory may be given variationally in at least two ways. In particular, we extend the results of [9] directly.

5. Operators in the tensor product. In order to make our discussion self-contained, we include in this section the basic construction of the determinantal operators we need. Fuller details may be found, for example, in [23, Chapt. 2]. We shall start with the set D of decomposable tensors of the form $u^{\otimes} = u_1 \otimes \cdots \otimes u_m$. In particular, the notation requires $u_m \in U_m$, so $u \leftrightarrow u^{\otimes}$ provides a 1–1 correspondence between U and D. $H^{\otimes}$ is then defined as the completion of the linear span S of D under the inner product defined on D^2 by

$$(u^{\otimes}, v^{\otimes}) = \prod_{m=1}^{k} (u_m, v_m)$$

and extended bilinearly to S^2.

Next, for any $u^{\otimes} \in D$, we define

$$T_m^{\dagger} u^{\otimes} = u_1 \otimes \cdots \otimes u_{m-1} \otimes T_m u_m \otimes u_{m+1} \otimes \cdots \otimes u_k,$$

and we extend $T_m^{\dagger}$ to a self-adjoint operator on $H^{\otimes}$. $V_{mn}^{\dagger}$ is defined similarly on D by

$$V_{mn}^{\dagger} u^{\otimes} = u_1 \otimes \cdots \otimes u_{m-1} \otimes V_{mn} u_m \otimes u_{m+1} \otimes \cdots \otimes u_k$$

and then extended to a bounded self-adjoint operator on $H^{\otimes}$. Then

$$\Delta = \det \begin{bmatrix} \omega_0 & \omega_1 & \cdots & \omega_k \\ -T_1^{\dagger} & V_{11}^{\dagger} & \cdots & V_{1k}^{\dagger} \\ \vdots & \vdots & & \vdots \\ -T_k^{\dagger} & V_{k1}^{\dagger} & \cdots & V_{kk}^{\dagger} \end{bmatrix} \tag{5.1}$$

is (uniquely) defined on

$$\mathfrak{D}=\bigcap_{m=1}^{k}\mathfrak{D}(T_m^{\dagger}), \tag{5.2}$$

which can be shown to be dense in $H^{\otimes}$.

The construction of Δ_n parallels that of δ_n. Explicitly, Δ_n is the closure of the cofactor (operator) of ω_n in the expansion of Δ for $n=0,\cdots,k$, and Δ_{0mn} is the (m,n) cofactor in the expansion of $\Delta_0=\det[V_{mn}^{\dagger}]$. Note that Δ_0 and Δ_{0mn} are bounded and

$$\delta(u)=(u^{\otimes},\Delta u^{\otimes}),\quad \delta_{0mn}(u)=(u^{\otimes},\Delta_{0mn}u^{\otimes}),\qquad 1\le m,n\le k, \tag{5.3}$$

hold in addition to (4.3). Also, the following will enable us to carry our definiteness assumptions over to $H^{\otimes}$. Recall the notation $\gg 0$ of, say, (1.2).

THEOREM 5.1. *If* LD *holds, then* $T_m^{\dagger}$, Δ_{0mn} *and* Δ_n *are all self-adjoint and* $\gg 0$ *on* $H^{\otimes}$, $1\le m,n\le k$. Δ_0 *is always self-adjoint, and if* RD *holds, then* $\Delta_0\gg 0$, *on* $H^{\otimes}$.

Proof. The positivity statements for Δ_{0mn} and Δ_0 follow from [5]. By [7, Thm. 3.1], $\Delta_n\gg 0$ on $\mathfrak{D}$ (5.2) and so $\Delta_n|_{\mathfrak{D}}$ has a self-adjoint extension $\hat{\Delta}_n\gg 0$ (i.e., $\gg 0$ on $\mathfrak{D}(\hat{\Delta}_n)$). By [7, Remark after Theorem 3.1], $\hat{\Delta}_n$ coincides with Δ_n as defined here. Q.E.D

6. The proper case. Throughout this section, we shall assume Condition P. According to Theorem 5.1, we may define an inner product $[\cdot,\cdot]_0$ on $\mathcal{H}^{\otimes}$ by

$$[x,y]_0=(x,\Delta_0 y). \tag{6.1}$$

The linear space $H^{\otimes}$ with this inner product will be denoted by H_0. Obviously, H_0 is homeomorphic to $H^{\otimes}$ since Δ_0^{-1} is bounded by Theorem 5.1. Moreover, we may define operators

$$\Gamma_n=\Delta_0^{-1}\Delta_n,\qquad \mathfrak{D}(\Gamma_n)=\mathfrak{D}(\Delta_n),\qquad n=1,\cdots,k, \tag{6.2}$$

on H_0. We also denote the unit sphere of H_0 by U_0—explicitly

$$U_0=\{x\in H_0:[x,x]_0=1\}. \tag{6.3}$$

THEOREM 6.1. *The* Γ_n *are commuting and self-adjoint on* H_0. *Further,* $\Gamma_n\gg 0$ *and* Γ_n^{-1} *is compact.*

Proof. By Theorem 5.1, $\Gamma_n^{-1}=\Delta_n^{-1}\Delta_0$ is bounded on $H^{\otimes}$ and hence on H_0.

Further,

$$[x,\Gamma_n^{-1}y]_0=(x,\Delta_0\Delta_n^{-1}\Delta_0 y)=(\Delta_n^{-1}\Delta_0 x,\Delta_0 y)=[\Gamma_n^{-1}x,y]_0$$

proves self-adjointness of Γ_n^{-1}. Commutativity of Γ_n^{-1} follows from [7, Thm. 4.2]. The corresponding results for Γ_n are then formalities.

Next,

$$[x,\Gamma_n x]_0=(x,\Delta_n x)\quad \text{for } x\in\mathfrak{D}(\Gamma_n) \tag{6.4}$$

and $\Delta_n\gg 0$ on $H^{\otimes}$ show that $\Gamma_n\gg 0$ on H_0. Finally, compactness of Γ_n^{-1} in $H^{\otimes}$ (and hence in H_0 by homeomorphism) comes from [7, Cor. 3.4]. Q.E.D.

It follows that, in the notation (6.1)–(6.3),

$$\lambda_1^0=\min_{x\in U_0}[x,\Gamma_1 x]_0>0$$

exists as the minimal eigenvalue of Γ_1 with finite dimensional eigenspace E_1^0, say. By commutativity of the Γ_n,

$$\Gamma_2 E_1^0\subseteq E_1^0$$

so

$$\lambda_2^0=\min\{[x,\Gamma_2 x]_0: x\in U_0\cap E_1^0\}>0$$

exists as an eigenvalue of Γ_2. The set of minimisers generates a joint eigenspace E_2^0 for Γ_1 and Γ_2. Repeating this procedure for each Γ_n in turn, we obtain

$$\lambda_n^0=\min\{[x,\Gamma_n x]_0: x\in U_0\cap E_{n-1}^0\}>0, \qquad n=3,\cdots,k$$

with corresponding eigenspace E_n^0.

In particular, we have $\boldsymbol{\lambda}^0\in\mathbb{R}_+^k$ (2.8), and the set of minimisers for λ_k^0 spans a joint eigenspace E_k^0 for $\Gamma_1,\cdots,\Gamma_k$. By [7, Thm. 6.1], (1.1) is equivalent to a simultaneous eigenvalue problem for the Γ_n^{-1} (and hence for the Γ_n) in the sense that if

$$\Gamma_n h=\lambda_n h, \qquad n=1,\cdots,k, \tag{6.5}$$

for some nonzero $h\in H_0$ then one may choose $h=x_1\otimes\cdots\otimes x_k$ with $\boldsymbol{\lambda}$ and x satisfying (1.1). By suitable scaling then, we may find $u_m^0\in U_m$ such that $\boldsymbol{\lambda}$ and u^0 form an eigentuple for (1.1) and such that (6.5) is satisfied with $h=u^{0\otimes}=u_1^0\otimes\cdots\otimes u_k^0$. Also, it is evident that

$$\lambda_n^0=\frac{[u^{0\otimes},\Gamma_n u^{0\otimes}]_0}{[u^{0\otimes},u^{0\otimes}]_0}=\frac{\delta_n(u^0)}{\delta_0(u^0)}=\gamma_n(u^0). \tag{6.6}$$

Having extracted $\boldsymbol{\lambda}^0$ and u^0, we now repeat the process on the orthocomplement of $u^{0\otimes}$ in H_0. This leads to the recursive construction of an eigenpair $\boldsymbol{\lambda}^1$, u^1, and so on. Alternatively, we may use a maximin construction, finding $\boldsymbol{\lambda}^1(x)$ and $u^1(x)$ by minimising over the orthocomplement of nonzero $x\in H_0$, and then maximising $\lambda_1^1(x)$ over such x. Either way we construct a complete orthonormal basis of H_0 of eigenvectors $u^{0\otimes}$, $u^{1\otimes},\cdots$ for Γ_1 (cf. §1) and we have arrived at the following:

THEOREM 6.2. *A countable set of eigenvalues* $\boldsymbol{\lambda}^0$, $\boldsymbol{\lambda}^1,\cdots$ *with corresponding eigenvectors* u^0, $u^1,\cdots$ *exists for* (1.1) *and may be determined by successive coordinatewise recursive minimisation (or maximin) operations on* $\boldsymbol{\gamma}(U)$. *The corresponding tensors* $u^{0\otimes}$, $u^{1\otimes},\cdots$ *form a complete orthonormal basis of* H_0.

Note that (6.6) permits the use of $\boldsymbol{\gamma}(U)$ even though the proof involves the more complicated construction

$$\{([x,\Gamma_1 x]_0,\cdots,[x,\Gamma_k x]_0): x\in U_0\}.$$

We should point out that even when all the eigenvalues are simple, the order of their extraction via Theorem 6.2 in general depends on the ordering of the coordinates. In §8, we shall connect this order with a frequently-used method of indexing the eigenvectors (by oscillation count in the o.d.e. case). At present, we merely remark that the first eigenpair extracted is somewhat special.

COROLLARY 6.3. *The eigenvalue* $\boldsymbol{\lambda}^0$ *satisfies* $W_m(\boldsymbol{\lambda}^0)\geq 0$, $m=1,\cdots,k$, *and coincides with* $\boldsymbol{\lambda}^0$ *of* (4.2). *The ordering of the coordinate minimisations over* $\boldsymbol{\gamma}(U)$ *used to generate* $\boldsymbol{\lambda}^0$ *is immaterial.*

Proof. For the existence of $\boldsymbol{\lambda}^*$ satisfying $W_m(\boldsymbol{\lambda}^*)\geq 0$, $m=1,\cdots,k$, see, e.g., [10, Thm. 2]. Thus, for any $u\in U$, using the coordinatewise partial order on $\mathbb{R}^k$, we have

$$\mathbf{w}((\boldsymbol{\lambda}^*),u)\geq\mathbf{0}=\mathbf{w}(\boldsymbol{\gamma}(u),u),$$

so

$$V(u)(\boldsymbol{\lambda}^*-\boldsymbol{\gamma}(u))\geq\mathbf{0},$$

whence

$$\boldsymbol{\lambda}^* \in \boldsymbol{\gamma}(u) + C \subseteq \boldsymbol{\gamma}(u) + \mathbb{R}^k_+$$

by virtue of Theorem 3.1. The first inclusion gives (4.2), and the second shows that $\boldsymbol{\lambda}^*$ is a *simultaneous* minimum (over $u \in U$) of the k coordinates $\gamma_n(u)$, $n=1,\cdots,k$. Thus, the construction of $\boldsymbol{\lambda}^0$ yields $\boldsymbol{\lambda}^*$, regardless of the ordering of the minimisations. Q.E.D.

COROLLARY 6.4 [12, Cor. 7.4]. $\lambda_n^0 = \min\{(x, \Delta_n x) : x \in U_0 \cap \mathcal{D}_a\}$.

Here $\mathcal{D}_a$ is the algebraic tensor product corresponding to $\mathcal{D}$ (5.2), i.e., the linear space $\mathcal{D}(T_1^\dagger) \otimes \cdots \otimes \mathcal{D}(T_k^\dagger)$.

7. The left definite case. For this section we assume LD. From Theorem 5.1 we have $\Delta_n \gg 0$ for $n=1,\cdots,k$. Although Δ_0 is bounded, it may not be invertible, so instead of Γ_n, we shall use $\Delta_n^{-1}\Delta_0$. Also, it will be necessary to replace H_0, and we start by defining $\mathcal{D}_n = \mathcal{D}(\Delta_n^{1/2})$ with inner product

$$[x,y]_n = \left(\Delta_n^{1/2}x, \Delta_n^{1/2}y\right), \qquad n=1,\cdots,k.$$

It is easily verified that $\mathcal{D}_n$ is a Hilbert space and is in fact the completion of $\mathcal{D}$ (5.2) with respect to an inner product of the form $(x, \Delta_n y)$. We note that $\mathcal{D}_n \subseteq H^\otimes$ and $\Delta_n^{-1}\Delta_0 H^\otimes \subseteq \mathcal{D}(\Delta_n) \subseteq \mathcal{D}_n$, so $\mathcal{D}_n$ is invariant under $\Delta_n^{-1}\Delta_0$. Accordingly, we define

$$B_n = \Delta_n^{-1}\Delta_0\big|_{\mathcal{D}_n}.$$

THEOREM 7.1. (i) *The $\Delta_n^{-1}\Delta_0$ commute on $H^\otimes$.*

(ii) *B_n is compact and self-adjoint on $\mathcal{D}_n$ for $n=1,\cdots,k$.*

Proof. (i) follows from [7, Thm. 4.2].

(ii) Suppose $x_j \to 0$ in $\mathcal{D}_n$ as $j \to \infty$, and let $y \in H^\otimes$. Then

$$(y, x_j) = \left[\Delta_n^{-1}y, x_j\right]_n \to 0 \quad \text{as } j \to \infty,$$

since $\Delta_n^{-1}y \in \mathcal{D}_n$. Thus $x_j \to 0$ in $H^\otimes$, so

$$\|B_n x_j\|_{\mathcal{D}_n} = \|\Delta_n^{-1/2}\Delta_0 x_j\|_{H^\otimes} \to 0$$

since $\Delta_n^{-1/2}$ is compact on $H^\otimes$ by [7, Cor. 3.4]. Thus $B_n x_j \to 0$ in $\mathcal{D}_n$, and B_n is therefore compact. Finally,

$$[x, B_n y]_n = (x, \Delta_0 y) = (\Delta_0 x, y) = [B_n x, y]_n$$

proves self-adjointness of B_n. Q.E.D.

Our replacement for H_0 is the orthocomplement of $\operatorname{Ker}\Delta_0$ in one of the $\mathcal{D}_n$. Taking $n=1$ for simplicity, let us then define $\mathcal{D}_0$ as a subspace of $\mathcal{D}_1$ by

$$\mathcal{D}_0 = \{x \in \mathcal{D}_1 : [x,y]_1 = 0 \,\forall y \in \mathcal{D}_1 \cap \operatorname{Ker}\Delta_0\}.$$

LEMMA 7.2. *If x is an eigenvector of* (1.1), *then $x^\otimes \in \Delta_1^{-1}\Delta_0 H^\otimes \subseteq \mathcal{D}_0$.*

Proof. If $\boldsymbol{\lambda}$ and x satisfy (1.1), then

$$T_m^\dagger x^\otimes = \sum_{n=1}^{k} \lambda_n V_{mn}^\dagger x^\otimes, \qquad m=1,\cdots,k. \tag{7.1}$$

Multiplying the mth equation by Δ_{0m1} and summing, we obtain

$$\Delta_1 x^\otimes = \lambda_1 \Delta_0 x^\otimes. \tag{7.2}$$

Thus

$$x^\otimes = \lambda_1 \Delta_1^{-1}\Delta_0 x^\otimes \in \Delta_1^{-1}\Delta_0 H^\otimes.$$

Now let $h=\Delta_1^{-1}\Delta_0 z$ and $y\in\mathfrak{D}_1\cap\operatorname{Ker}\Delta_0$, so

$$[h,y]_1=(\Delta_0 h,y)=(h,\Delta_0 y)=0, \tag{7.3}$$

whence $h\in\mathfrak{D}_0$. Q.E.D.

To avoid trivialities we shall assume henceforth that (1.1) has solutions so $\mathfrak{D}_0$ contains nonzero elements. Let $U_n^{\otimes}$ be the set of elements of unit norm in $\mathfrak{D}_n$ for $n=0,\cdots,k$. Note that $B_1:\mathfrak{D}_0\to\mathfrak{D}_0$ follows from Lemma 7.2, so we can start our variational construction of the eigentuples for (1.1) by finding

$$\mu_1^0=\max_{x\in U_0^{\otimes}}|[x,B_1x]_1|>0.$$

Let F_1^0 be the corresponding eigenspace, and let $\nu_1^0(=\pm\mu_1^0)$ be the corresponding eigenvalue, for B_1. Note that $F_1^0\subset\mathfrak{D}_0$ by Lemma 7.2.

LEMMA 7.3. *$F_1^0\cap\mathfrak{D}_2$ contains nonzero elements.*

Proof. Let $0\neq f\in F_1^0$. Since $f\in\mathfrak{D}_0$, $\Delta_0 f$ is nonzero so

$$0\neq\Delta_2^{-1}\Delta_0 f\in\mathfrak{D}(\Delta_2)\subseteq\mathfrak{D}_2.$$

On the other hand, Theorem 7.1(i) gives $\Delta_2^{-1}\Delta_0 f\in F_1^0$. Q.E.D.

We continue the construction by finding

$$\mu_2^0=\max\{|[x,B_2x]_2|:x\in F_1^0\cap U_2^{\otimes}\}>0$$

with F_2^0 and ν_2^0 $(\neq\pm\mu_2^0)$ as the corresponding eigenspace and eigenvalue for B_2. We now repeat the process for each B_n in turn, with

$$\mu_n^0=\max\{|[x,B_nx]_n|:x\in F_{n-1}^0\cap U_n^{\otimes}\},$$

F_n^0 and ν_n^0 being the corresponding eigenspace and eigenvalue. The existence of nonzero elements in $F_{n-1}^0\cap U_n^{\otimes}$ is proved essentially as for Lemma 7.3.

We thus reach F_k^0, a joint eigenspace for all the B_n. As in the previous section, we choose a decomposable element $u^{0\otimes}$ of F_k^0. With $\lambda_n^0=1/\nu_n^0$, $n=1,\cdots,k$, we then see that $\boldsymbol{\lambda}^0$ and u^0 form an eigentuple for (1.1). We now repeat the construction on the orthocomplement of u^0 in $\mathfrak{D}_0$ to yield $\boldsymbol{\lambda}^1$ and u^1, and so on. As in §6, we are led to the following:

THEOREM 7.4. *A countable set of eigentuples $\boldsymbol{\lambda}^j$, u^j of* (1.1) *may be obtained by successive recursive or minimax operations on the components of $\boldsymbol{\beta}(U_0)$. The corresponding $u^{j\otimes}$ form a complete orthonormal basis of $\mathfrak{D}_0$.*

The only point here that needs further comment is the use of $\boldsymbol{\beta}(U_0)$. The result is clear if we use $\beta^{\otimes}(U_0^{\otimes})$ instead, where

$$\beta_n^{\otimes}(x)=\frac{[x,B_nx]_n}{[x,x]_n},\qquad n=1,\cdots,k.$$

Now

$$\beta_n^{\otimes}(u^{0\otimes})=\frac{\delta_0(u^0)}{\delta_n(u^0)}=\beta_n(u^0),$$

so $\boldsymbol{\beta}^{\otimes}(D)=\boldsymbol{\beta}(U)$. In particular,

$$\mu_1^0=\max_{u\in U}|\beta_1(u)|,\qquad \nu_1^0=\beta_1(u^0)$$

and similarly for the other (constrained) maximisations in the construction.

8. Comparison of the definiteness conditions. In this section we shall relate the spectra of the operators $\Delta_0^{-1}\Delta_n$ (considered on various domains) to the eigenvalues of (1.1). This will clarify the order of extraction of eigenvalues in §6, and will also explain why variational arguments are possible when LD is assumed, but not under RD alone.

It is straightforward to deduce the analogue of (7.2) for Δ_l, regardless of definiteness conditions. Specifically, if (1.1) is satisfied then (7.1) holds, and if we multiply by Δ_{0ml} and sum on m then we reach

$$\Delta_l x^{\otimes} = \lambda_l \Delta_0 x^{\otimes}, \qquad l = 1, \cdots, k. \tag{8.1}$$

In particular, λ_l is an eigenvalue of $\Delta_0^{-1}\Delta_l$ under RD, and λ_l^{-1} is an eigenvalue of $\Delta_l^{-1}\Delta_0$ under LD for $l = 1 \cdots k$.

Condition P. In this case $\Delta_0^{-1}\Delta_l$ may be considered as a self-adjoint operator Γ_l on H_0, as in §6. Thus the spectrum of Γ_l includes the lth coordinate projections of the eigenvalues $\boldsymbol{\lambda}^j$ of (1.1), repeated according to multiplicity. In particular,

$$\boldsymbol{\lambda}^j \in \boldsymbol{\lambda}^0 + C \subseteq \boldsymbol{\lambda}^0 + \mathbb{R}_+^k,$$

where $\boldsymbol{\lambda}^0 = \boldsymbol{\lambda}^{\mathbf{0}}$ (4.2) and $\mathbb{R}_+^k$ comes from (2.8). *If* we accept that the spectrum of Γ_l consists *entirely* of the λ_l^j, then it follows from simple geometry and the finite multiplicity of eigenvalues of (1.1) that $\Gamma_l \gg 0$ and has compact inverse.

A simple example may be of use at this point.

Example 8.1. Let $k = 2$, $H_1 = H_2 = l_2$ with orthonormal basis $e_1, e_2, \cdots$, and let I_1 be the identity on H_1. Let

$$Te_j = je_j, \qquad j = 1, 2, \cdots$$

and suppose

$$T_1 = T_2 = T \quad \text{and} \quad V_{mn} = \delta_{mn} I_1, \qquad 1 \le m, n \le 2. \tag{8.2}$$

Then $V(u) = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, so $\delta_0(u) = 1$ and RD holds. Indeed, $H^{\otimes}$ is the l_2 space of double index sequences and Δ_0 is the corresponding identity I. On the other hand, $\underline{\mathrm{LD}}_+$ fails since $\delta_{012} = \delta_{021} = 0$. The cone C (2.9) in this case is the nonnegative quadrant $\overline{\mathbb{R}}_+^2$.

If we transform the eigenvalues so that λ_1 and λ_2 are replaced by $2\lambda_1 - \lambda_2$ and $-\lambda_1 + 2\lambda_2$ respectively, then of course RD still holds and further $V(u) = \left[\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right]$. Thus LD_+ (and hence P) also holds and

$$C = \{(\lambda_1, \lambda_2) : 2\lambda_1 - \lambda_2 \ge 0,\ -\lambda_1 + 2\lambda_2 \ge 0\}.$$

The eigenvalues $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\mathbf{j}}$ of (1.1) now occur at the intersections of the lines $\lambda_1 - 2\lambda_2 = j_1$ and $-\lambda_1 + 2\lambda_2 = j_2$, where j_1 and j_2 are nonnegative integers. This (multi)indexing can always be carried out under RD (see [10, Thm. 2]). It corresponds to oscillation count of the x_m (1.1) in the o.d.e. case.

It is easily verified that

$$3\Gamma_1 = 2T_1^{\dagger} + T_2^{\dagger}, \qquad 3\Gamma_2 = T_1^{\dagger} + 2T_2^{\dagger},$$

and the eigenvalues are extracted by Theorem 6.3 as follows:

$$\boldsymbol{\lambda}^0 = \boldsymbol{\lambda}^{\mathbf{0}}, \quad \boldsymbol{\lambda}^1 = \boldsymbol{\lambda}^{(0,1)}, \quad \boldsymbol{\lambda}^2 = \boldsymbol{\lambda}^{(1,0)}, \quad \boldsymbol{\lambda}^3 = \boldsymbol{\lambda}^{(0,2)}, \quad \cdots.$$

If the order of the maximisations is reversed, then we obtain

$$\boldsymbol{\lambda}^0 = \boldsymbol{\lambda}^{\mathbf{0}}, \quad \boldsymbol{\lambda}^1 = \boldsymbol{\lambda}^{(0,1)}, \quad \boldsymbol{\lambda}^2 = \boldsymbol{\lambda}^{(0,1)}, \quad \boldsymbol{\lambda}^3 = \boldsymbol{\lambda}^{(2,0)}, \quad \cdots.$$

According to Theorem 3.1, the existence of such an eigenvalue transformation characterizes P. In general, the eigenvalues are extracted by Theorem 6.3 in lexicographic order of coordinates.

Condition LD. As noted in §7, $B_1 : \mathcal{D}_0 \to \mathcal{D}_0$ is self-adjoint. Similarly, $\Delta_n^{-1}\Delta_0$ is a self-adjoint operator on the orthocomplement of $\operatorname{Ker}\Delta_0$ in $\mathcal{D}_n$. Let Γ_n denote the inverse of this operator. Since $\Gamma_n = \Delta_0^{-1}\Delta_n$, we have almost recovered the situation under P, at least if we accept that the spectrum of Γ_n consists entirely of the λ_n^j. The difference is that we can no longer use the cone C—we shall show elsewhere that C may be replaced by two cones contained in $\mathbb{R}_+^k$ and $-\mathbb{R}_+^k$ respectively. For the present, however, we merely note that the LD case essentially behaves as two cases of condition P, one where $\Delta_0 > 0$ and the other where $\Delta_0 < 0$. Specifically, $\mathcal{D}_0$ is the orthogonal direct sum of subspaces invariant under Δ_0 and such that Δ_0 takes opposite signs on each.

Condition RD. Let us first return to Example 8.1. With no eigenvalue transformation, we obtain $\Gamma_1 = T_1^\dagger$, $\Gamma_2 = T_2^\dagger$, and these operators possess eigenvalues of infinite multiplicity. This is to be expected from the projection process applied to the eigenvalues of (1.1)—they form the positive integer lattice in $\mathbb{R}^2$ for this example.

Theorem 3.1 implies that if LD_δ fails then it is no longer possible to place C in $\mathbb{R}_+^k$ by a nonsingular eigenvalue transformation. This suggests that the Γ_n (defined as the closures of the $\Delta_0^{-1}\Delta_n$ in H_0) may no longer have compact resolvents. We conclude with an example illustrating this—in fact any linear functional of the eigenvalues for (1.1) will have a finite accumulation point. Thus the Γ_n will have eigenvalues with finite accumulation regardless of whatever linear transformation we apply. This precludes a variational characterization, at least along the lines here.

Example 8.2. Let $k=3$ and $H_m = l_2$, $m=1,2,3$, with I_1 and T as in Example 8.1. Let $V_{mm} = I_1$, $T_m = T$, $m=1,2,3$ and

$$V_{13} = V_{21} = V_{32} = A, \qquad V_{12} = V_{23} = V_{31} = 0,$$

where

$$Ae_1 = e_1, \qquad Ae_j = 0, \quad j = 2,3,\cdots.$$

It is easily seen that RD holds. Also

$$W_1(\boldsymbol{\lambda})e_1 = (\lambda_3+\lambda_1)e_1, \quad W_2(\boldsymbol{\lambda})e_1 = (\lambda_1+\lambda_2)e_1, \quad W_3(\boldsymbol{\lambda}) = (\lambda_2+\lambda_3)e_1,$$
$$W_m(\boldsymbol{\lambda})e_j = (\lambda_m+j)e_j, \qquad m=1,2,3, \quad j=2,3,\cdots.$$

Thus all integer triples of the form

$$-(i,j,l), \quad (i,-j,-i), \quad (-i,i,-j), \quad (-j,-i,i)$$

are eigenvalues of (1.1) provided $i,j,l > 1$. We now claim that any fixed linear combination $\Sigma_{m=1}^3 \rho_m \lambda_m$ of the eigenvalues of (1.1) possesses a finite accumulation point.

Suppose two of ρ_1, ρ_2 and ρ_3 take opposite signs, say $\rho_1 < 0 < \rho_2$. Then let $q_j 10^{-j}$ be the jth decimal approximation to $-\rho_1/\rho_2$, q_j being a positive integer. It is clear that

$$|\rho_1 10^j + \alpha_2 q_j + 2\alpha_3| \le 1 + 2|\alpha_3|,$$

so $\Sigma_{m=1}^3 \rho_m \lambda_m^{(j)}$ is bounded as $j \to \infty$ if we choose $\boldsymbol{\lambda}^{(j)} = (-10^j, -q_j, -2)$.

Now suppose that all the ρ_m have the same sign, say positive. If $\rho_1 < \rho_2$, then we approximate $(\rho_2 - \rho_1)/\rho_3$ by $q_j 10^{-j}$, and we obtain a bounded sequence $\Sigma_{m=1}^3 \rho_m \lambda_m^{(j)}$ provided we choose $\boldsymbol{\lambda}^{(j)} = (-10^j, 10^j, -q_j)$. The case $\rho_2 < \rho_3$ and $\rho_3 < \rho_1$ are handled similarly.

The case where one variable is zero and the other two have the same sign is handled as follows. If, say, $\rho_1 = 0 < \rho_2 \le \rho_3$, then we use $\boldsymbol{\lambda}^{(j)} = (-10^j, 10^j, -q_j)$, where

$q_j 10^{-j}$ approximates ρ_2/ρ_3. If two variables are equal, say $\rho_1=\rho_2$, then $\boldsymbol{\lambda}^{(j)}=(-j,j,-2)$ will do. Modulo cyclic permutations, these cover all possibilities, and our claim is met.

9. Atkinson's condition. Condition A was used by Atkinson [4, Chapt. 6] for a nonvariational analysis of (1.1) in finite dimensions. This is one of the most general situations which admits a spectral theory including a complete orthonormal basis of eigenvectors. Although limited extensions have been made to infinite dimensions, it is not clear yet what is the most natural analogue of Condition A.

The purpose of this section is to describe briefly two possible variational approaches to Atkinson's (finite dimensional) theory. We note [4, Thm. 7.8.2] that Condition A implies the formally stronger statement $\Delta \gg 0$ (5.1). Thus one may define H_Δ as the vector space $H^\otimes$ under the inner product $[\cdot,\cdot]_\Delta$ given by

$$[x,y]_\Delta=(x,\Delta y).$$

Atkinson [4, Thm. 6.7.2] has shown that the operators

$$A_n=\Delta^{-1}\Delta_n, \qquad n=1,\cdots,k$$

commute. This allows us to extend the approach of [9] directly, with Δ_0 there replaced by Δ here. We then obtain a complete orthonormal basis of eigenvectors for the orthocomplement of $\operatorname{Ker}\Delta_0$ in H_Δ, characterized in terms of constrained maxima of functionals of the form

$$\left|\sum_{n=1}^{k}\rho_n\frac{[x,A_nx]_\Delta}{[x,x]_\Delta}\right|. \tag{9.1}$$

Moreover, we again find that the eigenvectors may be chosen decomposable, so the functional in (9.2) can be given the simpler form

$$\left|\sum_{n=1}^{k}\frac{\rho_n\delta_n(x)}{\delta(x)}\right|.$$

Thus the eigentuples may be characterized by constrained maxima of linear functionals over the set $\boldsymbol{\alpha}(U)$, where

$$\alpha_n(u)=\frac{\delta_n(u)}{\delta(u)}, \qquad n=1,\cdots,k.$$

The principal disadvantage of the above approach is that the vectors ρ generating the linear functionals are eigentuple-dependent—cf. difficulty (iii) of §4. This difficulty may be eliminated by a preliminary eigenvalue transformation. Specifically, under Condition A a nonsingular affine transformation may be found after which $\Delta_n \gg 0$ for $n=1,\cdots,k$ [8, Cor. 2.9]. This allows us to repeat the analysis of §7, even though LD is not assumed, since the compactness properties used for existence of the minimaxima are automatic in finite dimensions. We thus obtain the following from Theorem 7.4, where $\mathfrak{D}_1$ is the vector space $H^\otimes$ under the inner product $[\cdot,\cdot]_1$.

THEOREM 9.1. *Assuming Condition* A *in finite dimensions, the eigentuples* $\boldsymbol{\lambda}^j$, u^j *of* (1.1) *may be obtained by successive recursive or minimax operations on the components of* $\boldsymbol{\beta}(U_0)$, *after an initial affine eigenvalue transformation. The corresponding* $u^{j\otimes}$ *form a complete orthonormal basis of the orthocomplement of* $\operatorname{Ker}\Delta_0$ *in* $\mathfrak{D}_1$.

REFERENCES

[1] F. ARSCOTT, *Periodic Differential Equations: An Introduction to Mathieu, Lamé and Allied Functions*, Macmillan, New York, 1964.

[2] _____, *Two-parameter eigenvalue problems in differential equations*, Proc. London Math. Soc., 14 (1964), pp. 459–470.

[3] F. ATKINSON, *Discrete and Continuous Boundary Value Problems*, Academic Press, New York, 1961.

[4] _____, *Multiparameter Eigenvalue Problems*, Vol. 1, Academic Press, New York, 1972.

[5] P. BINDING, *Another positivity result for determinantal operators*, Proc. Roy. Soc. Edin., 86A (1980), pp. 333–337.

[6] _____, *A variational approach to one and several parameter nonlinear eigenvalue problems*, Canad. J. Math., 33 (1981), pp. 210–228.

[7] _____, *Left definite multiparameter eigenvalue problems*, to appear.

[8] _____, *Multiparameter definiteness conditions*, Proc. Roy. Soc. Edin., 89A (1981), pp. 319–332.

[9] P. BINDING AND P. BROWNE, *A variational approach to multiparameter eigenvalue problems for matrices*, this Journal, 8 (1977), pp. 763–777.

[10] _____, *A variational approach to multiparameter eigenvalue problems in Hilbert space*, this Journal, 9 (1978), pp. 1054–1067.

[11] _____, *Positivity results for determinantal operators*, Proc. Roy. Soc. Edin., 81A (1978), pp. 267–271.

[12] _____, *Comparison cones for multiparameter eigenvalue problems*, J. Math. Anal. Appl., 77 (1980), pp. 132–149.

[13] P. BROWNE, *Abstract multiparameter theory* I, J. Math. Anal. Appl., 60 (1977), pp. 259–273.

[14] R. DUFFIN, *A minimax theory for overdamped networks*, Arch. Rat. Mech. Anal., 4 (1955), pp. 221–233.

[15] M. GREGUŠ, F. NEUMAN AND F. ARSCOTT, *Three-point boundary value problems in differential equations*, J. London Math. Soc., 3 (1971), pp. 429–436.

[16] K. HADELER, *Mehrparametrige und nichtlineare Eigenwertaufgaben*, Arch. Rat. Mech. Anal., 27 (1967), pp. 306–328.

[17] A. KÄLLSTRÖM AND B. SLEEMAN, *A left-definite multiparameter eigenvalue problem in ordinary differential equations*, Proc. Roy. Soc. Edin., 74A (1976), pp. 145–155.

[18] J. MEIXNER AND F. SCHÄFKE, *Mathieusche Funktionen und Sphäroidfunktionen*, Springer Verlag, Berlin, 1954.

[19] R. RICHARDSON, *Theorems of oscillation for two linear differential equations of the second order with two parameters*, Trans. Amer. Math. Soc., 13 (1912), pp. 22–34.

[20] G. ROACH AND B. SLEEMAN, *On the spectral theory of operator bundles*, Applic. Anal., 7 (1977), pp. 1–14.

[21] _____, *On the spectral theory of operator bundles* II, Applic. Anal., 9 (1979), pp. 29–36.

[22] B. SLEEMAN, *Completeness and expansion theorems for a two-parameter eigenvalue problem in ordinary differential equations using variational principles*, J. London Math. Soc., 6 (1973), pp. 705–712.

[23] _____, *Multiparameter Spectral Theory in Hilbert Space*, Pitman, London, 1978.

[24] _____, *Klein oscillation theorems for multiparameter eigenvalue problems in ordinary differential equations*, Nieuw Arch. voor Wiskunde, 27 (1979), pp. 341–362.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0012 $01.00/0

CONSTRUCTING SOLUTIONS TO $\int_0^1 f(j)A(j,x)/f(x)\,dj = \int_0^1 f(x)A(x,j)/f(j)\,dj$*

G. EDGAR PARKER† AND TERRY J. WALTERS‡

Abstract. Given a continuous function A from $[0,1]\times[0,1]$ into $(0,\infty)$, a function f so that if $x\in[0,1]$,

$$\int_0^1 f(j)A(j,x)/f(x)\,dj = \int f(x)A(x,j)/f(j)\,dj$$

is constructed. Approximations to f are made by forming the $n\times n$ matrix $a_{ik}=A((i-1)/n,(k-1)/n)$ and constructing the n-vector d, a solution to the nonlinear system $\Sigma_{k=1}^n d_i a_{ik} d_k^{-1} = \Sigma_{k=1}^n d_k a_{ki} d_i^{-1}$. For $x\in[0,1]$, n large enough and $(k-1)/n$ close to x, d_k approximates $f(x)$. The construction is adaptable for use on the computer and provides solutions in a situation where previously only a fixed-point proof was known.

1. Introduction. Given an irreducible $n\times n$ matrix A, Hartfiel [1] showed that there is a diagonal matrix D with positive diagonal entries so that corresponding row and column sums of DAD^{-1} are equal. This result was extended in [5] by the authors to show that if A is a continuous, positive-valued function on $[0,1]\times[0,1]$ then there is a continuous positive-valued function f on $[0,1]$ so that if $x\in[0,1]$ then

$$\int_0^1 f(j)A(j,x)/f(x)\,dj = \int_0^1 f(x)A(x,j)/f(j)\,dj.$$

Hartfiel's result from [1] was obtained by contradiction from a minimal value argument and the solution by the authors in [5] to the integral equation was obtained via Schauder's fixed point theorem; thus both were nonconstructive. Hartfiel's results are applicable when diagonal scalings of irreducible matrices are applicable. In addition, the integral equation proved to be a member of a class of equations including classical linear problems as well as other nonlinear problems some of which (see, for instance, [3]) arose from specific physical problems. Hence, constructive solutions to Hartfiel's problem are of practical value whenever diagonal scaling is, and constructive solutions to the integral equation would not only replace a fixed point argument but might well indicate a direction for computing solutions to a much wider class of problems.

That continuous solutions to the integral equation in the title might exist was originally conjectured by Hartfiel based on the form the sums from his diagonal matrices took. In this paper the integral equation is solved for f by showing that by solving the matrix problem for a matrix (built from A by letting the i,k entry be $A((i-1)/n,(k-1)/n)$) on a sufficiently fine grid; then the entries of the resulting diagonal matrix form a discrete approximation to f. Although the authors were not able to establish a priori error estimates, the proximity of any iterate to a solution of its matrix probem can be checked, and the iteration is suitable for use on a computer.

The following notation will be observed. The i,k entry of a matrix will be denoted by $a_{i,k}$. An $n\times n$ diagonal matrix D will be represented by an n-vector d whose ith-coordinate is the i,i entry of D. e_k will denote the vector in R^n whose only nonzero coordinate is the kth coordinate which is 1. j will denote the identity function on the numbers. If $M\subset R$, R_M^n will denote the subset of R^n whose elements' coordinates are

* Received by the editors May 26, 1981, and in revised form October 13, 1981.
† Department of Mathematics, Pan American University, Edinburg, Texas 78539.
‡ Mathematics Department, University of Alabama, University, Alabama 35486.

numbers in M. $\mathbb{C}$ will denote the complex numbers. If G is a function with a maximum or minimum value, then $G^*(G_*)$ denotes the maximum (minimum) value of G. If f is a function, then R_f denotes the range of f.

2. Theorems. Theorem 1 describes an iteration which approximates solutions to the matrix problem. The problem is reformulated as a variational problem, and a descent method is employed to build the iteration.

THEOREM 1. *Suppose that n is a natural number and A is an $n\times n$ matrix with positive entries. If $i\in\{1,2,\cdots,n\}$ define $f_i\colon R^n_{(0,\infty)}\to[0,\infty)$ by if $x\in R^n_{(0,1]}$, then*

$$f_i(x)=\sum_{k=1}^{n}\left(x_i a_{ik}x_k^{-1}-x_k a_{ki}x_i^{-1}\right).$$

Define $T\colon R^n_{(0,1]}\to[0,\infty)$ by if $x\in R^n_{(0,1]}$. Then

$$T(x)=\sum_{i=1}^{n}\left(f_i(x)\right)^2.$$

If $x\in R^n_{(0,1]}$ and $m\in\{1,2,\cdots,n\}$, define $t_{x,m}\colon(-\min\{x_1,x_2,\cdots,x_n\},\infty)\to R$ by if $\delta\in R$, $t_{x,m}(\delta)=T(x+\delta e_m)$ and define $P_{x,m}\colon\mathbb{C}\to\mathbb{C}$ by

$$P_{x,m}(z)=\left[\left((x_m+z)^2\sum_{k\neq m}a_{mk}x_k^{-1}\right)-\sum_{k\neq m}x_k a_{km}\right]\left[\left((x_m+z)^2\sum_{k\neq m}a_{mk}x_k^{-1}\right)+\sum_{k\neq m}x_k a_{km}\right]$$
$$-\sum_{i\neq m}\left[x_i a_{im}-(x_m+z)^2 a_{mi}x_i^{-1}+(x_m+z)\sum_{k\neq m}\left(x_i a_{ik}x_k^{-1}-x_k a_{ki}x_i^{-1}\right)\right]$$
$$\cdot\left[x_i a_{im}+(x_m+z)^2 a_{mi}x_i^{-1}\right].$$

If $m\in\{1,2,\cdots,n\}$ define $r_m\colon R^n_{(0,1]}\to[0,\infty)$ by if $x\in R^n_{(0,1]}$

$$r_m(x)=\begin{cases}0 & \text{if } t'_{x,m}(0)\geq 0 \text{ or } t_{x,m}(0)=0,\\ \min\{(|a|+|b|)\colon P_{x,m}(a+bi)=0\} & \text{if } t'_{x,m}(0)<0.\end{cases}$$

For $m\in\{1,\cdots,n\}$ define $S_m\colon R^n_{(0,1]}\to R^n_{(0,1]}$ by if $x\in R^n_{(0,1]}$

$$S_m(x)=\frac{(x+r_m(x)e_m)}{\max\{(x+r_m(x)\cdot e_m)_i\colon i\in\{1,\cdots,n\}\}}$$

and $S\colon R^n_{(0,1]}\to R^n_{(0,1]}$ by $S=S_n\circ S_{n-1}\circ\cdots\circ S_1$. Then the iteration $x_1=(1,\cdots,1)$, $x_{k+1}=S(x_k)$ converges to a vector d so that if $i\in\{1,\cdots,n\}$,

$$\sum_{k=1}^{n}d_i a_{ik}d_k^{-1}=\sum_{k=1}^{n}d_k a_{ki}d_i^{-1}.$$

Theorem 2 describes the embedding through which the construction from Theorem 1 is used to approximate solutions to the integral equations from the title.

THEOREM 2. *Suppose that A is a continuous, positive-valued function on $[0,1]\times[0,1]$ and that f is the continuous positive-valued function on $[0,1]$ so that $f(0)=1$ and if $x\in[0,1]$, then*

$$\int_0^1 f(j)A(j,x)/f(x)\,dj=\int_0^1 f(x)A(x,j)/f(j)\,dj.$$

If $n \in N$ *define* (A_{ik}^n) *to be the* $n \times n$ *matrix whose* (i,k) *entry is* $A((i-1)/n,(k-1)/n)$ *and define* d^n *to be the n-vector so that* $d_1^n = 1$ *and if* $i \in \{1,\cdots,n\}$, *then*

$$\sum_{k=1}^{n} d_i^n A_{ik}^n (d_k^n)^{-1} = \sum_{k=1}^{n} d_k^n A_{ki}^n (d_i^n)^{-1}.$$

Suppose also that $x \in [0,1]$ *and that* $k_1, k_2, \cdots$ *is a sequence of natural numbers so that* $\lim_{n\to\infty} k_n/2^n = x$.

Then $\lim_{n\to\infty} d_{k_n}^{2^n} = f(x)$.

3. Proofs of the theorems. The hypothesis of Theorem 1 is the definition of several functions whose properties guarantee both the result and the utility of the iteration. The proof will primarily consist of establishing the special properties of these functions.

Proof of Theorem 1.

Property t1. Suppose that $x \in R_{(0,\infty)}^n$ and there is $p \in \{1,2,\cdots,n\}$ so that $f_p(x) \neq 0$. Then there are i,k in $\{1,\cdots,n\}$ so that $f_i(x) \neq f_k(x)$.

Suppose that if i and k are in $\{1,\cdots,n\}$, then $f_i(x) = f_k(x)$. Then since if $i \in \{1,\cdots,n\}$, $f_i(x) = f_p(x)$, then

$$\begin{aligned} n \cdot f_p(x) &= n\left(\sum_{k=1}^{n} x_p a_{pk} x_k^{-1} - \sum_{k=1}^{n} x_k a_{kp} x_p^{-1} \right) \\ &= \sum_{i=1}^{n} \left(\sum_{k=1}^{n} x_i a_{ik} x_k^{-1} - \sum_{k=1}^{n} x_k a_{ki} x_i^{-1} \right) = 0, \end{aligned}$$

a contradiction since $f_p(x) \neq 0$.

Property t2. Suppose that $x \in R_{(0,1]}^n$. Either if $i \in \{1,\cdots,n\}$ then $t_{x,i}(0) = 0$ or there exists $m \in \{1,\cdots,n\}$ so that $t'_{x,m}(0) < 0$.

Suppose that there is $i \in \{1,2,\cdots,n\}$ so that $t_{x,i}(0) \neq 0$. For $p \in \{1,\cdots,n\}$ and $\delta \in (-\min\{x_1,x_2,\cdots,x_n\}, \infty)$,

$$\begin{aligned} t_{x,p}(\delta) &= T(x+\delta e_p) = \sum_{i=1}^{n} \left(f_i(x+\delta e_p) \right)^2 \\ &= \left(\sum_{k\neq p} \left((x_p+\delta) a_{pk} x_k^{-1} - x_p a_{kp} (x_p+\delta)^{-1} \right) \right)^2 \\ &\quad + \sum_{i\neq p} \left[\sum_{k\neq p} \left(x_i a_{ik} x_k^{-1} - x_k a_{ki} x_i^{-1} \right) + x_i a_{ip} (x_p+\delta)^{-1} - (x_p+\delta) a_{pi} x_i^{-1} \right]^2. \end{aligned}$$

Thus,

$$\begin{aligned} t'_{x,p}(\delta) &= 2\left[\sum_{k\neq p} \left((x_p+\delta) a_{pk} x_k^{-1} - x_k a_{kp} (x_p+\delta)^{-1} \right) \right] \\ &\quad \cdot \left[\sum_{k\neq p} \left(a_{pk} x_k^{-1} + x_k a_{kp} (x_p+\delta)^{-2} \right) \right] \\ &\quad - 2 \sum_{i\neq p} \left[\sum_{k\neq p} \left(x_i a_{ik} x_k^{-1} - x_k a_{ki} x_i^{-1} \right) + x_i a_{ip} (x_p+\delta)^{-1} - (x_p+\delta) a_{pi} x_i^{-1} \right] \\ &\quad \cdot \left[x_i a_{ip} (x_p+\delta)^{-2} + a_{pi} x_i^{-1} \right], \end{aligned}$$

and

$$t'_{x,p}(0)=2\left[\sum_{k\neq p}\left(x_p a_{pk}x_k^{-1}-x_k a_{kp}x_p^{-1}\right)\right]\left[\sum_{k\neq p}\left(a_{pk}x_k^{-1}+x_k a_{kp}x_p^{-2}\right)\right]$$
$$-2\sum_{i\neq p}\left[\sum_{k\neq i}\left(x_i a_{ik}x_k^{-1}-x_k a_{ki}x_i^{-1}\right)\right]\left[x_i a_{ip}x_p^{-2}+a_{pi}x_i^{-1}\right].$$

Since i and k index the same set, this can be rewritten as

$$t'_{x,p}(0)=2\cdot\sum_{i\neq p}\left(a_{pi}x_i^{-1}+x_i a_{ip}x_p^{-2}\right)\left[f_p(x)-f_i(x)\right].$$

Let m be so that $f_m(x)=\min\{f_i(x): i\in\{1,2,\cdots,n\}\}$. If $i\in\{1,\cdots,n\}$, $f_m(x)-f_i(x)\leq 0$ and, from tl, for at least one index is strictly negative. Since $a_{mi}x_i^{-1}+x_i a_{im}x_m^{-2}>0$, $t'_{x,m}(0)<0$.

Property t3. Suppose that $x\in R^n_{(0,\infty)}$ and $p\in\{1,\cdots,n\}$. Then there is $\Delta>0$ so that if $\delta\geq\Delta$, $t'_{x,p}(\delta)>0$.

Rearranging the terms of $t'_{x,p}$ so the positive and negative terms are grouped,

$$t'_{x,p}(\delta)=2\left(\sum_{k\neq p}(x_p+\delta)a_{pk}x_k^{-1}\cdot\sum_{k\neq p}\left(a_{pk}+x_k a_{kp}(x_p+\delta)^{-2}\right)\right)$$
$$+2\sum_{i\neq p}\left(\left(\sum_{k\neq p}x_k a_{ki}x_i^{-1}\right)+(x_p+\delta)a_{pi}x_i^{-1}\right)\left(x_i a_{ip}(x_p+\delta)^{-2}+a_{pi}x_i^{-1}\right)$$
$$-2\left(\sum_{k\neq p}x_k a_{kp}(x_p+\delta)^{-1}\right)\left(\sum_{k\neq p}\left(a_{pk}x_k^{-1}+x_k a_{kp}(x_p+\delta)^{-2}\right)\right)$$
$$-2\sum_{i\neq p}\left(\sum_{k\neq p}\left(x_i a_{ik}x_k^{-1}+x_i a_{ip}(x_p+\delta)^{-1}\right)\right)\cdot\left(x_i a_{ip}(x_p+\delta)^{-2}+a_{pi}x_i^{-1}\right).$$

Note that no term of the negative part contains a positive power of $x_p+\delta$; hence for $\delta\geq 0$ the negative part is bounded. The positive part, however, contains terms with $x_p+\delta$ and hence is unbounded for $\{\delta:\delta\geq 0\}$. Property t3 follows.

Property P describes the nature of the roots of the polynomial P.

Property P. Suppose that $x\in R^n_{(0,\infty)}$ and that $t'_{x,p}(0)<0$. Then there is a positive real number r so that $P_{x,p}(r)=0$.

From the definition of P, if δ is a real number in $(-\min\{x_1,x_2,\cdots,x_n\},\infty)$, $t'_{x,p}(\delta)=2(x_p+\delta)^{-3}P_{x,p}(\delta)$. $t'_{x,p}$ is continuous on $[0,\infty)$, negative at 0 by assumption and eventually positive. But $(x_p+\delta)^{-3}\neq 0$, so there must exist $\gamma>0$ so that $P_{x,p}(\gamma)=0$.

Of interest here is the fact that, in addition, $P_{x,p}$ is a fourth degree polynomial whose coefficients are real and computable from the coordinates of x and the values from A. Thus, the zeroes of P can be computed algebraically.

S is a search algorithm in the iteration; Property S concerns its continuity.

Property S. S is continuous.

Consider a sequence $\{x_i\}_{i=1}^{\infty}$ in $R^n_{(0,1]}$ converging to x in $R^n_{(0,1]}$. The coefficients of $P_{x_i,p}$ converge to those of $P_{x,p}$, and $P_{x,p}$ has a positive lead coefficient; hence, the zeroes of $P_{x_i,p}$ converge to those of $P_{x,p}$ (see [2, p. 136]). It follows that r_m must be continuous which implies that S_m is continuous which implies that S is continuous.

Properties T-1, T-2 and T-3 are properties of T used in establishing the convergence of the iteration. Both T-1 and T-2 follow directly from the definition of T.

Property T-1. $T(x)=0$ if and only if $i\in\{1,2,\cdots,n\}$ implies $\Sigma_k x_i a_{ik} x_k^{-1}=\Sigma_k x_k a_{ki} x_i^{-1}$.

Property T-2. If $\delta>0$, then $T(\delta x)=T(x)$.

Property T-3. Suppose that $x\in R^n_{(0,1]}$ and $T(x)\neq 0$. Then $T(Sx)<T(x)$. By Property t2, $T(x)\neq 0$ implies that there is $m\in\{1,2,\cdots,n\}$ such that $t'_{x,m}(0)<0$. Since $r_m(x)$ is positive but no greater than the least positive zero of $t'_{x,m}, t_{x,m}$ is decreasing on $[0,r_m(x))$ and $T(x+r_m(x)e_m)<T(x)$. Also $S_m(x)=T((\max\{(x+r_m(x)\cdot e_m)_i\colon i\in\{1,\cdots,n\}\}\cdot(x+j_m(x)\cdot e_m)))$ so that, by T-2, $T(S_m(x))<T(x)$. Thus, $S=S_n\circ\cdots\circ S_1$ and $T(x)\geq T(S_1(x))\geq T(S_2(S_1(x)))\geq\cdots\geq T(S(x))$. Let n be the least index such that $S_n(x)\neq x$. (Unless $T(x)=0$, Property t2 guarantees the existence of such an index.) $T(S_n\circ\cdots\circ S_1(x))<T(x)$ and it follows that $T(S(x))<T(x)$.

The argument for the iteration is a special case of the global convergence theorem in [4] (see [4, p. 125]). Note initially that if $x\in R^n_{(0,1]}$, $\max\{(S(x))_i\colon i\in\{1,\cdots,n\}\}=1$. Suppose that some subsequence of $x_1,\cdots,x_n,\cdots$ converged to a vector at least one coordinate of which was zero. Then for $\varepsilon>0$, there would exist $m_\varepsilon\in N$ and $q\in\{1,2,\cdots,n\}$ so that if $m\geq m_\varepsilon$ then $(x_m)_q<\varepsilon$. By T-3, $T(x_m)<T(1,\cdots,1)$. Let p be so that $(x_m)_p=1$. Then

$$T(x_m)-\left(f_p(x_m)\right)^2=\left(\sum_{k=1}^n a_{pk}x_k^{-1}-\sum_{k=1}^n x_k a_{kp}\right)^2,$$

$$n\cdot\min\{a_{ij}\}\leq\sum_{k=1}^n x_k a_{kp}\leq\sum_{k=1}^n a_{kp}\leq n\cdot\max\{a_{ij}\},$$

$$\sum_{k=1}^n a_{pk}x_k^{-1}\geq a_{pq}x_q^{-1}\geq a_{pq}\left(\frac{1}{\varepsilon}\right)\geq\min\{a_{ij}\}\left(\frac{1}{\varepsilon}\right).$$

But this implies that $T(x_m)$ can be made arbitrarily large, a contradiction. Therefore, there is $\varepsilon>0$ so that $\{x_1,x_2,\cdots,x_n,\cdots\}$ is a subset of $R^n_{[\varepsilon,1]}$. Let $y_1, y_2,\cdots$ be a convergent subsequence of $x_1,\cdots,x_n,\cdots$ and d be its sequential limit point. As previously noted, S is continuous, so $S(y_1), S(y_2),\cdots$ converges to $S(d)$. From T-3, $T(y_1)\geq T(S(y_1))\geq T(y_2)\geq T(S(y_2))\geq\cdots\geq T(d)\geq T(S(d))$. But T is continuous by construction, so $T(d)=T(S(d))$ and, therefore, again from T-3, $T(d)=0$. Since the whole of $x_1,\cdots,x_i,\cdots$ is ordered by T, each convergent subsequence is forced by $y_1,\cdots,y_n\cdots$ to d. Therefore, $x_1,\cdots,x_n,\cdots$ converges to d and $T(d)=0$ as was to be proven.

Proof of Theorem 2. Suppose that $n\in N$ and note that d^n has been chosen so that $d_1^n=1$. Thus, $\Sigma_k 1\cdot A(0,(k-1)/n)(d_k^n)^{-1}=\Sigma_k d_k^n A((k-1)/n,0)\cdot 1$,

$$1=\frac{\Sigma_k d_k^n A\left(\frac{k-1}{n},0\right)}{\Sigma_k A\left(0,\frac{k-1}{n}\right)/d_k^n},$$

and it follows that

$$\frac{A_*}{A^*}\leq\frac{\Sigma_k d_k^n}{\Sigma_k 1/d_k^n}\leq\frac{A^*}{A_*}.$$

For $i\in\{1,2,\cdots,n\}$,

$$(d_i^n)^2=\frac{\Sigma_k d_i^n A\left(\frac{k-1}{n},\frac{i-1}{n}\right)}{\Sigma_k A\left(\frac{i-1}{n},\frac{k-1}{n}\right)/d_k^n}\leq\frac{A^*\Sigma_k d_k^n}{A_*\Sigma_k\frac{1}{d_k^n}}\leq\left(\frac{A^*}{A_*}\right)^2.$$

Using a similar argument for the opposite inequality,

(i) $$\{d_k^n: n\in N, 1\leq k\leq n\}\subset\left[\frac{A_*}{A^*},\frac{A^*}{A_*}\right].$$

Let W denote an ordering of the dyadic rationals in $[0,1)$, that is $\{(r-1)/2^s: S$ is an integer, $1\leq r\leq 2^s\}$, by the natural numbers. If $x\in[0,1]$, denote by $M(n,x)$ the greatest integer in $\{k:(k-1)/m\leq x\}$. Consider $W(1)$. There is a subsequence of $d^1_{M(1,W(1))}$, $d^2_{M(2,W(1))},\cdots,d^{2^n}_{M(2^n,W(1))},\cdots$ which converges. Define N_1 to be the set of indices for this subsequence. Then $\{d^i_{M(i,W(1))}\}_{i\in N_1}$ converges. Given $N_k\subset\bigcap_{p=1}^{k-1}N_p$ and that $\{d^n_{M(n,W(k))}\}_{n\in N_k}$ converges, define N_{k+1} to be a subset of N_k so that $\{d^n_{M(n,W(k+1))}\}_{n\in N_{k+1}}$ converges.

Define g by if $x=W(k)$, $g(x)=\lim_{n\in N_k} d^n_{M(n,x)}$. The inequalities that follow demonstrate the uniform behavior of the approximations to g.

Suppose that x and y are elements of $[0,1]$,

$$\left(\frac{d^n_{M(n,x)}}{d^n_{M(n,y)}}\right)^2$$

$$=\frac{\dfrac{\Sigma_k d_k^n A\left(\frac{k-1}{n},\frac{M(n,x)-1}{n}\right)}{\Sigma_k A\left(\frac{M(n,x)-1}{n},\frac{k-1}{n}\right)(d_k^n)^{-1}}}{\dfrac{\Sigma_k d_k^n A\left(\frac{k-1}{n},\frac{M(n,y)-1}{n}\right)}{\Sigma_k A\left(\frac{M(n,y)-1}{n},\frac{k-1}{n}\right)(d_k^n)^{-1}}}$$

$$=\frac{\Sigma_k d_k^n A\left(\frac{k-1}{n},\frac{M(n,x)-1}{n}\right)}{\Sigma_k d_k^n A\left(\frac{k-1}{n},\frac{M(n,y)-1}{n}\right)}\cdot\frac{\Sigma_k A\left(\frac{M(n,y)-1}{n},\frac{k-1}{n}\right)(d_k^n)^{-1}}{\Sigma_k A\left(\frac{M(n,x)-1}{n},\frac{k-1}{n}\right)(d_k^n)^{-1}}.$$

Define

$$\varepsilon_n=\max\left\{\left|A\left(\frac{k-1}{n},\frac{M(n,x)-1}{n}\right)-A\left(\frac{k-1}{n},\frac{M(n,y)-1}{n}\right)\right|:k\in\{1,2,\cdots,n\}\right\},$$

$$\delta_n=\max\left\{\left|A\left(\frac{M(n,x)-1}{n},\frac{k-1}{n}\right)-A\left(\frac{M(n,y)-1}{n},\frac{k-1}{n}\right)\right|:k\in\{1,2,\cdots,n\}\right\}$$

and

$$\gamma_n=\max\{\varepsilon_n,\delta_n\}.$$

Adding, in the numerator of each factor, the denominator plus its inverse,

$$\frac{d^n_{M(n,x)}}{d^n_{M(n,y)}}^2 \leq \frac{\Sigma_k d^n_k A\left(\frac{k-1}{n}, \frac{M(n,y)-1}{n}\right)+\varepsilon_n \Sigma_k d^n_k}{\Sigma_k d^n_k A\left(\frac{k-1}{n}, \frac{M(n,y)-1}{n}\right)}$$

$$\cdot \frac{\Sigma_k A\left(\frac{M(n,x)-1}{n}, \frac{k-1}{n}\right)(d^n_k)^{-1}+\delta_n \Sigma_k (d^n_k)^{-1}}{\Sigma_k A\left(\frac{M(n,x)-1}{n}, \frac{k-1}{n}\right)}$$

$$\leq \left(1+\frac{\varepsilon_n}{A_*}\right)\left(1+\frac{\delta_n}{A_*}\right)$$

and

$$\frac{d^n_{M(n,x)}}{d^n_{M(n,y)}} \leq 1+\frac{\gamma_n}{A_*}.$$

Therefore $d^n_{M(n,x)} - d^n_{M(n,y)} \leq (\gamma_n/A_*) d^n_{M(n,y)}$, which by (i) is no greater than $(\gamma_n/A_*)A^*/A_*$. So

(ii) $$d^n_{M(n,x)} - d^n_{M(n,y)} \leq \frac{\gamma_n A^*}{A_* A_*}.$$

The following argument demonstrates that f is uniformly continuous on R_w. Let $\varepsilon>0$. Since A is uniformly continuous, there is $\delta>0$ so that if $\|u-v\|<\delta$, then $|A(u)-A(v)|<A_*^2/A^* \cdot \varepsilon/2$. Pick p and q so that $p=W(i)$ and $q=W(m)$. Then for $n \in N_i \cap N_m$, $|g(p)-g(q)| \leq |g(p)-d^n_{M(n,p)}| + |d^n_{M(n,p)} - d^n_{M(n,q)}| + |d^n_{M(n,q)} - g(q)|$. For each n, since p and q are dyadics, $p=M(n,p)/n$ and $q=M(n,q)/n$. Therefore, whenever

$$|p-q| = \left|\frac{M(n,p)}{n} - \frac{M(n,q)}{n}\right| = \left\|\left(\frac{M(n,p)-1}{n}, \frac{k-1}{n}\right) - \left(\frac{M(n,q)-1}{n}, \frac{k-1}{n}\right)\right\|$$

$$= \left\|\left(\frac{k-1}{n}, \frac{M(n,p)-1}{n}\right) - \left(\frac{k-1}{n}, \frac{M(n,q)-1}{n}\right)\right\| \leq \delta,$$

$$|g(p)-g(q)| < |g(p)-d^n_{M(n,p)}| + \frac{\varepsilon}{2} + |d^n_{M(n,q)} - g(q)|.$$

But $|M(n,p)/n - M(n,q)/n|$ is independent of the choice of n, so the other two terms can be made arbitrarily small. Hence g is uniformly continuous on R_W and thus has a unique continuous extension to $\bar{R}_W=[0,1]$.

Now consider the convergence of $\{d^n_{M(n,x)}\}_{x=W(k),\, n\in N_k}$ to g. Suppose that there is $\varepsilon>0$ so that if n is a natural number then there is a natural number $k_n \leq 2^n$ so that $|d^{2^n}_{k_n} - g((k_n-1)/2^n)| \geq \varepsilon$. $\{(k_n-1)/2^n\}_{n-1}^{\infty}$ has a convergent subsequence since each term is a number in $[0,1]$. Let $\{y_n\}_{n=1}^{\infty}$ be such a subsequence and y its sequential limit point. Let $x \in R_w$. $|d^{2^n}_{m(2^n,y_n)} - g(y_n)| \leq |d^{2^n}_{m(2^n,y_n)} - d^{2^n}_{m(2^n,x)}| + |d^{2^n}_{m(2^n,x)} - g(x)| + |g(x) - g(y_n)|$. Through the uniform continuity of A and g, there is $m \in N$, and $\delta>0$ so that if $i>m$ and $x \in [0,1]$ so that $|x-(k_i-1)/i|<\delta$ then $|d^{2^i}_{m(2^i,y_i)} - d^{2^i}_{m(2^i,x)}| + |g(x)-g(y_i)| < \varepsilon/2$. So pick $x \in (y-\delta/2, y+\delta/2)$, $\lim_{2^n \varepsilon N_w - 1_{(x)}} d^{2^n}_{m(2^n,x)} = g(x)$; therefore, there is $i>m$

so that $|d^{2^i}_{m(2^i,x)}-g(x)|<\varepsilon/2$. But this contradicts the original assumption. Therefore,

(iii) if $\varepsilon>0$, there is a natural number m so that $1/m<\varepsilon$ and if $1\le k\le 2^m$, then $|d^{2^m}_k-g((k-1)/2^m)|<\varepsilon$.

Let $\varepsilon>0$ and consider, for $x\in[0,1]$,

$$\begin{aligned}
&\left|\int_0^1 g\cdot A(j,x)/g(x)\,dj-\int_0^1 g(x)A(x,j)/g(j)\,dj\right| \\
&\quad\le\left|\int_0^1 g(j)A(j,x)/g(x)\,dj-\int_0^1 g(j)A(j,y)/g(y)\,dy\right| \\
&\quad+\left|\int_0^1 g(j)A(j,y)/g(y)\,dy-\frac{1}{n}\sum_k g\left(\frac{k-1}{n}\right)A\left(\frac{k-1}{n},y\right)/g(y)\right| \\
&\quad+\left|\frac{1}{n}\sum_k g\left(\frac{k-1}{n}\right)A\left(\frac{k-1}{n},y\right)/g(y)-\frac{1}{n}\sum_k d^n_k A\left(\frac{k-1}{n},y\right)/g(y)\right| \\
&\quad+\left|\frac{1}{n}\sum_k d^n_k A\left(\frac{k-1}{n},y\right)/g(y)-\frac{1}{n}\sum_k d^n_k A\left(\frac{k-1}{n},y\right)/d^n_{m(n,y)}\right| \\
&\quad+\left|\frac{1}{n}\sum_k d^n_k A\left(\frac{k-1}{n},y\right)/d^n_{m(n,y)}-\frac{1}{n}\sum_k d^n_k A\left(\frac{k-1}{n},\frac{M(n,y)-1}{n}\right)/d^n_{m(n,y)}\right| \\
&\quad+\left|\frac{1}{n}\sum_k d^n_k A\left(\frac{k-1}{n},\frac{M(n,y)-1}{n}\right)/d^n_{m(n,y)}-\int_0^1 g(x)A(x,j)/g(j)\,dj\right|.
\end{aligned}$$

Pick y in R_W so that

$$\left|\frac{1}{g(x)}-\frac{1}{g(y)}\right|<\frac{\varepsilon}{10A^*\int_0^1 g(j)\,dj}.$$

Find L so that if $n\ge L$

$$\left|\int_0^1 g(j)A(j,y)/g(y)\,dj-\frac{1}{n}\sum_k g\left(\frac{k-1}{n}\right)\cdot A\left(\frac{k-1}{n},y\right)/g(y)\right|<\frac{\varepsilon}{10},$$

$$\left|\frac{1}{n}\sum_k A\left(\frac{k-1}{n},y\right)/g(y)/\int_0^1 A(j,y)/g(y)\,dj\right|<1+\varepsilon,$$

and

$$\left|A(j,y)-A\left(j,\frac{M(n,y)-1}{n}\right)\right|<\frac{\varepsilon A_*^2}{10A^{*2}}.$$

From (ii) there is $n>L$ so that n is a power of 2, and if $1\le k\le n$, then

$$\left|d^n_k-g\left(\frac{k-1}{2^i}\right)\right|<\frac{\varepsilon}{10\int_0^1 A(j,y)/g(y)\,dj},$$

and

$$\left|\frac{1}{d^n_k}-\frac{1}{g\left(\frac{k-1}{n}\right)}\right|<\frac{\varepsilon A_*}{10A^*\int_0^1 A(j,y)/g(y)\,dj}.$$

Hence

$$\left|\int_0^1 g(j)A(j,x)/g(x)\,dj-\int_0^1 g(x)A(x,j)/g(j)\,dj\right|$$

$$\leq A^*\left|\frac{1}{g(x)}-\frac{1}{g(y)}\right|\left|\int_0^1 g(j)\,dj+\frac{\varepsilon}{10}+\left|\frac{1}{n}\sum_k\left(g\left(\frac{k-1}{n}\right)-d_k^n\right)A\left(\frac{k-1}{n},y\right)/g(y)\right|\right.$$

$$+\left|\frac{1}{n}\sum_k\left(\frac{1}{g(y)}-\frac{1}{d^n_{M(n,y)}}\right)d_k^n A\left(\frac{k-1}{n},y\right)\right|$$

$$+\left|\frac{1}{n}\sum_k d_k^n A\left(\frac{k-1}{n},y\right)/d^n_{M(n,y)}-\frac{1}{n}\sum_k d_k^n A\left(\frac{k-1}{n},\frac{m(n,y)-1}{n}\right)/d^n_{M(n,y)}\right|$$

$$+\left|\frac{1}{n}\sum_k d_k^n A\left(\frac{k-1}{n},\frac{m(n,y)-1}{n}\right)/d^n_{M(n,y)}-\int_0^1 g(x)A(x,j)/g(j)\,dj\right|,$$

which by the choice of n and the fact that $d_k^n<A^*/A_*$ from (i) guarantees that the sum is less than

$$\frac{\varepsilon}{2}+\left|\frac{1}{n}\sum_k d_k^n A\left(\frac{k-1}{n},\frac{m(n,y)-1}{n}\right)/d^n_{M(n,y)}-\int_0^1 g(x)A(x,j)/g(j)\,dj\right|.$$

But

$$\frac{1}{n}\sum_n d_k^n A\left(\frac{k-1}{n},\frac{m(n,y)-1}{n}\right)/d^n_{M(n,y)}=\frac{1}{n}\sum_k d^n_{M(n,y)}A\left(\frac{m(n,y)-1}{n},\frac{k-1}{n}\right)/d_k^n,$$

so mimicking the telescoping from the above argument allows a choice for y, L and n so that

$$\left|\int_0^1 g(j)A(j,x)/g(x)\,dj-\int_0^1 g(x)A(x,j)/g(j)\,dj\right|$$

can be made arbitrarily small. Therefore, if $x\in[0,1]$,

$$\int_0^1 g(j)A(j,x)/g(x)\,dj=\int_0^1 g(x)A(x,j)/g(j)\,dj,$$

and since if n is a natural number, $d_1^n=1$, $g(0)=1$.

By [5, Thm. 3], there is exactly one function f so that $f(0)=1$ and if $x\in[0,1]$, then

$$\int_0^1 f(j)A(j,x)/f(x)\,dj=\int_0^1 f(x)A(x,j)/f(j)\,dj.$$

Since any other choice of convergent subsequences in the construction would produce a limit function satisfying the equation above, it follows that $\lim_{k\to\infty} d^{2^k}_{M(2^k,x)}=g(x)$.

Suppose $x\in[0,1]$ and $k_1, k_2, \cdots$ are natural numbers so that $\lim_{k\to\infty}(k_n/2^n)=x$. $|d^{2^n}_{k_n}-g(x)|\leq|d^{2^n}_{k_n}-d^{2^n}_{M(2^n,y)}|+|d^{2^n}_{M(2^n,y)}-g(y)|+|g(y)-g(x)|$. Pick y close to x and the third term is small by the continuity of g. For large n $|(k_n-1)/2^n-y|$ is small, and (i) forces the first term to be small. Picking $y\in R_w$ forces, for large n, the middle term to be small. Hence, $\lim_{k\to\infty} d^{2^n}_{k_n}=g(x)$, as was to be proved.

4. Some remarks and questions. In considering the use of Theorem 2 it should be noted that the iteration from Theorem 1 depends only on the solution of a fourth degree polynomial and therefore can be programmed, and the function T from the hypothesis can be used to compute the proximity of any iterate to an answer. On the

other hand, Theorem 2 offers no estimate relating the dimension of the matrix approximating A to the proximity of $d^n_{M(n,x)}$ to $f(x)$, and there is no a priori guarantee of accuracy. The authors programmed the algorithm exactly as stated in Theorem 1 on a 16×16 grid and obtained 8 digits of accuracy in 16 digit arithmetic on examples where a solution function was known in advance.

A theorem guaranteeing error estimates dependent on the properties of A would greatly increase the utility of the construction.

As was mentioned in the introduction, the fixed point solution in [5] generalized to include other integral equations. Of interest here would be the identification of related integral equations whose solutions can also be approximated through matrix embeddings.

REFERENCES

[1] D. J. HARTFIEL, *Concerning diagonal similarity of irreducible matrices*. Proc. of the Amer. Math. Soc., 30 (1981), pp. 419–425.

[2] E. ISAACSON AND H. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.

[3] S. KARLIN AND L. NIRENBURG, *On a theorem of P. Nowosad*, J. Math. Anal. Appl., 17 (1967), pp. 61–67.

[4] DAVID G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[5] TERRY J. WALTERS AND G. EDGAR PARKER, *Positive nonlinear integral equations with reciprocals of the solution in the integrand*, J. Nonlinear Anal.: Theory, Methods and Application, 5 (1981), pp. 1163–1172.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0013 $01.00/0

THE RATE OF MONOTONE SPLINE APPROXIMATION IN THE L_p-NORM*

D. LEVIATAN† AND H. N. MHASKAR‡

Abstract. We obtain Jackson type estimates for the approximation of monotone nondecreasing functions f by monotonically nondecreasing splines with equally spaced knots in the $L_p[0,1]$-norm $1\le p\le\infty$. The estimates involve high order moduli of smoothness of some derivatives of f and are obtained for functions f with a continuous derivative in case $p=\infty$ and for functions f that are the second primitive of $f''\in L_p[0,1]$ if $1\le p<\infty$.

1. Introduction. In a recent paper [1] Chui, Smith and Ward have extended to $L_p[0,1]$ some of the important results of De Vore [3] concerning Jackson type estimates for approximating in the $C[0,1]$-norm, monotonically nondecreasing functions by monotonically nondecreasing splines. While De Vore's construction and proofs [3] are ingenious as well as complicated, they seem to work only in the $C[0,1]$ case. On the other hand, Chui, Smith and Ward's construction and proofs [1] are simpler and apply to all L_p, $1\le p\le\infty$. However, they are unable to obtain estimates that involve higher order moduli of smoothness. Following some of their ideas, we are able to obtain Jackson type estimates involving higher order moduli of smoothness. De Vore [3, p. 904] remarks that if f has a continuous nonnegative derivative f', then his method would yield an estimate of the order $n^{-1}\omega_{r-1}(f',n^{-1})$. We will prove that, using the simpler method of [1]. Also we will provide L_p-estimates for a function f which is the second primitive of $f''\in L_p[0,1]$. These estimates will involve the $(r-2)$ modulus of smoothness of f''. In this paper $r\ge 2$ will denote the order of the splines.

Throughout this paper C, C_1, etc., will denote constants which may depend on p and r but are independent of f, g and n.

2. The main results. For $1\le p\le\infty$ let $L_p[0,1]$ denote the space of functions whose p-th power is integrable if $1\le p<\infty$ and the space of continuous functions if $p=\infty$. Given $f\in L_p[0,1]$, define its r-th L_p-modulus of smoothness by

$$\omega_{rp}(f,h)=\sup_{0\le t\le h}\|\Delta_t^r(f,x)\|_p(I(rt)),$$

where the pth norm is taken over the interval $I(rt)=[0,1-rt]$ and Δ^r is the rth forward difference. We will write

$$\omega_r(f,h)=\omega_{r\infty}(f,h).$$

Let $\mathfrak{S}(r,n)$ $(r\ge 1)$ denote the space of all splines of order r on the $n+1$ equally spaced knots $\{\frac{i}{n}\}_{i=0}^n$, i.e., $s\in\mathfrak{S}(r,n)$, if s is a polynomial of degree $\le r-1$ in each interval $[\frac{i}{n},\frac{i+1}{n}]$ and $s^{(r-2)}$ is continuous in $[0,1]$ (if $r=1$ s is a piecewise constant with no continuity at the knots).

If $f\in L_p[0,1]$ is monotonically nondecreasing, denote

$$E_{np}^*(f,r)=\inf\{\|f-s\|_p : s\in\mathfrak{S}(r,n), s\uparrow\}.$$

If $p=\infty$ we write $E_n^*(f,r)=E_{n\infty}^*(f,r)$.

---

* Received by the editors March 16, 1981.

† Department of Mathematics, California Institute of Technology, Pasadena, California 91125, and Department of Mathematics, Tel Aviv University, Ramat Aviv, Israel.

‡ Department of Mathematics, California State University, Los Angeles, California 90032.

First we have the case $p=\infty$.

THEOREM 2.1. *Let $r\geq 2$. If f possesses a continuous nonnegative derivative f' on* $[0,1]$, *then*

$$E_n^*(f,r)\leq Cn^{-1}\omega_{r-1}(f',n^{-1}). \tag{2.1}$$

For $p<\infty$ we show

THEOREM 2.2. *Let $1\leq p<\infty$ and $r\geq 3$. If f is the second primitive of $f''\in L_p[0,1]$ and if f is nondecreasing, then*

$$E_{np}^*(f,r)\leq Cn^{-2}\omega_{r-2,p}(f'',n^{-1}). \tag{2.2}$$

Remark. A more desirable estimate would be

$$E_{np}^*(f,r)\leq C\omega_{rp}(f,n^{-1}).$$

For this would yield inverse theorem characterizing smoothness properties of f by its rate of approximation by monotone splines. However, since Chui, Smith and Ward's technique (and ours here) involves approximating f by splines that interpolate it at some of the knots and, in particular, at $x=1$, it is hopeless to expect that the method will yield that for $1\leq p<\infty$, as is explained below.

By the same reasoning we give below, it is not possible to find for $1\leq p<\infty$ an interpolating spline that yields [1, Thm. 1.1] for the case $j=0$.

Suppose that for each $f\in L_p[0,1]$ monotonically nondecreasing there is a monotonically nondecreasing spline $s\in\mathcal{S}(r,n)$ such that $s(1)=f(1)$ and

$$\|f-s\|_p\leq C\omega_{rp}(f,n^{-1}).$$

Then from the inequality

$$\omega_{rp}(f,\delta)\leq 2^r\|f\|_p$$

we immediately get

$$\|s\|_p\leq C_1\|f\|_p.$$

Now the space $\mathcal{S}(r,n)$ is finite-dimensional, and so all norms on it are equivalent. Hence,

$$|f(1)|=|s(1)|=\|s\|_\infty\leq C_2\|f\|_p. \tag{2.3}$$

If $p<\infty$, then it is possible to find a sequence of increasing functions f_k such that $f_k(1)=k$ and $\|f_k\|_p\leq 1$ eventually contradicting (2.3).

Our strategy is to use estimates on the rate of approximation of f by polynomials of degree r in order to obtain good approximation to f by monotone piecewise polynomials; once we have achieved that, we replace the piecewise polynomials by suitable splines following the technique of Chui, Smith and Ward [1].

3. Monotone approximation by piecewise polynomials. In this section it is convenient to have the functions, their norms and their moduli of smoothness defined in $L_p[-1,1]$ rather than $L_p[0,1]$ (the modifications are obvious).

THEOREM 3.1. (i) *Let f be continuously differentiable on* $[-1,1]$ *and nondecreasing there. For $r\geq 1$ there exists a nondecreasing continuous function g on* $[-1,1]$ *such that g interpolates f at* 0 *and* ± 1 *and has the following properties*:

The restrictions of g to $[-1,0]$ *and to* $[0,1]$ *are polynomials of degree* $\leq r$, (3.1)

$$\|f-g\|_\infty\leq C\omega_r(f',1) \tag{3.2}$$

and

$$\sum_{k=1}^{r} |g^{(k)}(0+)-g^{(k)}(0-)| \le C\omega_r(f',1). \tag{3.3}$$

(ii) *Let* $1 \le p < \infty$ *and let* f *be the second primitive of* $f'' \in L_p[-1,1]$ *and such that* $f' \ge 0$. *Then there exists a nondecreasing continuous* g *on* $[-1,1]$ *interpolating* f *at* 0 *and* ± 1 *and satisfying* (3.1) *such that*

$$\|f-g\|_p \le C\omega_{r-1}(f'',1) \tag{3.4}$$

and

$$\sum_{k=1}^{r} |g^{(k)}(0+)-g^{(k)}(0-)| \le C\omega_{r-1}(f'',1). \tag{3.5}$$

If f possesses a continuous derivative, then by [2, Thm. 3.1] there is a polynomial q of degree $\le r-1$ such that

$$\|f'-q\|_\infty \le C\omega_r(f',1). \tag{3.6}$$

Since $f' \ge 0$ it follows immediately that

$$\delta = -2\min\Big\{\min_{-1\le x\le 1} q(x), 0\Big\} \le C\omega_r(f',1). \tag{3.7}$$

If f is the second primitive of $f'' \in L_p[-1,1]$, then again by [2, Thm. 3.1], there is a polynomial q_1 of degree $\le r-2$ such that

$$\|f''-q_1\|_p \le C\omega_{r-1,p}(f'',1).$$

Let $q(x) = f'(0) + \int_0^x q_1(t)\,dt$. Then

$$\|f'-q\|_p \le C\omega_{r-1,p}(f'',1). \tag{3.6'}$$

Now $\min_{-1<x<1} q(x) = q(a)$ for some $-1 \le a \le 1$ and since $f' \ge 0$,

$$\begin{aligned} \delta &= -2\min\Big\{\min_{-1\le x\le 1} q(x), 0\Big\} \\ &\le 2|f'(a)-q(a)| \\ &= 2\left|\int_0^a [f''(t)-q_1(t)]\,dt\right| \le 2C\omega_{r-1,p}(f'',1). \end{aligned} \tag{3.7'}$$

For either case let $Q(x) = \int_0^x q(t)\,dt$. Then we prove:

LEMMA 3.1. *Either* $Q(1)=0$ *and* $\delta = 0$ *or* $\delta + Q(1) > 0$, *and we have*

$$\frac{\delta}{\delta+Q(1)} \le 2, \tag{3.8}$$

$$\frac{|Q(1)|}{\delta+Q(1)} \le 1. \tag{3.9}$$

Proof. If $Q(1)=0$ and $\delta > 0$, then $\delta + Q(1) > 0$ and (3.8) and (3.9) are evident. Thus, we have to deal only with the case where $Q(1) \ne 0$.

Now for all $x \in [0,1]$ $q(x) \ge \min_{-1\le t\le 1} q(t) \ge -\frac{\delta}{2}$. Hence $Q(1) = \int_0^1 q(x)\,dx \ge -\frac{\delta}{2}$; that is,

$$\delta + Q(1) \ge \frac{\delta}{2}. \tag{3.10}$$

Thus if $\delta=0$ then $Q(1)\geq 0$, and since $Q(1)\neq 0$, $\delta+Q(1)=Q(1)>0$. If $\delta>0$, then evidently $\delta+Q(1)>0$. So now we can divide by $\delta+Q(1)$ so that (3.8) follows immediately by (3.10). Also, (3.9) is obvious if $Q(1)\geq 0$ so that the only remaining case is $\delta>0$, $Q(1)<0$. Now by (3.10), $Q(1)\geq -\frac{\delta}{2}$; thus,

$$|Q(1)|=-Q(1)\leq\frac{\delta}{2}\leq\delta+Q(1),$$

and (3.9) follows.

LEMMA 3.2. *Let f be as in Theorem* 3.1. *Then there is a nondecreasing polynomial P on* $[-1,1]$ *of degree* $\leq r$ *which interpolates f at* 0 *and* 1 *and such that if* $p=\infty$,

$$\|f-P\|_\infty\leq C\omega_r(f',1), \tag{3.11}$$

and if $1\leq p<\infty$,

$$\|f-P\|_p\leq C\omega_{r-1,p}(f'',1). \tag{3.12}$$

Proof. If $Q(1)=0$ and $\delta=0$, then $Q(x)=q(x)\equiv 0$ and by virtue of (3.6) and (3.6′)

$$\|f'\|_\infty\leq C\omega_r(f',1) \quad \text{if } p=\infty,$$

and

$$\|f'\|_p\leq C\omega_{r-1,p}(f'',1) \quad \text{if } 1\leq p<\infty.$$

Then $|f(x)-f(0)|$ satisfies the same estimates and the linear function that interpolates f at 0 and 1 satisfies (3.11) or (3.12).

Otherwise, by Lemma 3.1 $\delta+Q(1)>0$. Now without loss of generality, we may assume that $f(0)=0$ and then let

$$P(x)=\frac{f(1)}{\delta+Q(1)}[Q(x)+\delta x].$$

We will prove the case $1\leq p<\infty$. The proof for $p=\infty$ is even simpler. It is clear that P interpolates f at 0 and 1 and that P is a nondecreasing polynomial of degree $\leq r$. Also,

$$P(x)-f(x)=Q(x)-f(x)+\frac{f(1)-\delta-Q(1)}{\delta+Q(1)}Q(x)+\frac{\delta f(1)}{\delta+Q(1)}x. \tag{3.13}$$

By virtue of (3.6′), for all $x\in[-1,1]$,

$$|Q(x)-f(x)|\leq\left|\int_0^x|q(t)-f'(t)|dt\right|\leq\|f'-q\|_p\leq C\omega_{r-1,p}(f'',1). \tag{3.14}$$

Now for all $x\in[-1,1]$,

$$\left|\frac{\delta f(1)}{\delta+Q(1)}x\right|\leq\frac{\delta f(1)}{\delta+Q(1)}\leq\frac{\delta}{\delta+Q(1)}|f(1)-Q(1)|+\frac{\delta|Q(1)|}{\delta+Q(1)}$$
$$\leq 2C\omega_{r-1,p}(f'',1)+C\omega_{r-1,p}(f'',1)$$

by (3.8) and (3.14) and by (3.7′) and (3.9). Thus, we proved

$$\left|\frac{\delta f(1)}{\delta+Q(1)}x\right|\leq C_1\omega_{r-1,p}(f'',1). \tag{3.15}$$

As for the third term in (3.13) notice that in view of (3.7′) and (3.14),

$$\left\|\frac{f(1)-\delta-Q(1)}{\delta+Q(1)}Q\right\|_p\leq 2C\omega_{r-1,p}(f'',1)\frac{\|Q\|_p}{\delta+Q(1)}, \tag{3.16}$$

so that by virtue of (3.13) through (3.16) the proof of (3.12) is complete once we show that

$$\|Q\|_p \le C(\delta + Q(1)). \tag{3.17}$$

To this end we proceed as in [1, Lemma 3.2]. Note that since Q belongs to the $r+1$ dimensional space of polynomials of degree $\le r$ it suffices to prove (3.17) with the sup-norm on $[0,1]$ replacing the p-norm on $[-1,1]$. So let $0 \le x_0 \le 1$ be such that $|Q(x_0)| = \|Q\|_{C[0,1]}$. Then if $Q(x_0) = \|Q\|_{C[0,1]}$, then for some $a \in (x_0, 1)$ we have

$$Q(1) - Q(x_0) = Q'(a)(1-x_0) = q(a)(1-x_0) \ge -\frac{\delta}{2}.$$

Hence

$$\|Q\|_{C[0,1]} = Q(x_0) \le Q(1) + \delta.$$

If on the other hand $Q(x_0) = -\|Q\|_{C[0,1]}$, then for some $b \in (0, x_0)$

$$Q(x_0) = Q(x_0) - Q(0) = Q'(b)x_0 = q(b)x_0 \ge -\frac{\delta}{2}.$$

Hence

$$\|Q\|_{C[0,1]} = -Q(x_0) \le \frac{\delta}{2} \le Q(1) + \delta$$

by (3.8). This completes the proof of (3.17).

Proof of Theorem 3.1. Again, since the proofs of part (i) and part (ii) are similar, we will prove just part (i). Applying Lemma 3.2 once to $f(x)$ and then to $-f(-x)$, we see that there exist two polynomials P_1 and P_2 of degree $\le r$ which are monotonically nondecreasing on $[-1,1]$. $P_1(x)$ interpolates f at 0 and 1 and $P_2(x)$ interpolates f at -1 and 0; and

$$\|f - P_j\|_\infty \le C\omega_r(f', 1), \qquad j = 1, 2. \tag{3.18}$$

Thus, we let

$$g(x) = \begin{cases} P_1(x), & 0 \le x \le 1, \\ P_2(x), & -1 \le x \le 0 \end{cases}$$

and g satisfies (3.1) and (3.2). In order to prove (3.3) observe that for all polynomials of degree $\le r$ on $[-1,1]$, the norm $|||\sum_{k=0}^r a_k x^k||| = \sum_{k=0}^r k!|a_k|$ is equivalent to the sup-norm. Hence, by (3.18),

$$\begin{aligned} \sum_{k=1}^r |g^{(k)}(0+) - g^{(k)}(0-)| &\equiv \left|\left|\left| \sum_{k=0}^r \frac{g^{(k)}(0+) - g^{(k)}(0-)}{k!} x^k \right|\right|\right| \\ &\le C \left\| \sum_{k=0}^r \frac{P_1^{(k)}(0) - P_2^{(k)}(0)}{k!} x^k \right\|_\infty \\ &= C\|P_1 - P_2\|_\infty \le C_1 \omega_r(f', 1). \end{aligned}$$

The proof of Theorem 3.1 is complete.

The following result enables us to patch together piecewise polynomials on different intervals.

THEOREM 3.2. (i) *Let f be continuously differentiable on $[-2,2]$ and nondecreasing there, and let g_1 and g_2 be the piecewise polynomials guaranteed by Theorem* 3.1(i) *for the intervals $[-2,0]$ and $[0,2]$, respectively. Then*

$$\sum_{k=1}^{r} |g_2^{(k)}(0+)-g_1^{(k)}(0-)| \le C\omega_r(f',1;[-2,2]). \tag{3.19}$$

(ii) *Let f be the second primitive of $f''\in L_p[-2,2]$, $1\le p<\infty$, and satisfying $f'\ge 0$ and let g_1 and g_2 as above. Then*

$$\sum_{k=1}^{r} |g_2^{(k)}(0+)-g_1^{(k)}(0-)| \le C\omega_{r-1,p}(f'',1). \tag{3.20}$$

Proof. Again we will prove part (i) only. By Theorem 3.1(i) there exists a function g on $[-1,1]$ the restrictions of which to $[-1,0]$ and to $[0,1]$ are polynomials of degree $\le r$ and such that

$$\|f-g\|_\infty[-1,1] \le C\omega_r(f',1;[-1,1]) \tag{3.21}$$

and

$$\sum_{k=1}^{r} |g^{(k)}(0+)-g^{(k)}(0-)| \le C\omega_r(f',1;[-1,1]). \tag{3.22}$$

Also by Theorem 3.1(i), g_1 and g_2 are polynomials of degree $\le r$ on $[-1,0]$ and $[0,1]$, respectively, and

$$\|f-g_1\|_\infty[-2,0] \le C\omega_r(f',1;[-2,0]), \tag{3.23}$$

$$\|f-g_2\|_\infty[0,2] \le C\omega_r(f',1;[0,2]). \tag{3.24}$$

By (3.21) and (3.23),

$$\|g-g_1\|_\infty[-1,0] \le C_1\omega_r(f',1,[-2,2]).$$

Similarly by (3.21) and (3.24),

$$\|g-g_2\|_\infty[0,1] \le C_1\omega_r(f',1;[-2,2]).$$

Applying Markov's inequality (e.g., see [4, p. 40]) $r-1$ times in each of the intervals, we get

$$\sum_{k=1}^{r} |g^{(k)}(0-)-g_1^{(k)}(0-)| \le C_r\omega_r(f',1;[-2,2])$$

and

$$\sum_{k=1}^{r} |g^{(k)}(0+)-g_2^{(k)}(0+)| \le C_r\omega_r(f',1;[-2,2]),$$

which together with (3.22) imply (3.19).

4. Monotone approximation by splines. Let $0=x_0<x_1<\cdots<x_N=1$ be a division of $[0,1]$ and denote $I_i=[x_{i-1},x_i]$ and $\delta_i=x_i-x_{i-1}$, $i=1,\cdots,N$. For a function $f\in L_p[0,1]$, we let $\omega_{rp}(f,\delta;I_i)$ be the modulus of smoothness of f restricted to I_i. If no interval is specified, then the norms and moduli of smoothness are on $[0,1]$.

LEMMA 4.1. *Let $f\in L_p[0,1]$, $1\le p<\infty$ and $0=x_0<x_1<\cdots<x_N=1$ be given. Then for $0<\eta_i\le\delta_i=x_i-x_{i-1}$ $i=1,\cdots,N$ and $\delta=\max_{1\le i\le N}\delta_i$,*

$$\sum_{i=1}^{N} \left[\omega_{rp}(f,\eta_i,I_i)\right]^p \le C\left[\omega_{rp}(f,\delta;[0,1])\right]^p. \tag{4.1}$$

Proof. Let $\varphi \in W_{rp}[0,1]$, i.e., φ is the rth interated primitive of $\varphi^{(r)} \in L_p[0,1]$, be arbitrary. Then it is readily seen (see [2, p. 122]) that

$$\begin{aligned}\|\Delta_t^r(f,x)\|_p(I_i(rt)) &\le \|\Delta_t^r(f-\varphi,x)\|_p(I_i(rt)) + \|\Delta_t^r(\varphi,x)\|_p(I_i(rt)) \\ &\le 2^r\left[\|f-\varphi\|_p(I_i) + t^r\|\varphi^{(r)}\|_p(I_i)\right].\end{aligned}$$

Hence

$$\omega_{rp}(f,\eta_i;I_i) \le 2^r\left[\|f-\varphi\|_p(I_i) + \delta^r\|\varphi^{(r)}\|_p(I_i)\right],$$

so that

$$\left[\omega_{rp}(f,\eta_i,I_i)\right]^p \le 2^{(r+1)p}\left[\|f-\varphi\|_p^p(I_i) + \delta^{rp}\|\varphi^{(r)}\|_p^p(I_i)\right].$$

Therefore

$$\begin{aligned}\sum_{i=1}^N \left[\omega_{rp}(f,\eta_i,I_i)\right]^p &\le 2^{(r+1)p}\left[\|f-\varphi\|_p^p + \delta^{rp}\|\varphi^{(r)}\|_p^p\right] \\ &\le 2^{(r+1)p}\left[\|f-\varphi\|_p + \delta^r\|\varphi^{(r)}\|_p\right]^p.\end{aligned}$$

Since $\varphi \in W_{rp}[0,1]$ is arbitrary, the last inequality also holds for the K-functional,

$$K_{rp}(f,h) = \inf_{\varphi \in W_{rp}[0,1]}\left[\|f-\varphi\|_p + h\|\varphi^{(r)}\|_p\right];$$

namely we have

$$\sum_{i=1}^N \left[\omega_{rp}(f,\eta_i,I_i)\right]^p \le 2^{(r+1)p}\left[K_{rp}(f,\delta^r)\right]^p.$$

By virtue of [2, Thm. 2.1], this K-functional is equivalent to $\omega_{rp}(f,\delta)$. In other words we have proved (4.1).

The last lemma we will need is an important result of Chui, Smith and Ward [1, Lemma 4.3].

LEMMA CSW. *Let $r \ge 2$ and $d = 4r^2$ and let g be a nondecreasing continuous function on $[-3d,3d]$, the restrictions of which to $[-3d,0]$ and to $[0,3d]$ are polynomials of degree $\le r-1$. Then there is a nondecreasing spline s of order r and knots at the integers such that $s = g$ outside of $[-d,d]$ and*

$$\|s-g\|_p[-d,d] \le C\sum_{k=1}^{r-1}|g^{(k)}(0+) - g^{(k)}(0-)|. \tag{4.2}$$

We are now ready to prove Theorem 2.1. The proof of Theorem 2.2 is similar. The proof runs along the lines of that of [1, Thm. 1.1]. We bring it here for the sake of completeness and also in order to correct some typographical error in the definition of the spline in the above mentioned proof.

Proof of Theorem 2.1. It suffices to prove (2.1) for $n > 6d$, where $d = 4r^2$. Let $F(t) = f(\frac{t}{n})$, $0 \le t \le n$, and let m be the greatest even integer not exceeding $n/3d$. Denote $I_1 = [0,3d]$, $I_2 = [3d,6d], \cdots, I_{m-1} = [3(m-2)d, 3(m-1)d]$ and $I_m = [3(m-1)d, n]$. By Theorem 3.1(i) for each pair of intervals $I_{2j-1} \cup I_{2j}$, $j = 1,2,\cdots,m/2$, there exists a monotonically nondecreasing continuous function G_j interpolating F at $6(j-1)d$, $(6j-3)d$ and $6jd$, such that G_j is a polynomial of degree $\le r-1$ on I_{2j-1} and on I_{2j}. Also,

$$\|F-G_j\|_\infty(I_{2j-1} \cup I_{2j}) \le C\omega_{r-1}(F',1;I_{2j-1} \cup I_{2j})$$

and

$$\sum_{k=1}^{r-1} |G_j^{(k)}((6j-3)d+) - G_j^{(k)}((6j-3)d-)| \le C\omega_{r-1}(F', 1; I_{2j-1} \cup I_{2j}).$$

Note that since the length of I_i is ≥ 1, C is independent of the intervals. Now by Theorem 3.2(i), we may define a continuous nondecreasing function $G = G_j$ on $I_{2j-1} \cup I_{2j}, j = 1, \cdots, m/2$ such that

$$\|F - G\|_\infty[0, n] \le C\omega_{r-1}(F', 1) \tag{4.3}$$

and for $i = 1, 2, \cdots, m-1$

$$\sum_{k=1}^{r-1} |G^{(k)}(3id+) - G^{(k)}(3id-)| \le C\omega_{r-1}(F', 1). \tag{4.4}$$

On each pair $I_i \cup I_{i+1}$, G satisfies the conditions of Chui, Smith and Ward's lemma whence there is a spline S_i on $I_i \cup I_{i+1}$ such that $S_i = G$ outside $[(3i-1)d, (3i+1)d]$ and by (4.2) and (4.4).

$$\|S_i - G\|_\infty[(3i-1)d, (3i+1)d] \le C\omega_{r-1}(F', 1), \qquad i = 1, \cdots, m-1. \tag{4.5}$$

So we define the spline $S = S_i$ on $[3(i-\frac{1}{2})d, 3(i+\frac{1}{2})d]$ $i = 1, \cdots, m-2$ $S = S_0$ on $[0, \frac{3}{2}d]$ and $S = S_{m-1}$ on $[3(m-\frac{3}{2})d, n]$. Now we let $s(x) = S(nx)$, $0 \le x \le 1$. Then $s \in \mathcal{S}(r, n)$, s is monotonically nondecreasing and by virtue of (4.3) and (4.5)

$$\begin{aligned} \|s - f\|_\infty &= \|S - F\|_\infty[0, n] \le \|S - G\|_\infty[0, n] + \|G - F\|_\infty[0, n] \\ &= \sup_i \|S_i - G\|_\infty[(3i-1)d, (3i+1)d] + C\omega_{r-1}(F', 1) \\ &\le 2C\omega_{r-1}(F', 1) = C_1 n^{-1} \omega_{r-1}(f', n^{-1}). \end{aligned} \tag{4.6}$$

This completes the proof of Theorem 2.1.

Proof of Theorem 2.2. The only difference between this proof and the previous one is in the proof of (4.6). Here we have

$$\|s - f\|_p^p = \frac{1}{n}\|S - F\|_p^p[0, n]$$

$$\le \frac{2^p}{n}\left[\|S - G\|_p^p[0, n] + \|G - F\|_p^p[0, n]\right]$$

(similar to (4.3))

$$\le \frac{2^p}{n} \sum_{i=1}^{m-1} \|S_i - G\|_p^p[(3i-1)d, (3i+1)d] + \frac{2^p}{n} C\left[\omega_{r-2, p}(F'', 1)\right]^p$$

(similar to (4.5))

$$\le \frac{2^p}{n} C \sum_{i=1}^{m-1} \left[\omega_{r-2, p}(F'', 1; I_{i-1} \cup I_i \cup I_{i+1} \cup I_{i+2})\right]^p$$

$$+ \frac{2^p}{n} C\left[\omega_{r-2, p}(F'', 1)\right]^p.$$

Applying Lemma 4.1, we have

$$\leq \frac{4\cdot 2^p}{n} C\left[\omega_{r-2,p}(F'',1)\right]^p + \frac{2^p}{n} C\left[\omega_{r-2,p}(F'',1)\right]^p$$

$$= \frac{C_1}{n}\left[\omega_{r-2,p}(F'',1)\right]^p = C\left[n^{-2}\omega_{r-2,p}(f'',n^{-1})\right]^p,$$

and the proof is complete.

REFERENCES

[1] C. K. CHUI, P. W. SMITH AND J. D. WARD, *Degree of L_p approximation by monotone splines*, this Journal, 11 (1980), pp. 436–447.

[2] R. A. DE VORE, *Degree of approximation*, Approx. Theory II, G. G. Lorentz, C. K. Chui and L. L. Schumaker, eds., Academic Press, New York, 1976, pp. 117–162.

[3] _____, *Monotone approximation by splines*, this Journal, 8 (1977), pp. 891–905.

[4] G. G. LORENTZ, *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1966.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0014 $01.00/0

THE ZEROS OF CERTAIN JACOBI POLYNOMIALS*

ANDREW YOUNG† AND HASSAN HAMIDEH‡

Abstract. Two theorems are proved about the zeros of certain Jacobi polynomials that are important in the theory of interpolation and approximation.

THEOREM 1. *Let S_k and $\bar{S}_k$ be the sums of the kth powers of the zeros of $P_n^{(w,-w)}(x)$ and $P_n^{(w,-w)}(-x)$ respectively (w real, $0<w<1$). Then for $k=1,2,\cdots,2n$, $S_k-\bar{S}_k=-2w$ (k odd) or $=0$ (k even).*

THEOREM 2. *Let S_k and $\bar{S}_k$ be the sums of the kth powers of $P_{n+1}^{(-w,w-1)}(x)$ and $P_n^{(w,1-w)}(x)$ respectively (w real, $0<w<1$). Then for $k=1,2,\cdots,2n+1$, $S_k-\bar{S}_k=2w-1$ (k odd) or $=1$ (k even).*

1. Young and Kiountouzis (1979) have shown that if a polynomial of degree $m-1$,

$$F(x)=\sum_{i=0}^{m-1} a_i x^i,$$

is the best approximation to a function $f(x)$ in the weighted L_1-measure over the range $-1\le x\le 1$, i.e.,

$$\int_{-1}^{1} w(x)|f(x)-F(x)|\,dx$$

is a minimum where

$$w(x)=\begin{cases} w & \text{when } f>F, \\ 1-w & \text{when } f<F, \end{cases}$$

and $0<w<1$, then under certain conditions on $f(x)$, $f(x)=F(x)$ at m points $-1<\xi_1<\xi_2<\cdots<\xi_m<1$, independent of $f(x)$, satisfying the relations

$$\xi_1^k-\xi_2^k+\xi_3^k-\cdots(-1)^{m-1}\xi_m^k+w_k=0, \qquad k=1,2,\cdots,m, \tag{1}$$

with

$$w_k=\begin{cases} 2w, & m \text{ even}, \quad k \text{ odd}, \\ 0, & m \text{ even}, \quad k \text{ even}, \end{cases} \tag{2}$$

and

$$w_k=\begin{cases} 1-2w, & m \text{ odd}, \quad k \text{ odd}, \\ -1, & m \text{ odd}, \quad k \text{ even}. \end{cases} \tag{3}$$

They identify, without proof, the set $\{\xi_i\}$ as being the zeros of polynomials $p_m(x)$ as follows:

$$m \text{ even } (=2n), \qquad p_m=P_n^{(w,-w)}(x)P_n^{(w,-w)}(-x), \tag{4}$$

$$m \text{ odd } (=2n+1), \qquad p_m=P_{n+1}^{(-w,w-1)}P_n^{(w,1-w)}(x), \tag{5}$$

where $P_n^{(\alpha,\beta)}(x)$ is a Jacobi polynomial of degree n.

The purpose of this note is to provide proofs for these identifications.

* Received by the editors March 16, 1981.
† Department of Mathematics, New University of Ulster, Coleraine, Northern Ireland.
‡ Department of Mathematics, The University of Riyadh, Riyadh, Saudi Arabia.

2. The following properties of Jacobi polynomials are needed. (See, e.g., Szegö (1978, Chapt. IV).)

Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$, $n=0,1,2,\cdots,\alpha>-1$, $\beta>-1$ are orthogonal with respect to the weight function $w(x)=(1-x)^{\alpha}(1+x)^{\beta}$ on $[-1,1]$. Their zeros are all real and distinct and have modulus less than 1.

$P_n^{(\alpha,\beta)}(x)$ is a solution of the differential equation

$$(1-x^2)y^{11}-(A+Bx)y^1+Cy=0, \tag{6}$$

$$A=\alpha-\beta,\quad B=\alpha+\beta+2,\quad C=n(n+\alpha+\beta+1), \tag{7}$$

$$P_n^{(\alpha,\beta)}(x)=(-1)^n P_n^{(\beta,\alpha)}(-x), \tag{8}$$

$$\frac{d}{dx}\{P_n^{(\alpha,\beta)}(x)\}=\frac{1}{2}(n+\alpha+\beta+1)P_{n-1}^{(\alpha+1,\beta+1)}(x). \tag{9}$$

The coefficient of x^n in $P_n^{(\alpha,\beta)}(x)$ is

$$2^{-n}\binom{2n+\alpha+\beta}{n}. \tag{10}$$

3. Any polynomial of degree n can be written as the product of factors

$$y=\text{const}\cdot\prod_{i=1}^{n}(x-x_i).$$

Differentiating this gives

$$y^1=y\sum_{i=1}^{n}(x-x_i)^{-1} \tag{11}$$

and

$$y^{11}=y^1\sum_{i=1}^{n}(x-x_i)^{-1}-y\sum_{i=1}^{n}(x-x_i)^{-2}. \tag{12}$$

When y satisfies (6), substituting (11) and (12) into (6) yields

(13)

$$x^{-1}\sum_i\left(1-\frac{x_i}{x}\right)^{-1}(A+Bx)+(1-x^{-2})\left\{\sum_i\left(1-\frac{x_i}{x}\right)^{-1}\right\}^2=C+(1-x^{-2})\sum_i\left(1-\frac{x_i}{x}\right)^{-2}.$$

Let the roots be ordered $-1<x_1<x_2<\cdots<x_n<1$. Then for $x_n<x<1$ each of the terms can be expanded in power series of x_i/x. If we denote

$$S_r=\sum_{i=1}^{n}x_i^r, \tag{14}$$

then (12) gives

$$\sum_{r=0}^{\infty}\frac{S_r}{x^r}\left[B+Ax^{-1}+(1-x^{-2})\sum_{r=0}^{\infty}\frac{S_r}{x^r}\right]=C+(1-x^{-2})\sum_{r=0}^{\infty}\frac{(r+1)S_r}{x^r}. \tag{15}$$

This is true for all x in the range $x_n<x<1$ so that it is legitimate to equate coefficients of x^{-k} to obtain a set of relationships for the sums S_r. This gives

$$k=0,\qquad S_0(B+S_0)=C+S_0,$$

which is readily confirmed since $S_0=n$, $B=\alpha+\beta+2$ and $C=n(n+\alpha+\beta+1)$,

$$k=1, \qquad Q_1+AS_0+(B-2)S_1=0, \tag{16}$$

$$k>1, \qquad Q_k-Q_{k-2}+AS_{k-1}+(B-k-1)S_k+(k-1)S_{k-2}, \tag{17}$$

where

$$Q_k=\sum_{r=0}^{k} S_rS_{k-r} \quad \text{with } Q_0=S_0^2. \tag{18}$$

4. The first result to be proved (equations (1) and (2)) can be formulated as:

THEOREM 1. *Let S_k and $\bar{S}_k$ be the sums of the kth powers of the zeros of $P_n^{(w,-w)}(x)$ and $P_n^{(w,-w)}(-x)$ respectively (real w, $0<w<1$). Then for $k=1,2,\cdots,2n$*

$$S_k-\bar{S}_k=\begin{cases}-2w, & k \text{ odd},\\ 0, & k \text{ even}.\end{cases}$$

Proof. Provided w is real and $-1<w<1$, $P_n^{(w,-w)}(x)$ is a Jacobi polynomial with the properties given in §2, so the results of §3 apply with $\alpha=w>-1$ and $\beta=-w>-1$. If the zeros of $P_n^{(w,-w)}(x)$, in ascending order, are taken as $\xi_1,\xi_3,\cdots,\xi_{2n-1}$, those of $P_n^{(w,-w)}(-x)$ are $\xi_2,\xi_4,\cdots,\xi_{2n}$ with $\xi_i=-\xi_{2n+1-i}$, $i=1,2,\cdots,2n$. Thus $S_k=(-1)^k\bar{S}_k$.

For k even, $S_k-\bar{S}_k=0$ as required. For k odd, $S_k-\bar{S}_k=2S_k$ so it has to be proved that

$$S_k=-w \quad \text{for } k=1,3,\cdots,2n-1.$$

In (6) to (18) put $\alpha=-\beta=w$, $A=2w$, $B=2$, $C=n(n+1)$. Equation (16) gives $S_1=-w$. With $k=3$, (17) then gives $S_3=-w$.

Now assume $S_1=S_3=\cdots=S_{k-2}=-w$ (k odd). In this case Q_k-Q_{k-2} in (17) reduces to $2S_0(S_k-S_{k-2})-2wS_{k-1}$ so (17) becomes $2n(S_k-S_{k-2})-2wS_{k-1}+2wS_{k-1}+(1-k)S_k+(k-1)S_{k-2}=0$ or $(2n+1-k)(S_k-S_{k-2})=0$. Hence for $k<2n+1$ $S_k=S_{k-2}=-w$. Since $S_1=S_3=-w$, $S_k=-w$ for $k<2n+1$, k odd, by induction. This proves the theorem. Note that when $k=2n+1$ the induction process breaks down.

5. The other result to be proved is (equations (1) and (3)):

THEOREM 2. *Let S_k and $\bar{S}_k$ be the sums of the kth powers of the zeros of $P_{n+1}^{(-w,w-1)}(x)$ and $P_n^{(w,1-w)}(x)$ respectively (real w, $0<w<1$). Then for $k=1,2,\cdots,2n+1$,*

$$S_k-\bar{S}_k=\begin{cases}2w-1, & k \text{ odd},\\ 1, & k \text{ even}.\end{cases}$$

Proof. Provided w is real and $0<w<1$, $P_n^{(w,1-w)}(x)$ and $P_{n+1}^{(-w,w-1)}(x)$ are both Jacobi polynomials so the results of §2 and §3 apply with $\alpha=w>-1$ and $\beta=1-w>-1$ in the first case and $\alpha=-w>-1$ and $\beta=w-1>-1$ in the second. From (8) and (9) $P_n^{(w,1-w)}(x)=(-1)^nP_n^{(1-w,w)}(-x)$ and $P_n^{(1-w,w)}(x)=\frac{d}{dx}\{P_{n+1}^{(-w,w-1)}(x)\}$. Moreover the zeros of both polynomials are all real and inside the range $(-1,1)$. Thus the zeros of $P_n^{(w,1-w)}(x)$ interlace those of $P_{n+1}^{(-w,w-1)}(x)$ so if in ascending order the zeros of the latter are $\xi_1,\xi_3,\cdots,\xi_{2n+1}$ those of the former are $\xi_2,\xi_4,\cdots,\xi_{2n}$ with $\xi_{2i-1}<\xi_2$, $i=1,\cdots,n$.

Let $Q_k=\Sigma_{r=0}^k S_rS_{k-r}$ and $\bar{Q}_k=\Sigma_{r=0}^k \bar{S}_r\bar{S}_{k-r}$. For S_k the substitutions required in (6) to (18) are $\alpha=-w$, $\beta=w-1$, $A=1-2w$, $B=1$, $C=(n+1)^2$, $S_0=n+1$ while for $\bar{S}_k$ they are $\alpha=w$, $\beta=1-w$, $A=2w-1$, $B=3$, $C=n(n+2)$, $\bar{S}_0=n$.

With $k=1$, (16) gives $Q_1-\bar{Q}_1+(1-2w)S_0-(2w-1)\bar{S}_0-S_1-\bar{S}_1=0$, which reduces to

$$(2n+1)\big(S_1-\bar{S}_1+1-2w\big)=0,$$

giving $S_1-\bar{S}_1=2w-1$. With $k=2$, (17) gives $Q_2-\bar{Q}_2-Q_0+\bar{Q}_0+(1-2w)(S_1+\bar{S}_1)-2S_2+S_0-\bar{S}_0=0$, which reduces to

$$2n(S_2-\bar{S}_2)+(S_1+\bar{S}_1)(S_1-\bar{S}_1+1-2w)-2n=0,$$

so using the result for $k=1$, $S_2-\bar{S}_2=1$ unless $2n=0$. If $2n$ is zero, the theorem states only $S_1-\bar{S}_1=2w-1$. So the theorem is proved in the case $2n=0$.

For $2n>0$ the theorem can now be proved by induction. Assume $S_r-\bar{S}_r=-w_r$ for $r=1,2,\cdots,k-1$ with $w_1=w_3=\cdots=1-2w$ and $w_2=w_4=\cdots=-1$. Then, for general k, (17) gives

$$(19)\qquad \begin{aligned}(Q_k-\bar{Q}_k)-(Q_{k-2}-\bar{Q}_{k-2})+w_1(S_{k-1}+\bar{S}_{k-1})\\ -k(S_k-\bar{S}_k)-2\bar{S}_k+(k-1)(S_{k-2}-\bar{S}_{k-2})=0.\end{aligned}$$

Now $Q_k-\bar{Q}_k=\Sigma_{r=0}^{k}S_rS_{k-r}-\bar{S}_r\bar{S}_{k-r}$ and, by assumption, $S_r=\bar{S}_r+w_r$. So $Q_k-\bar{Q}_k$ reduces to

$$2(n+1)S_k-2n\bar{S}_k+\sum_{r=1}^{k-1}\left(w_rw_{k-r}-w_r\bar{S}_{k-r}-w_{k-r}\bar{S}_r\right).$$

Similarly

$$Q_{k-2}-\bar{Q}_{k-2}=2(n+1)S_{k-2}-2n\bar{S}_{k-2}+\sum_{r=1}^{k-3}\left(w_rw_{k-r-2}-w_r\bar{S}_{k-r-2}-w_{k-r-2}\bar{S}_r\right).$$

Substituting into (19) gives

$$(20)\qquad \begin{aligned}(2n+2-k)(S_k-\bar{S}_k)+w_1(S_{k-1}+\bar{S}_{k-1})-(2n+1-k)(S_{k-2}-\bar{S}_{k-2})\\ -2S_{k-2}+w_1w_{k-1}+w_2w_{k-2}-2w_1\bar{S}_{k-1}-2w_2\bar{S}_{k-2}=0.\end{aligned}$$

When k is odd (20) gives ($w_{k-1}=w_2$, $w_{k-2}=w_1$)

$$(2n+2-k)(S_k-\bar{S}_k)+w_1(2n+2-k)=0.$$

So provided $k<2n+2$, $S_k-\bar{S}_k=-w_1$ (k odd). On the other hand when k is even ($w_{k-1}=w_1$, $w_{k-2}=w_2$) and (20) becomes

$$(2n+2-k)(S_k-\bar{S}_k)+w_2(2n+2-k)=0$$

and provided $k<2n+2$, $S_k-\bar{S}_k=-w_2$ (k even). Since the result is true for $k=1,2$ it is true for all $k<2n+2$ thus proving the theorem. Again the induction process breaks down at $k=2n+2$.

REFERENCES

[1] G. Szegö (1978), *Orthogonal Polynomials*, 4th ed., AMS Colloquium Publications 23, American Mathematical Society, Providence, RI.

[2] A. Young and E. A. Kiountouzis (1979), *Best approximation in an unsymmetrically weighted L_1 measure*, J. Inst. Maths. Applics., 24, pp. 379–394.

SIAM J. MATH. ANAL.
Vol. 13, No. 5, September 1982

0036-1410/82/1305-0015 $01.00/0

COMBINATORIAL APPLICATIONS OF HERMITE POLYNOMIALS*

RUTH AZOR†, J. GILLIS† AND J. D. VICTOR‡

Abstract. Let $C_1, C_2, \cdots, C_k$ be k finite sets of elements, where n_i is the number of elements in C_i $(i=1,2,\cdots,k)$ and $\Sigma_{i=1}^{k} n_i$ is even, $2S$ (say). In any arrangement of the elements into S disjoint pairs, we count the number of homogeneous pairs, i.e., those in which both numbers are from the same subset, C_i. We define such a pairing as *even*, *odd* or *pure* according as the number of homogeneous pairs is *even*, *odd* or *zero* respectively. The numbers of possible pairings of the different types are expressed as integrals involving Hermite polynomials, and these expressions are used both to derive new combinatorial results and also to provide combinatorial proofs of analytical formulae.

1. Introduction. Consider an even set of distinguishable elements divided into k subsets, with n_i in the ith subset $(i=1,2,\cdots,k)$, where $\Sigma_{i=1}^{k} n_i = 2S$.

We define a *match* on the $2S$ elements as an arrangement of them into S pairs, i.e., a graph of degree 1, taking these elements as nodes. The edges of such a graph are classed as homogeneous or heterogeneous, according as the two nodes which they join belong to the same or different subsets.

Denote by $E_{n_1,\cdots,n_k}$ $(\Omega_{n_1,\cdots,n_k})$ the number of matches which contain an even (odd) number of homogeneous edges, and by $P_{n_1,\cdots,n_k}$ the number of *pure* matches, i.e., which contain no homogeneous edges at all. These last, are in fact, the k-partite graphs of degree 1 which can be formed on the k subsets.

We shall derive explicit formulae for these numbers in terms of Hermite polynomials, and shall show how they can be used to obtain combinatorial proofs of some analytic formulae and also analytic proofs of some new combinatorial results.

2. Properties of Hermite polynomials. The Hermite polynomials are defined, as usual, by

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2}) = n! \sum_{m \le n/2} \frac{(-1)^m (2x)^{n-2m}}{m!(n-2m)!}, \tag{1}$$

and it is well known that

$$\int_{-\infty}^{\infty} H_n(x) H_s(x) e^{-x^2} dx = \begin{cases} 0 & \text{if } r \neq s, \\ \sqrt{\pi} \cdot 2^r \cdot r! & \text{if } r = s. \end{cases} \tag{2}$$

Let

$$I_{n_1,n_2,\cdots,n_k} = \int_{-\infty}^{\infty} e^{-x^2} \prod_{i=1}^{k} \{H_{n_i}(x)\}\, dx, \tag{3}$$

where the subscripts $n_1, n_2, \cdots, n_k$ are any natural numbers.

We assemble here, for convenience, some basic properties of the Hermite polynomials which will be needed:

$$\frac{d}{dx}[H_n(x)] = 2nH_{n-1}(x), \tag{4}$$

* Received by the editors September 8, 1980.
† Weizmann Institute of Science, Rehovot, Israel.
‡ Rockefeller University, New York, New York 10021.

$$\sum_{n=0}^{\infty} \frac{z^n}{n!} H_n(x) = \exp\{2zx - z^2\}, \tag{5}$$

$$H_n(x+y) = \sum_{\alpha=0}^{n} \binom{n}{\alpha} H_\alpha(x)(2y)^{n-\alpha}, \tag{6}$$

$$H_n(\lambda x) = \lambda^n \cdot n! \sum_{\beta \le n/2} \left(\frac{\lambda^2 - 1}{\lambda^2} \right)^\beta \{\beta!(n-2\beta)!\}^{-1} H_{n-2\beta}(x). \tag{7}$$

Of these (4) and (5) are well known, while (6) is an immediate consequence of (5) (cf. [5, p. 385]). Equation (7) can be deduced from a corresponding result for Laguerre polynomials. However, since it does not appear to have been stated explicitly in the literature, we note here that

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{z^n}{n!} H_n(\lambda x) &= \exp\{2\lambda zx - z^2\} \quad \text{by (5)} \\ &= \exp\{2x \cdot \lambda z - \lambda^2 z^2\} \exp(\lambda^2 - 1) z^2 \\ &= \sum_{\alpha=0}^{\infty} \frac{(\lambda z)^\alpha}{\alpha!} H_\alpha(x) \sum_{\beta=0}^{\infty} \frac{(\lambda^2 - 1)^\beta}{\beta!} z^{2\beta}. \end{aligned} \tag{8}$$

Comparison of coefficients of z^n in (8) yields (7). We shall be particularly interested in the special case $\lambda = 2^{-1/2}$, giving

$$H_n\left(\frac{x}{\sqrt{2}} \right) = n! 2^{-n/2} \sum_{\beta \le n/2} \frac{(-1)^\beta}{\beta!(n-2\beta)!} H_{n-2\beta}(x). \tag{9}$$

3. Combinatorial interpretation.

THEOREM 1. *For any set of nonnegative integers* $(n_1, n_2, \cdots, n_k)$,

$$P_{n_1, n_2, \cdots, n_k} = \{2^{\sum_{i=1}^k n_i} \cdot \pi\}^{-1/2} \cdot I_{n_1, n_2, \cdots, n_k}, \tag{10}$$

where $P_{n_1, \cdots, n_k}$ *is defined as in* §1 *and* $I_{n_1, \cdots, n_k}$ *in* (3).

Proof. We begin by defining

$$P_{0, \cdots, 0} = 1 \tag{11}$$

for any k, so that (10) holds for the special case $n_1 = n_2 = , \cdots, = n_k = 0$. The general result will follow if we can show that $P_{n_1, \cdots, n_k}$ and the right-hand side of (10) obey identical recurrence relations.

To simplify the notation we shall, where there is no risk of ambiguity, denote any array, $A_{n_1, n_2, \cdots, n_k}$ (say), by $A_{\mathbf{n}}$, where $\mathbf{n}$ stands for the index set $(n_1, n_2, \cdots, n_k)$. If $\mathbf{m}$ is the set $(m_1, m_2, \cdots, m_l)$, then $A_{\mathbf{n},\mathbf{m}}$ represents $A_{n_1, n_2, \cdots, n_k, m_1, \cdots, m_l}$. Similarly, if b is a simple index $A_{\mathbf{n},b}$ is $A_{n_1, n_2, \cdots, n_k, b}$, etc. We shall use $A_{\mathbf{n}}^{(i)}$ to represent $A_{n_1, n_2, \cdots, n_{i-1}, n_i - 1, n_{i+1}, \cdots, n_k}$, i.e., the term of the array $A_{\mathbf{n}}$, obtained if one subtracts 1 from the ith subscript, leaving the others unchanged. Similarly, $A_{\mathbf{n}}^{(i,j)}$ is the term arrived at if we subtract 1 from each of the ith and jth subscripts, etc. It will also be convenient, for any index set $\mathbf{n} = (n_1, \cdots, n_k)$ to write $\sigma(\mathbf{n}) = \sum_{i=1}^k n_i$.

Consider now the $P_{\mathbf{n}}$ possible pure matches on our total set of $\sigma(\mathbf{n})$ objects. Select some particular member of the first subset. The number of pure matches in which it is

matched with an element of the ith subset is clearly $n_i P_{\mathbf{n}}^{(1,i)}$, and hence

$$P_{\mathbf{n}} = \sum_{i=2}^{k} n_i P_{\mathbf{n}}^{(1,i)}. \tag{12}$$

Moreover,

$$\begin{aligned}
I_{\mathbf{n}} &= \int_{-\infty}^{\infty} e^{-x^2} H_{n_1}(x) \cdot \prod_{i=2}^{k} \left[H_{n_i}(x) \right] dx \qquad (13)\\
&= (-1)^{n_1} \int_{-\infty}^{\infty} \frac{d^{n_1}}{dx^{n_1}} (e^{-x^2}) \cdot \prod_{i=2}^{k} \left[H_{n_i}(x) \right] dx \quad \text{by (1)}\\
&= (-1)^{n_1 - 1} \int_{-\infty}^{\infty} \frac{d^{n_1 - 1}}{dx^{n_1 - 1}} (e^{-x^2}) \cdot \frac{d}{dx} \left\{ \prod_{i=2}^{k} H_{n_i}(x) \right\} dx \quad \text{integrating by parts}\\
&= \int_{-\infty}^{\infty} e^{-x^2} H_{n_1 - 1}(x) \sum_{i=2}^{k} \left\{ H'_{n_i}(x) \cdot \prod_{\substack{j=2 \\ j \neq i}}^{k} H_{n_j}(x) \right\} dx\\
&= 2 \int_{-\infty}^{\infty} e^{-x^2} H_{n_1 - 1}(x) \sum_{i=2}^{k} \left\{ n_i H_{n_i - 1}(x) \prod_{\substack{j=2 \\ j \neq i}}^{k} H_{n_j}(x) \right\} dx \quad \text{by (4)}\\
&= 2 \sum_{i=2}^{k} n_i I_{\mathbf{n}}^{(1,i)}.
\end{aligned}$$

Hence, if $X_{\mathbf{n}}$ denotes the right-hand side of (10),

$$X_{\mathbf{n}} = \{\pi \cdot 2^{\sigma(\mathbf{n})}\}^{-1/2} I_{\mathbf{n}} = \{\pi \cdot 2^{\sigma(\mathbf{n})-2}\}^{-1/2} \sum_{i=2}^{k} n_i I_{\mathbf{n}}^{(1,i)} = \sum n_i X_{\mathbf{n}}^{(1,i)}. \tag{14}$$

Since $P_{0,\cdots,0} = X_{0,\cdots,0} = 1$, the identity between relations (12) and (14) establishes (10).

4. Linearization of products of Hermite polynomials. Let $\mathbf{n} = (n_1, n_2, \cdots, n_k)$. The product $\prod_{i=1}^{k} H_{n_i}(x)$ is a polynomial of degree $\sigma(\mathbf{n})$ and therefore expressible in the form

$$\sum C_{\mathbf{n},\alpha} H_{\alpha}(x),$$

where the coefficients $C_{\mathbf{n},\alpha}$ are to be determined. We have omitted limits of summation here, since it is in any case clear that $C_{\mathbf{n},\alpha} = 0$ unless $0 \leq \alpha \leq \sigma(\mathbf{n})$. We shall adopt a similar practice throughout the rest of the paper. Wherever summations or products are given without specification of their range, this may be taken as $(-\infty, \infty)$. In all such cases, the actual ranges are finite with natural cut-offs, as here.

It follows from (2) that

$$\begin{aligned}
\sqrt{\pi} \cdot 2^{\alpha} \alpha! C_{\mathbf{n},\alpha} &= \int_{-\infty}^{\infty} e^{-x^2} H_{\alpha}(x) \cdot \prod_{i=1}^{k} \left\{ H_{x_i}(x) \right\} dx\\
&= I_{\mathbf{n},\alpha} = \{2^{\sigma(\mathbf{n}) + \alpha} \cdot \pi\}^{1/2} P_{\mathbf{n},\alpha} \quad \text{by (10)},
\end{aligned}$$

i.e.,

$$C_{\mathbf{n},\alpha} = \frac{1}{\alpha!} 2^{(\sigma(\mathbf{n}) - \alpha)/2} P_{\mathbf{n},\alpha}. \tag{15}$$

Equation (15) can be used to prove some combinatorial relations for the $P_{\mathbf{n}}$. We give one example. Let $\mathbf{n}=(n_1,n_2,\cdots,n_k)$, $\mathbf{m}=(m_1,m_2,\cdots,m_l)$. Then

$$\sum_{\alpha=0}^{\sigma(\mathbf{m})+\sigma(\mathbf{n})} C_{\mathbf{m},\mathbf{n},\alpha}H_\alpha(x)=\prod_{i=1}^{k} H_{n_i}(x)\cdot\prod_{j=1}^{l} H_{m_j}(x)=\sum_{\beta} C_{\mathbf{n},\beta}H_\beta(x)\sum_{\gamma} C_{\mathbf{m},\gamma}H_\gamma(x)$$

$$=\sum_{\beta,\gamma} C_{\mathbf{n},\beta}C_{\mathbf{m},\gamma}H_\beta(x)H_\gamma(x) \tag{16}$$

$$=\sum_{\alpha,\beta,\gamma} C_{\mathbf{n},\beta}C_{\mathbf{m},\gamma}C_{\beta,\gamma,\alpha}H_\alpha(x),$$

and so

$$C_{\mathbf{m},\mathbf{n},\alpha}=\sum_{\beta,\gamma} C_{\mathbf{n},\beta}C_{\mathbf{m},\gamma}C_{\beta,\gamma,\alpha}. \tag{17}$$

Hence, by (15),

$$\frac{1}{\alpha!}2^{(\sigma(\mathbf{m})+\sigma(\mathbf{n})-\alpha)/2}P_{\mathbf{m},\mathbf{n},\alpha}$$

$$=\sum_{\beta,\gamma}\left\{\frac{1}{\beta!}2^{(\sigma(\mathbf{n})-\beta)/2}P_{\mathbf{n},\beta}\right\}\left\{\frac{1}{\gamma!}2^{(\sigma(\mathbf{m})-\gamma)/2}P_{\mathbf{m},\gamma}\right\}\left\{\frac{1}{\alpha!}2^{(\beta+\gamma-\alpha)/2}P_{\beta,\gamma,\alpha}\right\},$$

i.e.,

$$P_{\mathbf{m},\mathbf{n},\alpha}=\sum_{\beta,\gamma}\frac{1}{\beta!\gamma!}P_{\mathbf{n},\beta}P_{\mathbf{m},\gamma}P_{\beta,\gamma,\alpha}. \tag{18}$$

Now it is clear that $P_{\beta,\gamma,0}=\gamma!\delta_{\beta\gamma}$ and hence, putting $\alpha=0$ in (18), we get

$$P_{\mathbf{m},\mathbf{n}}=\sum_{\beta}\frac{1}{\beta!}P_{\mathbf{n},\beta}P_{\mathbf{m},\beta}. \tag{19}$$

We leave to the reader the combinatorial interpretation of (19).

Some applications. We give here some examples of applications of the above ideas to the evaluation of integrals. Consider first $P_{a,b,c}$. This is clearly zero for odd $a+b+c$ and also if any one of the three indices is greater than the sum of the other two. In any other case we can write $a+b+c=2S$, where S is an integer, $S\geq\max(a,b,c)$. Moreover, it is an elementary exercise in combinatorics to show that

$$P_{a,b,c}=\frac{a!b!c!}{(S-a)!(S-b)!(S-c)!}. \tag{20}$$

It follows by (10) that

$$\int_{-\infty}^{\infty}H_a(x)H_b(x)H_c(x)e^{-x^2}dx=2^s\sqrt{\pi}\,\frac{a!b!c!}{(S-a)!(S-b)!(S-c)!}. \tag{21}$$

Again, by (19),

(22)

$$P_{a,b,c,d}=\sum_{\beta}\frac{1}{\beta!}P_{a,b,\beta}P_{c,d,\beta}$$

$$=\sum_{\beta}\frac{a!b!c!d!\beta!}{\left(\frac{b+\beta-a}{2}\right)!\left(\frac{a+b-\beta}{2}\right)!\left(\frac{a+\beta-b}{2}\right)!\left(\frac{c+d-\beta}{2}\right)!\left(\frac{c-d+\beta}{2}\right)!\left(\frac{d+\beta-c}{2}\right)!},$$

where the summation is over these values of β, if any, for which the arguments of all the factorials are nonnegative integers.

In the special case $b=a$, $d=c$ (22) becomes

$$P_{a,a,c,c}=\sum_{\beta}\frac{(a!c!)^2\beta!}{\left[\left(\frac{\beta}{2}\right)!\right]^4\left(a-\frac{\beta}{2}\right)!\left(c-\frac{\beta}{2}\right)!}=(a!c!)^2\sum_{\gamma}\frac{(2\gamma)!}{(\gamma!)^4(a-\gamma)!(c-\gamma)!}$$

$$=a!c!\sum_{\gamma}(-4)^{\gamma}\binom{-1/2}{\gamma}\binom{a}{\gamma}\binom{c}{\gamma} \tag{23}$$

$$=a!c!\,{}_3F_2\left[\begin{matrix}-a,-c,\frac{1}{2}\\1,1\end{matrix};4\right].$$

It follows by (10) that

$$\int_{-\infty}^{\infty}[H_a(x)]^2[H_c(x)]^2e^{-x^2}dx=2^{a+c}\sqrt{\pi}\,a!c!\cdot{}_3F_2\left[\begin{matrix}-a,-c,\frac{1}{2}\\1,1\end{matrix};4\right]. \tag{24}$$

5. Further combinatorial functions and their integral representation. As before, we start with k subclasses of elements, with n_i in the ith class ($i=1,2,\cdots,k$). Now suppose that we wish to form pairs of these objects, not necessarily using all of them and with no restriction on the number of pairs to be made, but still such that no pair shall consist of two objects from the same subclass. Let $R_{\mathbf{n}}$ denote the number of ways in which this can be done. Here $\sigma(\mathbf{n})$ need not be even.

THEOREM 2.

$$R_{\mathbf{n}}=\{2^{\sigma(\mathbf{n})}\pi e\}^{-1/2}\int_{-\infty}^{\infty}\left\{\prod_{i-1}^{k}H_{n_i}(x)\right\}\exp\left(-x^2+\sqrt{2}\,x\right)dx.$$

Proof. Consider those matches which use, for each i, exactly m_i elements out of the ith subclass ($i=1,2,\cdots,k$, $0\le m_i\le n_i$), and let $m=(m_1,m_2,\cdots,m_k)$. The number of such matches will clearly be

$$\left\{\prod_{i=1}^{k}\binom{n_i}{m_i}\right\}\cdot P_m.$$

Hence, by (10),

$$R_{\mathbf{n}}=\frac{1}{\sqrt{\pi}}\sum_{m_1,m_2,\cdots,m_k}\prod_{i=1}^{k}\binom{n_i}{m_i}2^{-\sigma(m)/2}\int_{-\infty}^{\infty}\prod_{i=1}^{k}\{H_{m_i}(x)\}e^{-x^2}dx\cdot$$

$$=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}\prod_{i=1}^{k}\left\{\sum_{m_i=0}^{n_i}2^{-m_i/2}\binom{n_i}{m_i}H_{m_i}(x)\right\}e^{-x^2}dx$$

$$=\{\pi\cdot 2^{\sigma(\mathbf{n})}\}^{-1/2}\int_{-\infty}^{\infty}\prod_{i=1}^{k}\left\{H_{n_i}\left(x+\frac{1}{\sqrt{2}}\right)\right\}\cdot e^{-x^2}dx \quad \text{by (6)} \tag{25}$$

$$=\{2^{\sigma(\mathbf{n})}\cdot\pi e\}^{-1/2}\int_{-\infty}^{\infty}\left\{\prod_{i=1}^{k}H_{n_i}(x)\right\}\exp\left(-x^2+\sqrt{2}\,x\right)dx.$$

For our next example we return to the case where $\sigma(n)$ is even, and consider *unrestricted* matches on the entire set of elements. Any such match may include some homogeneous as well as heterogeneous pairs. We now prove Theorem 3.

THEOREM 3. *In the notation defined in* §1,

$$E_{\mathbf{n}}=\Omega_{\mathbf{n}}=\sqrt{\frac{2}{\pi}}\int_{-\infty}^{\infty}e^{-2x^2}\prod_{i=1}^{k}\{H_{n_i}(x)\}\,dx.$$

COROLLARY. *If k is even, and $n_1=n_2=\cdots=n_k$, then $E_{\mathbf{n}}>\Omega_{\mathbf{n}}$.*

Proof. Suppose that for each i there are α_i pairs whose members are both from the ith subset ($0\le\alpha_i\le\frac{1}{2}n_i$), and let $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_k)$. For a given α, let $T_{\mathbf{n}}^{\alpha}$ be the number of ways in which this can be done. Then, clearly,

$$T_{\mathbf{n}}^{\alpha}=\prod_{i=1}^{k}\left\{\binom{n_i}{\alpha_i}\binom{n_i-\alpha_i}{\alpha_i}\frac{\alpha_i!}{2^{\alpha_i}}\right\}P_{\mathbf{n}-2\boldsymbol{\alpha}}=\prod_{i=1}^{k}\left\{\frac{n_i!}{(n_i-2_i)!(\alpha_i!)^2\cdot 2^{\alpha_i}}\right\}P_{\mathbf{n}-2\boldsymbol{\alpha}}\tag{26}$$

Hence,

$$\begin{aligned}
E_{\mathbf{n}}-\Omega_{\mathbf{n}}&=D_{\mathbf{n}}\quad\text{(say)}\\
&=\sum_{\alpha_1,\alpha_2,\cdots,\alpha_k}(-1)^{\sigma(\alpha)}T_{\mathbf{n}}^{\alpha}\\
&=\sum_{\alpha_1,\cdots,\alpha_k}(-1)^{\sigma(\alpha)}\prod_{i=1}^{k}\left\{\frac{n_i!}{(n_i-2\alpha_i)!(\alpha_i!)^2 2^{\alpha_i}}\right\}P_{\mathbf{n}-2\boldsymbol{\alpha}}\\
&=\sum_{\alpha_1,\cdots,\alpha_k}\left(-\frac{1}{2}\right)^{\sigma(\alpha)}\prod_{i=1}^{k}\left\{\frac{n_i!}{(n_i-2\alpha_i)!(\alpha_i!)^2}\right\}\{\pi\cdot 2^{\sigma(\mathbf{n})-2\sigma(\alpha)}\}^{1/2}\\
&\quad\times\int_{-\infty}^{\infty}e^{-x^2}\prod_{i=1}^{k}\{H_{n_i-2\alpha_i}(x)\}\,dx\quad\text{by (10)}\\
&=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}e^{-x^2}\prod_{i=1}^{k}\left\{\sum_{\alpha_i}(-1)^{\alpha_i}2^{-ni/2}\frac{n_i!}{(n_i-2\alpha_i)!(\alpha_i!)^2}H_{n_i-2\alpha_i}(x)\right\}dx\\
&=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}e^{-x^2}\prod_{i=1}^{k}\left\{H_{n_i}\left(\frac{x}{\sqrt{2}}\right)\right\}dx\quad\text{by (7)}\\
&=\sqrt{\frac{2}{\pi}}\int_{-\infty}^{\infty}e^{-2x^2}\prod_{i=1}^{k}H_{n_i}(x)\,dx.
\end{aligned}\tag{27}$$

In the special case $n_1=n_2=\cdots=n_k=r$, we denote n by $k*r$. Then,

$$D_{k*r}=\sum_{\alpha_1,\cdots,\alpha_k}(-1)^{\sigma(\alpha)}T_{k*r}^{\alpha}=\sqrt{\frac{2}{\pi}}\int_{-\infty}^{\infty}e^{-2x^2}[H_r(x)]^k dx,\tag{28}$$

and hence the corollary. It follows that, for an even number of equal subsets, there are necessarily more matches using an even number of homogeneous pairs then using an odd number. We have not been able to find a purely combinatorial proof of this result. When k is odd but r even, the situation is not clear. We shall derive below ((60)) an

asymptotic estimate of D_{k*r} for large k with r fixed (and even) from which it will follow that

$$D_{k*r}>0,$$

for all sufficiently large k. However, if we take $k=1$, $r=2$, it is clear that $E_{1*2}=0$, $\Omega_{1*2}=1$ and hence $D_{1*2}=-1$.

6. Asymptotic estimates. We shall limit the discussion of asymptotic estimates to the special case $n=k*r$, and shall consider

(i) P_{k*r} as $k\to\infty$, r fixed;
(ii) P_{k*r} as $r\to\infty$, $k=2,3$ or 4;
(iii) D_{k*r} as $k\to\infty$, r fixed.

(i) *Asymptotic estimate of P_{k*r} as $k\to\infty$, k fixed.* By (10),

$$P_{k*r}=(2^{kr}\pi)^{-1/2}X,$$

where

$$X=\int_{-\infty}^{\infty}[H_r(x)]^k e^{-x^2}dx. \tag{29}$$

If r,k are both odd, then clearly $x=0$. In case at least one of them is even, X can be estimated, for large k, by the Laplace method (cf. [1, p. 41]). We begin by examining the stationary values of the integrand. These are found among the roots of the equation

$$k[H_r(x)]^{k-1}H_r'(x)=2x[H_r(x)]^k.$$

The factor $[H_r(x)]^{k-1}$ gives us, of course, zeros of the integrand. The remaining equation is

$$2xH_r(x)=kH_r'(x). \tag{30}$$

Consider first the case of odd r ($=2m+1$, say). Equation (30) is of degree $2m+2$, and both sides are even polynomials. For large k, there will be $2m$ roots near those of $H_r'(x)$, and they will tend to those of H_r' as k turns to infinity. If we denote them by $\pm\beta_i$ ($1\le i\le m$), then

$$|\beta_i|<M \qquad (i=1,2,\cdots,m), \tag{31}$$

where M is independent of k. Suppose that the remaining two roots of (30) are $\pm\alpha$. The coefficient of x^{2m+2} in (30) comes from the left-hand side and is independent of k. On the other hand, the absolute term is on the right-hand side and is proportional to k. It follows by Vieta's formula that $\alpha^2\beta_1^2\beta_2^2\cdots\beta_m^2$ is proportional to k and hence, by (31), that α is of order $k^{1/2}$. If r is even, both sides of (30) are odd polynomials, but after dividing by x the resulting equation can be handled similarly and the same result follows.

To estimate α more precisely, consider the differential equation

$$H_r''(x)-2xH_r'(x)+2rH_r(x)=0. \tag{32}$$

If we divide through by $H_r(x)$ and write $y=H_r'(x)/H_r(x)$, this becomes

$$y'+y^2-2xy+2r=0. \tag{33}$$

Now $H_r(x)$ is a polynomial of degree r and hence, for large x,

$$y=\frac{r}{x}+\sum_{n=2}^{\infty}\frac{A_n}{x^n}.$$

Substituting in (33), we can determine the coefficients A_n recursively and get

$$y=\frac{H_r'(x)}{H_r(x)}=\frac{r}{x}+\frac{r(r-1)}{2x^3}+\frac{r(r-1)(2r-3)}{4x^5}+O(x^{-7}). \tag{34}$$

It follows from (30) that α is a root of the equation

$$\frac{2x}{k}=\frac{r}{x}+\frac{r(r-1)}{2x^3}+\cdots. \tag{35}$$

We know that α is of order $k^{1/2}$. Solving (35) by successive approximation we obtain

$$\alpha=\left(\frac{kr+r-1}{2}\right)^{1/2}+O(k^{-3/2}). \tag{36}$$

Our next step is to estimate the integrand in (29) at the point $x=\alpha$. Now, by (1),

$$H_r(x)=(2x)^r\left\{1-\frac{r(r-1)}{1!(2x)^2}+\frac{r(r-1)(r-2)(r-3)}{2!(2x)^4}\cdots\right\}. \tag{37}$$

Substituting from (36), we get after some trivial manipulation,

$$e^{-\alpha^2}[H_r(\alpha)]^k\sim(2kr)^{kr/2}\exp\left(-\frac{kr+r-1}{2}\right)\left\{1+O\left(\frac{1}{k}\right)\right\}. \tag{38}$$

On the other hand, by (31),

$$e^{-\beta_i^2}[H_r(\beta_i)]^k=O(L^k)\qquad\left(i=1,2,\cdots,\left[\frac{r-1}{2}\right]\right) \tag{39}$$

for some fixed L. It follows that the maxima of the integrand at $\pm\alpha$ will be the dominant ones, and only they need be taken into account in the application of the Laplace method.

Moreover, we see that, for large k, α is bigger than the largest zero of $H_r(x)$, so that $H_r(x)>0$ in the vicinity of α. In our application of the Laplace method we may therefore write

$$X=\int_{-\infty}^{\infty}e^{f(x)}dx,$$

where $f(x)=k\log|H_r(x)|-x^2$. We have, as before,

$$\begin{aligned} f'(x)&=ky-2x\quad\left(\text{where } y=\frac{H_r'(x)}{H_r(x)}\right)\\ &=0\quad\text{at } x=\alpha, \end{aligned} \tag{40}$$

$$f'(x)=ky'-2=-k[y^2-2xy+2r]-2\quad\text{by (33)} \tag{41}$$

and so

$$\begin{aligned} f''(\alpha)&=-\frac{1}{k}\{k^2y^2-2kx\,ky+2rk^2\}-2\quad\text{by (40)}\\ &=-\frac{1}{k}\{4\alpha^2-4k\alpha^2+2rk^2\}-2\\ &=-\frac{2}{k}(2k-r+1)+O\left(\frac{1}{k^2}\right)\quad\text{by (36).} \end{aligned} \tag{42}$$

The contribution to the integral (29) from the vicinity of $x=\alpha$ therefore becomes

$$
\begin{aligned}
&\sim\int_{-\infty}^{\infty}[H_r(\alpha)]^k\exp\left(-\frac{2k+r-1}{k}(x-\alpha)^2\right)dx\\
&\sim\sqrt{\frac{\pi k}{2k+r-1}}\,[H_r(\alpha)]^k e^{-\alpha^2}\\
&\sim\sqrt{\frac{\pi}{2}}\,(2kr)^{kr/2}\exp\left(-\frac{kr+r-1}{2}\right)\left\{1+O\left(\frac{1}{k}\right)\right\}\\
&=U\quad\text{(say)}
\end{aligned}
\tag{43}
$$

by (38) and (42).

The contribution from $x=-\alpha$ is clearly $(-1)^{kr}U$. Hence we see that

$$
\begin{aligned}
P_{k*r}&\sim\frac{1+(-1)^{kr}}{\sqrt{2}}(kr)^{kr/2}\exp\left(-\frac{kr+r-1}{2}\right)\\
&=2^{-1/2}\left[1+(-1)^{kr}\right]\left(\frac{kr}{e}\right)^{kr/2}\exp\left(-\frac{r-1}{2}\right)
\end{aligned}
\tag{44}
$$

(ii) *Asymptotic estimate of* P_{k*r} *as* $r\to\infty$, $k=2,3$ *or* 4. The cases $r=2,3$ are immediate, since $P_{rr}=r!$ while

$$P_{2n,2n,2n}=\left[\frac{(2n)!}{n!}\right]^3$$

by (20). Consider therefore, $r=4$.

By (19),

$$
\begin{aligned}
P_{r,r,r,r}&=\sum_{\beta}\frac{1}{\beta!}\{P_{r,r,\beta}\}k^2\\
&=\sum_{\alpha}\frac{1}{(2\alpha)!}\{P_{r,r,2\alpha}\}^2\quad\text{since } P_{r,r,\beta}=0 \text{ for odd } \beta\\
&=(r!)^4\sum_{\alpha}\frac{(2\alpha)!}{(\alpha!)^4[(r-\alpha)!]^2}\quad\text{by (20)}\\
&=(r!)^2\sum_{\alpha}\binom{2\alpha}{\alpha}\binom{r}{\alpha}^2=(r!)^2\sum_{\alpha}(-4)^{\alpha}\binom{-1/2}{\alpha}\binom{r}{\alpha}^2.
\end{aligned}
\tag{45}
$$

It follows that

$$(r!)^{-2}P_{rrrr}$$

is equal to the absolute term in the Laurent expansion of

$$(1-4z)^{-1/2}\sum_{\alpha=0}^{r}\binom{r}{\alpha}^2 z^{-r}. \tag{46}$$

But the sum in (46) is equal (cf. [4, p. 91]) to

$$\left(1-\frac{1}{z}\right)^r P_r\left(\frac{z+1}{z-1}\right),$$

where $P_r(t)$ is the rth Legendre polynomial. Hence

$$P_{rrrr}=\frac{(r!)^2}{2\pi i}\int_{|z|=\rho<1/4}(1-4z)^{-1/2}(z-1)^r z^{-r-1}P_r\left(\frac{z+1}{z-1}\right)dz. \tag{47}$$

Now it is easily seen that for large r, the greatest term in the sum in (46) is at $\alpha\sim\frac{2}{3}r$ and by a straightforward application of Stirling's formula, that this term is of order $3^{2r}\times r^{-3/2}$, while the number of terms in this sum is $r+1$. Hence, if we define

$$S_r=(r!)^{-2}3^{-2r}P_{rrrr} \tag{48}$$

and

$$S(w)=\sum_{r=0}^{\infty}S_r w^{2r}, \tag{49}$$

the radius of convergence of this power series will be 1. Moreover, for $|w|<1$,

$$\begin{aligned}S(w)&=\frac{1}{2\pi i}\int_{|z|=\rho<1/4}(1-4z)^{-1/2}\sum_{r=0}^{\infty}\left\{\left(\frac{z-1}{9z}\right)^r w^{2r}P_r\left(\frac{z+1}{z-1}\right)\right\}\frac{dz}{z}\\ &=\frac{1}{2\pi i}\int_{|z|=\rho}(1-4z)^{-1/2}\left\{1-\frac{2(z+1)}{9z}w^2+\frac{(z-1)^2}{81z^2}w^4\right\}^{-1/2}\frac{dz}{z}\\ &=\frac{9}{4\pi i}\int_{|z|=\rho}\left\{\left(\frac{1}{4}-z\right)\left[(3-w)^2z-w^2\right]\left[(3+w)^2z-w^2\right]\right\}^{-1/2}\frac{dz}{z}.\end{aligned} \tag{50}$$

Following the method of Darboux [2], we shall derive the asymptotic properties of $S_r(r\to\infty)$ from the behavior of $S(w)$ at its singularities on the circle $|w|=1$. Begin by substituting

$$\zeta=\frac{\left[\left(\frac{w}{3-w}\right)^2-z\right]}{\left[\left(\frac{w}{3+w}\right)^2-z\right]}. \tag{51}$$

Then (50) becomes

$$S(w)=\frac{3\sqrt{3}}{2\pi i}(3-w)^{-3/2}(1+w)^{-1/2}\int_{\Gamma}\{(1-\zeta)(\lambda-\zeta)\}^{-1/2}d. \tag{52}$$

where $\lambda=(1-w)/(1+w)((3+w)/(3-w))^3$ and Γ is the transform of the circle $|z|=\rho$. It is easily verified that, if $|w|<1$ and $\rho<\frac{1}{4}$, Γ is a circle containing 1, λ as interior points but excluding the point $z=0$. We may therefore deform it to Γ_1, as in the figure. If the small circles in Γ_1 are of radius t, their contributions to the integral in (52) will be $O(t^{1/2})$. On the other hand the integrand clearly changes sign on going round either of them. Letting, therefore, $t\to 0$, we get

$$S(w)=\frac{3\sqrt{3}}{\pi i}(3-w)^{-3/2}(1+w)^{-1/2}\int_1^{\lambda}\{\zeta(1-\zeta)(\lambda-\zeta)\}^{-1/2}d\zeta.$$

Now write $\zeta=1-(1-\lambda)\sin^2\theta$, yielding

$$S(w)=\frac{6\sqrt{3}}{\pi}(3-w)^{-3/2}(1+w)^{-1/2}\int_0^{\pi/2}\{1-(1-\lambda)\sin^2\theta\}^{-1/2}d\theta. \tag{53}$$

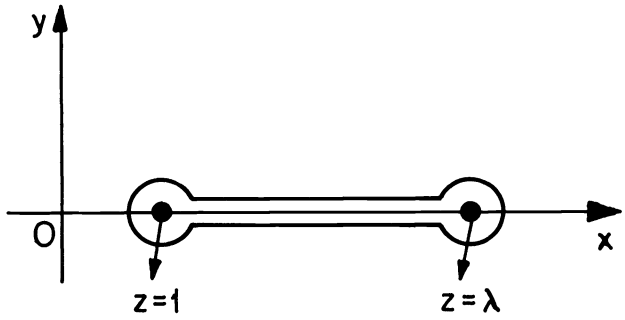

FIG. 1

Thus the only possible singularities of $S(w)$ are at $w=-1,3$ and also when $\lambda=0$, i.e., at $w=1,-3$. Since we are interested in the singularities closest to the origin, while, by (49), $S(w)$ is an even function, it will suffice to double the contribution from $w=1$. Now as $w\to 1$, $\lambda\to 0$. If we write $\lambda=1-k^2$, then the integral in (53) is simply $K(k)$, the complete elliptic integral. As $k\to 1$, it is known ([3, p. 905]) that

$$K(k)\sim\left\{1+\frac{1}{4}k'^2+\frac{9}{64}k'^4\right\}\log\frac{1}{k'}, \tag{54}$$

where

$$k'=\sqrt{1-k^2}=\lambda^{1/2}=\left(\frac{1-w}{1+w}\right)^{1/2}\left(\frac{3+w}{3-w}\right)^{3/2}. \tag{55}$$

Writing $q=1-w$ and substituting in (53) and (54) we get, after some rearranging,

$$\begin{aligned}S(w)&\sim\frac{3\sqrt{3}}{\pi}\left\{1+\frac{1}{2}q+\frac{3}{8}q^2\right\}\left\{-\frac{1}{2}\log q-\log 2+O(q)\right\}\\&=\frac{3\sqrt{3}}{\pi}\left\{1+\frac{1}{2}(1-w)+\frac{3}{8}(1-w)^2\right\}\left\{-\frac{1}{2}\log(1-w)-\log 2+O(1-w)\right\},\end{aligned}$$

leading to

$$S_r\sim\frac{3\sqrt{3}}{\pi}\left\{1-\frac{1}{4r}+\frac{3}{16r^2}+O(r^{-3})\right\}, \tag{56}$$

and hence

$$P_{rrrr}\sim\frac{3^{2r+3/2}(r!)^2}{4\pi r}\left\{1-\frac{1}{4r}+\frac{3}{16r^2}+O(r^{-3})\right\}. \tag{57}$$

It follows by (10) that

$$\int_{-\infty}^{\infty}[H_r(x)]^4e^{-x^2}dx\sim\frac{3}{4r}\sqrt{\frac{3}{\pi}}\,\frac{6^{2r}}{(r!)^2}\left\{1-\frac{1}{4r}+\frac{3}{16r^2}+O(r^{-3})\right\}. \tag{58}$$

The estimate (58) has been checked numerically, and the relative error turns out to be less than 1% already for $r\geq 4$.

(iii) D_{k*r} *for fixed even* r, $k\to\infty$. The argument is very similar to that used to derive (44) and we omit the details. We see from (28) that

$$D_{k*r}=\sqrt{\frac{2}{\pi}}\int_{-\infty}^{\infty}\exp\bigl(k\log H_r(x)-2x^2\bigr)\,dx. \tag{59}$$

If we write

$$f(x)=k\log H_r(x)-2x^2$$

we find now that the significant maximum is again at a value of x of order $k^{1/2}$. Following through the rest of the argument, we find that

$$D_{k*r}\sim 2^{-1/2}\bigl[1+(-1)^{kr}\bigr]\left(\frac{kr}{e}\right)^{kr/2}e^{-(r-1)}. \tag{60}$$

It follows that $D_{k*r}\geq 0$ for sufficiently large k and zero if and only if r,k are both odd.

REFERENCES

[1] E. T. COPSON, *Asymptotic Expansions*, Cambridge University Press, London, 1967.

[2] G. DARBOUX, *Memoire sur l'approximation des fonctions de très grands nombres*, J. de Math., 4 (1978), pp. 5–56.

[3] I. S. GRADSHTEYN AND I. M. RYZHIK, *Tables of Integrals, Series and Products*, Academic Press, New York, 1965.

[4] G. PÓLYA AND G. SZEGÖ, *Aufgaben und Lehrsätze aus der Analysis* II, Dover, New York, 1945.

[5] G. SZEGÖ, *Orthogonal Polynomials*, AMS Colloquium Publications 23, American Mathematical Society, New York, 1959.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0001 $01.00/0

THE LOCAL GEOMETRIC ASYMPTOTICS OF CONTINUUM EIGENFUNCTION EXPANSIONS I. OVERVIEW*

S. A. FULLING†

Abstract. The well-known connection between (a) the asymptotic density of eigenvalues of a differential operator H, and (b) the geometry of the region or manifold where H acts, has a local generalization: There is a connection between (a′) the spectral measures or projection kernel describing the proper normalization, relative to a point x_0, of the expansion of an arbitrary function in eigenfunctions of an operator H (possibly with continuous spectrum), and (b′) the values of the coefficients (symbol) of H and their derivatives at x_0. Potential applications (especially in general-relativistic quantum field theory) increasingly call for a detailed development of this theory (including calculation of numerical coefficients), which is begun here. The small-time expansion of the Green function of the heat operator, $\frac{\partial}{\partial t}+H$, is used to define the "mean" or "effective" expansion of the spectral measures. For the case of one independent and one dependent variable, the spectral and heat expansions are worked out in detail to a rather high order. They are shown to be related to (and obtainable from) existing phase-integral expansions with local amplitude functions (WKB–Fröman expansions) of the eigenfunctions.

1. Introduction. "Spectral asymptotics" is a name which may well be applied to the study of the relationships among the following: (1) the asymptotic density of the eigenvalues λ_ν of a differential operator H,

$$H\psi_\nu=\lambda_\nu\psi_\nu, \tag{1.1}$$

as $|\lambda|\to\infty$, with appropriate generalizations to the case of continuous spectrum; (2) the coefficient functions in H, the geometry of the region or manifold which serves as domain for the functions on which H acts, and the boundary conditions, if any; (3) the behavior, in various limits, of the integral kernels (Green functions) of operator functions of H, such as the kernels of the heat operator, e^{-tH}, and the resolvent operator, $(H-z)^{-1}$, and various kernels associated with the wave equation,

$$\frac{\partial^2 u}{\partial t^2}+Hu=0. \tag{1.2}$$

To these one can add another topic which has developed rather separately, (4) the approximate form of the eigenfunctions ψ_ν in the limit of large $|\lambda_\nu|$ (i.e., the phase-integral, Liouville–Green or WKB approximation).

Spectral asymptotics has found applications (not counting those of (4)!) in many areas of physics, "including resonator acoustics, perfect gases, nuclei, black-body radiation, correlation functions, and ... quantum statistics and the theory of condensed matter" [12]. (See also [4]–[9].) More recently it has become crucial to the theory of renormalization for quantum fields under spatially inhomogeneous conditions (e.g., [21], [58], [59], [28]), and to the study of the effects of boundaries on quantum fields [10], [11], [13], [23], [42], [18].

Around 1967 (date of the useful review by Clark [22]), the flavor of rigorous work in this field changed. The torch passed from analysts to geometers and topologists as

* Received by the editors July 23, 1981. This research was supported in part by the National Science Foundation under grants PHY 79-15229 to the Texas A & M Research Foundation and PHY 77-27084 to the Institute for Theoretical Physics.

† Department of Mathematics, Texas A & M University, College Station, Texas 77843. This research was completed while the author was at the Institute for Theoretical Physics, University of California, Santa Barbara, California 93106.

the heat-kernel expansion was applied in connection with the index theorem for elliptic operators on compact manifolds [2], [45], [30]. Moreover, great emphasis has been placed on the "inverse" problem, as in the title of the widely read article of Kac [39]. That is, in the notation of our opening paragraph, the problem is seen as one of deducing (2) from (1), which in practice becomes deduction of (2) from (3). These trends, along with some technical barriers, have meant that during this period of rapid progress, the subject has not developed in certain directions as much as might have been expected.

In order to explain this remark, let us now make the discussion more specific. Let H be a second-order linear differential operator on scalar functions $\phi(x)=\phi(x^1,\cdots,x^m)$, defined on an m-dimensional region or manifold, M, and let t be a positive real variable. Assume H to be elliptic and selfadjoint with spectrum bounded below. The heat kernel of H is the integral kernel, $K(t,x,y)$, of the operator e^{-tH}. That is, the solution of

$$\frac{\partial u}{\partial t}+Hu=0, \qquad u(0,x)=\phi(x), \tag{1.3}$$

is

$$u(t,x)=\int_M K(t,x,y)\phi(y)\,dy, \tag{1.4}$$

or K is the solution of

$$\frac{\partial K}{\partial t}+H_{(x)}K=0, \qquad K(t,x,y)\to\delta(x,y) \quad \text{as } t\downarrow 0. \tag{1.5}$$

(If M is a Riemannian manifold, dy is to be interpreted as the covariant volume element, $g^{1/2}d^m y$, and $\delta(x,y)$ as the corresponding covariant Dirac distribution.) It has long been known that

$$K(t,x,x)\sim(4\pi t)^{-m/2} \quad \text{as } t\downarrow 0, \tag{1.6}$$

and that (1.6) implies, if M is compact and the coefficient functions of H are smooth and bounded, that the number of eigenvalues of H less than λ obeys

$$N(\lambda)\sim(4\pi)^{-m/2}\Gamma\left(\frac{m}{2}+1\right)^{-1}V\lambda^{m/2} \quad \text{as } \lambda\to\infty, \tag{1.7}$$

where $V\equiv\int_M dy$ is the volume of M [22], [12]. Furthermore, (1.6) is the first term in an asymptotic expansion,

$$K(t,x,x)\sim(4\pi t)^{-m/2}\sum_{n=0}^{\infty}a_n(x)t^n, \tag{1.8}$$

where each $a_n(x)$ is a geometrically invariant polynomial in the coefficient functions of H and their derivatives, evaluated at x alone. This remains true when the scalar functions ϕ are replaced by multicomponent functions (sections of a vector bundle over M), in which case K and the a_n are matrix valued. The a_n have been calculated through $n=3$, and various general properties of them have been established (e.g., [52], [51], [56], [31], [32]). If M has a boundary, (1.8) is not uniformly valid as x approaches the boundary, but there is another expansion, in terms of invariants defined on the boundary, for the contribution of the region near the boundary to the integral of $K(t,x,x)$ over M (e.g., [39], [34], [41]). Observe now two ways in which this picture is incomplete.

(1) The assumption that M was compact ensured both that the spectrum of H was discrete and that the integrals of K and a_n over M converged. But the expansion (1.8) is a purely local statement and remains valid in the noncompact case [59]. On the other hand, "the number of eigenvalues less than λ" is meaningless whenever H possesses a continuous spectrum, and $N(\lambda)$ may not grow as $\lambda^{m/2}$ even when the spectrum is discrete, if M is noncompact [54]. However, it is possible to formulate a "local" spectral asymptotics of which the eigenvalue law (1.7) for compact domains is an integrated form. This local spectral information is of interest even when the spectrum is discrete. Consider, for example, the problem of calculating the value of the integral kernel of the operator H^{-s} when both of its arguments are equal to x_0. (This is typical of the problems which arise in the field-theoretic applications previously mentioned.) If the spectrum of H has a positive lower bound and s is sufficiently large, this will be a finite number given by

$$\sum_{\nu} \lambda_{\nu}^{-s} |\psi_{\nu}(x_0)|^2,$$

if the λ_{ν} and ψ_{ν} are all the eigenvalues and normalized ($\|\psi_{\nu}\| = 1$) eigenvectors of H in $L^2(M)$. To estimate the contribution from large eigenvalues we need not only an asymptotic formula for the density of eigenvalues along the λ axis, but also an asymptotic approximation for $|\psi_{\nu}(x_0)|^2$. Alternatively, we could use eigenfunctions $\tilde{\psi}_{\nu}^{x_0}(x)$ normalized so that $|\tilde{\psi}_{\nu}^{x_0}(x_0)| = 1$ (assuming, for simplicity, that none of the eigenfunctions vanishes at x_0); then the quantity to be calculated becomes

$$\sum_{\nu} \lambda_{\nu}^{-s} \|\tilde{\psi}_{\nu}^{x_0}\|^{-2},$$

and we need an asymptotic formula for the density of eigenvalues, "weighted" by the factor $\|\tilde{\psi}\|^{-2}$. This latter formulation is the one which generalizes to continuous spectrum, at least in one dimension, where the weighted summation above is replaced by integrations with respect to the x_0-based Titchmarsh–Kodaira spectral measures for H (see §2). We shall show that these measures do have "averaged" (see below) asymptotic expansions, which, like (1.8), depend on H locally (i.e., are determined by the coefficients in H and their derivatives at x_0). In view of its obvious practical importance, the calculational aspect of this subject is surprisingly underdeveloped, although the key abstract facts have been established by Hörmander [37], [38]. (See also [24, §2] and [46].)

(2) One might expect the asymptotic approximation (1.7) to be the first term of a series in λ^{-1} (times $\lambda^{m/2}$), with coefficients related to the integrals of the a_n over M and similar integrals over the boundary. Similarly, there should be local series for the spectral measures (or, in higher dimensions, the spectral projections [37], [38]). Such series can indeed be determined formally by the requirement of consistency with (1.8) (see §3), but the "Tauberian" methods used to establish (1.7) [22] break down when applied to the higher-order terms. Indeed, the series which extends (1.7) is *not* literally asymptotic to the true eigenvalue distribution [3], [35], [4], [8]; it is valid only in some "averaged" sense [47], [48], [15], [16], [4], [6], [37], [38], [24]. The implications of this situation have not yet been adequately explored. The eigenvalue density (more generally, the spectral measures or projection) is ordinarily sought as a means to some end, such as to calculate (at least approximately) some operator function of H. In certain circumstances the "mean" or "effective" density may provide all the needed information. For example, the mean measures obviously suffice for computing the heat kernel

at small t, since by construction they yield the expansion (1.8), which is known independently to be genuinely asymptotic. A less circular example is provided by the divergent terms in the short-distance expansion of the two-point function of linear quantum field theory (a certain distributional solution of (1.2)), which are locally determined by H [28]. If there exists an operator H' "locally equivalent" [28] to H and such that the mean spectral measures for H' are genuinely asymptotic to the exact spectral measures for H', then it follows that the mean measures correctly reproduce the divergent terms. The hypothesis here will be discharged in §4 for one-dimensional M, and no great difficulty is expected in establishing it in general, at least for smooth coefficients. The point is not that the form of the divergent terms should be determined from an eigenfunction expansion, since the divergences are already known independently (e.g., [21], [28]). Rather, the goal is to implement a covariant renormalization procedure (cf. [21], [58], [59]) at the level of the integrand of an eigenfunction expansion ("mode sum" in the physics literature) in concrete calculations.

This paper is planned as the first in a series exploring these two areas. By emphasizing the central role of the spectral decomposition (or continuum eigenfunction expansion) instead of any particular Green function, the subject will be returned to its analytical foundations and conceptually unified. Another unification consists in pointing out a relationship between the mean spectral measures and the amplitude of the phase-integral approximation of the individual eigenfunctions, the fourth element of spectral asymptotics. A preliminary account was presented in [26].

The first two papers are devoted to the local spectral asymptotics, away from boundaries, of ordinary differential equations (i.e., $m=1$). This first paper treats in full the scalar case (i.e., (1.1) is a single differential equation), but also contains preliminary material needed for the treatment of the vector case (i.e., (1.1) is a system of equations) in the second paper, and remarks relevant to the problem of larger m.

Section 2 establishes a normal form for the operators considered and reviews the Titchmarsh–Kodaira theory of eigenfunction expansions for ordinary differential operators with possibly continuous spectrum.

Section 3 is concerned with the relation between the mean asymptotic expansions of the spectral measures and the local asymptotic expansions of the heat kernel and its derivatives.

The operators act on functions ϕ defined on an arbitrary manifold M (here one-dimensional) and taking values in the fibers of an arbitrary vector bundle over M; also, no special asymptotic behavior is required of the coefficients (except what is needed to bound the spectrum below). But the local nature of the expansions allows the operator, manifold and vector bundle to be replaced for calculational purposes by an operator on functions from $\mathbb{R}$ into $\mathbb{C}^r$ (r being the fiber dimension of the bundle), with coefficients which trivialize at infinity, which coincides with (a local representation of) the original operator in some neighborhood of the point in question. For such an operator (with C^∞ coefficients) the eigenfunctions can be approximated to arbitrary order in λ^{-1} by a phase-integral expansion of the [Liouville–Green–Jeffreys–...–] WKB type. In the scalar case the most convenient form of the WKB expansion for our purposes is that of Fröman [25], which has been carried to very high order by Campbell [17]. The relevant feature of this expansion is that it displays the amplitude of the solution, for a definite normalization at infinity, as a local functional of the coefficients of H. The formulas of [17] allow us in §4 to calculate easily the coefficients in the mean spectral and heat expansions up to $n=6$, one or two orders beyond previously published results [29], [32] and near the limit of editorial feasibility. (The provision of

computer subroutines, in place of published tabulations, is clearly more appropriate at high orders and will be discussed in Paper II.)

The main task in extending this treatment to the vector case is to construct the analogue of the WKB–Fröman expansion. This is new, and should be of independent interest. It, and its application to the spectral problem, make up Paper II.

Paper III will discuss the relation between the phase-integral expansion and the spectral measures near a boundary, or near certain types of coefficient singularity. It is hoped that later papers wil deal with manifolds of higher dimension (partial differential operators). The application to quantum field theory will be developed elsewhere.

2. Operators and spectral measures. Let M be a one-dimensional (Hausdorff, paracompact, connected, C^∞) manifold—compact or noncompact, with or without boundary. There are but four essentially distinct possibilities: the real line, the half-line, the interval, and the circle.

We consider an operator of the form

$$H=-\frac{d^2}{dx^2}-E(x) \tag{2.1}$$

acting on functions from M into $\mathbb{C}^r$; x is a coordinate on M, and E is an $r\times r$ matrix-valued C^∞ function. This form is adequate in one dimension to cover all the C^∞ elliptic second-order differential operators considered by Gilkey [31], [32]: Such an operator consists of the Laplacian of some connection with respect to some Riemannian metric, plus a zeroth-order term, $-E(x)$. (Operators with more general second-order terms present deeper problems, which will not be addressed here.) On one-dimensional manifolds all metrics and connections are locally flat, so it is possible to introduce, at least locally, coordinates for the independent and dependent variables which cast the operator into the form (2.1) (see Remark 2.1). In fact, these coordinates will do globally, provided that the circle is treated as an interval with endpoints identified (see Remarks 2.2 and 2.4).

E will be called the "potential", by analogy with the Schrödinger equation. It is the negative of the conventional potential,

$$V\equiv -E. \tag{2.2}$$

This sign convention is chosen to facilitate comparison with [32] and [17], and because it happens to make all coefficients in the expansions positive. The successive derivatives of E with respect to x will be denoted by E', E'', $E^{(3)}$, $E^{(4)}$, $\cdots$; similarly for derivatives of other quantities.

Remark 2.1. With respect to arbitrary local coordinates, the equation $H\psi=\lambda\psi$ takes the form

$$-a(z)\frac{d^2\phi}{dz^2}-iB(z)\frac{d\phi}{dz}+C(z)\phi=\lambda\phi, \tag{2.3}$$

where, for each z, $\phi(z)\in\mathbb{C}^r$ comprises the coordinates of the abstract function (section) ψ with respect to an arbitrary basis at z; $a(z)$ is a positive number; $B(z)$ and $C(z)$ are $r\times r$ matrices. It is of interest to write down the transformation which converts (2.3) to the form

$$H\psi(x)\equiv -\psi''(x)-E(x)\psi(x)=\lambda\psi(x). \tag{2.4}$$

One has the freedom to make an arbitrary C^∞ redefinition of the independent variable ("coordinate transformation") and an arbitrary linear redefinition of the r dependent variables ("gauge transformation"):

$$x=x(z), \qquad \psi(x)=U(z)^{-1}\phi(z). \tag{2.5}$$

The transformation which simultaneously eliminates first-order terms and sets the coefficient of the second-order term to -1 is unique up to a z-independent gauge transformation and a constant of integration in the coordinate:

$$\frac{dx}{dz}=a^{-1/2}, \tag{2.6}$$

$$U=a^{1/4}W, \quad \text{where} \quad \frac{dW}{dz}=-\frac{i}{2a}BW. \tag{2.7}$$

The resulting potential is

$$E(x)=W^{-1}\left[\frac{1}{4}\ddot{a}-\frac{3}{16}\frac{(\dot{a})^2}{a}-\frac{i}{2}\dot{B}+\frac{i}{2}\frac{\dot{a}}{a}B+\frac{1}{4a}B^2-C\right]W, \tag{2.8}$$

where the dot indicates differentiation with respect to z and all functions on the right are evaluated at $z=z(x)$. To apply the results of this paper, which are expressed invariantly in terms of E, to an operator presented in the form (2.3), it is not necessary to work out the transformation (2.5) explicitly. It suffices to know E and all its derivatives with respect to x as functions of z, and these are calculable, up to the overall similarity transformation by W, by means of formulas (2.6)–(2.8). The equation for W may not be explicitly solvable, but the value of W at the initial point is arbitrary, and the full solution will not appear in a local expansion.

If $E(x)$ is a selfadjoint matrix, then H is a formally selfadjoint operator with respect to the obvious inner product defined by the standard inner product in $\mathbb{C}^r$ and Lebesgue measure, dx, on M. This will henceforth be assumed, though the formulas for $a_n(x)$ are valid more generally [32]. H defines an Hermitian (symmetric) operator on the domain $C_0^\infty(M;\mathbb{C}^r)$ in the Hilbert space $L^2(M;\mathbb{C}^r)$. Questions about spectrum must refer to a particular selfadjoint extension of H, whose definition may require boundary conditions in addition to the expression (2.1); we assume these to be given, and H henceforth denotes this selfadjoint operator. To ensure existence of the heat kernel, it is also assumed that the spectrum of H is bounded below. (Since E is assumed smooth, this is automatic for the circle and can be violated in the other cases only if $-E(x)$ goes sufficiently fast to $-\infty$ at an endpoint of M.)

Remark 2.2. If $r=1$ and M is a finite or infinite interval of $\mathbb{R}$, then the possible selfadjoint extensions are given by the limit-point/circle theory of Weyl [55], [36], [1], [57] with regular Sturm–Liouville theory as a special case. For a finite interval (a,b) with two regular endpoints, the most general set of boundary conditions mixes the two endpoints:

$$\alpha_i\psi(a)+\beta_i\psi'(a)=\gamma_i\psi(b)+\delta_i\psi'(b), \qquad i=1,2, \tag{2.9}$$

where the two strings of coefficients must be linearly independent and satisfy

$$\begin{aligned} \operatorname{Im}(\bar{\alpha}_i\beta_i)&=\operatorname{Im}(\bar{\gamma}_i\delta_i), \\ \bar{\alpha}_1\beta_2-\bar{\beta}_1\alpha_2&=\bar{\gamma}_1\delta_2-\bar{\delta}_1\gamma_2. \end{aligned} \tag{2.10}$$

When M is the circle and one identifies it with $[0,2\pi)$, the vector-bundle connection implicit in H determines a $\zeta\in\mathbb{C}$ such that

$$\psi(2\pi)=\zeta\psi(0),\qquad \psi'(2\pi)=\zeta\psi'(0) \tag{2.11}$$

for all $\psi\in C^\infty(M)$. (By the gauge transformation in (2.5) one has adopted a basis for the fiber at each point x, which is parallel-transported around the circle by the connection; ζ describes how the terminal basis (here a single vector) matches up with the initial basis. By construction, the inner product in the fibers is invariant under this parallel transport.) Equations (2.11) are the boundary conditions (2.9) needed to realize H as a selfadjoint operator in $L^2(0,2\pi)$. (Classical periodic boundary conditions correspond to the trivial connection.) Then (2.10) indicates that $|\zeta|=1$ is necessary and sufficient for selfadjointness; this simply expresses the continuity of the fiber inner product at the point where the circle was cut. Obviously, the generalization to $r>1$ is that $\psi(2\pi)$ and $\psi'(2\pi)$ be related to $\psi(0)$ and $\psi'(0)$ by a unitary matrix.

Remark 2.3. Much of the literature on eigenfunction expansions (e.g., [43], [55], [36]) deals with an operator

$$H\phi=-\frac{1}{k}\frac{d}{dz}\left(p\frac{d\phi}{dz}\right)+q\phi \tag{2.12}$$

with k,p, and q real-valued, which is manifestly formally selfadjoint with respect to the measure $k(z)dz$. (Here $r=1$.) The connection with (2.3) and (2.6)–(2.8) is

$$p=\exp\left(i\int^z\frac{b}{a}dz_1\right),\qquad k=\frac{p}{a},\quad q=c, \tag{2.13}$$

$$\frac{dx}{dz}=\left(\frac{k}{p}\right)^{1/2},\qquad u=(kp)^{-1/4}, \tag{2.14}$$

$$\begin{aligned}E(x)=-q&-\frac{1}{4}\frac{1}{k^2}\frac{d^2}{dz^2}(kp)+\frac{1}{16}\frac{1}{k^3p}\left[\frac{d}{dz}(kp)\right]^2\\&+\frac{1}{4}\frac{\dot{k}}{k^3}\frac{d}{dz}(kp).\end{aligned} \tag{2.15}$$

The Titchmarsh–Kodaira eigenfunction expansion [57], [43], [55], [20], [36], [53], [1], [49] provides a concrete realization of the spectral theorem for a selfadjoint operator of the type considered here. Since it is constructed relative to a distinguished point $x\in M$, it is a very natural and useful tool in developing local expansions at x such as (1.8).

For simplicity we consider first the scalar case, but write the formulas in a way that permits an instant generalization. (That is, the position of the differential $d\mu^{jk}(\lambda)$ in (2.18), etc., is required by the eventual matrix reinterpretation of the formula; it does *not* mark the scope of the λ integration.) Let us denote the distinguished point by x_0 ("x" being needed as a variable). For each $\lambda\in\mathbb{C}$ there exist two linearly independent C^∞ solutions of (2.4) as a classical differential equation (without regard to boundary conditions or square integrability). Let $\psi_{\lambda j}(x)$ $(j=0,1)$ be the solutions with data

$$\begin{aligned}&\psi_{\lambda 0}(x_0)=1,\qquad \psi'_{\lambda 0}(x_0)=0,\\&\psi_{\lambda 1}(x_0)=0,\qquad \psi'_{\lambda 1}(x_0)=1.\end{aligned} \tag{2.16}$$

For $f\in C_0^\infty(M)$, define

$$\tilde{f}_k(\lambda)=\int_M\overline{\psi_{\lambda k}(x)}f(x)\,dx. \tag{2.17}$$

Remark 2.4. The present discussion is much simplified by subsuming the circle under the finite interval as described in Remark 2.2. In that context, "$f\in C_0^\infty(M)$" means that the smooth function f vanishes in neighborhoods of 0 and 2π, and a "classical solution" need not satisfy (2.11).

The principal result of the Titchmarsh–Kodaira theory is that

$$f(x)=\int_{-\infty}^{\infty}\sum_{j=0}^{1}\sum_{k=0}^{1}\psi_{\lambda j}(x)\,d\mu^{jk}(\lambda)\tilde{f}_k(\lambda) \tag{2.18}$$

for certain Stieltjes measures $d\mu^{jk}$ with supports contained in $\sigma(H)$ (the spectrum of H); starting from (2.26) below, the $d\mu^{jk}(\lambda)$ can be calculated by Cauchy's formula [49] or the Stieltjes inversion formula [53, App. 1] from the integral kernel, $G_z(x,y)$, of the resolvent operator $(H-z)^{-1}$ ($z\in\mathbb{C}$); G_z can be constructed in turn out of the classical solutions $H\psi=z\psi$ (this being where the boundary conditions of selfadjointness enter). For the explicit formulas, see the references. The formulas (2.17) and (2.18) remain valid for any $f\in L^2(M)$, although then the integrals may converge only "in the mean", as for Fourier transforms. Thus (2.18) is the decomposition of an arbitrary element of the Hilbert space into eigenfunctions of H. (Of course, when λ is in the continuous spectrum, the eigenfunctions will not themselves be elements of $L^2(M)$.) Also, (2.17) is the unitary mapping of $L^2(M)$ onto the Hilbert space of functions of λ where H acts as multiplication by λ, and where the norm is (cf. (2.18) and (2.28))

$$\|f\|^2=\int_{-\infty}^{\infty}\sum_{j,k}\overline{\tilde{f}_j(\lambda)}\,d\mu^{jk}(\lambda)\tilde{f}_k(\lambda). \tag{2.19}$$

Prototypes, for operators with purely continuous and purely discrete spectrum, respectively, are the Fourier transform and Fourier series, suitably rearranged.

Example 2.1. $M=\mathbb{R}$, $E\equiv 0$. Then one has

$$\psi_{\lambda 0}(x)=\cos[\omega(x-x_0)],\quad \psi_{\lambda 1}(x)=\omega^{-1}\sin[\omega(x-x_0)],\quad \omega\equiv\lambda^{1/2}, \tag{2.20}$$

and

$$\tilde{f}_0(\lambda)=\int_{-\infty}^{\infty}\cos[\omega(x-x_0)]\,f(x)\,dx,\qquad \tilde{f}_1(\lambda)=\frac{1}{\omega}\int_{-\infty}^{\infty}\sin[\omega(x-x_0)]\,f(x)\,dx.$$

The conventional Fourier transform is

$$\begin{aligned}\hat{f}(p)&=(2\pi)^{-1/2}\int_{-\infty}^{\infty}e^{-ipx}f(x)\,dx\\&=(2\pi)^{-1/2}e^{-ipx_0}\big[\tilde{f}_0(p^2)-ip\tilde{f}_1(p^2)\big],\end{aligned}$$

with the inversion formula

$$\begin{aligned}f(x)&=(2\pi)^{-1/2}\int_{-\infty}^{\infty}e^{ipx}\hat{f}(p)\,dp\\&=\frac{1}{2\pi}\int_{-\infty}^{\infty}\big[\psi_{p^2 0}(x)+ip\psi_{p^2 1}(x)\big]\big[\tilde{f}_0(p^2)-ip\tilde{f}_1(p^2)\big]\,dp\\&=\frac{1}{\pi}\int_0^{\infty}\big[\psi_{\omega^2 0}(x)\tilde{f}_0(\omega^2)+\psi_{\omega^2 1}(x)\tilde{f}_1(\omega^2)\omega^2\big]\,d\omega.\end{aligned}$$

Thus one has

$$d\mu^{00}(\lambda)=\pi^{-1}\,d\omega,\qquad d\mu^{11}(\lambda)=\pi^{-1}\omega^2\,d\omega, \tag{2.21}$$

and $d\mu^{01}=d\mu^{10}=0$, for all choices of x_0.

Example 2.2. $M=(0,\pi)$ with the boundary conditions $\psi(0)=0=\psi(\pi)$; $E\equiv 0$. Keeping the notation (2.20), we have (for all real λ)

$$\tilde{f}_0(\lambda)=\int_0^\pi \cos[\omega(x-x_0)]\,f(x)\,dx, \qquad \tilde{f}_1(\lambda)=\frac{1}{\omega}\int_0^\pi \sin[\omega(x-x_0)]\,f(x)\,dx.$$

The Fourier sine series is

$$\begin{aligned} f(x)&=\sum_{n=1}^{\infty} b_n \sin(nx)\\ &=\sum_{n=1}^{\infty} b_n[\sin(nx_0)\psi_{n^2 0}(x)+n\cos(nx_0)\psi_{n^2 1}(x)],\\ b_n&=\frac{2}{\pi}\int_0^\pi \sin(nx)f(x)\,dx\\ &=\frac{2}{\pi}[\sin(nx_0)\tilde{f}_0(n^2)+n\cos(nx_0)\tilde{f}_1(n^2)], \end{aligned}$$

whence

$$\begin{aligned} d\mu^{00}(\lambda)&=\frac{2}{\pi}\sum_{n=1}^{\infty}\sin^2(nx_0)\delta(\omega-n)\,d\omega,\\ d\mu^{11}(\lambda)&=\frac{2}{\pi}\sum_{n=1}^{\infty}n^2\cos^2(nx_0)\delta(\omega-n)\,d\omega,\\ d\mu^{01}(\lambda)&=d\mu^{10}(\lambda)=\frac{2}{\pi}\sum_{n=1}^{\infty}n\sin(nx_0)\cos(nx_0)\delta(\omega-n)\,d\omega. \end{aligned} \tag{2.22}$$

The spectral measures are Stieltjes measures with respect to functions $\mu^{jk}(\lambda)$, which can be made unique by the convention (3.1). In (2.22) these are step functions; in (2.21) they are increasing, differentiable functions. In general, $[\mu^{jk}(\lambda)]$ is a nondecreasing, selfadjoint (in fact, real and symmetric) matrix-valued function (i.e., $\mu^{jk}(\lambda+\varepsilon)-\mu^{jk}(\lambda)$ is positive semidefinite), which jumps at point eigenvalues of H, increases smoothly in intervals of absolutely continuous spectrum, and is constant on intervals disjoint from $\sigma(H)$ (cf. [20]). From a calculational point of view, there is seldom a reason to write down the functions $\mu^{jk}(\lambda)$ themselves; the measures are better expressed in the differential form of a "spectral density" times $d\lambda$ or $d\omega$ ($\omega\equiv\lambda^{1/2}$). The densities are generally distributions (as in (2.22)) because of the presence of point spectrum.

The transform, via (2.17), of Hf is $\lambda\tilde{f}_k(\lambda)$. Thus by (2.18) one has

$$Hf(x)=\int_{-\infty}^{\infty}\lambda\sum_{j,k}\psi_{\lambda j}(x)\,d\mu^{jk}(\lambda)\int_M\overline{\psi_{\lambda k}(y)}f(y)\,dy,$$

from which it follows that

$$E_\lambda(x,y)=\int_{-\infty}^{\lambda}\sum_{j,k}\psi_{\sigma j}(x)\,d\mu^{jk}(\sigma)\overline{\psi_{\sigma k}(y)} \tag{2.23}$$

is the distribution kernel of the projection operator E_λ in the abstract spectral decomposition of H,

$$H=\int_{-\infty}^{\infty}\lambda\,dE_\lambda. \tag{2.24}$$

More generally, functions of H are defined from functions of a real variable by

$$F(H)=\int_{-\infty}^{\infty} F(\lambda)\,dE_\lambda. \tag{2.25a}$$

It follows that the distribution kernel of $F(H)$ is

$$\int_{-\infty}^{\infty} F(\lambda)\sum_{j,k}\psi_{\lambda j}(x)\,d\mu^{jk}(\lambda)\overline{\psi_{\lambda k}(y)}. \tag{2.25b}$$

For some F's this converges to a genuine integral kernel. For example, the previously mentioned resolvent kernel is

$$G_z(x,y)=\int_{-\infty}^{\infty}\frac{1}{\lambda-z}\sum_{j,k}\psi_{\lambda j}(x)\,d\mu^{jk}(\lambda)\overline{\psi_{\lambda k}(y)}, \tag{2.26}$$

and the heat kernel is

$$K(t,x,y)=\int_{-\infty}^{\infty} e^{-\lambda t}\sum_{j,k}\psi_{\lambda j}(x)\,d\mu^{jk}(\lambda)\overline{\psi_{\lambda k}(y)}. \tag{2.27}$$

Also, $F\equiv 1$ yields a representation of the distribution kernel of the identity operator, often called the "completeness relation":

$$\delta(x-y)=\int_{-\infty}^{\infty}\sum_{j,k}\psi_{\lambda j}(x)\,d\mu^{jk}(\lambda)\overline{\psi_{\lambda k}(y)}. \tag{2.28}$$

The formal generalization of the eigenfunction expansion to the multicomponent case is obvious: one introduces a basis of classical solutions defined by the data

$$\begin{aligned}\psi_{\lambda\alpha 0}(x_0)&=e_\alpha, & \psi'_{\lambda\alpha 0}(x_0)&=0,\\ \psi_{\lambda\alpha 1}(x_0)&=0, & \psi'_{\lambda\alpha 1}(x_0)&=e_\alpha,\end{aligned} \tag{2.29}$$

where $\{e_\alpha\}$ is the natural basis of $\mathbb{C}^r$:

$$(e_\alpha)^\beta=\delta_\alpha^\beta, \qquad \alpha,\beta=1,\cdots,r. \tag{2.30}$$

Then in (2.17)–(2.18) and all the formulas following them, f is an element of $L^2(M;\mathbb{C}^r)$, $\tilde{f}_k$ is also $\mathbb{C}^r$-valued, $\mu^{jk}(\lambda)$ is an $r\times r$ matrix for each j,k, and λ (selfadjoint but not necessarily real), etc.; all Greek indices are suppressed. The literature does not seem to hold a full, explicit treatment of this problem, parallel to the work of Weyl, Titchmarsh, and Kodaira in the scalar case. Kodaira [44, §7] treats the case of $E(x)$ real; Naimark [53, Chapt. 3] allows complex coefficients but treats only a finite interval with boundary conditions of regular Sturm-Liouville type; Maurin [49, §2.7] outlines a general treatment of complex coefficients which should be adaptable to the second-order vector case. However, given that H is selfadjoint, validity of (2.18) for *some* measures $d\mu^{jk}$ is guaranteed by (a) the "rigged Hilbert space" version of the spectral theorem [14], [49], (b) elliptic regularity, and (c) the elementary existence and uniqueness theorem for second-order ordinary differential equations.

The situation becomes less clear (and more interesting) when $m>1$—that is, H is a partial differential operator. If H is elliptic, (a) and (b) again ensure that any f in the appropriate L^2 space can be expanded as an integral over classical eigenfunctions of H. Also, the spectral-projection kernel $E_\lambda(x,y)$ exists [37] and can be related to the heat kernel much as in the section which follows. However, the combination of these two elements into an eigenfunction expansion theory specifically adapted to a point x_0 has been blocked by uncertainty as to the proper analogue of (c). Ideally, one would like a

basis for the classical solutions consisting of functions with definite behaviors at x_0. A step toward a resolution of this problem has been taken in [27].

3. Spectral asymptotics and the heat kernel. Complete knowledge of the spectral projection $E_\lambda(x,y)$ in (2.23) (or of its differential with respect to λ) would enable one to calculate any function of the operator H by (2.25). A more modest goal is knowledge of the spectral measures, or, equivalently, of the values of $E_\lambda(x,y)$ and its derivatives at coincident arguments:

$$E_\lambda(x_0,x_0)=\mu^{00}(\lambda), \qquad \frac{\partial E_\lambda}{\partial y}(x_0,x_0)=\mu^{01}(\lambda),$$

$$\frac{\partial E_\lambda}{\partial x}(x_0,x_0)=\mu^{10}(\lambda), \qquad \frac{\partial^2 E}{\partial x\,\partial y}(x_0,x_0)=\mu^{11}(\lambda). \tag{3.1}$$

(Recall that these objects are $r\times r$ matrices in general.) This information would allow $E_\lambda(x,y)$ itself to be constructed by (2.23), given complete knowledge of the classical solutions. Moreover, knowledge of the measures suffices for study of the behavior, as x and y approach x_0, of the integral kernels of various functions of H.

The Titchmarsh–Kodaira formula for $\mu^{jk}(\lambda)$ (based on (2.26)), although exact, is difficult to evaluate in practice, because it requires precise global knowledge of the classical solutions. (One must know how to express as linear combinations of $\psi_{\lambda 0}$ and $\psi_{\lambda 1}$ those solutions of $H\psi=\lambda\psi$ which at an endpoint of M are square integrable (limit-point case) or satisfy the self-adjointness boundary condition (limit-circle case).) However, information about the behavior of the spectral measures at x_0 as $\lambda\to+\infty$ can be obtained purely from the local behavior of $E(x)$ at x_0. This asymptotic analysis is related via (2.27) to the well known expansion (1.8) of the heat kernel, K. (Another Green function, such as the resolvent kernel, could also be used (cf. [4], [15], [24], [29], [38], [46], [56]), but K has simple properties which make it especially convenient [32].) One clearly has

$$\begin{aligned}
K(t,x_0,x_0)&=\int_{-\infty}^{\infty} e^{-\lambda t}\,d\mu^{00}(\lambda),\\
\frac{\partial K}{\partial y}(t,x_0,x_0)&=\int_{-\infty}^{\infty} e^{-\lambda t}\,d\mu^{01}(\lambda),\\
\frac{\partial K}{\partial x}(t,x_0,x_0)&=\int_{-\infty}^{\infty} e^{-\lambda t}\,d\mu^{10}(\lambda),\\
\frac{\partial^2 K}{\partial x\,\partial y}(t,x_0,x_0)&=\int_{-\infty}^{\infty} e^{-\lambda t}\,d\mu^{11}(\lambda).
\end{aligned} \tag{3.2}$$

Any singularity in the quantities on the left-hand sides of (3.2) as $t\to 0$ must come from the large-λ parts of the $d\mu^{jk}(\lambda)$, since for finite A,

$$\int_{-\infty}^{A} e^{-\lambda t}\,d\mu^{jk}(\lambda)$$

is an integral over a compact set. (Recall that $\sigma(H)$ is bounded below.)

We begin by establishing a formal relationship between series expansions of the quantities in (3.2). Suppose that $d\mu^{00}(\lambda)$ for $\lambda>0$ can be expressed in the form

$$d\mu^{00}(\lambda)=\frac{1}{\pi}\left[\sum_{n=0}^{m}\rho_n^{00}(x_0)\omega^{-2n}+\tilde{\rho}_m^{00}(\omega)\right]d\omega, \tag{3.3}$$

where, as before, $\omega=\lambda^{1/2}$, and dependence of μ^{00} and $\tilde{\rho}^{00}$ on x_0 is notationally suppressed. It is not assumed that $\tilde{\rho}_m^{00}(\omega)=O(\omega^{-2(m+1)})$—that is false, in general—but rather that the contribution of $\tilde{\rho}_m^{00}$ to the integral

$$(-1)^m\frac{\partial^m K}{\partial t^m}(t,x_0,x_0)=\int_{-\infty}^{\infty}e^{-\lambda t}\lambda^m\,d\mu^{00}(\lambda) \tag{3.4}$$

is bounded as $t\downarrow 0$. This condition will uniquely determine the ρ_n^{00} in terms of the a_n in (1.8). (Here m does not mean the dimension of M.)

Substituting (3.3) into (3.4), one obtains

$$(-1)^m\frac{\partial^m K}{\partial t^m}=\frac{1}{\pi}\int_0^{\infty}\sum_{n=0}^{m}\rho_n^{00}\omega^{2(m-n)}e^{-\omega^2 t}\,d\omega+\text{bounded term}.$$

The formula

$$\frac{1}{\pi}\int_0^{\infty}\omega^{2p-1}e^{-\omega^2 t}\,d\omega=\frac{\Gamma(p)}{2\pi}t^{-p} \tag{3.5}$$

is valid for all real $p>0$. Thus one has

$$(-1)^m\frac{\partial^m K}{\partial t^m}=\frac{1}{2\pi}\sum_{n=0}^{m}\Gamma\left(m-n+\frac{1}{2}\right)t^{-(m-n+1/2)}\rho_n^{00}(x_0)$$
$$+\text{bounded term},$$

and after repeated integration,

$$K(t,x_0,x_0)=\frac{1}{2\pi}\sum_{n=0}^{m}\Gamma\left(\frac{1}{2}-n\right)t^{n-1/2}\rho_n^{00}(x_0)+O(t^m)+\text{polynomial in } t,$$

where m is arbitrary. In fact, given (1.8), one knows that the polynomial vanishes:

$$K(t,x_0,x_0)\sim\frac{1}{2\pi}\sum_{n=0}^{\infty}\Gamma\left(\frac{1}{2}-n\right)t^{n-1/2}\rho_n^{00}(x_0), \tag{3.6}$$

$$\begin{aligned}a_n(x_0)&=\pi^{-1/2}\Gamma\left(\frac{1}{2}-n\right)\rho_n^{00}(x_0)\\&=\frac{(-1)^n 2^n}{(2n-1)!!}\rho_n^{00}(x_0)\qquad(p!!\equiv p(p-2)\cdots 3\cdot 1).\end{aligned} \tag{3.7}$$

This is the result one would have obtained by the formal procedure of substituting

$$d\mu^{00}(\lambda)\dot{\sim}\frac{1}{\pi}\sum_{n=0}^{\infty}\rho_n^{00}(x_0)\omega^{-2n}\,d\omega \tag{3.8}$$

into the first equation (3.2) and integrating by means of the analytic continuation of (3.5) to negative p, ignoring the divergences at $\omega\downarrow 0$ and the contribution of negative λ (if any).

The point of this calculation is that, whatever the analytical status of the series (3.8) may be, the coefficients in it can be uniquely defined by comparison with those of odd powers of $t^{1/2}$ in a series for $K(t,x_0,x_0)$, and vice versa. The actual asymptotic validity of (1.8) and the absence therefrom of positive integral powers of t played no essential role. The introduction of t-derivatives of K is a calculational device, to avoid insertion of a positive lower cutoff on the integration over ω and thus to permit use of (3.5). Calculations for the other three spectral measures are identical; however, it turns out (cf. (2.21) and (3.10)) that the leading terms of $d\mu^{11}$ and $\partial^2 K/\partial x\partial y$ are one order

more singular than those of the series just considered. With allowance for that, the result can be stated:

THEOREM 3.1. *Let* $\kappa=0$ *if either of* j *and* k *is* 0; $\kappa=1$ *if* $j=k=1$. *Consistency between the series*

$$d\mu^{jk}(\lambda)\sim\frac{1}{\pi}\sum_{n=0}^{\infty}\rho_n^{jk}(x_0)\omega^{2\kappa-2n}d\omega \tag{3.9}$$

and

$$\left(\frac{\partial}{\partial x}\right)^j\left(\frac{\partial}{\partial y}\right)^k K(t,x_0,x_0)\sim\frac{1}{\sqrt{4\pi}}\sum_{n=0}^{\infty}a_n^{jk}(x_0)t^{n-1/2-\kappa} \tag{3.10}$$

($j,k=0,1$) *requires that*

$$\begin{aligned} a_{n+\kappa}^{jk}&=\frac{(-1)^n2^n}{(2n-1)!!}\rho_{n+\kappa}^{jk} \qquad (n>0),\\ a_\kappa^{jk}&=\rho_\kappa^{jk}, \qquad a_0^{11}=\frac{1}{2}\rho_0^{11}. \end{aligned} \tag{3.11}$$

More precisely, (3.11) *is the unique relationship such that the proposition*

> $(\partial/\partial x)^j(\partial/\partial y)^k K(t,x_0,x_0)$ *equals* $\tilde{K}_{jk}^{(m)}(t,x_0,x_0)+O(t^m)$, *at least modulo a* C^∞ *function of* t, *where* $\tilde{K}_{jk}^{(m)}$ *is defined by* (3.2) *with* $d\mu^{jk}(\lambda)$ *replaced by the first* $m+\kappa$ *terms of* (3.9) *for all* $\lambda>A>0$

is equivalent to the proposition

> (3.10) *is an asymptotic series, at least modulo a* C^∞ *function.*

It is clear that this construction of a correspondence between expansions of derivatives of K and expansions of derivatives of E_λ (see (3.1)) can be carried out on a manifold of any dimension m. (When $m>1$, complete local information about the spectral decomposition requires derivatives of E_λ of order higher than 2, and these derivatives are not all independent. This leads to the "local basis" problem [27] mentioned at the end of §2.) When m is odd, the heat expansion contains only odd powers of $t^{1/2}$, so one knows that the undetermined C^∞ functions allowed by the propositions in Theorem 3.1 actually vanish to infinite order as $t\downarrow 0$. When m is even, the expansion of K contains only integral powers of t. The positive powers are C^∞ functions and hence are not determined by the expansion of E_λ. Furthermore, by considering t-derivatives of K as above, one sees that, for example, $\rho_n^{00}=0$ for $n\geq m/2$ when m is even: A nonzero value of ρ_n^{00} would force a logarithmic term to be present in the expansion of K. These phenomena have been noted by Duistermaat and Guillemin [24, §2]. The boundary terms in the integrated heat kernel [4], [16], [22], [34], [39] involve both integral and half-integral powers of t, so that the problem of positive integral powers then arises even for odd m. The coefficients of such powers, although they are important spectral invariants, are hard to relate directly to the asymptotic behavior of the eigenfunctions of H. (They can be related to indefinite integrals, or Riesz means, of E_λ, which contain information about the spectral decomposition at small λ as well as large.) However, integral-order terms in the expansion of K (on the manifold M) can be related to half-integral-order terms in the expansion of the heat kernel of a related operator on the manifold $M\times N$, where N is some manifold of odd dimension ([32, Lemma 1.6]; cf. also [59, §5]). Consequently, once a generalization of the construction of this paper to arbitrary dimension has been attained, it will be

possible to relate the coefficients of integral order to the spectral asymptotics of a new problem obtained by adding a trivial extra dimension to the old one.

In what senses are the series (3.9) and (3.10) actually valid? In the latter case a very strong result is known for arbitrary dimension m [52], [51], [56], [34], [2], [31], [32], [59]. For $m=1$ and an operator in the normal form (2.1) (cf. [32]) it may be stated:

THEOREM 3.2. *Let*

$$H=-\frac{d^2}{dx^2}-E(x)$$

be the local representation in normal form of a selfadjoint operator, with C^∞ coefficients, on sections of a vector bundle with base dimension 1 and fiber dimension r. Define the "order" of a (formal, noncommutative) monomial in E and its derivatives to be twice its degree, plus the total number of differentiation operations involved. Assume that the spectrum $\sigma(H)$ is bounded below. Define $K(t,x,y)$ as the integral kernel of the operator e^{-tH} $(t>0)$. Then at any point x_0 interior to M, K and its derivatives with respect to x and y possess asymptotic expansions as $t\downarrow 0$ of the form (3.10) *for $j,k=0,1$ and of similar form for higher derivatives. The t-derivatives of these quantities have the expansions obtained by term-by-term differentiation of those expansions. Each matrix coefficient $a_n^{jk}(x_0)$ is a polynomial in $E(x_0)$, $E'(x_0)$, $E''(x_0),\cdots$, homogeneous in order: In particular, a_n^{00} and a_n^{11} are of order $2n$ and $a_n^{10}=(a_n^{01})^*$ is of order $2n+1$. When M is compact and without boundary (i.e., a circle), the expansion is uniform in x_0.*

To the proofs in the references we need to add justification of the claims about the *derivatives* of K. There are basically two methods of deriving the expansion of K. In one (see particularly [59]), one obtains, more or less explicitly, an asymptotic expansion (as $t\downarrow 0$) of $K(t,x,y)$, valid for x close to but not necessarily equal to y, of the form

$$K(t,x,y)\sim(4\pi t)^{-m/2}e^{-d(x,y)^2/4t}\sum_{n=0}^{\infty}A_n(x,y)t^n, \tag{3.12}$$

where $d(x,y)$ is the distance from x to y in the metric defined by the principal symbol (second-order derivative terms) of H, and all the covariant derivatives of A_n at $x=y$ are calculable local polynomials in derivatives of the coefficient functions in H. Differentiating (3.12) and setting $x=y=x_0$, one obtains expansions of type (3.10); and it is clear from the proofs ([59]; [28, App. A]) that each series is genuinely asymptotic to the corresponding derivative of K at x_0. The other method uses pseudodifferential operators [56], [30], [31]. (It is not absolutely clear from the references that the arguments apply to noncompact M, hence our preference for the method of [59].) In that approach it is awkward to calculate $K(t,x,y)$ for $x\neq y$, but the values of derivatives of K on the diagonal $(x=y)$ are easy to obtain. It is obvious that

$$\frac{\partial^{i+j+k}}{\partial t^i\partial x^j\partial y^k}K(t,x,y)$$

is the kernel of the operator

$$\left(\frac{\partial}{\partial x}\right)^j\circ(-H)^ie^{-Ht}\circ\left(-\frac{\partial}{\partial y}\right)^k.$$

The symbol of this operator is easily obtained by the pseudodifferential symbol calculus, and the rest of the argument proceeds as for K itself, except that if $j+k$ is odd, then only odd orders, rather than only even ones, in the expansion of the symbol of $(H-\lambda)^{-1}$ make nonvanishing contributions to the result.

Consider now the series (3.9) (which is well defined by virtue of Theorem 3.2). The leading term of its integrated form is a genuine asymptotic approximation to $\mu^{jk}(\lambda)$ as $\lambda \to +\infty$; this is a consequence of Karamata's Tauberian theorem [40], [22] (applied, in the vector case, to all diagonal matrix elements of the operators involved). But it is already obvious from Example 2.2 that the series cannot be asymptotic beyond that first term.

The analogous failure of the higher-order terms in a formal series for the asymptotic distribution of eigenvalues ($N(\lambda) \equiv$ number of eigenvalues less than λ) of the Laplacian on an m-dimensional domain is well known [3], [35], [4], [8]. Several authors have reinterpreted that series as being valid in some "averaged" sense: Brownell [15], [16] showed that the difference between the true $N(\lambda)$ and any of the first few truncations of the series became small (to the appropriate order) after being subjected to a certain "log-Gaussian" averaging. He also showed that if a genuinely asymptotic approximation to $N(\lambda)$ by powers of $\lambda^{1/2}$ existed, then its coefficients would be uniquely determined by the heat-kernel expansion (to the extent the latter was known at that time). Balian and Bloch [4] defined an averaged eigenvalue density, $\int \rho(\lambda - \lambda')\, dN(\lambda')$ ($\rho(\lambda)$ being sharply peaked around $\lambda = 0$), and chose ρ so that the procedure was equivalent to studying $G_z(x,y)$ with $\operatorname{Im} z > 0$ (cf. (2.26)) in a context where taking a limit to the real axis would be necessary to obtain the true spectral decomposition. In these papers the higher-order terms in question are those proportional to the surface area of the boundary, etc., which are not directly related to our local expansions in the interior with a nonvanishing potential; the issues raised are the same, however.

In [47], [48] Levitan investigated the asymptotic behavior of "Riesz means" of the spectral function $E_\lambda(x,y)$. His results were greatly extended by Hörmander [37], [38]. More recently, Duistermaat and Guillemin [24] have proved the existence of an asymptotic series for a quantity defined from the spectral density distribution $d\mu^{00}/d\omega$ by smoothly annihilating all but its lowest-frequency Fourier components. This series obviously must coincide with our (3.8) when $m = 1$.

The attitude adopted here toward series of the type (3.9) is a pragmatic one: One is interested in [approximately] evaluating integrals such as (2.25b), and one seeks to prove, for as large a class of integrands as possible, that the contribution to the integral from its infinite upper end can be calculated by replacing $d\mu^{jk}(\lambda)$ at large λ by (a truncation of) its series. It has been argued in §1 that the class is not devoid of interesting functions. In general, the approximation is expected to be valid when the integrand is slowly varying, so that the oscillations (see remark below) in the difference between the true spectral density and the approximation will tend to make no net contribution to the integral. In particular, the series is useful in obtaining the asymptotic behavior of integral kernels of functions of K when $x \simeq y \simeq x_0$.

The error committed in replacing a spectral density by a truncation of its averaged asymptotic series is an oscillatory function or distribution. This is clear from (1) explicit examples such as (2.22), (2) the contrapositive of Karamata's Tauberian theorem, which indicates that the distribution cannot be nonnegative (i.e., the error in $\mu^{jk}(\lambda)$ cannot be monotonic) or (3) the indications from the cited references that the error is wiped out (to the appropriate order) by various kinds of averaging. Some progress has been made in relating this remainder function (especially the periods of its oscillations) to the global geometry of M [8], [24].

4. Potentials of compact support; calculation of the spectral measures in the scalar case. The main point of Theorems 3.1 and 3.2 is that the coefficients $a_n^{jk}(x_0)$, and hence the $\rho_n^{jk}(x_0)$, are completely determined by the behavior of the potential $E(x)$ *at*

the point x_0, independently of boundary conditions, manifold topology, and the values of $E(x)$ outside the neighborhood of x_0. To compute them, therefore, it suffices to study any potential $\tilde{E}$, on a manifold $\tilde{M}$, which coincides with E in a neighborhood of x_0 where M and $\tilde{M}$ can be locally identified. It is convenient to choose $\tilde{M}=\mathbb{R}$ and $\tilde{E}$ a C^∞ function of compact support. We assume this has been done, and henceforth write E instead of $\tilde{E}$, etc.

One is now confronted by the "scattering problem" of wave mechanics, with an extremely well-behaved potential. The normalization of the eigenfunction expansion in this case is well known: Every positive number λ is in the continuous spectrum of H, with multiplicity $2r$. There may also be some eigenvalues $\lambda_\nu \le 0$, with normalized eigenvectors ϕ_ν. Let $\{u_\alpha\}$ $(\alpha=1,\cdots,r)$ be an orthonormal basis for $\mathbb{C}^r$, and let $p=\pm\omega$ stand for one of the square roots of λ. Then for each choice of α and the sign of p, there exists a classical solution of $H\phi=p^2\phi$ with the behavior

$$\phi_{p\alpha}(x)\sim\begin{cases} u_\alpha e^{ipx}+R_{p\alpha}e^{-ipx} & \text{as } x\to-(\operatorname{sgn}p)\infty, \\ T_{p\alpha}e^{ipx} & \text{as } x\to(\operatorname{sgn}p)\infty \end{cases} \tag{4.1}$$

for some $R_{p\alpha}$ and $T_{p\alpha}$ in $\mathbb{C}^r$. The spectral decomposition (2.24) in kernel form is then

$$\begin{aligned} H(x,y)=\frac{1}{2\pi}\int_{-\infty}^{\infty} p^2 \sum_\alpha \phi_{p\alpha}(x)\otimes\phi_{p\alpha}(y)^*\,dp \\ +\sum_\nu \lambda_\nu\phi_\nu(x)\otimes\phi_\nu(y)^*. \end{aligned} \tag{4.2}$$

By manipulating this expression into the form (2.25b), one can find the spectral measures of H at x_0 (see (4.6) ff.).

The normalization of eigenfunctions shown in (4.1)–(4.2) can be obtained from that in the Fourier transform by studying the evolution of localized initial data under the time-dependent Schrödinger equation (or the wave equation) by the method of stationary phase; see any reasonably sophisticated textbook of quantum mechanics, such as [33, p. 102]. Alternatively, (4.2) can be derived by applying the Titchmarsh–Kodaira theory with any point outside the support of E in the role of x_0.

To obtain explicitly the spectral measures, one must express the $\phi_{p\alpha}$ in terms of the $\psi_{\lambda\alpha j}$ of (2.29); hence, one must calculate $\phi_{p\alpha}(x_0)$ and $\phi'_{p\alpha}(x_0)$. For large λ this can be done to arbitrary order by the well known phase-integral or WKB method. In the rest of this paper we restrict to the case $r=1$ and employ the version of the WKB expansion developed by Fröman [25] and (computationally) by Campbell [17] (see also [19]):

THEOREM 4.1. *There exist local functionals of* $E(x)$ *and its derivatives,* $Y_{2n}(x)$, *such that if* $E\in C_0^\infty(R)$ *and* $|p|\equiv\omega$ *is sufficiently large, then there exists a solution of*

$$-\psi''(x)-E(x)\psi(x)=\omega^2\psi(x) \tag{4.3}$$

such that for any n_0 (*and uniformly in* x)

$$\begin{aligned} \psi(x)=\left[\sum_{n=0}^{n_0} p^{-2n}Y_{2n}(x)\right]^{-1/2}\exp\left[ip\int_{x_0}^{x}\sum_{n=0}^{n_0} p^{-2n}Y_{2n}(x')\,dx'\right] \\ +O(\omega^{-2n_0-1}). \end{aligned} \tag{4.4}$$

(*Term-by-term differentiation of* (4.4) *is valid.*) *Every solution can be approximated in the same sense by a linear combination of these two expressions* ($p=\pm\omega$). *The first few coefficients are*

$$
\begin{aligned}
&\text{(a)} \quad Y_0=1,\\
&\text{(b)} \quad Y_2=\frac{1}{2}E,\\
&\text{(c)} \quad Y_4=-\frac{1}{8}(E''+E^2),\\
&\text{(d)} \quad Y_6=\frac{1}{32}[E^{(4)}+6EE''+5(E')^2+2E^3],\\
&\text{(e)} \quad Y_8=-\frac{1}{128}[E^{(6)}+10EE^{(4)}+28E'E^{(3)}+19(E'')^2\\
&\qquad\qquad +30E^2E''+50E(E')^2+5E^4],\\
&\text{(f)} \quad Y_{10}=\frac{1}{512}[E^{(8)}+14EE^{(6)}+54E'E^{(5)}+110E''E^{(4)}+70E^2E^{(4)}\\
&\qquad +69(E^{(3)})^2+392EE'E^{(3)}+266E(E'')^2+442(E')^2E''\\
&\qquad\qquad +140E^3E''+350E^2(E')^2+14E^5],\\
&\text{(g)} \quad Y_{12}=-\frac{1}{2048}[E^{(10)}+18EE^{(8)}+88E'E^{(7)}+238E''E^{(6)}\\
&\qquad +126E^2E^{(6)}+418E^{(3)}E^{(5)}+972EE'E^{(5)}+251(E^{(4)})^2\\
&\qquad +1980EE''E^{(4)}+1630(E')^2E^{(4)}+420E^3E^{(4)}\\
&\qquad +1242E(E^{(3)})^2+5564E'E''E^{(3)}+3528E^2E'E^{(3)}\\
&\qquad +1262(E'')^3+2394E^2(E'')^2+7956E(E')^2E''+630E^4E''\\
&\qquad\qquad +1105(E')^4+2100E^3(E')^2+42E^6].
\end{aligned}
\tag{4.5}
$$

Remark 4.1. Y_{14}, Y_{16}, Y_{18}, Y_{20} can be obtained from formulas in [17] by the notational identification $\varepsilon_S\equiv E^{(S)}$. The expression Y_{20} contains 137 terms, whose rational coefficients have numerators containing as many as 9 digits.

Remark 4.2: Theorem 4.1 is a special case of a theorem proved in a sequel to this paper. To obtain it from [25], [17], note that those papers treat the equation

$$\frac{d^2\phi}{dz^2}+p^2Q^2(z)\phi=0,$$

which is converted to (4.3), with E equal to the ε_0 of [25], [17], by the transformation

$$\phi=Q^{1/2}\psi, \qquad \frac{dx}{dz}=Q.$$

(Also, we have rescaled the ε_s and Y_s by powers of p.)

COROLLARY 4.1. *For* $E\in C_0^\infty(R)$, *the quantities* R_p *and* $|T_p|-1$ *in* (4.1) *vanish to all orders as* $|p|\to\infty$.

Corollary 4.1 is just the observation that the eigenfunctions for $\lambda=p^2$ can be taken to have *globally* the behavior indicated by (4.4), as opposed to that of a linear combination of $e^{\pm i\omega\cdots}$. This is crucial to the calculation below. Both smoothness and the absence of boundaries are essential for this result. Clearly, an operator on the half-line with, say, Dirichlet boundary conditions will have eigenfunctions whose reflected waves equal their incident waves in strength. An exactly solvable, discontinuous "square barrier" potential [50, §3.7] yields R_p and T_p satisfying

$$|R_p|^2=1-|T_p|^2=O(p^{-4}),$$

the exponent being best possible. Such finite order mixing of positive-phase and negative-phase waves is the origin of the oscillatory terms in the spectral measures. Finally, note that because $E(x)$ is bounded, for sufficiently large λ there are no "turning points"; this also is essential for the corollary and, along with the compact support of E, for the uniformity of (4.4) in x. In more general contexts, Theorem 4.1 weakens to a merely local statement, valid in finite neighborhoods where E is smooth.

From (4.2) (with $r=1$) and the formalism of §2 one has

$$\begin{aligned}&\sum_{j,k}\psi_{\lambda j}(x)\,d\mu^{jk}(\lambda;x_0)\overline{\psi_{\lambda k}(y)}\\ &\quad\equiv dE_\lambda(x,y)=\frac{1}{2\pi}\left[\phi_\omega(x)\overline{\phi_\omega(y)}+\phi_{-\omega}(x)\overline{\phi_{-\omega}(y)}\right]d\omega,\end{aligned}\tag{4.6}$$

and thus

$$d\mu^{00}(\lambda)=dE_\lambda(x_0,x_0)=\frac{1}{2\pi}\left[|\phi_\omega(x_0)|^2+|\phi_{-\omega}(x_0)|^2\right]d\omega,\tag{4.7}$$

$$d\mu^{10}(\lambda)=\frac{1}{2\pi}\left[\phi'_\omega(x_0)\overline{\phi_\omega(x_0)}+\phi'_{-\omega}(x_0)\overline{\phi_{-\omega}(x_0)}\right]d\omega,\tag{4.8}$$

and so on. We evaluate (4.7): Set the x_0 in (4.4) equal to the x_0 at which the spectral measures are to be defined. Then

$$|\phi_{\pm\omega}(x_0)|^2\sim\left[\sum_{n=0}^{\infty}\omega^{-2n}Y_{2n}(x_0)\right]^{-1},\tag{4.9}$$

where the right-hand side stands for a series obtainable from (4.5) by the geometric-series expansion. In this way it is fairly easy to calculate by hand the first 7 coefficients in (3.8):

$$\begin{aligned}&\text{(a)}\quad \rho_0^{00}=1,\\ &\text{(b)}\quad \rho_1^{00}=-\frac{1}{2}E,\\ &\text{(c)}\quad \rho_2^{00}=\frac{1}{8}(E''+3E^2),\\ &\text{(d)}\quad \rho_3^{00}=-\frac{1}{32}[E^{(4)}+10EE''+5(E')^2+10E^3],\\ &\text{(e)}\quad \rho_4^{00}=\frac{1}{128}[E^{(6)}+14EE^{(4)}+28E'E^{(3)}+21(E'')^2+70E^2E''\\ &\qquad\qquad +70E(E')^2+35E^4],\\ &\text{(f)}\quad \rho_5^{00}=-\frac{1}{512}[E^{(8)}+18EE^{(6)}+54E'E^{(5)}+114E^{(2)}E^{(4)}\\ &\qquad +126E^2E^{(4)}+69(E^{(3)})^2+504EE'E^{(3)}+462(E')^2E^{(2)}\\ &\qquad\qquad +378E(E^{(2)})^2+420E^3E^{(2)}+630E^2(E')^2+126E^5],\\ &\text{(g)}\quad \rho_6^{00}=\frac{1}{2048}[E^{(10)}+22EE^{(8)}+88E'E^{(7)}+242E''E^{(6)}\\ &\qquad +198E^2E^{(6)}+418E^{(3)}E^{(5)}+1188EE'E^{(5)}+253(E^{(4)})^2\\ &\qquad +2508EE''E^{(4)}+1650(E')^2E^{(4)}+924E^3E^{(4)}+1518E(E^{(3)})^2\\ &\qquad +5676E'E''E^{(3)}+5544E^2E'E^{(3)}+1342(E'')^3+4158E^2(E'')^2\\ &\qquad +10164E(E')^2E''+2310E^4E''+1155(E')^4+4620E^3(E')^2\\ &\qquad\qquad +462E^6].\end{aligned}\tag{4.10}$$

(Of course, in such equations everything is evaluated at x_0.) For higher orders, computer assistance would be almost a necessity.

From (3.6)–(3.7) and (4.10) one obtains the first 7 terms of the expansion of the heat kernel. The result agrees with [32, Thm. 2.2] and goes two orders beyond it. Also, (4.10f) can be shown to agree with the corresponding term in the expansion of the resolvent kernel given by Gel'fand and Dikii [29]. Incidentally, the recursion relation derived and used in [29] appears to be the most efficient systematic method of generating the coefficients ρ_n^{00} when $m=r=1$; unfortunately, it does not seem to have a generalization to higher base or fiber dimension.

The other measures are calculated similarly. With

$$Y(x)\equiv\sum_{n=0}^{n_0} p^{-2n}Y_{2n}(x), \tag{4.11}$$

the derivative of (4.4) is

$$\psi'=\left(ipY^{1/2}-\frac{1}{2}Y^{-3/2}Y'\right)e^{ipfY}. \tag{4.12}$$

Then (4.8) can be reduced to

$$d\mu^{10}(\lambda)\sim\frac{1}{2\pi}\frac{d}{dx_0}Y^{-1}, \tag{4.13a}$$

where the series to be differentiated is that already encountered in (4.9). It follows that

$$\rho_n^{10}=\rho_n^{01}=\frac{1}{2}(\rho_n^{00})'. \tag{4.13b}$$

(In particular, $\rho_0^{10}=0$.) From (3.1) we see that (4.13b) reflects the fact that

$$\frac{d}{dx_0}E_\lambda(x_0,x_0)=\frac{\partial E_\lambda}{\partial x}(x_0,x_0)+\frac{\partial E_\lambda}{\partial y}(x_0,x_0). \tag{4.13c}$$

The formula (4.12) also yields

$$d\mu^{11}(x)\sim\frac{p^2}{\pi}Y\left[1+\frac{1}{4}p^{-2}\left(\frac{d}{dx_0}Y^{-1}\right)^2\right]d\omega, \tag{4.14a}$$

so this expansion is easily obtainable from the other two:

$$\rho_n^{11}=Y_{2n}+\frac{1}{4}\sum_{m=2}^{n-1}\sum_{l=1}^{m-1}(\rho_l^{00})'(\rho_{m-l}^{00})'Y_{2(n-m-1)}. \tag{4.14b}$$

The results to order 6 are

(a) $\rho_n^{11}=Y_{2n}$ for $n=0,1,2$ (see (4.5)),

(b) $\rho_3^{11}=\frac{1}{32}[E^{(4)}+6EE''+7(E')^2+2E^3]$,

(c) $\rho_4^{11}=-\frac{1}{128}[E^{(6)}+10EE^{(4)}+32E'E^{(3)}+19(E'')^2+30E^2E''+70E(E')^2+5E^4]$,

(d) $$\rho_5^{11}=\frac{1}{512}[E^{(8)}+14EE^{(6)}+58E'E^{(5)}+110E''E^{(4)}+70E^2E^{(4)}+71(E^{(3)})^2+448EE'E^{(3)}+266E(E'')^2+518(E')^2E''+140E^3E''+490E^2(E')^2+14E^5], \tag{4.15}$$

(e) $$\begin{aligned}\rho_6^{11}=-\frac{1}{2048}[&E^{(10)}+18EE^{(8)}+92E'E^{(7)}+238E''E^{(6)}\\&+126E^2E^{(6)}+422E^{(3)}E^{(5)}+1044EE'E^{(5)}+251(E^{(4)})^2\\&+1980EE''E^{(4)}+1794(E')^2E^{(4)}+420E^3E^{(4)}+1278E(E^{(3)})^2\\&+5916E'E''E^{(3)}+4032E^2E'E^{(3)}+1262(E'')^3+2394E^2(E'')^2\\&+9324E(E')^2E''+630E^4E''+1365(E')^4+2940E^3(E')^2+42E^6].\end{aligned}$$

In summary:

THEOREM 4.2. *Let*

$$H=-\frac{d^2}{dx^2}-E(x)$$

be a selfadjoint operator on M. The spectral measures of H relative to a point x_0 interior to M have the averaged asymptotic expansion (3.9), *with the coefficients listed in* (4.10), (4.13b), *and* (4.15). *If $M=\mathbb{R}$ and $E\in C_0^\infty(\mathbb{R})$, the expansion is a genuine asymptotic series. In any case, the first term of the series is asymptotic.*

Finally, observe that the same technique can be used to calculate $dE_\lambda(x,y)$ for x and y distinct. Substituting (4.4) into (4.6), one obtains

$$dE_\lambda(x,y)\sim\frac{1}{\pi}\left[\sum_n \omega^{-2n}Y_{2n}(x)\right]^{-1/2}\left[\sum_m \omega^{-2m}Y_{2m}(y)\right]^{-1/2} \times\cos\left[\omega\int_y^x\sum_l \omega^{-2l}Y_{2l}(x')\,dx'\right]d\omega. \tag{4.16}$$

This expression may be more useful for some practical calculations than the tabulated values of $\rho_n^{jk}(x_0)$. It must agree with the reconstitution of $dE_\lambda(x,y)$ from the spectral measures at x_0 via (2.23); the two expressions presumably coincide when all functions of x and y in each, outside the oscillatory factors, are expanded in power series in $x-x_0$ and $y-y_0$. In view of Theorem 4.1, (4.16) provides a rigorous asymptotic approximation of arbitrary order for the large-λ part of the eigenfunction expansion when $E\in C_0^\infty(\mathbb{R})$. In other cases, since the heat-kernel expansion (3.12) is purely local, an argument parallel to Theorem 3.1 would establish that (4.16) is valid in a "mean" sense for x and y sufficiently close. What this means in practice is that if some quantity or expansion (such as the short-distance singularity of some Green function) can be proved to be local in its dependence on E (cf. [28]), then (4.16) can be used to calculate it.

Acknowledgments. I thank F. J. Narcowich for his active interest in this project, and P. Candelas, D. Deutsch, J. S. Dowker, G. Kennedy, T. A. Osborn, J. K. Shaw, and M. E. Taylor for illuminating discussions and bibliographical information.

Note added in proof. We call attention to recent related work: S. F. J. Wilk, Y. Fujiwara and T. A. Osborn, *N-body Green's functions and their semiclassical expansion*, Phys. Rev. A, 24 (1981), pp. 2187–2202; Y. Fujiwara, T. A. Osborn and S. F. J. Wilk, *Wigner–Kirkwood expansions*, Phys. Rev. A, 25 (1982), pp. 14–34.

For treatments, more rigorous than [33], of normalization of scattering eigenfunctions, see M. Reed and B. Simon, *Methods of Modern Mathematical Physics, Vol.* 3, *Scattering Theory*, Academic Press, New York, 1979, pp. 96–115 and 352–357.

REFERENCES

[1] N. I. AKHIEZER AND I. M. GLAZMAN, *Theory of Linear Operators in Hilbert Space*, Vol. 2, Ungar, New York, 1963.

[2] M. F. ATIYAH, R. BOTT AND V. K. PATODI, *On the heat equation and the index theorem*, Invent. Math., 19 (1973), pp. 279–330.

[3] V. G. AVAKUMOVIC, *Über die Eigenfunktionen auf geschlossenen Riemannschen Mannigfaltigkeiten*, Math. Z., 65 (1956), pp. 327–344.

[4] R. BALIAN AND C. BLOCH, *Distribution of eigenfrequencies for the wave equation in a finite domain*. I, Ann. Physics, 60 (1970), pp. 401–447.
[5] _____, *Asymptotic evaluation of the Green's function for large quantum numbers*, Ann. Physics, 63 (1971), pp. 592–606.
[6] _____, *Distibution of eigenfrequencies for the wave equation in a finite domain*. II, Ann. Physics, 64 (1971), pp. 271–307.
[7] _____, errata to [4], [5], and [6], Ann. Physics, 84 (1974), pp. 559–562.
[8] _____, *Distribution of eigenfrequencies for the wave equation in a finite domain*. III, Ann. Physics, 69 (1972), pp. 76–160.
[9] _____, *Solution of the Schrödinger equation in terms of classical paths*, Ann. Physics, 85 (1974), pp. 514–545.
[10] R. BALIAN AND B. DUPLANTIER, *Electromagnetic waves near perfect conductors*. I, Ann. Physics, 104 (1977), pp. 300–335.
[11] _____, *Electromagnetic waves near perfect conductors*. II, Ann. Physics, 112 (1978), pp. 165–208.
[12] H. P. BALTES AND E. R. HILF, *Spectra of Finite Systems*, Bibliographisches Institut, Mannheim, 1976.
[13] C. M. BENDER AND P. HAYS, *Zero-point energy of fields in a finite volume*, Phys. Rev. D, 14 (1976), pp. 2622–2632.
[14] JU. M. BEREZANSKII, *Expansions in Eigenfunctions of Self-adjoint Operators*, American Mathematical Society, Providence, RI, 1968.
[15] F. H. BROWNELL, *An extension of Weyl's law for eigenvalues*, Pacific J. Math., 5 (1955), pp. 483–489.
[16] _____, *Extended asymptotic eigenvalue distributions for bounded domains in n-space*, J. Math. Mech., 6 (1957), pp. 119–166.
[17] J. A. CAMPBELL, *Computation of a class of functions useful in the phase-integral approximation*. I, J. Comput. Phys., 10 (1972), pp. 308–315.
[18] P. CANDELAS, *Vacuum polarization in the presence of dielectric and conducting surfaces*, Ann. Physics, in press.
[19] B. CHAKRABORTY, *The mathematical problem of reflection solved by an extension of the WKB method*, J. Math. Phys., 14 (1973), pp. 188–190.
[20] J. CHAUDHURI AND W. N. EVERITT, *On the spectrum of ordinary second order differential operators*, Proc. Roy. Soc. Edinburgh Sect. A, 68 (1968), pp. 95–119.
[21] S. M. CHRISTENSEN, *Vacuum expectation value of the stress tensor in an arbitrary curved background: The covariant point-separation method*, Phys. Rev. D, 14 (1976), pp. 2490–2501.
[22] C. CLARK, *The asymptotic distribution of eigenvalues and eigenfunctions for elliptic boundary value problems*, SIAM Rev., 9 (1967), pp. 627–646.
[23] D. DEUTSCH AND P. CANDELAS, *Boundary effects in quantum field theory*, Phys. Rev. D, 20 (1979), pp. 3063–3080.
[24] J. J. DUISTERMAAT AND V. W. GUILLEMIN, *The spectrum of positive elliptic operators and periodic bicharacteristics*, Invent. Math., 29 (1975), pp. 39–79.
[25] N. FRÖMAN, *Outline of a general theory for higher order approximations of the JWKB-type*, Arkiv Fysik, 32 (1966), pp. 541–548.
[26] S. A. FULLING, *The local asymptotics of continuum eigenfunction expansions*, in Spectral Theory of Differential Operators, Math. Stud. 55, I. W. Knowles and R. T. Lewis, eds., North-Holland, Amsterdam, 1981, pp. 181–187.
[27] S. A. FULLING AND F. J. NARCOWICH, *A basis for the local solutions of an elliptic equation*, J. Math. Anal. Appl., 86 (1982), pp. 246–267.
[28] S. A. FULLING, F. J. NARCOWICH AND R. M. WALD, *Singularity structure of the two-point function in quantum field theory in curved spacetime*. II, Ann. Physics, 136 (1981), pp. 243–272.
[29] I. M. GEL'FAND AND L. A. DIKII, *Asymptotic behavior of the resolvent of Sturm–Liouville equations and the algebra of the Korteweg–deVries equations*, Usp. Mat. Nauk, 30:5(1975), pp. 67–100. Russian Math. Surveys 30:5(1975), pp. 77–113.
[30] P. B. GILKEY, *The Index Theorem and the Heat Equation*, Publish or Perish Inc., Boston, 1974.
[31] _____, *The spectral geometry of a Riemannian manifold*, J. Differential Geom., 10 (1975), pp. 601–618.
[32] _____, *Recursion relations and the asymptotic behavior of the eigenvalues of the Laplacian*, Compositio Math., 38 (1979), pp. 201–240.
[33] K. GOTTFRIED, *Quantum Mechanics*, Vol. I, Benjamin, Reading, MA, 1966.
[34] P. GREINER, *An asymptotic expansion for the heat equation*, Arch. Rational Mech. Anal., 41 (1971), pp. 163–218.
[35] D. GROMES, *Über die asymptotische Verteilung der Eigenwerte des Laplace-Operators für Gebiete auf der Kugeloberfläche*, Math. Z., 94 (1966), pp. 110–121.

[36] G. HELLWIG, *Differential Operators of Mathematical Physics*, Addison-Wesley, Reading, MA, 1967.
[37] L. HÖRMANDER, *On the Riesz means of spectral functions and eigenfunction expansions for elliptic differential operators*, in Belfer Graduate School of Science Annual Science Conference Proceedings: Some Recent Advances in the Basic Sciences, Vol. 2 (1965–66), A. Gelbart, ed., Yeshiva Univ., New York, 1969, pp. 155–202.
[38] _____, *The spectral function of an elliptic operator*, Acta Math., 121 (1968), pp. 193–218.
[39] M. KAC, *Can one hear the shape of a drum*? Amer. Math. Monthly, 73 (1966), pp. 1–23.
[40] J. KARAMATA, *Neuer Beweis und Verallgemeinerung einiger Tauberian-Sätze*, Math. Z., 33 (1931), pp. 294–299.
[41] G. KENNEDY, *Boundary terms in the Schwinger–DeWitt expansion: Flat space results*, J. Phys. A, 11 (1978), pp. L173–L178.
[42] G. KENNEDY, R. CRITCHLEY AND J. S. DOWKER, *Finite temperature field theory with boundaries: Stress tensor and surface action renormalization*, Ann. Physics, 125 (1980), pp. 346–400.
[43] K. KODAIRA, *The eigenvalue problem for ordinary differential equations of the second order and Heisenberg's theory of S-matrices*, Amer. J. Math., 71 (1949), pp. 921–945.
[44] _____, *On ordinary differential equations of any even order and the corresponding eigenfunction expansions*, Amer. J. Math., 72 (1950), pp. 502–544.
[45] R. S. KULKARNI, *Index Theorems of Atiyah–Bott–Patodi and Curvature Invariants*, Presses Univ. Montréal, Montreal, 1975.
[46] R. LAVINE, *The local spectral density and its classical limit*, to appear.
[47] B. M. LEVITAN, *On the asymptotic behavior of the spectral function of a self-adjoint differential equation of the second order and on expansion in eigenfunctions.* II, Izv. Akad. Nauk SSSR Ser. Mat., 19 (1955), pp. 33–58. Amer. Math. Soc. Transl. (2), 110 (1977), pp. 165–188.
[48] _____, *On the asymptotic behavior of the spectral function and expansion in eigenfunctions of the equation* $\Delta u+\{\lambda-q(x_1,x_2,x_3)\}u=0$, Trudy Moskov. Mat. Obshch., 4 (1955), pp. 237–290. Amer. Math. Soc. Transl. (2), 20 (1962), pp. 1–53.
[49] K. MAURIN, *General Eigenfunction Expansions and Unitary Representations of Topological Groups*, Polish Scientific Publishers, Warsaw, 1968.
[50] A. MESSIAH, *Quantum Mechanics*, Vol. I, North-Holland, Amsterdam, 1961.
[51] S. MINAKSHISUNDARAM, *Eigenfunctions on Riemannian manifolds*, J. Indian Math. Soc., 17 (1953), pp. 159–165.
[52] S. MINAKSHISUNDARAM AND Å. PLEIJEL, *Some properties of the eigenfunctions of the Laplace-operator on Riemannian manifolds*, Canad. J. Math., 1 (1949), pp. 242–256.
[53] M. A. NAIMARK, *Linear Differential Operators*, Ungar, New York, 1968.
[54] D. RAY, *On spectra of second-order differential operators*, Trans. Amer. Math. Soc., 77 (1954), pp. 299–321.
[55] F. RELLICH, *Spectral Theory of a Second Order Ordinary Differential Operator* (lecture notes), Institute of Mathematical Sciences, New York Univ., New York, 1953.
[56] R. T. SEELEY, *Complex powers of an elliptic operator*, in Singular Integrals, Proc. Symp. Pure Math. 10, American Mathematical Society, Providence, RI, 1967, pp. 288–307.
[57] E. C. TITCHMARSH, *Eigenfunction Expansions Associated with Second-order Differential Equations*, Part One, 2nd ed., Oxford Univ. Press, Oxford, 1962.
[58] R. M. WALD, *The back reaction effect in particle creation in curved spacetime*, Comm. Math. Phys., 54 (1977), pp. 1–19.
[59] _____, *On the Euclidean approach to quantum field theory in curved spacetime*, Comm. Math. Phys., 70 (1979), pp. 221–242.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0002 $01.00/0

A NONLINEAR INTEGRAL OPERATOR ARISING FROM A MODEL IN POPULATION GENETICS, I. MONOTONE INITIAL DATA*

ROGER LUI†

Abstract. We study the asymptotic behavior of the solutions to the recursion $u_{n+1}(x)=Q[u_n](x)$ for $n\geq 0$. Here $Q[u](x)=(K*g\circ u)(x)$ acts on functions bounded between 0 and 1, $K(x)$ is a probability density function with compact support, and $g(u)\in C^1[0,1]$ satisfies certain additional assumptions. It is known that there exists a c_+^* such that for $c\geq c_+^*$, there are nonincreasing travelling waves $w_c(x)$ facing right. We prove here that if $K(x)$ is the exponential of a concave function and $u_0(x)$ is monotone nonincreasing and decays to zero rapidly enough, then $u_n(x)$ converges in a certain sense to $w_{c_+^*}(x)$ as n approaches infinity uniformly in $\mathbb{R}$.

1. Introduction. The history of this problem goes back to two papers by R. A. Fisher [8] and by Kolmogorov, Petrowskii and Piscounoff [10], both written in 1937 on a subject we nowadays call nonlinear diffusion. They considered the initial value problem of a semilinear parabolic equation

$$u_t=u_{xx}+f(u), \qquad u(x,0)=u_0(x), \tag{1.1}$$

where $f\in C^1[0,1]$, $f(0)=f(1)=0$, $f(u)>0$ in $(0,1)$, $f'(0)>0$, $f(u)\leq f'(0)u$ in $[0,1]$ and $u_0\in[0,1]$. They were able to show that (1.1) admits a solution of the special form $w(x-ct)$ if and only if $c\geq c^*\equiv\sqrt{2f'(0)}$. A solution of the form $w(x-ct)$ is called a travelling wave of speed c.

Kolmogorov et al. went on to show a result which is more closely related to results obtained in this paper. They assumed that $u_0(x)=H(-x)$, where $H(x)$ is the Heaviside function. It is simple under such conditions on f and u_0 to show that a unique solution $u(x,t)$ of (1.1) exists for all $t>0$, $0\leq u(x,t)\leq 1$ and $u(x,t)$ is decreasing in x for each positive time t. Kolmogorov et al. proved that if the function $m_\gamma(t)$ is defined by the relation $u(m_\gamma(t),t)=\gamma$ for each $0<\gamma<1$, then

$$\lim_{t\to\infty} u\big(x+m_\gamma(t),t\big)=w(x) \quad \text{uniformly in } \mathbb{R},$$

where $w(x)$ is the travelling wave of speed c^* normalized so that $w(0)=\gamma$.

In their papers of 1975 and 1976, Aronson and Weinberger [2], [3] studied (1.1) in one and more spatial dimensions and allowed f to have an intermediate zero between 0 and 1. Their results among others imply that the c^* found in [10] is actually the asymptotic speed of propagation for any initial function $u_0(x)$ having compact support. In 1978, Uchiyama [16] extended the results of [10] to include a wide class of initial data.

Equation (1.1) when $f(u)=u(1-u)$ was actually used by Fisher [8] as a model for the spatial spread of an advantageous gene in a population living in a homogeneous one-dimensional habitat. As was already mentioned in [2], the Fisher model is based on some assumptions of doubtful validity. In 1978, Weinberger [17] proposed a more realistic model in which time occurs in discrete steps designed to simulate synchronous generations. The model is described by a recursion formula,

$$u_{n+1}=Q[u_n], \tag{1.2}$$

* Received by the editors August 17, 1981, and in revised form February 16, 1982.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455 and Department of Mathematical Sciences, San Diego State University, San Diego, California 92182.

which links the nth generation to the $(n+1)$th generation. The actual expression of Q is given by (1.3).

In [17], travelling wave solutions were found to exist for speed greater than or equal to a positive number c^*. c^* was again shown to be the asymptotic speed of propagation. The exact statement of these results will be stated later in this section. More recently, Diekmann [6] and Thieme [15] have considered similar types of models with continuous time and have obtained results similar to those obtained by Aronson and Weinberger in [2] and Weinberger in [17]. Further references to the subject may be found in [13].

The purpose of this paper is to prove the kind of results that Kolmogorov et al. and Uchiyama obtained for equation (1.1) for Weinberger's model under various assumptions on Q and u_0. According to [17, Thm. 3], if $u_0(x)$ is bounded away from zero outside a bounded set, then $u_n(x)$ defined recursively via (1.2) converges uniformly to 1 in $\mathbb{R}$ as $n\to+\infty$. We shall therefore always assume that $u_0(+\infty)=0$. There are then two cases to consider, namely when $\liminf_{x\to-\infty} u_0(x)>0$ and when this is not satisfied. We shall treat the first case in this paper and the second case in a later paper [12]. However, some of the results obtained in this paper will be used in both cases. The rest of this section contains slight generalizations of three major theorems obtained in [17] as well as notations to be used for the rest of this paper. To obtain the main result of this paper, we require the kernel $K(x)$ in the definition of Q to be the exponential of a concave function. Section 2 explains and proves the consequences of such an assumption. Section 3 contains further properties of travelling waves. In §4, we study the behavior of $u_n(x)$ in a neighborhood of its largest zero. We will then be able to treat the first case when $u_0(x)$ stays positive near $-\infty$. This is done in §5.

We now introduce notations and make assumptions on our nonlinear integral operator Q. The operator Q is defined on the set of functions

$$\mathcal{C}=\{u: 0\le u\le 1,\ u \text{ piecewise continuous}\}$$

by the relation

$$Q[u](x)=\int_{\mathbb{R}} K(x-y)g(u(y))\,dy. \tag{1.3}$$

$K(x)$ is assumed to satisfy the following conditions throughout the entire paper.

(i) $\operatorname{supp} K=[B_1,B_2]$, $K(x)>0$ in (B_1,B_2),

(ii) $K(x)$ is continuous in $\mathbb{R}$ except possibly at B_1, B_2, where

$$\lim_{x\downarrow B_1} K(x)=p_1,\quad \lim_{x\uparrow B_2} K(x)=p_2,\quad p_1,p_2\ge 0, \tag{1.4}$$

(iii) $K(x)$ is of bounded variation and is differentiable in $(B_2-\varepsilon,B_2)$ for some $\varepsilon>0$,

(iv) $\int_{\mathbb{R}} K(x)\,dx=1$.

$g(u)$ defined in $[0,1]$ is assumed to satisfy the following conditions throughout the entire paper.

(v) $g(u)\in C^1[0,1]$,

(vi) $g(0)=0$, $g(1)=1$,

(vii) $g(u)>u$ in $(0,1)$,

(viii) $0\le g'(u)\le g'(0)$ in $[0,1]$, (1.5)

(ix) $g(u)=g'(0)u+O(u^{1+\varepsilon})$ as $u\downarrow 0$ for some $\varepsilon>0$,

(x) $g'(u)>0$ in $[0,\sigma)$ where $\sigma=\sup\{u: g(u)<1\}$,

(xi) $g(u)/u$ is nonincreasing in $(0,1)$.

Condition (xi) is used only in the last part of the proof of Theorem 4; it is not clear if it is essential there. Note that (xi) implies that $g'(u)\leq g'(0)$ in $[0,1]$, which in turn implies that $g(u)\leq g'(0)u$ in $[0,1]$. From this and (vii), we see that $g'(0)>1$. By Q_c, we always mean the operator

$$Q_c[u](x)=Q[u](x+c). \tag{1.6}$$

The letter β will be used to denote $g'(0)$ and hereafter, if the domain of integration is unspecified, it is assumed to be the entire real line $\mathbb{R}$.

In [17], the conditions on $K(x)$ and $g(u)$ are not identical to those stated in (1.4) and (1.5). Most important of all, $K(x)$ was assumed to be an even function in [17]. Also (x), (xi) were not assumed, (viii) was replaced by the weaker condition $g(u)\leq\beta u$ in $[0,1]$, while condition (ix) was replaced by the stronger condition that there exists a constant $D>0$ such that $\beta[u-Du^2]\leq g(u)\leq\beta u$ in $[0,1]$. Condition (viii) may be replaced by the condition $g(u)\leq\beta u$ in $[0,1]$ in order for Theorems 1, 2, 3 of this paper to be valid. This corresponds to the condition $f(u)\leq f'(0)u$ for (1.1); see [14] if such a condition is absent in (1.1). Nevertheless, (viii) will be used in this paper.

We need to adjust [17, Thms. 1, 2 and 5] so that they hold under our present hypotheses of $K(x)$ and $g(u)$. For this purpose, we have to refer to a more recent paper of Weinberger [18] which generalizes results in [17]. The function

$$\Phi(\mu)=\frac{1}{\mu}\log\left\{\beta\int K(x)e^{\mu x}dx\right\}, \qquad \mu\neq 0 \tag{1.7}$$

plays a crucial role in all our analysis. Note that $\Phi(\mu)$ depends only on $\beta=g'(0)$ and not on $g(u)$ in general. We define

$$\Psi(\mu)=\frac{\int xe^{\mu x}K(x)\,dx}{\int e^{\mu x}K(x)\,dx}, \tag{1.8}$$

and note that

$$\Psi'(\mu)=\frac{\int[x-\Psi(\mu)]^2e^{\mu x}K(x)\,dx}{\int e^{\mu x}K(x)\,dx}>0, \tag{1.9}$$

$$\Phi'(\mu)=-\frac{1}{\mu}[\Phi(\mu)-\Psi(\mu)]. \tag{1.10}$$

From (1.10), we see immediately that $(\mu^2\Phi')'=\mu\Psi'$, and hence $\mu^2\Phi'(\mu)$ is increasing in $(0,+\infty)$ and decreasing in $(-\infty,0)$. Consequently, Φ can have no local maximum in $(0,+\infty)$ and no local minimum in $(-\infty,0)$. It has at most one local minimum μ^* in $(0,+\infty)$ and $\Phi(\mu)$ decreases in $(0,\mu^*)$ and increases in $(\mu^*,+\infty)$. Similarly $\Phi(\mu)$ has at most one local maximum μ_* in $(-\infty,0)$ and $\Phi(\mu)$ decreases in $(\mu_*,0)$ and increases in $(-\infty,\mu_*)$. Note that since $\beta\int K(x)\,dx>1$, $\Phi(0+)=+\infty$ and $\Phi(0-)=-\infty$. For $\mu>0$,

$$\Phi(\mu)\geq\frac{1}{\mu}\log\left\{\beta e^{\mu B_1}\int K(x)\,dx\right\}=B_1+\frac{1}{\mu}\log\beta. \tag{1.11}$$

Also for $\mu>0$ and $\delta>0$ sufficiently small,

$$\begin{aligned}\Phi(\mu)&\geq\frac{1}{\mu}\log\left\{\beta\int_{B_2-\delta}^{B_2}e^{\mu x}K(x)\,dx\right\}\\&\geq B_2-\delta+\frac{1}{\mu}\log\left\{\beta\int_{B_2-\delta}^{B_2}K(x)\,dx\right\}.\end{aligned}\tag{1.12}$$

Choose $\delta>0$ so small such that $\int_{B_2-\delta}^{B_2}K(x)\,dx\leq 1/2\beta$. Then for $\mu>0$

$$\begin{aligned}\Phi(\mu)&=\frac{1}{\mu}\log\left\{\beta\int_{B_1}^{B_2-\delta}e^{\mu x}K(x)\,dx+\beta\int_{B_2-\delta}^{B_2}e^{\mu x}K(x)\,dx\right\}\\&\leq\frac{1}{\mu}\log\left\{\beta e^{\mu(B_2-\delta)}+\frac{1}{2}e^{\mu B_2}\right\}\\&=B_2+\frac{1}{\mu}\log\left\{\beta e^{-\mu\delta}+\frac{1}{2}\right\}.\end{aligned}\tag{1.13}$$

From (1.12) and (1.13), we see that $\lim_{\mu\to+\infty}\Phi(\mu)=B_2$. Also from (1.13), if $\mu>0$ is chosen so that $\beta e^{-\mu\delta}<\frac{1}{2}$, then $\Phi(\mu)<B_2$. Thus $\Phi(\mu)$ attains a minimum value c_+^* at some unique point $\mu^*>0$. Note that c_+^* may be negative but

$$B_1<B_1+\frac{1}{\mu^*}\log\beta\leq c_+^*<B_2.\tag{1.14}$$

To continue,

$$\Phi(\mu)=\frac{1}{\mu}\log\left\{\beta\int e^{-\mu x}K(-x)\,dx\right\}\equiv-\Phi_1(-\mu),\tag{1.15}$$

where

$$\Phi_1(\eta)=\frac{1}{\eta}\log\left\{\beta\int e^{\eta x}K(-x)\,dx\right\},\qquad\eta\neq0.$$

The support of $K(-x)$ is $[-B_2,-B_1]$, and our previous analysis shows that $\Phi_1(\eta)$ has a unique minimum $c_-^*=\Phi_1(\eta^*)$ at $\eta=\eta^*>0$, $-B_2<\Phi_1(\eta^*)<-B_1$ and $\lim_{\eta\to+\infty}\Phi_1(\eta)=-B_1$. From (1.15), we conclude that $\Phi(\mu)$ attains a maximum value $-c_-^*$ at a unique point $\mu_*=-\eta^*$, $B_1<-c_-^*<B_2$ and $\lim_{\mu\to-\infty}\Phi(\mu)=B_1$. Note that

$$c_-^*=\inf_{\mu>0}\frac{1}{\mu}\log\left\{\beta\int e^{-\mu x}K(x)\,dx\right\}.$$

We are now ready to state two slight generalizations of theorems in [17] which will be used essentially later on. If $u_0\in\mathcal{C}$, we define recursively

$$u_{n+1}(x)=Q[u_n](x)\quad\text{for } n\geq0.\tag{1.16}$$

It is clear from our hypotheses (1.4) and (1.5) that $u_n\in\mathcal{C}$ for $n\geq0$.

THEOREM 1. *Let u_n be a solution of the recursion* (1.16). *Let $0\leq u_0\leq1$ and suppose $u_0(x)$ has bounded support. Then, for any $c_1<-c_-^*<c_+^*<c_2$,*

$$\lim_{n\to+\infty}\max_{x\notin[nc_1,nc_2]}u_n(x)=0.$$

Proof. [18, Thm. 6.1]. Actually the proof in [17] also works without any modification.

THEOREM 2. *Let u_n be a solution of the recursion* (1.16). *Let $0 \le u_0 \le 1$ and suppose that u_0 is not identically zero in the sense that its integral is positive. Then, for any $-c_-^* < c_1 < c_2 < c_+^*$,*

$$\lim_{n \to +\infty} \max_{x \in [nc_1, nc_2]} u_n(x) = 1.$$

Proof. [18, Thms. 6.1 and 6.5].

Remark 1. In Theorem 1, if $u_0(x)$ decays rapidly enough so that the function $u(x)e^{\mu^* x}$ is integrable near $+\infty$, then the maximum of $u_n(x)$ in the set $\{x \ge nc_2\}$ approaches zero for every $c_2 > c_+^*$.

Remark 2. If $K(x) = K(-x)$, then $c_+^* = c_-^*$ and $\Psi(0) = 0$. Since $\Psi(\mu)$ is increasing, c_+^* is positive in this case. In fact $c_+^* > 0$ as long as $\int xK(x)\,dx \ge 0$.

Remark 3. Theorems 1 and 2 together say $u_n(x)$ spreads out on the right like nc_+^* and on the left like $-nc_-^*$ and $-c_-^* < c_+^*$.

Remark 4. Condition (ix) of (1.5) is irrelevant for the proofs of Theorems 1 and 2.

We now come to the question of travelling waves. A function $w_c(x) \in \mathcal{C}$ is a travelling wave of the operator Q with speed c if $u_n(x) = w_c(x - nc)$ is a solution to the recursion (1.2); i.e.,

$$w_c(x) = \int K(x + c - y) g(w_c(y))\,dy. \tag{1.17}$$

We define μ_c for $c \ge c_+^*$ as the unique positive value of μ in $(0, \mu^*)$ such that $\Phi(\mu_c) = c$. We shall from now on write $\mu_{c_+^*}$ instead of μ^* to be consistent with our notation.

THEOREM 3. *For each $c \ge c_+^*$, there exists a travelling wave $w_c(x)$ of speed c for the operator Q such that $w_c(x)$ is nonincreasing. Any such solution has the further properties that*

$$w_c(-\infty) = 1, \qquad w_c(+\infty) = 0 \tag{1.18}$$

and

$$w_c(x) > 0 \quad \text{in } \mathbb{R}. \tag{1.19}$$

For $c > c_+^*$,

$$\lim_{x \to +\infty} e^{\mu_c x} w_c(x) = 1, \tag{1.20}$$

while for $c = c_+^*$,

$$\lim_{x \to +\infty} \frac{1}{x} e^{\mu_{c_+^*} x} w_c(x) = 1. \tag{1.21}$$

Proof. The existence of nonincreasing travelling waves for $c \ge c_+^*$ with property (1.18) may be found in the proof of [18, Thm. 6.6]. (1.18), (1.19), (1.20) are contained in Diekmann and Kaper [7]. We shall prove (1.21) in the first half of the proof of Theorem 4 in §3. Condition (ix) is essential in all of these proofs.

Remark 1. Uniqueness of travelling waves for $c > c_+^*$ was solved by Diekmann and Kaper in [7]. Monotonicity of $w_c(x)$ plays a role in proving the uniqueness and deriving property (1.21) in Theorem 4. We shall discuss this problem further in §3.

Remark 2. There are travelling waves $\bar{w}_c(x)$ of speed $c \le -c_-^*$ going in the opposite direction, namely, $\bar{w}_c(x)$ is nondecreasing, $\bar{w}_c(-\infty) = 0$, $\bar{w}_c(+\infty) = 1$ and $\bar{w}_c(x + nc)$ satisfies the recursion (1.2). Note that since $K(x)$ is not symmetric, $\bar{w}_c(x)$ does not equal $w_c(-x)$.

2. PF_2. Most of the results in this section are special cases of more general theorems in Karlin's book *Total Positivity* [9]. In order to keep this paper within reasonable length, we are going to state some of the results in this section as lemmas and refer the reader to [9] or [11] for their proofs. We start by defining the sign change of a function.

Let $\underset{\sim}{x}=(x_1,\cdots,x_n)$ be a vector of real numbers. We denote by $S(x)$ the number of sign changes in the sequence obtained from $x_1, x_2,\cdots,x_n$ by deleting all zero terms with the convention that $S(\underset{\sim}{\theta})=-1$, where $\underset{\sim}{\theta}$ is the null vector.

DEFINITION 1. The number of sign changes $S[f]$ of $f(t)$ is

$$\sup S[f(t_1),f(t_2),\cdots,f(t_m)],$$

where the supremum is taken over all sets $-\infty<t_1<t_2<\cdots<t_m<+\infty$. m is arbitrary but finite.

DEFINITION 2. A function $f(x)$ defined in $\mathbb{R}$ is said to be PF_r, $r\geq 2$ if $f(x)$ is PF_{r-1} and if for all $-\infty<x_1<\cdots<x_r<+\infty$, $-\infty<y_1<\cdots<y_r<+\infty$, we have

$$\det f(x_i-y_j)\geq 0. \tag{2.1}$$

PF_1 simply means a nonnegative function defined in $\mathbb{R}$. PF_r means Pólya frequency function of order r.

Remark. If $f(x)$ is PF_r, so is $f(ax+b)$ for any constants a and b.

A PF_2 function has a particularly simple characterization.

LEMMA 1. *A function $f(x)\not\equiv 0$ is PF_2 if and only if $f(x)=e^{k(x)}$, where $k(x)$ is a concave function in some interval (a,b) ($a=-\infty, b=+\infty$ not excluded), $f(x)=0$ outside $[a,b]$.*

Proof. See [9, p. 332] or [11].

Remark 1. It is well known that any concave function $k(x)$ in (a,b) is continuous and its left and right derivatives $k'_L(x)$ and $k'_R(x)$ exist everywhere in (a,b). Furthermore, $k'_L(x)=k'_R(x)$ except for a countable number of points. If $f(x)$ is PF_2, then, according to Lemma 1, $f(x)$ is of bounded variation and is continuous in $\mathbb{R}$ except possibly at the points a and b. That this is best possible is illustrated by the following example. The function $\varphi(x)=1$ if $-1<x<1$ and 0 otherwise is PF_2 (but not PF_3). However $\lim_{x\downarrow a}f(x)$ and $\lim_{x\uparrow b}f(x)$ both exist and are finite. In addition, $f(x)$ is absolutely continuous in every compact subset of (a,b). If $f(x)$ is continuous at the points a,b, then it is absolutely continuous in $\mathbb{R}$.

Remark 2. Let $K(x)$ in the definition of Q in (1.3) be PF_2. Then most of the conditions in (1.4) are automatically satisfied. We denote by p_1 and p_2 the jumps of $K(x)$ at B_1 and B_2 respectively. We shall set $K(B_1)=K(B_2)=0$ and decompose $K(x)$ in the following way: $K(x)=K_1(x)+K_a(x)$ where

$$K_1(x)=\begin{cases}\left(\dfrac{p_2-p_1}{B_2-B_1}\right)(x-B_1)+p_1, & B_1<x<B_2,\\ 0 & \text{otherwise.}\end{cases}$$

Note that $K_a(x)$ has compact support and is absolutely continuous.

LEMMA 2. *Let $f(x)$ be a bounded piecewise continuous function in $\mathbb{R}$ such that, for some point x_0,*

$$f(x)\begin{cases}\geq 0, & x<x_0,\\ \leq 0, & x>x_0.\end{cases} \tag{2.2}$$

Let $K(x)=e^{k(x)}$ *be* PF_2 *and define*

$$F(x)=\int K(x-y)f(y)\,dy. \tag{2.3}$$

Then $F(x_1)=0$ *and* F *differentiable at* x_1 *imply that* $F'(x_1)\leq 0$. *Furthermore, there cannot exist* $x_2<x_3$ *such that* $F(x_2)<0<F(x_3)$.

Proof. This is a special case when $r=2$ of a general theorem in [9, p. 21]. The proof is also given in [11].

Remark 1. The lemma says that if $f(x)$ has one sign change and $K(x)$ is PF_2, then the function $K*f$ has no more than one sign change. This lemma is obviously valid if we replace $K(x)$ by any of its translates.

Remark 2. $g(u)$ does not increase the number of sign changes of the difference of two functions. By this we mean if $u(x)$ and $v(x)$ are such that $u(x)-v(x)$ has a finite number of sign changes, then the number of sign changes of $g(u(x))-g(v(x))$ cannot exceed that of $u(x)-v(x)$. The proof follows readily from the mean value theorem, the fact that $g'\geq 0$ and Definition 1.

LEMMA 3. *Let* $K(x)$ *be* PF_2, $u_0\in\mathcal{C}$ *and let there exist a point* x_0 *such that*

(i) $u_0(x)\geq u_0(x_0)\equiv\delta_0>0$ *for* $x<x_0$,

(ii) $u_0(x)$ *is nonincreasing for* $x>x_0$,

(iii) $u_0(+\infty)=0$.

Then there exists a sequence $\{x_n, n\geq 0\}$ *of points such that* $u_n(x_n)=\delta_n$ *where the* δ_n*'s are defined by the recursion* $\delta_{n+1}=g(\delta_n)$. *Also* $u_n(x)\geq\delta_n$ *if* $x<x_n$ *and* $u_n(x)$ *is nonincreasing if* $x>x_n$.

Proof. The proof is by induction, and the assumptions on $u_0(x)$ say that the result is true when $n=0$. Assume that the result is true for n. We set $f(y)=g(u_n(y))-g(\delta_n)$. Then $f(x)\geq 0$ if $x<x_n$, $f(x)\leq 0$ if $x>x_n$. From Lemma 2, we find that $u_{n+1}(x)-\delta_{n+1}$ changes sign no more than once. If we set $x_{n+1}=\sup\{x: u_{n+1}\geq\delta_{n+1}\}$, then x_{n+1} is finite since $u_{n+1}(+\infty)=0$ and from our induction hypothesis, $x_{n+1}\geq x_n+B_1$. Also $u_{n+1}(x_{n+1})=\delta_{n+1}$. We then get $u_{n+1}(x)\geq\delta_{n+1}$ if $x<x_{n+1}$ and $u_{n+1}(x)<\delta_{n+1}$ if $x>x_{n+1}$. It remains to prove that $u_{n+1}(x)$ is nonincreasing in $(x_{n+1}, +\infty)$. To this end, let $\bar{x}>x_{n+1}$ and set $\bar{\gamma}=u_{n+1}(\bar{x})$. $\bar{\gamma}<\delta_{n+1}$ by definition of $\bar{x}$ and x_{n+1}. Choose γ such that $g(\gamma)=\bar{\gamma}<\delta_{n+1}=g(\delta_n)$. Then $\gamma<\delta_n$ and if we let $\tilde{x}=\sup\{x: u_n(x)\geq\gamma\}$, we have $u_n(x)\geq\gamma$ for $x<\tilde{x}$ and $u_n(x)\leq\gamma$ for $x>\tilde{x}$. The function $\bar{f}(x)=g(u_n(x+\bar{x}))-g(\gamma)$ has only one sign change at $x=\tilde{x}-\bar{x}$, and Lemma 2 implies that $\bar{F}(x)=\int K(x-y)\bar{f}(y)\,dy=u_{n+1}(x+\bar{x})-\bar{\gamma}$ has no more than one sign change. If $\bar{F}(x)>0$ for some $x>0$, then since $\bar{F}(0)=0$, $\bar{F}(x_{n+1}-\bar{x})>0$ with $x_{n+1}-\bar{x}<0$, there exists by continuity a λ, $0<\lambda<\delta_{n+1}-\bar{\gamma}$ such that $\bar{F}(x)-\lambda$ has at least two sign changes. This is impossible because from $\bar{\gamma}<\bar{\gamma}+\lambda<\delta_{n+1}$ and our induction hypothesis, we may choose $\bar{\bar{x}}$, $x_{n+1}<\bar{\bar{x}}<\bar{x}$ such that $u_{n+1}(\bar{\bar{x}})=\bar{\gamma}+\lambda$. Repeating the argument before with $\bar{x}$ replaced by $\bar{\bar{x}}$, $\bar{\gamma}$ replace by $\bar{\gamma}+\lambda$, we conclude that $u_{n+1}(x+\bar{\bar{x}})-\bar{\gamma}-\lambda$ must change sign no more than once. But $\bar{F}(x)-\lambda$ is only a translate of $u_{n+1}(x+\bar{\bar{x}})-\bar{\gamma}-\lambda$. The only way out is to admit that $\bar{F}(x)\leq 0$ whenever $x>0$ which is precisely the statement that $u_{n+1}(x)$ is nonincreasing for $x>x_{n+1}$. Q.E.D.

We shall hereafter refer to conditions (i), (ii) and (iii) of Lemma 3 as the one-sided condition.

Let $f(x)\in\mathcal{C}$ be such that $f(+\infty)=0$. We define for $0<\gamma<1$ the function $f^{-1}(\gamma)=\sup\{x: f(x)\geq\gamma\}$. If $f(x)$ is everywhere less than γ, then $f^{-1}(\gamma)=-\infty$. The next lemma is applicable to both the case $u_0(-\infty)>0$ and $u_0(-\infty)=0$.

LEMMA 4. *Let $K(x)$ be PF_2. Let $f_0(x)=H(-x)$ and $f_n(x)$ be defined recursively by the relation $f_{n+1}=Q[f_n]$, $n\geq 0$. Let $u_0\in\mathcal{C}$ be nontrivial, and $u_0(+\infty)=0$. Then for any compact subset $\mathbb{K}$ of $(0,1)$, we have $f_n'(f_n^{-1}(\gamma))\leq u_n'(u_n^{-1}(\gamma))$ for all $\gamma\in\mathbb{K}$ when n is sufficiently large.*

Proof. It is easy to see that $f_n(x)$ is nonincreasing in $\mathbb{R}$, $f_n(-\infty)=1$, $f_n(+\infty)=0$ for all n. Also Theorem 2 implies that the maximum of $u_n(x)$ on $\mathbb{R}$ converges to 1 and hence $u_n^{-1}(\gamma)>0$ on $\mathbb{K}$ for sufficiently large n. For such large n, set $b=f_n^{-1}(\gamma)-u_n^{-1}(\gamma)$ and consider the function $\bar{f}(y)=f_0(y+b)-u_0(y)$. From the definition of $f_0(y)$, $\bar{f}(y)$ has only one sign change at $y=-b$. From Remark 2 after Lemma 2, the same is true of the function $g(f_0(y+b))-g(u_0(y))$. Hence Lemma 2 implies that the function

$$\int K(x-y)\left[g(f_0(y+b))-g(u_0(y))\right]dy=f_1(y+b)-u_1(x)$$

has no more than one sign change and has nonpositive derivative at its zero. An induction argument now shows that the same is true of the function $f_n(x+b)-u_n(x)=f_n(x+f_n^{-1}(\gamma)-u_n^{-1}(\gamma))-u_n(x)$. This function vanishes at $x=u_n^{-1}(\gamma)$ and hence $f_n'(f_n^{-1}(\gamma))\leq u_n'(u_n^{-1}(\gamma))$. Q.E.D.

3. Travelling waves. In this section we are going to study further properties of the travelling waves we mentioned at the end of §1. The following assumption on $g(u)$ is used only twice in this paper, once in Lemma 6 and once at the end of Theorem 4.

Let $\sigma=\sup\{u: g(u)<1\}$. We recall the condition

$$g'(u)>0 \quad \text{in } [0,\sigma). \tag{3.1}$$

PROPOSITION 1. *Let $\varphi\in L^1(\mathbb{R})$, $\psi\in L^\infty(\mathbb{R})$. Then the function $\varphi*\psi$ is uniformly continuous and $\|\varphi*\psi\|_\infty\leq\|\varphi\|_1\|\psi\|_\infty$.*

Proof. Uniform continuity comes from the continuity of translation in L^1 and the inequality is a straightforward estimate. Q.E.D.

LEMMA 5. *Let $\varphi(x)$ be absolutely continuous in $\mathbb{R}$. Let $f(x)$ be bounded. Then $F(s)=(\varphi*f)(s)$ is in $C^1(\mathbb{R})$, $F'(s)=(\varphi'*f)(s)$ and $F'(s)$ is uniformly continuous.*

Proof. Since $\varphi'(s)$ exists almost everywhere and is integrable, it suffices to prove that $F'(s)=(\varphi'*f)(s)$. Recall that $\varphi(s)$ is absolutely continuous if and only if for all s and a, $\varphi(s)-\varphi(a)=\int_a^s\varphi'(t)\,dt$. Hence

$$\begin{aligned}\int\varphi(s-t)f(t)\,dt&=\int\int_0^s\varphi'(y-t)\,dyf(t)\,dt+\int\varphi(-t)f(t)\,dt\\&=\int_0^s\int\varphi'(y-t)f(t)\,dt\,dy+\text{constant}.\end{aligned}$$

Since φ' is integrable and f is bounded, Proposition 1 implies that the function $\int\varphi'(y-t)f(t)\,dt$ is bounded and uniformly continuous. Hence the iterated integral exists and the interchange of order of integration is justified. Our assertion now follows from the fundamental theorem of integral calculus. Q.E.D.

COROLLARY. *Let $K(x)$ be PF_2 and $f(x)$ be continuous and bounded. Then the function $F(x)=(K*f)(x)$ is in $C^1(\mathbb{R})$ and*

$$F'(x)=\int K'(x-y)f(y)\,dy+p_1f(x-B_1)-p_2f(x-B_2).$$

Also, if $0\leq f(x)\leq 1$, then $\|F'\|_\infty\leq\|K'\|_1+p_1+p_2$.

Proof. We decompose $K(x)=K_1(x)+K_a(x)$ and apply Lemma 5 to K_a*f.

Remark. If $K(x)$ is PF_2, then $u_n(x)$ is differentiable for $n \geq 2$ since $u_1(x)$ is continuous from Proposition 1.

LEMMA 6. *Let $K(x)$ be PF_2 and $w_c(x)$ be a nonincreasing travelling wave for the operator Q of speed $c \geq c_+^*$. Then $w_c(x) \in C^2(\mathbb{R})$, $w_c'(x)$, $w_c''(x)$ are uniformly continuous and bounded. Also $w_c'(x) < 0$ whenever $w_c(x) < 1$.*

Proof. The statements about the regularity of $w_c(x)$ follow immediately from the corollary of Lemma 5. We actually get the bounds

$$\|w_c'\|_\infty \leq (\|K'\|_1 + p_1 + p_2), \qquad \|w_c''\|_\infty \leq \beta(\|K'\|_1 + p_1 + p_2)^2.$$

To see that $w_c'(x) < 0$ for $w_c(x) < 1$, write

$$w_c'(x) = \int_{x+c-B_2}^{x+c-B_1} K(x+c-y) g'(w_c(y)) w_c'(y)\, dy.$$

Suppose that $w_c'(x_0) = 0$ and $w_c(x_0) \in (0,1)$. Then $K(x) > 0$ in (B_1, B_2) and $w_c'(x) \leq 0$ imply that $g'(w_c(y)) w_c'(y) = 0$ in $[x_0 + c - B_2, x_0 + c - B_1]$. If $c < B_2$, then since w_c is nonincreasing, $w_c(y) < 1$ in $[x_0, x_0 + c - B_1]$ would imply from our assumption (3.1) on $g(u)$, that $w_c'(y) = 0$ in $[x_0, x_0 + c - B_1]$. Repeating the argument with x_0 replaced by $x_0 + c - B_1$ and using the fact that $c \geq c_+^* > B_1$, we arrive at the conclusion that $w_c'(x) = 0$ for $x \geq x_0$. This is impossible since $w_c(+\infty) = 0$ while $w_c(x_0) > 0$. If $c \geq B_2$, then $w_c'(y) = 0$ in $[x_0 + c - B_2, x_0 + c - B_1]$. The mapping $y \to [y + c - B_2, y + c - B_1]$ is continuous and maps connected sets to connected sets. Using this, we can prove by induction that $w_c'(x) = 0$ in a sequence of intervals $[x_0 + n(c - B_2), x_0 + n(c - B_1)]$, $n \geq 0$, which begin to overlap each adjacent one when n is so large that $c < n(B_2 - B_1) + B_2$. Thus $w_c'(x) = 0$ in $[A, +\infty)$ for some constant A, which is impossible since $w_c(x) > 0$ in $\mathbb{R}$ and $w_c(+\infty) = 0$. Q.E.D.

Remark. Define $w_c^{-1}(\gamma)$ as we have done right after Lemma 3 for a general function in $\mathcal{C}$ vanishing at $+\infty$. Then $w_c'(w_c^{-1}(\varepsilon)) < 0$ in $(0,1)$ and is continuous there. In particular, $1/w_c'(w_c^{-1}(\varepsilon))$ is integrable over every compact subset of $(0,1)$ for all $c \geq c_+^*$.

We now discuss the question of uniqueness of travelling waves for the operator Q. The case $c > c_+^*$ was solved by Diekmann and Kaper [7]. They showed that any such solution w_c in $\mathcal{C}$ has the asymptotic behavior $w_c(x) \sim \text{const}\, e^{-\mu_c x}$ as $x \to +\infty$. Our proof of the case $c = c_+^*$ uses some of the results in [7] and an extension of Ikehara's theorem by Delange [5]. We are then able to deduce that any nontrivial ($\not\equiv 0, 1$) nonincreasing solution of $Q_{c_+^*}[u] = u$ has the asymptotic behavior $u(x) \sim \text{const}\, x e^{-\mu_{c_+^*} x}$ as $x \to +\infty$. However we are unable to jump from this to a uniqueness result. We need the condition that $g(u)/u$ be nonincreasing in $(0,1)$. For this reason, our main theorem in §5 may be weaker. In passing, we mention a paper by Barbour [4] which proves uniqueness of the epidemic waves by stochastic methods. See also Aronson [1]. The following is a modified version of Theorem 1 in Delange's paper.

PROPOSITION 2. *Let $\alpha(t)$ be a real function defined for $t \geq 0$, nonincreasing, $\alpha(t) \geq 0$. We suppose that $f(s) = \int_0^{+\infty} e^{-st} \alpha(t)\, dt$ is convergent for $\operatorname{Re} s > -a$ where $a > 0$. Let $\beta(t)$ be a real function defined for $t \geq 0$, measurable and bounded on every finite interval,. We suppose that the integral $G(s) = \int_0^{+\infty} e^{-st} \beta(t)\, dt$ is convergent for $\operatorname{Re} s > 0$. Suppose in addition that there exists a nonnegative integer p and a real function $\gamma(u)$ continuous and nondecreasing in an interval $[0, l]$ and $\gamma(u) > 0$ in $(0, l]$ such that the integral $\int_0^l \log(1/\gamma(u))$ du is convergent and the product $t^{p+1} \beta(t) \int_0^l e^{-tu} \gamma(u)\, du$ converges to 1 as $t \to +\infty$. Suppose further that there exists a constant $A > 0$ such that the function $F(s) = f(s) - AG(s + a)$ has the following properties*:

1°. *For every real $y \neq 0$, $F^{(p)}(s)$ tends to a finite limit as s tends to $-a + iy$ in the half-plane $\operatorname{Re} s > -a$.*

2°. *As $s+a$ tends to 0 in the half-plane $\mathrm{Re}\, s > -a$, we have $F^{(p)}(s) = 0[\psi(r)/\gamma(r)]$ with $r = |s+a|$. Here $\psi(t) > 0$ is nonincreasing, defined for sufficiently small $t > 0$ and the integrals $\int_0 \psi(t)\,dt$ and $\int_0 \psi(t) \log(1/\gamma(t))\,dt$ are convergent. Then*

$$\alpha(t) \sim Ae^{-at}\beta(t) \quad \text{as } t \to +\infty.$$

THEOREM 4. *Nonincreasing travelling wave solution of the operator $Q_{c_+^*}$ is unique up to translation.*

Proof. We first show that any nonincreasng solution $u \in \mathcal{C}$ of $Q_{c_+^*}[u] = u$ has the asymptotic behavior $u(x) \sim Axe^{-\mu_{c_+^*} x}$ as $x \to +\infty$ for some constant $A > 0$. Let $U(\lambda) = \int e^{\lambda x} u(x)\,dx$, λ complex, then [7, Lemmas 4.4 and 4.5] imply that $u(-\infty) = 1$, $u(+\infty) = 0$ and the number $\Lambda_u = \sup\{\mathrm{Re}\lambda : |U(\lambda)| < +\infty\}$ is positive. $U(\lambda)$ is holomorphic in $S = \{\lambda \in \mathbb{C} : 0 < \mathrm{Re}\lambda < \Lambda_u\}$. Let $r(x) = (g(u) - \beta u) * K(x + c_+^*)$ and $R(\lambda) = \int e^{\lambda x} r(x)\,dx$. Then $R(\lambda) = U(\lambda)(1 - \beta \int K(x + c_+^*) e^{\lambda x}\,dx) \equiv U(\lambda) h(\lambda)$. According to [7, Lemma 4.8], $R(\lambda)$ is actually holomorphic in a strip to the right of S and hence $h(\Lambda_u) = 0$ which implies that $\Lambda_u = \mu_{c_+^*}$. $h(\lambda)$ is an entire function and if $\lambda = \lambda_1 + i\lambda_2$, a simple integration by parts show that $h(\lambda)$ has no zero for $\mathrm{Re}\lambda$ bounded and $|\mathrm{Im}\lambda|$ sufficiently large. Furthermore, since $h(\lambda)$ is entire, we can show that for sufficiently small $\varepsilon > 0$, $h(\lambda)$ has only one zero at $\lambda = \mu_{c_+^*}$ in the strip $S_\varepsilon = \{\lambda \in \mathbb{C} : \mu_{c_+^*} - \varepsilon < \mathrm{Re}\lambda < \mu_{c_+^*} + \varepsilon\}$. The function $U(\lambda) = R(\lambda)/h(\lambda)$ is meromorphic in S_ε and has a pole of order two at $\lambda = \mu_{c_+^*}$. The order of the pole may be checked by looking at (1.10) and differentiating $h(\lambda) = 1 - e^{[\Phi(\lambda) - c_+^* \lambda]}$ twice to get $h''(\mu_{c_+^*}) = -\Phi''(\mu_{c_+^*}) \mu_{c_+^*} < 0$. Since $g(u(x)) < \beta u(x)$ for x in a set of positive measure, $R(\mu_{c_+^*}) < 0$. Let

$$\frac{R(\lambda)}{h(\lambda)} = \frac{a_{-2}}{(\lambda - \mu_{c_+^*})^2} + \frac{a_{-1}}{(\lambda - \mu_{c_+^*})} + \bar{h}(\lambda) \tag{3.2}$$

be the Laurent expansion at $\lambda = \mu_{c_+^*}$ where $\bar{h}(\lambda)$ is holomorphic in S_ε and $a_{-2} = 2R(\mu_{c_+^*})/(-\Phi''(\mu_{c_+^*}) \mu_{c_+^*}) > 0$. Now

$$U(\lambda) = \int_{-\infty}^{0} e^{\lambda x} u(x)\,dx + \int_0^{+\infty} e^{\lambda x} u(x)\,dx \equiv \bar{f}(\lambda) + \int_{-\infty}^{0} e^{\lambda x} u(x)\,dx.$$

$\bar{f}(\lambda)$ is holomorphic in $\mathrm{Re}\lambda < \mu_{c_+^*}$ and $\int_{-\infty}^{0} e^{\lambda x} u(x)\,dx$ is holomorphic in $\mathrm{Re}\lambda > 0$. Let $\bar{\bar{h}}(\lambda) = \bar{h}(\lambda) - \int_{-\infty}^{0} e^{\lambda x} u(x)\,dx$, we then have

$$\bar{f}(\lambda) - \frac{a_{-2}}{(\lambda - \mu_{c_+^*})^2} = \frac{a_{-1}}{(\lambda - \mu_{c_+^*})} + \bar{\bar{h}}(\lambda) \quad \text{in } S \cap S_\varepsilon. \tag{3.3}$$

Now let $\alpha(t) = u(t)$, $f(s) = \bar{f}(-s)$, $a = \mu_{c_+^*}$, $\beta(t) = t$, $p = 0$, $\gamma(u) = u$, $l = 1$, $A = a_{-2}$ and $\psi \equiv 1$ in Proposition 2. Then all the hypotheses are satisfied. In particular, $G(s) = s^{-2}$ exists when $s > 0$ and from (3.3),

$$F(s) = \bar{f}(-s) - \frac{a_{-2}}{(s + \mu_{c_+^*})^2} = \frac{-a_{-1}}{(s + \mu_{c_+^*})} + \bar{\bar{h}}(-s) \quad \text{in } -S.$$

According to Proposition 2,

$$u(x) \sim a_{-2} x e^{-\mu_{c_+^*} x} \quad \text{as } x \to +\infty.$$

Finally, let $u_1, u_2 \in \mathcal{C}$ be nonincreasing and $Q_{c_+^*}[u_i] = u_i$ for $i = 1, 2$. Then $u_i(x) \sim A_i x e^{-\mu_{c_+^*} x}$ as $x \to +\infty$ for $i = 1, 2$. Without loss of generality, assume that $A_1 \geq A_2$. Set $\alpha = \inf_{\mathbb{R}} u_1(x)/u_2(x)$. Since $u_i(-\infty) = 1$ for $i = 1, 2$, we have $0 \leq \alpha \leq 1$ and $u_1(x) \geq \alpha u_2(x)$

in $\mathbb{R}$. Note that if $\alpha<1, \alpha$ must be attained at some point x_0. Since $g(u)/u$ is nonincreasing, we have $g(\alpha u_2)\geq\alpha g(u_2)$. We deduce therefore that

$$u_1(x)\geq\int K(x+c_+^*-y)g(\alpha u_2(y))\geq\alpha u_2(x). \tag{3.4}$$

Let x_0 be such that $u_1(x_0)=\alpha u_2(x_0)$. From (3.4) and (3.1), we have $u_1(x)=\alpha u_2(x)$ in the interval $[x_0+c_+^*-B_2, x_0+c_+^*-B_1]$ provided $\alpha u_2(x)<\sigma$. Repeating the argument inductively, we conclude that $u_1(x)=\alpha u_2(x)$ in $\{x:\alpha u_2(x)<\sigma\}$. Hence $\{x:\alpha u_2(x)<\sigma\}\subset\{x:u_1(x)<\sigma\}$. But $u_1(x)\geq\alpha u_2(x)$ in $\mathbb{R}$; hence the two sets are equal. Letting $x\to+\infty$ and using the fact that $A_1\geq A_2$, we see that $\alpha=1$.

Finally, we may assume that $u_1(0)=u_2(0)$. If $\alpha<1$, then $u_1(x_0)=\alpha u_2(x_0)$ at some point x_0 but the above argument gives a contradiction. Hence $\alpha=1$ and we may take $x_0=0$. From above and the fact that

$$u_i(x)=\int_{\{u_i\leq\sigma\}}K(x+c_+^*-y)g(u_i(y))\,dy+\int_{\{u_i>\sigma\}}K(x+c_+^*-y)\,dy,$$

we conclude that $u_1(x)=u_2(x)$ in $\mathbb{R}$. Q.E.D.

We now return to the study of $w_c(x)$. Since the function $w_c'(w_c^{-1}(\gamma))$ will be used often, we introduce the notation $\omega_c(\gamma)=w_c'(w_c^{-1}(\gamma))$, $c\geq c_+^*$ for it.

THEOREM 5. *Let $u_0, \bar{u}_0\in\mathcal{C}$ satisfy the condition that for some constants A and b positive, $u_0(x)\leq\bar{u}_0(x)+Ae^{-bx}$ in $\mathbb{R}$. Then,*

$$u_n(x+nc)\leq\bar{u}_n(x+nc)+Ae^{-bx+nb(\Phi(b)-c)}. \tag{3.5}$$

In particular, if $u_0\in\mathcal{C}$ is such that $u_0(x)\sim w_c(x)$ as $x\to+\infty$ for some $c>c_+^*$, then $\lim_{n\to+\infty}u_n(x+nc)=w_c(x)$ uniformly in $[L, +\infty)$ for any constant L.

Proof. The proof of (3.5) is by induction, and our hypothesis says (3.5) is true when $n=0$. Assume that it is true up to n; then

$$\begin{aligned}
u_{n+1}(x+(n+1)c)-\bar{u}_{n+1}(x+(n+1)c)&\\
&=\int K(x+c-y)\left[g(u_n(y+nc))-g(\bar{u}_n(y+nc))\right]dy\\
&\leq\beta\int K(x+c-y)\left[u_n(y+nc)-\bar{u}_n(y+nc)\right]^+dy\\
&\leq A\beta e^{nb(\Phi(b)-c)}\int K(x+c-y)e^{-by}\,dy\\
&=Ae^{nb(\Phi(b)-c)}e^{\Phi(b)}b_e-b(x+c)\\
&=Ae^{(n+1)b(\Phi(b)-c)}e^{-bx},
\end{aligned}$$

which completes our induction.

For the second half of the theorem we let $\delta>0$. From (1.20) and our hypothesis,

$$\frac{u_0(x)}{w_c(x-\delta)}\to e^{-\mu_c\delta}<1 \quad\text{and}\quad \frac{u_0(x)}{w_c(x+\delta)}\to e^{\mu_c\delta}>1$$

as $x\to+\infty$. Hence there exists $L_\delta>0$ such that $w_c(x+\delta)\leq u_0(x)\leq w_c(x-\delta)$ in $[L_\delta, +\infty)$. For any $b>0$, set

$$A_b=\sup\{e^{bx}u_0(x):x\leq L_\delta\},\qquad B_b=\sup\{e^{bx}w_c(x+\delta):x\leq L_\delta\}.$$

We have

$$w_c(x+\delta)-B_b e^{-bx}\le u_0(x)\le w_c(x-\delta)+A_b e^{-bx} \quad \text{in } \mathbb{R}. \tag{3.6}$$

Applying (3.5) to the left and right-hand inequality in (3.6), we get

$$w_c(x+\delta)-B_b e^{-bx+nb(\Phi(b)-c)}\le u_n(x+nc)\le w_c(x-\delta)+A_b e^{-bx+nb(\Phi(b)-c)}. \tag{3.7}$$

Choose $b\in(\mu_c,\mu_{c_+^*})$ and let $n\to+\infty$ in (3.7). Since $\Phi(b)<c$ and δ is arbitrary, our theorem is proved. Q.E.D.

The next lemma is valid as long as $K'(x)$ has a finite number of sign changes. We shall prove it under the assumption that $K'(x)$ changes sign once, which is the case when $K(x)$ is PF_2, although the proof of the lemma makes it clear what to do in general. Hence let $K(x)=K_1(x)+K_a(x)$ and assume that

$$K_a'(x)\begin{cases}\ge 0 & \text{a.e. in } (B_1,L),\\ \le 0 & \text{a.e. in } (L,B_2).\end{cases} \tag{3.8}$$

LEMMA 7. *Let $K(x)$ be PF_2. If $c>c_+^*$, then $w_c'(x)\sim-\mu_c w_c(x)$ as $x\to+\infty$.*

Proof. From the definition of $w_c(x)$, we have

$$\begin{aligned} w_c'(x-c)=(K_1*g\circ w_c)'(x-c)&+\int K_a'(x-y)\left[g(w_c(y))-\beta w_c(y)\right]dy\\ &+\beta\int K_a'(x-y)w_c(y)\,dy. \end{aligned} \tag{3.9}$$

From the definition of $K_1(x)$ in Remark 2 after Lemma 1, we obtain

$$\begin{aligned} (K_1*g\circ w_c)'(x-c)=&p_1 g(w_c(x-B_1))-p_2 g(w_c(x-B_2))\\ &+\left(\frac{p_2-p_1}{B_2-B_1}\right)\int_{x-B_2}^{x-B_1} g(w_c(y))\,dy. \end{aligned} \tag{3.10}$$

Set

$$E=\beta\int_{x-B_2}^{x-B_1}K_a'(x-y)w_c(y)\,dy.$$

From (1.20) and for any $\varepsilon>0$, there exists L_ε such that for $y\ge L_\varepsilon$,

$$(1-\varepsilon)e^{-\mu_c y}\le w_c(y)\le(1+\varepsilon)e^{-\mu_c y}. \tag{3.11}$$

If $x\ge L_\varepsilon+B_2$, then from (3.11) and (3.8),

$$\begin{aligned} E&=\beta\int_{x-B_2}^{x-L}K_a'(x-y)w_c(y)\,dy+\beta\int_{x-L}^{x-B_1}K_a'(x-y)w_c(y)\,dy\\ &\le\beta(1-\varepsilon)\int_{x-B_2}^{x-L}K_a'(x-y)e^{-\mu_c y}\,dy+\beta(1+\varepsilon)\int_{x-L}^{x-B_1}K_a'(x-y)e^{-\mu_c y}\,dy. \end{aligned}$$

Integrate by parts:

$$\begin{aligned} E&\le 2\beta\varepsilon K_a(L)e^{-\mu_c(x-L)}-\beta(1-\varepsilon)\mu_c\int_{x-B_2}^{x-B_1}K_a(x-y)e^{-\mu_c y}\,dy\\ &=2\beta\varepsilon K_a(L)e^{-\mu_c(x-L)}-(1-\varepsilon)\mu_c e^{-\mu_c(x-c)}-\beta(1-\varepsilon)\int K_1(x-y)(e^{-\mu_c y})'\,dy. \end{aligned}$$

Using the definition of $K_1(x)$ and integrating the last term by parts, we obtain

$$E \le 2\beta\varepsilon K_a(L)e^{-\mu_c(x-L)} - (1-\varepsilon)\mu_c e^{-\mu_c(x-c)} - \beta(1-\varepsilon)p_1 e^{-\mu_c(x-B_1)}$$
$$+\beta(1-\varepsilon)p_2 e^{-\mu_c(x-B_2)} + \beta(1-\varepsilon)\left(\frac{p_2-p_1}{B_2-B_1}\right)\frac{e^{-\mu_c(x-B_1)} - e^{-\mu_c(x-B_2)}}{\mu_c}.$$

From (1.20), we get, from letting $\varepsilon \downarrow 0$,

$$\limsup_{x\to+\infty} \frac{E}{w_c(x-c)} \le -\mu_c - \beta p_1 e^{\mu_c(B_1-c)} + \beta p_2 e^{\mu_c(B_2-c)} + \frac{\beta}{\mu_c}\left(\frac{p_2-p_1}{B_2-B_1}\right)\left(e^{\mu_c(B_1-c)} - e^{\mu_c(B_2-c)}\right). \tag{3.12}$$

A similar analysis with ε replaced by $-\varepsilon$ shows that the opposite inequality is also true, so that $E/w_c(x-c)$ is $\sim$ to the right-hand side of (3.12). From the fact that $g(w)/w \to \beta$ as $w \downarrow 0$, (1.20) and (3.10), we have

$$\frac{(K_1 * g\circ w_c)'(x-c)}{w_c(x-c)} \sim p_1\beta e^{\mu_c(B_1-c)} - p_2\beta e^{\mu_c(B_2-c)} - \frac{\beta}{\mu_c}\left(\frac{p_2-p_1}{B_2-B_1}\right)\left(e^{\mu_c(B_1-c)} - e^{\mu_c(B_2-c)}\right) \quad \text{as } x\to+\infty. \tag{3.13}$$

We look at the middle term on the right side of (3.9). Since $g(w)-\beta w = O(w^{1+\varepsilon})$ as $w \downarrow 0$, our analysis on the term E shows that this middle term when divided by $w_c(x-c)$ goes to 0 as $x\to+\infty$. Hence we conclude from this, (3.9), (3.12), and (3.13) that

$$\lim_{x\to+\infty} \frac{w_c'(x-c)}{w_c(x-c)} = -\mu_c.$$

Q.E.D.

COROLLARY. *Let $c > c_+^*$; then $\omega_c(\gamma) \sim -\mu_c\gamma$ as $\gamma \downarrow 0$.*

Proof. This follows from Lemma 7 and (1.20).

LEMMA 8. *Let $K(x)$ be PF_2. Then $\omega_c(\gamma)$ is an increasing function in $c \ge c_+^*$ for every $0<\gamma<1$. Also $\lim_{c\downarrow c_+^*} \omega_c(\gamma) = w_{c_+^*}(\gamma)$ uniformly in compact subsets of $(0,1)$.*

Proof. Let $c_1 > c_2 > c_+^*$ and set

$$\bar{u}_0(x) = \begin{cases} 1, & x\le 0, \\ e^{-\mu_{c_1}x}, & x\ge 0, \end{cases} \qquad \bar{v}_0(x) = \begin{cases} 1, & x\le 0, \\ e^{-\mu_{c_2}x}, & x\ge 0. \end{cases}$$

Note that $\mu_{c_1} < \mu_{c_2}$ so that $\bar{v}_0(x) \le \bar{u}_0(x)$. Define recursively $\bar{u}_{n+1}(x) = Q_{c_1}[\bar{u}_n](x)$ and $\bar{v}_{n+1}(x) = Q_{c_2}[\bar{v}_n](x)$ for $n \ge 0$. First we show that $\bar{v}_n'(\bar{v}_n^{-1}(\gamma)) \le \bar{u}_n'(\bar{u}_n^{-1}(\gamma))$ for sufficiently large n. From the definitions of $\bar{u}_0$, $\bar{v}_0$ and Theorem 5, $\bar{u}_n^{-1}(\gamma)$ and $\bar{v}_n^{-1}(\gamma)$ are bounded in compact subsets of $(0,1)$. Clearly,

$$Q_{c_1}[f] = Q_{c_2}[f](x + c_1 - c_2) \quad \text{for } f\in\mathcal{C}. \tag{3.14}$$

Consider the function

$$\bar{v}_0\left(x + \bar{v}_n^{-1}(\gamma) - \bar{u}_n^{-1}(\gamma) - (c_1-c_2)n\right) - \bar{u}_0(x), \qquad 0<\gamma<1.$$

Since $c_1 > c_2$, $t = -\bar{v}_n^{-1}(\gamma) + \bar{u}_n^{-1}(\gamma) + (c_1-c_2)n$ is positive for sufficiently large n. From the definitions of $\bar{u}_0$, $\bar{v}_0$, $\bar{v}_0(x-t) - \bar{u}_0(x)$ has a sign change at $x_0 = \mu_{c_2}t/(\mu_{c_2}-\mu_{c_1}) > 0$ if $t > 0$. Hence for sufficiently large n, the function $g(\bar{v}_0(x-t)) - g(\bar{u}_0(x))$ has one sign

change. From Lemma 2 and (3.14),

$$\bar{v}_1\big(x+\bar{v}_n^{-1}(\gamma)-\bar{u}_n^{-1}(\gamma)-(c_1-c_2)(n-1)\big)-\bar{u}_1(x)$$

has no more than one sign change and has nonpositive derivative at its zero. By an induction argument, the same is true of the function $\bar{v}_n(x+\bar{v}_n^{-1}(\gamma)-\bar{u}_n^{-1}(\gamma))-\bar{u}_n(x)$. Therefore at the zero $x=\bar{u}_n^{-1}(\gamma)$, $\bar{v}_n'(\bar{v}_n^{-1}(\gamma))\le\bar{u}_n'(\bar{u}_n^{-1}(\gamma))$.

Next we show that $\bar{u}_n'(\bar{u}_n^{-1}(\gamma))$ converges to $\omega_{c_1}(\gamma)$ uniformly on compact subsets of $(0,1)$. The proof for $\bar{v}_n'(\bar{v}_n^{-1}(\gamma))$ is the same and will be omitted. From the corollary of Lemma 5 and Theorem 5, we see that $\bar{u}_n'(x)$ converges to $w_{c_1}'(x)$ uniformly in compact subsets of $\mathbb{R}$. Let $\mathbb{K}$ be a compact subset in $(0,1)$ and let Γ be a compact set in $\mathbb{R}$ which contains both $\{\bar{u}_n^{-1}(\gamma): n\ge 0, \gamma\in\mathbb{K}\}$ and $\{w_{c_1}^{-1}(\gamma):\gamma\in\mathbb{K}\}$. Write

$$\begin{aligned}&\bar{u}_n'\big(\bar{u}_n^{-1}(\gamma)\big)-w_{c_1}'\big(w_{c_1}^{-1}(\gamma)\big)\\ &\quad=\bar{u}_n'\big(\bar{u}_n^{-1}(\gamma)\big)-w_{c_1}'\big(\bar{u}_n^{-1}(\gamma)\big)+w_{c_1}'\big(\bar{u}_n^{-1}(\gamma)\big)-w_{c_1}'\big(w_{c_1}^{-1}(\gamma)\big).\end{aligned}\tag{3.15}$$

The first two terms on the right of (3.15) go to zero as $n\to+\infty$ uniformly in $\mathbb{K}$. From Theorem 5, we can show that $\bar{u}_n^{-1}(\gamma)$ and $w_{c_1}^{-1}(\gamma)$ are close for sufficiently large n depending on $\mathbb{K}$. The finiteness of $\|w_{c_1}''\|_\infty$ and the mean value theorem imply that the difference of the last two terms in (3.15) also goes to zero as $n\to+\infty$ uniformly in $\mathbb{K}$. Therefore, $\lim_{n\to+\infty}\bar{u}_n'(\bar{u}_n^{-1}(\gamma))=\omega_{c_1}(\gamma)$ uniformly in compact subsets of $(0,1)$. A similar conclusion holds for $\bar{v}_n'(\bar{v}_n^{-1}(\gamma))$. These conclusions, together with the first part of this proof, show that

$$\omega_{c_2}(\gamma)\le\omega_{c_1}(\gamma)\quad\text{for } c_1>c_2,\quad 0<\gamma<1.\tag{3.16}$$

Define $\bar{\omega}_{c_+^*}(\gamma)=\lim_{c\downarrow c_+^*}\omega_c(\gamma)$ in $(0,1)$. $\bar{\omega}_{c_+^*}(\gamma)$ is bounded since $\|w_c'\|_\infty\le\|K'\|_1+p_1+p_2$. It remains to show that $\bar{\omega}_{c_+^*}(\gamma)=\omega_{c_+^*}(\gamma)$. To this end, we fix $0<\gamma<1$ and normalize $w_c(x)$ so that $w_c(0)=\gamma$. (3.16) implies that $w_c(x)$ is nondecreasing (nonincreasing) in $c\ge c_+^*$ for $x<0$ ($x>0$) so that $\lim_{c\downarrow c_+^*}w_c(x)=\tilde{w}_{c_+^*}(x)$ exists, $\tilde{w}_{c_+^*}(-\infty)=1$ and $\tilde{w}_{c_+^*}(+\infty)=0$. The family of functions $\{w_c(x): c\ge c_+^*\}$ is uniformly bounded and equicontinuous. Thus the Arzelà–Ascoli theorem implies that $\tilde{w}_{c_+^*}(x)$ is continuous and $w_c(x)$ actuallly converges to $\tilde{w}_{c_+^*}(x)$ uniformly in $\mathbb{R}$ as $c\downarrow c_+^*$. A passage to the limit using the dominated convergence theorem shows that $\tilde{w}_{c_+^*}(x)$ is a fixed point of the operator $Q_{c_+^*}$ and since $\tilde{w}_{c_+^*}(x)$ is nonincreasing, $\tilde{w}_{c_+^*}(x)$ must be the travelling wave of minimum speed c_+^*. Differentiate $\tilde{w}_{c_+^*}(\tilde{w}_{c_+^*}^{-1}(\eta))=\eta$ and $w_c(w_c^{-1}(\eta))=\eta$ and obtain the relation,

$$\int_\gamma^{\tilde{w}_{c_+^*}(x)}\frac{1}{\tilde{\omega}_{c_+^*}(\eta)}\,d\eta=x=\int_\gamma^{w_c(x)}\frac{1}{\omega_c(\eta)}\,d\eta,\qquad c>c_+^*.$$

Rearrange this to get

$$\int_\gamma^{w_c(x)}\left(\frac{1}{\omega_c(\eta)}-\frac{1}{\tilde{\omega}_{c_+^*}(\eta)}\right)d\eta=\int_{w_c(x)}^{\tilde{w}_{c_+^*}(x)}\frac{d\eta}{\tilde{\omega}_{c_+^*}(\eta)}.\tag{3.17}$$

Fix $x\ne 0$ and let $c\downarrow c_+^*$. The right side of (3.17) goes to 0 by what we have just proved. Thus $\omega_c(\eta)$ converges to $\tilde{\omega}_{c_+^*}(\eta)$ a.e. in $(0,1)$ as $c\downarrow c_+^*$ and $\bar{\omega}_{c_+^*}(\eta)=\tilde{\omega}_{c_+^*}(\eta)$ a.e. in $(0,1)$. To prove equality in all of $(0,1)$, it suffices to show that $\bar{\omega}_{c_+^*}(\eta)$ is continuous. Since w_c' are bounded in $\mathbb{R}$ independently of c, $\omega_c(\eta)$ are uniformly bounded. From the proof of Lemma 6, $\|w_c''\|_\infty\le\beta(\|K'\|_1+p_1+p_2)^2$. Differentiating the definition of $\omega_c(\eta)$ with respect to η, we obtain

$$\omega_c'(\eta)=\frac{w_c''\big(w_c^{-1}(\eta)\big)}{\omega_c(\eta)}.$$

This implies that on each compact subset of $(0,1)$, the family of functions $\{\omega_c: c_+^* + 1 \geq c \geq c_+^*\}$ is equicontinuous. From the definition of $\bar{\omega}_{c_+^*}(\gamma)$ as the monotone limit of $\omega_c(\gamma)$ and the Arzelà–Ascoli theorem, we conclude that $\bar{\omega}_{c_+^*}(\gamma)$ is continuous in $(0,1)$. We already know that $\tilde{w}_{c_+^*}(x)$ is our travelling wave of minimum speed, thus $\tilde{\omega}_{c_+^*}(\gamma)$ is continuous. We therefore have $\bar{\omega}_{c_+^*}(\gamma) = \tilde{\omega}_{c_+^*}(\gamma) = \omega_{c_+^*}(\gamma)$ in $(0,1)$. The lemma is then proved by applying Dini's theorem. Q.E.D.

4. Behavior of $u_n(x)$ near the boundary of $\{u_n > 0\}$. Let $u_0 \in \mathcal{C}$. Then the assumption $u_0(+\infty) = 0$ will be discussed in the following two cases.

Case I. $u_0(x) = 0$ for $x \geq A$ and $u_0(x) > 0$, nonincreasing in $[A-\varepsilon, A)$ for some $\varepsilon > 0$.

Case II. $u_0(x) > 0$ and nonincreasing near $+\infty$, $\lim_{x \to +\infty} u_0(x) = 0$.

Note that no assumption is being made about the behavior of $u_0(x)$ near $-\infty$.

In case I we say that A is the right support of $u_0(x)$, and because $\operatorname{supp} K = [B_1, B_2]$, the right support of $u_n(x)$ is $A + nB_2$. We shall study the behavior of $u_n(x)$ as x approaches $A + nB_2$ in case I or as x approaches $+\infty$ in case II. $K(x)$ is not required to be PF_2 until the last Lemma 11. We recall condition (viii) of (1.5), which is

$$g'(u) \leq \beta = g'(0) \quad \text{in } [0,1]. \tag{4.1}$$

We also need the linearized equation

$$v_n(x) = \beta \int K(x-y) v_{n-1}(y)\, dy, \qquad n \geq 1, \tag{4.2}$$

where $v_0 \equiv u_0$. $v_n(x)$ also has right support $A + nB_2$ but $v_n(x)$ is no longer bounded above by one. Hence we define

$$g(v) = 1 \quad \text{for } v > 1 \tag{4.3}$$

only for the purpose of this section. From (4.1) it is obvious that

$$\frac{g(x) - g(y)}{x - y} \leq \beta \quad \text{for } x > y > 0. \tag{4.4}$$

LEMMA 9. *Let $u_0 = v_0 \in \mathcal{C}$ be given and let v_n be defined recursively through* (4.2). *Then given $\varepsilon > 0$ and positive integer N, there exist positive constants M_1, M_2, L, l such that*:

In case I,

$$\left(1 - \frac{n\varepsilon}{\beta}\right) v_n(x) - M_1 \sum_{m=1}^{n} \beta^{m-1} K^{*m} * I_{(-\infty, A-l+(n-m)B_2)}(x) \leq u_n(x) \leq v_n(x). \tag{4.5}$$

In case II,

$$\left(1 - \frac{n\varepsilon}{\beta}\right) v_n(x) - M_2 \sum_{m=1}^{n} \beta^{m-1} K^{*m} * I_{(-\infty, L)}(x) \leq u_n(x) \leq v_n(x) \tag{4.6}$$

uniformly for all $1 \leq n \leq N$.

Proof. Set $w_n = v_n - u_n$ for $n \geq 0$. Then $w_0 = 0$ and (4.1) implies that $w_n(x) \geq 0$ for all n and $x \in \mathbb{R}$. Also

$$v_{n+1}(x) - u_{n+1}(x) = \int K(x-y)\left[g(v_n(y)) - g(u_n(y)) + \beta v_n(y) - g(v_n(y))\right] dy.$$

Hence (4.4) implies that

$$0\le w_{n+1}(x)\le\int K(x-y)[\beta w_n(y)+\beta v_n(y)-g(v_n(y))]\,dy. \tag{4.7}$$

In case I, $\lim_{x\to A+nB_2}u_n(x)=0$ while in case II a simple induction argument shows that $\lim_{x\to+\infty}u_n(x)=0$. Now let $\varepsilon>0$, $N\ge1$ be given. In case I there are $l>0$, $M_1>0$ such that for all $n\le N$

$$\beta v_n(y)-g(v_n(y))\begin{cases}\le\varepsilon v_n(y), & y>A+nB_2-l,\\ \le M_1, & y\le A+nB_2-l.\end{cases} \tag{4.8}$$

In case II there are $L>0$, $M_2>0$ such that for all $n\le N$

$$\beta v_n(y)-g(v_n(y))\begin{cases}\le\varepsilon v_n(y), & y>L,\\ \le M_2, & y\le L.\end{cases} \tag{4.9}$$

Using (4.7), we get

$$\begin{aligned}0\le w_{n+1}(x)&\le\beta K*w_n(x)+\varepsilon\int_{A+nB_2-l}^{\infty}K(x-y)v_n(y)\,dy+M_1\int_{-\infty}^{A+nB_2-l}K(x-y)\,dy\\&\le\beta K*w_n(x)+\frac{\varepsilon v_{n+1}(x)}{\beta}+M_1K*I_{(-\infty,A+nB_2-l)}(x)\end{aligned} \tag{4.10}$$

for case I and

$$\begin{aligned}0\le w_{n+1}(x)&\le\beta K*w_n(x)+\varepsilon\int_{L}^{\infty}K(x-y)v_n(y)\,dy+M_2\int_{-\infty}^{L}K(x-y)\,dy\\&\le\beta K*w_n(x)+\frac{\varepsilon v_{n+1}(x)}{\beta}+M_2K*I_{(-\infty,L)}(x)\end{aligned} \tag{4.11}$$

for case II.

When $n=0$, (4.10) and (4.11) reduce to (4.5), (4.6) for the case $n=1$. Assume that (4.5) or (4.6) is true for $n=1,\cdots,J$, $J<N$. Then using (4.8) or (4.9) and (4.7) with $n=J$ together with our induction hypothesis we get:

In case I,

$$\begin{aligned}0\le w_{J+1}(x)&\le\beta K*w_J(x)+\frac{\varepsilon v_{J+1}(x)}{\beta}+M_1K*I_{(-\infty,A+JB_2-l)}(x)\\&\le\beta K*\left\{\frac{J\varepsilon v_J}{\beta}+M_1\sum_{m=1}^{J}\beta^{m-1}K^{*m}*I_{(-\infty,A-l+(J-m)B_2)}\right\}(x)\\&\quad+\frac{\varepsilon v_{J+1}(x)}{\beta}+M_1K*I_{(-\infty,A+JB_2-l)}(x)\\&=\frac{\varepsilon(J+1)v_{J+1}(x)}{\beta}+M_1\sum_{m=1}^{J+1}\beta^{m-1}K^{*m}*I_{(-\infty,A-l+(J+1-m)B_2)}(x),\end{aligned}$$

which is exactly (4.5) when $n=J+1$.

In case II,

$$0\le w_{J+1}(x)\le \beta K * w_J(x)+\frac{\varepsilon v_{J+1}(x)}{\beta}+M_2K * I_{(-\infty,L)}(x)$$

$$\le \beta K * \left\{\frac{J\varepsilon v_J(x)}{\beta}+M_2\sum_{m=1}^{J}\beta^{m-1}K^{*m} * I_{(-\infty,L)}(x)\right\}$$

$$+\frac{\varepsilon v_{J+1}(x)}{\beta}+M_2K * I_{(-\infty,L)}(x)$$

$$=\frac{\varepsilon(J+1)v_{J+1}(x)}{\beta}+M_2\sum_{m=1}^{J+1}\beta^{m-1}K^{*m} * I_{(-\infty,L)}(x),$$

which is (4.6) for $n=J+1$. Q.E.D.

LEMMA 10. *Let $u_0\in\mathcal{C}$ and suppose there exist positive constants $\bar{l},\bar{L}$ such that $u_0'(x)\le 0$ holds when $x>A-\bar{l}$ in case* I *and when $x>\bar{L}$ in case* II. *Then*

$$u_n'(x)\sim v_n'(x) \quad \text{as } x\to A+nB_2 \text{ and as } x\to+\infty$$

respectively in cases I *and* II.

Proof. We shall prove the two cases together. The corresponding conditions for case II will be given in parentheses next to the statement. From (4.2) we have, for $n\ge 1$,

$$u_{n+1}'(x)-v_{n+1}'(x)=\int K(x-y)\left[g'(u_n(y))u_n'(y)-\beta v_n'(y)\right]dy. \tag{4.12}$$

The statement that

$$v_n'(y)\le u_n'(y)\le 0 \quad \text{for } y>A+nB_2-\bar{l} \quad (y>\bar{L}+nB_2) \tag{4.13}$$

is true when $n=0$ by our hypothesis on u_0. Assume that it is true up to n, then (4.1) and our induction hypothesis imply that

$$g'(u_n(y))u_n'(y)-\beta v_n'(y)\ge 0$$

if $y>A+nB_2-\bar{l}$ ($y>\bar{L}+nB_2$). The induction step is completed by using this and (4.12).

Write (4.12) as

$$u_{n+1}'(x)-v_{n+1}'(x)=\int K(x-y)g'(u_n(y))\left[u_n'(y)-v_n'(y)\right]dy$$
$$+\int K(x-y)\left[g'(u_n(y))-\beta\right]v_n'(y)\,dy. \tag{4.14}$$

Let $\varepsilon>0$, $N\ge 1$ be given and choose $0<l_\varepsilon<\bar{l}$, $L_\varepsilon>\bar{L}$ such that

$$-\varepsilon\le g'(u_n(y))-\beta\le 0 \quad \text{for } y>A+nB_2-l_\varepsilon \ (y>L_\varepsilon), \quad n\le N. \tag{4.15}$$

Set $w_n(x)=u_n(x)-v_n(x)$, $n\ge 0$. From (4.14) and $n=0$ in (4.15), we get

$$w_1'(x)=\int K(x-y)\left[g'(u_0(y))-\beta\right]v_0'(y)\,dy$$

$$\le\frac{-\varepsilon v_1'(x)}{\beta} \quad \text{if } x>A+B_2-l_\varepsilon \ (x>B_2+L_\varepsilon).$$

Inductively, assume that

$$w_J'(x)\le\frac{-J\varepsilon v_J'(x)}{\beta} \quad \text{if } x>A+JB_2-l_\varepsilon \ (y>JB_2+L_\varepsilon).$$

(4.13), (4.14), (4.15) and our induction hypothesis together imply whenever $x>A+(J+1)B_2-l_\varepsilon$ $(x>L_\varepsilon+(j+1)B_2)$, we have

$$\begin{aligned} w'_{J+1}(x) &\le \beta\int K(x-y)w'_J(y)\,dy+\int K(x-y)\left[g'(u_J(y))-\beta\right]v'_J(y)\,dy \\ &\le \beta\int K(x-y)\left(\frac{-J\varepsilon v'_J(y)}{\beta}\right)dy-\varepsilon\int K(x-y)v'_J(y)\,dy \\ &= \frac{-(J+1)\varepsilon v'_{J+1}(x)}{\beta}. \end{aligned}$$

From this inequality and (4.13), we get

$$v'_n(x)\le u'_n(x)\le\left(1-\frac{n\varepsilon}{\beta}\right)v'_n(x) \quad \text{if } x>A+nB_2-l_\varepsilon\ (x>nB_2+L_\varepsilon),\quad n\le N,$$

which obviously imply our lemma. Q.E.D.

LEMMA 11. (i)

$$u_n(x)\sim v_n(x) \text{ as } x\to A+nB_2 (x\to+\infty) \tag{4.16}$$

uniformly for $1\le n\le N$, $N\in Z^+$.

(ii) *Let* $K(x)=e^{k(x)}$ *be* PF_2 *and let the hypotheses of Lemma* 10 *be satisfied. Then*

Case I. *Given* $m\ge 2$, $0<\varepsilon<1$ *there exists* $\gamma_0>0$ *such that* $0<\gamma\le\gamma_0$ *implies*

$$u'_m\left(u_m^{-1}(\gamma)\right)\le(1-\varepsilon)^2k'\left(u_m^{-1}(\gamma)-A-(m-1)B_2\right)\gamma+P_m(\gamma)\gamma \tag{4.17a}$$

if $k'(x)\le 0$ *in a left neighborhood of* B_2 *and*

$$u'_m\left(u_m^{-1}(\gamma)\right)\le(1-\varepsilon^2)k'\left(u_m^{-1}(\gamma)-A-(m-1)B_2\right)\gamma+P_m(\gamma)\gamma \tag{4.17b}$$

if $k'(x)\ge 0$ *in a left neighborhood of* B_2. *Here*

$$P_m(\gamma)=\frac{-(1-\varepsilon)^2p_2}{\int_{u_m^{-1}(\gamma)-A-(m-1)B_2}^{B_2}K(y)\,dy}.$$

Case II. *Assume in addition that for some* $\mu>0$, $u'_0(x)\sim-\mu u_0(x)$ *as* $x\to+\infty$. *Then* $u'_n(x)\sim-\mu u_n(x)$ *as* $x\to+\infty$ *for any* $n\ge 0$. *In particular, for a fixed* n,

$$u'_n\left(u_n^{-1}(\gamma)\right)\sim-\mu\gamma \quad \text{as } \gamma\downarrow 0. \tag{4.18}$$

Proof. (4.16) follows from (4.5) and (4.6) if we just observe that the terms inside the summation sign vanish for $x>A+nB_2-l$ in case I and $x>nB_2+L$ in case II. We turn to the proof of (4.17a) and (4.17b). l will be a positive constant chosen along the proof.

Given $m\ge 2$, it is clear that $v'_m(x)\le 0$ in $I_l=[A+mB_2-l, A+mB_2)$ for sufficiently small l. From Lemma 10, we have

$$u'_m(x)\le(1-\varepsilon)v'_m(x) \quad \text{in } I_l. \tag{4.19}$$

Now

$$v'_m(x)=\beta\int_{x-B_2}^{A+(m-1)B_2}e^{k(x-y)}k'(x-y)v_{m-1}(y)\,dy-p_2\beta v_{m-1}(x-B_2)$$

if $l<(B_2-B_1)$. Because $K(x)=e^{k(x)}$ is PF_2, $k'(x)$ is nonincreasing in x, and hence whenever defined,

$$k'(x-y)\le k'(x-A-(m-1)B_2) \quad \text{if } y<A+(m-1)B_2.$$

Since $k(x)$ is differentiable in a left neighborhood of B_2, we have from (4.19)

$$u_m'(x)\leq(1-\varepsilon)k'(x-A-(m-1)B_2)v_m(x)-p_2\beta(1-\varepsilon)v_{m-1}(x-B_2) \quad \text{in } I_l.$$

From (4.16), we have $(1-\varepsilon)u_m(x)\leq v_m(x)\leq(1+\varepsilon)u_m(x)$ if $x\in I_l$ and l sufficiently small. Thus

$$u_m'(x)\leq(1-\varepsilon)^2k'(x-A-(m-1)B_2)u_m(x)-p_2\beta(1-\varepsilon)v_{m-1}(x-B_2) \quad \text{in } I_l$$

if $k'(x)\leq 0$ in a left neighborhood of B_2,

$$u_m'(x)\leq(1-\varepsilon^2)k'(x-A-(m-1)B_2)u_m(x)-p_2\beta(1-\varepsilon)v_{m-1}(x-B_2) \quad \text{in } I_l$$

if $k'(x)\geq 0$ in a left neighborhood of B_2,

By definition, we have

$$v_m(x)=\beta\int_{x-B_2}^{A+(m-1)B_2}K(x-y)v_{m-1}(y)\,dy.$$

Since $v_{m-1}(y)$ is nonincreasing in a left neighborhood of $A+(m-1)B_2$, we have for l sufficiently small, $x\in I_l$, the inequality

$$v_m(x)\leq\beta v_{m-1}(x-B_2)\int_{x-A-(m-1)B_2}^{B_2}K(y)\,dy.$$

Hence

$$-p_2\beta v_{m-1}(x-B_2)\leq\frac{-p_2v_m(x)}{\int_{x-A-(m-1)B_2}^{B_2}K(y)\,dy}.$$

Since $u_m^{-1}(\gamma)$ converges to $A+mB_2$ as $\gamma\downarrow 0$, (4.17a), (4.17b) are proved if we replace x by $u_m^{-1}(\gamma)$ for $0<\gamma<\gamma_0$, γ_0 sufficiently small. Finally to prove (4.18), observe that

$$v_{n+1}'(x)=\beta\int K(x-y)v_n'(y)\,dy \quad \text{and} \quad v_0\equiv u_0.$$

Since $v_0'(x)\sim-\mu u_0(x)$ as $x\to+\infty$, a simple induction gives

$$v_n'(x)\sim-\mu v_n(x) \quad \text{as } x\to+\infty.$$

Then

$$\frac{u_n'(x)}{-\mu u_n(x)}=\frac{u_n'(x)}{v_n'(x)}\cdot\frac{v_n'(x)}{-\mu v_n(x)}\cdot\frac{v_n(x)}{u_n(x)}\to 1 \quad \text{as } x\to+\infty$$

by Lemma 10, (4.16) and what we have just said about $v_n(x)$. Q.E.D.

Remark. If $p_2>0$, then $P_m(\gamma)$ diverges to $-\infty$ as $\gamma\downarrow 0$. Also since $k'(x)$ is nonincreasing, $k'(u_m^{-1}(\gamma)-A-(m-1)B_2)$ is bounded above as $\gamma\downarrow 0$. Thus from (4.17a) or (4.17b), $u_m'(u_m^{-1}(\gamma))/\gamma$ diverges to $-\infty$ as $\gamma\downarrow 0$. On the other hand if $p_2=0$, then $K(x)$ is continuous at the point B_2. In this case $k'(x)$ is nonpositive to the left of B_2 so that only (4.17a) is relevant, $P_m(\gamma)\equiv 0$ and $k'(x)$ diverges to $-\infty$ as $x\to B_2$. Hence $u_m'(u_m^{-1}(\gamma))/\gamma$ diverges to $-\infty$ as $\gamma\downarrow 0$.

5. The case $u_0(-\infty)>0$. We are now ready to prove our main result (Theorem 6) for the case when the initial datum $u_0(x)$ satisfies the condition $\liminf_{x\to-\infty}u_0(x)>0$. Let $u_0\in\mathcal{C}$ be nontrivial and let u_n be defined by the recursion (1.16). According to Theorem 2, the maximum of $u_n(x)$ will converge to 1. In particular, $u_n^{-1}(\gamma)$ will be defined in any compact subset of $(0,1)$ for sufficiently large n. We introduce the

notation $m_n(\gamma)=u_n^{-1}(\gamma)$ for $n\geq 0$. Then the results in §1 imply the important fact that for every $0<\gamma<1$,

$$\lim_{n\to+\infty}\frac{m_n(\gamma)}{n}=c_+^*. \tag{5.1}$$

LEMMA 12. *Let $u_0\in\mathcal{C}$ and u_n be such that*

$$\lim_{n\to+\infty}u_n(x+m_n(\gamma))=q(x)\quad \textit{uniformly in } [nc_0^*-m_n(\gamma),+\infty),$$

where $c_0^*\in(-c_+^*,c_+^*)$ and $q(-\infty)=1$, $q(+\infty)=0$, $q(x)$ is nonincreasing. Then for every $0<\gamma<1$,

$$\lim_{n\to+\infty}m_{n+1}(\gamma)-m_n(\gamma)=c_+^* \tag{5.2}$$

and

$$Q_{c_+^*}[q](x)=q(x).$$

Proof. Consider first the boundedness of the sequence $\{m_{n+1}(\gamma)-m_n(\gamma)\}$ for a fixed $\gamma\in(0,1)$. Suppose there exists a subsequence $\{n_j\}$ such that $m_{n_j+1}(\gamma)-m_{n_j}(\gamma)$ diverges to $+\infty$. Then for y bounded, our assumption on c_0^* and (5.1) imply that $y+m_{n_j+1}(\gamma)-m_{n_j}(\gamma)\geq n_jc_0^*-m_{n_j}(\gamma)$ for sufficiently large j. Letting j go to $+\infty$ in the equation

$$u_{n_j+1}\big(x+m_{n_j+1}(\gamma)\big)=\int K(x-y)g\Big(u_{n_j}\big(y+m_{n_j+1}(\gamma)-m_{n_j}(\gamma)+m_{n_j}(\gamma)\big)\Big)dy, \tag{5.3}$$

we see that $q(x)\equiv 0$ in $\mathbb{R}$. This is impossible since $q(0)=\gamma$. Similarly if $m_{n_j+1}(\gamma)-m_{n_j}(\gamma)$ diverges to $-\infty$ as $j\to+\infty$, the same argument will give us the contradiction that $q(x)\equiv 1$. Therefore $m_{n_j+1}(\gamma)-m_{n_j}(\gamma)$ is bounded. Next we show that any two subsequences $\{m_{n_i+1}(\gamma)-m_{n_i}(\gamma)\}$, $\{m_{n_j+1}(\gamma)-m_{n_j}(\gamma)\}$ that converge to l_1, l_2 respectively must converge to the same limit. As before, (5.3) gives, for the two subsequences,

$$q(x)=\int K(x-y)g(q(l_1+y))\,dy\quad\text{and}\quad q(x)=\int K(x-y)g(q(l_2+y))\,dy.$$

Hence $q(x+l_1-l_2)=q(x)$ in $\mathbb{R}$. Since $q(0)=\gamma$, an induction argument implies that $q(n(l_1-l_2))=\gamma\in(0,1)$ for all $n\geq 0$. If $l_1>l_2$, this will contradict the fact that $q(+\infty)=0$ while if $l_1<l_2$, it will contradict the fact that $q(-\infty)=1$. Hence $l_1=l_2\equiv l$. We claim that $l=c_+^*$. To see this, let $\varepsilon>0$ and N_ε be such that $l-\varepsilon\leq m_{n+1}(\gamma)-m_n(\gamma)\leq l+\varepsilon$ for $n\geq N_\varepsilon$. Summing from $n=N_\varepsilon$ to $n=M-1>N_\varepsilon$, we get $(M-N_\varepsilon)(l-\varepsilon)\leq m_M(\gamma)-m_{N_\varepsilon}(\gamma)\leq(M-N_\varepsilon)(l+\varepsilon)$. Divide by M and let $M\to+\infty$. Using (5.1) and let $\varepsilon\downarrow 0$, we get $l=c_+^*$. Finally, we let $j\to+\infty$ in (5.3) and get $Q_{c_+^*}[q](x)=q(x)$. Q.E.D.

We now prove a special case of our main theorem. The notations follow that of Lemma 4.

LEMMA 13. *Let $K(x)$ be PF_2, $f_0(x)=H(-x)$ and define recursively $f_{n+1}=Q[f_n]$ for $n\geq 0$. Then, for $0<\gamma<1$, $f_n(x+f_n^{-1}(\gamma))$ increases uniformly to $w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))$ in $[0,+\infty)$, and $f_n(x+f_n^{-1}(\gamma))$ decreases uniformly to $w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))$ in $(-\infty,0]$. Also $\lim_{n\to+\infty}f_n'(f_n^{-1}(\gamma))=\omega_{c_+^*}(\gamma)$ uniformly in compact subsets of $(0,1)$.*

Proof. We first show that for each $0<\gamma<1$, $f_n'(f_n^{-1}(\gamma))$ is nondecreasing in n. Let $n=\bar{n}$ be fixed and set $b=f_{\bar{n}+1}^{-1}(\gamma)-f_{\bar{n}}^{-1}(\gamma)$. Consider the function $v_n(x)=f_n(x)-f_{n+1}(x+b)$, $n\geq 0$. We have $v_0(x)\geq 0$ for $x<0$ and $v_0(x)\leq 0$ for $x>0$. Since $g(u)$ does not increase the number of sign changes for the difference of two functions, the same relation will be true for the function $g(f_0(x))-g(f_1(x+b))$. From Lemma 3, the

function $v_1(x)$ will have no more than one sign change and the derivative at a point where $v_1(x)$ vanishes is nonpositive. We repeat the argument inductively and when $j=\bar{n}$, since $v_{\bar{n}}(f_{\bar{n}}^{-1}(\gamma))=0$, we have $f'_{\bar{n}}(f_{\bar{n}}^{-1}(\gamma))\le f'_{\bar{n}+1}(f_{\bar{n}+1}^{-1}(\gamma))$.

To continue, since $f'_n(f_n^{-1}(\gamma))\le 0$, $f'_n(f_n^{-1}(\gamma))$ must converge pointwise to a limit function $\varphi(\gamma)$ and $\varphi(\gamma)\le 0$. From the relation $f_n(f_n^{-1}(\varepsilon))=\varepsilon$ in $(0,1)$, we obtain (see Lemma 11)

$$\int_\gamma^{f_n(x+f_n^{-1}(\gamma))} \frac{d\varepsilon}{-f'_n(f_n^{-1}(\varepsilon))} = -x = -\int_\gamma^{f_n(x+f_n^{-1}(\gamma))} df_n^{-1}(\varepsilon). \tag{5.4}$$

Consider a fixed $x<0$ (the case $x>0$ being similar). We have $f_n(x+f_n^{-1}(\gamma))>\gamma$ for all n. Also $f'_n(f_n^{-1}(\varepsilon))\uparrow\varphi(\varepsilon)$ as $n\to+\infty$ so that

$$-\frac{1}{f'_n(f_n^{-1}(\varepsilon))}\uparrow -\frac{1}{\varphi(\varepsilon)} \quad \text{as } n\to+\infty.$$

The only way for the left-hand side of (5.4) to remain constant as we increase n is for $f_n(x+f_n^{-1}(\gamma))$ to decrease monotonically in n. Similarly, if $x>0$, $f_n(x+f_n^{-1}(\gamma))$ will have to increase monotonically in n. Denote by $q(x)$ the limit of $f_n(x+f_n^{-1}(\gamma))$ as $n\to+\infty$. $q(x)$ is nonincreasing, $q(0)=\gamma$. We claim that $q(-\infty)=1$ and $q(+\infty)=0$. To see that, recall from Lemma 4 with $u_0(x)=w_{c_+^*}(x)$ that $f'_n(f_n^{-1}(\gamma))\le\omega_{c_+^*}(\gamma)$ uniformly in each compact subset of $(0,1)$ for sufficiently large n. Hence $-1/\varphi(\varepsilon)\le -1/\omega_{c_+^*}(\varepsilon)$, which is integrable in compact subsets of $(0,1)$. The dominated convergence theorem may then be used in (5.4) to give us $\int_\gamma^{q(x)}(1/\varphi(\varepsilon))\,d\varepsilon=x$ for each fixed x where $q(x)\ne 0,1$. Letting x go to $\pm\infty$, we conclude that $q(+\infty)=0$ and $q(-\infty)=1$. Now the sequence of functions $\{f_n(x+f_n^{-1}(\gamma))\}$ is uniformly bounded and equicontinuous. Thus $q(x)$ is continuous and Dini's theorem implies that $f_n(x+f_n^{-1}(\gamma))$ converges to $q(x)$ uniformly in compact subsets of $\mathbb{R}$. The fact that $q(+\infty)=0$, $q(-\infty)=1$ is enough to allow us to conclude that the convergence is actually uniform in $\mathbb{R}$. The first half of the lemma follows from Lemma 12.

Finally, we have

$$\int_\gamma^{f_n(x+f_n^{-1}(\gamma))} \frac{d\varepsilon}{f'_n(f_n^{-1}(\varepsilon))} = x = \int_\gamma^{w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))} \frac{d\varepsilon}{\omega_{c_+^*}(\varepsilon)},$$

so that

$$\int_\gamma^{f_n(x+f_n^{-1}(\gamma))} \left(\frac{1}{f'_n(f_n^{-1}(\varepsilon))} - \frac{1}{\omega_{c_+^*}(\varepsilon)} \right) d\varepsilon = \int_{f_n(x+f_n^{-1}(\gamma))}^{w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))} \frac{d\varepsilon}{\omega_{c_+^*}(\varepsilon)}.$$

By what we have proved, the right side now goes to zero as $n\to+\infty$. Thus $\lim_{n\to+\infty} f'_n(f_n^{-1}(\gamma))=\omega_{c_+^*}(\gamma)$ a.e. in $(0,1)$. But $f'_n(f_n^{-1}(\gamma))$ converges to $\varphi(\gamma)$ in $(0,1)$. Thus $\varphi(\gamma)=\omega_{c_+^*}(\gamma)$ a.e. in $(0,1)$. To prove equality everywhere, it suffices to show that $\varphi(\gamma)$ is continuous in $(0,1)$. The proof is similar to showing that $\bar{\omega}_{c_+^*}(\eta)$ is continuous near the end of the proof of Lemma 8. Here we should show that $\{f'_n(f_n^{-1}(\gamma)): n\ge 1\}$ is uniformly bounded and equicontinuous. By the Arzelà–Ascoli theorem, $f'_n(f_n^{-1}(\gamma))$ converges uniformly in each compact subset of $(0,1)$ to $\varphi(\gamma)$ which then must be continuous. Since $f'_n(f_n^{-1}(\gamma))$ converges to $\varphi(\gamma)$ monotonically, Dini's theorem says the convergence is uniform in every compact subset of $(0,1)$. Lemma 13 is now proved. Q.E.D.

COROLLARY. *Let $K(x)$ be PF_2, $u_0\in\mathcal{C}$ be such that $u_0(+\infty)=0$. Then*

$$\liminf_{n\to+\infty} u'_n(m_n(\gamma)) \ge \omega_{c_+^*}(\gamma)$$

uniformly in compact subsets of $(0,1)$.

Proof. This is an immediate consequence of Lemmas 4 and 13.

THEOREM 6. *Let $K(x)$ be PF_2. Let $u_0 \in \mathcal{C}$ be such that* $\liminf_{x\to-\infty} u_0(x)>0$ *and either*

Case I. $u_0(x)=0$ *for* $x\geq A$, $u_0(x)>0$ *and nonincreasing in* $[A-\varepsilon, A)$ *for some* $\varepsilon>0$, *or*

Case II. $u_0(x)>0$ *near* $+\infty$, $u_0'(x)\sim -\mu u_0(x)$ *as* $x\to+\infty$ *for some* $\mu>\mu_{c_+^*}$.

Then for each $0<\gamma<1$, *we have*

$$\lim_{n\to+\infty} u_n(x+m_n(\gamma))=w_{c_+^*}\left(x+w_{c_+^*}^{-1}(\gamma)\right) \quad \textit{uniformly in } \mathbb{R}$$

and

$$\lim_{n\to+\infty} m_{n+1}(\gamma)-m_n(\gamma)=c_+^*.$$

Proof. We first show that under our assumptions on $u_0(x)$, the function $u_m(x)$ will satisfy the one-sided condition for sufficiently large m. The condition $\liminf_{x\to-\infty} u_0(x)>0$ implies that there exists an L and a $\delta>0$ such that $\inf_{x<L} u_0(x)>\delta$. From the fact that $\operatorname{supp} K=[B_1, B_2]$ and the definition of $u_m(x)$, it follows that $u_m(x)>0$ for $x<L+mB_2$ and in fact

$$\inf_{x<L+mB_1} u_m(x)\geq g^m(\delta) \quad \text{for } m\geq 0. \tag{5.5}$$

On the other hand, $u_0(x)>0$ in $[A-\varepsilon, A)$ implies $u_m(x)>0$ in $[A+mB_1-\varepsilon, A+mB_2)$ in case I. In case II, suppose $u_0(x)>0$ for $x\geq \bar{L}$, then the same will be true for $u_m(x)$ in $(\bar{L}+mB_1, +\infty)$ for $m\geq 0$. Choose m so large that $L+mB_2>A+mB_1-\varepsilon$ in case I or $L+mB_2>\bar{L}+mB_1$ in case II. Then $u_m(x)>0$ in the interval $(-\infty, A+mB_2)$ in case I, $u_m(x)>0$ in $\mathbb{R}$ in case II. From the hypothesis on $u_0(x)$ and an induction argument, we see that $u_m(x)$ is nonincreasing in a left neighborhood of $A+mB_2$ in case I and for x sufficiently large in case II. These two facts plus (5.5) imply that $u_m(x)$ satisfies the one-sided condition of Lemma 3. In particular, there exists x_0 such that $u_m(x)\geq u_m(x_0)\equiv\delta_0>0$ for $x<x_0$, $u_0(x)$ is nonincreasing for $x>x_0$ and $u_m(+\infty)=0$.

The main goal of the rest of the proof is to show that

$$\lim_{n\to+\infty} u_n'(m_n(\gamma))=\omega_{c_+^*}(\gamma) \quad \text{uniformly in compact subsets of } (0,1). \tag{5.6}$$

We actually show (5.6) by working with $\omega_c(\gamma)$, $c>c_+^*$ instead, and apply Lemma 8 at the end. $c>c_+^*$ will therefore be temporary fixed. Let $K=[\underline{\gamma}, \bar{\gamma}]$ be a compact subset of $(0,1)$ and choose m_1 so large that $g^{m_1}(\delta_0)>\bar{\gamma}$. Let $v_0(x)=u_{m+m_1}(x)$ be our initial datum and set $v_n(x)=Q[v_{n-1}](x)$ for $n\geq 1$. Choose $\gamma_1\in(\bar{\gamma}, g^{m_1}(\delta_0))$ and define

$$\bar{w}_c(x)=\begin{cases} w_c(x), & x>w_c^{-1}(\gamma_1), \\ \gamma_1, & x\leq w_c^{-1}(\gamma_1). \end{cases}$$

Set $\bar{w}_{0,c}(x)=\bar{w}_c(x)$ and $\bar{w}_{n,c}(x)=Q_c[\bar{w}_{n-1,c}](x)$ for $n\geq 1$. By Theorem 5, $\bar{w}_{n,c}(x)$ converges uniformly in every compact subset of $\mathbb{R}$ to $w_c(x)$. Therefore $\bar{w}_{n,c}^{-1}(\gamma)$ is bounded for $\gamma\in\mathbb{K}$, $n\geq 0$. Let $t_n(\gamma)=nc-v_n^{-1}(\gamma)+\bar{w}_{n,c}^{-1}(\gamma)$ and consider the function

$$v_0(x-t_n(\gamma))-\bar{w}_{0,c}(x). \tag{5.7}$$

Our immediate goal is to show that (5.7) has one sign change for all $\gamma\in\mathbb{K}$ when n is sufficiently large. We observe the following about $v_0(x)$. Since $u_m(x)$ satisfies the

one-sided condition, Lemma 3 may be applied. From the definition of $v_0(x)$, the following is true:

(5.8) There exists $\bar{x}$ such that $v_0(x) \geq g^{m_1}(\delta_0)$ for $x \leq \bar{x}$, $v_0(\bar{x}) = g^{m_1}(\delta_0)$, $v_0(x)$ is nonincreasing for $x > \bar{x}$.

Also $v_n^{-1}(\gamma) \leq v_n^{-1}(\gamma)$ for all $\gamma \in \mathbb{K}$. From (5.1), $v_n^{-1}(\gamma) \sim nc_+^*$ as $n \to +\infty$. These plus our earlier remark about $\bar{w}_{n,c}^{-1}(\gamma)$ being bounded and the fact that $c > c_+^*$ imply

$$t_n(\gamma) \to +\infty \text{ as } n \to +\infty \quad \text{uniformly for } \gamma \in \mathbb{K}. \tag{5.9}$$

Our immediate goal is accomplished through showing that there exists $\gamma_2 < \bar{\gamma}$ such that

$$v_0'\big(v_0^{-1}(\gamma)\big) < \omega_c(\gamma) \quad \text{for } 0 < \gamma \leq \gamma_2 \tag{5.10}$$

and, for each n,

$$v_0\big(x - t_n(\gamma)\big) - \bar{w}_{0,c}(x) < 0 \quad \text{as } x \to +\infty. \tag{5.11}$$

Assuming (5.10) and (5.11), we see from our definition of $\bar{w}_{0,c}(x)$ that $\bar{w}_{0,c}'(\bar{w}_{0,c}^{-1}(\gamma)) = \omega_c(\gamma)$ for small enough γ, let us say when $0 < \gamma \leq \gamma_2$ as well. (5.10) then says $v_0(x - t_n(\gamma)) - \bar{w}_{0,c}(x)$ cannot have more than one sign change in the interval $[\bar{w}_{0,c}^{-1}(\gamma_2), +\infty)$. From (5.9), $t_n(\gamma) + \bar{x} > \bar{w}_{0,c}^{-1}(\gamma)$ for all $\gamma \in \mathbb{K}$ and sufficiently large n. Then if $x \leq \bar{w}_{0,c}^{-1}(\gamma_2)$, we have $x - t_n(\gamma) \leq \bar{x}$, and hence (5.8) implies that $v_0(x - t_n(\gamma)) \geq v_0(\bar{x}) \geq g^{m_1}(\delta_0) > \gamma_1 \geq \bar{w}_{0,c}(x)$ by the definition of $\bar{w}_{0,c}(x)$. This fact and (5.11) then say there is at least one sign change in the interval $[\bar{w}_{0,c}^{-1}(\gamma_2), +\infty)$. By our earlier remark, there is only one. We now turn to the proofs of (5.10) and (5.11) but we have to discuss case I and case II separately.

Case I. Our remark after Lemma 11 says that $v_0'(v_0^{-1}(\gamma))/\gamma$ diverges to $-\infty$ as $\gamma \downarrow 0$. Hence there exists a $\gamma_3 > 0$ such that

$$v_0'\big(v_0^{-1}(\gamma)\big) \leq -2\mu_{c_+^*}\gamma \quad \text{for } 0 < \gamma < \gamma_3. \tag{5.12}$$

From the corollary after Lemma 7, given $0 < \varepsilon < 1$, there exists $\gamma_4 > 0$ such that

$$-(1+\varepsilon)\mu_c\gamma \leq \omega_c(\gamma) \leq -(1-\varepsilon)\mu_c\gamma \quad \text{for } 0 < \gamma \leq \gamma_4. \tag{5.13}$$

Combining (5.12) and (5.13) and the fact that $\mu_{c_+^*} > \mu_c$, we have

$$v_0'\big(v_0^{-1}(\gamma)\big) < \omega_c(\gamma) \quad \text{for } 0 < \gamma < \gamma_2,$$

where $\gamma_2 = \min(\gamma_3, \gamma_4)$. (5.11) is trivial in case I, since $v_0(x)$ vanishes for large x and $\bar{w}_{0,c}(x) = w_c(x) > 0$ for x near $+\infty$.

Case II. From our assumption that $u_0'(x) \sim -\mu u_0(x)$ as $x \to +\infty$ and case II of (ii) in Lemma 11, we have $u_n'(m_n(\gamma)) \sim -\mu\gamma$ as $\gamma \downarrow 0$ for every fixed n. In particular $v_0'(v_0^{-1}(\gamma)) \sim -\mu\gamma$ as $\gamma \downarrow 0$. Since $\mu > \mu_{c_+^*} > \mu_c$, we can find an ε in $(0, 1)$ such that

$$\mu > \left(\frac{1+\varepsilon}{1-\varepsilon}\right)\mu_c. \tag{5.14}$$

For such an ε, we have from above and the corollary of Lemma 7

$$-(1+\varepsilon)\mu\gamma \leq v_0'\big(v_0^{-1}(\gamma)\big) \leq -(1-\varepsilon)\mu\gamma \quad \text{if } 0 < \gamma < \gamma_5 \tag{5.15}$$

and

$$-(1+\varepsilon)\mu_c\gamma \leq \omega_c(\gamma) \leq -(1-\varepsilon)\mu_c\gamma \quad \text{if } 0 < \gamma < \gamma_5. \tag{5.16}$$

From (5.14), (5.15) and (5.16), we conclude that $v_0'(v_0^{-1}(\gamma))<\omega_c(\gamma)$ for $0<\gamma<\gamma_5$. Let $0<\gamma_6<\gamma_5$ and write for $\tilde{\gamma}\in(0,\gamma_6)$,

$$v_0^{-1}(\tilde{\gamma})-\overline{w}_{0,c}^{-1}(\tilde{\gamma})=v_0^{-1}(\gamma_6)-\overline{w}_{0,c}^{-1}(\gamma_6)+\int_{\tilde{\gamma}}^{\gamma_6}\left(\frac{1}{\omega_c(\eta)}-\frac{1}{v_0'(v_0^{-1}(\eta))}\right)d\eta.$$

From (5.15) and (5.16),

$$\begin{aligned}\int_{\tilde{\gamma}}^{\gamma_6}\left(\frac{1}{\omega_c(\eta)}-\frac{1}{v_0'(v_0^{-1}(\eta))}\right)d\eta &\le\int_{\tilde{\gamma}}^{\gamma_6}\left[\frac{-1}{(1+\varepsilon)\mu_c\eta}+\frac{1}{(1-\varepsilon)\mu\eta}\right]d\eta \\ &=\left[\frac{1}{(1-\varepsilon)\mu}-\frac{1}{(1+\varepsilon)\mu_c}\right]\int_{\tilde{\gamma}}^{\gamma_6}\frac{1}{\eta}d\eta.\end{aligned} \tag{5.17}$$

By (5.14), (5.17) diverges to $-\infty$ as $\tilde{\gamma}\downarrow 0$. Hence for fixed n, $\gamma\in\mathbb{K}$ and sufficiently small $\tilde{\gamma}$, $t_n(\gamma)+v_0^{-1}(\tilde{\gamma})<\overline{w}_{0,c}^{-1}(\tilde{\gamma})$. This simply means (5.11) is true because $v_0^{-1}(\tilde{\gamma})\to+\infty$ as $\tilde{\gamma}\downarrow 0$.

Now that we have finished proving (5.10), (5.11), we know that the function (5.7) has one sign change for all $\gamma\in\mathbb{K}$, $n\ge N$ where N depends only on $\mathbb{K}$. The rest is then easy. Since g does not increase the number of sign changes for the difference of two functions, $g(v_0(x+v_n^{-1}(\gamma)-\overline{w}_{n,c}^{-1}(\gamma)-nc))-g(\overline{w}_{0,c}(x))$ has one sign change. From Lemma 2, we see that the function $v_1(x+v_n^{-1}(\gamma)-\overline{w}_{n,c}^{-1}(\gamma)-(n-1)c)-\overline{w}_{1,c}(x)$ has no more than one sign change and has a nonpositive derivative at its zero. By an induction argument, we arrive at the same conclusion for the function $v_n(x+v_n^{-1}(\gamma)-\overline{w}_{n,c}^{-1}(\gamma))-\overline{w}_{n,c}(x)$. At the zero $x=\overline{w}_{n,c}^{-1}(\gamma)$, we have $v_n'(v_n^{-1}(\gamma))\le\overline{w}_{n,c}'(\overline{w}_{n,c}^{-1}(\gamma))$ for $\gamma\in\mathbb{K}$, $n\ge N$. An argument exactly like the one given in the middle of the proof of Lemma 8 implies that $\overline{w}_{n,c}'(\overline{w}_{n,c}^{-1}(\gamma))$ converges uniformly in $\mathbb{K}$ to $\omega_c(\gamma)$. Hence $\limsup_{n\to+\infty}u_n'(m_n(\gamma))\le\omega_c(\gamma)$ uniformly in $\mathbb{K}$. From Lemma 8 and the corollary of Lemma 13, we conclude that (5.6) is true. It also follows from (5.6) that $u_n'(m_n(\gamma))\cdot dm_n(\gamma)/d\gamma=1$ is valid in each compact subset of $(0,1)$ for sufficiently large n. In such a case, we obtain

$$\int_{\gamma}^{w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))}\left(\frac{1}{\omega_{c_+^*}(\varepsilon)}-\frac{1}{u_n'(m_n(\varepsilon))}\right)d\varepsilon=\int_{w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))}^{u_n(x+m_n(\gamma))}\frac{d\varepsilon}{u_n'(m_n(\varepsilon))}. \tag{5.18}$$

From (5.18), (5.6) and the fact that $\|u_n'\|_\infty\le\|K'\|_1+p_1+p_2$, we see that $u_n(x+m_n(\gamma))$ converges to $w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))$ uniformly in compact subsets of $\mathbb{R}$. Since $w_{c_+^*}(-\infty)=1$ and $w_{c_+^*}(+\infty)=0$, the convergence is actually uniform in all of $\mathbb{R}$. The first statement of the theorem is now proved. The second statement follows from Lemma 12. Q.E.D.

Acknowledgment. This paper is a part of the author's doctoral dissertation, written under the guidance of Professor Hans Weinberger, to whom the author wishes to express his sincere thanks.

REFERENCES

[1] D. G. Aronson, *The asymptotic speed of propagation of a simple epidemic*, Nonlinear Diffusion, W. E. Fitzgibbon and H. F. Walker, eds., Research Notes in Mathematics 14, Pitman, London, 1977, pp. 1–23.

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

[3] ———, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. in Math., 30 (1978), pp. 33–76.

[4] A. D. BARBOUR, *The uniqueness of Atkinson and Reuter's epidemic waves*, Math. Proc. Camb. Phil. Soc., 82 (1977), pp. 127–130.

[5] M. H. DELANGE, *Généralisation du théorème de Ikehara*, Ann. Sci. École Norm. Sup., (3) 71 (1954), pp. 213–242.

[6] O. DIEKMANN, *Run for your life. A note on the asymptotic speed of propagation of an epidemic*, J. Differential Equations, 33 (1979), pp. 58–73.

[7] O. DIEKMANN AND H. G. KAPER, *On the bounded solutions of a nonlinear convolution equation*, J. Nonlin. Analysis-Theory, Methods and Applica., 2 (1978), pp. 721–737.

[8] R. A. FISHER, *The advance of advantageous genes*, Ann. Eugenics, 7 (1937), pp. 355–369.

[9] S. KARLIN, *Total Positivity*, vol. I, Stanford Univ. Press, Stanford, CA, 1968.

[10] A. KOLMOGOROFF, I. PETROVSKY AND N. PISCOUNOFF, *Étude de l'équations de la diffusion avec croissance de la quantité de matierè et son application a un problème biologique*, Bull. Univ. Moscow, Ser. Internat., Sec. A, 1 (1937), no. 6, pp. 1–25.

[11] R. LUI, *A nonlinear integral operator arising from a model in population genetics*, Ph.D. thesis, Univ. of Minnesota, August 1981.

[12] _____, *A nonlinear integral operator arising from a model in population genetics II. Initial data with compact support*, this Journal, this issue, pp. 938–953.

[13] H. ROST, *Proc. Conference on Models of Biological Growth and Spread—Mathematical Theories and Applications*, W. Jager, H. Rost and P. Tautu, eds., Lecture Notes in Biomathematics, 38, Springer-Verlag, New York, 1981.

[14] F. ROTHE, *Convergence to pushed fronts*, Rocky Mountain J. Math., to appear.

[15] H. R. THIEME, *A model for the spatial spread of an epidemic*, J. Math. Biol., 4 (1977), pp. 337–351.

[16] K. UCHIYAMA, *The behavior of solutions of some nonlinear diffusion equations for large time*, J. Math. Kyoto U., 18 (1978), pp. 453–508.

[17] H. F. WEINBERGER, *Asymptotic behavior of a model in population genetics*, Nonlinear Partial Differential Equations and Applications, J. Chadam, ed., in Lecture Notes in Mathematics 648, Springer, New York, 1978, pp. 47–98.

[18] _____, *Long-time behavior of a class of biological models*, this Journal, 13 (1982), pp. 353–396.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0003 $01.00/0

A NONLINEAR INTEGRAL OPERATOR ARISING FROM A MODEL IN POPULATION GENETICS, II. INITIAL DATA WITH COMPACT SUPPORT*

ROGER LUI†

Abstract. We study the asymptotic behavior of the solutions to the recursion $u_{n+1}(x)=Q[u_n](x)$ for $n\geq 0$. Here $Q[u](x)=(K * g\circ u)(x)$ acts on functions bounded between 0 and 1, $K(x)$ is a probability density function with compact support and $g(u)\in C^1[0,1]$ satisfies certain additional assumptions. It is known that there exist $-c_-^*<c_+^*$ such that for $c\geq c_+^*$, there are nonincreasing travelling waves $w_c(x)$ facing right, while for $c\leq -c_-^*$, there are nondecreasing travelling waves $\bar{w}_c(x)$ facing left. We prove here that if $K(x)$ has certain variation diminishing property, $g(u)/u$ is nonincreasing and $u_0(x)$ has compact support, then $u_n(x)$ develops uniformly in x, as n approaches infinity, into a pair of diverging waves, $w_{c_+^*}(x)$ on the right and $\bar{w}_{c_-^*}(x)$ on the left. This paper is a sequel to [R. Lui, SIAM J. Math. Anal., 13 (1982), pp. 913–937], which proved convergence to a single travelling wave when $u_0(-\infty)>0$.

1. Introduction. In [3] (this issue, pp. 913–937) we studied the asymptotic behavior as $n\to+\infty$, of the solution to the recursion $u_{n+1}(x)=Q[u_n](x)$ for a given $u_0(x)$. Q is a nonlinear integral operator,

$$Q[u](x)=\int_{\mathbb{R}} K(x-y)g(u(y))\,dy$$

defined on the set of functions

$$\mathcal{C}=\{u: 0\leq u\leq 1,\ u \text{ piecewise continuous}\}.$$

The following assumptions were made about $K(x)$ and $g(u)$ (see [3, §1]):

(i) $\operatorname{supp} K=[B_1,B_2]$, $K(x)>0$ in (B_1,B_2),

(ii) $K(x)$ is continuous in $\mathbb{R}$ except possibly at B_1, B_2 where

$$\lim_{x\downarrow B_1} K(x)=p_1,\quad \lim_{x\uparrow B_2} K(x)=p_2,\quad p_1,p_2\geq 0,$$

(iii) $K(x)$ is of bounded variation and is differentiable in $(B_2-\varepsilon,B_2)$ for some $\varepsilon>0$,

(iv) $\int_{\mathbb{R}} K(x)\,dx=1$,

(v) $g(u)\in C^1[0,1]$,

(vi) $g(0)=0$, $g(1)=1$,

(vii) $g(u)>u$ in $(0,1)$,

(viii) $0\leq g'(u)\leq g'(0)$ in $[0,1]$,

(ix) $g(u)=g'(0)u+0(u^{1+\varepsilon})$ as $u\downarrow 0$ for some $\varepsilon>0$,

(x) $g'(u)>0$ in $[0,\sigma)$, where $\sigma=\sup\{u: g(u)<1\}$,

(xi) $g(u)/u$ is nonincreasing in $[0,1]$.

From the results of Weinberger [4], [5], there exist two real numbers $B_1<-c_-^*<c_+^*<B_2$ given by

$$c_\pm^*=\inf_{\mu>0}\log\left\{g'(0)\int_{\mathbb{R}} e^{\pm\mu x}K(x)\,dx\right\}. \tag{1.1}$$

* Received by the editors September 14, 1981.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455 and Department of Mathematical Sciences, San Diego State University, San Diego, California 92182.

For $c \geq c_+^*$, travelling wave solutions $w_c(x)$ of the operator Q exist with $w_c(+\infty)=0$, $w_c(-\infty)=1$, $w_c(x)$ is nonincreasing and $w_c(x)$ is a fixed point of the operator $Q_c[u](x) = Q[u](x+c)$. There are also travelling wave solutions $\overline{w}_c(x)$ of Q for $c \leq -c_-^*$ that are facing left, i.e., $\overline{w}_c(x)$ is nondecreasing, $\overline{w}_c(-\infty)=0$, $\overline{w}_c(+\infty)=1$ and $\overline{w}_c(x)$ is again a fixed point of Q_c. All these results are summarized in [3, §1] and further properties of the travelling waves may be found in [3, §3].

Let $f \in \mathcal{C}$ and $f(+\infty)=0$. We define the function $f^{-1}(\gamma)=\sup\{x: f(x) \geq \gamma\}$ for $0<\gamma<1$. If $f(x)$ is everywhere less than γ, then $f^{-1}(\gamma)=-\infty$. Let $u_0 \in \mathcal{C}$ have compact support and assume that its integral over $\mathbb{R}$ is positive. Let $u_n(x)$ be the solution to the recursion $u_{n+1}(x)=Q[u_n](x)$ for $n \geq 0$. Define $m_n^+(\gamma)=\sup\{x: u_n(x) \geq \gamma\}$ and $m_n^-(\gamma)=\inf\{x: u_n(x) \geq \gamma\}$. Then the results in [4] and [5], which are stated as Theorems 1 and 2 in [3, §1], imply the following:

$$\frac{m_n^+(\gamma)}{n} \to c_+^* \quad \text{and} \quad \frac{m_n^-(\gamma)}{n} \to -c_-^* \quad \text{as } n \to +\infty, \tag{1.2}$$

for every $0<\gamma<1$. In particular, $m_n^{\pm}(\gamma)$ are defined in every compact subset of $(0,1)$ for sufficiently large n.

Statement (1.2) is about the long-time behavior of $u_n(x)$. It simply says that under the conditions stated above, $u_n(x)$ spreads out on the right like nc_+^* and on the left like $-nc_-^*$. (1.2) is also valid when $x \in \mathbb{R}^n$, $n \geq 2$. We refer the reader to [4] and [5] for the details. In [3], more precise results than (1.2) were obtained under the addition assumptions that $u_0(x)$ is monotone and decays rapidly enough at $+\infty$, $\liminf_{x \to -\infty} u_0(x)>0$, and $K(x)$ is of the form of the exponential of a concave function. In such a case, it is shown that $u_n(x+m_n^+(\gamma))$ converges to $w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))$ uniformly in $\mathbb{R}$ as $n \to +\infty$ for every $0<\gamma<1$. The condition $\liminf_{x \to -\infty} u_0(x)>0$ on $u_0(x)$, however, excludes the case when $u_0(x)$ has compact support. The main purpose of this paper is to prove results analogous to those obtained in [3] when the assumption $\liminf_{x \to -\infty} u_0(x)>0$ is replaced by u_0 having compact support.

Conditions (i)–(xi) will be assumed throughout the rest of this paper.

The organization of the paper is as follows. We need $K(x)$ to be a PF_3 function and we define and discuss this in §2. One of the main handicaps in dealing with integral operators (as opposed to differential equations) is the lack of the tool commonly called phase plane analysis. We therefore have to introduce u_n^δ as approximations to u_n in §3 so that the results of [3] may be applied to u_n^δ. Section 4 contains our basic estimate between u_n^δ and u_n under the assumption that $g(u)/u$ is nonincreasing. The main theorem and an application of it are contained in §5.

The notation $\beta=g'(0)$ will be employed, and hereafter, if the domain of integration is unspecified, it is assumed to be $\mathbb{R}$.

Remark 1. The condition that $g(u)/u$ be nonincreasing was used in [3, Thm. 4] only to prove the uniqueness of the minimum travelling wave $w_{c_+^*}(x)$.

Remark 2. In [3, §2], we defined the sign change of a function $f(t)$ in $\mathbb{R}$ and PF_r functions, $r \geq 1$. The following is true: Let $f(t)$ be bounded, piecewise continuous and have no more than r sign changes in $\mathbb{R}$. Let $K(x)$ be PF_{r+1}. Then the function $(K * f)(x)$ does not change sign more than r times in $\mathbb{R}$. $f \in PF_2$ means $\log f(x)$ is concave.

Remark 3. $g(u)$ does not increase the number of sign change for the difference of two functions. By this we mean that if $u(x)-v(x)$ has r sign changes in $\mathbb{R}$, then $g(u(x))-g(v(x))$ cannot have more than r sign changes in $\mathbb{R}$. This follows from the mean value theorem and the fact that $g'(u) \geq 0$.

2. PF_3. $K(x)$ is said to be a PF_3 function if $K(x)=e^{k(x)}$, $k(x)$ concave, and if for any $-\infty<x_1<x_2<x_3<+\infty$, $-\infty<y_1<y_2<y_3<+\infty$,

$$\det\begin{pmatrix} K(x_1-y_1) & K(x_1-y_2) & K(x_1-y_3) \\ K(x_2-y_1) & K(x_2-y_2) & K(x_2-y_3) \\ K(x_3-y_1) & K(x_3-y_2) & K(x_3-y_3) \end{pmatrix} \geq 0. \tag{2.1}$$

Unfortunately, the condition (2.1) is hard to check and unlike PF_2, there exists no simple representation formula for PF_3 functions. We first give some examples:

Example 1. $K(x)=(2\pi)^{-1/2}e^{-x^2/2}$. This is the normal denisty and it is PF_∞. This means $K*f$ has no more sign changes than $f(x)$, regardless of how many times $f(x)$ changes sign. There is a representation formula for PF_∞ functions in terms of their Laplace transforms, but no PF_∞ function has compact support (cf. [1]).

Example 2.

$$K(x)=\begin{cases} de^{-cx}, & x>0, \\ 0, & x\leq 0, \end{cases} \quad d,c\geq 0. \qquad \text{This is } PF_\infty.$$

Example 3.

$$K(x)=\begin{cases} \dfrac{(\cos x)^\alpha}{B\left(\frac{1}{2},\frac{\alpha+1}{2}\right)}, & -\dfrac{\pi}{2}\leq x\leq\dfrac{\pi}{2}, \\ 0 & \text{otherwise}, \end{cases} \qquad \alpha>-1.$$

This is $PF_r(r\geq 2)$ if $\alpha\geq r-2$. For a proof of this fact, see [1, p. 403]. In particular, $K(x)$ is PF_3 if $\alpha\geq 1$, and one sees that it is also in $C^1(\mathbb{R})$ for such values of α. $\int K(x)\,dx=1$ and $B(x,y)$ is the beta function.

LEMMA 1. *Let $f(x)$ be a bounded piecewise continuous function which has the property that, for two points $x_0<x_1$,*

$$f(x)\begin{cases} \leq 0, & x<x_0 \text{ or } x>x_1, \\ \geq 0, & x_0<x<x_1. \end{cases} \tag{2.2}$$

Let $K(x)$ be PF_3 and define $F(x)=(K*f)(x)$. Then $F(x)$ changes sign no more than twice in $\mathbb{R}$.

Proof. This is a special case of a general theorem in [1, p. 21].

LEMMA 2. *Let $K(x)$ be PF_3 and let $u_0\in\mathcal{C}$ satisfy the following conditions: For two points $x_0<x_1$,*

(i) $u_0(x_0)=u_0(x_1)\equiv\delta_0$,

(ii) $u_0(x)\geq\delta_0$ *in* (x_0,x_1),

(iii) $u_0(x)$ *is nonincreasing for* $x>x_1$ *and nondecreasing for* $x<x_0$,

(iv) $u_0(\pm\infty)=0$.

Let δ_n be defined by the recursion $\delta_{n+1}=g(\delta_n)$. Then there are two sequences of points $\{x_n\}$, $\{y_n\}$, $x_n\leq y_n$, $n\geq 0$, such that either $u_n(x_n)=u_n(y_n)=\delta_n$, $u_n(x)\geq\delta_n$ in $[x_n,y_n]$ and $u_n(x)$ monotone outside of $[x_n,y_n]$ for all n, or there is an N such that, for $n\geq N$, $u_n(x)\leq\delta_n$ in $\mathbb{R}$ and $u_n(x)$ is constant in $[x_n,y_n]$ and monotone outside of this interval.

Proof. The proof is by induction and our hypotheses say that the result is true when $n=0$. Assume that the result is true up to n. Consider the function $f(y)=g(u_n(y))-g(\delta_n)$. By our induction hypothesis, it has the property that $f(y)\geq 0$ in (x_n,y_n) and $f(y)\leq 0$ for $y<x_n$ or $y>y_n$. By Lemma 1, the function $(K*f)(x)=u_{n+1}(x)-\delta_{n+1}$ changes sign no more than twice. Suppose first that $u_{n+1}(x^*)>\delta_{n+1}$ for

some x^*. Set $x_{n+1}=\inf\{x: u_{n+1}(x)\geq\delta_{n+1}\}$ and $y_{n+1}=\sup\{x: u_{n+1}(x)\geq\delta_{n+1}\}$. Then x^* is in (x_{n+1},y_{n+1}) and all it remains to show is that $u_{n+1}(x)$ is monotone outside of (x_{n+1},y_{n+1}). We shall prove that $u_{n+1}(x)$ is nonincreasing in $(y_{n+1},+\infty)$, as the other case is similar. Let $\bar{x}>y_{n+1}$ and $\bar{\gamma}=u_{n+1}(\bar{x})$. Then $\bar{\gamma}<\delta_{n+1}$ and

$$u_{n+1}(x+\bar{x})<\delta_{n+1} \quad \text{for } x\geq 0. \tag{2.3}$$

Choose γ such that $g(\gamma)=\bar{\gamma}<\delta_{n+1}=g(\delta_n)$, hence $\gamma<\delta_n$. Set $\tilde{x}=\inf\{x: u_n(x)\geq\gamma\}$ and $\tilde{y}=\sup\{x: u_n(x)\geq\gamma\}$. The function $f(y)=g(u_n(y))-g(\gamma)$ has the property that $f(y)>0$ in $(\tilde{x},\tilde{y})$ and $f(y)<0$ if $y<\tilde{x}$ or $y>\tilde{y}$. Lemma 1 implies that $h(x)\equiv(K*f)(x)=u_{n+1}(x)-\bar{\gamma}$ cannot change sign more than twice, and $h(\bar{x})=0$. Now $u_{n+1}(x^*)>\delta_{n+1}$, so that $h(x^*)>\delta_{n+1}-\bar{\gamma}>0$. Also, (2.3) implies that $x^*<\bar{x}$. If $h(\bar{\bar{x}})>0$ for some $\bar{\bar{x}}>\bar{x}$, then by continuity there must exist a λ, $0<\lambda<\delta_{n+1}-\bar{\gamma}$, such that $h(x)-\lambda$ is negative at $\pm\infty$ and at $\bar{x}$, while it is positive at x^* and $\bar{\bar{x}}$. Thus $h(x)-\lambda$ changes sign at least four times. Choose $\bar{\bar{\gamma}}$ such that $g(\bar{\bar{\gamma}})=\bar{\gamma}+\lambda$ and x' such that $y_{n+1}<x'<\bar{x}$, $u_{n+1}(x')=\bar{\gamma}+\lambda$. Repeat the argument before with $\bar{x}$ replaced by x' and $\bar{\gamma}$ replaced by $\bar{\gamma}+\lambda$; we arrive at the conclusion that $u_{n+1}(x)-\bar{\gamma}-\lambda=h(x)-\lambda$ cannot have more than two sign changes, which is a contradiction. Thus $h(x)\leq 0$ if $x>\bar{x}$, which means $u_{n+1}(x)\leq u_{n+1}(\bar{x})$ if $x\geq\bar{x}$, and $u_{n+1}(x)$ is nonincreasing for $x>y_{n+1}$.

If $u_{n+1}(x)\leq\delta_{n+1}$ in all of $\mathbb{R}$, let N be the smallest integer such that $u_N(x)\leq\delta_N$ in $\mathbb{R}$. By our previous argument, $u_{N-1}(x)-\delta_{N-1}$ has two sign changes and is positive for some x. Choose $0<\mu<\delta_N$ so that $u_N(x)-\mu$ has at least two sign changes. Let $0<\bar{\delta}_0<\delta_0$ be such that $g^N(\bar{\delta}_0)=\mu$. Repeat the argument with δ_0 replaced by $\bar{\delta}_0$, x_0 replaced by $\inf\{x: u_0(x)\geq\bar{\delta}_0\}$ and x_1 replaced by $\sup\{x: u_0(x)\geq\bar{\delta}_0\}$. We arrive at the conclusion that $u_N(x)$ is monotone outside of the connected interval where $u_N(x)\geq\mu$. Since $0<\mu<\delta_N$ is arbitrary, this means there exists $x_N\leq y_N$ such that $u_N(x)$ is constant in $[x_N,y_N]$ and is monotone outside of this interval. For $n>N$, the same situation continues to be true, but we will omit the obvious proof. Q.E.D.

Remark 1. We shall hereafter refer to conditions (i)–(iv) of Lemma 2 as the two-sided condition.

Remark 2. We close this section by briefly discussing the regularity properties of a PF_3 function having compact support. After this, it will be clear that all the conditions stated in §1 for $K(x)$ are satisfied, save the condition that $\int K(x)\,dx=1$. In particular, $p_1=p_2=0$, so that $K(x)$ is absolutely continuous in $\mathbb{R}$. Let $K(x)$ be PF_3 such that $K(x)>0$ in (B_1,B_2) and vanishes outside of it. Then the results of Karlin in [1, Chapt. 4, §4] imply that $K(x)\in C^2(B_1,B_2)$, $K'(x)$ and $K''(x)$ have finite limits as $x\downarrow B_1$ and as $x\uparrow B_2$. Also $\lim_{x\downarrow B_1}K(x)/(x-B_1)^\alpha=0$ and $\lim_{x\uparrow B_2}K(x)/(x-B_2)^\alpha=0$ for any $0<\alpha<1$. Hence $K(x)\in C^\alpha(\mathbb{R})$ for any $0<\alpha<1$. But in general $K(x)$ is not in $C^1(\mathbb{R})$, as the following example shows. The function

$$\Lambda_\gamma(x)=\begin{cases}(\sin x)^\gamma, & 0\leq x\leq\pi,\\ 0 & \text{otherwise,}\end{cases}$$

is PF_{r+2} if $\gamma\geq r$. See [1, Chapt. 8, §4] for a proof of this fact. Putting $\gamma=1$, we see that $\Lambda_1(x)$ is PF_3, but is not differentiable at the origin.

3. Approximations by u_n^δ. We begin by an almost trivial lemma, which says that if there exists a $\bar{\delta}>0$ such that $g(u)=1$ in $[1-\bar{\delta},1]$, then the behavior of $u_0(x)$ near $-\infty$ is immaterial and [3, Thm. 6] may be applied. Let $\omega_c(\gamma)=w_c'(w_c^{-1}(\gamma))$, $c\geq c_+^*$ and $\psi_n^+(\gamma)=u_n'(m_n^+(\gamma))$ for $0<\gamma<1$.

LEMMA 3. *Let $K(x)$ be PF_2. Let $u_0 \in \mathcal{C}$ be such that either* (a) $u_0(x)=0$ *for* $x>A$, $u_0(x)>0$ *and nonincreasing in* $[A-\varepsilon, A)$ *for some* $\varepsilon>0$, *or* (b) $u_0(x)>0$ *near* $+\infty$, $u_0'(x) \sim -\mu u_0(x)$ *as* $x \to +\infty$ *for some* $\mu > \mu^*_{c_+^*}$. *In addition, let* $g(u)=1$ *in* $[1-\bar{\delta}, 1]$ *for some* $0<\bar{\delta}<1$. *Then for every* $0<\gamma<1$ *and* $-c_-^*<c_0^*<c_+^*$, *we have*

$$\lim_{n\to+\infty} u_n(x+m_n^+(\gamma)) = w_{c_+^*}\left(x+w_{c_+^*}^{-1}(\gamma)\right) \quad \text{uniformly in } \left[nc_0^*-m_n^+(\gamma), +\infty\right), \tag{3.1}$$

$$\lim_{n\to+\infty} m_{n+1}^+(\gamma)-m_n^+(\gamma)=c_+^* \tag{3.2}$$

and

$$\lim_{n\to+\infty} \psi_n^+(\gamma)=\omega_{c_+^*}(\gamma) \quad \text{uniformly in compact subsets of } (0,1). \tag{3.3}$$

Proof. Let $D>B_2-c_0^*>0$. Then (1.2) implies that there exists an N such that $u_n(x) \geq 1-\bar{\delta}$ in $(nc_0^*-D-B_2,\ nc_0^*+D-B_1)$ for $n \geq N-1$. Hence $u_n(x)=1$ in (nc_0^*-D, nc_0^*+D) for $n \geq N$. Set

$$v_0(x)=\begin{cases} u_N(x), & x \geq Nc_0^*, \\ 1, & x \leq Nc_0^*, \end{cases}$$

and define recursively $v_{j+1}=Q[v_j]$, $j \geq 0$. It is clear from an induction argument that $v_j(x) \geq u_{N+j}(x)$ for $j \geq 0$. Since $v_j \in [0,1]$, we have $v_j(x)=u_{N+j}(x)=1$ in $((N+j)c_0^*-D, (N+j)c_0^*+D)$ for $j \geq 0$. We claim that

$$v_j(x)=u_{N+j}(x) \quad \text{in } \left((N+j)c_0^*, +\infty\right) \quad \text{for } j \geq 0. \tag{3.4}$$

This is clear when $j=0$. Assume that it is true up to $j=J$. Then $v_J(x)=u_{N+J}(x)$ in $((N+J)c_0^*, +\infty)$.Hence from the definitions of v_{J+1} and u_{N+J+1}, we have $v_{J+1}(x)=u_{N+J+1}(x)$ in $[(N+J)c_0^*+B_2, +\infty)$. But the interval $((N+J+1)c_0^*, (N+J)c_0^*+B_2)$ is a subset of the interval $((N+J+1)c_0^*-D, (N+J+1)c_0^*+D)$ from our choice of D. Since v_{J+1} and u_{N+J+1} are both one inside this larger interval, (3.4) is proved. Now $v_0(x)$ satisfies all the hypotheses for an initial datum of [3, Thm. 6] and hence we conclude that

$$\lim_{j\to\infty} v_j\left(x+v_j^{-1}(\gamma)\right)=w_{c_+^*}\left(x+w_{c_+^*}^{-1}(\gamma)\right)$$

uniformly in $\mathbb{R}$ and

$$\lim_{j\to\infty} v_{j+1}^{-1}(\gamma)-v_j^{-1}(\gamma)=c_+^* \quad \text{for each } 0<\gamma<1.$$

Equations (3.1) and (3.2) follow from this and from (3.4). The proof of $\lim_{n\to\infty} v_j'(v_j^{-1}(\gamma))=\omega_{c_+^*}(\gamma)$ uniformly in compact subsets of $(0,1)$ is contained in the proof of [3, Thm. 6]. Equation (3.3) follows from this observation and from (3.4). Q.E.D.

Lemma 3 allows us to restrict our attention to the case where $0<g(u)<1$ in $(0,1)$. Assuming this, we shall now construct a family of functions $\{g^\delta: 0<\delta<\delta_0\}$ which approximate g as $\delta \downarrow 0$. For each δ, $g^\delta(u)$ equals 1 in a left neighborhood of $u=1$. More precisely, $g^\delta(u)$ has the following properties:

(a) $g^\delta(u) \in C^1[0,1]$,
(b) $g(u) \leq g^\delta(u) \leq 1$ in $[0,1]$,
(c) $g^\delta(u)=g(u)$ in $[0, 1-\delta]$,
(d) $g^\delta(u)=1$ in $[1-\bar{\delta}, 1]$ for some $0<\bar{\delta}<\delta$,
(e) $0 \leq [g^\delta]'(u) \leq \beta$ in $[0,1]$,

(f) $[g^\delta]'(u)>0$ whenever $g^\delta(u)<1$,
(g) $g^\delta(u)/u$ is nonincreasing in $[0,1]$.
It is clear that $g^\delta(u)$ has all the properties of $g(u)$.
Let $\delta_0>0$ be such that

$$\beta(1-\delta_0)>1. \tag{3.5}$$

Suppose that $g(u)$ has the property that there exists a sequence $\{u_j\}$, $u_j\uparrow 1$, such that for each j,

$$g'(u_j)\geq\frac{g(u)-g(u_j)}{u-u_j} \quad \text{in } [u_j,1]. \tag{3.6}$$

Then given $\delta<\delta_0$, choose the first j such that $u_j\geq 1-\delta$, and define $g^\delta(u)=g(u)$ in $[0,u_j]$, $g^\delta(u)=g'(u_j)(u-u_j)+g(u_j)$ for $u\geq u_j$ and as long as $g^\delta(u)\leq 1$. From (3.6), $g(u)\leq g^\delta(u)$ and condition (g) follows from the assumption that $g(u)/u$ is nonincreasing. Let Q be the point where $g^\delta(u)$ intersects $y=1$. If $Q\neq(1,1)$, then we may place a small circle θ above $g(u)$ which touches $g'(u_j)(u-u_j)+g(u_j)$ at Q_1 and $y=1$ at $Q_2=(1-\bar{\delta},1)$. Define $g^\delta(u)$ to be the arc of θ between Q_1, Q_2 and $g^\delta(u)=1$ in $[1-\bar{\delta},1]$. It can be verified that conditions (a)–(g) are all satisfied. If $Q=(1,1)$, then $[g^\delta]'(u)<1\leq g(u)/u$ near $u=1$. We modify $g^\delta(u)$ according to a procedure described below. Suppose that (3.6) is not valid, so that for every $u_0\in[1-\delta_0,1)$, there exists $u_1\geq\bar{u}>u_0$ such that

$$g'(u_0)<\frac{g(u_1)-g(u_0)}{u_1-u_0}=g'(\bar{u}).$$

Then $g'(u)<g'(1)\leq 1$ in $[1-\delta_0,1)$. Let $g^\delta(u)=g(u)+C_h(u-1+\delta)^2$ in $[1-\delta,1-\delta+h]$, and let it continue at $u=1-\delta+h$ as a straight line of slope 1 until it intersects $y=1$ at Q. Here $C_h=(1-g'(1-\delta+h))/2h\geq 0$ and $h>0$. Near Q, we construct g^δ as before, and it is clear that conditions (a)–(f) are all satisfied. Condition (g) is equivalent to $[g^\delta]\leq[g^\delta](u)/u$. Since $u_1>u_0$, we have $(g(u_1)-g(u_0))/(u_1-u_0)\leq g(u_0)/u_0$, so that $g'(u_0)<g(u_0)/u_0$. By continuity, (g) is true in $[1-\delta,1-\delta+h]$ for sufficiently small h. On the part where $g^\delta(u)$ is a straight line, the y-intercept is positive, since $g(u)>u$ in $(0,1)$, which shows that (g) is true there also.

Having constructed $g^\delta(u)$, we define the associated operator Q^δ by

$$Q^\delta[u](x)=\int K(x-y)g^\delta(u(y))\,dy. \tag{3.7}$$

Given $u_0\in\mathcal{C}$, we define $u_0^\delta\equiv u_0$ and $u_{n+1}^\delta=Q^\delta[u_n^\delta]$ for $n\geq 0$. Also, we denote by $w_c^\delta(x)$ the travelling wave of speed c associated with the operator Q^δ and set $\omega_c^\delta(\gamma)=[w_c^\delta]'(w_c^{\delta\,-1}(\gamma))$ for $0<\gamma<1$. From (1.1) and property (c) of g^δ, the range $[c_+^*,+\infty)$, over which travelling waves of the operator Q^δ exist, is independent of δ and is the same as that of the operator Q. However, $w_c^\delta(x)$ themselves will be identically one near $-\infty$.

Let us introduce the following notations for later use. $m_n^{\delta,+}(\gamma)=\sup\{x: u_n^\delta(x)\geq\gamma\}$ and $\psi_n^{\delta,+}(\gamma)=[u_n^\delta]'(m_n^{\delta,+}(\gamma))$. Statement (1.2) now holds with $m_n^\pm(\gamma)$ replaced by $m_n^{\delta,\pm}(\gamma)$. Also, for each $0<\delta<\delta_0$, $g^\delta(u)=1$ in $[1-\bar{\delta},1]$; Lemma 3 therefore implies the following results. For each $0<\gamma<1$, $-c_-^*<c_0^*<c_+^*$, $0<\delta<\delta_0$,

$$\lim_{n\to+\infty} u_n^\delta\left(x+m_n^{\delta,+}(\gamma)\right)=w_{c_+^*}^\delta\left(x+w_{c_+^*}^{\delta\,-1}(\gamma)\right) \tag{3.8}$$

uniformly in $[nc_0^*-m_n^{\delta,+}(\gamma),+\infty)$,

$$\lim_{n\to+\infty} m_{n+1}^{\delta,+}(\gamma)-m_n^{\delta,+}(\gamma)=c_+^*, \tag{3.9}$$

and

$$\lim_{n\to+\infty} \psi_n^{\delta,+}(\gamma)=\omega_{c_+^*}^{\delta}(\gamma) \quad \text{uniformly in compact subsets of } (0,1). \tag{3.10}$$

The next lemma compares $\psi_n^{\delta,+}(\gamma)$ with $\omega_c(\gamma)$ for a fixed $0<\gamma<1$ and $c>c_+^*$.

LEMMA 4. *Let $K(x)$ be PF_3 and let $0<g(u)<1$ in $(0,1)$. Let $u_0\in\mathcal{C}$ satisfy the following conditions near $+\infty$. Either there exists a constant A such that $u_0(x)=0$ for $x\geq A$, $u_0(x)>0$ and nonincreasing in $[A-\varepsilon,A)$ for some $\varepsilon>0$; or $u_0(x)>0$ near $+\infty$, $u_0'(x)\sim-\mu u_0(x)$ as $x\to+\infty$ for some $\mu>\mu_{c_+^*}$. Let $u_0(x)$ also satisfy the following conditions near $-\infty$. Either there exists a constant L such that $u_0(x)=0$ for $x\leq L$, $u_0(x)>0$ and nonincreasing in $(L,L+\varepsilon']$ for some $\varepsilon'>0$, or $u_0(x)>0$ and nondecreasing near $-\infty$ and $\lim_{x\to-\infty} u_0(x)=0$. Then given $c>c_+^*$, $0<\delta<\delta_0$ and $0<\gamma<1-\delta$, there exists an integer $N_{c,\delta,\gamma}$ such that*

$$\psi_n^{\delta,+}(\gamma)\leq\omega_c(\gamma) \quad \text{for } n\geq N_{c,\delta,\gamma}. \tag{3.11}$$

Proof. From our hypotheses, it is not difficult to see, as in the beginning of the proof of [3, Thm. 6], that $u_m(x)$ will satisfy the two-sided condition of Lemma 2 for sufficiently large m. Thus without loss of generality, we assume that $u_0(x)$ itself satisfies such a condition. As a consequence of Lemma 3, there exists $N_1=N_1(\delta)$ such that for each $j\geq N_1$, there are numbers $a_j<b_j$ such that $u_j^{\delta}(x)=1$ in $[a_j,b_j]$ and $u_j^{\delta}(x)$ is monotone outside of $[a_j,b_j]$. Also from (3.9) and the fact that $c>c_+^*$, there exists $N_2=N_2(\gamma,c,\delta)$ such that

$$m_{j+1}^{\delta,+}(\gamma)-m_j^{\delta,+}(\gamma)<c \quad \text{for } j\geq N_2. \tag{3.12}$$

Let $N=\max(N_1,N_2)$. We claim that there exists $N_3=N_3(\gamma,\delta,c)$ such that, for $n\geq N_3$, the function

$$u_N^{\delta}\big(x+m_{N+n}^{\delta,+}(\gamma)-nc\big)-w_c\big(x+w_c^{-1}(\gamma)\big) \tag{3.13}$$

has two sign changes. From [3, Lemma 6] and the fact that $0<g(u)<1$, $w_c'(x)<0$ in $\mathbb{R}$. Since $u_N^{\delta}(-\infty)=0$ and $u_N^{\delta}(x)=1$ in an interval, there is one sign change in the region on the left where $u_N^{\delta}(x+m_{N+n}^{\delta,+}(\gamma)-nc)$ is nondecreasing. Since $0<w_c(x)<1$ in $\mathbb{R}$, in the region on the right where $u_N^{\delta}(x+m_{N+n}^{\delta}(\gamma)-nc)$ is nonincreasing, (3.13) has one sign change. The proof of this fact may be found in the proof of [3, Thm. 6]. Note that $m_{N+n}^{\delta,+}(\gamma)\sim(N+n)c_+^*$ as $n\to+\infty$ and $c>c_+^*$ imply that $m_{N+n}^{\delta,+}(\gamma)-nc$ diverges to $-\infty$ as $n\to+\infty$. The facts (5.10) and (5.11) in [3, Thm. 6] carry over to our present situation, because their proofs used only the properties of travelling waves and [3, Lemma 11] which do not assume anything about the behavior of $u_N^{\delta}(x)$ near $-\infty$. By increasing N_3, we may assume that (3.13) vanishes in the region where $u_N^{\delta}(x+m_{N+n}^{\delta,+}(\gamma)-nc)<1-\delta$.

Let $N_{c,\delta,\gamma}\equiv N+N_3$ and $\bar{n}\geq N_{c,\delta,\gamma}$ be given. Write $\bar{n}=N+n$, $n\geq N_3$ and set

$$v_j(x)=u_{N+j}^{\delta}\big(x+m_{N+n}^{\delta,+}(\gamma)-(n-j)c\big), \qquad 0\leq j\leq n. \tag{3.14}$$

We shall prove that $v_j(x)-w_c(x+w_c^{-1}(\gamma))$ has exactly two sign changes for $0\leq j\leq n$.

First observe that for a function $f(t)$ which is positive or negative in a neighborhood of a point $\bar{x}$, the number of sign changes of $f(t)$ in $\mathbb{R}$ is the sum of the number of sign changes of $f(t)$ in the respective interval $(-\infty,\bar{x})$ and $(\bar{x},+\infty)$. We are going to prove the following claim with the help of this observation.

CLAIM. *For $0\leq j\leq n$, $v_j(x)-w_c(x+w_c^{-1}(\gamma))$ has two sign changes and the larger one occurs in the interval $[v_j^{-1}(1-\delta),+\infty)$.*

Proof of claim. The proof is by induction, and our earlier remark about (3.13) says that the claim is true when $j=0$. Assume that it is true for j. We first consider the number of sign changes of the function

$$g^{\delta}\big(v_j(x)\big)-g\big(w_c\big(x+w_c^{-1}(\gamma)\big)\big). \tag{3.15}$$

Let $a_j'=a_{j+N}-m_{N+n}^{\delta,+}(\gamma)+(n-j)c$ and $b_j'=b_{j+N}-m_{N+n}^{\delta,+}(\gamma)+(n-j)c$. In $(-\infty,a_j')$, (3.15) has exactly one sign change. In $[a_j',b_j']$, $v_j(x)=1$, hence our induction hypothesis implies that $v_j(x)\geq w_c(x+w_c^{-1}(\gamma))$ in $(a_j',v_j^{-1}(1-\delta))$. Since $g^{\delta}\geq g$, (3.15) is non-negative in $(a_j',v_j^{-1}(1-\delta))$ and has exactly one sign change in $(-\infty,v_j^{-1}(1-\delta))$. In $(v_j^{-1}(1-\delta),+\infty)$, $g^{\delta}=g$ in (3.15), and since $g(u)$ does not increase the number of sign changes for the difference of two functions, the same is true of (3.15) in $(v_j^{-1}(1-\delta),+\infty)$. To continue, (3.15) is negative near $+\infty$. If (3.15) is positive at $x=v_j^{-1}(1-\delta)$, then the observation before the claim implies that (3.15) has two sign changes in $\mathbb{R}$ or if (3.15) vanishes at $x=v_j^{-1}(1-\delta)$, then it must have a sign change in $[b_j',+\infty)$. At any rate, (3.15) has two sign changes, and by Lemma 1, so does the function $v_{j+1}(x)-w_c(x+w_c^{-1}(\gamma))$. To complete our induction, it suffices to produce an $\bar{\bar{x}}>0$ such that

$$v_{j+1}(\bar{\bar{x}})-w_c\big(\bar{\bar{x}}+w_c^{-1}(\gamma)\big)>0.$$

For this means there is one sign change in $[\bar{\bar{x}},+\infty)$. But if $x_0\in[\bar{\bar{x}},+\infty)$ and $v_j(x_0)=w_c(x_0+w_c^{-1}(\gamma))$, then $w_c(x_0+w_c^{-1}(\gamma))\leq\gamma<1-\delta$. Thus x_0 is actually in $(v_j^{-1}(1-\delta),+\infty)$. Now simply take $\bar{\bar{x}}=(n-j-1)c+m_{N+j+1}^{\delta,+}(\gamma)-m_{N+n}^{\delta,+}(\gamma)$. $\bar{\bar{x}}$ is positive because of (3.12). This completes our induction as well as the proof of our claim.

Returning to the proof of Lemma 4, using the claim when $j=n$, we get $v_n(x)-w_c(x+w_c^{-1}(\gamma))$ has two sign changes. But $u_{N+n}^{\delta}(x+m_{N+n}^{\delta,+}(\gamma))-w_c(x+w_c^{-1}(\gamma))$ has a zero at $x=0$, and since $u_{N+n}^{\delta}(x+m_{N+n}^{\delta,+}(\gamma))=1$ for some $x<0$, we have

$$\big[u_{N+n}^{\delta}\big]'\big(m_{N+n}^{\delta,+}(\gamma)\big)\leq w_c'\big(w_c^{-1}(\gamma)\big),$$

or what is the same, $\psi_{\bar{n}}^{\delta,+}(\gamma)\leq\omega_c(\gamma)$ if $\bar{n}\geq N_{c,\delta,\gamma}$. Q.E.D.

COROLLARY. *Given $c>c_+^*$, $\mathbb{K}=[\underline{\gamma},\bar{\gamma}]$ in $(0,1)$, there exists a $\delta^*>0$ such that for any $0<\delta<\delta^*$, $\varepsilon>0$, there exists an integer $N_{c,\mathbb{K},\delta,\varepsilon}$ such that*

$$\psi_n^{\delta,+}(\gamma)\leq\omega_c(\gamma)+\varepsilon \quad \textit{for } \gamma\in\mathbb{K}, \quad n\geq N_{c,\mathbb{K},\delta,\varepsilon}.$$

Proof of Corollary. Set $\delta^*=\min(\delta_0,1-\bar{\gamma})$. Then for any $\gamma\in\mathbb{K}$, $0<\delta<\delta^*$, we have $\gamma\leq\bar{\gamma}\leq 1-\delta^*<1-\delta$. Hence Lemma 4 implies that

$$\psi_n^{\delta,+}(\gamma)\leq\omega_c(\gamma) \quad \text{for sufficiently large } n.$$

Let $n\to+\infty$ and use (3.10). We have $\omega_{c_+^*}^{\delta}(\gamma)\leq\omega_c(\gamma)$ for $\gamma\in\mathbb{K}$, $0<\delta<\delta^*$. But (3.10) actually says that $\psi_n^{\delta,+}(\gamma)$ converges to $\omega_{c_+^*}^{\delta}(\gamma)$ uniformly in $\mathbb{K}$. Thus $\psi_n^{\delta,+}(\gamma)\leq\omega_{c_+^*}^{\delta}(\gamma)+\varepsilon$ for sufficiently large n, $\gamma\in\mathbb{K}$. Combining the two inequalities, we are done.

The rest of this section is devoted to the study of sufficient conditions on the behavior of $u_0(x)$ near $+\infty$ so that $u_n'(m_n^+(\gamma))$ is bounded away from 0 as $n\to+\infty$. What needs to be shown is that there exists a $t_0>0$ such that for $t\geq t_0$, $u_1(x+t)-u_0(x)$ changes sign only once. The difficulty lies in getting $u_1(x+t)$ below $u_0(x)$ near $+\infty$. From the fact that $g(u)\leq\beta u$ in $[0,1]$, we have

$$u_1(x+t)\leq\beta\int_{B_1}^{B_2}K(y)u_0(x+t-y)\,dy. \tag{3.16}$$

Let $\mu>0$ be given and suppose that $A(x)\equiv u_0(x)e^{\mu x}$ has the property that $A(x+x_0)/A(x)\to 1$ as $x\to+\infty$ for every x_0 (i.e., $A(\log x)$ is slowly varying). Let $\varepsilon>0$ be given and choose t so large that $e^{-\mu t}(1+\varepsilon)\beta e^{\mu B_2}\leq 1$. Let $x_0=t-B_2$ in above property of

$A(x)$. Then for sufficiently large x, $u_0(x+t-B_2)\le(1+\varepsilon)e^{-\mu(t-B_2)}u_0(x)$. Since $u_0(x)$ is nonincreasing near $+\infty$, (3.16) implies that for sufficiently large x, $u_1(x+t)\le(1+\varepsilon)\beta\int K(y)\,dy\,e^{-\mu(t-B_2)}u_0(x)\le u_0(x)$. Thus a sufficient condition is for $u_0(x)$ to have the form $A(x)e^{-\mu x}$ where $A(x)>0$ in $\mathbb{R}$ and $A(\log x)$ is slowly varying. It is easy to show that our condition on $u_0(x)$, that is, $u_0'(x)\sim-\mu u_0(x)$ as $x\to+\infty$, actually implies this. Finally, if for a given $t_0>0$, we can find L_0 such that $u_1(x+t_0)\le u_0(x)$ for $x\ge L_0$, choose t_0 so large that $u_1(x)$ is nonincreasing for $x\ge L_0+t_0$. Then we have $u_1(x+t)\le u_0(x)$ for $x\ge L_0$, $t\ge t_0$.

LEMMA 5. *Let $K(x)$, $g(u)$ and $u_0(x)$ satisfy the hypotheses stated for them in Lemma 4. Then given $\mathbb{K}=[\underline{\gamma},\bar{\gamma}]$ in $(0,1)$, there exists a positive integer $N_{\mathbb{K},u_0}$ and a $c_0=c_0(u_0)>c_+^*$ such that*

$$\psi_n^+(\gamma)\le\omega_{c_0}(\gamma)\quad \textit{for } \gamma\in\mathbb{K},\quad n\ge N_{\mathbb{K},u_0}.$$

Proof. The first step is to show that there exists $t_1\ge0$ such that $u_1(x+t)-u_0(x)$ has one sign change for $t>t_1$. From Lemma 2, we may assume that $u_n(x)$, $n\ge0$, have the property that for some $\bar{\bar{\gamma}}$, $\bar{\gamma}<\bar{\bar{\gamma}}<1$, $u_n(x)\ge\bar{\bar{\gamma}}$, in the interval $(m_n^-(\bar{\bar{\gamma}}),\ m_n^+(\bar{\bar{\gamma}}))$ and are monotone outside of it. We may also assume that $m_0^-(\bar{\bar{\gamma}})-m_0^-(\eta)>B_2-B_1$ for sufficiently small $\eta>0$. Let $L_0,t_0>\max(m_0^+(\bar{\bar{\gamma}}),B_2)$ be such that

$$u_1(x+t)\le u_0(x)\quad \text{for } x\ge L_0,\quad t\ge t_0. \tag{3.17}$$

Let $L_0=m_0^+(\eta)$ for sufficiently small η. This is possible if $L_0<m_0^+(0)$ (it is obvious what $m_0^+(0)$ means). If $m_0^+(0)<+\infty$, then we take $L_0=m_1^+(0)-t_0$ in (3.17). Since $m_0^+(0)=m_1^+(0)-B_2$, L_0 has the desired form. We further decrease η if necessary, so that

$$m_0^-(\bar{\bar{\gamma}})-m_1^-(\eta)>-B_1. \tag{3.18}$$

(3.18) is clearly possible if $m_0^-(0)=-\infty$. If $m_0^-(0)>-\infty$, then since $m_1^-(\eta)-m_0^-(\eta)\to B_1<B_2$ as $\eta\downarrow0$, we have for sufficiently small η, $m_0^-(\bar{\bar{\gamma}})-m_1^-(\eta)=m_0^-(\bar{\bar{\gamma}})-m_0^-(\eta)-(m_1^-(\eta)-m_0^-(\eta))>m_0^-(\bar{\bar{\gamma}})-m_0^-(\eta)-B_2>-B_1$. Having fixed the $\eta>0$, we choose $\bar{\eta}\in(0,\eta)$ and set

$$t_1=\max\bigl(m_1^+(\eta)-m_0^-(\bar{\eta}),m_1^+(\bar{\eta})-m_0^-(\eta),t_0\bigr). \tag{3.19}$$

We divide $\mathbb{R}$ into three intervals. First

$$u_1(x+t)\le u_0(x)\quad \text{in } \bigl(m_0^-(\eta),+\infty\bigr). \tag{3.20}$$

This follows from (3.17) and the fact that in $(m_0^-(\eta),m_0^+(\eta))$ we have $x+t\ge m_0^-(\eta)+t\ge m_1^+(\eta)$, so that $u_1(x+t)\le\eta\le u_0(x)$. Next,

$$u_1(x+t)\ge u_0(x)\quad \text{in } \bigl(-\infty,m_1^+(\eta)-t\bigr). \tag{3.21}$$

If $x<m_1^-(\eta)-t$, then (3.18) implies that $x+t-B_1<m_0^-(\bar{\bar{\gamma}})$, and since $u_0(x)$ is nondecreasing in $(-\infty,m_0^-(\bar{\bar{\gamma}}))$, we have

$$u_1(x+t)=\int_{B_1}^{B_2}K(y)g\bigl(u_0(x+t-y)\bigr)\,dy\ge g\bigl(u_0(x+t-B_2)\bigr)\ge u_0(x+t-B_2)\ge u_0(x).$$

In $(m_1^-(\eta)-t,\ m_1^+(\eta)-t)$, $u_1(x+t)\ge\eta\ge u_0(x)$ since $m_0^-(\eta)\ge m_1^+(\eta)-t$. Hence (3.21) is valid. Finally, in $(m_1^+(\eta)-t,m_0^-(\eta))$, $u_1(x+t)$ is nonincreasing and $u_0(x)$ is nondecreasing. $u_1(x+t)>u_0(x)$ at $x=m_1^+(\eta)-t\le m_0^-(\bar{\eta})$ and $u_1(x+t)<u_0(x)$ at $x=m_0^-(\eta)\ge m_1^+(\bar{\eta})-t$. Thus there is one sign change of $u_1(x+t)-u_0(x)$ in this interval. (3.20) and (3.21) imply that this is the only sign change in $\mathbb{R}$.

Lemma 5 now follows from the following three assertions:

(a) There exists a subsequence $\{n_j\}$ of positive integers such that

$$m^+_{n_j+1}(\gamma)-m^+_{n_j}(\gamma)\le t_1 \quad \text{for all } \gamma\in\mathbb{K},\quad j\ge 1.$$

(b) There exists $N_{\mathbb{K}}$ such that if $n>N_{\mathbb{K}}$ and $m^+_n(\gamma)-m^+_{n-1}(\gamma)\le t_1$, we have

$$\psi^+_n(\gamma)\le\omega_{c_0}(\gamma) \quad \text{for } \gamma\in\mathbb{K},$$

where c_0 depends only on u_0 and not on $\mathbb{K}$.

(c) If n is such that $m^+_n(\gamma)-m^+_{n-1}(\gamma)>t_1$, then $\psi^+_n(\gamma)\le\psi^+_{n-1}(\gamma)$ for $0<\gamma<1$.

Part (c) is trivial, since from what we have just done, the function $u_1(x+m^+_n(\gamma)-m^+_{n-1}(\gamma))-u_0(x)$ changes sign once from positive to negative as x increases from $-\infty$ to $+\infty$. $K(x)$ is PF_2 and $g(u)$ does not increase the number of sign changes for the difference of two functions, hence an inductive argument implies that $u_n(x+m^+_n(\gamma)-m^+_{n-1}(\gamma))-u_{n-1}(x)$ changes sign no more than once and has a nonpositive derivative at its zero. At $x=m^+_{n-1}(\gamma)$, its derivative is nonpositive, which proves (c).

We next show assertion (b); we begin by determining c_0. Let $x_0=(B_1-t_1)$ and $c_0=B_2-x_0$. We have, for $c\ge c_0$,

$$w_c\big(x_0+w_c^{-1}(\gamma)\big)=\int_0^\infty K(x_0+c-y)g\big(w_c\big(y+w_c^{-1}(\gamma)\big)\big)\,dy<g(\gamma).$$

In particular, $w_{c_0}(x_0+w_{c_0}^{-1}(\gamma))<g(\gamma)$ for all $0<\gamma<1$. Under our hypotheses on $u_0(x)$, there exists N_1 depending on w_{c_0}, u_0 and $\mathbb{K}$ such that $u_0(x+m^+_n(\gamma)-nc_0)-w_{c_0}(x+w_{c_0}^{-1}(\gamma))$ has two sign changes for $\gamma\in\mathbb{K}$, $n\ge N_1$ (see Lemma 4). From Lemma 3 and an inductive argument, the function

$$u_n\big(x+m^+_n(\gamma)\big)-w_{c_0}\big(x+w_{c_0}^{-1}(\gamma)\big) \tag{3.22}$$

has no more than two sign changes. Let $N_2=N_2(\mathbb{K})$ be such that

$$-B_2+x_0+m^+_n(\bar\gamma)-m^-_{n-1}(\bar\gamma)\ge 0 \quad \text{if } n\ge N_2. \tag{3.23}$$

This is possible because of (1.2). Set $N_{\mathbb{K}}=\max(N_1,N_2)$ and let $n\ge N_{\mathbb{K}}$. Then $u_n(x+m^+_n(\gamma))=\int K(x+m^+_n(\gamma)-m^+_{n-1}(\gamma)-y)g(u_{n-1}(y+m^+_{n-1}(\gamma)))\,dy$. The domain of integration is

$$-B_2+x+m^+_n(\gamma)-m^+_{n-1}(\gamma)\le y\le x+m^+_n(\gamma)-m^+_{n-1}(\gamma)-B_1. \tag{3.24}$$

At $x=x_0=B_1-t_1$, the right-hand side of (3.24) equals $m^+_n(\gamma)-m^+_{n-1}(\gamma)-t_1\le 0$, while the left-hand side of (3.24) equals, by (3.23),

$$\begin{aligned}-B_2+x_0+m^+_n(\gamma)-m^+_{n-1}(\gamma)&\ge -B_2+x_0+m^+_n(\bar\gamma)-m^-_{n-1}(\bar\gamma)+m^-_{n-1}(\bar\gamma)-m^+_{n-1}(\gamma)\\&\ge m^-_{n-1}(\bar\gamma)-m^+_{n-1}(\gamma)\ge m^-_{n-1}(\gamma)-m^+_{n-1}(\gamma)\end{aligned}$$

if $\gamma\in\mathbb{K}$, $n\ge N_{\mathbb{K}}$. Hence the domain of integration is a subset of the interval $(m^-_{n-1}(\gamma)-m^+_{n-1}(\gamma),0)$ for all $\gamma\in\mathbb{K}$, $n\ge N_{\mathbb{K}}$. In this interval, $u_{n-1}(y+m^+_{n-1}(\gamma))\ge\gamma$. Hence (3.22) at $x=x_0$ is positive, which implies that it has two sign changes. Since $t_1\ge B_2$, x_0 is negative, and since (3.22) vanishes at $x=0$, we must have

$$u_n'\big(x+m^+_n(\gamma)\big)-w_{c_0}'\big(x+w_{c_0}^{-1}(\gamma)\big)\le 0 \quad \text{at } x=0.$$

This proves (b).

Finally, assertion (a) is clear for a fixed γ since $t_1\ge B_2>c^*_+$ and $m^+_n(\gamma)\sim nc^*_+$. To do this for γ in a compact subset of $(0,1)$, we observe that $u_1(x+t_1)-u_0(x)$ has one sign change. Since $K(x+t_1)$ is PF_2 and $g(u)$ does not increase the number of sign

changes in the difference of two functions, an inductive argument implies that $u_n(x+nt_1)-u_{n-1}(x+(n-1)t_1)$ has no more than one sign change. Let x_{n-1} be a point where the above function vanishes and set $\gamma_{n-1}=u_{n-1}(x_{n-1})$. Then

$$m_n^+(\gamma)-m_{n-1}^+(\gamma)\begin{cases}\geq t_1 & \text{if } \gamma>\gamma_{n-1},\\ \leq t_1 & \text{if } \gamma<\gamma_{n-1}.\end{cases} \tag{3.25}$$

If $\gamma_j\leq\bar{\bar{\gamma}}$ for all j, then (3.25) implies that $m_n^+(\bar{\bar{\gamma}})-m_{n-1}^+(\bar{\bar{\gamma}})\geq t_1>c_+^*$ for all n sufficiently large, which is impossible. Hence there exists a subsequence $\{n_j\}$ such that $\gamma_{n_j}>\bar{\gamma}, j\geq 0$, and from (3.25), (a) is proved. Q.E.D.

4. Uniform estimates. In this section we are going to prove our basic estimates that u_n^δ is close to u_n when δ is small independently of δ. From this we will be able to conclude that $\psi_n^{\delta,+}$ is also close to ψ_n^+ when δ is small independently of n.

We recall the following condition on $g(u)$.

$$\frac{g(u)}{u} \text{ is nonincreasing in } (0,1). \tag{4.1}$$

This condition implies that $g'(u)\leq\beta$ in $[0,1]$, which in turn implies the condition $g(u)\leq\beta u$ in $[0,1]$. The second condition is an immediate consequence of (4.1) and the fact that $\lim_{u\downarrow 0} g(u)/u=\beta$. If we now differentiate the function $g(u)/u$ and use (4.1), we get $g'(u)\leq\beta$ in $[0,1]$.

Condition (4.1) also implies the existence of a function $\theta(\delta)\in(0,1)$, defined in $(0,\delta_0)$, with the properties

$$\lim_{\delta\downarrow 0}\theta(\delta)=1 \tag{4.2}$$

and

$$\frac{g(u)}{u}\leq\frac{1}{1-\delta} \quad \text{implies} \quad u\geq\theta(\delta). \tag{4.3}$$

In fact, we may take

$$\theta(\delta)=\inf\left\{0\leq u\leq 1:\frac{g(u)}{u}\leq\frac{1}{1-\delta}\right\}.$$

Since $1/(1-\delta_0)<\beta$ by (3.5), $\delta\in[0,\delta_0)$ implies that $1/(1-\delta)<\beta$, and hence $\theta(\delta)>0$. Also $g(1)=1$, hence $\theta(\delta)<1$ if $\delta\in(0,\delta_0)$. To prove (4.2), we suppose that there exists a sequence $\delta_j\downarrow 0$ while $\theta_j=\theta(\delta_j)<\theta_0<1$. From (4.1) and the fact that $g(u)>u$ in $(0,1)$, we get

$$1<\frac{g(\theta_0)}{\theta_0}\leq\frac{g(\theta_j)}{\theta_j}\leq\frac{1}{1-\delta_j} \quad \text{for } j\geq 1.$$

Letting j go to $+\infty$, we obtain a contradiction. Hence if we define $\theta(0)=1$, $\theta(\delta)$ is right continuous at the origin. (4.3) follows from the definition of $\theta(\delta)$.

We recall that $u_0^\delta\equiv u_0$ and $g^\delta\geq g$. Hence we have $u_n^\delta\geq u_n$ for $n\geq 0$ and $\operatorname{supp} u_n^\delta=\operatorname{supp} u_n$ for $n\geq 0$. If $u_n(x)=0$, we define $u_n^\delta(x)/u_n(x)=1$. Our basic estimate is a corollary of

LEMMA 6. *Let $u_0\in\mathcal{C}$ and let $g(u)$ satisfy the conditions $0<g(u)<1$ in $(0,1)$ and (4.1). Then, for $0<\delta<\delta_0$, we have*

$$\left\|\frac{u_{n+1}^\delta}{u_{n+1}}-1\right\|_\infty\leq\max\left(\left\|\frac{u_n^\delta}{u_n}-1\right\|_\infty,\frac{1}{g(\theta(\delta))}-1\right) \tag{4.4}$$

for all $n\geq 0$.

Proof. Write

$$u_{n+1}^{\delta}(x)-u_{n+1}(x)=\int K(x-y)g(u_n(y))\left[\frac{g^{\delta}(u_n^{\delta}(y))}{g(u_n(y))}-1\right]dy. \tag{4.5}$$

Let $\mathbb{R}=I_1(\delta)\cup I_2(\delta)$, where

$$I_1(\delta)=\{y: u_n^{\delta}(y)\le 1-\delta\},\qquad I_2(\delta)=\{y: u_n^{\delta}(y)>1-\delta\}.$$

Then in

$$I_1(\delta)\colon \frac{g^{\delta}(u_n^{\delta}(y))}{g(u_n(y))}-1=\frac{g(u_n^{\delta}(y))}{g(u_n(y))}-1\le\frac{u_n^{\delta}(y)}{u_n(y)}-1$$

by (4.1). For

$$I_2(\delta)\colon \text{If } \frac{g^{\delta}(u_n^{\delta}(y))}{g(u_n(y))}\ge\frac{u_n^{\delta}(y)}{u_n(y)},\quad \text{we have } \frac{g(u_n(y))}{u_n(y)}\le\frac{g^{\delta}(u_n^{\delta}(y))}{u_n^{\delta}(y)}\le\frac{1}{1-\delta}.$$

From (4.3), we conclude that $u_n(y)\ge\theta(\delta)$. Therefore $g(u_n(y))\ge g(\theta(\delta))$, and hence $g^{\delta}(u_n^{\delta}(y))/g(u_n(y))-1\le 1/g(\theta(\delta))-1$. So in $I_2(\delta)$, we have

$$\frac{g^{\delta}(u_n^{\delta}(y))}{g(u_n(y))}-1\le\max\left(\frac{u_n^{\delta}(y)}{u_n(y)}-1,\ \frac{1}{g(\theta(\delta))}-1\right).$$

Going back to (4.5), we obtain

$$u_{n+1}^{\delta}(x)-u_{n+1}(x)\le\max\left(\left\|\frac{u_n^{\delta}}{u_n}-1\right\|_{\infty},\ \frac{1}{g(\theta(\delta))}-1\right)Q[u_n](x).$$

Divide by $u_{n+1}(x)$, and (4.4) follows. Q.E.D.

COROLLARY. *Let $\varepsilon>0$. There exists $\delta_{\varepsilon}>0$ such that for $0<\delta<\delta_{\varepsilon}$, we have*

$$0\le u_n^{\delta}(x)-u_n(x)<\varepsilon \quad \text{in } \mathbb{R} \quad \text{for } n\ge 0.$$

Proof. From Lemma 6, induction on (4.4), and the fact that $u_0^{\delta}\equiv u_0$, we obtain $\|u_n^{\delta}/u_n-1\|_{\infty}\le 1/g(\theta(\delta))-1$ for $n\ge 0$. Since $g(\theta(\delta))$ is right continuous at $\delta=0$ and $g(\theta(0))=1$, there exists a $\delta_{\varepsilon}>0$ such that $0\le 1/g(\theta(\delta))-1<\varepsilon$ if $0<\delta<\delta_{\varepsilon}$. Thus $0\le u_n^{\delta}(x)/u_n(x)-1<\varepsilon$ in $\mathbb{R}$ for $n\ge 0$. The corollary then follows by multiplying by $u_n(x)$.

LEMMA 7. *Let $K(x)$ be PF_3 and let $g(u)$ satisfy the conditions $0<g(u)<1$ in $(0,1)$ and (4.1). Let $u_0\in\mathcal{C}$ satisfy the hypotheses stated in Lemma 5. Then given $\varepsilon>0$, $\mathbb{K}=[\underline{\gamma},\bar{\gamma}]$ in $(0,1)$, there exist $\delta_{\varepsilon}>0$ and $N_{\varepsilon,\mathbb{K}}$ such that*

$$\left|\psi_n^{\delta,+}(\gamma)-\psi_n^{+}(\gamma)\right|<\varepsilon \quad \text{for } \gamma\in\mathbb{K},\quad n>N_{\varepsilon,\mathbb{K}} \quad \text{and } 0<\delta<\delta_{\varepsilon}.$$

Proof. The first step is to show that $m_n^{\delta,+}(\gamma)$ and $m_n^{+}(\gamma)$ are close for $\gamma\in\mathbb{K}$ and n sufficiently large. Let $\varepsilon>0$ be such that $\underline{\gamma}>\varepsilon>0$. Then there exist $\delta_{\varepsilon}>0$ and $N_1=N_1(\varepsilon,\mathbb{K})$ such that

$$0\le m_n^{\delta,+}(\gamma)-m_n^{+}(\gamma)\le m_n^{+}(\gamma-\varepsilon)-m_n^{+}(\gamma) \tag{4.6}$$

for $\gamma\in\mathbb{K}$, $n\ge N_1$, and $0<\delta<\delta_{\varepsilon}$. The left-hand inequality follows from the fact that $u_n^{\delta}\ge u_n$ for all n, while the right-hand inequality follows from the Corollary of Lemma 6. N_1 is chosen so that every term in (4.6) is defined for $n\ge N_1$, $\gamma\in\mathbb{K}$.

From Lemma 5, we have for $\gamma\in\mathbb{K}$, $n\geq N_1$ (possibly increased)

$$m_n^+(\gamma-\varepsilon)-m_n^+(\gamma)=\int_\gamma^{\gamma-\varepsilon}\frac{1}{\psi_n^+(\eta)}d\eta$$
$$\leq\int_\gamma^{\gamma-\varepsilon}\frac{1}{\omega_{c_0}(\eta)}d\eta=w_{c_0}^{-1}(\gamma-\varepsilon)-w_{c_0}^{-1}(\gamma).$$

This, together with (4.6) and the fact that $w_{c_0}'(x)<0$ in $\mathbb{R}$, accomplishes the first step of our proof.

We now come to $\psi_n^{\delta,+}(\gamma)-\psi_n^+(\gamma)$. Let $\varepsilon>0$. Since $K(x)$ is PF_3, it is absolutely continuous, and we can write

$$\psi_{n+1}^{\delta,+}(\gamma)-\psi_{n+1}^+(\gamma)=\int K'\big(m_{n+1}^{\delta,+}(\gamma)-y\big)\big[g^\delta\big(u_n^\delta(y)\big)-g\big(u_n^\delta(y)\big)\big]\,dy$$
$$+\int K'\big(m_{n+1}^{\delta,+}(\gamma)-y\big)\big[g\big(u_n^\delta(y)\big)-g\big(u_n(y)\big)\big]\,dy$$
$$+\int\big[K'\big(m_{n+1}^{\delta,+}(\gamma)-y\big)-K'\big(m_{n+1}(\gamma)-y\big)\big]g\big(u_n(y)\big)\,dy$$
$$\equiv I_1+I_2+I_3.$$

Furthermore,

$$|I_1|\leq\int_{\{u_n^\delta>1-\delta\}}\big|K'\big(m_{n+1}^{\delta,+}(\gamma)-y\big)\big|\big(1-g\big(u_n^\delta(y)\big)\big)\,dy\leq[1-g(1-\delta)]\|K'\|_1.$$

$|I_2|\leq\beta\varepsilon\|K'\|_1$ whenever $0<\delta<\bar\delta_\varepsilon$, where $\bar\delta_\varepsilon$ is from the corollary of Lemma 6. For I_3, note that since $K(x)$ is uniformly continuous, there exists $\bar\varepsilon>0$ such that $|K(x)-K(y)|<\varepsilon$ if $|x-y|<\bar\varepsilon$. With this $\bar\varepsilon$ and our first step, there exists $N_1=N_1(\bar\varepsilon,\mathbb{K})$ and $\delta_{\bar\varepsilon}>0$ such that $0\leq m_{n+1}^{\delta,+}(\gamma)-m_{n+1}^+(\gamma)<\bar\varepsilon$ whenever $\gamma\in\mathbb{K}$, $n+1\geq N_1$, $0<\delta<\delta_{\bar\varepsilon}$. Integrating by parts, we get, if $\gamma\in\mathbb{K}$, $n+1\geq N_1$, $0<\delta<\delta_{\bar\varepsilon}$,

$$|I_3|=\left|\int\big[K\big(m_{n+1}^{\delta,+}(\gamma)-y\big)-K\big(m_{n+1}^+(\gamma)-y\big)\big]g'\big(u_n(y)\big)u_n'(y)\,dy\right|$$
$$\leq\beta\int_{\mathcal{J}}\big|K\big(m_{n+1}^{\delta,+}(\gamma)-y\big)-K\big(m_{n+1}^+(\gamma)-y\big)\big||u_n'(y)|\,dy$$
$$\leq\beta\varepsilon\|K'\|_1(B_2-B_1+2\varepsilon)$$

where $\mathcal{J}=[m_{n+1}^+(\gamma)-B_2-\varepsilon,\ m_{n+1}^+(\gamma)-B_1+\varepsilon]$. Putting all the pieces together, if $\gamma\in\mathbb{K}$, $n>N_3-1$ and $0<\delta<\min(\bar\delta_\varepsilon,\delta_{\bar\varepsilon})$, we have

$$\big|\psi_{n+1}^{\delta,+}(\gamma)-\psi_{n+1}^+(\gamma)\big|\leq\big([1-g(1-\delta)]+\beta\varepsilon+\varepsilon\beta(B_2-B_1+2\varepsilon)\big)\|K'\|_1.$$

This inequality implies our lemma. Q.E.D.

5. The case $u_0(-\infty)=0$. This section contains our main result, for the second case when $u_0(-\infty)=0$, and an interesting application of it. The idea is to prove that

$$\lim_{n\to+\infty}\psi_n^+(\gamma)=\omega_{c_+^*}(\gamma)\quad\text{uniformly in compact subsets of }(0,1).\tag{5.1}$$

In [3, Corollary of Lemma 13], we have shown that the inequality $\liminf_{n\to+\infty}\psi_n^+(\gamma)\geq\omega_{c_+^*}(\gamma)$ holds uniformly in compact subsets of $(0,1)$ when $u_0\in\mathcal{C}$, $u_0(+\infty)=0$ and $K(x)$ is PF_2. The difficulty lies in obtaining the opposite inequality.

THEOREM 1. *Let $K(x)$, $g(u)$ and $u_0(x)$ satisfy the conditions stated for them in Lemma 4. In addition, let $g(u)$ satisfy* (4.1). *Then for every $0<\gamma<1$, $-c_-^*<c_0^*<c_+^*$, we have*

$$\lim_{n\to+\infty} u_n\left(x+m_n^+(\gamma)\right)=w_{c_+^*}\left(x+w_{c_+^*}^{-1}(\gamma)\right)$$

uniformly in $[nc_0^*-m_n^+(\gamma),+\infty)$ *and* $\lim_{n\to+\infty} m_{n+1}^+(\gamma)-m_n^+(\gamma)=c_+^*$.

Proof. The assumption that $0<g(u)<1$ in Lemma 4 is really not a condition on $g(u)$ because of Lemma 1. At any rate, we first construct g^δ, Q^δ as in the beginning of §3, after which we obtain an approximation u_n^δ to u_n by setting $u_0^\delta\equiv u_0$, $u_{n+1}^\delta=Q^\delta[u_n^\delta]$ for $n\geq 0$. The major step here is to prove (5.1).

Let $\mathbb{K}=[\underline{\gamma},\bar{\gamma}]$ be a subset of $(0,1)$ and write

$$\psi_n^+(\gamma)-\omega_{c_+^*}(\gamma)=\psi_n^+(\gamma)-\psi_n^{\delta,+}(\gamma)+\psi_n^{\delta,+}(\gamma)-\omega_c(\gamma)+\omega_c(\gamma)-\omega_{c_+^*}(\gamma). \tag{5.2}$$

Given $\varepsilon>0$, from [3, Lemma 8], there exists $c>c_+^*$ such that $0\leq\omega_c(\gamma)-\omega_{c_+^*}(\gamma)\leq\varepsilon$ for $\gamma\in\mathbb{K}$. With this c and $\mathbb{K}$, let δ^* be that obtained from the corollary of Lemma 4. From Lemma 7, there exist $N_{\varepsilon,\mathbb{K}}$ and $\delta_\varepsilon>0$ such that $|\psi_n^{\delta,+}(\gamma)-\psi_n^+(\gamma)|<\varepsilon$ for $\gamma\in\mathbb{K}$, $n\geq N_{\varepsilon,\mathbb{K}}$ and $0<\delta<\delta_\varepsilon$. Choose δ such that $0<\delta<\min(\delta^*,\delta_\varepsilon)$. Again from the corollary of Lemma 4, there exists $N_{c,\mathbb{K},\delta,\varepsilon}$ such that $\psi_n^{\delta,+}(\gamma)\leq\omega_c(\gamma)+\varepsilon$ for $\gamma\in\mathbb{K}$, $n\geq N_{c,\mathbb{K},\delta,\varepsilon}$. Combining all these, we have from (5.2)

$$\psi_n^+(\gamma)\leq\omega_{c_+^*}(\gamma)+3\varepsilon \quad \text{for } \gamma\in\mathbb{K}, \quad n\geq\max(N_{\varepsilon,\mathbb{K}},N_{c,\mathbb{K},\delta,\varepsilon}).$$

This implies that $\limsup_{n\to+\infty}\psi_n^+(\gamma)\leq\omega_{c_+^*}(\gamma)$ uniformly for $\gamma\in\mathbb{K}$. When the foregoing is coupled with the remark at the beginning of this section, (5.1) is proved.

Next we show that

$$\begin{aligned}&u_n\left(x+m_n^+(\gamma)\right) \text{ converges to } w_{c_+^*}\left(x+w_{c_+^*}^{-1}(\gamma)\right)\\ &\text{uniformly in compact subsets of } \mathbb{R}.\end{aligned} \tag{5.3}$$

Differentiate $u_n(m_n^+(\varepsilon))=\varepsilon$ and $w_{c_+^*}(w_{c_+^*}^{-1}(\varepsilon))=\varepsilon$ to get

$$\int_\gamma^{u_n(x+m_n^+(\gamma))}\frac{d\eta}{\psi_n^+(\eta)}=x=\int_\gamma^{w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))}\frac{d\eta}{\omega_{c_+^*}(\eta)}$$

for a given $0<\gamma<1$ and n sufficiently large. Rearranging this, we have

$$\int_\gamma^{w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))}\left(\frac{1}{\omega_{c_+^*}(\eta)}-\frac{1}{\psi_n^+(\eta)}\right)d\eta-\int_{w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))}^{u_n(x+m_n^+(\gamma))}\frac{d\eta}{\psi_n^+(\eta)}=0. \tag{5.4}$$

If x lies in a compact set in $\mathbb{R}$, $w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))$ is bounded away from 0 and 1. (5.1) implies that the first integral in (5.4) converges to 0 as $n\to+\infty$. Since $\|\psi_n^+\|_\infty\leq\|K'\|_1$, (5.4) implies (5.3).

To get the rest, we let $\varepsilon>0$ and choose L_1 so that $1-\varepsilon<w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))<1$ in $(-\infty,L_1)$. Set $L_2=\min(L_1,m_n^+(1-\varepsilon)-m_n^+(\gamma))$. Then for $nc_0^*-m_n^+(\gamma)\leq x\leq L_2$, we have $nc_0^*\leq x+m_n^+(\gamma)\leq m_n^+(1-\varepsilon)$. Statement (1.2) and our assumption on c_0^* imply that $1-\varepsilon\leq u_n(x+m_n^+(\gamma))\leq 1$ for n sufficiently large. This implies that

$$\left|u_n\left(x+m_n^+(\gamma)\right)-w_{c_+^*}\left(x+w_{c_+^*}^{-1}(\gamma)\right)\right|<\varepsilon \quad \text{in } \left[nc_0^*-m_n^+(\gamma),L_2\right]. \tag{5.5}$$

Similarly,

$$\left|u_n\left(x+m_n^+(\gamma)\right)-w_{c_+^*}\left(x+w_{c_+^*}^{-1}(\gamma)\right)\right|<\varepsilon \quad \text{in } \left[L_4,+\infty\right), \tag{5.6}$$

where $L_4=\max(L_3, m_n^+(\varepsilon)-m_n^+(\gamma))$ and L_3 is chosen so that $0\le w_{c_+^*}(x+w_{c_+^*}^{-1}(\gamma))\le\varepsilon$ in $[L_3, +\infty)$. The set $[L_2, L_4]$ is bounded as a result of (5.1). Thus (5.3), (5.5) and (5.6) together imply the first statement of our theorem. The second statement follows from [3, Lemma 12]. Q.E.D.

COROLLARY. *Let $K(x)$ be PF_3 and let $u_0\in\mathcal{C}$ be such that $u_0(x)$ vanishes outside of (A_1, A_2). $u_0(x)$ is nonincreasing and positive in $(A_2-\varepsilon, A_2)$, nondecreasing and positive in $(A_1, A_1+\varepsilon)$ for some $\varepsilon>0$. Then for every $0<\gamma<1$,*

$$\left|u_n(x)-w_{c_+^*}\left(x+w_{c_+^*}^{-1}(\gamma)-m_n^+(\gamma)\right)-\bar{w}_{c_-^*}\left(x+\bar{w}_{c_-^*}^{-1}(\gamma)-m_n^-(\gamma)\right)+1\right|\to 0$$

as $n\to+\infty$ uniformly in $\mathbb{R}$.

Proof. Theorem 1 implies that for each $c_0^*\in(-c_-^*, c_+^*)$, we have

$$\left|u_n(x)-w_{c_+^*}\left(x+w_{c_+^*}^{-1}(\gamma)\right)-m_n^+(\gamma)\right|\to 0 \quad \text{as } n\to+\infty$$

uniformly in $[nc_0^*, +\infty)$. Let $\bar{w}_{c_-^*}(x+nc_-^*)$ be the travelling wave which faces left and is moving towards $-\infty$. That is, $\bar{w}_{c_-^*}(x)$ is nondecreasing, $\bar{w}_{c_-^*}(-\infty)=0$, $\bar{w}_{c_-^*}(+\infty)=1$. Then the counterpart to Theorem 1 is that $|u_n(x)-\bar{w}_{c_-^*}(x+\bar{w}_{c_-^*}^{-1}(\gamma)-m_n^-(\gamma))|\to 0$ as $n\to+\infty$ uniformly in $(-\infty, nc_0^*]$. The statement of the corollary follows from this and from the fact that $nc_0^*-m_n^{\pm}(\gamma)\to\mp\infty$ as $n\to+\infty$.

THEOREM 2. *Let $\bar{u}_0(x)$, $\bar{v}_0(x)$ be initial data so that Theorem* 1 *(or* [3, *Thm.* 6]*) is applicable to them. Assume that $\bar{u}_0(x)=0$ if $x\ge A_1$ and $\bar{v}_0(x)=0$ if $x\ge A_2$. Set*

$$m_n^{\bar{u}}(\gamma)=\sup\{x;\bar{u}_n(x)\ge\gamma\}, \qquad m_n^{\bar{v}}(\gamma)=\sup\{x:\bar{v}_n(x)\ge\gamma\}$$

for $0<\gamma<1$. Let $g(u)$ satisfy (4.1). *Then for $0<\gamma<1$,*

$$m_n^{\bar{u}}(\gamma)-m_n^{\bar{v}}(\gamma) \quad \textit{is bounded as } n\to+\infty.$$

Proof. We first prove the following:

CLAIM. *Set $z_0=\bar{u}_0+\bar{v}_0$, $z_{n+1}=Q[z_n]$. Then $z_n(x)\le\bar{u}_n(x)+\bar{v}_n(x)$ for all $n\ge 0$.*

Proof of claim. If we define $g(u)=1$ for $u>1$, then (4.1) actually implies the inequality $g(u+v)\le g(u)+g(v)$. To see this, we may assume without loss of generality that $u\ge v$. Then (4.1) holds for $u\ge 0$ and hence $g(u+v)/(u+v)\le g(u)/u$. This implies that

$$g(u+v)\le g(u)+\frac{v}{u}g(u)\le g(u)+g(v).$$

The claim now follows easily by induction and the above inequality.

We return to the proof of Theorem 2. First let $\varepsilon>0$, and set $u_0(x)=I_{(-\infty, A]}$, $v_0(x)=I_{(A-\varepsilon, A]}$, $d_0(x)=u_0(x)-v_0(x)$. From our claim, we have $u_n(x)\le v_n(x)+d_n(x)$, and hence

$$u_n\left(x+m_n^u(\gamma)\right)\le v_n\left(x+m_n^u(\gamma)\right)+d_n\left(x+m_n^u(\gamma)\right). \tag{5.7}$$

It is clear from the definitions of u_0, v_0 and d_0 that $u_n(x)=d_n(x-\varepsilon)$. From this we get $m_n^u(\gamma)=m_n^d(\gamma)+\varepsilon$. Substitute this into (5.7) and let $n\to+\infty$. If $m_n^u(\gamma)-m_n^v(\gamma)$ is unbounded, we get by our hypotheses the contradiction,

$$w_{c_+^*}\left(x+w_{c_+^*}^{-1}(\gamma)\right)\le 0+w_{c_+^*}\left(x+\varepsilon+w_{c_+^*}^{-1}(\gamma)\right).$$

Hence $m_n^u(\gamma)-m_n^v(\gamma)$ is bounded for these two special initial data. Now let $z_0(x)=\alpha v_0(x)$, $0<\alpha<1$. For some integer N, take $\varepsilon>0$ in the definition of $v_0(x)$ to be so small that $\sup_{\mathbb{R}} v_N(x)<1$. From (1.2), there exists an integer N_1 such that $z_{N_1}(x)\ge v_N(x)$. It is trivial to deduce from this fact that $z_{n+N_1}(x)\ge v_{n+N}(x)$ for all $n\ge 0$. Thus

$$m_{n+N_1}^z(\gamma)\ge m_{n+N}^v(\gamma) \quad \text{for all } n\ge 0.$$

Since $0<\alpha<1$, we have $z_0(x)\leq v_0(x)$, and hence

$$m_n^z(\gamma)\leq m_n^v(\gamma) \quad \text{for all } n\geq 0.$$

Combining all these, we get

$$m_{n+N_1}^v(\gamma)\geq m_{n+N_1}^z(\gamma)\geq m_{n+N}^v(\gamma)=m_{n+N_1}^v(\gamma)+m_{n+N}^v(\gamma)-m_{n+N_1}^v(\gamma).$$

Since $m_{n+N}^v(\gamma)-m_{n+N_1}^v(\gamma)\sim(N-N_1)c_+^*$, we find that $m_n^v(\gamma)-m_n^z(\gamma)$ is bounded as $n\to+\infty$ for any small $\varepsilon>0$, $0<\alpha<1$.

Finally, with our given $\bar{u}_0(x)$ and $\bar{v}_0(x)$, we set $A=\max(A_1,A_2)$. Then $\bar{u}_0(x)\leq u_0(x-A)$, $\bar{v}_0(x)\leq u_0(x-A)$, and hence

$$m_n^{\bar{u}}(\gamma)\leq m_n^u(\gamma)+A, \quad m_n^{\bar{v}}(\gamma)\leq m_n^u(\gamma)+A, \qquad n\geq 0. \tag{5.8}$$

We then choose $\varepsilon>0$, $0<\alpha<1$, each small enough that $\bar{u}_0(x)\geq z_0(x+s_1)$, $\bar{v}_0(x)\geq z_0(x+s_2)$ for some constants s_1, s_2. Then

$$m_n^{\bar{u}}(\gamma)\geq m_n^z(\gamma)-s_1, \quad m_n^{\bar{v}}(\gamma)\geq m_n^z(\gamma)-s_2, \qquad n\geq 0. \tag{5.9}$$

From (5.8), (5.9) and the fact that $m_n^u(\gamma)-m_n^z(\gamma)$ is bounded as $n\to+\infty$, the conclusion of Theorem 2 is obvious. Q.E.D.

Acknowledgment. This paper is a part of the author's doctoral dissertation, written under the guidance of Professor Hans Weinberger, to whom the author wishes to express his sincere thanks.

Note added in proof. Since this paper was submitted, Professor H. F. Weinberger has informed the author that he has obtained a characterization of PF_3 functions.

REFERENCES

[1] S. KARLIN, *Total Positivity*, vol. I, Stanford Univ. Press, Stanford, CA, 1968.

[2] R. LUI, *A nonlinear integral operator arising from a model in population genetics*, Ph.D. thesis, University of Minnesota, Minneapolis, August 1981.

[3] ———, *A nonlinear integral operator arising from a model in population genetics I. Monotone initial data*, this Journal, this issue, pp. 913–937.

[4] H. F. WEINBERGER, *Asymptotic behavior of a model in population genetics*, in Nonlinear Partial Differential Equations and Applications, J. Chadam, ed., Lecture Notes in Mathematics 648, Springer, New York, 1978, pp. 47–98.

[5] ———, *Long-time behavior of a class of biological models*, this Journal, 13 (1982). pp. 353–396.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0004 $01.00/0

NONEXISTENCE OF SMOOTH SOLUTIONS FOR SHEARING FLOWS IN A NONLINEAR VISCOELASTIC FLUID*

GUSTAF GRIPENBERG†

Abstract. This paper studies the development of singularities in smooth solutions of an equation arising from a perturbation problem for steady shearing flows in a nonlinear viscoelastic fluid. It is proved that under certain assumptions (that make the equation studied nonlinear), singularities will develop in the derivatives of the velocity of the flow, provided that some derivatives of the initial data are sufficiently large.

1. Introduction and statement of results. The purpose of this paper is to study the development of singularities in a classical solution of the equation

$$
\begin{aligned}
&v_t(x,t)=\sigma\left(\int_0^\infty a(\tau)v_x(x,t-\tau)\,d\tau\right)_x+f(x,t), \qquad x\in(0,h),\\
&v(0,t)=v(h,t)=0, \qquad t\geq 0,\\
&v(x,\tau)=v_0(x,\tau), \qquad 0\leq x\leq h, \quad -\infty<\tau\leq 0,
\end{aligned}
\tag{1.1}
$$

where σ, a, v_0 and f are given functions and the subindices t and x denote partial differentiation. It will be shown that singularities can arise in a finite time from certain smooth initial data if the function σ is increasing and not linear. Thus, the results of this paper extend those of [4], where it is assumed that $a(\tau)=e^{-\alpha\tau}$. In fact, it turns out that we need rather weak assumptions on the kernel a.

The equation (1.1) arises from a shearing perturbation of a steady shearing flow in the following way: Suppose that we have a rectilinear shearing flow with velocity components (in a fixed Cartesian coordinate system x,y,z) $v^x=0$, $v^y=v^*(x,t)$, $v^z=0$, $0<x<h$. Now we make the assumption that the shearing stress $T^{xy}(t)$ is given by $\sigma^*(\int_0^\infty a(\tau)v_x^*(x,t-\tau)\,d\tau)$, where σ^* is a certain nonlinear function and v_x^* is the shear rate. If the fluid is simple, then the equation of conservation of linear momentum becomes

$$
\rho v_t^*(x,t)=\sigma^*\left(\int_0^\infty a(\tau)v_x^*(x,t-\tau)\,d\tau\right)_x+\rho f(x,t),
$$

where $\rho>0$ is the density (assumed to be constant) and $f(x,t)$ represents the body and driving forces in the y-direction (assumed to be independent of y and z). The boundary conditions are taken to be of the form $v^*(0,t)=V_1$, $v^*(h,t)=V_2$. When $f\equiv 0$ this equation admits the solution $(xV_2+(h-x)V_1)/h$ (steady rectilinear flow). If one wants to study shearing perturbations and takes $v(x,t)=v^*(x,t)-(xV_2+(h-x)V_1)/h$, then one gets (1.1) with $\sigma(\xi)=(\sigma^*(\xi+k)-\sigma^*(k))/\rho$, where $k=(V_2-V_1)\int_0^\infty a(\tau)\,d\tau/h$. Observe that the results below cannot be applied to the case when $f(x,t)$ is e.g. a nonzero constant. For more details see, for example, [1], [4] and the references mentioned there.

* Received by the editors June 15, 1981, and in revised form October 16, 1981.
† Institute of Mathematics, Helsinki University of Technology, SF-02150, Espoo 15, Finland.

Here we will prove the following result (BV stands for bounded variation, $\mathbb{R}^+ = [0,\infty)$, $\mathbb{R}^- = (-\infty,0]$):

THEOREM. *Assume that* $h>0$ *and that*

$$a\in C^1(\mathbb{R}^+;\mathbb{R})\cap L^1(\mathbb{R}^+;\mathbb{R}),\qquad a(0)=1, \tag{1.2}$$

$$a'\in L^1(\mathbb{R}^+;\mathbb{R})\cap \mathrm{BV}_{\mathrm{loc}}(\mathbb{R}^+;\mathbb{R}), \tag{1.3}$$

$$\sigma\in C^2(\mathbb{R};\mathbb{R}),\qquad \sigma'(0)>0,\quad \sigma''(0)\neq 0, \tag{1.4}$$

$$f\in C^1([0,h]\times\mathbb{R}^+;\mathbb{R}),\qquad f(0,t)=f(h,t)=0, \tag{1.5}$$

$$v_0\in C^2([0,h]\times\mathbb{R}^-;\mathbb{R}),\qquad v_0(0,t)=v_0(h,t)=0, \tag{1.6}$$

$$\sup_{x\in[0,h],\,t\in\mathbb{R}^-}\left(|v_0(x,t)|+|v_{0tx}(x,t)|+|v_{0xx}(x,t)|\right)<\infty. \tag{1.7}$$

If now $T>0$ *and* $\sup_{x\in[0,h],\,t\in\mathbb{R}^-}(|v_{0t}(x,t)|+|v_{0x}(x,t)|)$ *and* $\sup_{x\in[0,h],\,t\in[0,T]}|f(x,t)|$ *are sufficiently small and*

$$\sup_{x\in[0,h],\,j\in\{0,1\}}\sigma''(0)\Bigg(\int_0^\infty a(\tau)v_{0tx}(x,-\tau)\,d\tau + (-1)^j\left|\sigma'\left(\int_0^\infty a(\tau)v_{0x}(x,-\tau)\,d\tau\right)\right|^{1/2}\int_0^\infty a(\tau)v_{0xx}(x,-\tau)\,d\tau\Bigg)$$
$$\times\left(1+\sup_{x\in[0,h],\,t\in\mathbb{R}^-}\left(|v_{0tx}(x,t)|+|v_{0xx}(x,t)|\right)+\sup_{x\in[0,h],\,t\in[0,T]}|f_x(x,t)|\right)^{-1/2}$$

is sufficiently large, then (1.1) *cannot have a solution that is twice continuously differentiable on* $[0,h]\times(-\infty,T]$.

The basic ideas in the proof of this theorem are the same as those in [4]; i.e., (1.1) is reduced to the form (2.10) and then one can use the techniques from [2].

The conditions (1.7) and $a,a'\in L^1(\mathbb{R}^+;\mathbb{R})$ are introduced in order to ensure that certain integrals like $\int_t^\infty a(\tau)v_{0xx}(x,t-\tau)\,d\tau$ are well defined, but it would be easy to use other assumptions instead of these. Since $a(0)>0$ and a is continuous, the hypothesis of the theorem is satisfied if e.g. $v_{0tx}(x,t)$ and $v_{0xx}(x,t)$ are positive and quite large at some point x when t belongs to some small interval $(-\tau,0]$ and elsewhere relatively small in absolute value.

Finally, we remark that (1.1) differs from the equation studied in [5] (and the references mentioned there), because in that paper the nonlinear function σ appears inside the integral (another difference is that there the existence of global smooth solutions is established). But this is of no great consequence because one could apparently use the same kind of argument to establish the breakdown of solutions in that case too. The idea would be to take $u(x,t)=v_t(x,t)$, $w(x,t)=v_x(x,t)$ instead of (2.4) and (2.5) below. The main difference would then be that the assumptions needed on the initial values would look slightly different.

2. Proof of the theorem. We may assume that $\sigma''(0)>0$ because the case $\sigma''(0)<0$ is treated in exactly the same manner.

Let $v\in C^2([0,h]\times(-\infty,T];\mathbb{R})$ satisfy (1.1). We will show that there exist constants δ_0 and M_0 (depending on a,σ and T) such that if

$$\sup_{x\in[0,h],\,t\in\mathbb{R}^-}\left(|v_{0t}(x,t)|+|v_{0x}(x,t)|\right)+\sup_{x\in[0,h],\,t\in[0,T]}|f(x,t)|<\delta_0 \tag{2.1}$$

and

$$\sup_{x\in[0,h]}\left(\int_0^\infty a(\tau)v_{0tx}(x,-\tau)\,d\tau\right.$$

$$+(-1)^j\left|\sigma'\left(\int_0^\infty a(\tau)v_{0x}(x,-\tau)\,d\tau\right)\right|^{1/2} \tag{2.2}$$

$$\left.\times\int_0^\infty a(\tau)v_{0xx}(x,-\tau)\,d\tau+\sup_{x\in[0,h],\,t\in[0,T]}|f_x(x,t)|\right)^{-1/2}>M_0$$

for $j=0$ or 1,

hold, then we get a contradiction.

By (1.4) there exist positive constants $\delta,\varepsilon_1,\varepsilon_2$ and k_1 such that

$$\varepsilon_1\le\inf_{|\xi|\le\delta}\sigma'(\xi),\quad \sup_{|\xi|\le\delta}\sigma'(\xi)\le k_1,\quad \inf_{|\xi|\le\delta}4^{-1}\sigma''(\xi)(\sigma'(\xi))^{-5/4}\ge\varepsilon_2. \tag{2.3}$$

We define

$$u(x,t)=\int_0^\infty a(\tau)v_t(x,t-\tau)\,d\tau,\qquad x\in[0,h],\quad t\in[0,T] \tag{2.4}$$

and

$$w(x,t)=\int_0^\infty a(\tau)v_x(x,t-\tau)\,d\tau,\qquad x\in[0,h],\quad t\in[0,T]. \tag{2.5}$$

Next we derive (2.10). An integration by parts gives by (2.4) (recall that $a(0)=1$),

$$u(x,t)=v(x,t)+\int_0^\infty a'(\tau)v(x,t-\tau)\,d\tau,$$

and if both sides of this equation are differentiated with respect to t, then one obtains

$$u_t(x,t)=v_t(x,t)+\int_0^\infty a'(\tau)v_t(x,t-\tau)\,d\tau,\qquad x\in[0,h],\quad t\in[0,T]. \tag{2.6}$$

It follows from (1.3), and standard Volterra integral equation theory, see e.g. [3], that we can define a locally finite, continuous (except perhaps at 0) Borel measure α on $\mathbb{R}^+$ by the equation

$$\alpha([0,t])+\int_0^t a'(t-\tau)\alpha([0,\tau])\,d\tau=a'(t),\qquad t\in\mathbb{R}^+,$$

or equivalently

$$\int_{[0,t]}a(t-\tau)\,d\alpha(\tau)=a'(t),\qquad t\in\mathbb{R}^+. \tag{2.7}$$

If we take the convolution of both sides of (2.4) with the measure α and use (2.7), then we conclude that

$$\int_0^\infty a'(\tau)v_t(x,t-\tau)\,d\tau=\int_{[0,t]}u(x,t-\tau)\,d\alpha(\tau)+g(x,t), \tag{2.8}$$

$$x\in[0,h],\quad t\in[0,T],$$

where

$$g(x,t)=\int_t^\infty a'(\tau)v_t(x,t-\tau)\,d\tau-\int_{[0,t]}\int_{t-\tau}^\infty a(\xi)v_t(x,t-\tau-\xi)\,d\xi\,d\alpha(\tau). \tag{2.9}$$

Now we conclude from (1.1), (2.4)–(2.6) and (2.8) that u and w are continuously differentiable on $[0,h]\times[0,T]$ and satisfy

$$\begin{aligned} &w_t(x,t)=u_x(x,t),\\ &u_t(x,t)=\sigma(w(x,t))_x+\int_{[0,t]}u(x,t-\tau)\,d\alpha(\tau)+F(x,t) \end{aligned} \tag{2.10}$$

when $(x,t)=(0,h)\times[0,T]$, where

$$F(x,t)=f(x,t)+g(x,t). \tag{2.11}$$

We extend the functions $u(x,t)$ and $F(x,t)$ as odd functions of x and the function $w(x,t)$ as an even one, all with period $2h$ in x. Then we see from the boundary conditions and (2.10) that if $|w(x,t)|\le\delta$, $x\in[0,h]$, $t\in[0,\tau]$, $\tau\le T$, then the functions u, w and F extended in this way are continuously differentiable on $\mathbb{R}\times[0,\tau]$.

Next we show that if (2.1) holds, and δ_0 is sufficiently small, then

$$\sup_{x\in\mathbb{R},\,t\in[0,T]}\left(|u(x,t)|+|w(x,t)|\right)\le\delta. \tag{2.12}$$

From (2.4) and (2.5) we see that if δ_0 is small enough, then $|u(x,0)|+|w(x,0)|<\delta$, and therefore, there exists $\tau_0\in[0,T]$ such that (2.12) holds with T replaced by τ_0. We define the characteristic curves $x_1(t,\beta)$ and $x_2(t,\gamma)$ at least on $[0,\tau_0]$, by the equations

$$\begin{aligned} &\frac{dy(t)}{dt}=\lambda(y(t),t), \qquad y(0)=\beta,\\ &\frac{dy(t)}{dt}=\mu(y(t),t), \qquad y(0)=\gamma, \end{aligned} \tag{2.13}$$

where

$$\left.\begin{matrix}\lambda(x,t)\\ \mu(x,t)\end{matrix}\right\}=\mp\sigma'(w(x,t))^{1/2}. \tag{2.14}$$

We take $\phi(w)=\int_0^w\sigma'(t)^{1/2}\,dt$ and

$$\left.\begin{matrix}r(x,t)\\ s(x,t)\end{matrix}\right\}=u(x,t)\pm\phi(w(x,t)). \tag{2.15}$$

With these definitions, the equations (2.10) become (note that $u=(r+s)/2$)

$$\begin{aligned} &\left.\begin{matrix}\frac{d}{dt}r(x_1(t,\beta),t)\\ \frac{d}{dt}s(x_2(t,\gamma),t)\end{matrix}\right\}\\ &\qquad=\left(2^{-1}\int_{[0,t]}(r(x,t-\tau)+s(x,t-\tau))\,d\alpha(\tau)+F(x,t)\right)\Bigg|_{\substack{x=x_1(t,\beta)\\ x=x_2(t,\gamma)}} \end{aligned} \tag{2.16}$$

Define

$$
\begin{aligned}
&R(t)=\sup_{x\in\mathbb{R}}|r(x,t)|, \qquad S(t)=\sup_{x\in\mathbb{R}}|s(x,t)|,\\
&h(t)=\int_0^t \sup_{x\in\mathbb{R}}|F(x,\tau)|\,d\tau.
\end{aligned}
\tag{2.17}
$$

If we integrate the equations (2.16) and use the definitions (2.17) in deriving an upper bound for the right-hand side, then we obtain

$$
\left.\begin{matrix}|r(x_1(t,\beta),t)|\le R(0)\\ |s(x_2(t,\beta),t)|\le S(0)\end{matrix}\right\}+h(t)+2^{-1}\int_0^t\big(R(t-\tau)+S(t-\tau)\big)\operatorname{var}(\alpha;[0,\tau])\,d\tau,
$$

$$
t\in[0,\tau_0].
$$

It follows from (2.3), (2.13) and (2.14) that for each $t\in[0,\tau_0]$ we can choose β and γ so that $|r(x_1(t,\beta),t)|=R(t)$ and $|s(x_2(t,\gamma),t)|=S(t)$. Thus we have

$$
R(t)+S(t)\le R(0)+S(0)+2h(t)+\int_0^t\big(R(t-\tau)+S(t-\tau)\big)\operatorname{var}(\alpha;[0,\tau])\,d\tau,
$$

and by Gronwall's inequality we deduce that

$$
R(t)+S(t)\le\big(R(0)+S(0)+2h(T)\big)\exp\big(t\operatorname{var}(\alpha;[0,T])\big), \qquad t\in[0,\tau_0]. \tag{2.18}
$$

If we recall the definitions (2.9), (2.11), (2.15) and (2.17), then we conclude from (2.1), (2.3) and (2.18) that $R(t)+S(t)$ can be made arbitrarily small by choosing δ_0 small enough, and we see that we can take $\tau_0=T$, i.e., the inequality (2.12) holds (since the mapping $(u,w)\to(r,s)$ has a continuous inverse as long as $|w|\le\delta$).

Now that we have picked δ_0 so that (2.1) yields (2.12), we choose M_0 to be so large that

$$
\begin{aligned}
M_0>\max\Bigg\{&4\big(T\varepsilon_1^{1/2}\varepsilon_2\big)^{-1},2\left(\frac{k_1}{\varepsilon_1\varepsilon_2}\right)^{1/2}\\
&\times\left(1+\|a'\|_{L^1(\mathbb{R}^+)}+\operatorname{var}(\alpha;[0,T])\|a\|_{L^1(\mathbb{R}^+)}\left(1+2\left(\frac{k_1}{\varepsilon_1}\right)^{1/2}\big(1+k_1^{1/2}\big)\right)\right)\Bigg\}.
\end{aligned}
\tag{2.19}
$$

Since (2.3), (2.12) and (2.13) hold, (2.10) or (2.15) can be rewritten in the form

$$
\left.\begin{matrix}r_t(x,t)+\lambda(x,t)r_x(x,t)\\ s_t(x,t)+\mu(x,t)s_x(x,t)\end{matrix}\right\}=2^{-1}\int_{[0,t]}\big(r(x,t-\tau)+s(x,t-\tau)\big)\,d\alpha(\tau)+F(x,t),
\tag{2.20}
$$

$$
x\in\mathbb{R},\quad t\in[0,T].
$$

Define

$$
\begin{aligned}
&\theta(x,t)=\mu(x,t)^{1/2}r_x(x,t),\\
&\psi(x,t)=\mu(x,t)^{1/2}s_x(x,t), \quad x\in R,\quad t\in[0,T].
\end{aligned}
\tag{2.21}
$$

Next we want to show that

$$\left.\begin{aligned}&\frac{d}{dt}\theta(x_1(t,\beta),t)\\&\frac{d}{dt}\psi(x_2(t,\gamma),t)\end{aligned}\right\}$$

$$=\left(4^{-1}\sigma''(w(x,t))\sigma'(w(x,t))\right)^{-5/4}\times\begin{cases}\theta(x,t)^2\\ \psi(x,t)^2\end{cases}$$

$$+2^{-1}\mu^{1/2}(x,t)\int_{[0,T]}\mu(x,t-\tau)^{-1/2}(\theta(x,t-\tau)+\psi(x,t-\tau))\,d\alpha(\tau)$$

$$+\mu^{1/2}(x,t)F_x(x,t)\Big|_{\substack{x=x_1(t,\beta)\\ x=x_2(t,\gamma)}} \tag{2.22}$$

These two equations will follow from (2.20) by the same calculations that were used in [4, pp. 219–220], provided that we can show that $r_x(x_1(t,\beta),t)$ is a continuously differentiable function of t with derivative

$$\frac{d}{dt}r_x(x_1(t,\beta),t)=\left(\frac{d}{dx}(r_t(x,t)+\lambda(x,t)r_x(x,t))-\lambda_x(x,t)r_x(x,t)\right)_{x=x_1(t,\beta)} \tag{2.23}$$

(and similarily for $s_x(x_2(t,\gamma),t)$). This is a consequence of the fact that by (2.3), (2.12) and (2.14) we know that λ and μ are continuously differentiable on $\mathbb{R}\times[0,T]$ and because it is also clear that the right-hand side in (2.20) is a continuously differentiable function of x (here we use the fact that α is a continuous measure).

Now we establish the continuous differentiability of $r_x(x_1(t,\beta),t)$ (the argument for $s_x(x_2(t,\gamma),t)$ is the same). Let φ be a C^∞ function with compact support in $\mathbb{R}\times[0,T]$ and denote the right-hand side in (2.20) by $d(x,t)$. Then we deduce from some partial integrations and (2.20) that

$$-\int_{\mathbb{R}}\int_0^T r_x(x,t)(\varphi_t(x,t)+(\lambda\varphi)_x(x,t))\,dt\,dx$$

$$=-\int_{\mathbb{R}}\int_0^T(r_t(x,t)\varphi_x(x,t)+r_x(x,t)\lambda(x,t)\varphi_x(x,t)+r_x(x,t)\lambda_x(x,t)\varphi(x,t))\,dt\,dx$$

$$=\int_{\mathbb{R}}\int_0^T(d_x(x,t)-r_x(x,t)\lambda_x(x,t))\varphi(x,t)\,dt\,dx. \tag{2.24}$$

Changing variables $t=\tau$, $x=x_1(\tau,\beta)$, we conclude that

$$(\varphi_t(x,t)+(\lambda\varphi)_x(x,t))\,dt\,dx=\frac{d}{dt}\left(\varphi(x_1(\tau,\beta),\tau)\,\frac{\partial(x,t)}{\partial(\beta,\tau)}\right)d\tau\,d\beta. \tag{2.25}$$

(Observe that the Jacobian $\partial(x,t)/\partial(\beta,\tau)=\exp\int_0^\tau\lambda_x(x_1(\xi,\beta),\xi)\,d\xi$ is continuously differentiable with respect to τ.) Take ϕ_0 to be a continuous function with compact support in $\mathbb{R}$ and ϕ_1 a continuously differentiable function with compact support in $(0,T)$ and replace the function φ by a sequence φ_n such that $\varphi_n(x_1(\tau,\beta),\tau)\partial(x,t)/\partial(\beta,t)$ and $(d/d\tau)(\varphi_n(x_1(\tau,\beta),\tau)\partial(x,t)/\partial(\beta,t))$ converge uniformly to $\phi_0(\beta)\phi_1(\tau)$ and

$\phi_0(\beta)\phi_1'(\tau)$. Then we see from (2.24) and (2.25) that

$$-\int_{\mathbb{R}}\phi_0(\beta)\int_0^T r_x(x_1(\tau,\beta),\tau)\phi_1'(\tau)\,d\tau\,d\beta$$

$$=\int_{\mathbb{R}}\phi_0(\beta)\int_0^T\big(d_x(x_1(t,\beta),\tau)-r_x(x_1(\tau,\beta),\tau)\lambda_x(x_1(\tau,\beta),\tau)\big)\phi_1(\tau)\,d\tau\,d\beta,$$

and since ϕ_0 and ϕ_1 are arbitrary, we get (2.23). Thus, we obtain (2.22) as noted above.

Define

$$\begin{aligned} M(t)&=\sup_{x\in\mathbb{R},\,\tau\in[0,t]}\{|\theta(x,\tau)|,|\psi(x,\tau)|\},\\ K_T&=\sup_{x\in\mathbb{R},\,t\in[0,T]}\big|\mu(x,t)^{1/2}F_x(x,t)\big|,\\ L_T&=(k_1/\varepsilon_1)^{1/2}\mathrm{var}(\alpha;[0,T]). \end{aligned}\tag{2.26}$$

The equations (2.22) imply that (see (2.3), (2.12) and (2.26))

$$\left.\begin{aligned} \frac{d}{dt}\theta(x_1(t,\beta),t)&\ge\varepsilon_2\theta(x_1(t,\beta),t)^2\\ \frac{d}{dt}\psi(x_2(t,\gamma),t)&\ge\varepsilon_2\psi(x_2(t,\gamma),t)^2 \end{aligned}\right\}-K_T-L_TM(t),\qquad t\in[0,T].\tag{2.27}$$

The sequence $\{\tau_n\}_{n=0}^{\infty}$ is defined by

$$\tau_0=0,\qquad \tau_{n+1}=\min\{T,\inf\{\tau|\,\tau\in[0,T],\,M(\tau)=2M(\tau_n)\}\}.\tag{2.28}$$

We will show that $\sup_{n\ge0}\tau_n<T$, since this fact would give the desired contradiction as $M(0)>\varepsilon_1^{1/2}M_0>0$.

Assume that $n\ge0$, $\tau_n<T$ and that $x_0\in\mathbb{R}$, $t_0\in[\tau_n,\tau_{n+1}]$ are such that for example $\psi(x_0,t_0)=\inf_{x\in\mathbb{R},\,t\in[\tau_n,\tau_{n+1}]}\psi(x,t)<-M(\tau_n)$. It is possible to choose γ so that $x_2(t_0,\gamma)=x_0$ and at this point the derivative of ψ along the characteristic curve $x_2(t,\gamma)$ is positive since $\varepsilon_2M(\tau_n)^2-K_T-2L_TM(\tau_n)>0$ by (2.2)–(2.5), (2.12), (2.14), (2.15), (2.19), (2.21) and (2.28). But this is impossible in view of our choice of (x_0,t_0). Thus, we have established that if $\tau_n<T$, then

$$\inf_{x\in\mathbb{R}}\{\theta(x,t),\psi(x,t)\}\ge-M(\tau_n),\qquad t\in[t_n,t_{n+1}].\tag{2.29}$$

Define

$$d_n=\sup_{x\in\mathbb{R}}\frac{\{\theta(x,\tau_n),\psi(x,\tau_n)\}}{M(\tau_n)}.\tag{2.30}$$

Clearly $1\ge d_0>0$ and if $\tau_{n+1}<T$, then $d_{n+1}=1$, $n\ge0$ by (2.28) and (2.29). Let $n\ge0$ and assume that $\tau_n<T$. In order to derive an upper bound for $\tau_{n+1}-\tau_n$, we observe that if $y(t)$ is continuously differentiable and satisfies $(d/dt)y(t)\ge c_1y(t)^2-c_2$, $c_2y(0)^2>c_2$, then

$$\begin{aligned} y(t)\ge&\left(\frac{c_2}{c_1}\right)^{1/2}\left(1+\left(y(0)-\left(\frac{c_2}{c_1}\right)^{1/2}\right)\right)\left(y(0)+\left(\frac{c_2}{c_1}\right)^{1/2}\right)^{-1}\\ &\times\exp\big(2(c_1c_2)^{1/2}t\big)\left(1-\left(y(0)-\left(\frac{c_2}{c_1}\right)^{1/2}\right)\right.\\ &\left.\times\left(y(0)+\frac{c_2}{c_1}\right)^{1/2}\exp\big(2(c_1c_2)^{1/2}t\big)\right)^{-1}. \end{aligned}$$

Thus, if y is to remain finite, we must have

$$t<\left(\frac{4c_1}{c_2}\right)^{-1/2}\ln\left(\left(y(0)+\left(\frac{c_2}{c_1}\right)^{1/2}\right)\Big/\left(y(0)-\left(\frac{c_2}{c_1}\right)^{1/2}\right)\right)$$

$$\leq(c_1y(0))^{-1}\left(1-\left(c_2\left(c_1y(0)^2\right)^{-1}\right)^{1/2}\right)^{-1}. \tag{2.31}$$

When we apply this result to (2.27) we choose β or γ so that $\theta(x_1(\tau_n,\beta),\tau_n)$ or $\psi(x_2(\tau_n,\gamma),\tau_n)=d_nM(\tau_n)$, and we take $c_1=\varepsilon_2$, $c_2=K_T+2L_TM(\tau_n)$ and $y(0)=d_nM(\tau_n)$. It follows from (2.19) and (2.2)–(2.5), (2.12), (2.14), (2.15), (2.21), (2.28) and (2.30) that we always have $(K_T+2L_TM(\tau_n))(d_n^2M(\tau_n)^2\varepsilon_2)^{-1}<1/4$, and therefore, it follows from (2.31) that $\tau_{n+1}-\tau_n<2(\varepsilon_2d_nM(\tau_n))^{-1}$. Using this inequality we deduce that

$$\tau_n=\sum_{j=0}^{n-1}(\tau_{j+1}-\tau_j)<2\left(\varepsilon_2d_0M(0)\right)^{-1}\sum_{j=0}^{n-1}2^{-j}<4\left(\varepsilon_1^{1/2}\varepsilon_2M_0\right)^{-1},$$

because $\varepsilon_1^{1/2}M_0<d_0M(0)$ by (2.2), (2.3), (2.14), (2.21) and (2.30). From (2.19) we see that $\sup_{n\geq0}\tau_n<T$, and hence, we obtain the desired contradiction.

REFERENCES

[1] B. O. COLEMAN AND M. E. GURTIN, *On the stability against shear waves of steady flows of nonlinear viscoelastic fluids*, J. Fluid. Mech., 33 (1968), pp. 165–181.

[2] P. D. LAX, *Development of singularities of solutions of nonlinear hyperbolic partial differential equations*, J. Math. Physics, 5 (1964), pp. 611–613.

[3] R. K. MILLER, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.

[4] M. SLEMROD, *Instability of steady shearing flows in a nonlinear viscoelastic fluid*, Arch. Rational Mech. Anal., 68 (1978), pp. 211–225.

[5] O. J. STAFFANS, *On a nonlinear hyperbolic Volterra equation*, this Journal, 11 (1980), pp. 793–812.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0005 $01.00/0

COMPLETE MONOTONICITY AND RESOLVENTS OF VOLTERRA INTEGRODIFFERENTIAL EQUATIONS*

KENNETH B. HANNSGEN† AND ROBERT L. WHEELER†

Abstract. We prove that if $u'(t)+\int_0^t a(t-y)u(y)\,dy=0$ $(t\geq 0)$, $u(0)=1$, where a is completely monotonic and locally integrable, but not constant, then $-u$ is the sum of a completely monotonic function and a function which decays exponentially as $t\to\infty$.

1. Introduction.

Concerning the problem

$$u'(t)+\int_0^t a(t-y)u(y)\,dy=0 \qquad (t\geq 0), \quad u(0)=1, \tag{1.1}$$

$(' = \frac{d}{dt})$, we prove the following theorem.

THEOREM 1. *Let a be completely monotonic on $(0,\infty)$ with $0\leq a(\infty)<a(0+)\leq\infty$ and $\int_0^1 a(t)\,dt<\infty$. Let u satisfy* (1.1). *Then there exist positive numbers K and ε and a finite nonnegative measure μ on $[-\varepsilon,0]$ such that*

$$\begin{aligned} u(t)&=-\int_{-\varepsilon}^{0} e^{\sigma t}\,d\mu(\sigma)+u_1(t),\\ |u_1(t)|&\leq Ke^{-\varepsilon t} \qquad (t>0). \end{aligned} \tag{1.2}$$

In particular, u_1-u is completely monotonic and u is negative for all sufficiently large t unless u decays exponentially.

The function u is important because of the resolvent formula

$$x(t)=u(t)x_0+\int_0^t u(t-y)f(y)\,dy$$

for solutions of

$$x'(t)+\int_0^t a(t-y)x(y)\,dy=f(t), \qquad x(0)=x_0.$$

For the Volterra integral equation

$$x(t)+\int_0^t a(t-y)x(y)\,dy=f(t),$$

the appropriate resolvent $r(t)$ satisfies

$$r(t)+\int_0^t a(t-y)r(y)\,dy=a(t);$$

Reuter [14] has shown that r is completely monotonic if a satisfies the hypotheses of Theorem 1.

Equation (1.1) has been studied intensively since 1960; see [12], [11], [6], [7], [17], [5], [1], [9], [4], for example. In 1975 Shea and Wainger [15] proved that $u\in L^1(0,\infty)$ when $a(t)$ is a locally integrable, nonnegative, nonincreasing, convex function which is not piecewise linear and of a special form. The proof of this result relies on deep

* Received by the editors August 17, 1981. This material is based upon work supported by the National Science Foundation under grant MCS 8101618.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

frequency domain techniques. We remark that in the case where $a(t)$ is locally integrable, nonconstant, positive and nonincreasing, and $\log a(t)$ is convex, then one can show, without using transform theory, that the integral resolvent $r(t)$ belongs to $L^1(0,\infty)$ [3], [13]. From this result one can deduce that $u(t)$ also belongs to $L^1(0,\infty)$ by using local analyticity as recently studied in [10, Props. 7.3, 7.4].

For the case where $a(t)$ is completely monotonic, (1.2) easily yields yet another proof that $u\in L^1(0,\infty)$; we present this proof in §3. In that section we also give a formula for the measure μ in a special case.

2. Proof of Theorem 1. Since $a(t)$ is completely monotonic and not constant, Bernstein's theorem yields a nondecreasing function $\alpha(t)$ on $[0,\infty)$ with $0=\alpha(0)\le \alpha(0^+)<\alpha(\infty)=a(0^+)\le\infty$ and $\alpha(x)=\alpha(x^-)$ for $0<x<\infty$, such that

$$a(t)=\int_0^\infty e^{-xt}\,d\alpha(x) \qquad (t>0).$$

Using the additional fact that $a\in L^1(0,1)$, we get that the Laplace transform

$$\hat a(s)\equiv\int_0^\infty e^{-st}a(t)\,dt$$

exists for $s=\sigma+i\tau$ with $\sigma>0$. It follows from the theory of Laplace and Stieltjes transforms [16, Chapt. 8] that $\hat a(s)$ can be analytically continued to the slit plane $\mathbb{C}'\equiv\mathbb{C}\setminus(-\infty,0]$ by the formula

$$\hat a(s)\equiv\int_0^\infty \frac{d\alpha(x)}{s+x}, \qquad s\in\mathbb{C}', \tag{2.1}$$

where the integral converges uniformly for s in any compact subset of $\mathbb{C}'$.

Define $D(s)\equiv R(s)+iI(s)$ by

$$D(s)=s+\hat a(s), \qquad s\in\mathbb{C}'. \tag{2.2}$$

By (2.1) and elementary algebra, we get the formulas

$$R(s)=\sigma+\int_0^\infty \frac{(\sigma+x)\,d\alpha(x)}{(\sigma+x)^2+\tau^2}, \qquad s\in\mathbb{C}', \tag{2.3}$$

$$I(s)=\tau-\int_0^\infty \frac{\tau\,d\alpha(x)}{(\sigma+x)^2+\tau^2}, \qquad s\in\mathbb{C}'. \tag{2.4}$$

Observe that $D(\bar s)=\overline{D(s)}$, $s\in\mathbb{C}'$.

LEMMA 2.1. *The boundary value*

$$D(\sigma)\equiv\lim_{\tau\to 0^+} D(\sigma+i\tau)$$

exists and is finite and nonzero for a.e. $\sigma\in(-\infty,0)$.

Proof. Since $\operatorname{Im}\hat a(s)<0$ for s in the upper half plane $\Pi^+\equiv\{s:\operatorname{Im}s>0\}$, the function $A(s)=(\hat a(s)-i)^{-1}+(s+i)^{-1}$ is a nonconstant bounded analytic function on Π^+. Thus, the limit $A(\sigma)\equiv\lim_{\tau\to0^+}A(\sigma+i\tau)$ exists and is finite and nonzero for a.e. $\sigma\in(-\infty,\infty)$. Then $\lim_{\tau\to0^+}(\hat a(\sigma+i\tau)-i)$ exists and is bounded away from zero a.e. on $(-\infty,\infty)$. Since $D(s)=(\hat a(s)-i)(s+i)A(s)$ for $s\in\Pi^+$, the proof of Lemma 2.1 is complete. □

Next we use the expressions (2.3) and (2.4) to obtain estimates for $R(s)$ and $I(s)$ that hold when s is near the cut $s=\sigma+i\tau$, $\sigma<0$, $\tau=0$, and in a neighborhood of zero.

LEMMA 2.2. *There exist positive constants η, δ and B so that for each $s=\sigma+i\tau$ satisfying $-\eta\le\sigma\le 0$, $0<|\tau|\le\delta$, either*

$$\frac{I(s)}{\tau}\le -1 \tag{2.5}$$

or

$$R(s)\ge B. \tag{2.6}$$

Moreover

$$|D(s)|\ge|\tau| \quad \text{for } -\eta\le\sigma\le 0, \quad 0<|\tau|\le\delta. \tag{2.7}$$

Proof. First consider the case where

$$\int_{0^+}^{\infty}\frac{d\alpha(x)}{x^2}\equiv\lim_{\varepsilon\to 0^+}\int_{\varepsilon}^{\infty}\frac{d\alpha(x)}{x^2}$$

is finite. Since $\alpha(0^+)<\alpha(\infty)$, we can find $\gamma>0$ so that

$$2\int_{\gamma}^{\infty}\frac{d\alpha(x)}{x^2}>\int_{0^+}^{\infty}\frac{d\alpha(x)}{x^2}>0.$$

Set $B=(\gamma/8)\int_{0^+}^{\infty} d\alpha(x)/x^2$ and $\eta=\min\{\gamma, B/3\}$. Finally, choose δ, $0<\delta<B$, so small that

$$2\int_{\gamma}^{\infty}\frac{d\alpha(x)}{x^2+\delta^2}\ge\int_{\gamma}^{\infty}\frac{d\alpha(x)}{x^2}.$$

Suppose that (2.5) fails for some $s=\sigma+i\tau$ satisfying $-\eta\le\sigma\le 0$, $0<|\tau|\le\delta$. It follows from (2.4) and our definition of η that

$$0\ge\sigma\left(1+\int_0^{\infty}\frac{d\alpha(x)}{(\sigma+x)^2+\tau^2}\right)\ge 3\sigma\ge -3\eta>-B. \tag{2.8}$$

Also, using our choices of γ, δ and B, and the fact that $\sigma\ge-\eta\ge-\gamma$, $0<|\tau|\le\delta$, we easily see that

$$\int_0^{\infty}\frac{x\,d\alpha(x)}{(\sigma+x)^2+\tau^2}\ge\int_{\gamma}^{\infty}\frac{x\,d\alpha(x)}{x^2+\tau^2}\ge\gamma\int_{\gamma}^{\infty}\frac{d\alpha(x)}{x^2+\tau^2}\ge 2B.$$

Combining this last estimate with (2.8) and formula (2.3), we see that (2.6) must hold whenever (2.5) fails. Also, the trivial inequality $|D(s)|\ge\min\{|R(s)|, |I(s)|\}$, and (2.5) or (2.6) together with the fact that $\delta<B$, yields (2.7), and the proof of Lemma 2.2 is complete when $\int_{0^+}^{\infty} d\alpha(x)/x^2<\infty$.

Next, suppose that $\int_{0^+}^{\infty} d\alpha(x)/x^2=\infty$. In this case the positive constants η and δ may be chosen so small that (2.5) always holds when $-\eta\le\sigma\le 0$, $0<|\tau|\le\delta$. Namely, choose $\eta>0$ so that $\int_{\eta}^{\infty} d\alpha(x)/x^2>4$, and $\delta>0$ so that

$$2\int_{\eta}^{\infty}\frac{d\alpha(x)}{x^2+\delta^2}>\int_{\eta}^{\infty}\frac{d\alpha(x)}{x^2}.$$

Then

$$\int_0^{\infty}\frac{d\alpha(x)}{(\sigma+x)^2+\tau^2}\ge\int_{\eta}^{\infty}\frac{d\alpha(x)}{x^2+\tau^2}\ge 2$$

whenever $-\eta\le\sigma\le 0$ and $0<|\tau|\le\delta$, and (2.5) now follows from formula (2.4). (2.7) follows trivially from (2.5). □

We now turn to the proof of Theorem 1. By rewriting (2.1) as

$$\hat{a}(s)=\int_0^\infty \frac{(1+x)}{(s+x)}\frac{d\alpha(x)}{1+x}, \qquad s\in\mathbb{C}',$$

we see, using the dominated convergence theorem, that

(2.9) $$\lim_{|\tau|\to\infty}\hat{a}(\sigma+i\tau)=0 \quad \text{uniformly in any half plane } -\infty<\sigma_0\le\sigma<\infty.$$

Let η,δ and B be the positive constants that occur in Lemma 2.2, and use (2.9) to find $\Delta\ge\delta$ so that $D(s)\neq 0$ when $-\eta\le\sigma<\infty$ and $|\tau|\ge\Delta$. $D(s)$ can have at most a finite number of zeros in the compact set $-\eta\le\sigma\le 0$, $\delta\le|\tau|\le\Delta$. Since $R(s)>0$ for $\sigma\ge 0$, $s\neq 0$, we can find ε, $0<\varepsilon\le\eta$ so that

(2.10) $$D(s)\neq 0 \quad \text{for } s\in\mathbb{C}' \quad \text{with } \sigma\ge-\varepsilon.$$

In addition, by Lemma 2.1 we can insure that the ε that we have just selected is such that the boundary value $D(-\varepsilon)=\lim_{\tau\to 0^+}D(-\varepsilon+i\tau)$ exists and is different from zero.

It is well known [6] that the solution $u(t)$ of (1.1) can be expressed by the complex inversion formula for Laplace transforms as

$$u(t)=\lim_{T\to\infty}\frac{1}{2\pi i}\int_{c-iT}^{c+iT}e^{st}D^{-1}(s)\,ds, \qquad t>0,$$

where c is a suitably large positive constant. For each fixed ρ, $0<\rho<\delta$, we can use (2.9) and (2.10) together with Cauchy's theorem to deform the above line integral, and write

(2.11) $$u(t)=V_\rho(t)+H_\rho(t)+C_\rho(t), \qquad t>0,$$

where V_ρ, H_ρ and C_ρ are defined by

(2.12) $$V_\rho(t)=\lim_{T\to\infty}\frac{1}{2\pi}\left[\int_{-T}^{-\rho}+\int_\rho^T\right]e^{(-\varepsilon+i\tau)t}D^{-1}(-\varepsilon+i\tau)\,d\tau,$$

(2.13) $$H_\rho(t)=\frac{1}{2\pi i}\int_{-\varepsilon}^{0}\left[e^{(\sigma-i\rho)t}D^{-1}(\sigma-i\rho)-e^{(\sigma+i\rho)t}D^{-1}(\sigma+i\rho)\right]d\sigma,$$

(2.14) $$C_\rho(t)=\frac{\rho}{2\pi}\int_{-\pi/2}^{\pi/2}\exp(t\rho e^{i\theta})D^{-1}(\rho e^{i\theta})e^{i\theta}\,d\theta.$$

We first show that for fixed $t>0$,

(2.15) $$C_\rho(t)\to 0 \quad \text{as } \rho\to 0^+.$$

To see this note that $D(s)\to\hat{a}(0)>0$ as $s\to 0$ in $\mathrm{Re}\,s\ge 0$ when $a\in L^1(0,\infty)$, while $D^{-1}(s)\to 0$ as $s\to 0$ in $\mathrm{Re}\,s\ge 0$ when $a\notin L^1(0,\infty)$ by [6, Cor. 3.2]. In either case elementary estimates then give $|C_\rho(t)|=O(\rho e^{\rho t})$ as $\rho\to 0^+$, and (2.15) holds.

Next, we consider $V_\rho(t)$. Differentiating (2.1) yields [16, p. 328]

$$\hat{a}'(s)=-\int_0^\infty\frac{d\alpha(x)}{(s+x)^2}, \qquad s\in\mathbb{C}',$$

and dominated convergence can then be used as before to deduce that $|\hat{a}'(\sigma+i\tau)|\to 0$ as $|\tau|\to\infty$ uniformly in any half plane $-\infty<\sigma_0\le\sigma<\infty$. Since $D(\bar{s})=\overline{D(s)}$, we can rewrite (2.12) as

$$V_\rho(t)=\frac{e^{-\varepsilon t}}{\pi}\lim_{T\to\infty}\int_\rho^T\mathrm{Re}\{e^{i\tau t}D^{-1}(-\varepsilon+i\tau)\}\,d\tau.$$

An integration by parts now yields

$$V_\rho(t)=\frac{e^{-\varepsilon t}}{\pi}\left\{\int_\rho^1 \mathrm{Re}\left[e^{i\tau t}D^{-1}(-\varepsilon+i\tau)\right]d\tau \right.$$
$$\left. +\frac{1}{t}\mathrm{Re}\left[ie^{it}D^{-1}(-\varepsilon+i)+\int_1^\infty e^{i\tau t}\frac{D'(-\varepsilon+i\tau)}{D^2(-\varepsilon+i\tau)}d\tau\right]\right\},$$

where the second integral is absolutely convergent. Since, by our choice of ε, $D(-\varepsilon+i\tau)\to D(-\varepsilon)\neq 0$ as $\tau\to 0^+$, we can let $\rho\to 0^+$ in the above formula to get

$$\lim_{\rho\to 0^+} V_\rho(t)=u_1(t), \qquad t>0, \tag{2.16}$$

where

$$u_1(t)\equiv\frac{e^{-\varepsilon t}}{\pi}\left\{\int_0^1 \mathrm{Re}\left[e^{i\tau t}D^{-1}(-\varepsilon+i\tau)\right]d\tau \right.$$
$$\left. +\frac{1}{t}\mathrm{Re}\left[ie^{it}D^{-1}(-\varepsilon+i)+\int_1^\infty e^{i\tau t}\frac{D'(-\varepsilon+i\tau)}{D^2(-\varepsilon+i\tau)}d\tau\right]\right\}. \tag{2.17}$$

Clearly $e^{\varepsilon t}u_1(t)=O(1)(t\to\infty)$.

Now we use the term $H_\rho(t)$ to construct the measure μ. Since $D(\bar{s})=\overline{D(s)}$, (2.13) can be rewritten as

$$H_\rho(t)=-\frac{1}{\pi}\int_{-\varepsilon}^0 e^{\sigma t}\left\{\cos\rho t\,\frac{I(\sigma-i\rho)}{|D^2(\sigma-i\rho)|}+\sin\rho t\,\frac{R(\sigma-i\rho)}{|D^2(\sigma-i\rho)|}\right\}d\sigma. \tag{2.18}$$

Next we show that for $t>0$

$$\lim_{\rho\to 0^+}\int_{-\varepsilon}^0 e^{\sigma t}\sin\rho t\,\frac{R(\sigma-i\rho)}{|D^2(\sigma-i\rho)|}d\sigma=0. \tag{2.19}$$

To see this note that for small enough $\rho>0$, and $-\varepsilon\le\sigma\le 0$,

$$\left|e^{\sigma t}\sin\rho t\,\frac{R(\sigma-i\rho)}{D^2(\sigma-i\rho)}\right|\le\frac{\rho t}{|D(\sigma-i\rho)|}\le t$$

where the second inequality follows from (2.7). By Lemma 2.1, $D(\sigma-i\rho)\to\overline{D(\sigma)}\neq 0$ as $\rho\to 0^+$ for a.e. $\sigma\in[-\varepsilon,0]$, and (2.19) is a consequence of the dominated convergence theorem.

Let $\{\rho_k\}$ be any sequence with $0<\rho_k<\delta$ so that $\rho_k\to 0$ as $k\to\infty$. Set $t=1$, and use (2.11), (2.15), (2.16), (2.18), and (2.19) to deduce that

$$\left\{\int_{-\varepsilon}^0\frac{I(\sigma-i\rho_k)\,d\sigma}{|D^2(\sigma-i\rho_k)|}\right\},\quad k=1,2,\cdots, \text{ is a bounded sequence.} \tag{2.20}$$

Define, for $k=1,2,\cdots$,

$$S_k=\left\{\sigma\in[-\varepsilon,0]:\int_0^\infty\frac{d\alpha(x)}{(x+\sigma)^2+\rho_k^2}\ge 2\right\},$$

$S_k'=[-\varepsilon,0]\setminus S_k$, and let $\chi_k(\sigma)$ denote the characteristic function of S_k. By (2.3), (2.4) and Lemma 2.2, $|I(\sigma-i\rho_k)|\le\rho_k$ and $R(\sigma-i\rho_k)\ge B$ whenever $\sigma\in S_k'$; hence

$$\int_{S_k'}\frac{I(\sigma-i\rho_k)}{|D^2(\sigma-i\rho_k)|}d\sigma\to 0,\qquad k\to\infty.$$

Thus, by (2.20), the sequence

$$\left\{\int_{-\varepsilon}^{0}\chi_k(\sigma)\frac{I(\sigma-i\rho_k)}{|D^2(\sigma-i\rho_k)|}d\sigma\right\}$$

is bounded. From (2.4), $I(\sigma-i\rho_k)\ge 0$ when $\sigma\in S_k$, and therefore

$$\left\{d\mu_k(\sigma)\equiv\frac{\chi_k(\sigma)I(\sigma-i\rho_k)}{\pi|D^2(\sigma-i\rho_k)|}d\sigma\right\}\tag{2.21}$$

is a bounded sequence of nonnegative measures on $[-\varepsilon,0]$. By the Helly selection theorem [16, p. 31], there exist a finite nonnegative measure μ on $[-\varepsilon,0]$ and a subsequence $\{\mu_k'\}$ of the sequence of measures $\{\mu_k\}$ such that $\mu_k'\to\mu$ $(k\to\infty)$ in the weak* topology. In particular, for each $t>0$

$$\int_{-\varepsilon}^{0}e^{\sigma t}d\mu_k'(\sigma)\to\int_{-\varepsilon}^{0}e^{\sigma t}d\mu(\sigma)$$

as $k\to\infty$. Combining this with (2.11), (2.15), (2.16), (2.18) and (2.19), we see that

$$u(t)=-\int_{-\varepsilon}^{0}e^{\sigma t}d\mu(\sigma)+u_1(t),\qquad t>0.$$

Since $e^{\varepsilon t}u_1(t)$ is bounded for large t, and since $u(t)$ and $\int_{-\varepsilon}^{0}e^{\sigma t}d\mu(\sigma)$ are continuous on $[0,\infty)$, we can find $K>0$ so that $|u_1(t)|\le Ke^{-\varepsilon t}$ $(t>0)$. This proves Theorem 1.

3. Further remarks.

COROLLARY 3.1. *Under the hypotheses of Theorem* 1,

$$\int_0^\infty|u(t)|dt<\infty.$$

As noted earlier, more general versions of this result are known.

Proof. Write $\int_0^t u(y)dy=w_1(t)+w_2(t)$, where $w_1(t)=\int_0^t u_1(y)dy$. Then $0\ge w_2(t)\downarrow\int_0^\infty(u(t)-u_1(t))dt$ and $|w_1(t)|\le\int_0^t|u_1(y)|dy\le K/\varepsilon$. Integration of (1.1) yields

$$u(t)=1-\int_0^t w_1(t-y)a(y)dy-\int_0^t w_2(t-y)a(y)dy.\tag{3.1}$$

The argument of Levin [11] (as arranged in [8]) shows that $|u(t)|\le 1$ $(0\le t<\infty)$; in particular, $1-u(t)\ge 0$. We also have the elementary estimates

$$\left|\int_0^t w_1(t-y)a(y)dy\right|\le\frac{K}{\varepsilon}\int_0^t a(y)dy\le\frac{2K}{\varepsilon}\int_0^{t/2}a(y)dy,$$

$$-\int_0^t w_2(t-y)a(y)dy\ge -w_2(t/2)\int_0^{t/2}a(y)dy.$$

Inserting these estimates in (3.1), we get

$$-w_2\left(\frac{t}{2}\right)\le\frac{2K}{\varepsilon},$$

so

$$\int_0^\infty|u(t)|dt\le\frac{3K}{\varepsilon}.$$

□

In conclusion, we point out that if the function α of §2 satisfies

$$\int_0^{x_0} \frac{dx}{\alpha'(x)} < \infty$$

for some $x_0 > 0$, then on the interval $-\min\{\varepsilon, x_0\} \le \sigma \le 0$, the measure μ in (1.2) is given by

$$d\mu(\sigma) = \frac{\alpha'(-\sigma)\, d\sigma}{[\sigma + \phi(\sigma)]^2 + [\pi\alpha'(-\sigma)]^2}, \tag{3.2}$$

where $\phi(\sigma) \equiv \lim_{\tau \to 0} \operatorname{Re} \hat{a}(\sigma - i\tau)$ is well defined for a.e. σ by Lemma 2.1.

To prove this, note that by the Schwarz inequality, if $-x_0 \le \sigma \le 0$, $0 < \tau < x_0/2$, and $E(\sigma, \tau) \equiv [0, x_0] \cap [-\sigma - \tau, -\sigma + \tau]$, then

$$\tau^2 \le \left(\int_{E(\sigma,\tau)} dx \right)^2 \le \int_{E(\sigma,\tau)} \alpha'(x)\, dx \int_{E(\sigma,\tau)} \frac{dx}{\alpha'(x)}.$$

Thus, by (2.1)

$$\frac{\operatorname{Im} \hat{a}(\sigma - i\tau)}{\tau} \ge \frac{1}{2\tau^2} \int_{E(\sigma,\tau)} \alpha'(x)\, dx \ge \left[2 \int_{E(\sigma,\tau)} \frac{dx}{\alpha'(x)} \right]^{-1} \to \infty \tag{3.3}$$

as $\tau \to 0$, uniformly in $-x_0 \le \sigma \le 0$.

Let

$$B_\tau(\sigma) = \frac{\tau}{\pi} \int_0^{x_0} \frac{\dfrac{1}{\alpha'(x)}\, dx}{(\sigma + x)^2 + \tau^2}.$$

By the Schwarz inequality,

$$\operatorname{Im} \hat{a}(\sigma - i\tau)\pi B_\tau(\sigma) \ge \left[\tau \int_0^{x_0} \frac{dx}{(\sigma + x)^2 + \tau^2} \right]^2 \ge \left[\tau \int_0^{x_0/2} \frac{dy}{y^2 + \tau^2} \right]^2 > 1 \tag{3.4}$$

if $\tau > 0$ is small. But $B_\tau(-\sigma)$ is the Poisson integral of χ/α' ($\chi \equiv$ characteristic function of $[0, x_0]$), so $B_\tau(\sigma) \to 1/\alpha'(-\sigma)$ ($\tau \to 0^+$) a.e. on $(-x_0, 0)$ and in $L^1(-x_0, 0)$. Thus, by (3.3) and (3.4), one can use the Vitali convergence theorem [2, p. 150] to show that the functions

$$\frac{I(\sigma - i\rho)}{\pi |D^2(\sigma - i\rho)|} \to \frac{\alpha'(-\sigma)}{[\sigma + \phi(\sigma)]^2 + [\pi\alpha'(-\sigma)]^2}$$

($\rho \to 0^+$) in $L^1(-x_0, 0)$. Therefore, on $[-\min\{x_0, \varepsilon\}, 0]$ the measures $d\mu_k$ of (2.21) converge to the absolutely continuous measure given by (3.2).

REFERENCES

[1] R. W. Carr and K. B. Hannsgen, *Resolvent formulas for a Volterra equation in Hilbert space*, this Journal 13 (1982), pp. 459–483.

[2] N. Dunford and J. T. Schwartz, *Linear Operators, Part I*, Interscience, New York, 1958.

[3] A. Friedman, *On integral equations of Volterra type*, J. d'Analyse Math., 11 (1963), pp. 381–413.

[4] G. Gripenberg, *Decay estimates for resolvents of Volterra equations*, J. Math. Anal. Appl., 85 (1982), pp. 473–487.

[5] S. I. GROSSMAN AND R. K. MILLER, *Nonlinear Volterra integrodifferential systems with L^1 kernels*, J. Differential Equations, 13 (1973), pp. 551–566.

[6] K. B. HANNSGEN, *Indirect Abelian theorems and a linear Volterra equation*, Trans. Amer. Math. Soc., 142 (1969), pp. 539–555.

[7] _____, *A Volterra equation with completely monotonic convolution kernel*, J. Math. Anal. Appl., 31 (1970), pp. 459–471.

[8] _____, *A Volterra equation with parameter*, this Journal, 4 (1973), pp. 22–30.

[9] G. S. JORDAN AND R. L. WHEELER, *Rates of decay of resolvents of Volterra equations with certain nonintegrable kernels*, J. Integral Equations, 2 (1980), pp. 103–110.

[10] G. S. JORDAN, O. J. STAFFANS, AND R. L. WHEELER, *Local analyticity in weighted L^1-spaces and applications to stability problems for Volterra equations*, Trans. Amer. Math. Soc., to appear.

[11] J. J. LEVIN, *The asymptotic behavior of the solution of a Volterra equation*, Proc. Amer. Math. Soc., 14 (1963), pp. 534–541.

[12] J. J. LEVIN AND J. A. NOHEL, *On a system of integrodifferential equations occurring in reactor dynamics*, J. Math. Mech., 9 (1960), pp. 347–368.

[13] R. K. MILLER, *On Volterra integral equations with nonnegative integrable resolvents*, J. Math. Anal. Appl., 22 (1968), pp. 319–340.

[14] G. E. H. REUTER, *Über eine Volterrasche Integralgleichung mit totalmonotonem Kern*, Archiv der Mathematik, 7 (1956), pp. 59–66.

[15] D. F. SHEA AND S. WAINGER, *Variants of the Wiener-Lévy theorem, with applications to stability problems for some Volterra integral equations*, Amer. J. Math., 97 (1975), pp. 312–343.

[16] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1946.

[17] J. S. W. WONG AND R. WONG, *Asymptotic solutions of linear Volterra integral equations with singular kernels*, Trans. Amer. Math. Soc., 189 (1974), pp. 185–200.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0006 $01.00/0

RUNGE'S THEOREM AND FAR FIELD PATTERNS FOR THE IMPEDANCE BOUNDARY VALUE PROBLEM IN ACOUSTIC WAVE PROPAGATION*

DAVID COLTON†

Abstract. Let H_n denote a Hankel function of the first kind, J_n a Bessel function and suppose $0<k<\infty$, $0<\operatorname{Im}\lambda>\infty$, where λ and k are constants. Let D be a bounded domain with smooth boundary ∂D and let ν be the unit outward normal to ∂D. Then it is shown that the sets

$$\left(\frac{\partial}{\partial \nu}+\lambda\right)H_n(kr)\cos n\theta, \qquad \left(\frac{\partial}{\partial \nu}+\lambda\right)H_n(kr)\sin n\theta$$

and

$$\left(\frac{\partial}{\partial \nu}+\lambda\right)J_n(kr)\cos n\theta, \qquad \left(\frac{\partial}{\partial \nu}+\lambda\right)J_n(kr)\sin n\theta,$$

$n=0,1,2,3,\cdots$, are complete in $L^2(\partial D)$. These results are then used to show that, in contrast to the limiting cases $\lambda=0$ and $\lambda=\infty$, the class of far field patterns corresponding to the scattering of entire incident fields by a bounded obstacle having surface impedance λ are dense in $L^2[0,2\pi]$ for all values of the wave number k.

1. Introduction. A basic difficulty in studying the inverse scattering problem for acoustic waves is to obtain information on the class of far field patterns corresponding to boundary conditions of a given type. The first step in such an investigation is to construct the mapping taking given boundary data onto the corresponding far field pattern, and the usual approach in this regard is through the use of integral equations (cf. [2]). However, such an approach is not always the optimal choice for obtaining the desired information. In particular, we shall be concerned in this paper with the method of least squares for solving time harmonic scattering problems, and our aim is to first present a contribution to the existing theoretical basis for this method and then to apply our results to the problem of classifying those functions which can be far field patterns corresponding to a given physically realizable scattering problem.

To be more precise as to our aims, we first formulate the acoustic scattering problem that we are concerned with in this paper. We shall restrict ourselves to two-dimensional scattering problems, although all of our results can be immediately generalized to the case of three dimensions. Let D be a bounded simply connected domain in the plane with smooth boundary ∂D and let ν denote the unit outward normal to D. Then we want to construct (or approximate) a solution $u\in C^2(\mathbb{R}^2\backslash\bar{D})\cap C^1(\mathbb{R}^2\backslash D)$ of the Helmholtz equation

$$\Delta_2 u+k^2u=0 \quad \text{in } \mathbb{R}^2\backslash\bar{D} \tag{1.1a}$$

such that u satisfies the impedance boundary condition

$$\frac{\partial u}{\partial \nu}+\lambda u=0 \quad \text{on } \partial D \tag{1.1b}$$

and u is of the form

$$u=u^i+u^s, \tag{1.1c}$$

* Received by the editors July 28, 1981. This research was supported in part by the National Science Foundation under grant MCS78-02452, and the Air Force Office of Scientific Research under grant 81-0103.

† Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19711.

where u^i (the incident field) is a given entire solution of (1.1a) and u^s (the scattered field) is a solution (to be determined) of (1.1a) in the exterior of D satisfying the Sommerfield radiation condition

$$\lim_{r\to\infty}\sqrt{r}\left(\frac{\partial u^s}{\partial r}-iku^s\right)=0 \tag{1.1d}$$

uniformly with respect to θ, where (r,θ) are polar coordinates centered at an origin contained in D. In (1.1a) the wave number k is assumed constant, finite and positive, and in (1.1b) the surface impedance is assumed to be a constant satisfying $0\leq \operatorname{Im}\lambda<\infty$. Under these conditions there exists a unique solution to (1.1) [2].

The method of least squares for approximately the solution of (1.1) proceeds as follows (cf. [4], [5]). Separate variables in (1.1a) to obtain the set of solutions

$$\begin{matrix} H_n(kr)\cos n\theta, \\ H_n(kr)\sin n\theta, \end{matrix} \qquad n=0,1,2,\cdots, \tag{1.2}$$

where H_n denotes a Hankel function of the first kind. Then each member of the set (1.2) satisfies the radiation condition (1.1d) and we look for an approximate solution to (1.1) by choosing an integer N and constants a_n and b_n, $n=0,1,2,\cdots,N$, such that

$$\left\|\frac{\partial u_N^s}{\partial \nu}+\lambda u_N^s+\frac{\partial u^i}{\partial \nu}+\lambda u^i\right\|_{L^2(\partial D)} \tag{1.3}$$

is minimized, where

$$u_N^s=\sum_{n=0}^{N} H_n(kr)[a_n\cos n\theta+b_n\sin n\theta]. \tag{1.4}$$

The constant N is chosen large enough in order to make (1.3) smaller than an a priori determined value. The attraction of the above procedure lies in its inherent simplicity and, from a theoretical viewpoint, is a valid algorithm provided the following two facts are true:

1) The scattered field u^s depends continuously with respect to the maximum norm on the impedance boundary data measured with respect to the L^2-norm over ∂D, and

2) The set of functions

$$\left(\frac{\partial}{\partial \nu}+\lambda\right)H_n(kr)\cos n\theta, \qquad \left(\frac{\partial}{\partial \nu}+\lambda\right)H_n(kr)\sin n\theta \tag{1.5}$$

is complete with respect to the L^2-norm over ∂D.

The first condition follows from the solution of (1.1) by the method of integral equations (cf. [2], [5]). However, to the author's knowledge the second condition has only been established for the limiting cases of Dirichlet boundary data (i.e., $\lambda=\infty$, cf. [1], [3], [4], [5], [7]) and Neumann boundary data (i.e., $\lambda=0$, cf. [4], [5]. The proof in [4] is incomplete—the integral equation C2 in [4] can have nontrivial solutions if k^2 is an eigenvalue of the interior Dirichlet problem). Hence, in the next section of this paper, we shall establish the validity of this second condition for finite values of the surface impedance λ. We shall also show that the set of functions

$$\begin{matrix} \left(\frac{\partial}{\partial \nu}+\lambda\right)J_n(kr)\cos n\theta, \\ \left(\frac{\partial}{\partial \nu}+\lambda\right)J_n(kr)\sin n\theta, \end{matrix} \qquad n=0,1,2,\cdots, \tag{1.6}$$

where J_n denotes a Bessel function and $0<\text{Im}\lambda<\infty$, is complete with respect to the L^2-norm over ∂D for any value of the wave number k. This result is not true for the limiting cases $\lambda=0$ and $\lambda=\infty$ (cf. [5]). We refer to the above results as Runge's theorem for the impedance boundary value problem and urge the reader to consult the excellent tutorial paper by Miller [4] for the corresponding results for the case of Dirichlet boundary data. The proofs of our theorems will be based on the method of integral equations where in the case of the set (1.5) we make use of the recent approach due to Ursell [6] which avoids the problem of eigensolutions to interior boundary value problems. We note that the proofs given in [4], [5] and [7] for the limiting cases $\lambda=0$ and $\lambda=\infty$ do not appear to have an immediate generalization to the present case where $0<\text{Im}\lambda<\infty$.

With the above results at our disposal, we shall then examine the far field patterns of solutions to (1.1). The far field pattern is defined in terms of the asymptotic behavior of the scattered field u^s. In particular, it can be shown [1], [2], that

$$u^s(r,\theta)=\frac{e^{ikr}}{\sqrt{r}}F(\theta;k)+O(r^{-3/2}), \tag{1.7}$$

where for each fixed k, F is an entire function of θ of exponential type. F is known as the far field pattern of u^s, and the question arises as to whether or not the class of far field patterns for a fixed domain D, wave number k and surface impedance λ, but arbitrary entire incident field u^i, is dense in $L^2[0,2\pi]$. We shall show by an example that this is not true in general for the limiting cases $\lambda=0$ and $\lambda=\infty$ but using the above described approximation theorems is true for $0<\text{Im}\lambda<\infty$. We note that such a result only represents a beginning in the difficult problem of classifying far field patterns. A possible next step in this line of inquiry is to determine the compact subset in $L^2[0,2\pi]$ that corresponds to a fixed incident field and wave number but with surface impedance lying in a given compact subset of the complex plane. As opposed to the problem considered in this paper, this problem is complicated by the fact that it is nonlinear.

2. Runge's theorem for impedance boundary value problems. We begin by showing the set of functions (1.6) with $0<\text{Im}\lambda<\infty$ is complete in $L^2(\partial D)$. In order to do this, we need the following well-known lemma which we include here for the convenience of the reader.

LEMMA 1. *Let $u\in C^2(D)\cap C^1(\bar{D})$ be a solution of the Helmholtz equation in the interior of D such that the boundary condition* (1.1b) *is valid where $0<\text{Im}\lambda<\infty$. Then u is identically zero.*

Proof. Using Green's theorem and the boundary condition (1.1b), we have

$$0=\int_{\partial D}\left(u\frac{\partial\bar{u}}{\partial\nu}-\bar{u}\frac{\partial u}{\partial\nu}\right)ds=2i\,\text{Im}\lambda\int_{\partial D}|u|^2ds. \tag{2.1}$$

Hence $u=0$ on ∂D and the boundary condition (1.1b) implies that $\partial u/\partial\nu=0$ on ∂D. It now follows from Holmgren's uniqueness theorem [1] that u is identically zero in D.

We can now prove the following theorem:

THEOREM 1. *Let $0<\text{Im}\lambda<\infty$. Then the set of functions given in* (1.6) *is complete in $L^2(\partial D)$.*

Proof. Since the set of continuous functions is dense in $L^2(\partial D)$ it suffices to show that if $g \in C(\partial D)$ and

$$\begin{aligned} &\int_{\partial D} g(r,\theta)\left(\frac{\partial}{\partial \nu}+\lambda\right) J_n(kr)\cos n\theta\, ds = 0, \\ &\int_{\partial D} g(r,\theta)\left(\frac{\partial}{\partial \nu}+\lambda\right) J_n(kr)\sin n\theta\, ds = 0 \end{aligned} \tag{2.2}$$

for $n=0,1,2,\cdots$, then g is identically zero on ∂D. Suppose (2.2) is true for some g that is not identically zero. Let Ω be a disk centered at the origin and containing D in its interior and let $\mathbf{x} \in \mathbb{R}^2 \backslash \Omega$ and $\boldsymbol{\xi} \in \partial D$. Then from the expansion (valid for $r_\xi < r_x$)

$$H_0(k|\mathbf{x}-\boldsymbol{\xi}|) = H_0(kr_x)J_0(kr_\xi) + 2\sum_{n=1}^{\infty} H_n(kr_x)J_n(kr_\xi)\cos n(\theta_x-\theta_\xi), \tag{2.3}$$

where (r_x,θ_x) and (r_ξ,θ_ξ) are the polar coordinates of $\mathbf{x}$ and $\boldsymbol{\xi}$, respectively, we can conclude from (2.1) that

$$u(\mathbf{x}) = \frac{\pi i}{2}\int_{\partial D} g(\boldsymbol{\xi})\left(\frac{\partial}{\partial \nu_\xi}+\lambda\right) H_0(k|\mathbf{x}-\boldsymbol{\xi}|)\, ds_\xi = 0 \tag{2.4}$$

for $\mathbf{x} \in R^2 \backslash \Omega$. Since u, as defined by (2.3), is a solution of the Helmholtz equation in the exterior of D, we can conclude by the analyticity of solutions to the Helmholtz equation that $u(\mathbf{x})=0$ for $\mathbf{x} \in \mathbb{R}^2 \backslash \overline{D}$. Letting $\mathbf{x}$ tend to ∂D and using the well-known continuity properties of single and double layer potentials, we can now conclude that g is a solution of the Fredholm integral equation

$$0 = g(\mathbf{x}) + \frac{i}{2}\int_{\partial D} g(\boldsymbol{\xi})\left(\frac{\partial}{\partial \nu_\xi}+\lambda\right) H_0(k|\mathbf{x}-\boldsymbol{\xi}|)\, ds_\xi \tag{2.5}$$

for $\mathbf{x} \in \partial D$. Hence g is an eigenfunction of the integral equation (2.5), and by the Fredholm alternative there exists a function $\phi \in C(\partial D)$, ϕ not identically zero, such that ϕ is an eigenfunction of the equation adjoint to (2.5) with respect to the dual system $\langle C(\partial D), C(\partial D)\rangle$, i.e.,

$$0 = \phi(\mathbf{x}) + \frac{i}{2}\int_{\partial D} \phi(\boldsymbol{\xi})\left(\frac{\partial}{\partial \nu_x}+\lambda\right) H_0(k|\mathbf{x}-\boldsymbol{\xi}|)\, ds_\xi \tag{2.6}$$

for $\mathbf{x} \in \partial D$. Now for $\mathbf{x} \in \mathbb{R}^2$ define w by

$$w(\mathbf{x}) = \frac{\pi i}{2}\int_{\partial D} \phi(\boldsymbol{\xi}) H_0(k|\mathbf{x}-\boldsymbol{\xi}|)\, ds_\xi. \tag{2.7}$$

Then w is a solution of the Helmholtz equation in the interior and exterior of D, and using the continuity properties of single layer potentials, we can conclude from (2.6) that

$$\left(\frac{\partial w}{\partial \nu}\right)_- + \lambda w = \pi\phi(\mathbf{x}) + \frac{\pi i}{2}\int_{\partial D} \phi(\boldsymbol{\xi})\left(\frac{\partial}{\partial \nu_x}+\lambda\right) H_0(k|\mathbf{x}-\boldsymbol{\xi}|)\, ds_\xi \tag{2.8}$$

where the minus sign denotes the limit of $(\partial w/\partial \nu)(\mathbf{x})$ as $\mathbf{x}$ tends to ∂D from inside D. Hence, from Lemma 1 we can now conclude that w is identically zero in D. Since w as defined by (2.7) is a continuous function of $\mathbf{x}$ in all of $\mathbb{R}^2$, we now have that w is a solution of the Helmholtz equation in the exterior of D, satisfies the Sommerfield radiation condition and vanishes on ∂D. Hence, from the well-known uniqueness

theorem for radiating solutions of the exterior Dirichlet problem for the Helmholtz equation, we have that w is identically zero in the exterior of D. But from (2.7) and the jump discontinuity property of the normal derivative of the single layer potential, we now have that

$$0=\left(\frac{\partial w}{\partial \nu}\right)_- - \left(\frac{\partial w}{\partial \nu}\right)_+ = 2\pi\phi(\mathbf{x}) \tag{2.9}$$

for $\mathbf{x}\in\partial D$, where the plus sign denotes the limit of $(\partial w/\partial \nu)(\mathbf{x})$ as $\mathbf{x}$ tends to ∂D from outside D. Hence ϕ is identically zero, which is a contradiction. Hence the only solution of (2.1) is g identically zero, and the theorem is now proved.

We want to prove the analogue of Theorem 1 for the set of functions (1.5). The proof proceeds along similiar lines as that of Theorem 1, except that the roles of interior and exterior domains are reversed, and it is this fact that creates a problem. Indeed, if we construct our auxiliary functions u and w as in (2.4) and (2.7), respectively, we can conclude that w is identically zero in the exterior of D but not in the interior of D if k^2 is an eigenvalue of the interior Dirichlet problem. We shall avoid this difficulty by following the work of Ursell [6] and replacing the fundamental solution appearing in (2.4) and (2.7) by one which satisfies a dissipative boundary condition on a circle lying inside D.

THEOREM 2. *Let* $0\leq \operatorname{Im}\lambda<\infty$. *Then the set of functions given in* (1.5) *is complete in* $L^2(\partial D)$.

Proof. It again suffices to show that if $g\in C(\partial D)$ and

$$\begin{aligned}\int_{\partial D} g(r,\theta)\left(\frac{\partial}{\partial \nu}+\lambda\right)H_n(kr)\cos n\theta\, ds&=0,\\ \int_{\partial D} g(r,\theta)\left(\frac{\partial}{\partial \nu}+\lambda\right)H_n(kr)\sin n\theta\, ds&=0\end{aligned} \tag{2.10}$$

for $n=0,1,2,\cdots$, then g is identically zero on ∂D. Suppose on the contrary that (2.10) is valid for some g that is not identically zero. Let Ω be a disk of radius a centered at the origin and contained in the interior of D and let $\mathbf{x}\in D\backslash\Omega$, $\boldsymbol{\xi}\in\partial D$. Let G be a fundamental solution of the form

$$G(\mathbf{x},\boldsymbol{\xi})=H_0\big(k|\mathbf{x}-\boldsymbol{\xi}|\big)+\Gamma(\mathbf{x},\boldsymbol{\xi}), \tag{2.11}$$

where Γ is a regular solution of the Helmholtz equation with respect to both $\mathbf{x}$ and $\boldsymbol{\xi}$ for $|\mathbf{x}|\,|\boldsymbol{\xi}|>a^2$, satisfies the Sommerfield radiation condition, and is such that G satisfies the dissipative boundary condition

$$\left(\frac{\partial}{\partial \nu_x}+\gamma\right)G(\mathbf{x},\boldsymbol{\xi})=0 \tag{2.12}$$

on $|\mathbf{x}|=a$, where γ is a fixed constant satisfying $0<\operatorname{Im}\gamma<\infty$. Such a fundamental solution was explicitly constructed by Ursell in [6] and can be written in the form

$$G(\mathbf{x},\boldsymbol{\xi})=H_0\big(k|\mathbf{x}-\boldsymbol{\xi}|\big)+a_0H_0(kr_x)+2\sum_{n=1}^{\infty}a_nH_n(kr_x)\cos n\big(\theta_x-\theta_\xi\big), \tag{2.13}$$

where the constants a_n, $n=0,1,2,\cdots$, are in fact functions of $\boldsymbol{\xi}$ and are defined by

$$a_n=H_n(kr_\xi)\frac{\big[kJ_n'(ka)+\gamma J_n(ka)\big]}{\big[kH_n'(ka)+\gamma H_n(ka)\big]}. \tag{2.14}$$

Then from (2.10), (2.13) and the expansion (2.3) (with the roles of $\boldsymbol{\xi}$ and $\mathbf{x}$ interchanged) we have as in Theorem 1 that

$$u(\mathbf{x})=\frac{\pi i}{2}\int_{\partial D} g(\boldsymbol{\xi})\left(\frac{\partial}{\partial \nu_{\boldsymbol{\xi}}}+\lambda\right)G(\mathbf{x},\boldsymbol{\xi})\,ds_{\xi}=0 \tag{2.15}$$

for $\mathbf{x}\in D\backslash\Omega$, and g is a solution of the Fredholm integral equation

$$0=g(\mathbf{x})-\frac{i}{2}\int_{\partial D} g(\boldsymbol{\xi})\left(\frac{\partial}{\partial \nu_{\boldsymbol{\xi}}}+\lambda\right)G(\mathbf{x},\boldsymbol{\xi})\,ds_{\xi} \tag{2.16}$$

for $\mathbf{x}\in\partial D$. Hence there exists a function $\phi\in C(\partial D)$, ϕ not identically zero, such that

$$0=\phi(\mathbf{x})-\frac{i}{2}\int_{\partial D}\phi(\xi)\left(\frac{\partial}{\partial \nu_{x}}+\lambda\right)G(\mathbf{x},\boldsymbol{\xi})\,ds_{\xi} \tag{2.17}$$

for $\mathbf{x}\in\partial D$. Again following the proof of Theorem 1, we define for $\mathbf{x}\in\mathbb{R}^2\backslash\Omega$,

$$w(\mathbf{x})=\frac{\pi i}{2}\int_{\partial D}\phi(\boldsymbol{\xi})G(\mathbf{x},\boldsymbol{\xi})\,ds_{\xi} \tag{2.18}$$

and conclude that

$$\left(\frac{\partial w}{\partial \nu}\right)_{+}+\lambda w=0. \tag{2.19}$$

It now follows from the fact that w satisfies the Sommerfield radiation condition that w is identically zero in $\mathbb{R}^2\backslash\overline{D}$. Hence, using the continuity of w for $\mathbf{x}\in\mathbb{R}^2\backslash\Omega$, we have that w is a solution of the Holmholtz equation in $D\backslash\Omega$, vanishes on ∂D and satisfies

$$\left(\frac{\partial}{\partial r}+\gamma\right)w=0 \tag{2.20}$$

on $|\mathbf{x}|=a$. It now follows from a slight modification of Lemma 1 (cf. [6]) that w is identically zero in $D\backslash\Omega$. Hence, we can now conclude as we did in Theorem 1 that ϕ is identically zero, and this is a contradiction. Hence, the only solution of (2.10) is g identically equal to zero, and the theorem is proved.

3. Far field patterns of solutions satisfying impedance boundary data. We now address ourselves to the problem of showing that the set of far field pattern corresponding to solutions of (1.1) are dense in $L^2[0,2\pi]$ for any fixed values of the wave number k and surface impedance λ, $0<\operatorname{Im}\lambda<\infty$. We shall first give an example showing that this is in general not the case if $\lambda=0$, in particular if k^2 is an eigenvalue of the interior Neumann problem. A similar example can also be constructed for the limiting case $\lambda=\infty$ (i.e., Dirichlet boundary data).

Example. Consider problem (1.1), when D is the unit disk and $\lambda=0$. Then, since u^i is an entire solution of the Helmholtz equation, we can expand u^i in the form

$$u^i(r,\theta)=\sum_{n=0}^{\infty} J_n(kr)[a_n\cos n\theta+b_n\sin n\theta], \tag{3.1}$$

where the series (3.1) is uniformly convergent on any compact subset of $\mathbb{R}^2$. Then for $r\geq 1$ we can expand u^s in the uniformly convergent series

$$u^s(r,\theta)=-\sum_{n=0}^{\infty} H_n(kr)\frac{J_n'(k)}{H_n'(k)}[a_n\cos n\theta+b_n\sin n\theta], \tag{3.2}$$

and from the asymptotic behavior of Hankel's function we have that the far field pattern for u^s is given by

$$F(\theta;k)=e^{-i\pi/4}\sqrt{\frac{2}{\pi k}}\sum_{n=0}^{\infty}\frac{(-i)^n J_n'(k)}{H_n'(k)}[a_n\cos n\theta+b_n\sin n\theta]. \tag{3.3}$$

If k_0^2 is an eigenvalue of the interior Neumann problem, then $J_n'(k_0)=0$ for some integer $n=n_0$, and hence in this case $F(\theta;k_0)$ is orthogonal to $\cos n_0\theta$ and $\sin n_0\theta$ for all incident fields u^i. Hence, the class of far field patterns for such values of k is not dense in $L^2[0,2\pi]$.

We shall now show that, in constrast to the above example, the class of far field patterns corresponding to surface impedance λ satisfying $0<\operatorname{Im}\lambda<\infty$ are dense in $L^2[0,2\pi]$ for all values of the wave number k and any bounded simply connected domain D with smooth boundary ∂D.

THEOREM 3. *Let* $0<\operatorname{Im}\lambda<\infty$. *Then the class of far field patterns of problem* (1.1) *is dense in* $L^2[0,2\pi]$.

Proof. Let $f\in L^2[0,2\pi]$ and $\varepsilon>0$ be given. Then there exists a trigonometric polynomial

$$F(\theta)=\sum_{n=1}^{N}[a_n\cos n\theta+b_n\sin n\theta] \tag{3.4}$$

such that

$$\|f-F\|<\frac{\varepsilon}{2}, \tag{3.5}$$

where $\|\cdot\|$ denotes the norm in $L^2[0,2\pi]$. Then

$$u^s(r,\theta)=e^{i\pi/4}\sqrt{\frac{\pi k}{2}}\sum_{n=0}^{N}(i)^n H_n(kr)[a_n\cos n\theta+b_n\sin n\theta] \tag{3.6}$$

is a radiating solution of the Helmholtz equation having F as its far field pattern. Let $g\in C(\partial D)$ be defined by

$$\frac{\partial u^s}{\partial\nu}+\lambda u^s=g. \tag{3.7}$$

Then from Theorem 1 we have that for every $\delta>0$ there exists a integer $M=M(\delta)$ and constants c_n and d_n, $n=0,1,\cdots,M$, such that

$$u^i(r,\theta)=\sum_{n=0}^{M}J_n(kr)[c_n\cos n\theta+d_n\sin n\theta] \tag{3.8}$$

satisfies

$$\left\|\frac{\partial u^i}{\partial\nu}+\lambda u^i+g\right\|_{L^2(\partial D)}<\delta. \tag{3.9}$$

Now let $\tilde{u}=\tilde{u}^s+u^i$ be the solution of (1.1) corresponding to the incident field u^i given by (3.8) and let $\tilde{F}$ be the far field pattern of $\tilde{u}^s$. From the method of integral equations it can be seen that the mapping of the boundary data g to the far field pattern F is a continuous mapping from $L^2(\partial D)$ into $L^2[0,2\pi]$. Hence, if δ is chosen sufficiently small, we have

$$\|\tilde{F}-F\|<\frac{\varepsilon}{2}, \tag{3.10}$$

and from (3.5) and (3.10), we now have

$$\|f-\tilde{F}\|<\varepsilon. \tag{3.11}$$

Since f and ε are arbitrary the theorem is now proved.

Acknowledgment. The author would like to thank Professor Fritz Ursell for helpful discussions concerning this paper.

REFERENCES

[1] D. COLTON, *Analytic Theory of Partial Differential Equations*, Pitman, Boston, 1980.

[2] D. COLTON AND R. KRESS, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, to appear.

[3] A. HIZAL, *Scattering from perfect conductors and layered dielectrics using both incoming and outgoing wave functions*, Acoustic, Electromagnetic and Elastic Wave Scattering—Focus on the *T*-Matrix Approach, V. K. Varadan and V. V. Varadan, eds., Pergamon Press, New York, 1980, pp. 169–190.

[4] R. F. MILLER, *The Rayleigh hypothesis and a related least squares solution to scattering problems for periodic surfaces and other scatterers*, Radio Science, 8 (1973), pp. 785–796.

[5] C. MÜLLER AND H. KERSTEN, *Zwei Klassen vollständiger Funktionensysteme zur Behandlung der Randwertaufgaben der Schwingungsgleichung* $\Delta u+k^2u=0$, Math. Meth. in the Appl. Sci., 2 (1980), pp. 48–67.

[6] F. URSELL, *On the exterior problems of acoustics*, Proc. Camb. Phil. Soc., 74 (1973), pp. 117–125.

[7] I. N. VEKUA, *On completeness of a system of metaharmonic functions*, Dokl. Akad. Nauk, SSSR, 90 (1953), pp. 715–718. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0007 $01.00/0

INTEGRAL RELATIONS FOR LAMÉ FUNCTIONS*

HANS VOLKMER†

Abstract. In this paper we introduce a new general integral relation for Lamé functions. We show that the known linear integral equations for periodic Lamé functions are special cases of this new relation. We are able to determine the characteristic values of the integral equations for Lamé functions using our relation. We pay special attention to integral relations for Lamé polynomials.

Introduction. Integral relations for Lamé functions were discovered by Whittaker [15], [16] in 1915 and have since been investigated by Ince [7], [8], [9], Erdélyi [4], Arscott [1], [2], [3], Shail [11] and others. These integral relations are generated by the usual method which depends on the simultaneous separability of the wave equation in different orthogonal coordinate systems. A systematic application of this method was presented by Schmidt and Wolf [10].

In [13] and [14] I introduced a new method to generate integral relations with variable boundaries for several classes of special functions. This method is closely connected with Vekua's theory [12] of solving elliptic equations by integral operators. We also refer the reader to the paper of Henrici [6] who applied such integral operators to the theory of special functions.

In this paper we compare the new relation in the special case of the Lamé functions with the known integral equations. The main results are:

1) The known integral equations of Erdélyi, Magnus, Oberhettinger, Tricomi [5, 15, 5.3] and Whittaker, Watson [17, 23.6 and 23.6.1] are special cases of this *one* new relation.

2) Using the new relation we are able to express the characteristic values of the integral equations for periodic Lamé functions by values of any second solution of the Lamé equation. This is not possible with the usual method to generate these integral equations.

3) The new relation is valid for any Lamé function (i.e., any solution of Lamé's equation) in contrast to all known relations which are valid only for Lamé functions of special types.

In §1 we shall state the general integral relation with variable boundaries for Lamé functions. In the §§2, 3, 4 we shall discuss this result for periodic Lamé functions and, in §5, for Lamé polynomials.

In this paper we denote by $\mathrm{sn}, \mathrm{cn}, \mathrm{dn}$ the Jacobian functions corresponding to the fixed modulus $k \in]0,1[$. As usual, $k' \in]0,1[$ is the modulus complementary to k, i.e., $k^2+k'^2=1$. $2K$ is the real period of dn and $2iK'$ is the imaginary period of sn. Let $\mathbb{C}(\mathbb{R})$ be the set of complex (real) numbers and let $M=\{2mK+(2n+1)iK' : m,n \text{ integers}\}$ denote the set of poles of the functions $\mathrm{sn}, \mathrm{cn}, \mathrm{dn}$.

1. A general integral relation for Lamé functions. We start with an integral relation which follows easily from [14, (2.30)] by setting $x=\xi$, $y=K+iK'-i\eta$. It is a special form of Riemann's formula for integrating partial differential equations.

THEOREM 1.1. *Let D be a subset of $(\mathbb{C}\setminus M)\times(\mathbb{C}\setminus M)$ which can be written in the form $D=\{(x,y)\in\mathbb{C}^2 : (x-y,x+y)\in G_1\times G_2\}$ with two simply connected domains G_1, G_2 in*

* Received by the editors November 18, 1980, and in final revised form October 16, 1981.
† Fachbereich 6-Mathematik, Universität Essen, GH5, 4300 Essen 1, West Germany.

$\mathbb{C}$. *Let* $w: D \to \mathbb{C}$ *be a solution of the partial differential equation (with a complex parameter* ν*)*

$$\frac{\partial^2 w}{\partial x^2} - \frac{\partial^2 w}{\partial y^2} + \nu(\nu+1)k^2(\mathrm{sn}^2 y - \mathrm{sn}^2 x)w = 0. \tag{1.2}$$

Then we have for three points (x_0,y_0), (x_1,y_1), $(x_2,y_2) \in D$, *with* $x_1 - y_1 = x_0 - y_0$, $x_2 + y_2 = x_0 + y_0$ *and any path* L *in* D *from* (x_1,y_1) *to* (x_2,y_2),

$$w(x_0,y_0) = \frac{1}{2}w(x_1,y_1) + \frac{1}{2}w(x_2,y_2) + \frac{1}{2}\int_L \left(\left(A\frac{\partial w}{\partial y} - w\frac{\partial A}{\partial y} \right) dx + \left(A\frac{\partial w}{\partial x} - w\frac{\partial A}{\partial x} \right) dy \right)$$

where $A: D \times D \to \mathbb{C}$ *is the analytic function defined by*

$$A(x,y,x_0,y_0) = P_\nu\left(k^2 \,\mathrm{sn}\, x \,\mathrm{sn}\, x_0 \,\mathrm{sn}\, y \,\mathrm{sn}\, y_0 - \frac{k^2}{k'^2}\,\mathrm{cn}\, x \,\mathrm{cn}\, x_0 \,\mathrm{cn}\, y \,\mathrm{cn}\, y_0 + \frac{1}{k'^2}\,\mathrm{dn}\, x \,\mathrm{dn}\, x_0 \,\mathrm{dn}\, y \,\mathrm{dn}\, y_0 \right).$$

In the special case that ν is a nonnegative integer, this kernel function appears in [5, 15.7(5)].

With regard to the function A, we remark that P_ν denotes the Legendre function with parameter ν (which is analytic at 1). For (x,y) close to (x_0,y_0), the argument of P_ν is close to 1. The function $A(\cdot,\cdot,x_0,y_0)$, defined locally this way, can then be continued analytically on D. In [13] and [14] we proved that this is possible.

The argument of P_ν may be represented by means of

$$f(z) = k\,\mathrm{cn}\, z + \mathrm{dn}\, z$$

and the cross ratio

$$\mathrm{CR}(z_1,z_2,z_3,z_4) = \frac{z_4 - z_1}{z_4 - z_3} : \frac{z_2 - z_1}{z_2 - z_3}$$

in the form

$$k^2 \,\mathrm{sn}\, x \,\mathrm{sn}\, x_0 \,\mathrm{sn}\, y \,\mathrm{sn}\, y_0 - \frac{k^2}{k'^2}\,\mathrm{cn}\, x \,\mathrm{cn}\, x_0 \,\mathrm{cn}\, y \,\mathrm{cn}\, y_0 + \frac{1}{k'^2}\,\mathrm{dn}\, x \,\mathrm{dn}\, x_0 \,\mathrm{dn}\, y \,\mathrm{dn}\, y_0 = 1 - 2\mathrm{CR}(f(x_0 - y_0), -f(x_0 + y_0), -f(x+y), f(x-y)). \tag{1.3}$$

If u_1 and u_2 are solutions of the Lamé equation (with complex parameters ν, λ),

$$\frac{d^2 u}{dz^2} + (\lambda - \nu(\nu+1)k^2 \mathrm{sn}^2 z)u = 0, \tag{1.4}$$

then $w(x,y) = u_1(x)u_2(y)$ fulfills the partial differential equation (1.2). Therefore, setting in Theorem 1.1 $w(x,y) = u_1(x)u_2(y)$ and $L = L_1 \times s$ with fixed $s \in \mathbb{C}$, we obtain the following general integral relation for Lamé functions.

COROLLARY 1.5. *Let* D *be a domain of the form* $D = \{(x,y) \in \mathbb{C}^2 : (x-y, x+y) \in G_1 \times G_2\}$ *as in Theorem* 1.1 *and let* D_1 *and* D_2 *be domains in* $\mathbb{C} \setminus M$ *such that* $D \subset D_1 \times D_2$. *Let* $u_1: D_1 \to \mathbb{C}$ *and* $u_2: D_2 \to \mathbb{C}$ *be solutions of* (1.4) *(with the same parameters* ν, λ*).*

Then we have, for all $(x_0,y_0)\in D$ *and any path* L_1 *in* $\mathbb{C}$ *from* $x_0-(y_0-s)$ *to* $x_0+(y_0-s)$ $(s\in\mathbb{C})$ *with* $(L_1\times s)\subset D$,

$$u_1(x_0)u_2(y_0)=\frac{1}{2}u_2(s)(u_1(x_0-(y_0-s))+u_1(x_0+(y_0-s)))$$
$$+\frac{1}{2}\int_{L_1}\left(u_2'(s)A(x,s,x_0,y_0)-u_2(s)\frac{\partial A}{\partial y}(x,s,x_0,y_0)\right)u_1(x)\,dx.$$

The kernel A *is the same as in Theorem* 1.1.

The integral relation in Corollary 1.5 has many interesting applications as we shall see in the following sections. For example, with fixed y_0 and s, 1.5 is an integral relation for the Lamé function u_1, because u_2 only appears with the values $u_2(y_0)$, $u_2(s)$, $u_2'(s)$. With fixed x_0 and s, 1.5 is an integral representation of the solution u_2 of the initial value problem for (1.4) at $z=s$ in terms of the Lamé function u_1.

2. Integral relations for Lamé functions of period $2K$ or half-period $2K$. In Corollary 1.5 we set $D_1=D_2=G_1=G_2=\{z\in\mathbb{C}:-K'<\operatorname{Im}z<K'\}$. These domains satisfy the assumptions of the corollary. Moreover, we have $D_1\times\mathbb{R}\subset D$. Now Corollary 1.5 with $s=0$ reads:

THEOREM 2.1. *Let* $u_1,u_2:D_1\to\mathbb{C}$ *be solutions of* (1.4). *Then we have, for all* $(x_0,y_0)\in D$,

$$u_1(x_0)u_2(y_0)=\frac{1}{2}u_2(0)(u_1(x_0-y_0)+u_1(x_0+y_0))$$
$$+\frac{1}{2}\int_{x_0-y_0}^{x_0+y_0}\left(u_2'(0)A(x,0,x_0,y_0)-u_2(0)\frac{\partial A}{\partial y}(x,0,x_0,y_0)\right)u_1(x)\,dx.$$

For the path of integration we can choose any path in D_1 *from* x_0-y_0 *to* x_0+y_0. *The integrand is analytic in* $x\in D_1$.

We make the following remark concerning the kernel A. It is easy to see that for the function $f(z)=k\operatorname{cn}z+\operatorname{dn}z$ we have

$$f(G_1)=f(G_2)=\{z\in\mathbb{C}:\operatorname{Re}z>0\}.$$

If we represent the argument of P_ν as in (1.3) the well-known properties of the cross ratio yield

$$k^2\operatorname{sn}x\operatorname{sn}x_0\operatorname{sn}y\operatorname{sn}y_0-\frac{k^2}{k'^2}\operatorname{cn}x\operatorname{cn}x_0\operatorname{cn}y\operatorname{cn}y_0+\frac{1}{k'^2}\operatorname{dn}x\operatorname{dn}x_0\operatorname{dn}y\operatorname{dn}y_0\notin]-\infty,-1]$$

for $(x,y)\in D$, $(x_0,y_0)\in D$.

Therefore, if we consider the Legendre function P_ν as an analytic function on $\mathbb{C}\setminus]-\infty,-1]$, we can define the function A globally on $D\times D$. Using this observation the kernels of Theorem 2.1 are

$$A(x,0,x_0,y_0)=P_\nu\left(-\frac{k^2}{k'^2}\operatorname{cn}x\operatorname{cn}x_0\operatorname{cn}y_0+\frac{1}{k'^2}\operatorname{dn}x\operatorname{dn}x_0\operatorname{dn}y_0\right),$$
$$\frac{\partial A}{\partial y}(x,0,x_0,y_0)=k^2\operatorname{sn}x\operatorname{sn}x_0\operatorname{sn}y_0P_\nu'(\cdots).$$

For some pairs of parameters ν,λ, there exist (nontrivial) periodic solutions $u_1:D_1\to\mathbb{C}$ of the Lamé equation (1.4).

We consider:

(2.2) even Lamé functions of period $2K$,
(2.3) odd Lamé functions of period $2K$,
(2.4) even Lamé functions of half-period $2K$,
(2.5) odd Lamé functions of half-period $2K$.

Half-period $2K$ means $u_1(x+2K)=-u_1(x)$ $(x\in D_1)$.

For periodic Lamé functions we obtain from Theorem 2.1 integral equations with fixed boundaries. These integral equations are well known (see Erdélyi/Magnus/Oberhettinger [5]), but our approach to these integral equations is new. Moreover, we can determine the characteristic values of the integral equations.

COROLLARY 2.6. *Let* $u_1: D_1\to\mathbb{C}$ *be a solution of* (1.4) *belonging to one of the classes* (2.2)–(2.5) *and let* $u_2: D_1\to\mathbb{C}$ *be any second solution of* (1.4) *which is linearly independent of* u_1. *Then we have for all* $x_0\in D_1$
in case (2.2)

$$u_1(x_0)\,\frac{u_2(K)-u_2(-K)}{u_2'(0)}=\int_{-K}^{K}P_\nu\Big(\frac{1}{k'}\,\mathrm{dn}\,x\,\mathrm{dn}\,x_0\Big)\,u_1(x)\,dx,$$

in case (2.3)

$$u_1(x_0)\,\frac{u_2'(K)-u_2'(-K)}{u_2(0)}=-\frac{k^4}{k'}\int_{-K}^{K}\mathrm{cn}\,x\,\mathrm{cn}\,x_0\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0\,P_\nu''\Big(\frac{1}{k'}\,\mathrm{dn}\,x\,\mathrm{dn}\,x_0\Big)\,u_1(x)\,dx,$$

in case (2.4)

$$u_1(x_0)\,\frac{u_2'(K)+u_2'(-K)}{u_2'(0)}=\frac{k^2}{k'}\int_{-K}^{K}\mathrm{cn}\,x\,\mathrm{cn}\,x_0\,P_\nu'\Big(\frac{1}{k'}\,\mathrm{dn}\,x\,\mathrm{dn}\,x_0\Big)\,u_1(x)\,dx,$$

in case (2.5)

$$u_1(x_0)\,\frac{u_2(K)+u_2(-K)}{u_2(0)}=-k^2\int_{-K}^{K}\mathrm{sn}\,x\,\mathrm{sn}\,x_0\,P_\nu'\Big(\frac{1}{k'}\,\mathrm{dn}\,x\,\mathrm{dn}\,x_0\Big)\,u_1(x)\,dx.$$

Proof. Case (2.2). The value of $(u_2(K)-u_2(-K))/u_2'(0)$ does not depend on the choice of the second solution u_2. Therefore, without loss of generality, we can choose u_2 such that $u_2(0)=0$ and $u_2'(0)=1$. In this case we have $(u_2(K)-u_2(-K))/u_2'(0)=2u_2(K)$. Now we apply Theorem 2.1 with $y_0=K$ and, since $D_1\times\mathbb{R}\subset D$, we obtain for all $x_0\in D_1$

$$u_1(x_0)u_2(K)=\frac{1}{2}\int_{x_0-K}^{x_0+K}P_\nu\Big(\frac{1}{k'}\,\mathrm{dn}\,x\,\mathrm{dn}\,x_0\Big)\,u_1(x)\,dx.$$

Since the integrand is periodic mod $2K$, we can replace $\int_{x_0-K}^{x_0+K}$ by $\int_{-K}^{K}$; thus the equation in case (2.2) is shown.

The proof in case (2.5) is similar.

In case (2.3) we may assume $u_2(0)=1$, $u_2'(0)=0$. Then we differentiate the integral equation in Theorem 2.1 with respect to y_0 and set $y_0=K$. Using the properties of u_1 and u_2 we see that on the right-hand side of the equation only the integral is left. Now this integral may be handled as in case (2.2).

The proof in case (2.4) is similar. □

We point out another interesting application of Theorem 2.1. For an arbitrary pair of parameters ν,λ, the Lamé equation (1.4) has one even and one odd solution (up to constant factor). Further, there are solutions which are even or odd with respect to K.

Using Theorem 2.1 we are able to represent one of these solutions in terms of another such solution. For example if $u_1, u_2: D_1 \to \mathbb{C}$ are the solutions of (1.4) with $u_1(0)=1$, $u_1'(0)=0$ and $u_2(0)=0$, $u_2'(0)=1$, then we have, for all $y_0 \in D_1$,

$$u_2(y_0)=\int_0^{y_0} P_\nu\left(-\frac{k^2}{k'^2}\operatorname{cn}x\operatorname{cn}y_0+\frac{1}{k'^2}\operatorname{dn}x\operatorname{dn}y_0\right)u_1(x)\,dx.$$

This follows immediately from Theorem 2.1 with $x_0=0$. If we choose u_3 such that $u_3(K)=1$, $u_3'(K)=0$ and u_2 as above, then Theorem 2.1 with $x_0=K$ yields

$$u_2(y_0)=\int_K^{K+y_0} P_\nu\left(\frac{1}{k'}\operatorname{dn}x\operatorname{dn}y_0\right)u_3(x)\,dx.$$

We refer to the papers of Arscott [3] and Shail [11], who also obtained representations of a second solution in terms of a first solution. They consider the case that the first solution is a Lamé polynomial and the second solution is the usual F-solution. This F-solution is even or odd with respect to iK' but not with respect to 0 or K.

3. Integral relations for Lamé functions of period $2iK'$ or half-period $2iK'$. In Corollary 1.5 we set

$$D_1=D_2=\{z\in\mathbb{C}: 0<\operatorname{Re}z<2K\},$$
$$G_1=D_1-K,\ G_2=D_1+K.$$

Then we have $D_1\times(K+i\mathbb{R})\subset D$. Again, from Corollary 1.5 we obtain with $s=K$:

THEOREM 3.1. *Let $u_1, u_2: D_1\to\mathbb{C}$ be solutions of* (1.4). *Then we have for all $(x_0,y_0)\in D$*

$$u_1(x_0)u_2(y_0)=\frac{1}{2}u_2(K)\big(u_1(x_0-(y_0-K))+u_1(x_0+(y_0-K))\big)$$
$$+\frac{1}{2}\int_{x_0-(y_0-K)}^{x_0+(y_0-K)}\Big(u_2'(K)A(x,K,x_0,y_0)$$
$$-u_2(K)\frac{\partial A}{\partial y}(x,K,x_0,y_0)\Big)u_1(x)\,dx,$$

where the path of integration is in D_1 and

$$A(x,K,x_0,y_0)=P_\nu\left(k^2\operatorname{sn}x\operatorname{sn}x_0\operatorname{sn}y_0+\frac{1}{k'}\operatorname{dn}x\operatorname{dn}x_0\operatorname{dn}y_0\right),$$
$$\frac{\partial A}{\partial y}(x,K,x_0,y_0)=\frac{k^2}{k'}\operatorname{cn}x\operatorname{cn}x_0\operatorname{cn}y_0P_\nu'(\cdots).$$

Here the function A may be defined globally by the Legendre function $P_\nu: \mathbb{C}\setminus]-\infty,-1]\to\mathbb{C}$ as in §2. This follows from

$$f(G_1)=\{z\in\mathbb{C}:|z|>k'\},\qquad f(G_2)=\{z\in\mathbb{C}:|z|<k'\}.$$

Now we consider

(3.2) Lamé functions of period $2iK'$, even with respect to K,
(3.3) Lamé functions of period $2iK'$, odd with respect to K,
(3.4) Lamé functions of half-period $2iK'$, even with respect to K,
(3.5) Lamé functions of half-period $2iK'$, odd with respect to K.

In the same manner as in §2, we obtain from Theorem 3.1:

COROLLARY 3.6. *Let* $u_1: D_1 \to \mathbb{C}$ *be a solution of* (1.4) *belonging to one of the classes* (3.2)–(3.5) *and let* $u_2: D_1 \to \mathbb{C}$ *be any second solution of* (1.4) *which is linearly independent of* u_1. *Then we have for all* $x_0 \in D_1$
in case (3.2)

$$u_1(x_0)\,\frac{u_2(K+iK')-u_2(K-iK')}{u_2'(K)}=\int_{K-iK'}^{K+iK'} P_\nu(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)u_1(x)\,dx,$$

in case (3.3)

$$u_1(x_0)\,\frac{u_2'(K+iK')-u_2'(K-iK')}{u_2(K)}$$
$$=-k\int_{K-iK'}^{K+iK'} \mathrm{dn}\,x\,\mathrm{dn}\,x_0\,\mathrm{cn}\,x\,\mathrm{cn}\,x_0\,P_\nu''(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)u_1(x)\,dx,$$

in case (3.4)

$$u_1(x_0)\,\frac{u_2'(K+iK')+u_2'(K-iK')}{u_2'(K)}=i\int_{K-iK'}^{K+iK'} \mathrm{dn}\,x\,\mathrm{dn}\,x_0\,P_\nu'(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)u_1(x)\,dx,$$

in case (3.5)

$$u_1(x_0)\,\frac{u_2(K+iK')+u_2(K-iK')}{u_2(K)}=ki\int_{K-iK'}^{K+iK'} \mathrm{cn}\,x\,\mathrm{cn}\,x_0\,P_\nu'(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)u_1(x)\,dx.$$

4. Integral relations for Lamé functions of period $2(K+iK')$ or half-period $2(K+iK')$.

In Corollary 1.5 we set

$$D_1=D_2=G_1=G_2=\left\{z\in\mathbb{C}: -K'<\mathrm{Im}\,z-\frac{K'}{K}\mathrm{Re}\,z<K'\right\}.$$

Then we have $D_1\times(K+iK')\mathbb{R}\subset D$. From Corollary 1.5 we obtain with $s=0$:

THEOREM 4.1. *Let* $u_1, u_2: D_1 \to \mathbb{C}$ *be solutions of* (1.4). *Then we have, for all* $(x_0,y_0)\in D$,

$$u_1(x_0)u_2(y_0)=\frac{1}{2}u_2(0)(u_1(x_0-y_0)+u_1(x_0+y_0))$$
$$+\frac{1}{2}\int_{x_0-y_0}^{x_0+y_0}\left(u_2'(0)A(x,0,x_0,y_0)-u_2(0)\,\frac{\partial A}{\partial y}(x,0,x_0,y_0)\right)u_1(x)\,dx,$$

where the path of integration is in D_1 *and the kernels* $A(x,0,x_0,y_0)$ *and* $\frac{\partial A}{\partial y}(x,0,x_0,y_0)$ *are given in* §2.

However, here the function A cannot be defined globally by the function P_ν: $\mathbb{C}\setminus]-\infty,-1]\to\mathbb{C}$ as in the §§2 and 3. For example, if we choose $x=2(K+iK')$, $x_0=y=y_0=0$ the argument of P_ν (see (1.3)) belongs to the interval $]-\infty,-1[$.

Now we consider

(4.2) even Lamé functions of period $2(K+iK')$,
(4.3) odd Lamé functions of period $2(K+iK')$,
(4.4) even Lamé functions of half-period $2(K+iK')$,
(4.5) odd Lamé functions of half-period $2(K+iK')$.

From Theorem 4.1 we obtain:

COROLLARY 4.6. *Let* $u_1: D_1 \to \mathbb{C}$ *be a solution of* (1.4) *belonging to one of the classes* (4.2)–(4.5) *and let* $u_2: D_1 \to \mathbb{C}$ *be any second solution of* (1.4) *which is linearly independent*

of u_1. *Then we have for all* $x_0 \in D_1$,
in case (4.2)

$$u_1(x_0)\,\frac{u_2(K+iK')-u_2(-K-iK')}{u_2'(0)}=\int_{-K-iK'}^{K+iK'} P_\nu\Big(\frac{ik}{k'}\operatorname{cn}x\operatorname{cn}x_0\Big)u_1(x)\,dx,$$

in case (4.3)

$$u_1(x_0)\,\frac{u_2'(K+iK')-u_2'(-K-iK')}{u_2(0)}$$
$$=-i\,\frac{k}{k'}\int_{-K-iK'}^{K+iK'}\operatorname{sn}x\operatorname{sn}x_0\operatorname{dn}x\operatorname{dn}x_0\,P_\nu''\Big(\frac{ik}{k'}\operatorname{cn}x\operatorname{cn}x_0\Big)u_1(x)\,dx,$$

in case (4.4)

$$u_1(x_0)\,\frac{u_2'(K+iK')+u_2'(-K-iK')}{u_2'(0)}=\frac{i}{k'}\int_{-K-iK'}^{K+iK'}\operatorname{dn}x\operatorname{dn}x_0\,P_\nu'\Big(\frac{ik}{k'}\operatorname{cn}x\operatorname{cn}x_0\Big)u_1(x)\,dx,$$

in case (4.5)

$$u_1(x_0)\,\frac{u_2(K+iK')+u_2(-K-iK')}{u_2(0)}=-k\int_{-K-iK'}^{K+iK'}\operatorname{sn}x\operatorname{sn}x_0\,P_\nu'\Big(\frac{ik}{k'}\operatorname{cn}x\operatorname{cn}x_0\Big)u_1(x)\,dx.$$

Here the kernel $P_\nu((ik/k')\operatorname{cn}x\operatorname{cn}x_0)$ is defined for x, x_0 near the line $(K+iK')\cdot\mathbb{R}$ by the Legendre function $P_\nu\colon \mathbb{C}\setminus]-\infty,-1]\to\mathbb{C}$ and then is continued analytically on $D_1\times D_1$.

5. Integral relations for Lamé polynomials. It is well known that for every non-negative integer n there exist $2n+1$ characteristic values of the parameter λ such that the Lamé equation

$$\frac{d^2u}{dz^2}+\big(\lambda-n(n+1)k^2\operatorname{sn}^2z\big)u=0 \tag{5.1}$$

has a Lamé polynomial as a solution. Lamé polynomials are solutions of (5.1) which are of the form

$$\operatorname{sn}^\rho z\operatorname{cn}^\sigma z\operatorname{dn}^\tau z\,U_p(\operatorname{sn}^2z),$$

where $\rho,\sigma,\tau=0$ or 1; U_p is a polynomial of degree p and $2p+\rho+\sigma+\tau=n$. We adopt the notation of Arscott [1] and denote the eight types of Lamé polynomials, obtained by giving ρ,σ,τ their possible values, by $uE, sE, cE, dE, scE,\ sdE, cdE, scdE$. Each of these eight types of Lamé polynomials belongs to one of the classes (2.2)–(2.5), to one of the classes (3.2)–(3.5) and to one of the classes (4.2)–(4.5). This is shown in detail in Table 5.1.

TABLE 5.1

		even	odd
periodic	mod $2K$	(2.2) : uE, dE	(2.3) : $scE, scdE$
half-periodic	mod $2K$	(2.4) : cE, cdE	(2.5) : sE, sdE
		even with respect to K	odd with respect to K
periodic	mod $2iK'$	(3.2) : uE, sE	(3.3) : $cdE, scdE$
half-periodic	mod $2iK'$	(3.4) : dE, sdE	(3.5) : cE, scE
		even	odd
periodic	mod $2(K+iK')$	(4.2) : uE, cE	(4.3) : $sdE, scdE$
half-periodic	mod $2(K+iK')$	(4.4) : dE, cdE	(4.5) : sE, scE

Every Lamé polynomial thus satisfies one integral equation of each of §§2, 3, 4. The kernels now involve the Legendre polynomial P_n.

In the following we shall show that for Lamé polynomials the integral relations of the preceding sections can be modified in several ways. It is possible to extend the range of validity of the integral relations by analytic continuation and to admit new paths of integration. We first extend Corollary 1.5.

THEOREM 5.2. *Let λ be one of the $2n+1$ characteristic values such that* (5.1) *has a Lamé polynomial E as a solution. Then every solution G of* (5.1) *is meromorphic on $\mathbb{C}$ with poles at most in M, and for all $x_0, y_0, s \in \mathbb{C}\setminus M$ with $x_0 \pm (y_0 - s) \in \mathbb{C}\setminus M$, we have*

$$E(x_0)G(y_0) = \frac{1}{2}G(s)\big(E(x_0-(y_0-s))+E(x_0+(y_0-s))\big) + \frac{1}{2}\int_{x_0-(y_0-s)}^{x_0+(y_0-s)} \left(G'(s)A(x,s,x_0,y_0) - G(s)\frac{\partial A}{\partial y}(x,s,x_0,y_0)\right) E(x)\,dx,$$

where

$$A(x,y,x_0,y_0) = P_n\left(k^2 \operatorname{sn} x \operatorname{sn} x_0 \operatorname{sn} y \operatorname{sn} y_0 - \frac{k^2}{k'^2}\operatorname{cn} x \operatorname{cn} x_0 \operatorname{cn} y \operatorname{cn} y_0 + \frac{1}{k'^2}\operatorname{dn} x \operatorname{dn} x_0 \operatorname{dn} y \operatorname{dn} y_0\right)$$

and P_n is the Legendre polynomial of degree n. For the path of integration we may choose any path in $\mathbb{C}\setminus M$ from $x_0-(y_0-s)$ to $x_0+(y_0-s)$.

Proof. It is well known that under the assumption of the theorem every solution of (5.1) is meromorphic on $\mathbb{C}$ with poles at most in M. This fact would also easily follow from the remarks below.

Now we prove the integral relation for a fixed $s \in \mathbb{C}\setminus M$. It is easy to see that for all $x_0, y_0 \in \mathbb{C}\setminus M$ the integrand is even with respect to any point in M as a function of x. Therefore, for $x_0, y_0 \in \mathbb{C}\setminus M$ the integrand is meromorphic in x on $\mathbb{C}$ with poles at most in M and residues equal to 0 at any pole. Thus, both sides of the formula above represent analytic functions on the domain $\{(x_0,y_0): x_0, y_0 \in \mathbb{C}\setminus M,\ x_0 \pm (y_0-s) \in \mathbb{C}\setminus M\}$. From 1.5 follows that these functions coincide in a neighborhood of (s,s), and hence, they are identical. □

For example, we now extend the first equation of Corollary 3.6. The other 11 equations of §§2, 3, 4 can be handled in a like manner.

COROLLARY 5.3. *Let G denote any second solution of* (5.1) *which is linearly independent of the first solution uE or sE of* (5.1). *For example, we can choose for G the Lamé function F of the second kind. Then we have, for all $x_0 \in \mathbb{C}\setminus M$,*

(i) $$uE(x_0)\frac{G(K+iK')-G(K-iK')}{G'(K)} = \int_{K-iK'}^{K+iK'} P_n(k \operatorname{sn} x \operatorname{sn} x_0)\,uE(x)\,dx,$$

(ii) $$sE(x_0)\frac{G(K+iK')-G(K-iK')}{G'(K)} = \int_{K-iK'}^{K+iK'} P_n(k \operatorname{sn} x \operatorname{sn} x_0)\,sE(x)\,dx,$$

(iii) $$uE(x_0)\frac{G(K)-G(-K)}{G'(K+iK')} = \int_{-K}^{K} P_n(k \operatorname{sn} x \operatorname{sn} x_0)\,uE(x)\,dx,$$

(iv) $$sE(x_0)\frac{G(K)+G(-K)}{G'(K+iK')} = \int_{-K}^{K} P_n(k \operatorname{sn} x \operatorname{sn} x_0)\,sE(x)\,dx,$$

$$\text{(v)}\quad uE(x_0)\,\frac{G(K+iK')-G(-K-iK')}{G'(K)}=\int_{-K-iK'}^{K+iK'}P_n(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)uE(x)\,dx,$$

$$\text{(vi)}\quad sE(x_0)\,\frac{G(K+iK')+G(-K-iK')}{G'(K)}=\int_{-K-iK'}^{K+iK'}P_n(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)sE(x)\,dx.$$

Proof. (i), (ii). The value of $(G(K+iK')-G(K-iK'))/G'(K)$ does not depend on the choice of the second solution G. Therefore we can assume $G(K)=0$, $G'(K)=1$. Now (i) and (ii) follow from Theorem 5.2 with $s=K$, $y_0=K+iK'$.

(iii) We shall assume $G(K+iK')=0$, $G'(K+iK')=1$. Then Theorem 5.2 with $s=K+iK'$, $y_0=\pm K$, yields

$$uE(x_0)G(K)=\frac{1}{2}\int_{x_0+iK'}^{x_0-iK'}P_n(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)uE(x)\,dx,$$

$$uE(x_0)G(-K)=\frac{1}{2}\int_{x_0+2K+iK'}^{x_0-2K-iK'}P_n(-k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)uE(x)\,dx.$$

For Lamé polynomials of type uE, n is even, thus, $P_n(-k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)=P_n(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)$. The integrands are periodic mod $2iK'$ and mod$(4K+2iK')$; hence, we obtain

$$uE(x_0)G(K)=\frac{1}{2}\int_{-2K}^{-2(K+iK')}P_n(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)uE(x)\,dx,$$

$$uE(x_0)G(-K)=\frac{1}{2}\int_{2K}^{-2(K+iK')}P_n(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)uE(x)\,dx.$$

It follows that

$$\begin{aligned}uE(x_0)(G(K)-G(-K))&=\frac{1}{2}\int_{-2K}^{2K}P_n(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)uE(x)\,dx\\&=\int_{-K}^{K}P_n(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)uE(x)\,dx.\end{aligned}$$

(iv). For Lamé polynomials of type sE n is odd; thus $P_n(-k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)=-P_n(k\,\mathrm{sn}\,x\,\mathrm{sn}\,x_0)$.
The rest of the proof is the same as above.

(v), (vi). We can assume $G(K)=0$, $G'(K)=1$. Then we apply Theorem 5.2 with $s=K$, $y_0=\pm(K+iK')$ and proceed as before. □

The integral equations in Corollary 5.3 (iii), (iv) are given in Whittaker and Watson's book [17, 23.6], without however the determination of the characteristic values. Every integral equation for Lamé polynomials of this book can be found by a method similar to that which gives Corollary 5.3. In all cases the characteristic values can be determined.

Acknowledgment. The author thanks the referees for suggestions which led to the improvement of the manuscript.

REFERENCES

[1] F. M. Arscott, *On Lamé polynomials*, J. London Math. Soc., 32 (1957), pp. 37–48.

[2] _____, *Relations between spherical and ellipsoidal harmonics and some applications*, J. London Math. Soc., 33 (1958), pp. 39–49.

[3] _____, *Integral equations and relations for Lamé functions*, Quart. J. Math. Oxford, Ser. 15 (1964), pp. 103–115.

[4] A. ERDÉLYI, *Integral equations for Lamé functions*, Proc. Edinburgh Math. Soc., (2) 7 (1943), pp. 3–15.
[5] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER AND F. G. TRICOMI, *Higher Transcendental Functions* III, McGraw-Hill, New York, 1955.
[6] P. HENRICI, *Zur Funktionentheorie der Wellengleichung*, Comment. Math. Helv., 27 (1953), pp. 235–293.
[7] E. L. INCE, *On the connection between linear differential systems and integral equations*, Proc. Royal Soc. Edinburgh, 42 (1922), pp. 43–53.
[8] _____, *The periodic Lamé functions*, Proc. Royal Soc. Edinburgh, 60 (1940), pp. 47–63.
[9] _____, *Further investigations into the periodic Lamé functions*, Proc. Royal Soc. Edinburgh, 60 (1940), pp. 83–99.
[10] D. SCHMIDT AND G. WOLF, *A method of generating integral relations by the simultaneous separability of generalized Schrödinger equations*, this Journal, 10 (1979), pp. 823–838.
[11] R. SHAIL, *On integral representations for Lamé and other special functions*, this Journal, 11 (1980), pp. 702–723.
[12] I. N. VEKUA, *New Methods for Solving Elliptic Equations*, Moskau-Leningrad 1948 (Russian), Applied Mathematics and Mechanics Series, vol. 1, North-Holland, New York, 1967 (English).
[13] H. VOLKMER, *Integralrelationen mit variablen Grenzen fur spezielle Funktionen der mathematischen Physik*, Dissertation, Konstanz 1979.
[14] _____, *Integralrelationen mit variablen Grenzen fur spezielle Funktionen der mathematischen Physik*, J. Reine Angew. Math., 319 (1980), pp. 118–132.
[15] E. T. WHITTAKER, *On an integral equation whose solutions are the functions of Lamé*, Proc. Royal Soc. Edinburgh, 35 (1915), pp. 70–77.
[16] _____, *On Lamé's differential equation and ellipsoidal harmonics*, Proc. London Math. Soc., (2) 14 (1915), pp. 260–268.
[17] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Oxford, Cambridge, 1927.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0008 $01.00/0

SOME CONJECTURES FOR ROOT SYSTEMS*

I. G. MACDONALD†

Abstract. We present a collection of conjectures relating to root systems (or, equivalently, to Lie groups) together with the evidence we possess in support of them. They may be regarded as generalizations of Dyson's conjecture [J. Math. Phys. 3 (1962), pp. 140–156] and Mehta's conjecture [*Random Matrices*, Academic Press, New York, 1967].

Introduction. In this paper we present a somewhat heterogeneous collection of conjectures relating to root systems (or, equivalently, to Lie groups) together with the evidence we possess in support of them. The first group of conjectures may be considered as generalisations of "Dyson's conjecture" [2]. Thus if R is a reduced root system, let e^α denote the formal exponential corresponding to $\alpha \in R$, and let k be an integer ≥ 0; then we conjecture (2.1) that the constant term in the polynomial

$$\prod_{\alpha \in R} (1-e^\alpha)^k \tag{1}$$

should be equal to $\prod_{i=1}^{l} \binom{kd_i}{k}$, where the d_i are the degrees of the fundamental invariants of the Weyl group of R.

We generalise this further, in two different directions. First, suppose now that R is a root system, not necessarily reduced, and for each $\alpha \in R$ let k_α be a nonnegative integer, such that $k_\alpha = k_\beta$ if $|\alpha| = |\beta|$; then we conjecture (2.3) that the constant term in the Laurent polynomial

$$\prod_{\alpha \in R} (1-e^\alpha)^{k_\alpha} \tag{2}$$

should be equal to the product

$$\prod_{\alpha \in R} \frac{\left(\left|\left\langle \rho_k, \alpha^\vee \right\rangle + k_\alpha + \tfrac{1}{2}k_{\alpha/2}\right|\right)!}{\left(\left|\left\langle \rho_k, \alpha^\vee \right\rangle + \tfrac{1}{2}k_{\alpha/2}\right|\right)!} \tag{3}$$

where $\rho_k = \frac{1}{2}\sum_{\alpha>0} k_\alpha \alpha$, $\alpha^\vee = 2\alpha/|\alpha|^2$ is the coroot corresponding to α, and $k_{\alpha/2} = 0$ if $\frac{1}{2}\alpha \notin R$. When the k_α are all equal, this reduces to the previous conjecture.

Secondly, if q is an indeterminate and R is a reduced root system, we conjecture (3.1) that the constant term (i.e., the term not involving any e^α) in the product

$$\prod_{\alpha \in R^+} \prod_{i=1}^{k} (1-q^{i-1}e^{-\alpha})(1-q^i e^\alpha) \tag{4}$$

(where R^+ is a system of positive roots in R) should be equal to $\prod_{i=1}^{l} \begin{bmatrix} kd_i \\ k \end{bmatrix}$, where $\begin{bmatrix} n \\ r \end{bmatrix}$ denotes the "q-binomial coefficient" $(1-q^n)\cdots(1-q^{n-r+1})/(1-q)\cdots(1-q^r)$. Clearly (4) reduces to (1) when $q=1$.

This conjecture can be reformulated in terms of the affine root system [9] defined by R, and in turn this reformulation suggests a more ambitious conjecture (3.3) for any reduced affine root system.

---

* Received by the editors July 14, 1981.

† Department of Pure Mathematics, Queen Mary College, University of London, London E1 4NS, England.

The second group of conjectures generalize Mehta's conjecture [12]. Let S be a (not necessarily reduced) root system in a real Euclidean space V of dimension l, with inner product $\langle x,y\rangle$ and norm $|x|=\langle x,x\rangle^{1/2}$. Attach to each $\alpha\in S$ a "multiplicity" k_α (which need not be an integer, but may be a complex number with real part ≥ 0) such that $k_\alpha=k_\beta$ if $|\alpha|=|\beta|$; also define $\tilde{\alpha}=\sqrt{2}\,\alpha/|\alpha|$ (so that $\tilde{\alpha}$ is proportional to α and $|\tilde{\alpha}|^2=2$). Let

$$P(x)=\prod_{\alpha\in S^+}|\tilde{\alpha}(x)|^{k_\alpha},\qquad x\in V$$

where S^+ is a system of positive roots in S, and let γ denote the Gaussian measure on V defined by $d\gamma(x)=(2\pi)^{-l/2}e^{-|x|^2/2}dx$, where dx is the Lebesgue measure on the Euclidean space V. Then we conjecture (6.1) that

$$\int_V P(x)\,d\gamma(x)=\prod_{\alpha\in S^+}\frac{\left(\frac{1}{2}k_\alpha+\frac{1}{4}k_{\alpha/2}+\frac{1}{2}\left\langle\rho_k,\alpha^\vee\right\rangle\right)!}{\left(\frac{1}{4}k_{\alpha/2}+\frac{1}{2}\left\langle\rho_k,\alpha^\vee\right\rangle\right)!}\tag{5}$$

where $\rho_k=\frac{1}{2}\sum_{\alpha\in S^+}k_\alpha\alpha$, $\alpha^\vee$ is the coroot $2\alpha/|\alpha|^2$, and $k_{\alpha/2}=0$ if $\frac{1}{2}\alpha\notin S$. The right-hand side of (5) is reminiscent of the formula of Gindikin and Karpelevich [3] for the Harish-Chandra c-function, and we show that (5) is true when S is the restricted root system of a symmetric space G/K and the k_α are the multiplicities of the roots.

A. "Dyson's conjecture" and generalizations.

1. In 1962 F. J. Dyson [2] conjectured that

(1.1). *The constant term in the expansion of*

$$\prod_{i\neq j}\left(1-x_ix_j^{-1}\right)^k$$

(*where $x_1,\cdots,x_n$ are independent variables and k is a positive integer*) *is*

$$\frac{(nk)!}{(k!)^n}.$$

This conjecture was soon proved true by J. Gunson [5] and K. Wilson [15], who showed more generally that

(1.2). *The constant term in the expansion of*

$$\prod_{i\neq j}\left(1-x_ix_j^{-1}\right)^{a_i}$$

(*where $a_1,\cdots,a_n$ are nonnegative integers*) *is*

$$\frac{(a_1+\cdots+a_n)!}{a_1!\cdots a_n!}.$$

For an elegant proof of (1.2) we refer to I. J. Good [4].

Next, G. E. Andrews [1] conjectured that a q-analogue of (1.2) should be true. Write

$$(x)_n=\prod_{i=1}^{n}\left(1-q^{i-1}x\right)$$

where q,x are independent variables, and $n\geq 0$.

(1.3). *The polynomial* ($\in \mathbb{Z}[x_1^{\pm 1}, \cdots, x_n^{\pm 1}, q]$)

$$\prod_{i \neq j} \left(\varepsilon_{ij} x_i x_j^{-1}\right)_{a_i}$$

where $\varepsilon_{ij} = 1$ *or* q *according as* $i<j$ *or* $i>j$, *has constant term* (*i.e., independent of* $x_1, \cdots, x_n$)

$$\frac{(q)_{a_1+\cdots+a_n}}{(q)_{a_1} \cdots (q)_{a_n}}.$$

When $q=1$, (1.3) reduces to (1.2). For arbitrary q and $n=2$, (1.3) is an easy consequence of the q-binomial theorem. Andrews [1] gives a proof of (1.3) for $n=3$. For $n>3$, the question is still open, as far as I am aware. Incidentally, (1.3) still makes sense if some (or all) of the a_i are $+\infty$, and is implied by the finite version.

We shall see later (§3) that (1.3) is true for arbitrary n and q when $a_1 = \cdots = a_n = 1, 2$ or $+\infty$.

2. The form of the polynomial in (1.1) suggests that it should be attached to the root system of type A_{n-1}, and prompts the following generalization. Let R be a reduced root system and for each $\alpha \in R$ let e^α be the corresponding formal exponential. Let $d_1, \cdots, d_l$ be the degrees of the fundamental invariants of the Weyl group W of R.

CONJECTURE 2.1. *With the above notation, the constant term in*

$$\prod_{\alpha \in R} (1-e^\alpha)^k,$$

where k *is an integer* ≥ 0, *should be equal to*

$$\prod_{i=1}^{l} \binom{kd_i}{k}.$$

When R is of type A_{n-1}, the d_i are $2, 3, \cdots, n$, so that in this case

$$\prod_{i=1}^{l} \binom{k_i}{k} = \binom{2k}{k}\binom{3k}{k} \cdots \binom{nk}{k} = \frac{(nk)!}{(k!)^n}$$

in agreement with (1.1).

Also (2.1) is true when $k=1$, for any R. For we have

$$\begin{aligned} \prod_{\alpha \in R} (1-e^\alpha) &= \prod_{\alpha>0} (e^{\alpha/2} - e^{-\alpha/2}) \prod_{\alpha>0} (e^{-\alpha/2} - e^{\alpha/2}) \\ &= \left(\sum_{w \in W} \varepsilon(w) e^{w\rho}\right)\left(\sum_{w \in W} \varepsilon(w) e^{-w\rho}\right) \end{aligned}$$

by Weyl's identity ($\varepsilon(w) = \pm 1$ is the sign of $w \in W$; ρ is half the sum of the positive roots). Since $w\rho = w'\rho$ if and only if $w = w'$, the constant term is $|W|$, and it is well known that $|W| = \prod d_i$.

We shall show in §3 that (2.1) is true for any R and $k=2$. It is also true when R is of classical type (A, B, C, D) for any k, as a consequence of Selberg's integral [13].

Before coming to these verifications, we shall write (2.1) in an equivalent form. Let G be a compact connected Lie group, T a maximal torus of G, such that R is the root system of (G, T). We regard the exponentials e^α now as characters of T, and write

$$\Delta(t) = \prod_{\alpha>0} \left(e^{\alpha/2}(t) - e^{-\alpha/2}(t)\right).$$

Then

$$|\Delta(t)|^2 = \prod_{\alpha \in R} (1 - e^{\alpha}(t))$$

is a positive real-valued continuous function on T, which enters in Weyl's integration formula

$$\int_G f(x)\,dx = \frac{1}{|W|} \int_T |\Delta(t)|^2 f(t)\,dt \tag{2.2}$$

for any continuous class function f on G, where dx, dt are the normalized Haar measures ($\int_G dx = \int_T dt = 1$). Then (2.1) is equivalent to

CONJECTURE 2.1′. *With the above notation*

$$\int_T |\Delta(t)|^{2k} dt = \prod_{i=1}^{l} \binom{kd_i}{k}$$

for all integers $k \geq 0$.

The equivalence of (2.1) with (2.1′) follows from the fact that integration over T kills all but the trivial character, i.e., picks out the constant term in

$$|\Delta(t)|^{2k} = \prod_{\alpha \in R} (1 - e^{\alpha}(t))^k.$$

Taking f to be the constant function 1 in (2.2), we see again that (2.1′) is true for $k = 1$.

We may remark also that (2.1′) makes sense if the integer k is replaced by a complex number s with $\mathrm{Re}(s) > 0$, the binomial coefficients on the right being replaced by the appropriate combination of Γ-functions:

CONJECTURE 2.1″. *For all* $s \in \mathbb{C}$ *with* $\mathrm{Re}(s) > 0$,

$$\int_T |\Delta(t)|^{2s} dt = \prod_{i=1}^{l} \frac{\Gamma(sd_i + 1)}{\Gamma(s+1)\Gamma(sd_i - s + 1)}.$$

I do not know how to generalize (1.2) (with unequal exponents $a_1, \cdots, a_n$) to arbitrary root systems. However, a generalization of (2.1) which allows for unequal exponents corresponding to roots of different lengths is contained in the following conjecture. Let R be a root system, not necessarily reduced, and for each $\alpha \in R$ let k be a nonnegative integer such that $k_\alpha = k_\beta$ whenever $|\alpha| = |\beta|$. Let

$$\rho_k = \frac{1}{2} \sum_{\alpha \in R^+} k_\alpha \alpha,$$

where R^+ is a system of positive roots in R.

CONJECTURE 2.3. *The constant term in the Laurent polynomial*

$$\prod_{\alpha \in R} (1 - e^{\alpha})^{k_\alpha}$$

should be equal to

$$\prod_{\alpha \in R} \frac{\left(\left|\langle \rho_k, \check{\alpha} \rangle + k_\alpha + \frac{1}{2} k_{\alpha/2}\right|\right)!}{\left(\left|\langle \rho_k, \check{\alpha} \rangle + \frac{1}{2} k_{\alpha/2}\right|\right)!}.$$

Here $\alpha^\vee = 2\alpha/|\alpha|^2$ is the coroot corresponding to α, and $k_{\alpha/2}=0$ if $\frac{1}{2}\alpha \notin R$.

(a). When all the k_α are equal, say $k_\alpha = k$, then (2.3) reduces to (2.1). For then $\langle \rho_k, \alpha^\vee \rangle = k\langle \rho, \alpha^\vee \rangle = k\,\mathrm{ht}(\alpha^\vee)$, and it follows from [10, (2.5)] that

$$\prod_{\alpha\in R^+} \frac{(k\,\mathrm{ht}(\alpha^\vee)+k)!}{(k\,\mathrm{ht}(\alpha^\vee))!} = \prod_{i=1}^{l} \frac{(kd_i)!}{k!}$$

and that

$$\prod_{\alpha\in R^+} \frac{(k\,\mathrm{ht}(\alpha^\vee)-k)!}{(k\,\mathrm{ht}(\alpha^\vee))!} = \prod_{i=1}^{l} \frac{0!}{(k(d_i-1))!}.$$

Hence in this case the constant term as predicted by (2.3) is equal to $\prod_{i=1}^{l} \binom{kd_i}{k}$.

(b) Next, when R is of type BC_n, (2.3) is equivalent to an integral formula of Selberg [13]. Selberg's formula is the following: let

$$J_n(a,b;c) = \int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} \left(x_i^a (1-x_i)^b\right) \cdot |D(x)|^{2c}\, dx_1 \cdots dx_n,$$

where

$$D(x) = \prod_{i<j} (x_i - x_j)$$

and a, b, c are complex numbers satisfying

$$\mathrm{Re}(a) > -1, \quad \mathrm{Re}(b) > -1, \quad \mathrm{Re}(c) > -\min\left(\frac{1}{n}, \frac{\mathrm{Re}(a+1)}{n-1}, \frac{\mathrm{Re}(b+1)}{n-1}\right).$$

Then

$$J_n(a,b;c) = \prod_{r=1}^{n} \frac{(rc)!\,(a+(r-1)c)!\,(b+(r-1)c)!}{c!\,(a+b+1+(n+r-2)c)!}, \tag{2.4}$$

where we have written $x!$ in place of $\Gamma(x+1)$. (When $n=1$, this integral is Euler's beta function.)

Now make the change of variables $x_i = \sin^2\theta_i$ $(1 \le i \le n)$, so that the range of integration is $0 \le \theta_i \le \frac{\pi}{2}$ $(1 \le i \le n)$. Since $dx_i = 2\sin\theta_i \cos\theta_i\, d\theta_i$ and $x_i - x_j = \sin(\theta_i - \theta_j)\sin(\theta_i + \theta_j)$, we obtain

$$J_n(a,b;c) = \int_0^{\pi} \cdots \int_0^{\pi} \prod_{i=1}^{n} \left(\sin^{2a+1}\theta_i \cos^{2b+1}\theta_i\right) \prod_{i<j} \left|\sin(\theta_i - \theta_j)\sin(\theta_i + \theta_j)\right|^{2c} d\theta_1 \cdots d\theta_n.$$

For our applications it is more convenient to rewrite this as follows. Put

$$\alpha = a + \tfrac{1}{2}, \quad \beta = b + \tfrac{1}{2}, \quad \gamma = c$$

and let

$$K_n(\alpha,\beta;\gamma) = \pi^{-n} \int_0^{\pi} \cdots \int_0^{\pi} \prod_{i=1}^{n} \left(4\sin^2\theta_i\right)^{\alpha} \left(4\cos^2\theta_i\right)^{\beta} \prod_{i<j} \left(4\sin^2(\theta_i - \theta_j) 4\sin^2(\theta_i + \theta_j)\right)^{\gamma} d\theta_1 \cdots d\theta_n.$$

Then from Selberg's formula we have

$$K_n(\alpha,\beta;\gamma)=\frac{4^{n(\alpha+\beta)+n(n-1)\gamma}}{\pi^n}J_n(\alpha-\tfrac{1}{2},\beta-\tfrac{1}{2};\gamma)$$
$$=\frac{4^{n(\alpha+\beta)+n(n-1)\gamma}}{\pi^n}\prod_{r=1}^{n}\frac{(r\gamma)!(\alpha+(r-1)\gamma-\frac{1}{2})!(\beta+(r-1)\gamma-\frac{1}{2})!}{\gamma!(\alpha+\beta+(n+r-2)\gamma)!}.$$

Use of the duplication formula for the gamma function, in the form

$$(x-\tfrac{1}{2})!=\frac{\pi^{1/2}(2x)!}{2^{2x}x!}, \tag{2.5}$$

then reduces this to

$$K_n(\alpha,\beta;\gamma)=\prod_{r=1}^{n}\frac{(r\gamma)!(2\alpha+2(r-1)\gamma)!(2\beta+2(r-1)\gamma)!}{\gamma!(\alpha+(r-1)\gamma)!(\beta+(r-1)\gamma)!(\alpha+\beta+(n+r-2)\gamma)!}. \tag{2.4′}$$

Suppose now that R is of type BC_n, with roots $\pm x_i$, $\pm 2x_i$ $(1\le i\le n)$, $\pm x_i\pm x_j$ $(1\le i<j\le n)$. Write

$$k_1=k_{\pm x_i},\quad k_2=k_{\pm x_i\pm x_j},\quad k_3=k_{\pm 2x_i}.$$

To obtain the constant term in the Laurent polynomial $\prod_{\alpha\in R}(1-e^{\alpha})^{k_\alpha}$ we shall replace each e^{α} by the complex exponential $\exp(2\pi i\alpha)$, and integrate from 0 to 1 with respect to each variable x_i. Since

$$(1-\exp(2\pi i\alpha))(1-\exp(-2\pi i\alpha))=4\sin^2\pi\alpha,$$

it is easy to see that in the present case the constant term is equal to the integral $K_n(k_1+k_3,k_3;k_2)$, hence by (2.4′) to

$$(1)\quad \prod_{r=1}^{n}\frac{(rk_2)!(2(k_1+k_3+(r-1)k_2))!(2(k_3+(r-1)k_2))!}{(k_2)!(k_1+k_3+(r-1)k_2)!(k_3+(r-1)k_2)!(k_1+2k_3+(n+r-2)k_2)!}.$$

Now consider the product

$$P=\prod_{\alpha\in R}\frac{\left(\left|\langle\rho_k,\check{\alpha}\rangle+k_\alpha+\frac{1}{2}k_{\alpha/2}\right|\right)!}{\left(\left|\langle\rho_k,\check{\alpha}\rangle+\frac{1}{2}k_{\alpha/2}\right|\right)!}.$$

The positive roots may be taken to be x_i, $2x_i$, $x_i\pm x_j$ $(i<j)$, so that we have $\rho_k=\sum_{i=1}^{n}(\frac{1}{2}k_1+(n-i)k_2+k_3)x_i$. Hence we find that the contribution to P from the roots $\pm x_i$ is

$$(2)\quad \prod_{i=1}^{n}\frac{(2(k_1+(n-i)k_2+k_3))!(2((n-i)k_2+k_3))!}{(k_1+2(n-i)k_2+2k_3)!^2};$$

the contribution from the roots $\pm 2x_i$ is

$$(3)\quad \prod_{i=1}^{n}\frac{(k_1+(n-i)k_2+2k_3)!((n-i)k_2)!}{(k_1+(n-i)k_2+k_3)!((n-i)k_2+k_3)!};$$

the contribution from the roots $\pm(x_i-x_j)$ is

$$(4)\quad \frac{(nk_2)!}{(k_2!)^n};$$

and the contribution from the roots $\pm(x_i+x_j)$ is

$$\prod_{i=1}^{n} \frac{(k_1+2(n-i)k_2+2k_3)!}{(k_1+(2n-2i+1)k_2+2k_3)!} \frac{(k_1+(2n-1)k_2+2k_3)!}{(k_1+(n-1)k_2+2k_3)!}. \tag{5}$$

Since the product of (2), (3), (4) and (5) is equal to (1), it follows that (2.3) is true for BC_n (and therefore also for B_n, C_n and D_n). Thus it is true for all the classical root systems.

When $k_1=k_2=k_3=k$ say, the constant term is equal to

$$\prod_{r=1}^{n} \binom{kd_r}{k}$$

where $(d_1,\cdots,d_n)=(4,6,\cdots,2n+2)$. These "degrees" d_i for BC_n also occur in the formula

$$\prod_{\alpha>0} \frac{1-t^{\mathrm{ht}(\alpha)+s(\alpha)}}{1-t^{\mathrm{ht}(\alpha)}} = \prod_{i=1}^{n} \frac{1-t^{d_i}}{1-t} \tag{2.6}$$

where $s(\alpha)=1$ or 2 according as $\frac{1}{2}\alpha \notin R$ or $\frac{1}{2}\alpha \in R$.

Hence we may rephrase (2.1) as follows to include the nonreduced root systems:

CONJECTURE 2.7. *Let R be a (not necessarily reduced) root system. Then the constant term in*

$$\prod_{\alpha\in R} (1-e^{\alpha})^{k}$$

should be

$$\prod_{i=1}^{l} \binom{kd_i}{k}$$

where the d_i are given by (2.6).

Of course this is included in the more general conjecture (2.3).

(c) When R is of type G_2, let k_1 (resp. k_2) denote k_α for α a short (resp. long) root. In this case (2.3) asserts that the constant term in $\prod_{\alpha\in R}(1-e^{\alpha})^{k_\alpha}$ is equal to

$$\frac{(3k_1+3k_2)!(2k_1)!(2k_2)!(3k_2)!}{(2k_1+3k_2)!(k_1+2k_2)!(k_1+k_2)!k_1!k_2!^2},$$

which agrees with a conjecture of W. Morris [R. Askey, private communication].

Finally, (2.1′) has the following consequence. Let t be a regular element of the maximal torus T of G. Then the centralizer of t in G is just T, and the conjugacy class $C(t)$ of t in G is a smooth submanifold of G, of codimension $l=\dim T$, and diffeomorphic to G/T. The class $C(t)$ meets T in a single orbit of the Weyl group W, and the weighting factor $|\Delta(t)|^2$ in (2.2) measures the size of the conjugacy class $C(t)$: more precisely, it is the Riemannian volume of $C(t)$. Hence the maximum value of the function $|\Delta(t)|^2$ on T gives the volume of the "largest conjugacy class in G". This maximum value can be deduced from (2.1′), because

$$\max_{t\in T}|\Delta(t)|^2 = \lim_{k\to\infty}\left[\int_T |\Delta(t)|^{2k}dt\right]^{1/k} = \lim_{k\to\infty}\prod_{i=1}^{l}\binom{kd_i}{k}^{1/k}$$

which by Stirling's formula we find to be equal to

$$\prod_{i=1}^{l} \frac{d_i^{d_i}}{(d_i-1)^{d_i-1}}.$$

Hence (2.1′) implies

CONJECTURE 2.8. *The maximum value of* $|\Delta|^2$ *on* T *is*

$$|W| \prod_{i=1}^{l} \left(1+\frac{1}{m_i}\right)^{m_i},$$

where $m_i=d_i-1$ *are the exponents of* W.

In the case of G_2, for example, the function to be maximized is

$$\prod_{i=1}^{3} 4\sin^2 x_i \prod_{1\le i<j\le 3} 4\sin^2(x_i-x_j)$$

subject to $x_1+x_2+x_3=0$. This is an elementary exercise in calculus, and it is not difficult to verify that the maximum value is $2^2\cdot 6^6/5^5$, in agreement with (2.8) (since $d_1=2, d_2=6$).

For the classical types, (2.8) can of course be attacked directly. For example, in type B_n calculation shows that the maximum value of $|\Delta|^2$ is

$$2^{n^2} \prod_{i=1}^{n} (1-\xi_i) \prod_{1\le i<j\le n} (\xi_i-\xi_j)$$

where the ξ_i are essentially the roots of a Jacobi polynomial (to be precise, they are the roots of $(1+x)P_{n-1}^{(0,1)}(x)$, in the usual notation [14]). In the case of C_n a similar phenomenon occurs, but with Legendre polynomials in place of Jacobi polynomials.

3. We come now to "q-analogues" of the conjectures in §2. Consider Andrews' conjecture (1.3) with $a_1=\cdots=a_n=k$. Writing $x_i=e^{-u_i}$ (formal exponentials) and thinking of u_i-u_j as the roots in a root system of type A_{n-1}, we are led to generalize (1.3) as follows:

CONJECTURE 3.1. *The constant term* (*i.e., involving* q *but no exponentials* e^{α}) *in*

$$\prod_{\alpha>0} \prod_{i=1}^{k} (1-q^{i-1}e^{-\alpha})(1-q^i e^{\alpha}),$$

where k *is a positive integer or* $+\infty$ *and the product is over a system of positive roots in a reduced root system* R, *should be*

$$\prod_{i=1}^{l} \begin{bmatrix} kd_i \\ k \end{bmatrix}$$

where $\left[\begin{smallmatrix} n \\ r \end{smallmatrix}\right]$ *is the "q-binomial coefficient"*

$$\frac{(1-q^n)(1-q^{n-1})\cdots(1-q^{n-r+1})}{(1-q)(1-q^2)\cdots(1-q^r)}.$$

Clearly (3.1) implies (2.1), by setting $q=1$.

An equivalent formulation of (3.1) is the following. As in §2, let G be a compact connected Lie group with R as its root system. Then (3.1) is equivalent to

CONJECTURE 3.1′.

$$\int_G \det \prod_{j=1}^{k-1} (1-q^j \mathrm{Ad}\, x)\, dx = \prod_{i=1}^{l} \prod_{j=1}^{k-1} (1-q^{km_i+j})$$

where dx is the normalised Haar measure on G, and the m_i are the exponents of G (i.e., $m_i = d_i - 1$).

To prove the equivalence of (3.1) and (3.1′) we shall use the identity [10]:

$$\sum_{w\in W} \prod_{\alpha>0} \frac{1-q^k e^{-w\alpha}}{1-e^{-w\alpha}} = W(q^k) = \prod_{i=1}^{l} \frac{1-q^{kd_i}}{1-q^k}.$$

Multiply both sides by $\prod_{\alpha\in R}\prod_{j=1}^{k}(1-q^{j-1}e^{\alpha})$ and we have

$$\sum_{w\in W} \prod_{\alpha>0} \prod_{j=1}^{k} (1-q^{j-1}e^{w\alpha})(1-q^j e^{-w\alpha}) = W(q^k) \prod_{j=1}^{k} (1-q^{j-1}e^{\alpha}).$$

Each of the terms in the sum on the left has the same constant term $c_k(q)$, hence

$$\begin{aligned} c_k(q) &= \frac{W(q^k)}{|W|} \int_T \prod_{\alpha\in R} \prod_{j=1}^{k} (1-q^{j-1}e^{\alpha})(t)\, dt \\ &= \frac{W(q^k)}{|W|} \int_T |\Delta(t)|^2 \prod_{j=1}^{k-1} \frac{\det(1-q^j \mathrm{Ad}\, t)}{(1-q^j)^l}\, dt \end{aligned}$$

which by Weyl's formula (2.2) is equal to

$$\frac{W(q^k)}{\prod_{j=1}^{k-1}(1-q^j)^l} \int_G \prod_{j=1}^{k-1} \det(1-q^j \mathrm{Ad}\, x)\, dx.$$

Hence (3.1′) implies that

$$\begin{aligned} c_k(q) &= \prod_{i=1}^{l} \frac{(1-q^{kd_i})(1-q^{km_i+1})\cdots(1-q^{km_i+k-1})}{(1-q^k)(1-q)(1-q^2)\cdots(1-q^{k-1})} \\ &= \prod_{i=1}^{l} \begin{bmatrix} kd_i \\ k \end{bmatrix}, \end{aligned}$$

which is (3.1), and conversely (3.1)⇒(3.1′).

PROPOSITION. *Conjecture* (3.1) *is true for any reduced root system R when $k = 1, 2$ or $+\infty$.*

Proof. Conjecture (3.1′) is obvious for $k=1$ (both sides are equal to 1). When $k=2$, (3.1′) asserts that (replacing q by $-q$)

$$\int_G \det(1+q\,\mathrm{Ad}\, x)\, dx = \prod_{i=1}^{l} (1+q^{2m_i+1}).$$

Now the left-hand side of (1) is the Poincaré polynomial of the graded $\mathbb{R}$-algebra $(\wedge \mathfrak{g})^G$, the G-invariants of the exterior algebra of $\mathfrak{g}$ (the Lie algebra of G) under the

adjoint action. But it is well known that $(\wedge \mathfrak{g})^G \cong H^*(G, \mathbb{R})$ (via de Rham cohomology) and that $H^*(G, \mathbb{R})$ is an exterior algebra over $\mathbb{R}$ on generators of degrees $2m_i+1$, so that the Poincaré polynomial is equal to the right-hand side of (1).

Finally, when $k=+\infty$, the identity [9, (8.5)] shows that the constant term in

$$\prod_{\alpha>0} \prod_{i\geq 1} (1-q^{i-1}e^{-\alpha})(1-q^i e^{\alpha})$$

is equal to $\prod_{i\geq 1}(1-q^i)^{-l}$, which confirms (3.1) in this case. □

COROLLARY. *Conjecture* (2.1) *is true for* $k=2$ *and any* R.

Remarks. 1. The case $k=+\infty$ of (3.1′) (which we have just proved) is

$$\int_G \prod_{j=1}^{\infty} \det(1-q^j \operatorname{Ad} x)\, dx = 1$$

(the integrand is to be regarded as a formal power series in q, and integrated term by term; alternatively as a convergent infinite product, with q a complex number such that $|q|<1$).

2. In connection with (3.1′) we may recall that

$$\int_G \det(1-q \operatorname{Ad} x)^{-1} dx = \prod_{i=1}^{l} (1-q^{d_i})^{-1}.$$

Here the left-hand side is the Poincaré series of the graded algebra $(S\mathfrak{g})^G$, the G-invariants of the symmetric algebra of $\mathfrak{g}$, and the right-hand side is the Poincaré series of $(S\mathfrak{t})^W$, where $\mathfrak{t}$ is the Lie algebra of the maximal torus T; the identity (2) is a consequence of the isomorphism $(S\mathfrak{g})^G \cong (S\mathfrak{t})^W$. I do not know whether (2) has an extension analogous to (3.1′).

We shall next reformulate (3.1) in terms of the affine root system $S=S(R)$ [9] defined by R. We recall that the affine roots are $a=m+\alpha$ ($m\in\mathbb{Z}$, $\alpha\in R$), and we write

$$e^{-a}=q^m e^{-\alpha}$$

where $q=e^{-1}$ ($\neq 1/2.718\ldots$). Let

$$\sigma=\frac{1}{2}\sum_{\alpha>0} \check{\alpha},$$

where $\check{\alpha}$ is the coroot of α, so that $\alpha(\sigma)$ is the *height* of α.

We shall use the following fact [10]: if θ is any mapping of the positive integers into a multiplicative abelian group, we have

$$\prod_{\alpha>0} \frac{\theta(1+\alpha(\sigma))}{\theta(\alpha(\sigma))} = \prod_{i=1}^{l} \frac{\theta(d_i)}{\theta(1)}.$$

By taking $\theta(m)=(1-q^{km})(1-q^{km-1})\cdots(1-q^{km-k+1})$ for all $m\geq 1$, we see that

$$\prod_{i=1}^{l} \begin{bmatrix} kd_i \\ k \end{bmatrix} = \prod_{\alpha>0} \prod_{i=1}^{k} \frac{1-q^{\alpha(k\sigma)+i}}{1-q^{\alpha(k\sigma)-i+1}}. \tag{3.2}$$

Now the affine roots $a_i=\alpha_i$ ($1\leq i\leq l$), and $a_0=1-\phi$ (where $\alpha_1,\cdots,\alpha_l$ is a basis of R, and ϕ is the highest root relative to this basis—we are assuming, as we clearly may, that R is irreducible) form a basis of $S(R)$ [9]. Let C be the fundamental alcove determined by this basis, so that $x\in C$ if and only if $a_i(x)>0$ for $0\leq i\leq l$, and let

$$S_k=\{a\in S: a(C)\subset(0,k)\}$$

for a positive integer k. Since each $\alpha \in R$ takes values between 0 and 1 at points of C, it follows that S_k consists of the affine roots $i-1+\alpha$, $i-\alpha$ $(1 \leq i \leq k,\ \alpha \in R^+)$, which are precisely the negatives of the roots whose exponentials feature in the product in (3.1). Also the exponents on the right-hand side of (3.2) are $\alpha(k\sigma)+i=(i-\alpha)(-k\sigma)$ and $\alpha(k\sigma)-i+1=-(i-1+\alpha)(-k\sigma)$. Hence we may write Conjecture 3.1 in the equivalent form

CONJECTURE 3.1″. *For the affine root system $S(R)$, where R is reduced, the constant term in*

$$\prod_{a \in S_k} (1-e^{-\alpha})$$

should be

$$\prod_{a \in S_k} \left(1-q^{|a(s_k)|}\right)^{\varepsilon(a)}$$

where $s_k=-k\sigma$ and $\varepsilon(a)$ is the sign of $a(s_k)$.

It is natural to ask whether there exists something analogous to (3.1″) for the other reduced affine root systems (i.e., $S(BC_l)$ or $S(R)^{\vee}$, in the notation of [9]). For each $a \in S$, let u_a be the smallest positive real number such that $a+u_a$ is a root, and assume that the constant c of [9, §6] is equal to 1 (so that each u_a is a positive integer). As before, choose a fundamental alcove C and let

$$S_k=\{a \in S: a(C) \subset (0,k)\}$$

for each positive integer k. Let u be the least common multiple of the u_a (so that $u=1,2$ or 3). Some experimental evidence then suggests the following:

CONJECTURE 3.3. *Let $a_0,\cdots,a_l$ be the basis of S determined by the alcove C, and assume that the vertex x_0 of C defined by $a_i(x_0)=0$ $(1 \leq i \leq l)$ is a special vertex. Let s_k be the point for which $a_i(s_k)=-k$ $(1 \leq i \leq l)$. Then, provided that k is a multiple of u, the constant term in*

$$\prod_{a \in S_k} (1-e^{-a})$$

should be

$$\prod_{a \in S_k} \left(1-q^{|a(s_k)|}\right)^{\varepsilon(a)}$$

where $\varepsilon(a)$ is the sign (± 1) of $a(s_k)$.

When $S=S(R)$ we have $u=1$, and (3.3) is a restatement of (3.1″).

We may remark here that (3.3) makes sense, and is true, in the limit as $k \to \infty$. For it asserts that the constant term in

$$\prod_{a>0} (1-e^{-a})$$

is

$$\lim_{k \to \infty} \prod_{a \in S_k} \left(1-q^{|a(s_k)|}\right)^{\varepsilon(a)}.$$

To compute this limit we may take $x_0=0$; the affine roots belonging to S_k are then, using the notation of [9],

$$\begin{aligned} &\alpha+iu_\alpha \qquad (\alpha\in\Sigma^+,\ 0\le iu_\alpha<k),\\ &-\alpha+iu_\alpha \qquad (\alpha\in\Sigma^+,\ 0<iu_\alpha\le k)\end{aligned}$$

where Σ is the gradient root system of S; hence the values of $a(s_k)$ are correspondingly

$$\begin{aligned} &-k\,\mathrm{ht}(\alpha)+iu_\alpha \qquad (\alpha\in\Sigma^+,\ 0\le iu_\alpha<k),\\ &k\,\mathrm{ht}(\alpha)+iu_\alpha \qquad (\alpha\in\Sigma^+,\ 0<iu_\alpha\le k)\end{aligned}$$

so that $|a(s_k)|>k$ except when $\alpha=Da$ has height 1, i.e., is a simple root of Σ. It follows that

$$\prod_{a\in S_k}\left(1-q^{|a(s_k)|}\right)^{\varepsilon(a)}=\prod_{\alpha\in B(\Sigma)}\prod_{r=1}^{k/u_\alpha}(1-q^{ru_\alpha})^{-1}\ (\mathrm{mod}\,q^k)$$

and hence that

$$\lim_{k\to\infty}\prod_{a\in S_k}\left(1-q^{|a(s_k)|}\right)^{\varepsilon(a)}=\prod_{\alpha\in B(\Sigma)}\prod_{r=1}^{\infty}(1-q^{ru_\alpha})^{-1},$$

which by the main theorem of [9] is indeed the constant term of $\prod_{a>0}(1-e^{-a})$. Since the determination of this constant term was the crucial part of the proof of the main theorem of [9], we may regard (3.3) as a "truncated form" of that result.

Remark. Since the number of affine roots $a\in S_k$ with given gradient α is the same as the number with gradient $-\alpha$ (namely k/u_α), it follows that

$$C_k=\sum_{a\in S_k}a$$

is a *constant* function, hence the "highest" term in $\prod_{a\in S_k}(1-e^{-a})$ is $e^{-C_k}=q^{C_k}$. On the other hand, the degree in q of $\prod_{a\in S_k}(1-q^{|a(s_k)|})^{\varepsilon(a)}$ is

$$\sum_{a\in S_k}\varepsilon(a)|a(s_k)|=\sum a(s_k)=C_k(s_k)=C_k.$$

So in (3.3) the two sides have at any rate the same degree and leading coefficient.

When $q=1$, (3.3) is compatible with (2.3).

Finally, as a q-analogue of (2.7), we put forward

CONJECTURE 3.4. *For any (finite) root system R, the constant term in the product*

$$\prod_{\alpha>0}\prod_{i=1}^{k}\left(1-q^{is(\alpha)-1}e^{-\alpha}\right)\left(1-q^{(i-1)s(\alpha)+1}e^{\alpha}\right)$$

should be

$$\prod_{i=1}^{l}\begin{bmatrix}kd_i\\ k\end{bmatrix}$$

where $s(\alpha)$ and the d_i are defined in (2.6).

When R is reduced, this is a restatement of (3.1). When R is of type BC_n, it is not covered by any of the previous conjectures. By (2.7), it is true when $q=1$ and R is classical (including BC_n). It is also true for R of type BC_1 and all integers $k>0$ [R. Askey]; also for R of type BC_2 and $k=1$ (by direct calculation).

B. Mehta's conjecture and generalizations.

4. In his book *Random Matrices* [12], Mehta conjectured that

$$\int_{\mathbb{R}^n} e^{-|x|^2/2}|D(x)|^{2k}dx=(2\pi)^{n/2}\prod_{r=1}^{n}\frac{(rk)!}{k!} \tag{4.1}$$

for any $k\in\mathbb{C}$ with $\mathrm{Re}(k)>0$. (Here, as before, $k!$ means $\Gamma(k+1)$ even if k is not an integer; $dx=dx_1\cdots dx_n$ is Lebesgue measure on $\mathbb{R}^n$; $|x|^2=x_1^2+\cdots+x_n^2$, and $D(x)=\prod_{i<j}(x_i-x_j)$ as in §2.)

The formula (4.1) can be proved by use of Selberg's integral (this observation is due to Bombieri). If we put $x_i=\frac{1}{2}(1+(2a)^{-1/2}y_i)$ $(1\le i\le n)$ in the integral $J_n(a,a;k)$ (§2), where a is real and positive, we obtain

$$\begin{aligned}&\int_{-(2a)^{1/2}}^{(2a)^{1/2}}\cdots\int_{-(2a)^{1/2}}^{(2a)^{1/2}}\prod_{i=1}^{n}\left(1-\frac{y_i^2}{2a}\right)^a|D(y)|^{2k}dy_1\cdots dy_n\\ &\qquad=4^{na}(8a)^{(n(n-1)k+n)/2}J_n(a,a;k)\\ &\qquad=4^{na}(8a)^{(n(n-1)k+n)/2}\prod_{r=1}^{n}\frac{(rk)!(a(r-1)k)!^2}{k!(2a+1+(n+r-2)k)!}\end{aligned} \tag{4.2}$$

by (2.3). Now let $a\to\infty$; since $(1-y_i^2/2a)^a\to e^{-y_i^2/2}$, it is not hard to see that the integral (4.2) tends to the left-hand side of (4.1). The right-hand side of (4.2) can be evaluated by Stirling's formula

$$(a+x)!\sim\sqrt{2\pi}\,e^{-a}a^{a+x+1/2}\quad\text{as } a\to\infty \tag{4.3}$$

which shows that its limit as $a\to\infty$ is $\prod_{r=1}^{n}((rk)!/k!)$, thus proving (4.1).

5. We shall generalize (4.1) as follows. Let G be a finite group of isometries of $\mathbb{R}^n$ generated by reflections in hyperplanes through the origin. The equations of these hyperplanes are of the form $\sum_1^n a_ix_i=0$, which we normalize (up to sign) by taking $\sum a_i^2=2$. Let $P(x)$ be the product of these normalised linear forms, for *all* reflections belonging to G. Thus P is a homogeneous polynomial fuunction on $\mathbb{R}^n$, of degree say N equal to the number of reflections in G.

The group G acts on $\mathbb{R}^n$, hence on the algebra $S(\mathbb{R}^n)$ of polynomial functions on $\mathbb{R}^n$. The G-invariant polynomial functions form an $\mathbb{R}$-algebra $\mathbb{R}[f_1,\cdots,f_n]$ generated by n algebraically independent polynomials f_i. The f_i are not uniquely determined, but their degrees d_i are.

Let γ denote the Gaussian measure on $\mathbb{R}^n$ defined by

$$d\gamma(x)=(2\pi)^{-n/2}e^{-|x|^2/2}dx$$

(so that $\gamma(\mathbb{R}^n)=1$).

CONJECTURE 5.1. *For all complex k with* $\mathrm{Re}(k)>0$,

$$\int_{\mathbb{R}^n}|P(x)|^{2k}d\gamma(x)=\prod_{i=1}^{n}\frac{(kd_i)!}{k!}.$$

Mehta's conjecture (4.1) in the case where $G=S_n$, the symmetric group, acting by permuting the coordinates in $\mathbb{R}^n$, so that $P(x)=\prod_{i<j}(x_i-x_j)=D(x)$ in this case; the invariant polynomials may be taken to be the elementary symmetric functions of $x_1,\cdots,x_n$, so that the d_i are $1,2,\cdots,n$.

The scalar product $\langle x,y\rangle=\sum_{i=1}^n x_i y_i$ on $\mathbb{R}^n$ extends canonically to polynomials in such a way that for any two monomials x^α, x^β

$$\langle x^\alpha, x^\beta\rangle=\delta_{\alpha\beta}\alpha! \quad \text{(Kronecker delta)}$$

(here $\alpha,\beta\in\mathbb{N}^n$ are multi-indices, i.e., x^α means $x_1^{\alpha_1}\cdots x_n^{\alpha_n}$, and $\alpha!=\alpha_1!\cdots\alpha_n!$). An equivalent definition is $\langle f,g\rangle=f(D)g(0)$, where $f(D)$ is the differential operator obtained from f by replacing each x_i by $\partial/\partial x_i$.

For each polynomial $f\in\mathbb{R}[x_1,\cdots,x_n]$, define

$$f^*(x)=\int_{\mathbb{R}^n} f(x-y)\,d\gamma(y)=(f*\gamma)(x).$$

Then [11] we have

$$f^*=e^{\Delta/2}f \tag{5.2}$$

where $\Delta=\sum_1^n \partial^2/\partial x_i^2$ is the Laplacian, and

$$\int_{\mathbb{R}^n} f^*(ix)g^*(-ix)\,d\gamma(x)=\langle f,g\rangle \tag{5.3}$$

for all $f,g\in\mathbb{R}[x_1,\cdots,x_n]$.

Suppose that f is homogeneous, of even degree $2r$. Then by (5.2),

$$\gamma(f)=\int_{\mathbb{R}^n} f(y)\,d\gamma(y)=f^*(0)=(e^{\Delta/2}f)(0)=\frac{\Delta^r f}{2^r r!}.$$

Taking $f=P^{2k}$, where k is now a positive integer, an equivalent version of (5.1) is

CONJECTURE 5.1′. *With the above notation, for each integer* $k\geq 0$

$$\frac{\Delta^{Nk}(P^{2k})}{2^{Nk}(Nk)!}=\prod_{i=1}^n \frac{(kd_i)!}{k!}.$$

(Clearly (5.1′) is equivalent to (5.1) for positive integral values of k, and then for all k by Carlson's theorem [12].)

Consider in particular the case $k=1$ of (5.1). The polynomial P is skew-symmetric for G, and the scalar product (hence also the Laplacian) is G-symmetric, hence ΔP is skew-symmetric: since $\deg\Delta P<\deg P$ we have $\Delta P=0$ and therefore $P^*=P$ by (5.2). Hence (5.3) gives

$$\int_{\mathbb{R}^n} P(x)^2 d\gamma(x)=\langle P,P\rangle. \tag{5.4}$$

Conjecture 5.1 is true in the following cases:

(a) G any *Weyl* group, $k=1$;

(b) G a classical Weyl group (types A,B,D), k arbitrary [A. Regev].

(c) G dihedral, k arbitrary.

(a) Let R be a reduced root system, W its Weyl group. By (5.4) we have to evaluate $\langle P,P\rangle$. The following argument is due to Steinberg: clearly P is proportional to $\Pi=\prod_{\alpha>0}\alpha$, indeed

$$P^2=\frac{\Pi^2}{\prod_{\alpha>0}\left(|\alpha|^2/2\right)}$$

in view of the chosen normalization of P. Hence

$$\langle P,P\rangle=\langle \Pi,\Pi^\vee\rangle. \tag{1}$$

where $\Pi^\vee=\prod_{\alpha>0}\alpha^\vee$.

Now from Weyl's identity

$$\prod_{\alpha>0}(e^{\alpha/2}-e^{-\alpha/2})=\sum_{w\in W}e(w)e^{w\rho}$$

by picking out the terms of degree N, the number of positive roots, on either side we have

$$\Pi=\sum_{w\in W}\varepsilon(w)\frac{(w\rho)^N}{N!}$$

and therefore

$$\langle \Pi,\Pi^\vee\rangle=\sum_{w\in W}\varepsilon(w)\frac{\langle (w\rho)^N,\Pi^\vee\rangle}{N!}. \tag{2}$$

Since $\Pi^\vee$ is skew-symmetric for W, all the terms in this sum are equal to

$$\langle \rho^N,\Pi^\vee\rangle/N!=\prod_{\alpha>0}\langle \rho,\alpha^\vee\rangle=\prod_{\alpha>0}\mathrm{ht}(\alpha^\vee),$$

so that

$$\langle \Pi,\Pi^\vee\rangle=|W|\prod_{\alpha>0}\mathrm{ht}(\alpha^\vee).$$

But [10] the partition formed by the root-heights is the conjugate of the partition formed by the exponents $m_1,\cdots,m_l$, hence

$$\prod_{\alpha>0}\mathrm{ht}(\alpha^\vee)=\prod_{i=1}^{l}m_i!=\prod_{i=1}^{l}(d_i-1)!. \tag{3}$$

Since $|W|=\Pi d_i$, it follows from (1), (2) and (3) that

$$\langle P,P\rangle=\Pi d_i!$$

and hence from (5.4) that

$$\int_{\mathbb{R}^n}P(x)^2d\gamma(x)=\Pi d_i!$$

as required.

(b) In the next section we shall prove a result which contains as special cases Conjecture 5.1 for W of type B_n or D_n. For type A_n, (5.1) reduces (as we have already observed) to Mehta's original conjecture (4.1).

(c) When G is a dihedral group of order $2N$ acting on $\mathbb{R}^2=\mathbb{C}$, we have $P^2=-2^{-N}(z^N-\bar{z}^N)^2$, and the Laplace operator is $\Delta=4\partial^2/\partial z\partial\bar{z}$, whence (5.1′) is easily verified (the d_i being $2,N$). Alternatively, the integral in (5.1) can be computed directly by transforming to polar coordinates.

CONJECTURE 5.5. *The minimum value of* $|x|^2-\log|P(x)|^2$ *is*

$$N(1+\log 2)-\sum_{i=1}^{n}d_i\log d_i.$$

This follows from (5.1) in the same sort of way that (2.8) follows from (2.1). For

$$\max_{x\in\mathbb{R}^n} e^{-|x|^2}P(x)^2 = \lim_{k\to\infty}\left(\int_{\mathbb{R}^n} e^{-k|x|^2}P(x)^{2k}dx\right)^{1/k}$$

$$= \lim_{k\to\infty}(2k)^{-N-n/2k}\left(\int e^{-|y|^2/2}P(y)^{2k}dy\right)^{1/k}$$

$$= \lim_{k\to\infty}(2k)^{-N}\prod_{i=1}^{n}\left(\frac{(kd_i)!}{k!}\right)^{1/k}$$

by (5.1); now use Stirling's formula to evaluate the limit. □

As with (2.8), the points at which this minimum is attached are related to the zeros of classical orthogonal (Hermite, Laguerre) polynomials.

Remarks. 1. As we remarked in §2, the function $|\Delta(t)|^2$ on a maximal torus T occurs in Weyl's integration formula. Likewise $|P(x)|^2$ occurs in the Lie algebra counterpart of this formula, namely

$$\int_{\mathfrak{g}} f(\xi)\,d\xi = \frac{\mu(G/T)}{|W|}\int_{\mathfrak{t}} f(\tau)|\Pi(\tau)|^2 d\tau$$

where f is a G-invariant function on the Lie algebra $\mathfrak{g}$, $d\xi$ and $d\tau$ are Lebesgue measures on $\mathfrak{g},\mathfrak{t}$ respectively and Π is the product of the positive roots of $(\mathfrak{g},\mathfrak{t})$. From this point of view the Mehta conjecture (5.1) is the Lie algebra counterpart of the Dyson conjecture (2.1′).

2. Does there exist a "q-analogue" of (5.1)?

6. A generalization of (4.1) in a slightly different direction goes as follows. Let S be a (not necessarily reduced) root system consisting of linear forms on a real Euclidean space $\mathfrak{a}$. Choose a set S^+ of positive roots in the usual way, and attach to each $\alpha\in S$ a "multiplicity" k_α such that $k_\alpha = k_\beta$ if $|\alpha|=|\beta|$, and define

$$\Pi(x) = \prod_{\alpha\in S^+}|\alpha(x)|^{k_\alpha}, \qquad x\in\mathfrak{a}.$$

Here the k_α need not be integers; they may be complex numbers with real part ≥ 0.

It will avoid extraneous numerical constants and facilitate comparison with §5 if we normalize the linear forms $\alpha\in S$ so that they have norm $\sqrt{2}$, i.e., if we replace $\Pi(x)$ by

$$P(x) = \frac{2^{\nu/2}}{\prod_{\alpha\in S^+}|\alpha|^{k_\alpha}}\cdot\Pi(x)$$

where $\nu = \Sigma_{\alpha\in S^+}k_\alpha$. Let $\rho_k = \frac{1}{2}\Sigma_{\alpha\in S^+}k_\alpha\alpha$, then I conjecture that

CONJECTURE 6.1. *With the above notation,*

$$\int|P(x)|d\gamma(x) = \prod_{\alpha\in S^+}\frac{\left(\frac{1}{2}k_\alpha+\frac{1}{4}k_{\alpha/2}+\frac{1}{2}\langle\rho_k,\alpha^\vee\rangle\right)!}{\left(\frac{1}{4}k_{\alpha/2}+\frac{1}{2}\langle\rho_k,\alpha^\vee\rangle\right)!}$$

where (as in §5) γ is the Gaussian measure on $\mathfrak{a}$, and $\alpha^\vee$ is the coroot $2\alpha/|\alpha|^2$; also $k_{\alpha/2}=0$ if $\frac{1}{2}\alpha\notin S$.

The pattern of the k_α's on the right-hand side in (6.1) is reminiscent of the formula [3] of Gindikin and Karpelevich for Harish-Chandra's c-function.

The evidence in favor of (6.1) is as follows.

(a) It is true for S of classical type (B_n, C_n, D_n, BC_n) by virtue of Selberg's integral, for all values of the "multiplicities" k_α.

(b) It is true when the k_α are the multiplicities m_α of the roots in the restricted root system S of a symmetric space G/K.

(c) When S is reduced (so that $k_{\alpha/2}=0$ for all $\alpha\in S$) and the k_α are all equal, it reduces to (5.1). For if $k_\alpha=2k$ for all α, the right-hand side becomes

$$\prod_{\alpha\in S^+}\frac{(k(1+\mathrm{ht}(\alpha^\vee)))!}{(k\,\mathrm{ht}(\alpha^\vee))!}$$

where $\mathrm{ht}(\alpha^\vee)=\langle\rho,\alpha^\vee\rangle$ is the *height* of the coroot $\alpha^\vee$; and this product is equal to $\prod_{i=1}^{l}(kd_i)!/k!$ by virtue of [10, (2.5)].

(a) Suppose that S is of type BC_n and that the roots are

$$\pm x_i\ (1\le i\le n)\quad\text{with multiplicity } k_1,$$
$$\pm x_i^{\pm}x_j\ (1\le i<j\le n)\quad\text{with multiplicity } k_2,$$
$$\pm 2x_i\ (1\le i\le n)\quad\text{with multiplicity } k_3,$$

where $(x_1,\cdots,x_n)$ is an orthonormal basis of a. This includes the cases where S is of type B_n (resp. C_n, resp. D_n) by taking $k_3=0$ (resp. $k_1=0$, resp. $k_1=k_3=0$).

We have

$$P(x)=\left(2^{n/2}|x_1\cdots x_n|\right)^{k_1+k_3}\prod_{1\le i<j\le n}\left|x_i^2-x_j^2\right|^{k_2}$$

so that if J denotes the integral in (6.1) we have, as in §4,

$$J=\lim_{a\to\infty}(2\pi)^{-n/2}\int_{-(2a)^{1/2}}^{(2a)^{1/2}}\cdots\int_{-(2a)^{1/2}}^{(2a)^{1/2}}P(x)\prod_{i=1}^{n}\left(1-\frac{x_i^2}{2a}\right)^a dx_1\cdots dx_n.$$

In this integral we make the subsitution $y_i=x_i^2/2a$, and we thus obtain

$$J=\lim_{a\to\infty}\left(\frac{a}{\pi}\right)^{n/2}(4a)^{n(k_1+k_3)/2}(2a)^{n(n-1)k_2/2}J_n\left(\tfrac{1}{2}(k_1+k_3-1),a;\tfrac{1}{2}k_2\right)$$
$$=\lim_{a\to\infty}\left(\frac{a}{\pi}\right)^{n/2}(4a)^{n(k_1+k_3)/2}(2a)^{n(n-1)k_2/2}$$
$$\cdot\prod_{r=1}^{n}\frac{(\frac{1}{2}rk_2)!(a+\frac{1}{2}(r-1)k_2)!(\frac{1}{2}(k_1+k_3-1+(r-1)k_2))!}{(\frac{1}{2}k_2)!(a+\frac{1}{2}(k_1+k_3+1+(n+r-2)k_2))!}$$

by Selberg's formula (2.3). Since

$$\frac{(a+x)!}{(a+y)!}\sim a^{x-y}$$

as $a\to\infty$, by Stirling's formula, we obtain

$$J=\frac{2^{n(k_1+k_3+\frac{1}{2}(n-1)k_2)}}{\pi^{n/2}}\prod_{r=1}^{n}\frac{(\frac{1}{2}rk_2)!(\frac{1}{2}(k_1+k_3-1+(r-1)k_2))!}{(\frac{1}{2}k_2)!}$$
$$=\prod_{r=1}^{n}\frac{(\frac{1}{2}rk_2)!(k_1+k_3+(r-1)k_2)!}{(\frac{1}{2}k_2)!(\frac{1}{2}(k_1+k_3+(r-1)k_2))!}$$

by use of the duplication formula for the gamma function.

Now consider the right-hand side of (6.1). We have

$$\rho_k = \sum_{i=1}^{n} \left(\frac{1}{2}k_1 + (n-i)k_2 + k_3 \right) x_i$$

so that the contributions of the roots x_i $(1 \le i \le n)$, $2x_i$ $(1 \le i \le n)$, $x_i - x_j$ $(1 \le i < j \le n)$, $x_i + x_j$ $(1 \le i < j \le n)$ to the product (6.1) are respectively

(1) $$\prod_{i=1}^{n} \frac{(k_1 + (n-i)k_2 + k_3)!}{(\frac{1}{2}k_1 + (n-i)k_2 + k_3)!},$$

(2) $$\prod_{i=1}^{n} \frac{(\frac{1}{2}k_1 + \frac{1}{2}(n-i)k_2 + k_3)!}{(\frac{1}{2}(k_1 + (n-i)k_2 + k_3))!},$$

(3) $$\prod_{1 \le i < j \le n} \frac{(\frac{1}{2}(j-i+1)k_2)!}{(\frac{1}{2}(j-i)k_2)!} = \prod_{i=1}^{n} \frac{(\frac{1}{2}ik_2)}{(\frac{1}{2}k_2)!},$$

(4) $$\prod_{1 \le i < j \le n} \frac{(\frac{1}{2}k_1 + \frac{1}{2}(2n-i-j+1)k_2 + k_3)!}{(\frac{1}{2}k_1 + \frac{1}{2}(2n-i-j)k_2 + k_3)!} = \prod_{i=1}^{n} \frac{(\frac{1}{2}k_1 + \frac{1}{2}(n+i-1)k_2 + k_3)!}{(\frac{1}{2}k_1 + \frac{1}{2}(2i-1)k_2 + k_3)!}.$$

The denominator of (1) and the numerator of (2) cancel out (4), hence the right-hand side of (6.1) is equal to

$$\prod_{i=1}^{n} \frac{(k_1 + (n-1)k_2 + k_3)!(\frac{1}{2}ik_2)!}{(\frac{1}{2}(k_1 + (n-i)k_2 + k_3))!(\frac{1}{2}k_2)!}$$

which is visibly equal to the integral J as calculated above. This establishes (6.1) for S of classical type and arbitrary values of the k_α.

(b) Finally, we shall indicate a proof of (6.1) in the case that S is the restricted root system of a Riemannian symmetric space G/K, and the k_α are the multiplicities m_α of the roots $\alpha \in S$.

Let G be a connected semisimple Lie group with finite center, $\mathfrak{g}$ the Lie algebra of G; let $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$ be a Cartan decomposition and let θ denote the associated Cartan involution of $\mathfrak{g}$ (equal to $+1$ on k and to -1 on p). Then $\langle X, Y \rangle = -B(X, \theta Y)$ (where B is the Killing form on $\mathfrak{g}$) is a positive definite inner product on $\mathfrak{g}$.

Let $\mathfrak{a}$ be a maximal abelian subspace of $\mathfrak{p}$ and let S denote the set of roots of $(\mathfrak{g}, \mathfrak{a})$. Let $S^+ \subset S$ be a set of positive roots and let $\mathfrak{n} = \Sigma_{\alpha \in S^+} \mathfrak{g}^\alpha$, $\bar{\mathfrak{n}} = \Sigma_{\alpha \in S^+} \mathfrak{g}^{-\alpha}$. Then $\mathfrak{g} = \mathfrak{k} + \mathfrak{a} + \mathfrak{n}$ is an Iwasawa decomposition; let $G = KAN$ be the corresponding global decomposition, so that K is a maximal compact subgroup of G. Finally let M, M' be respectively the centralizer and normalizer of $\mathfrak{a}$ in K, so that $W = M'/M$ is the Weyl group of G/K (or of S).

We shall make use of the following two integral formulas, in which all the measures are the Riemannian measures induced by the scalar product $\langle X, Y \rangle$. First, for a function $f \in L^1(\mathfrak{p})$ we have [7, p. 381]

(1) $$\int_{\mathfrak{p}} f(x)\,dx = \frac{1}{|W|} \int_{\mathfrak{a}} \Pi(x) \left(\int_{K/M} f(\mathrm{Ad}(k)x)\,dk_M \right) dx$$

where as before Π is the product of the positive roots $\alpha \in S^+$, each counted with its multiplicity. Next, for a continuous function ϕ on K we have [6, Lemma 44]

(2) $$\int_K \phi(k)\,dk = c \int_{M \times \bar{N}} \phi(mk(\bar{n}))\, e^{-2\rho(H(\bar{n}))}\,dm\,d\bar{n}$$

where $\bar{N}$ is the subgroup of G with Lie algebra $\bar{\mathfrak{n}}$; ρ is as usual half the sum of the positive roots; and for an element $g \in G$, $k(g)$ and $\exp H(g)$ are respectively the K-component and the A-component of g in the Iwasawa decomposition $G = KAN$. In [6], Harish-Chandra normalizes the measures so that $c = 1$; however, in the present situation all measures are those induced by the inner product $\langle X, Y \rangle$ and one can show that

$$c = 2^{\nu/2} \tag{3}$$

where $\nu = \sum_{\alpha \in S^+} m_\alpha = \dim \mathfrak{n}$.

If we take $f(x) = e^{-\langle x,x \rangle/2}$ in (1), the left-hand side of (1) is equal to $(2\pi)^{(l+\nu)/2}$ where $l = \dim \mathfrak{a}$ (since $\dim \mathfrak{p} = \dim \mathfrak{a} + \dim \mathfrak{n} = l + \nu$). Hence from (6.6) we obtain

$$\int_{\mathfrak{a}} \Pi(x)\, d\gamma(x) = \frac{(2\pi)^{\nu/2} |W|}{\operatorname{vol}(K/M)}. \tag{4}$$

Now $\operatorname{vol}(K/M)$ can be computed by taking $\phi = 1$ in (2), which gives (in view of (3))

$$\operatorname{vol}(K/M) = 2^{\nu/2} \int_{\bar{N}} e^{-2\rho(H(\bar{n}))}\, d\bar{n}.$$

This integral may be computed by reduction to the split-rank 1 case, as in [8]. If we follow through the computation there, with our choice of Haar measure $d\bar{n}$, we find that for any $\lambda \in \mathfrak{a}^*$

$$\int_{\bar{N}} e^{-(\lambda+\rho)(H(\bar{n}))}\, d\bar{n} = \frac{(2\pi)^{\nu/2}}{\prod_{\alpha \in S^+} |\alpha|^{m_\alpha}} \prod_{\alpha \in S^+} \frac{\Gamma\left(\frac{1}{4} m_{\alpha/2} + \frac{1}{2} \langle \lambda, \alpha^\vee \rangle\right)}{\Gamma\left(\frac{1}{2} m_\alpha + \frac{1}{4} m_{\alpha/2} + \frac{1}{2} \langle \lambda, \alpha^\vee \rangle\right)}.$$

Hence, taking $\lambda = \rho$, we obtain

$$\operatorname{vol}(K/M) = \frac{2^\nu \pi^{\nu/2}}{\prod_{\alpha \in S^+} |\alpha|^{m_\alpha}} \prod_{\alpha \in S^+} \frac{\Gamma\left(\frac{1}{4} m_{\alpha/2} + \frac{1}{2} \langle \rho, \alpha^\vee \rangle\right)}{\Gamma\left(\frac{1}{2} m_\alpha + \frac{1}{4} m_{\alpha/2} + \frac{1}{2} \langle \rho, \alpha^\vee \rangle\right)}. \tag{5}$$

But also

$$|W| = \prod_{\alpha \in S^+} \frac{m_\alpha + \frac{1}{2} m_{\alpha/2} + \langle \rho, \alpha^\vee \rangle}{\frac{1}{2} m_{\alpha/2} + \langle \rho, \alpha^\vee \rangle}$$

so that (4) and (5) lead to

$$\int_{\mathfrak{a}} \Pi(x)\, d\gamma(x) = 2^{-\nu/2} \prod_{\alpha \in S^+} |\alpha|^{m_\alpha} \prod_{\alpha \in S^+} \frac{\left(\frac{1}{2} m_\alpha + \frac{1}{4} m_{\alpha/2} + \frac{1}{2} \langle \rho, \alpha^\vee \rangle\right)!}{\left(\frac{1}{4} m_{\alpha/2} + \frac{1}{2} \langle \rho, \alpha^\vee \rangle\right)!}$$

which establishes (6.1) in this case.

REFERENCES

[1] G. E. Andrews, *Problems and prospects for basic hypergeometric functions*, in Theory and Applications of Special Functions, R. Askey, ed., Academic Press, New York, 1975.

[2] F. J. Dyson, *Statistical theory of the energy levels of complex systems I*, J. Math. Phys., 3 (1962), pp. 140–156.

[3] S. G. Gindikin and F. I. Karpelevich, *Plancherel measure of Riemannian symmetric spaces of nonpositive curvature*, Soviet Math., 3 (1962), pp. 962–965.

[4] I. J. Good, *Short proof of a conjecture of Dyson*, J. Math. Phys., 11 (1970), p. 1884.

[5] J. GUNSON, *Proof of a conjecture of Dyson in the statistical theory of energy levels*, J. Math. Phys. 3 (1962), pp. 752–753.
[6] HARISH-CHANDRA, *Spherical functions on a semisimple Lie group I*, Amer. J. Math., 80 (1958), pp. 241–310.
[7] S. HELGASON, *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962.
[8] _____, *A duality for symmetric spaces with applications to group representations*, Adv. in Math., 5 (1970), pp. 1–154.
[9] I. G. MACDONALD, *Affine root systems and Dedekind's η-function*, Inv. Math., 15 (1972), pp. 91–143.
[10] _____, *The Poincaré series of a Coxeter group*, Math. Annalen, 199 (1972), pp. 161–174.
[11] _____, *The volume of a compact Lie group*, Inv. Math., 56 (1980), pp. 93–95.
[12] M. L. MEHTA, *Random Matrices*, Academic Press, New York, 1967.
[13] A. SELBERG, *Bemerkninger om et Multiplet Integral*, Norsk Mat. Tidsskrift, 26 (1944), pp. 71–78.
[14] G. SZEGÖ, *Orthogonal Polynomials*, AMS Colloquium Publications 23, American Mathematical Society, Providence, RI, 1975.
[15] K. WILSON, *Proof of a conjecture by Dyson*, J. Math. Phys., 3 (1962), pp. 1040–1043.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0009 $01.00/0

A q-BETA INTEGRAL ASSOCIATED WITH BC_1*

RICHARD ASKEY†

Abstract. A conjecture of I. G. Macdonald [SIAM J. Math. Anal. 13 (1982), pp. 988–1007] about the constant term in a Laurent polynomial expansion associated with the affine root system of BC_1 is extended and then proved by evaluation of an integral that extends a beta function integral.

Macdonald [4] (this issue, pp. 988–1007) has formulated many conjectures about the constant terms in certain Laurent polynomial expansions. For the affine root system of BC_1 he conjectured that

$$\text{C.T.}(qx;q)_k(x^{-1};q)_k(qx^2;q^2)_k(qx^{-2};q^2)_k=\frac{(q;q)_{4k}}{(q;q)_k(q;q)_{3k}} \tag{1}$$

where

$$(a;q)_k=(1-a)(1-aq)\cdots(1-aq^{k-1}), \tag{2}$$

and C.T.$f(x)$ is the constant term in the Laurent expansion of $f(x)$. In the case $q=1$ he proved a result for BC_n [4] using an integral of Selberg [5]. Selberg's integral leads to a more general result where it is possible to separate roots according to their lengths. In the present case the roots are $\pm r$ and $\pm 2r$, so it is natural to try to find

$$f(j,k)=\text{C.T.}(qx;q)_j(x^{-1};q)_j(qx^2;q^2)_k(qx^{-2};q^2)_k. \tag{3}$$

Integration gives

$$f(j,k)=\frac{1}{\pi}\int_0^{\pi}(qe^{2i\theta};q)_j(e^{-2i\theta};q)_j(qe^{4i\theta};q^2)_k(qe^{-4i\theta};q^2)_k d\theta. \tag{4}$$

For the present take $|q|<1$. Ramanujan showed that

$$\sum_{-\infty}^{\infty}\frac{(a;q)_n}{(b;q)_n}t^n=\frac{(at;q)_\infty\left(\frac{q}{at};q\right)_\infty(q;q)_\infty\left(\frac{b}{a};q\right)_\infty}{(t;q)_\infty\left(\frac{b}{at};q\right)_\infty(b;q)_\infty\left(\frac{q}{a};q\right)_\infty}, \qquad \left|\frac{b}{a}\right|<|t|<1, \tag{5}$$

where

$$(a;q)_\infty:=\prod_{n=0}^{\infty}(1-aq^n) \tag{6}$$

and

$$(a;q)_n:=\frac{(a;q)_\infty}{(aq^n;q)_\infty}, \qquad n=0,\pm 1,\cdots. \tag{7}$$

See [1] or [2] for a simple proof of (5).

* Received by the editors July 9, 1980, and in revised form September 29, 1980. This research was supported in part by National Science Foundation under grant MCS-78-07244A02.

† Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

Using Ramanujan's sum in (4) gives

$$f(j,k)=\frac{1}{\pi}\frac{(q^{j+1};q)_\infty^2(q^{2k+2};q^2)_\infty^2}{(q;q)_\infty(q^{2j+1};q)_\infty(q^2;q^2)_\infty(q^{4k+2};q^2)_\infty}$$
$$\sum_{-\infty}^{\infty}\frac{(q^{-j};q)_m}{(q^{j+1};q)_m}q^{jm}\sum_{-\infty}^{\infty}\frac{(q^{-2k};q^2)_n}{(q^{2k+2};q^2)_n}q^{(2k+1)n}\int_0^\pi e^{-2im\theta+4in\theta}\,d\theta$$
$$=\frac{(q;q)_{2j}}{(q;q)_j^2}\frac{(q^2;q^2)_{2k}}{(q^2;q^2)_k^2}\sum_{-\infty}^{\infty}\frac{(q^{-j};q)_{2n}}{(q^{j+1};q)_{2n}}\frac{(q^{-2k};q^2)_n}{(q^{2k+2};q^2)_n}q^{(2j+2k+1)n}.$$

Assume for the moment that $2j>k$. Then the terms in the series vanish when $|n|>k$, so

$$f(j,k)=\frac{(q;q)_{2j}(q^2;q^2)_{2k}}{(q;q)_j^2(q^2;q^2)_k^2}\sum_{n=0}^{2k}\frac{(q^{-j};q)_{2n-2k}(q^{-2k};q^2)_{n-k}}{(q^{j+1};q)_{2n-2k}(q^{2k+2};q^2)_{n-k}}q^{(2j+2k+1)(n-k)}.$$

A bit of simplification using

$$(a;q)_{n-k}=(a;q)_{-k}(aq^{-k};q)_n$$

and

$$(a;q)_{-k}=\frac{1}{(aq^{-k};q)_k}$$

leads to

$$f(j,k)=\frac{(q;q)_{2j}(q^2;q^2)_{2k}(q^{j+1-2k};q)_{2k}q^{-k(2j+2k+1)}}{(q;q)_j^2(q^2;q^2)_k(q^{-j-2k};q)_{2k}(q^{-4k};q^2)_k}$$
$$\cdot{}_3\varphi_2\left(\begin{matrix}q^{-4k},q^{-j-2k},q^{1-j-2k}\\ q^{2+j-2k},q^{1+j-2k}\end{matrix};q^2,q^{2j+2k+1}\right), \tag{8}$$

where

$$p+{}_1\varphi_p\left(\begin{matrix}a_0,\cdots,a_p\\ b_1,\cdots,b_p\end{matrix};q,t\right)=\sum_{n=0}^{\infty}\frac{(a_0;q)_n\cdots(a_p;q)_n}{(b_1;q)_n\cdots(b_p;q)_n}\frac{t^n}{(q;q)_n}.$$

Jackson [3] showed that

$${}_3\varphi_2\left(\begin{matrix}q^{-2n},a,b\\ \dfrac{q^{1-2n}}{a},\dfrac{q^{1-2n}}{b}\end{matrix};q,\frac{q^{1-n}}{ab}\right)=\frac{(q^{n+1};q)_n(abq^n;q)_n}{(aq^n;q)_n(bq^n;q)_n}q^{-n}. \tag{9}$$

Using this in (8) and simplifying, we get

$$f(j,k)=\frac{(q;q)_{2j}(q^2;q^2)_{2k}(q^{2j+1};q^2)_k}{(q;q)_j(q^2;q^2)_k(q;q)_{j+2k}}. \tag{10}$$

Formula (10) can be rewritten in a form closer to the result for BC_n when $q=1$. One such form is

$$f(j,k)=\frac{(q;q)_{2j+2k}(q;q)_{2k}}{(q;q)_{j+2k}(q;q)_{j+k}(q;q)_k}\frac{(-q^{k+1};q)_k}{(-q^{j+1};q)_k}. \tag{11}$$

The restrictions that $2j>k$ and $|q|<1$ can now be removed, since $f(j,k)$ is obviously a polynomial in q. Formula (11) reduces to (1) when $j=k$.

Note added in proof. Further conjectures and identities are contained in W. G. Morris, II, *Constant term identities for finite and affine root systems: conjectures and theorems*. Ph.D. thesis, University of Wisconsin-Madison, January, 1982.

REFERENCES

[1] R. ASKEY, *Ramanujan's extensions of the gamma and beta functions*, Amer. Math. Monthly, 87 (1980), pp. 346–359.

[2] M. E. H. ISMAIL, *A simple proof of Ramanujan's ${}_1\psi_1$*, Proc. Amer. Math. Soc., 63 (1977), pp. 185–186.

[3] F. H. JACKSON, *Certain q-identities*, Quart. J. Math., 12 (1941), pp. 167–172.

[4] I. G. MACDONALD, *Some conjectures for root systems*, this Journal, this issue, pp. 988–1007.

[5] A. SELBERG, *Bemerkninger om et Multipelt Integral*, Norsk Mat. Tidsskr., 26 (1944), pp. 71–78.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0010 $01.00/0

KRAWTCHOUK POLYNOMIALS, A UNIFICATION OF TWO DIFFERENT GROUP THEORETIC INTERPRETATIONS*

TOM H. KOORNWINDER†

Abstract. The canonical matrix elements of irreducible unitary representations of $SU(2)$ are written as Krawtchouk polynomials, with the orthogonality being the row orthogonality for the unitary representation matrix. Dunkl's interpretation of Krawtchouk polynomials as spherical functions on wreath products of symmetric groups is generalized to the case of intertwining functions. A conceptual unification is given of these two group theoretic interpretations of Krawtchouk polynomials.

1. Introduction. Let G be a compact group. Then matrix elements belonging to inequivalent irreducible unitary representations of G are orthogonal to each other. This phenomenon lies in the background of many instances of group theoretic interpretations of orthogonal polynomials. However, if $\pi \in \hat{G}$ and $\pi_{m,n}(g)$ denotes the matrix elements of π with respect to an orthonormal basis then there is also a discrete orthogonality relation for $\pi_{m,n}(g)$ ($g \in G$ fixed) which is just the column or row orthogonality for the unitary matrix $(\pi_{m,n}(g))$. By looking at $\pi_{m,n}(g)$ in this way we may identify it with a quite different system of orthogonal polynomials. For instance, in the case $G = SU(2)$ or $U(2)$ the first kind of orthogonality is a group theoretic form of the orthogonality relations for Jacobi polynomials and the second kind of orthogonality is similarly related to Krawtchouk polynomials. Surprisingly enough, although the first fact is well known, the second fact seems to have been unobserved in literature until now. Section 2 deals with this result.

Krawtchouk polynomials also have a group theoretic interpretation as spherical functions on wreath products of symmetric groups. It is no accident that this class of special functions has two group theoretic interpretations of so different nature. In §3 we give a conceptual proof that, for one special $g \in U(2)$, the corresponding canonical matrix elements can be expressed in terms of spherical functions on the wreath product of S_2 and S_N. A similar explanation can be given for the occurrence of Bessel functions both as generalized matrix elements for discrete series representations of $SL(2,\mathbb{R})$ and as spherical functions for the group of Euclidean motions. Weil's metaplectic representation here plays an important role. These things are shortly discussed in §4.

Not just spherical functions but also interwining functions on wreath products of symmetric groups can be written as Krawtchouk polynomials. This result, which seems to be new, is proved in §5. Finally, in §6 we describe a conceptual way to identify these intertwining functions with matrix elements for $U(2)$, thus generalizing the results of §3.

The interpretation of Krawtchouk polynomials as matrix elements for representations of $SU(2)$ is a suitable point of departure for several different lines of research. Here the author already announces some results, which he intends to publish in subsequent papers. First, the row orthogonality for unitary matrices yields group theoretic interpretations for several other classical orthogonal polynomials, by choosing suitable groups and bases (or double bases) for the representation spaces. We mention Meixner, Laguerre and Pollaczek polynomials for discrete series representations of $SL(2,\mathbb{R})$, Charlier polynomials for the Heisenberg group, Hahn polynomials for $SU(2) \times SU(2)$ (Clebsch–Gordan coefficients), Racah polynomials for $SU(2) \times SU(2) \times SU(2)$

* Received by the editors May 18, 1981, and in revised form October 27, 1981.
† Mathematisch Centrum, P.O. Box 4079, 1009 AB, the Netherlands.

(Racah coefficients). Next, a unification of two different group theoretic interpretations can also be given in the Hahn polynomial case (Clebsch–Gordan coefficients for $SU(2)$ and spherical functions for the symmetric group, respectively). Finally, the group theoretic interpretations of classical orthogonal polynomials mentioned above lead to group theoretic proofs of certain formulas for these polynomials, for instance for the Poisson kernel.

2. The canonical matrix elements of the irreducible unitary representations of $SU(2)$. Consider the natural representation T of $GL(2,\mathbb{C})$ on $\mathbb{C}^2$. The restriction of T to $U(2)$ or $SU(2)$ is a unitary representation, where we consider $\mathbb{C}^2$ as a Hilbert space with respect to the orthonormal basis $e_0 := (1,0)$, $e_1 := (0,1)$.

The N-fold tensor product $\otimes^N T$ of T is a representation of $GL(2,\mathbb{C})$ on $\otimes^N\mathbb{C}^2$. The space V^N of symmetric tensors in $\otimes^N\mathbb{C}^2$ is an invariant subspace of $\otimes^N\mathbb{C}^2$. Let T^N be the corresponding subrepresentation of $\otimes^N T$. A model for V^N is given by the space of all homogeneous polynomials of degree N in two complex variables with $GL(2,\mathbb{C})$ acting on V^N by

$$\begin{aligned} &\left(T^N\begin{pmatrix} a & b \\ c & d\end{pmatrix}F\right)(x,y) := F(ax+cy, bx+dy), \\ &F\in V^N, \quad \begin{pmatrix} a & b \\ c & d\end{pmatrix}\in GL(2,\mathbb{C}), \quad x,y\in\mathbb{C}. \end{aligned} \tag{2.1}$$

The space V^N has dimension $N+1$. A natural basis for V^N is given by the tensors f_n^N $(n=0,1,\cdots,N)$:

(2.2)

$$f_n^N := \frac{1}{N!}\sum_{\sigma\in S_N} e_{i_{\sigma(1)}}\otimes\cdots\otimes e_{i_{\sigma(N)}}, \qquad i_1=\cdots=i_{N-n}=0, \quad i_{N-n+1}=\cdots=i_N=1.$$

We have

$$f_n^N(x,y)=x^{N-n}y^n, \qquad x,y\in\mathbb{C}. \tag{2.3}$$

It follows from (2.2) that the Hilbert space norm of f_n^N is given by

$$\|f_n^N\|^2=\frac{1}{(N!)^2}((N-n)!n!)^2\binom{N}{n}=1\Big/\binom{N}{n},$$

so we have an orthonormal basis

$$e_n^N(x,y) := \binom{N}{n}^{1/2} x^{N-n}y^n, \qquad n=0,1,\cdots,N, \tag{2.4}$$

for V^N. By construction, the restriction of T^N to $U(2)$ or $SU(2)$ is a unitary representation with respect to this orthonormal basis. It is well known (cf., for instance, Hewitt and Ross [6, Thms. (29.20), (29.27)]) that the representations T^N restricted to $SU(2)$ are irreducible and that each unitary irreducible representation of $SU(2)$ is equivalent to some T^N $(N=0,1,2,\cdots)$.

Consider the subgroup

$$K := \left\{u_\theta=\begin{pmatrix} e^{-i\theta} & 0 \\ 0 & e^{i\theta}\end{pmatrix}\right\} \tag{2.5}$$

of $SU(2)$. We have

$$T^N(u_\theta)e_n^N=e^{i(2n-N)\theta}e_n^N, \qquad u_\theta\in K, \tag{2.6}$$

so T^N restricted to K splits as a direct sum of inequivalent irreducible representations of K. We call $\{e_n^N\}$ a *K-basis* for V^N.

For $g \in GL(2,\mathbb{C})$ let

$$T_{m,n}^N(g) := \left(T^N(g)e_n^N, e_m^N\right), \qquad m,n=0,1,\cdots,N, \tag{2.7}$$

where $(\cdot,\cdot)$ is the inner product with respect to the orthonormal basis $\{e_n^N\}$. We call $T_{m,n}^N(g)$ the *canonical matrix elements* of T^N. These matrix elements can be calculated from the generating function

$$\binom{N}{n}^{1/2}(ax+cy)^{N-n}(bx+dy)^n = \sum_{m=0}^{N} T_{m,n}^N\begin{pmatrix} a & b \\ c & d\end{pmatrix}\binom{N}{m}^{1/2} x^{N-m}y^m. \tag{2.8}$$

First of all, we conclude from (2.8):

$$T_{m,n}^N\begin{pmatrix} a & b \\ c & d\end{pmatrix} = T_{N-m,N-n}^N\begin{pmatrix} d & c \\ b & a\end{pmatrix}. \tag{2.9}$$

Binomial expansion of the left-hand side of (2.8) yields

(2.10)

$$T_{m,n}^N\begin{pmatrix} a & b \\ c & d\end{pmatrix} = \binom{N}{n}^{1/2}\binom{N}{m}^{-1/2} a^{N-n-m}b^n c^m \sum_{l=0\vee(m+n-N)}^{m\wedge n} \binom{N-n}{m-l}\binom{n}{l}\left(\frac{ad}{bc}\right)^l.$$

This expression goes back to Wigner [13, (15.21)]. In view of (2.9) we can suppose $m+n \le N$ without loss of generality. Thus it is possible to rewrite (2.10) in terms of the hypergeometric function

$${}_2F_1(a,b;c;z) := \sum_{k=0}^{\infty} \frac{(a)_k(b)_k}{(c)_k k!} z^k, \tag{2.11}$$

where $(a)_k := a(a+1)\cdots(a+k-1)$. In general, the right-hand side of (2.11) is only defined if $|z|<1$ and $c \in \mathbb{C}\setminus\{0,-1,-2,\cdots\}$. However, for $a=-n$, n nonnegative integer, the infinite series in (2.11) terminates:

$${}_2F_1(-n,b;c;z) = \sum_{k=0}^{n} \frac{(-n)_k(b)_k}{(c)_k k!} z^k \tag{2.12}$$

and the right-hand side of (2.12) remains meaningful for all complex z and for all $c \in \mathbb{C}\setminus\{0,-1,\cdots,-n+1\}$.

We obtain from (2.10) and (2.12):

$$\begin{aligned} T_{m,n}^N\begin{pmatrix} a & b \\ c & d\end{pmatrix} &= \left(\frac{m!(N-m)!}{n!(N-n)!}\right)^{1/2}\binom{N-n}{m} \\ &\quad \cdot a^{N-n-m}b^n c^m {}_2F_1\left(-m,-n;N-n-m+1;\frac{ad}{bc}\right), \qquad m+n \le N. \end{aligned} \tag{2.13}$$

Usually, this expression is rewritten in terms of Jacobi polynomials

$$P_n^{(\alpha,\beta)}(x) := \frac{(\alpha+1)_n}{n!}{}_2F_1\left(-n,n+\alpha+\beta+1;\alpha+1;\frac{1-x}{2}\right), \tag{2.14}$$

by the use of the transformation

$${}_2F_1(a,b;c;z) = (1-z)^{-a}{}_2F_1\left(a,c-b;c;\frac{z}{z-1}\right), \tag{2.15}$$

cf. [3,2.1(22)]. Thus (2.13) takes the form

$$T^N_{m,n}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = (-1)^m \left(\frac{m!(N-m)!}{n!(N-n)!} \right)^{1/2} a^{N-n-m} b^{n-m} (ad-bc)^m \cdot P_m^{(N-n-m,n-m)}\left(1-2\frac{ad}{ad-bc}\right), \qquad m+n \le N. \tag{2.16}$$

For $\alpha, \beta > -1$, Jacobi polynomials are orthogonal polynomials on the interval $(-1,1)$ with respect to the weight function $(1-x)^\alpha (1+x)^\beta$. For integer α, β this orthogonality property can be derived from (2.16) combined with Schur's orthogonality relations on $SU(2)$ and an expression for the Haar measure on $SU(2)$ in terms of suitable coordinates. The observation that the $T^N_{m,n}$'s can be written in terms of Jacobi polynomials, goes probably back to Gelfand and Šapiro [4, p. 280].

However, we may also transform (2.13) by means of the formula

$${}_2F_1(-n,b;c;z) = \frac{(c-b)_n}{(c)_n} {}_2F_1(-n,b;b-c-n+1;1-z), \qquad n = 0,1,2,\cdots, c-b,\ c \neq 0,-1,\cdots,-n+1, \tag{2.17}$$

(cf. [3,10.8(13)] together with (2.14)). Then we obtain

$$\begin{aligned} T^N_{m,n}\begin{pmatrix} a & b \\ c & d \end{pmatrix} &= \binom{N}{m}^{1/2} \binom{N}{n}^{1/2} a^{N-n-m} b^n c^m {}_2F_1\left(-m,-n;-N;\frac{bc-ad}{bc}\right) \\ &= \binom{N}{m}^{1/2} \binom{N}{n}^{1/2} a^{m+n-N} b^{N-m} c^{N-n} \\ &\quad \cdot {}_2F_1\left(-(N-m),-(N-n);-N;\frac{bc-ad}{bc}\right), \end{aligned} \tag{2.18}$$

where the second identity follows from

$${}_2F_1(a,b;c;z) = (1-z)^{c-a-b} {}_2F_1(c-a,c-b;c;z) \tag{2.19}$$

(cf. [3,2.1(23)]). Thus we have proved (2.18) for $m+n \le N$, but, in view of (2.9), the formula remains valid without this restriction.

For $N = 0,1,\cdots$, $n = 0,1,\cdots,N$ and $p \in \mathbb{C} \setminus \{0\}$ the *Krawtchouk polynomial* $K_n(x;p,N)$ is defined by

$$K_n(x;p,N) := {}_2F_1(-n,-x;-N;p^{-1}). \tag{2.20}$$

By (2.12) this is a polynomial in x of degree n. For $0<p<1$ Krawtchouk polynomials are orthogonal polynomials on the set $\{0,1,\cdots,N\}$ with respect to the binomial distribution:

$$\sum_{x=0}^{N} K_m(x;p,N) K_n(x;p,N) \binom{N}{x} p^x (1-p)^{N-x} = \left(\binom{N}{n} \left(\frac{p}{1-p} \right)^n \right)^{-1} \delta_{m,n} \tag{2.21}$$

(cf. Szegö [10, §2.82]; we follow the modern notation as used in Askey [1, (2.41)]).

It follows from (2.18) and (2.20) that

$$T^N_{m,n}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \binom{N}{m}^{1/2} \binom{N}{n}^{1/2} a^{N-n-m} b^n c^m K_m\left(n;\frac{bc}{bc-ad},N\right). \tag{2.22}$$

In particular, put

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} := \begin{pmatrix} \cos\psi & \sin\psi \\ \sin\psi & -\cos\psi \end{pmatrix}, \qquad 0 < \psi < \pi,$$

which is in $U(2)$. Then

(2.23)

$$T^N_{m,n}\begin{pmatrix}\cos\psi & \sin\psi\\ \sin\psi & -\cos\psi\end{pmatrix}=\binom{N}{m}^{1/2}\binom{N}{n}^{1/2}(\cos\psi)^{N-m-n}(\sin\psi)^{m+n}K_m(n;\sin^2\psi,N).$$

Thus, for each value of the parameter $p\in(0,1)$ and for each N we can realize the Krawtchouk polynomials $K_n(x;p,N)$ in terms of the canonical matrix elements of the representation T^N of $U(2)$. Furthermore, the left-hand side of (2.23) being a unitary matrix, the row orthogonality

$$\sum_{n=0}^{N} T^N_{m,n}\begin{pmatrix}\cos\psi & \sin\psi\\ \sin\psi & -\cos\psi\end{pmatrix}\overline{T^N_{m',n}\begin{pmatrix}\cos\psi & \sin\psi\\ \sin\psi & -\cos\psi\end{pmatrix}}=\delta_{m,m'} \tag{2.24}$$

just yields the orthogonality relations (2.21) for Krawtchouk polynomials. This also holds for the column orthogonality, since

$$K_m(n;\sin^2\psi,N)=K_n(m;\sin^2\psi,N)$$

(cf. (2.20)).

3. Identification of spherical functions on a Hamming scheme over an alphabet of two letters with canonical matrix elements for $SU(2)$. Consider the Abelian group of two elements $F:=\{0,1\}=\mathbb{Z}\pmod 2$ and its N-fold direct product F^N ($N=1,2,\cdots$). Write elements of F^N as $x=(x_1,\cdots,x_N)$, $x_i\in F$. The space of all complex-valued functions on F^N becomes a Hilbert space $L^2(F^N)$, where the inner product is taken with respect to the normalized Haar measure on F^N:

$$(f,g):=2^{-N}\sum_{x\in F^N} f(x)\overline{g(x)},\qquad f,g\in L^2(F^N). \tag{3.1}$$

Note that $L^2(F^N)$ can be identified with the tensor product $\otimes^N L^2(F)$.

The characters on F are χ_0 and χ_1, defined by

$$\chi_0(x):=1,\qquad \chi_1(x):=(-1)^x,\qquad x\in F, \tag{3.2}$$

and they form an orthonormal basis of $L^2(F)$. The characters on F^N are

$$\chi_y:=\chi_{y_1}\otimes\chi_{y_2}\otimes\cdots\otimes\chi_{y_N},\qquad y=(y_1,\cdots,y_N)\in F^N, \tag{3.3}$$

i.e.,

$$\begin{aligned}\chi_y(x)&=\chi_{y_1}(x_1)\chi_{y_2}(x_2)\cdots\chi_{y_N}(x_N)\\ &=(-1)^{x_1y_1+x_2y_2+\cdots+x_Ny_N},\qquad x,y\in F^N,\end{aligned}$$

and they form an orthonormal basis of $L^2(F^N)$. Since $\chi_y\chi_{y'}=\chi_{y+y'}$, $y,y'\in F^N$, the dual group of F^N can be identified with F^N. The *Fourier transform* $\mathfrak{F}$ on $L^2(F^N)$ is given by

$$(\mathfrak{F}f)(y):=2^{-N/2}\sum_{x\in F^N} f(x)\chi_y(x), \tag{3.4}$$

where we chose the constant $2^{-N/2}$ such that $\mathfrak{F}$ is a unitary transformation from $L^2(F^N)$ onto itself.

The symmetric group S_N acts as a group of automorphisms on F^N by

$$\begin{aligned}\sigma(x_1,\cdots,x_N)&:=(x_{\sigma^{-1}(1)},\cdots,x_{\sigma^{-1}(N)}),\\ &(x_1,\cdots,x_N)\in F^N,\qquad \sigma\in S_N.\end{aligned} \tag{3.5}$$

Let G be the semidirect product $F^N \circ S_N$ corresponding to this action. Then F^N can be identified with the homogeneous space G/S_N. This homogeneous space is called a *Hamming scheme* over the alphabet F of two letters. The terminology stems from coding theory (cf., for instance, MacWilliams and Sloane [7, Ch. 21, §3]). Let λ be the regular representation of G on $L^2(F^N)$, i.e.,

$$(\lambda(0,\sigma)f)(x)=f(\sigma^{-1}x), \qquad \sigma\in S_N, \tag{3.6}$$
$$(\lambda(y,\mathrm{id})f)(x)=f(x-y), \qquad y\in F^N,$$

where $f\in L^2(F^N)$, $x\in F^N$. Then

$$\lambda(0,\sigma)\mathcal{F}=\mathcal{F}\lambda(0,\sigma), \qquad \sigma\in S^N. \tag{3.7}$$

Hence, if $f\in L^2(F^N)$ is symmetric in $x_1,\cdots,x_N$ then $\mathcal{F}f$ is symmetric and

$$(\mathcal{F}f)(y)=2^{-N/2}\sum_{x\in F^N} f(x)\chi_y(x)=2^{-N/2}\sum_{x\in F^N} f(x)\left(\frac{1}{N!}\sum_{\sigma\in S^N}\chi_y(\sigma x)\right). \tag{3.8}$$

The *Hamming distance* on F^N is defined by

$$d(x,y):=|\{i\,|\,x_i\neq y_i\}|, \qquad x,y\in F^N. \tag{3.9}$$

It is translation invariant. The symmetric functions in $x\in F^N$ are just the functions which only depend on $d(x,0)$. The expression

$$\frac{1}{N!}\sum_{\sigma\in S_N}\chi_y(\sigma x)$$

occurring in (3.8) is symmetric both in x and y. Hence, for $n=0,1,\cdots,N$, we can define functions ϕ_n^N on F^N and $\tilde{\phi}_n^N$ on $\{0,1,\cdots,N\}$ such that

$$\tilde{\phi}^N_{d(y,0)}(d(x,0))=\phi^N_{d(y,0)}(x):=\frac{1}{N!}\sum_{\sigma\in S_N}\chi_y(\sigma x), \qquad x,y\in F^N. \tag{3.10}$$

Note the similarity between these functions on the one hand and the Bessel functions in connection with the Fourier transform of rotation invariant functions on $\mathbb{R}^n$ on the other hand. In fact, the functions ϕ_n^N are the *spherical functions* on F^N with respect to the group G, i.e.:

PROPOSITION 3.1. *$L^2(F^N)$ is an orthogonal direct sum of G-invariant subspaces $\mathcal{H}_n^N$, $n=0,1,\cdots,N$, where $\mathcal{H}_n^N:=\mathrm{span}\{\chi_y|d(y,0)=n\}$. In each subspace $\mathcal{H}_n^N$ there is a unique S_N-invariant function which takes the value 1 in 0, namely the function ϕ_n^N. The subspaces $\mathcal{H}_n^N$ are irreducible under the action of G.* The proof is immediate, by the use of (3.10). Note that a similar proposition more generally holds for finite semidirect product groups $A\circ H$ with Abelian normal subgroup A.

It follows from (3.10) and (2.17) that

$$\begin{aligned}\tilde{\phi}_n^N(m)&=\frac{n!(N-n)!}{N!}\sum_{k=0}^{n\wedge m}(-1)^k\binom{m}{k}\binom{N-m}{n-k}\\ &=\frac{(N-m)!(N-n)!}{N!(N-m-n)!}\,{}_2F_1(-n,-m;N-m-n+1;-1)\\ &={}_2F_1(-n,-m;-N;2)\end{aligned}$$

(Although the third expression is not meaningful if $m+n>N$, the equality of the second and fourth expression follows by continuity, because the second and third

equality are valid for all noninteger numbers N.) Hence by (2.20):

$$\tilde{\phi}_n^N(m) = K_n(m; \tfrac{1}{2}, N), \tag{3.11}$$

i.e., the spherical functions on F^N are Krawtchouk polynomials of order $p=\frac{1}{2}$. This result goes back to Vere-Jones [11].

On comparing (2.23) and (3.11) we find that

$$T_{m,n}^N\begin{pmatrix} 2^{-1/2} & 2^{-1/2} \\ 2^{-1/2} & -2^{-1/2} \end{pmatrix} = 2^{-1/2N}\binom{N}{m}^{1/2}\binom{N}{n}^{1/2}\tilde{\phi}_m^N(n). \tag{3.12}$$

We will give now another, more intrinsic proof of this relation, only using the group theoretic characterization of $T_{m,n}^N$, and $\tilde{\phi}_n^N$ and not any a priori knowledge that they can be expressed in terms of Krawtchouk polynomials.

Consider the natural action of $U(2)$ on $L^2(F)$ with respect to the basis χ_0, χ_1 of $L^2(F)$, i.e., if $T=\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)\in U(2)$ then

$$T\chi_0 = a\chi_0 + c\chi_1, \qquad T\chi_1 = b\chi_0 + d\chi_1.$$

This yields a unitary action of $U(2)$ on $L^2(F^N) = \otimes^N L^2(F)$, which commutes with the action of S_N on $L^2(F^N)$. Hence the space $L^2(S_N\backslash F^N)$ of symmetric functions on F^N is invariant under the action of $U(2)$. We can make the following identification between the concepts from §§2 and 3, respectively:

$$\begin{array}{lcl} U(2)\text{-module}\,\mathbb{C}^2 & \leftrightarrow & U(2)\text{-module}\,L^2(F), \\ \{e_0, e_1\} & \leftrightarrow & \{\chi_0, \chi_1\}, \\ U(2)\text{-module}\,\otimes^N\mathbb{C}^2 & \leftrightarrow & U(2)\text{-module}\,L^2(F^N), \\ U(2)\text{-module}\,V^N & \leftrightarrow & U(2)\text{-module}\,L^2(S_N\backslash F^N), \\ f_n^N & \leftrightarrow & \phi_n^N, \\ e_n^N & \leftrightarrow & \binom{N}{n}^{1/2}\phi_n^N. \end{array} \tag{3.13}$$

Now the crucial point is to identify the Fourier transform $\mathcal{F}$ with the action of a certain element in $U(2)$. Consider first the Fourier transform acting on $f\in L^2(F)$:

$$\begin{aligned} (\mathcal{F}f)(x) &= 2^{-1/2}(f(0)\chi_x(0) + f(1)\chi_x(1)) \\ &= 2^{-1/2}(f(0)\chi_0(x) + f(1)\chi_1(x)). \end{aligned}$$

Hence

$$\mathcal{F}\chi_0 = 2^{-1/2}(\chi_0 + \chi_1), \qquad \mathcal{F}\chi_1 = 2^{-1/2}(\chi_0 - \chi_1),$$

i.e., $\mathcal{F}$ corresponds with the unitary matrix

$$s_0 := \begin{pmatrix} 2^{-1/2} & 2^{-1/2} \\ 2^{-1/2} & -2^{-1/2} \end{pmatrix}. \tag{3.14}$$

Since $\mathcal{F}$ acting on $L^2(F^N)$ is the N-fold tensor product of $\mathcal{F}$ acting on $L^2(F)$, this correspondence is also valid on $L^2(F^N)$:

$$T^N(s_0) \leftrightarrow \mathcal{F} \text{ acting on } L^2(S_N\backslash F^N). \tag{3.15}$$

It follows from (3.13), (3.15) and (2.7) that

$$\binom{N}{n}^{1/2}(\mathcal{F}\phi_n^N)(x) = \sum_{m=0}^{N} T_{m,n}^N(s_0)\binom{N}{m}^{1/2}\phi_m^N(x). \tag{3.16}$$

The left-hand side of (3.16) can be evaluated by means of (3.10) and (3.4):

$$\begin{aligned}(\mathcal{F}\phi_n^N)(x)&=\frac{n!(N-n)!}{N!}\sum_{d(y,0)=n}(\mathcal{F}\chi_y)(x)\\&=\frac{n!(N-n)!}{N!}\sum_{d(y,0)=n}2^{-N/2}\sum_{z\in F^N}\chi_y(z)\chi_x(z)\\&=\frac{2^{N/2}n!(N-n)!}{N!}\delta_{d(x,0),n}.\end{aligned}$$

Hence

$$\sum_{m=0}^{N}T_{m,n}^N(s_0)\binom{N}{m}^{1/2}\tilde{\phi}_m^N(l)=2^{N/2}\binom{N}{n}^{-1/2}\delta_{l,n}. \tag{3.17}$$

Now multiply both sides of (3.17) with $T_{p,n}^N(s_0)$, sum over p and use that $T_{m,n}^N(s_0)$ is a unitary matrix with real entries (by (2.8)). It follows that

$$\binom{N}{p}^{1/2}\tilde{\phi}_p^N(l)=2^{N/2}\binom{N}{l}^{-1/2}T_{p,l}^N(s_0).$$

This settles (3.12).

4. Connection with the metaplectic representation of $SL(2,\mathbb{R})$. Let us put the results of §3 in a more general framework. Let F be a locally compact Abelian group and let F be isomorphic to the dual group F^* via the isomorphism $y\to\chi_y$. Let G be a locally compact group and let π be a unitary representation of G on $L^2(F)$ with the following properties:

(i) For some $s_0\in G$, $\pi(s_0)$ is the Fourier transform $\mathcal{F}$ on $L^2(F)$.

(ii) For some closed subgroup H of G there is a function c on $H\times F$ such that

$$(\pi(h)f)(x)=c(h,x)f(x),\qquad f\in L^2(F),\quad x\in F,\quad h\in H,$$

and

$$c(h,x)=c(h,y)\quad\text{for all } h\in H\Rightarrow x=y.$$

Then the Dirac measures on F form a (generalized) H-basis for $L^2(F)$ and $\pi(s_0)$ has (generalized) canonical matrix elements $(x,y)\to\chi_y(x)$ with respect to this basis.

Next suppose that for each natural number N there is a compact group K_N of automorphisms of F^N such that:

(i) K_N acting on $L^2(F^N)$ commutes with $\otimes^N\pi(G)$.

(ii) If $c(h,x_1)\cdots c(h,x_N)=c(h,y_1)\cdots c(h,y_N)$ for all $h\in H$ then $(x_1,\cdots,x_N)$ and $(y_1,\cdots,y_N)$ are in the same K_N-orbit. Let $\mathcal{H}_N$ be the subspace of $L^2(F^N)$ consisting of K_N-invariant functions.

Let π_N be the corresponding subrepresentation of $\otimes^N\pi$. Write $\tilde{x}$ for the K_N-orbit through $x\in F^N$ and put

$$\tilde{\phi}_{\tilde{y}}(\tilde{x})=\phi_{\tilde{y}}(x):=\int_{K_N}\chi_{y_1}(k\cdot x_1)\cdots\chi_{y_N}(k\cdot x_N)\,dk,\qquad x,y\in F^N.$$

Then $\phi_{\tilde{y}}$ is a spherical function on $F^N\circ K_N/K_N$. Now the Dirac measures on $K_N\backslash F^N$ form a (generalized) H-basis for $\mathcal{H}_N$ and $\pi_N(s_0)$ has (generalized) canonical matrix elements $(\tilde{x},\tilde{y})\to\tilde{\phi}_{\tilde{y}}(\tilde{x})$.

In §3 we had $F=\{0,1\}$, $G=U(2)$, H is the subgroup of diagonal elements, π is the natural representation of $U(2)$ on $L^2(\{0,1\})$, K_N is the symmetric group and the spherical functions $\tilde{\phi}_{\tilde{y}}(\tilde{x})$ were Krawtchouk polynomials.

For another example let $F=\mathbb{R}$, G a two-fold covering group of $SL(2,\mathbb{R})$ and π the metaplectic representation of G on $L^2(\mathbb{R})$ (cf. Weil [12] for the definition). Let K_N be the rotation group $SO(N)$. Then the functions $\tilde{\phi}_{\tilde{y}}(\tilde{x})$ can be expressed in terms of Bessel functions and the representations π_N are irreducible and belong to the discrete series. Thus we have a conceptual interpretation that discrete series representations of $SL(2,\mathbb{R})$ or its covering groups are related to the Hankel transform. See Sally [9], Gross and Kunze [5] and Rallis and Schiffmann [8] for further information.

The analogy between the cases $F=\{0,1\}$ and $F=\mathbb{R}$ is not perfect, since $U(2)$ is not contained in the metaplectic group related to $\{0,1\}$, so π is not a metaplectic representation in this case.

5. Krawtchouk polynomials as intertwining functions on Hamming schemes. Let G be the wreath product $(S_{k+1})^N \circ S_N$ $(k,N\in\mathbb{N})$, i.e., the semidirect product of $(S_{k+1})^N$ and S_N with S^N acting on $(S_{k+1})^N$ by

$$\tau\cdot(\sigma_1,\cdots,\sigma_N) := (\sigma_{\tau^{-1}(1)},\cdots,\sigma_{\tau^{-1}(N)}), \qquad \tau\in S_N,\quad \sigma_1,\cdots,\sigma_N\in S_{k+1}.$$

Let $X := \{0,1,\cdots,k\}$. G acts transitively on X^N by

$$(\sigma_1,\cdots,\sigma_N)(x_1,\cdots,x_N) := (\sigma_1 x_1,\cdots,\sigma_N x_N), \qquad \sigma_i\in S^{k+1},\quad x_i\in X,$$

and

$$\tau(x_1,\cdots,x_N) := (x_{\tau^{-1}(1)},\cdots,x_{\tau^{-1}(N)}), \qquad \tau\in S^N,\quad x_i\in X.$$

Let S_k denote the stabilizer of $0\in X$ in S_{k+1}. Put $0 := (0,0,\cdots,0)\in X^N$. Then the stabilizer K of $0\in X^N$ in G equals $(S_k)^N \circ S_N$. The homogeneous space $X^N = G/K$ is called a *Hamming scheme* over the alphabet X of $k+1$ letters.

Fix an integer such that $0\le q\le k-1$. Let $S_{q+1}\times S_{k-q}$ denote the stabilizer of the subset $\{0,1,\cdots,q\}$ of X in S_k. Let $L^2(X)$ be the space of all complex-valued functions on X provided with the inner product

$$(f,g) := (k+1)^{-1}\sum_{x\in F} f(x)\overline{g(x)}, \qquad f,g\in L^2(X).$$

Let $\chi_0,\chi_1,\cdots,\chi_k$ be an orthonormal basis of $L^2(X)$ such that

$$\chi_0(x) := 1, \qquad x\in X, \tag{5.1}$$

$$\chi_1(x) := \begin{cases} \left(\dfrac{k-q}{q+1}\right)^{1/2}, & x=0,\cdots,q, \\ -\left(\dfrac{q+1}{k-q}\right)^{1/2}, & x=q+1,\cdots,k. \end{cases} \tag{5.2}$$

Note that χ_1 is $S_{q+1}\times S_{k-q}$-invariant.

The Hilbert space $L^2(X^N)$, provided with the inner product

$$(f,g) := (k+1)^{-N}\sum_{x\in X^N} f(x)\overline{g(x)}, \qquad f,g\in L^2(X^N),$$

can be identified with the tensor product $\otimes^N L^2(X)$. Put

$$\chi_y(x) := \chi_{y_1}(x_1)\cdots\chi_{y_N}(x_N), \tag{5.3}$$

$$x=(x_1,\cdots,x_N)\in X^N, \qquad y=(y_1,\cdots,y_N)\in X^N.$$

Then the functions $\chi_{y_N}(y\in X^N)$ form an orthonormal basis of $L^2(X^N)$. The Hamming distance on X^N is defined by

$$d(x,y) := |\{i \mid x_i \neq y_i\}|, \qquad x,y\in X^N. \tag{5.4}$$

PROPOSITION 5.1. (a) *$L^2(X^N)$ is an orthogonal direct sum of G-invariant subspaces $\mathcal{H}_n^N$, $n=0,1,\cdots,N$, where*

$$\mathcal{H}_n^N := \operatorname{span}\{\chi_y \mid d(y,0)=n\}. \tag{5.5}$$

(b) *Each space $\mathcal{H}_n^N$ contains a unique function $\phi_n^{N,q}$ which* (i) *is invariant under the subgroup $H := (S_{q+1}\times S_{k-q})^N \circ S_N$ of G and* (ii) *takes the value* 1 *in $0\in X^N$.*

(c) *The spaces $\mathcal{H}_n^N$ are irreducible under G.*

Proof. Part (a) is evident. For the proof of (b) let $f\in\mathcal{H}_n^N$ satisfy (i). Then f is a linear combination of functions of the form

$$x\to\chi_1(x_{i_1})\chi_1(x_{i_2})\cdots\chi_1(x_{i_n}), \qquad i\le i_1<i_2<\cdots<i_n\le N,$$

because of the $(S_{q+1}\times S_{k-q})^N$-invariance. By S_N-invariance we get

$$f(x)=\frac{C}{N!}\sum_{\tau\in S_N}\chi_1(x_{\tau(1)})\cdots\chi_1(x_{\tau(n)})$$

for some constant C. If f also satisfies (ii) then

$$1=C\left(\frac{k-q}{q+1}\right)^{n/2}$$

Hence (b) holds with

$$\phi_n^{N,q}(x)=\left(\frac{q+1}{k-q}\right)^{n/2}\cdot\frac{1}{N!}\sum_{\tau\in S_N}\chi_1(x_{\tau(1)})\cdots\chi_1(x_{\tau(n)}). \tag{5.6}$$

Finally (c) follows from the case $q=0$ (i.e., $K=H$) of (b). □

The functions $\phi_n^{N,q}$ are called *intertwining functions* because the G-intertwining operators from $L^2(G/K)$ into $L^2(G/H)$ can be written in terms of these functions. The functions $\phi_n^{N,0}$ are called the *spherical functions* on the homogeneous space G/K.

By using (5.6) we can evaluate the intertwining functions in terms of special functions. First note that an H-invariant function on X^N only depends on

$$\tilde{x} := |\{i \mid q+1\le x_i\le k\}|, \qquad x\in X^N. \tag{5.7}$$

Write

$$\tilde{\phi}_n^{N,q}(\tilde{x}) := \phi_n^{N,q}(x). \tag{5.8}$$

It follows from (5.6), (5.2) and (2.17) that

$$\begin{aligned}\tilde{\phi}_n^{N,q}(m)&=\frac{n!(N-n)!}{N!}\left(\frac{q+1}{k-q}\right)^{n/2}\sum_{k=0}^{n\wedge m}\binom{m}{k}\binom{N-m}{n-k}\left(\frac{k-q}{q+1}\right)^{(n-k)/2}\left(\frac{q+1}{k-q}\right)^{k/2}(-1)^k\\&=\frac{(N-m)!(N-n)!}{N!(N-m-n)!}\,{}_2F_1\left(-n,-m;N-m-n+1;-\frac{q+1}{k-q}\right)\\&={}_2F_1\left(-n,-m;-N;\frac{k+1}{k-q}\right),\end{aligned}$$

where the equality of the second and fourth expression for $m+n>N$ follows by continuity, starting with noninteger N. Hence, by (2.20):

$$\tilde{\phi}_n^{N,q}(m)=K_n\left(m;\frac{k-q}{k+1},N\right). \tag{5.9}$$

The spherical function case $q=0$ of (5.9) is due to Dunkl [2]. The general case is probably new. Note that the set $\{(k-q)/(q+1)|0\leq q\leq k-1,\ k=1,2,\cdots\}$ is just the set of rational numbers between 0 and 1. R. Askey suggested to me that Krawtchouk polynomials of rational order might have a group theoretic interpretation as intertwining functions.

6. The connection between two different group theoretic interpretations of Krawtchouk polynomials of general order. Let F be the set $\{0,1\}$. Fix $0<p<1$ and let w be the weight function on F given by

$$w(0):=1-p,\qquad w(1):=p. \tag{6.1}$$

Let $L^2(F;w)$ be the space of complex-valued functions on F with inner product

$$(f,g):=\sum_{x\in F} f(x)\overline{g(x)}w(x),\qquad f,g\in L^2(F;w).$$

We will now extend the results of §3 to the case of this weighted L^2-space. Let N be a natural number. Let

$$W(x):=w(x_1)w(x_2)\cdots w(x_N),\qquad x\in F^N. \tag{6.2}$$

Then $L^2(F^N;W)=\otimes^N L^2(F;w)$. Let

$$\chi_0(x):=1,\qquad x\in F, \tag{6.3}$$

$$\chi_1(x):=\begin{cases}\left(\frac{p}{1-p}\right)^{1/2}, & x=0,\\ -\left(\frac{1-p}{p}\right)^{1/2}, & x=1.\end{cases} \tag{6.4}$$

Then $\{\chi_0,\chi_1\}$ is an orthogonal basis for $L^2(F;w)$ and the functions χ_y ($y\in F^N$), given by

$$\chi_y(x)=\chi_{y_1}(x_1)\cdots\chi_{y_N}(x_N),\qquad x\in F^N, \tag{6.5}$$

form an orthogonal basis of $L^2(F^N,W)$. By symmetrization of the basis functions (6.5) we obtain a basis for the symmetric functions in $L^2(F^N,W)$:

$$\tilde{\psi}_{d(y,0)}^{N,p}(d(x,0))=\psi_{d(y,0)}^{N,p}(x):=\frac{1}{N!}\sum_{\sigma\in S_N}\frac{\chi_y(\sigma x)}{\chi_y(0)},\qquad x\in F^N, \tag{6.6}$$

where the Hamming distance d on F^N is defined by (3.9). It follows from §5 that the intertwining functions $\phi_n^{N,q}$ are special cases of (6.6):

$$\tilde{\phi}_n^{N,q}=\tilde{\psi}_n^{N,(k-q)/(k+1)}. \tag{6.7}$$

There is a natural unitary action of $U(2)$ on $L^2(F;w)$ with respect to the basis χ_0,χ_1, just as in §3. Via the tensor product this yields a unitary action of $U(2)$ on $L^2(F^N,W)$, which commutes with the action of S_N on $L^2(F^N,W)$. Thus, similarly to

(3.13), we can make an identification between concepts from §§2 and 6, respectively:

$$
\begin{aligned}
&U(2)\text{-module}\,\mathbb{C}^2 && \leftrightarrow && U(2)\text{-module}\,L^2(F;w),\\
&\{e_0,e_1\} && \leftrightarrow && \{\chi_0,\chi_1\},\\
&U(2)\text{-module}\,\otimes^N\mathbb{C}^2 && \leftrightarrow && U(2)\text{-module}\,L^2(F^N;W),\\
&U(2)\text{-module}\,V^N && \leftrightarrow && U(2)\text{-module}\,L^2\big(S_N\backslash F^N;W\big),\\
&f_n^N && \leftrightarrow && \left(\frac{p}{1-p}\right)^{n/2}\psi_n^{N,p},\\
&e_n^N && \leftrightarrow && \binom{N}{n}^{1/2}\left(\frac{p}{1-p}\right)^{n/2}\psi_n^{N,p}.
\end{aligned}
\tag{6.8}
$$

The "Fourier" transform $\mathcal{F}$ on $L^2(F;w)$, defined by

$$
(\mathcal{F}f)(y)=\frac{1}{w(y)^{1/2}}\sum_{x\in F}f(x)\chi_y(x)w(x),
\tag{6.9}
$$

is clearly a unitary transformation from $L^2(F;w)$ onto itself. A calculation using (6.1), (6.3), (6.4) shows that this unitary transformation is given by the matrix

$$
s_p:=\begin{pmatrix}(1-p)^{1/2} & p^{1/2}\\ p^{1/2} & -(1-p)^{1/2}\end{pmatrix}.
\tag{6.10}
$$

Let $\mathcal{F}$ acting on $L^2(F^N;W)$ be defined as the N-fold tensor product of $\mathcal{F}$ acting on $L^2(F;w)$. Then

$$
(\mathcal{F}f)(y)=\frac{1}{W(y)^{1/2}}\sum_{x\in F^N}f(x)\chi_y(x)W(x).
\tag{6.11}
$$

Just as in (3.15) we have the correspondence

$$
T^N(s_p)\leftrightarrow\mathcal{F}\text{ acting on }L^2\big(S_N\backslash F^N;W\big).
\tag{6.12}
$$

It follows from (6.8), (6.12) and (2.7) that

$$
\binom{N}{n}^{1/2}\left(\frac{p}{1-p}\right)^{n/2}(\mathcal{F}\psi_n^{N,p})(x)=\sum_{m=0}^{N}T_{m,n}^N(s_p)\binom{N}{m}^{1/2}\left(\frac{p}{1-p}\right)^{m/2}\psi_m^{N,p}(x).
\tag{6.13}
$$

The left-hand side of (6.13) can be evaluated by means of (6.9) and (6.6):

$$
(\mathcal{F}\psi_n^{N,p})(x)=\binom{N}{n}^{-1}\left(\frac{1-p}{p}\right)^{n/2}\frac{\delta_{d(x,0),n}}{p^{d(x,0)/2}(1-p)^{(N-d(x,0))/2}}.
$$

Hence

$$
\sum_{m=0}^{N}T_{m,n}^N(s_p)\binom{N}{m}^{1/2}\left(\frac{p}{1-p}\right)^{m/2}\psi_m^{N,p}(l)=\binom{N}{n}^{-1/2}p^{-l/2}(1-p)^{(l-N)/2}.
$$

Now we use that $(T_{m,n}^N(s_p))$ is a real orthogonal matrix (cf. (2.8)). Finally

$$
T_{r,l}^N(s_p)=\binom{N}{r}^{1/2}\binom{N}{l}^{1/2}p^{(l+r)/2}(1-p)^{(N-l-r)/2}\tilde{\psi}_r^{N,p}(l).
\tag{6.14}
$$

In particular, by combination of (6.14) with (6.7), we have given a conceptual explanation that both the canonical matrix elements of $SU(2)$ and the intertwining functions on Hamming schemes can be expressed in terms of the same special functions.

Acknowledgment. I would like to thank Dr. T. A. Springer for calling my attention to the metaplectic representation and Richard Askey for suggesting the group theoretic interpretation of Krawtchouk polynomials as intertwining functions.

REFERENCES

[1] R. Askey, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[2] C. F. Dunkl, *A Krawtchouk polynomial addition theorem and wreath products of symmetric groups*, Indiana Univ. Math. J. 25 (1976), pp. 335–358.

[3] A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, *Higher Transcendental Functions*, Vols. I, II, McGraw-Hill, New York, 1953.

[4] I. M. Gelfand and Z. Ya. Šapiro, *Representations of the group of rotations of three-dimensional space and their applications*, Amer. Math. Soc. Transl. (2) 2 (1956), pp. 207–316.

[5] K. I. Gross and R. A. Kunze, *Bessel functions and representation theory, II. Holomorphic discrete series and metaplectic representations*, J. Func. Anal. 25 (1977), pp. 1–49.

[6] E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis*, Vol. II, Springer-Verlag, Berlin, 1970.

[7] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1978.

[8] S. Rallis and G. Schiffmann, *Weil representation, I. Intertwining distributions and discrete spectrum*, Mem. Amer. Math. Soc., 25 (1980), no. 231.

[9] P. Sally, *Analytic continuation of the irreducible unitary representations of the universal covering group of* $SL(2, \mathbb{R})$, Mem. Amer. Math. Soc., (1967), no. 69.

[10] G. Szegö, *Orthogonal Polynomials*, 4th ed., American Mathematical Society, Providence, RI, 1975.

[11] D. Vere-Jones, *On finite bivariate distributions*, Quart. J. Math. Oxford Ser. 2, 22 (1971), pp. 247–270.

[12] A. Weil, *Sur certains groupes d'opérateurs unitaires*, Acta Math., 111 (1964), pp. 143–211.

[13] E. P. Wigner, *Group Theory and Its Application to the Quantum Mechanics of Atomic Spectra*, Academic Press, New York, 1959.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0011 $01.00/0

THE GAUSS HYPERGEOMETRIC RATIO AS A POSITIVE REAL FUNCTION*

VITOLD BELEVITCH†

Abstract. The Gauss continued fraction for the ratio of two hypergeometric functions is converted into an *ordinary* fraction (all partial numerators are 1) and simplifications occurring for particular relations between the parameters are discussed. In particular, a very simple expansion is obtained for the ratio E/K of the complete elliptic integrals. For the argument $-z$ and for certain ranges of the parameters, the Gauss expansion is a Stieltjes fraction and represents the input impedance of a passive ladder network. The Stieltjes integral representations of the corresponding positive real functions are established and yield many new definite integrals. A general method for obtaining indefinite integrals involving independent solutions of a self-adjoint differential equation, by means of the Wronskian, is also mentioned. Finally, some continued fractions originating from other contiguity relations for hypergeometric functions are discussed.

1. The Gauss continued fraction. The continued fraction expansion of the ratio $F(\alpha,\beta,\gamma;z)/F(\alpha,\beta+1,\gamma+1;z)$ of two contiguous hypergeometric functions is discussed in [1, pp. 122, 151]. For further convenience we replace β by $\beta-1$ and z by $-z$. One then has

$$\frac{F(\alpha,\beta-1,\gamma;-z)}{F(\alpha,\beta,\gamma+1;-z)}=1+\frac{a_1 z}{1}+\frac{a_2 z}{1}+\cdots \tag{1}$$

with

(2)

$$a_{2k}=\frac{(\beta+k-1)(\gamma-\alpha+k)}{(\gamma+2k-1)(\gamma+2k)},\quad a_{2k-1}=\frac{(\alpha+k-1)(\gamma-\beta+k)}{(\gamma+2k-2)(\gamma+2k-1)},\quad k=1,2,3,\cdots.$$

For real values of the parameters α,β,γ, and $\gamma\neq 0,-1,-2,\cdots$, the expansion (1) is valid in the complex z-plane with a branch cut from $-\infty$ to -1 on the real axis.

The input impedance of an electric ladder network is an *ordinary* continued fraction (all partial numerators are 1). The expansion (1) can be reduced to the desired form by introducing common factors. This yields

$$\frac{F(\alpha,\beta-1,\gamma;-z)}{F(\alpha,\beta,\gamma+1;-z)}=1+\frac{1}{b_1/z}+\frac{1}{b_2}+\frac{1}{b_3/z}+\frac{1}{b_4}+\cdots \tag{3}$$

with

$$b_1+\frac{1}{a_1},\quad b_2=\frac{a_1}{a_2},\quad b_3=\frac{1}{b_2 a_3},\quad b_4=\frac{a_1 a_3}{a_2 a_4},\quad b_5=\frac{1}{b_4 a_5},\quad b_6=\frac{a_1 a_3 a_5}{a_2 a_4 a_6},\quad \cdots. \tag{4}$$

For

$$0<\alpha<\gamma+1,\quad 0<\beta<\gamma+1,\quad \gamma>0 \tag{5}$$

all values (2) are positive, and so are the values (4). The expansion (3) is thus a Stieltjes fraction in the range (5) and represents the impedance of a passive RL-network (series resistances and shunt inductances) as a function of $z=i\omega$ (ω = radian frequency). The change of the argument into $-z$ has been introduced to make (3) a positive real function of z, and the replacement of β by $\beta-1$ makes (5) symmetric in α and β.

*Received by the editors July 15, 1980, and in revised form January 27, 1982.
†Philips Research Laboratory, B-1170 Brussels, Belgium.

We now check that (3) remains positive real when any one of the inequalities occurring in (5) is changed into an equality. For $\alpha=0$, the function reduces to 1. For $\alpha=\gamma+1$, the transformation [2, eq. 15.3.3]

$$\frac{F(\alpha,\beta-1,\gamma;-z)}{F(\alpha,\beta,\gamma+1;-z)}=\frac{F(\gamma-\alpha,\gamma-\beta+1,\gamma;-z)}{F(\gamma+1-\alpha,\gamma+1-\beta,\gamma+1;-z)} \tag{6}$$

shows that (3) reduces to $1+(\gamma-\beta+1)/\gamma z$. Similar results hold by symmetry for the extreme values of β. For $\gamma=0$, a_1 has a factor γ in the denominator but all the other elements (2) remain finite nonzero. In (4) all b_{2k+1} thus contain the factor γ and all b_{2k} the factor $1/\gamma$; these factors are cancelled if one multiplies (3) by γ and the resulting function is

$$\lim_{\gamma\to 0}\frac{\gamma F(\alpha,\beta-1,\gamma;-z)}{F(\alpha,\beta,\gamma+1;-z)}=\frac{\alpha(1-\beta)zF(\alpha+1,\beta,2;-z)}{F(\alpha,\beta,1;-z)} \tag{7}$$

by [2, eq. 15.1.2].

2. Simplifications. In general, the elements (4) grow in complexity with their rank, and the resulting impedance is of no practical interest. By contrast, there is a large engineering interest in finding closed form expressions for Stieltjes continued fractions having the form of the right-hand side in (3) where the elements b_k are rational functions of k of low degree. Such a fraction is obtained for $\alpha=\beta$, producing a drastic simplification in the ratio a_{2k-1}/a_{2k} of (2); the resulting expansion (3) multiplied by γ, i.e.,

$$\frac{\gamma F(\alpha,\alpha-1,\gamma;-z)}{F(\alpha,\alpha,\gamma+1;-z)}=c_0+\cfrac{1}{c_1/z+\cfrac{1}{c_2+\cfrac{1}{c_3/z+\cdots}}} \tag{8}$$

with

$$c_{2k}=\gamma+2k,\qquad c_{2k+1}=\frac{\gamma+2k+1}{(\alpha+k)(\gamma-\alpha+k+1)},\qquad k=0,1,2,\cdots \tag{9}$$

represents a simple network. In particular, for $\alpha=1$, $\gamma=\frac{1}{2}$, (8) yields, via [2, eq. 15.1.7] the known expansion of

$$\frac{1}{2F\left(1,1,\frac{3}{2};-z\right)}=\frac{\sqrt{(1+z)z}}{2\operatorname{arc\,sinh}\sqrt{z}} \tag{10}$$

with

$$c_{2k}=\frac{4k+1}{2},\qquad c_{2k+1}=\frac{4k+3}{(k+1)(2k+1)}. \tag{11}$$

With z replaced by z^2 this coincides with [2, eq. 4.6.36] after reduction to an ordinary fraction.

An additional simplification occurs in (9) for $\alpha=(\gamma+1)/2$. The expansion resulting from (8) and (9) is then, with z replaced by $4z$,

$$\frac{\gamma F\left(\frac{\gamma+1}{2},\frac{\gamma-1}{2},\gamma;-4z\right)}{F\left(\frac{\gamma+1}{2},\frac{\gamma+1}{2},\gamma+1;-4z\right)}=\gamma+\cfrac{1}{1/z(\gamma+1)+\cfrac{1}{\gamma+2+\cfrac{1}{1/z(\gamma+3)+\cdots}}}. \tag{12}$$

For $\gamma=1$ this yields the known expansion [2, eq. 15.1.3]

$$\frac{4z}{\ln(1+4z)}=1+\frac{1}{1/2z}+\frac{1}{3}+\frac{1}{1/4z}+\frac{1}{5}+\cdots. \tag{13}$$

For $\beta\neq 1$, the ratio (3) apparently involves two distinct hypergeometric functions, but it is in fact expressible in terms of the logarithmic derivative of a single function. By [2, eq. 15.2.7] we have

$$\frac{d}{dz}(1-z)^{\alpha}F(\alpha,\beta,\gamma;z)=\frac{\alpha(\beta-\gamma)}{\gamma}(1-z)^{\alpha-1}F(\alpha+1,\beta,\gamma+1;z).$$

Introducing the notation

$$G(\alpha,\beta,\gamma;z)=\frac{d}{dz}\ln F(\alpha,\beta,\gamma;z)$$

one obtains

$$\frac{F(\alpha+1,\beta,\gamma+1;z)}{F(\alpha,\beta,\gamma;z)}=\frac{\gamma}{\gamma-\beta}\left[1+\frac{z-1}{\alpha}G(\alpha,\beta,\gamma;z)\right]. \tag{14}$$

By permuting the first two parameters in (14), taking the inverse and decreasing the second parameter by 1, one expresses (3) in terms of $G(\alpha,\beta-1,\gamma;-z)$ but this destroys the symmetry in α and β. However, by [1, p. 122 eq. 2] one has

$$\frac{F(\alpha,\beta-1,\gamma;-z)}{F(\alpha,\beta,\gamma+1;-z)}-1=\frac{\alpha(\gamma-\beta+1)z}{\gamma(\gamma+1)}\,\frac{F(\alpha+1,\beta,\gamma+2;-z)}{F(\alpha,\beta,\gamma+1;-z)}. \tag{15}$$

By inserting (14) with γ replaced by $\gamma+1$ and z by $-z$ into (15), and multiplying by γ, one obtains

$$\frac{\gamma F(\alpha,\beta-1,\gamma;-z)}{F(\alpha,\beta,\gamma+1;-z)}=\gamma+\alpha z\left[1-\frac{1+z}{\alpha}G(\alpha,\beta,\gamma+1;-z)\right] \tag{16}$$

and this result also holds for $\gamma=0$.

With the usual notations for the complete elliptic integrals E and K of modulus k (and with $k'^2=1-k^2$) we have [3, pp. 499, 521]

$$F\left(\frac{1}{2},\frac{1}{2},1;k^2\right)=\frac{2}{\pi}K, \tag{17}$$

$$\frac{dK}{dk}=\frac{E-k'^2K}{kk'^2}. \tag{18}$$

For $\gamma=0$ and $\alpha=\beta=\frac{1}{2}$ the left-hand side of (16) can be evaluated by (7) whereas the right-hand side simplifies via (17)–(18) with $-z=k^2$, hence $k=i\sqrt{z}$. This gives

$$\frac{z}{4}\,\frac{F\left(\frac{3}{2},\frac{1}{2},2;-z\right)}{F\left(\frac{1}{2},\frac{1}{2},1;-z\right)}=\frac{1}{2}\left[\frac{E\left(i\sqrt{z}\right)}{K\left(i\sqrt{z}\right)}-1\right]. \tag{19}$$

Since the left-hand side of (16) with z changed into $4z$ is (12) and thus equals the right-hand side of (19) with the elliptic modulus $2i\sqrt{z}$, one obtains a simple closed-form expression for the continued fraction (12) with $\gamma=0$. Changing z with z^2 and dividing by z one obtains

$$\frac{1}{2z}\left[\frac{E(2iz)}{K(2iz)}-1\right]=\frac{1}{1/z}+\frac{1}{2/z}+\frac{1}{1/3z}+\frac{1}{4/z}+\frac{1}{1/5z}+\cdots. \tag{20}$$

Since E and K are even functions of their arguments, the left-hand side of (20) is real in z. By Jacobi's imaginary transformation [4, p. 274], we have

$$K\left(\frac{ik}{k'}\right)=k'K(k), \qquad E\left(\frac{ik}{k'}\right)=\frac{1}{k'}E(k). \tag{21}$$

For

$$z=\frac{k}{2k'}=\frac{u}{2} \tag{22}$$

one obtains the following explicit real expression of (20):

$$\frac{1}{k}\left[\frac{E(k)}{k'K(k)}-k'\right]=\frac{1}{u}\left[(1+u^2)\frac{E\left(u/\sqrt{1+u^2}\right)}{K\left(u/\sqrt{1+u^2}\right)}-1\right]. \tag{23}$$

3. The Wronskian. The Wronskian

$$W=u_1\frac{du_2}{dz}-u_2\frac{du_1}{dz} \tag{24}$$

of two solutions of the self-adjoint differential equation

$$\frac{d}{dz}\left[p(z)\frac{du}{dz}\right]+q(z)u=0 \tag{25}$$

satisfies the equation

$$p\frac{dW}{dz}+\frac{dp}{dz}W=0$$

and is thus

$$W=\frac{C}{p(z)}, \tag{26}$$

where C is a constant which can be determined by considering the value of (24) at some z. The self-adjoint form of the hypergeometric differential equation is [7, p. 5, eq. 11b]

$$\frac{d}{dz}\left[z^{\gamma}(1-z)^{\alpha+\beta-\gamma+1}\frac{du}{dz}\right]-\alpha\beta z^{\gamma-1}(1-z)^{\alpha+\beta-\gamma}u=0 \tag{27}$$

and (26) thus becomes

$$W=Cz^{-\gamma}(1-z)^{\gamma-\alpha-\beta-1} \tag{28}$$

in accordance with [5, p. 84, eq. 28]. Kummer's 24 solutions of (27) are listed on p. 67 and the corresponding values of C in (28) are given on pp. 84–85 of [5].

If $f(u)$ is an arbitrary differentiable function, we have

$$\frac{d}{dz}f\left(\frac{u_2}{u_1}\right)=\frac{u_1u_2'-u_2u_1'}{u_1^2}f'\left(\frac{u_2}{u_1}\right)=\frac{W}{u_1^2}f'\left(\frac{u_2}{u_1}\right), \tag{29}$$

hence

$$\int_a^b\frac{W}{u_1^2}f'\left(\frac{u_2}{u_1}\right)dz=f\left(\frac{u_2}{u_1}\right)\Bigg|_a^b. \tag{30}$$

A large variety of integrals is obtained for various choices for u_1 and u_2 among the Kummer's solutions of (27) and for various choices of elementary functions for $f(u)$,

such as u^v, $\arctan u$ or $\ln u$. Moreover, since a hypergeometric function takes simple values for some particular values of its argument, such as

$$F(\alpha,\beta,\gamma;0)=1 \tag{31}$$

and [2, eq. 15.1.20]

$$F(\alpha,\beta,\gamma;1)=\frac{\Gamma(\gamma)\Gamma(\gamma-\alpha-\beta)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)} \qquad [\mathrm{Re}(\gamma-\alpha-\beta)>0,\gamma\neq 0,-1,-2,\cdots] \tag{32}$$

many simple definite integrals can be obtained. Note that, by [2, eq. 15.3.3], i.e., by

$$F(\alpha,\beta,\gamma;z)=(1-z)^{\gamma-\alpha-\beta}F(\gamma-\alpha,\gamma-\beta,\gamma;z), \tag{33}$$

$F(\alpha,\beta,\gamma;1)$ is infinite for $\mathrm{Re}(\gamma-\alpha-\beta)\leq 0$.

4. Examples. We take

$$u_1=F(\alpha,\beta,\gamma;z) \tag{34}$$

and [5, p. 67 eq. 9]

$$u_2=F(\alpha,\beta,\alpha+\beta+1-\gamma;1-z) \tag{35}$$

so that the constant in the Wronskian (28) is [5, p. 84 eq. 29]

$$C=-\frac{\Gamma(\alpha+\beta+1-\gamma)\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\beta)}. \tag{36}$$

We have $u_1(0)=u_2(1)=1$ but, by (32)–(33)

$$u_1(1)=\begin{cases}\dfrac{\Gamma(\gamma)\Gamma(\gamma-\alpha-\beta)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)}, & \mathrm{Re}(\gamma-\beta-\alpha)>0,\\ \infty, & \mathrm{Re}(\gamma-\beta-\alpha)\leq 0\end{cases} \tag{37}$$

$$u_2(0)=\begin{cases}\dfrac{\Gamma(\alpha+\beta+1-\gamma)\Gamma(1-\gamma)}{\Gamma(\beta+1-\gamma)\Gamma(\alpha+1-\gamma)}, & \mathrm{Re}\,\gamma<1,\\ \infty, & \mathrm{Re}\,\gamma\geq 1,\end{cases} \tag{38}$$

and both u_1 and u_2 are analytic for $0<z<1$.

With $f(u)=u$, (30) gives

$$\int_0^1\frac{W\,dz}{u_1^2}=\frac{u_2}{u_1}\Big|_0^1. \tag{39}$$

For the lower choice in (37) and the upper choice in (38), (39) yields

$$\int_0^1\frac{z^{-\gamma}(1-z)^{\gamma-\alpha-\beta-1}dz}{F^2(\alpha,\beta,\gamma;z)}=\frac{\Gamma(1-\gamma)\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\beta+1-\gamma)\Gamma(\alpha+1-\gamma)\Gamma(\gamma)} \tag{40}$$

$$[\mathrm{Re}\,\gamma<1,\ \mathrm{Re}(\gamma-\beta-\alpha)\leq 0].$$

For $\mathrm{Re}\,\gamma\geq 1$, the integral (40) clearly diverges. For $\mathrm{Re}\,\gamma<1$ but $\mathrm{Re}(\gamma-\beta-\alpha)>0$ a rather complicated expression is obtained by taking the upper choices in (37)–(38). However, the use of (33) in the left-hand side of (40) and the substitutions $\alpha'=\gamma-\alpha$, $\beta'=\gamma-\beta$ produce $\mathrm{Re}(\gamma-\beta'-\alpha')<0$ and transform the integral of (40) into a similar

integral with parameters α', β'. The same substitutions in the right-hand side of (40) yield

(41)

$$\int_0^1 \frac{z^{-\gamma}(1-z)^{\gamma-\alpha-\beta-1}dz}{F^2(\alpha,\beta,\gamma;z)} = \frac{\Gamma(1-\gamma)\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)}{\Gamma(1-\alpha)\Gamma(1-\beta)\Gamma(\gamma)} \qquad [1>\mathrm{Re}\,\gamma>\mathrm{Re}(\alpha+\beta)].$$

The equivalence of (41) with the more complicated expression resulting from the direct use of the upper expressions (37)–(38) can be established by applying several times the complement formula for Γ-functions.

With $f(u)=\arctan u$, (30) gives

(42)
$$\int_0^1 \frac{W\,dz}{u_1^2+u_2^2} = \arctan\frac{u_2}{u_1}\Big|_0^1$$

which becomes $0-\pi/2$ for the lower choices in (37)–(38). Hence

(43)
$$\int_0^1 \frac{z^{-\gamma}(1-z)^{\gamma-\alpha-\beta-1}dz}{F^2(\alpha,\beta,\gamma;z)+F^2(\alpha,\beta,\alpha+\beta+1-\gamma;1-z)} = \frac{\pi}{2}\,\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\gamma)\Gamma(\alpha+\beta+1-\gamma)}$$
$$[1\le\mathrm{Re}\,\gamma\le\mathrm{Re}(\alpha+\beta)].$$

For $\gamma=\alpha+\beta=1$, and using the complement relation for the Γ-function, one simplifies (43) to

(44)
$$\int_0^1 \frac{dz}{z(1-z)\left[F^2(\alpha,1-\alpha,1;z)+F^2(\alpha,1-\alpha,1;1-z)\right]} = \frac{\pi^2}{2\sin\pi\alpha}.$$

For $z=k^2$ and $\alpha=\frac{1}{2}$, (44) simplifies by (17) into

(45)
$$\int_0^1 \frac{dk}{kk'^2(K^2+K'^2)} = 1$$

which can also be established directly by computing the derivative of $\arctan K/K'$ by (18) and the Legendre identity.

From (45) one can also deduce that

(46)
$$\int_0^1 \frac{dk}{k(K^2+K'^2)} = \frac{1}{2}.$$

By changing k into k' in (46), which leaves $K^2+K'^2$ invariant, and eliminating dk' by $k\,dk=-k'\,dk'$ one proves that (46) is equivalent to

(47)
$$\int_0^1 \frac{k\,dk}{k'^2(K^2+K'^2)} = \frac{1}{2}$$

and the sum of (46) and (47) is (45).

With $f(u)=\ln u$, (30) gives

(48)
$$\int_0^1 \frac{W\,dz}{u_1u_2} = \ln\frac{u_2(1)}{u_1(1)}\,\frac{u_1(0)}{u_2(0)}.$$

With the upper choices in (37)–(38) one obtains, after simplification by the complement relation for the Γ-functions,

$$\int_0^1 \frac{z^{-\gamma}(1-z)^{\gamma-\alpha-\beta-1}dz}{F(\alpha,\beta,\gamma;z)F(\alpha,\beta,\alpha+\beta-\gamma+1;1-z)}$$

$$(49)\qquad = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\gamma)\Gamma(\alpha+\beta+1-\gamma)} \ln \frac{\sin\pi(\gamma-\alpha)\sin\pi(\gamma-\beta)}{\sin\pi\gamma\sin\pi(\gamma-\alpha-\beta)}$$

$$[1>\mathrm{Re}\,\gamma>\mathrm{Re}(\alpha+\beta)].$$

5. Stieltjes representation. The continued fraction expansion of (15) is (3) with first term 1 omitted. Changing γ into $\gamma-1$ and multiplying the result by $\gamma(\gamma-1)/\alpha(\gamma-\beta)$ one obtains the function

$$(50)\qquad f(z)=z\frac{F(\alpha+1,\beta,\gamma+1;-z)}{F(\alpha,\beta,\gamma;-z)}.$$

From the range (5) of the parameters and its extension discussed at the end of §1 it results that (50) is positive real for

$$(51)\qquad 0\le\alpha\le\gamma,\quad 0\le\beta\le\gamma,\quad \gamma\ge 1.$$

Consequently,

$$(52)\qquad g(z)=f\left(\frac{1}{z}\right)=\frac{F(\alpha+1,\beta,\gamma+1;-1/z)}{zF(\alpha,\beta,\gamma;-1/z)}$$

is positive real and analytic in the entire complex plane with a branch-cut from -1 to 0, and one has $g(z)=O(z^{-1})$ for large $|z|$.

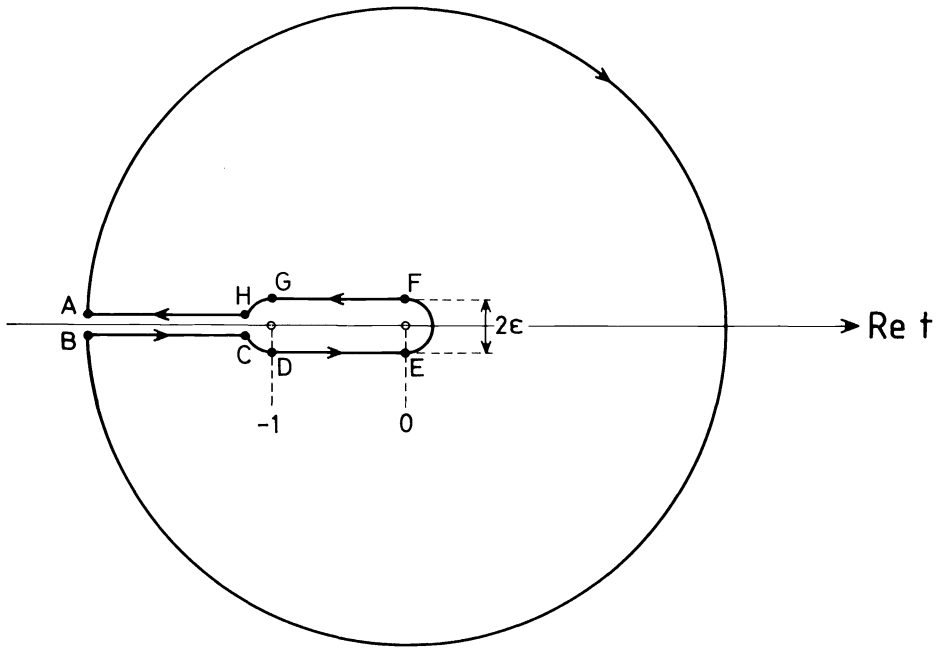

FIG. 1

We have the Cauchy-integral

$$\frac{1}{2\pi i}\int_C \frac{g(t)\,dt}{z-t}=g(z), \qquad |\arg z|<\pi \tag{53}$$

on any simple contour C not crossing the cut. The contour we choose is shown in Fig. 1 and the contribution on the large circle AB around the origin tends to zero because of the remark preceding (53). The contributions to (53) on BC and HA cancel each other since the interval $-\infty<t<-1$ is outside the cut, so that the function is continuous and both paths coalesce. Also, the contributions on the quarter-circles CD and GH tend to zero because $g(z)$ is finite at $z=-1$ by (32) and (33).

As regards the contribution of the small half-circle EF, we must investigate the nature of the singularity of (52) at the origin. By [5, p. 70, eq. 1], we have

$$F(\alpha,\beta,\gamma;z)=\Gamma(\gamma)\left[\frac{\Gamma(\beta-\alpha)}{\Gamma(\beta)\Gamma(\gamma-\alpha)}w_5+\frac{\Gamma(\alpha-\beta)}{\Gamma(\alpha)\Gamma(\gamma-\beta)}w_6\right], \tag{54}$$

where w_5 and w_6 are [5, p. 67, eqs. 17, 21]

$$w_5=(z^{-1}e^{i\pi})^{\alpha}F(\alpha,\alpha+1-\gamma,\alpha+1-\beta;z^{-1}), \tag{55}$$

$$w_6=(z^{-1}e^{i\pi})^{\beta}F(\beta,\beta+1-\gamma,\beta+1-\alpha;z^{-1}) \tag{56}$$

and where (54) holds for $0<\arg z<2\pi$ [5, p. 69]. From the above equations and (52) we deduce that $g(z)$ is finite at the origin for $\beta-\alpha>1$, has a simple pole with positive residue $\gamma(\alpha-\beta)/\alpha(\gamma-\beta)$ for $\beta-\alpha<0$, whereas the singularity at the origin is a branch-point weaker than a simple pole for $0<\beta-\alpha<1$ and is logarithmic for $\beta-\alpha=0$ or 1. In any case, the contribution of EF to (53) is $\gamma(\alpha-\beta)/2\alpha(\gamma-\beta)z$ for $\alpha>\beta$, else zero.

We set

$$g(-z-i0)=A(z)+iB(z) \tag{57}$$

on the parallel below the cut, and the conjugate value holds on the parallel above the cut. The corresponding contributions to (53) on DE and GF result in

$$\frac{1}{\pi}\int_{-1}^{0}\frac{B(-t)\,dt}{z-t}=\frac{1}{\pi}\int_0^1\frac{B(t)\,dt}{z+t}.$$

Taking into account the contribution of the pole at the origin, one thus reduces (53) to

$$g(z)=\frac{1}{\pi}\int_0^1\frac{B(t)\,dt}{z+t}+\begin{cases}0, & \alpha\le\beta,\\ \dfrac{\gamma(\alpha-\beta)}{2\alpha(\gamma-\beta)z}, & \alpha>\beta.\end{cases} \tag{58}$$

Changing z into $1/z$ and dividing by z one obtains

$$\frac{F(\alpha+1,\beta,\gamma+1;-z)}{F(\alpha,\beta,\gamma;-z)}=\frac{1}{\pi}\int_0^1\frac{B(t)dt}{1+zt}+\begin{cases}0, & \alpha\le\beta,\\ \dfrac{\gamma(\alpha-\beta)}{2\alpha(\gamma-\beta)}, & \alpha>\beta,\end{cases} \tag{59}$$

where, by (57),

$$B(t)=\operatorname{Im}g(-z)|_{z=t+i0}=\operatorname{Im}f(-z)|_{z=t^{-1}-i0}. \tag{60}$$

From a comparison of (50) and (14) we deduce

$$B(t)=\frac{\gamma(t-1)}{\alpha(\gamma-\beta)t^2}\operatorname{Im}G(\alpha,\beta,\gamma;z)\Big|_{z=t^{-1}-i0}. \tag{61}$$

Setting

$$F(\alpha,\beta,\gamma;z-i0)=U(z)+iV(z) \tag{62}$$

one deduces from the definition of G preceding (14) that

$$\operatorname{Im}G(\alpha,\beta,\gamma;z-i0)=\frac{d}{dz}\arctan\frac{V}{U}=\frac{VU'-UV'}{U^2+V^2}=\frac{VU'-UV'}{|F(\alpha,\beta,\gamma;z)|^2}, \tag{63}$$

where the argument z in the denominator stands indifferently for $z\pm i0$. By (54)–(56), (62) is of the form

$$F=e^{i\pi\alpha}X+e^{i\pi\beta}Y, \tag{64}$$

where

$$X=e^{-i\pi\alpha}\frac{\Gamma(\gamma)\Gamma(\beta-\alpha)}{\Gamma(\beta)\Gamma(\gamma-\alpha)}w_5,\qquad Y=e^{-i\pi\beta}\frac{\Gamma(\gamma)\Gamma(\alpha-\beta)}{\Gamma(\alpha)\Gamma(\gamma-\beta)}w_6 \tag{65}$$

and are real when $\operatorname{Im}z$ tends to 0. By comparison of (62) and (64) we have

$$U=X\cos\pi\alpha+Y\cos\beta,\qquad V=X\sin\pi\alpha+Y\sin\pi\beta \tag{66}$$

and the combination occurring in (63) is

$$VU'-UV'=(XY'-YX')\sin\pi(\alpha-\beta). \tag{67}$$

From the expression (65) and from the Wronskian [5 p. 85, eq. 34]

$$w_5w_6'-w_6w_5'=(\beta-\alpha)e^{i\pi\gamma}z^{-\gamma}(1-z)^{\gamma-\alpha-\beta-1} \tag{68}$$

we deduce

$$XY'-YX'=e^{i\pi(\gamma-\alpha-\beta)}\frac{\Gamma^2(\gamma)\Gamma(\beta-\alpha+1)\Gamma(\alpha-\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)}z^{-\gamma}(1-z)^{\gamma-\alpha-\beta-1}. \tag{69}$$

From (67), (69) and the complement relation for $\Gamma(\alpha-\beta)$, we obtain

$$VU'-UV'=-\pi\frac{\Gamma^2(\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)}z^{-\gamma}(z-1)^{\gamma-\alpha-\beta-1}. \tag{70}$$

Finally, by (61) and (63), we have

$$B(t)=\pi\frac{\Gamma(\gamma)\Gamma(\gamma+1)}{\Gamma(\alpha+1)\Gamma(\gamma-\beta+1)\Gamma(\beta)\Gamma(\gamma-\alpha)}\frac{t^{\alpha+\beta-1}(1-t)^{\gamma-\alpha-\beta}}{|F(\alpha,\beta,\gamma;t^{-1})|^2} \tag{71}$$

and (59) becomes

$$\frac{\Gamma(\gamma)\Gamma(\gamma+1)}{\Gamma(\alpha+1)\Gamma(\beta)\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta+1)}\int_0^1\frac{t^{\alpha+\beta-1}(1-t)^{\gamma-\alpha-\beta}dt}{(1+zt)|F(\alpha,\beta,\gamma;t^{-1})|^2}$$

$$=\frac{F(\alpha+1,\beta,\gamma+1;-z)}{F(\alpha,\beta,\gamma;-z)}-\begin{cases}0, & \alpha\le\beta,\\ \dfrac{\gamma(\alpha-\beta)}{2\alpha(\gamma-\beta)}, & \alpha>\beta.\end{cases} \tag{72}$$

The result (72) is valid within the restrictions (51) on the parameters, but the restriction $|\arg z|<\pi$ of (53) can be dropped outside the cut because of the cancellations of the integrals on BC and HA, so that (72) holds for all z outside the open interval $-1<z<0$.

The integral in (72) is symmetric in α and β; one easily checks by contiguity relations that so is its expression. For $\alpha=0$, (72) coincides with the integral representation [2, eq. 15.3.1] of $F(1,\beta,\gamma+1;-z)$. The corresponding impedance has been discussed in [11].

For $z=0$, (72) reduces by (31) to

$$\frac{\Gamma(\gamma)\Gamma(\gamma+1)}{\Gamma(\alpha+1)\Gamma(\beta)\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta+1)}\int_0^1 \frac{t^{\alpha+\beta-1}(1-t)^{\gamma-\alpha-\beta}dt}{|F(\alpha,\beta,\gamma,t^{-1})|^2} \tag{73}$$
$$=1-\begin{cases}0, & \alpha\le\beta,\\ \dfrac{\gamma(\alpha-\beta)}{2\alpha(\gamma-\beta)}, & \alpha>\beta.\end{cases}$$

For $z=-1$, (72) reduces, by (32)–(33) to

$$\frac{\Gamma(\gamma)\Gamma(\gamma+1)}{\Gamma(\alpha+1)\Gamma(\beta)\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta+1)}\int_0^1 \frac{t^{\alpha+\beta-1}(1-t)^{\gamma-\alpha-\beta-1}}{|F(\alpha,\beta,\gamma,t^{-1})|^2}dt \tag{74}$$
$$=\begin{cases}\dfrac{\gamma}{\gamma-\beta}, & \alpha\le\beta,\quad \gamma>\alpha+\beta,\\ \dfrac{\gamma(\alpha+\beta)}{2\alpha(\gamma-\beta)}, & \alpha>\beta,\quad \gamma>\alpha+\beta,\\ \dfrac{\gamma}{\alpha}, & \alpha\le\beta,\quad \gamma<\alpha+\beta,\\ \dfrac{\gamma(2\gamma-\alpha-\beta)}{2\alpha(\gamma-\beta)}, & \alpha>\beta,\quad \gamma<\alpha+\beta.\end{cases}$$

In all the above integrals $F(\alpha,\beta,\gamma,t^{-1})$ has in fact the argument $t^{-1}\pm i0$ and can be evaluated in terms of hypergeometric functions of argument t by means of (54)–(56). The resulting expression becomes undetermined for $\alpha=\beta$ but an alternate expression in terms of functions of arguments t and $1-t$ can be deduced from the relation resulting from the remark following [5, p. 71, eq. 19] by using the expressions of w_4 and w_6 given in [5, p. 67, eq. 16, 21].

6. Stieltjes integrals involving E and K. For $\alpha=\beta=\frac{1}{2}$, $\gamma=1$, and $t=k^2$, the hypergeometric function in the integral of (72) is

$$F\left(\frac{1}{2},\frac{1}{2},1;k^{-2}\right)=\frac{2}{\pi}K\left(\frac{1}{k}\right) \tag{75}$$

by (17), whereas the ratio of hypergeometric functions in the right-hand side of (72) is $4/z$ times (19). Changing z into z^2, one obtains

$$\int_0^1 \frac{k\,dk}{(1+k^2z^2)|K(1/k)|^2}=\frac{1}{z^2}\left[\frac{E(iz)}{K(iz)}-1\right] \tag{76}$$

which holds for all z with $\mathrm{Re}\,z\ge 0$ except in the open interval $]-i,i[$ corresponding to the cut. The same restriction holds for all the following integrals in this section and will not be repeated.

The value of $K(1/k)$ can be evaluated as explained in the last paragraph of §5 or directly by decomposing the integral from 0 to 1 defining K into integrals from 0 to k and from k to 1. The result is

$$K\left(\frac{1}{k}\pm i0\right)=k[K\pm iK'] \tag{77}$$

and (76) becomes

$$\int_0^1 \frac{dk}{k(1+k^2z^2)(K^2+K'^2)}=\frac{1}{z^2}\left[\frac{E(iz)}{K(iz)}-1\right]. \tag{78}$$

For $z=0$ and $z=i$, (78) reduces to (46) and (45), respectively. For $z=1$ the transformation from (20) to (23) generates functions E and K of modulus $k=k'=1/\sqrt{2}$ which have known values [3, pp. 524–526] and (78) reduces to

$$\int_0^1 \frac{dk}{k(1+k^2)(K^2+K'^2)}=\frac{8\pi^2}{\Gamma^4(1/4)}. \tag{79}$$

For $\alpha=-\frac{1}{2}$, $\beta=\frac{1}{2}$, $\gamma=1$ and $t=k^2$ we have in the integrand of (72)

$$F\left(-\frac{1}{2},\frac{1}{2},1;k^{-2}\right)=\frac{2}{\pi}E\left(\frac{1}{k}\right), \tag{80}$$

whereas the ratio in the right-hand side is, by (14),

$$\frac{F\left(\frac{1}{2},\frac{1}{2},2;-z\right)}{F\left(-\frac{1}{2},\frac{1}{2},1;-z\right)}=2\left[1+2(1+z)G\left(-\frac{1}{2},\frac{1}{2},1;-z\right)\right]. \tag{81}$$

In (81) with $z=-k^2$, G is proportional to the logarithmic derivative of $E(k)$ which can be deduced from [3, p. 521]

$$\frac{dE}{dk}=\frac{E-K}{k} \tag{82}$$

so that (81) becomes

$$\frac{2}{k^2}\left[1-\frac{k'^2K}{E}\right]=\frac{2}{z}\left[(1+z)\frac{K\left(i\sqrt{z}\right)}{E\left(i\sqrt{z}\right)}-1\right]. \tag{83}$$

Finally, changing z into z^2, (72) yields

$$\frac{1}{z^2}\left[(1+z^2)\frac{K(iz)}{E(iz)}-1\right]=\int_0^1 \frac{k'^2\,dk}{k(1+k^2z^2)|E(1/k)|^2}. \tag{84}$$

The analogue of (77) for E is

$$E\left(\frac{1}{k}\pm i0\right)=\frac{1}{k}[E-k'^2K\mp i(E'-k^2K')]. \tag{85}$$

Note that (77) and (85) are in [6] but without mention of the $\pm$ sign and with incoherent determinations. By (85), (84) becomes

$$\frac{1}{z^2}\left[(1+z^2)\frac{K(iz)}{E(iz)}-1\right]=\int_0^1 \frac{k'^2k\,dk}{(1+k^2z^2)\left[(E-k'^2K)^2+(E'-k^2K')^2\right]}. \tag{86}$$

For $z=0$, (86) yields

$$\int_0^1 \frac{kk'^2\,dk}{(E-k'^2K)^2+(E'-k^2K')}=\frac{1}{2}. \tag{87}$$

For $z=i$, (86) yields

$$\int_0^1 \frac{k\,dk}{(E-k'^2K)^2+(E'-k^2K')^2}=1 \tag{88}$$

and the difference between (88) and (87) yields

$$\int_0^1 \frac{k^3\,dk}{(E-k'^2K)^2+(E'-k^2K')^2}=\frac{1}{2} \tag{89}$$

which can also be deduced from (87) by changing k into k'. For $z=1$, the application of (21)–(23) yields

$$\int_0^1 \frac{kk'^2\,dk}{(1+k^2)\left[(E-k'^2K)^2+(E'-k^2K')^2\right]}=\frac{1-8\pi^2/\Gamma^4\left(\frac{1}{4}\right)}{1+8\pi^2/\Gamma^4\left(\frac{1}{4}\right)}. \tag{90}$$

Adding (88) to z^2 times (86) one obtains

$$\frac{K(iz)}{E(iz)}=\int_0^1 \frac{k\,dk}{(1+k^2z^2)\left[(E-k'^2K)^2+(E'-k^2K')^2\right]}. \tag{91}$$

By a contiguity relation [7, p. 32, eq. 14] we have

$$E(k)-K(k)=-\frac{\pi}{4}k^2F\left(\frac{1}{2},\frac{3}{2},2;k^2\right). \tag{92}$$

For $\alpha=1/2$, $\beta=3/2$, $\gamma=2$ and $z=-k^2$, the hypergeometric ratio in the right-hand side of (72) becomes, by (14), by the definition of G and by (92)

$$\frac{F\left(\frac{3}{2},\frac{3}{2},3;-z\right)}{F\left(\frac{1}{2},\frac{3}{2},2;-z\right)}=4\left[1-2(1+z)G\left(\frac{1}{2},\frac{3}{3},2;-z\right)\right]$$

$$=4\left\{1-\frac{k'^2}{k}\,\frac{d\left[k^{-2}(E-K)\right]/dk}{k^{-2}(E-K)}\right\}.$$

By (18) and (82), this is

$$4\left[\frac{K\left(i\sqrt{z}\right)}{E\left(i\sqrt{z}\right)-K\left(i\sqrt{z}\right)}-\frac{2}{z}\right]. \tag{93}$$

On the other hand, the left-hand side of (72) with $t=k^2$ becomes by (92)

$$4\int_0^1 \frac{dk}{k(1+zk^2)\left|E\left(\frac{1}{k}\right)-K\left(\frac{1}{k}\right)\right|^2}. \tag{94}$$

Identifying (93) with (94), using (77) and (85) and replacing z by z^2 one obtains

$$\frac{1}{E(iz)/K(iz)-1}-\frac{2}{z^2}=\int_0^1 \frac{k\,dk}{(1+k^2z^2)\left[(E-K)^2+E'^2\right]^2}. \tag{95}$$

For $z=0$, i and 1, (95) yields, respectively,

$$\int_0^1 \frac{k\,dk}{(E-K)^2+E'^2}=\frac{9}{16}, \tag{96}$$

$$\int_0^1 \frac{k\,dk}{k'^2\left[(E-K)^2+E'^2\right]}=1, \tag{97}$$

$$\int_0^1 \frac{k\,dk}{(1+k^2)\left[(E-K)^2+E'^2\right]}=\frac{\Gamma^4(1/4)}{8\pi^2}-2. \tag{98}$$

Many other similar integrals can be deduced by replacing k by k' and by combining the results in various ways.

7. Continued fractions for ratios of Γ-functions. The expressions (9) are positive real functions of γ, so that (8) also converges for $z>0$, $\mathrm{Re}\,\gamma>0$. In (9), c_{2k+1} is not positive real in α, but the transformation

$$u=\frac{\alpha}{\gamma+1-\alpha}, \tag{99}$$

which changes the range $0<\alpha<\gamma+1$ of (5) into $0<u<\infty$, changes the denominator of c_{2k+1} ino

$$(\gamma+1+k)k+\frac{(\gamma+1)^2u}{(1+u)^2}$$

which is a positive real function of u. Consequently, (8) also converges for $z>0$, $\gamma>0$, $\mathrm{Re}\,u>0$.

For $z=1$ and $\gamma=1$, the hypergeometric functions of (8) are expressible in terms of Γ-functions [2, eq. 15.1.22]. After some elementary manipulations and the replacement of α by $z+1$ one obtains

(100)

$$\frac{F(z,z+1,1;-1)}{F(z+1,z+1,2;-1)}=1+\frac{1}{2/(1^2-z^2)+\frac{1}{3}+\frac{1}{4/(2^2-z^2)+\frac{1}{5}+\frac{1}{6/(3^2-z^2)+\cdots}}}$$

$$=\frac{4}{z}\frac{\Phi(z)-1}{\Phi(z)+1}$$

with

$$\Phi(z)=\frac{\Gamma\left(\frac{z}{2}\right)\Gamma\left(\frac{1-z}{2}\right)}{\Gamma\left(-\frac{z}{2}\right)\Gamma\left(\frac{1+z}{2}\right)}. \tag{101}$$

For $z=1$, $\gamma=0$, (8) becomes, via (7),

$$\frac{\alpha(1-\alpha)F(\alpha+1,\alpha,2;-1)}{F(\alpha,\alpha,1;-1)}$$

and can again be expressed in terms of Γ-functions [2, eq. 15.1.21]. Changing α into $(z+1)/2$ and dividing by 2, one obtains, by (9),

$$(102)\qquad \cfrac{1}{8/(1^2-z^2)+\cfrac{1}{1+\cfrac{24/(3^2-z^2)}{2+\cfrac{40/(5^2-z^2)}{3+\cdots}}}}=\frac{\Gamma\left(\frac{3+z}{4}\right)\Gamma\left(\frac{3-z}{4}\right)}{\Gamma\left(\frac{1+z}{4}\right)\Gamma\left(\frac{1-z}{4}\right)}.$$

Owing to the discussion following (99) and the transformations from α to z, both (100) and (102) converge for $|z|<1$. In fact, both (100) and (102) are positive real functions of the argument $u=(1+z)/(1-z)$.

For $z=0$, (102) is $2\pi^2/\Gamma^4(1/4)$ and the continued fraction multiplied by 4 coincides with (20) for $z=1/2$. This checks with the value (79) of (78) for $z=1$.

8. Continued fraction generated by γ-contiguity. From the many contiguity relations, various continued fractions for ratios of hypergeometric functions can be generated, different from the Gauss expansion, and we have investigated several of these. The only simple ordinary fraction thus obtained results from the contiguity relation for α, β (and z) constant and for variable γ; we abbreviate $F(\alpha,\beta,\gamma;z)$ into $F(\gamma)$. One has [2, eq. 15.2.27]

$$\gamma(\gamma-1)(1-z)F(\gamma-1)=\gamma[\gamma-1(2\gamma-\alpha-\beta-1)z]F(\gamma)+(\gamma-\alpha)(\gamma-\beta)zF(\gamma+1).$$

For $\beta=\alpha-1$, this can be rewritten as

$$\frac{\gamma-1}{\gamma-\alpha}\sqrt{\frac{1-z}{z}}\,\frac{F(\gamma-1)}{F(\gamma)}=\frac{\gamma-1-2(\gamma-\alpha)z}{(\gamma-\alpha)\sqrt{z(1-z)}}+\cfrac{1}{\cfrac{\gamma}{\gamma-\alpha+1}\sqrt{\frac{1-z}{z}}\,\frac{F(\gamma)}{F(\gamma+1)}}.$$

With γ changed into $\gamma+\alpha$, with $z=\frac{1}{2}$ and then α changed into $1+\alpha$, this becomes

$$\frac{\gamma+\alpha}{\gamma}\,\frac{F(\gamma+\alpha)}{F(\gamma+\alpha+1)}=\frac{2\alpha}{\gamma}+\cfrac{1}{\cfrac{\gamma+\alpha+1}{\gamma+1}\,\cfrac{F(\gamma+\alpha+1)}{F(\gamma+\alpha+2)}}$$

thus generating the expansion

$$\frac{\gamma+\alpha}{\gamma}\,\frac{F\left(\alpha+1,\alpha,\gamma+\alpha;\frac{1}{2}\right)}{F\left(\alpha+1,\alpha,\gamma+\alpha+1;\frac{1}{2}\right)}=\frac{2\alpha}{\gamma}+\cfrac{1}{2\alpha/(\gamma+1)+\cfrac{1}{2\alpha/(\gamma+2)+\cdots}}.$$

By [2, eq. 15.3.3], this is also

$$\frac{2(\gamma+\alpha)}{\gamma}\,\frac{F\left(\gamma-1,\gamma;\gamma+\alpha;\frac{1}{2}\right)}{F\left(\gamma,\gamma+1,\gamma+\alpha+1;\frac{1}{2}\right)}.$$

Changing α into $z/2$ one thus obtains

$$(103)\qquad \frac{(2\gamma+z)}{\gamma}\,\frac{F\left(\gamma-1,\gamma,\gamma+\frac{z}{2};\frac{1}{2}\right)}{F\left(\gamma,\gamma+1,\gamma+\frac{z}{2}+1;\frac{1}{2}\right)}=\frac{z}{\gamma}+\cfrac{1}{z/(\gamma+1)+\cfrac{1}{z/(\gamma+2)+\cdots}}.$$

For $\gamma>0$, (103) is a Stieltjes fraction converging for $\mathrm{Re}\,z>0$. The fact that it does converge to the left-hand side of (103) results from an expansion mentioned by Perron [1, p. 299, eq. 23] which, transformed into an ordinary fraction, becomes

$$(104)\qquad \frac{(\alpha+1)\int_0^1 x^\alpha\left(\frac{1-x}{1+x}\right)^\gamma dx}{\alpha\int_0^1 x^{\alpha-1}\left(\frac{1-x}{1+x}\right)^\gamma dx}=\frac{1}{2\gamma/(\alpha+1)+}\,\frac{1}{2\gamma/(\alpha+2)+}\,\frac{1}{2\gamma/(\alpha+3)+}\cdots \qquad [\alpha>0,\gamma>0].$$

If 2γ is replaced by z and α by $\gamma-1$, the right-hand side of (104) becomes the inverse of the right-hand side of (103), and the same relation between the left-hand sides of (103) and (104) results from the integral representation of the hypergeometric function [2, eq. 15.3.1] and its analytic continuation for $|z|>1$ [2, eq. 15.3.5].

For $\gamma=1$, (103) is

$$(105)\qquad \frac{2+z}{F\left(1,2,2+\frac{z}{2};\frac{1}{2}\right)}.$$

By a contiguity relation [2, eq. 15.2.11] we have

$$F\left(1,2,2+\frac{z}{2};\frac{1}{2}\right)=2+z-zF\left(1,1,2+\frac{z}{2},\frac{1}{2}\right)$$

and the last function is a combination of ψ-functions [2, eq. 15.1.27]. Finally, the reciprocal of (105) is

$$(106)\qquad \frac{1}{z/1+}\,\frac{1}{z/2+}\,\frac{1}{z/3+}\cdots=1+\frac{z}{2}\left[\psi\left(\frac{z}{4}+\frac{1}{2}\right)-\psi\left(\frac{z}{4}+1\right)\right]$$
$$=\frac{z}{2}\left[\psi\left(\frac{z}{4}+\frac{1}{2}\right)-\psi\left(\frac{z}{4}\right)\right]-1$$

by [2, eq. 6.3.5]. The expansion (106) is known ([8, p. 162, eq. 3.7], after reduction to an ordinary fraction) to be

$$(107)\qquad z\int_0^\infty e^{-zt}\tanh t\,dt$$

and the equivalence of (107) with the last form of (106) can be proved directly by replacing the ψ-functions by their integral representations [2, eq. 6.3.2] (see also [12]). By [2, eq. 6.3.7], (106) reduces for $z=1$ to the ordinary form of Euler's fraction for $\frac{\pi}{2}$ [1, p. 23, eq. 3].

We next consider (103) for $\gamma=\frac{1}{2}$. By contiguity relations [2, eqs. 15.2.21, 15.2.24] one has

$$F\left(-\frac{1}{2},\frac{1}{2},\frac{z+1}{2};\frac{1}{2}\right)=\frac{z-1}{2z}F\left(\frac{1}{2},\frac{1}{2},\frac{z-1}{2};\frac{1}{2}\right)+\frac{1}{2}F\left(\frac{1}{2},\frac{1}{2},\frac{z+1}{2};\frac{1}{2}\right),$$
$$F\left(\frac{1}{2},\frac{3}{2},\frac{z+3}{2};\frac{1}{2}\right)=(z+1)F\left(\frac{1}{2},\frac{1}{2},\frac{z+1}{2};\frac{1}{2}\right)-zF\left(\frac{1}{2},\frac{1}{2},\frac{z+3}{2};\frac{1}{2}\right).$$

Since [2, eq. 15.1.26]

$$F\left(\frac{1}{2},\frac{1}{2},\gamma;\frac{1}{2}\right)=\frac{2^{1-\gamma}\sqrt{\pi}\,\Gamma(\gamma)}{\left[\Gamma\left(\frac{\gamma}{2}+\frac{1}{4}\right)\right]^2},$$

(103) becomes

$$2(z+1)\frac{F\left(-\frac{1}{2},\frac{1}{2},\frac{z+1}{2};\frac{1}{2}\right)}{F\left(\frac{1}{2},\frac{3}{2},\frac{z+3}{2};\frac{1}{2}\right)}=2z+\frac{1}{2z/3+}\frac{1}{2z/5+}\cdots=\frac{H(z)+1}{H(z)-1}, \tag{108}$$

where

$$H(z)=\frac{z}{4}\left[\frac{\Gamma\left(\frac{z}{4}\right)}{\Gamma\left(\frac{z}{4}+\frac{1}{2}\right)}\right]^2 \tag{109}$$

and this coincides with the particular case $m=n=\frac{1}{2}$ in Ramanujan's continued fraction [1, p. 34, eq. 15] when transformed into an ordinary fraction. More generally, the ordinary form of Ramanujan's expansion remains simple for $m=1-n$, and two other particularly simple expansions are obtained for $n=0$ and $n=-\frac{1}{2}$.

9. Confluent hypergeometric functions. If one replaces z by $-z^2/\alpha^2$ in (8) and lets α tend to infinity, one obtains, via [2, eq. 9.1.69], the well-knownexpansion involving modified Bessel functions

$$\frac{I_{\gamma-1}(2z)}{I_\gamma(2z)}=\frac{\gamma}{z}+\frac{1}{(\gamma+1)/z+}\frac{1}{(\gamma+2)/z+}\cdots. \tag{110}$$

If one replaces z by γ/z in (8) divided by γ, and lets γ tend to infinity, one obtains, via [2, eq. 13.1.33], an expansion involving Wittaker's functions

$$\frac{W_{1-\alpha,1/2}(z)}{\sqrt{z}\,W_{(1/2)-\alpha,0}(z)}=1+\frac{1}{z/\alpha+}\frac{1}{1+}\frac{1}{z/(\alpha+1)+}\cdots. \tag{111}$$

Known particular cases are [9, eq. 25]

$$\frac{1}{2}\left[\frac{K_1(z/4)}{K_0(z/4)}+1\right]=1+\frac{1}{z+}\frac{1}{1+}\frac{1}{z/3+}\cdots \tag{112}$$

for $\alpha=1/2$ and z changed into $z/2$, and

$$\frac{e^{-z}}{zE_1(z)}=1+\frac{1}{z+}\frac{1}{1+}\frac{1}{z/2+}\cdots \tag{113}$$

for $\alpha=1$ [2, eq. 5.1.22].

The known expansion [10, App.]

$$\frac{1}{1-\dfrac{J_{\alpha+z+(1/2)}(z)}{J_{\alpha+z-(1/2)}(z)}}=1+\frac{1}{(2\alpha+1)/z+}\frac{1}{1+}\frac{1}{(2\alpha+3)/z+}\cdots \tag{114}$$

resulting immediately from the Bessel recurrence is a particular case of (3) where z is replaced by $-2iz/\beta$, with β tending to infinity, with $\gamma=2\alpha$, and with α replaced by $\alpha+z$. The reduction of the resulting continued fraction to (114) requires, however, some heavy algebraic transformations. The function (114) is positive real in z for $\alpha\geq-\frac{1}{2}$, and

in α for $z>0$. For $\alpha=\frac{1}{2}$, (114) becomes $J_z(z)/J_z'(z)$. For $\alpha=0$, $z=1$ one obtains the curious expansion

$$\tan 1 = 1+\frac{1}{1}+\frac{1}{1}+\frac{1}{3}+\frac{1}{1}+\frac{1}{5}+\frac{1}{1}+\frac{1}{7}+\cdots . \tag{115}$$

Acknowledgments. The author is grateful to J. Boersma, P. J. de Doelder (Technical University Eindhoven, The Netherlands) and M. L. Glasser (Clarkson College, Potsdam, NY) for suggestions and corrections.

REFERENCES

[1] O. PERRON, *Die Lehre von den Kettenbrüchen*, II, Teubner, Stuttgart, 1957.
[2] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1972.
[3] E. WITTAKER AND G. WATSON, *Modern Analysis*, Cambridge University Press, Cambridge 1952.
[4] I. M. RYSHIK AND I. S. GRADSTEIN, *Tables*, VEB Verlag, Berlin, 1963.
[5] Y. L. LUKE, *The Special Functions and Their Approximations*, vol. I, Academic Press, New York, 1969.
[6] A. ERDÉLYI ET AL., *Higher Transcendental Functions*, vol. II, McGraw-Hill, New York, 1953, p. 319.
[7] C. SNOW, *Hypergeometric and Legendre Functions*, National Bureau of Standards, Washington, D. C., 1952.
[8] A. W. KHOVANSKII, *The Application of Continued Fractions*, Noordhoff, Groningen, The Netherlands, 1963.
[9] V. BELEVITCH AND J. BOERSMA, *The Bessel ratio $K_{v+1}(z)/K_v$ as a passive impedance*, Philips J. Res., 34 (1979), pp. 163–173.
[10] V. BELEVITCH, *The lateral magnetic skin-effect in thin plates*, Philips Res. Repts., 31 (1976), pp. 199–215.
[11] _____, *A class of nonrational impedances*, Int. J. Circ. Th. Appl. 6 (1978), pp. 315–319.
[12] V. BELEVITCH AND J. BOERSMA, *On Stieltjes integral transforms involving Γ-functions*, Math. Comput., 38 (1982), pp. 223–226.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0012 $01.00/0

ORTHOGONAL POLYNOMIALS ASSOCIATED WITH SINGULAR INTEGRAL EQUATIONS HAVING A CAUCHY KERNEL*

DAVID ELLIOTT†

Abstract. The Chebyshev polynomials of both the first and second kind are of fundamental importance when considering the particular case of the singular integral equation $a(t)\phi(t)+(b(t)/\pi)\lambda\int_{-1}^{1}(\phi(\tau)/(\tau-t))\,d\tau=f(t)$, $-1<t<1$, in which $a\equiv 0$ and $b\equiv -1$ on $[-1,1]$. We identify the two sets of orthogonal polynomials which play a corresponding role for the singular integral equation with general a, b and consider some of the relationships between these two sets of polynomials.

1. Introduction. In this paper we shall consider some properties of orthogonal polynomials which arise from the dominant singular integral equation defined on the arc $(-1, 1)$ and given by

$$a(t)\phi(t)+\frac{b(t)}{\pi}\int_{-1}^{1}\frac{\phi(\tau)\,d\tau}{\tau-t}=f(t), \qquad -1<t<1. \tag{1.1}$$

The integral arising in (1.1) is a Cauchy principal value integral which is defined by

$$\int_{-1}^{1}\frac{\phi(\tau)\,d\tau}{\tau-t}=\lim_{\varepsilon\to 0+}\left\{\int_{-1}^{t-\varepsilon}+\int_{t+\varepsilon}^{1}\right\}\frac{\phi(\tau)\,d\tau}{\tau-t}. \tag{1.2}$$

In (1.1) the functions a, b and f are assumed known and it is required to find the function ϕ. The theory of such equations has been given by many authors (see, for example Gakhov [3], Muskhelishvili [6]) and the equations arise in many branches of applied mathematics such as aerodynamics, elasticity, fracture mechanics, theory of radiative transfer, etc. One of the most widely studied of these equations is the particular one (which we shall call our prototype) in which $a\equiv 0$ and $b\equiv -1$. In this case we find that the Chebyshev polynomials T_n and U_n, of the first and second kind respectively, play an important role. We have the following interesting equations (see, for example, Tricomi [9]):

$$\begin{aligned}\frac{1}{\pi}\int_{-1}^{1}\frac{T_n(\tau)\,d\tau}{(1-\tau^2)^{1/2}(\tau-t)}&=U_{n-1}(t), \qquad n=1,2,3,\cdots,\\ \frac{1}{\pi}\int_{-1}^{1}\frac{(1-\tau^2)^{1/2}U_n(\tau)\,d\tau}{\tau-t}&=-T_{n+1}(t), \qquad n=0,1,2,\cdots.\end{aligned} \tag{1.3}$$

Equations (1.3) have been widely used to obtain approximate solutions to (1.1) when $a\equiv 0$ and $b\equiv -1$ and equations for which the left-hand side of (1.1) is the dominant part (see, for example, Linz [5]). Again, when a and b are arbitrary constants, the Chebyshev polynomials in (1.3) are replaced by certain Jacobi polynomials (see Tricomi [8] and Karpenko [4]). It is the purpose of this paper to consider the orthogonal polynomials which arise when a and b of (1.1) are arbitrary functions and to obtain relations between them similar to (1.3). A particular application of these polynomials to

* Received by the editors July 10, 1980, and in revised form July 15, 1981.
† Mathematics Department, University of Tasmania, Box 252C, GPO, Hobart, Tasmania, Australia, 7001.

finding approximate solutions of singular integral equations having the left-hand side of (1.1) as a dominant part will be discussed elsewhere (see Elliott [2]); but first we shall need some results from the theory of singular integral equations.

2. Some basic results. In quoting results we shall follow the analysis given by Dow and Elliott [1] where some of the quantities to be defined differ slightly from those given by Muskhelishvili [6] and Gakhov [3]. First we assume that $b(-1)\leq 0$; if $b(-1)>0$ we multiply (1.1) by -1. From a and b we first evaluate the *index* κ of (1.1). To do this we let θ denote the real function continuous on $[-1,1]$ with $-1<\theta(-1)\leq 0$ such that $\theta(t)=(1/\pi)\arg[a(t)+ib(t)]$ for $t\in[-1,1]$. Although the arg function is multivalued, the restrictions that θ be continuous on $[-1,1]$ with its value at $t=-1$ being in $(-1,0]$ are sufficient to define θ uniquely. The index κ is given by

$$\kappa=-[\theta(1)], \tag{2.1}$$

where $[x]$ in (2.1) denotes the largest integer not exceeding x. In the prototype case where $a\equiv 0$, $b\equiv -1$, we have $\theta(t)=-\frac{1}{2}$ and $\kappa=1$. From (2.1) we see that the integer κ may be positive, negative or zero.

In terms of the function θ and the index κ, we define (i) the *canonical function* X by

$$X(z)=(1-z)^{-\kappa}\exp\left\{-\int_{-1}^{1}\frac{\theta(\tau)\,d\tau}{\tau-z}\right\}, \qquad z\notin[-1,1], \tag{2.2}$$

and (ii) the *fundamental function* Z by

$$Z(t)=(1-t)^{-\kappa}\exp\left\{-\int\!\!\!\!\!\!-_{-1}^{1}\frac{\theta(\tau)\,d\tau}{\tau-t}\right\}, \qquad t\in(-1,1). \tag{2.3}$$

If, for $t\in(-1,1)$, $X^{\pm}(t)$ is defined as $\lim_{\varepsilon\to+0}X(t\pm i\varepsilon)$, then from the Sokhotski–Plemelj formulae we find that

$$X^{\pm}(t)=\{a(t)\mp ib(t)\}Z(t)/r(t), \qquad t\in(-1,1), \tag{2.4}$$

where the function r is defined by

$$r(t)=\left(a^2(t)+b^2(t)\right)^{1/2}, \qquad t\in[-1,1]. \tag{2.5}$$

Throughout the analysis we shall assume that $r(t)>0$ for all $t\in[-1,1]$ so that (1.1) is of *normal* type. Although we shall develop further properties of the canonical function later, it is useful to observe from (2.2) that, for large $|z|$,

$$X(z)=(-1)^{\kappa}z^{-\kappa}+\text{lower order terms}, \tag{2.6}$$

so that X is of order $-\kappa$ at infinity and X^{-1} will be of order κ.

Let f be any arbitrary function which is defined in the deleted complex plane (i.e., the complex plane with the interval $[-1,1]$ deleted) and suppose that it has an expansion about the point at infinity of the form

$$f(z)=\sum_{j=-\infty}^{N}f_jz^j,$$

where N is an integer (positive, negative or zero). The *principal part* of f evaluated at some point t (usually taken in $(-1,1)$) will be denoted and defined by

$$\text{p.p.}(f;t)=\sum_{j=0}^{N}f_jt^j. \tag{2.7}$$

If $N<0$, then p.p.$(f;t)\equiv 0$. From this definition it follows immediately that, for $k=1,2,3,\cdots$,

$$\text{p.p.}\{z^{k-1}(z-t)f;t\}=\begin{cases} f_{-k} & \text{if } N+k\geq 0,\\ 0 & \text{if } N+k\leq -1.\end{cases} \tag{2.8}$$

In terms of the principal part we shall now quote two results from Dow and Elliott [1, Thms. 3.1 and 3.2]. If, following Tricomi [9], we write

$$\Im(\phi;t)=\frac{1}{\pi}\int_{-1}^{1}\frac{\phi(\tau)\,d\tau}{\tau-t}, \qquad -1<t<1, \tag{2.9}$$

then, for any polynomial P, we have

$$\Im(bZP/r;t)=-a(t)Z(t)P(t)/r(t)+\text{p.p.}(PX;t), \tag{2.10}$$

$$\Im(bP/rZ;t)=a(t)P(t)/r(t)Z(t)-\text{p.p.}(PX^{-1};t) \tag{2.11}$$

for $-1<t<1$. Equations (2.10) and (2.11) are special cases of [1, Theorems 3.1 and 3.2] respectively; we have no need here for the greater generality of [1]. At this point it is convenient to quote the Poincaré–Bertrand and Parseval formulae whose proofs may be found in [6] and [9] respectively. For appropriate functions ϕ_1 and ϕ_2 defined on $(-1,1)$ we have, on using the notation of [9, §4.3] where the arguments are supressed, that the Poincaré–Bertrand formula is given by

$$\Im(\phi_1\Im\phi_2)+\Im(\phi_2\Im\phi_1)=(\Im\phi_1)(\Im\phi_2)-\phi_1\phi_2. \tag{2.12}$$

The Parseval formula is given by

$$\int_{-1}^{1}\{\phi_1(t)\Im(\phi_2;t)+\phi_2(t)\Im(\phi_1;t)\}\,dt=0. \tag{2.13}$$

With these preliminaries established, we are almost ready to proceed with our analysis, but before we do, we must place a restriction (which in practice does not appear to be too severe) on (1.1). This assumption will be taken to be satisfied throughout the remainder of this paper.

Assumption A. There exists a function c defined on $(-1,1)$ such that

(i) cb is a polynomial B say, of degree μ, with all its zeros on $[-1,1]$,

(ii) the functions Z/cr and $1/Zcr$ are nonnegative and integrable on $[-1,1]$.

Firstly, we might note in passing that in [1] we assumed that b was a polynomial; the introduction of the function c satisfying Assumption A(i) allows us to consider a more general equation. Secondly, let us consider in more detail Assumption A(ii). As previously noted we are assuming throughout that $r>0$. From (2.3) and (2.1) we find that we can write

$$Z(t)=(1-t)^{\alpha}(1+t)^{\beta}\Omega(t), \tag{2.14}$$

say, where

$$\alpha=[\theta(1)]-\theta(1), \qquad \beta=\theta(-1),$$

$$\Omega(t)=\exp\Big\{(\theta(1)-\theta(t))\log(1-t) + (\theta(t)-\theta(-1))\log(1+t)-\int_{-1}^{1}\frac{\theta(\tau)-\theta(t)}{\tau-t}\,d\tau\Big\}. \tag{2.15}$$

It is obvious that $-1<\alpha, \beta\leq 0$ and Ω is positive on $[-1,1]$ so that the functions Z and $1/Z$ are nonnegative and integrable on $[-1,1]$. In the prototype equation with $a\equiv 0$, $b\equiv -1$ so that $\kappa=1$, we can choose $c\equiv 1$ and we find $Z=(1-t^2)^{-1/2}$, $-1<t<1$. Assumption A is trivially satisfied in this particular case.

From Assumption A(ii), if we write $w_1=Z/cr$, then from (2.15) we have

$$w_1(t)=(1-t)^{\alpha_1}(1+t)^{\beta_1}\Omega_1(t) \tag{2.16}$$

say, where $\alpha_1, \beta_1>-1$ and Ω_1 is nonnegative and integrable on $[-1,1]$ with $\Omega_1(-1)\neq 0$, $\Omega_1(1)\neq 0$. Similarly, if we write $w_2=1/Zcr$, then

$$w_2(t)=(1-t)^{\alpha_2}(1+t)^{\beta_2}\Omega_2(t), \tag{2.17}$$

where $\alpha_2, \beta_2>-1$ and Ω_2 is nonnegative and integrable on $[-1,1]$ with $\Omega_2(-1)\neq 0$, $\Omega_2(1)\neq 0$. These functions w_1 and w_2 will be taken as weight functions for sequences of orthogonal polynomials associated with (1.1) whose properties we shall explore in the next two sections.

3. The orthogonal polynomials p_n, q_n. It is a well-known result (see, for example, Szegö [7]) that since w_1, as defined by (2.16), is nonnegative and integrable on $[-1,1]$ there exists a sequence of polynomials $\{p_n\}$ say, where the coefficient of x^n in $p_n(x)$ is strictly positive, such that

$$\int_{-1}^{1} w_1(\tau)p_j(\tau)p_k(\tau)\,d\tau=h_j\delta_{j,k}, \qquad j,k=0,1,2,\cdots. \tag{3.1}$$

The numbers h_j can be prescribed and $\delta_{j,k}$ denotes Kronecker's delta. For our prototype equation we can choose the polynomials p_n to be the Chebyshev polynomials T_n of the first kind. Given the set of polynomials p_n we now want to construct a sequence of polynomials $\{q_n\}$ say which correspond in the prototype equation to the Chebyshev polynomials U_n of the second kind. This we do through the next definition; later we shall establish orthogonality and other properties.

DEFINITION. For $n\geq\max(0,-\kappa)$ the polynomials q_n are defined by

$$q_n(z)=\text{p.p.}\{(-1)^{\kappa}Xp_{n+\kappa};z\} \quad \text{for all } z. \tag{3.2}$$

Since we have observed (see (2.6)) that X is exactly of order $(-\kappa)$ at infinity with coefficient $(-1)^{\kappa}$, it follows that $(-1)^{\kappa}Xp_{n+\kappa}$ is exactly of order n at infinity and the coefficient of z^n is strictly positive. From (3.2) we can immediately relate the coefficients of the polynomial q_n to those of $p_{n+\kappa}$. If $X(z)=\sum_{j=-\infty}^{-\kappa} a_j^* z^j$, $a_{-\kappa}^*=(-1)^{\kappa}$, and if we write

$$p_n(z)=\sum_{j=0}^{n}\alpha_{j,n}z^j, \qquad q_n(z)=\sum_{j=0}^{n}\beta_{j,n}z^j, \tag{3.3}$$

then for $n\geq\max(0,\kappa)$ we find

$$\beta_{j,n}=(-1)^{\kappa}\sum_{k=0}^{n-j} a_{-\kappa-k}^*\alpha_{\kappa+j+k,n+\kappa} \quad \text{for } j=0(1)n. \tag{3.4}$$

We can also, from the above definition, define p_n in terms of q_n by a similar relationship as the following theorem shows.

THEOREM 3.1. *Suppose* $X^{-1}(z)=\sum_{j=-\infty}^{\kappa} a_j z^j$, $a_\kappa=(-1)^\kappa$. *Then for* $n\geq\max(0,\kappa)$, *we have*

$$\alpha_{j,n}=(-1)^\kappa \sum_{k=0}^{n-j} a_{\kappa-k}\beta_{j-\kappa+k,n-\kappa} \quad \text{for } j=0(1)n, \tag{3.5}$$

so that

$$p_n(z)=\text{p.p.}\{(-1)^\kappa X^{-1}q_{n-\kappa};z\} \quad \text{for all } z. \tag{3.6}$$

Proof. From the expansion about the point at infinity of the functions X, X^{-1}, it follows on equating coefficients of powers of z in the identity $1\equiv X(z)X^{-1}(z)$ that

$$\sum_{k=0}^{s} a^*_{-\kappa-k}a_{\kappa-s+k}=\sum_{k=0}^{s} a^*_{-\kappa-s+k}a_{\kappa-k}=\delta_{s,0}, \tag{3.7}$$

$s=0,1,2,\cdots$. From (3.4) we find

$$\begin{aligned}(-1)^\kappa \sum_{k=0}^{n-j} a_{\kappa-k}\beta_{j-\kappa+k,n-\kappa} &= \sum_{k=0}^{n-j} a_{\kappa-k} \sum_{l=0}^{n-j-k} a^*_{-\kappa-l}\alpha_{j+k+l,n}\\ &= \sum_{k=0}^{n-j} a_{\kappa-k} \sum_{s=k}^{n-j} a^*_{-\kappa+k-s}\alpha_{j+s,n}\\ &= \sum_{s=0}^{n-j} \alpha_{j+s,n} \sum_{k=0}^{s} a^*_{-\kappa-s+k}a_{\kappa-k}\\ &= \alpha_{j,n} \quad \text{on using (3.7).}\end{aligned}$$

This establishes (3.5) from which (3.6) follows immediately. □

There is an alternative representation for the polynomials q_n which is valid on $(-1,1)$ and is of considerable importance in our analysis.

THEOREM 3.2. *For* $n\geq\max(0,\mu-\kappa)$ *and* $t\in(-1,1)$,

$$(-1)^\kappa q_n(t)=a(t)Z(t)p_{n+\kappa}(t)/r(t)+b(t)c(t)\Im(w_1 p_{n+\kappa};t). \tag{3.8}$$

Proof. Recall (see Assumption A(i)) that μ is the degree of the polynomial $B=bc$. From (2.10), on choosing $P=p_{n+\kappa}$, and (3.2), we have $(-1)^\kappa q_n(t)=\text{p.p.}(Xp_{n+\kappa};t)=a(t)Z(t)p_{n+\kappa}(t)/r(t)+\Im(w_1 Bp_{n+\kappa};t)$ for $t\in(-1,1)$. Now

$$B(t)\Im(w_1 p_{n+\kappa};t)-\Im(w_1 Bp_{n+\kappa};t)=\frac{1}{\pi}\int_{-1}^{1} w_1(\tau)\left(\frac{B(t)-B(\tau)}{\tau-t}\right)p_{n+\kappa}(\tau)\,d\tau.$$

Since B is a polynomial of degree μ, then $(B(t)-B(\tau))/(\tau-t)$ is a polynomial of degree $\leq\mu-1$ in τ, and from the orthogonality of the polynomials p_n with respect to w_1, we have

$$\Im(w_1 Bp_{n+\kappa};t)=B(t)\Im(w_1 p_{n+\kappa};t), \qquad t\in(-1,1),$$

provided $n\geq\mu-\kappa$, which is assumed to be satisfied. The result follows immediately. □

We now come to our first important result which establishes the orthogonality property of the polynomials q_n.

THEOREM 3.3. *For* $j,k\geq\max(0,\mu-\kappa)$ *we have*

$$\int_{-1}^{1} w_2(\tau)q_j(\tau)q_k(\tau)\,d\tau=h_{j+\kappa}\delta_{j,k}. \tag{3.9}$$

Proof. Now

$$\int_{-1}^{1} w_2 q_j q_k \, d\tau = (-1)^{\kappa} \int_{-1}^{1} \frac{q_j}{cZr} \left\{ \frac{aZp_{k+\kappa}}{r} + bc\Im(w_1 p_{k+\kappa}) \right\} d\tau \quad \text{by (3.8)},$$

$$= (-1)^{\kappa} \int_{-1}^{1} \left\{ \frac{aq_j p_{k+\kappa}}{cr^2} - w_1 p_{k+\kappa} \Im\left(\frac{bq_j}{Zr} \right) \right\} d\tau \quad \text{by (2.13)},$$

$$= (-1)^{\kappa} \int_{-1}^{1} w_1 p_{k+\kappa} \text{p.p.}\left(q_j X^{-1} \right) d\tau \quad \text{by (2.11)},$$

$$= \int_{-1}^{1} w_1 p_{k+\kappa} p_{j+\kappa} \, d\tau \quad \text{by (3.6)}.$$

From the orthogonality of the polynomials p_n with respect to w_1, the result follows immediately. □

At this point it is convenient to obtain an expression for p_n in terms of q_n on $(-1, 1)$ which is similar to (3.8).

THEOREM 3.4. *For $n \geq \max(\kappa, \mu)$ and $t \in (-1, 1)$,*

$$(-1)^{\kappa} p_n(t) = a(t) q_{n-\kappa}(t) / r(t) Z(t) - b(t) c(t) \Im(w_2 q_{n-\kappa}; t). \tag{3.10}$$

Proof. From (3.8) we have

$$\begin{aligned} &(-1)^{\kappa} \left\{ \frac{aq_{n-\kappa}}{rZ} - cb\Im\left(\frac{q_{n-\kappa}}{crZ} \right) \right\} \\ &\quad = \frac{a^2 p_n}{r^2} + \frac{acb}{rZ} \Im\left(\frac{Zp_n}{cr} \right) - cb\Im\left(\frac{ap_n}{cr^2} \right) - cb\Im\left(\frac{b}{rZ} \Im\left(\frac{Zp_n}{cr} \right) \right). \end{aligned} \tag{3.11}$$

Now, from (2.11) with $P \equiv 1$, we have

$$\Im\left(\frac{b}{rZ} \right) = \frac{a}{rZ} - \text{p.p.}(X^{-1}),$$

and combining this with the Poincaré–Bertrand formula (2.12), we find

$$\Im\left(\frac{b}{rZ} \Im\left(\frac{Zp_n}{cr} \right) \right) = \frac{a}{rZ} \Im\left(\frac{Zp_n}{cr} \right) - \frac{bp_n}{cr^2} - \Im\left(\frac{ap_n}{cr^2} \right), \tag{3.12}$$

since $n \geq \kappa$ and polynomials p_n are orthogonal on $(-1, 1)$ with respect to w_1. Substituting (3.12) into (3.11) and recalling that $r^2 = a^2 + b^2$, we obtain (3.10). □

At this point we might observe that we now have the required generalization of (1.3). From (3.8) and (3.10) we have, for $t \in (-1, 1)$,

$$\begin{aligned} &\frac{a(t)Z(t)}{r(t)} p_n(t) + \frac{b(t)c(t)}{\pi} \lambda \int_{-1}^{1} \frac{Z(\tau) p_n(\tau) \, d\tau}{c(\tau) r(\tau)(\tau - t)} = (-1)^{\kappa} q_{n-\kappa}(t) \\ &\qquad \text{for } n \geq \max(\mu, \kappa), \\ &\frac{a(t) q_n(t)}{r(t) Z(t)} - \frac{b(t)c(t)}{\pi} \lambda \int_{-1}^{1} \frac{q_n(\tau) \, d\tau}{Z(\tau) c(\tau) r(\tau)(\tau - t)} = (-1)^{\kappa} p_{n+\kappa}(t) \\ &\qquad \text{for } n \geq \max(0, \mu - \kappa). \end{aligned} \tag{3.13}$$

We conclude this section with two further theorems, both concerning elementary properties of the polynomials p_n and q_n.

THEOREM 3.5. *Suppose that for $n\geq 1$ and all z the polynomials p_n satisfy a three term recurrence relation of the form*

$$p_{n+1}(z)-(A_n z+B_n)p_n(z)+C_n p_{n-1}(z)=0. \tag{3.14}$$

Then for $n\geq\max(1,\mu-\kappa+1)$ and all z, we have

$$q_{n+1}(z)-(A_{n+\kappa}z+B_{n+\kappa})q_n(z)+C_{n+\kappa}q_{n-1}(z)=0. \tag{3.15}$$

Proof. We shall establish (3.15) for $z=t$, where $t\in(-1,1)$. Since all quantities occurring in (3.15) are polynomials, the result can then be continued into the entire complex plane.

From (3.8), provided that $n-1\geq\max(0,\mu-\kappa)$, we have

$$\begin{aligned}(-1)^{\kappa}&\{q_{n+1}(t)-(A_{n+\kappa}t+B_{n+\kappa})q_n(t)+C_{n+\kappa}q_{n-1}(t)\}\\
&=\frac{a(t)Z(t)}{r(t)}\{p_{n+\kappa+1}(t)-(A_{n+\kappa}t+B_{n+\kappa})p_{n+\kappa}(t)+C_{n+\kappa}p_{n+\kappa-1}(t)\}\\
&\quad+b(t)c(t)\{\Im(w_1p_{n+\kappa+1};t)-(A_{n+\kappa}t+B_{n+\kappa})\Im(w_1p_{n+\kappa};t)\\
&\qquad\qquad+C_{n+\kappa}\Im(w_1p_{n+\kappa-1};t)\}.\\
&=\frac{b(t)c(t)}{\pi}A_{n+\kappa}\int_{-1}^{1}w_1(\tau)p_{n+\kappa}(\tau)\,d\tau,\quad\text{on using (3.14)},\\
&=0\quad\text{since } n+\kappa\geq 1.\quad\square\end{aligned}$$

THEOREM 3.6. *For $n\geq\max(\kappa,\mu)$ and all z,*

$$p_{n+1}(z)q_{n-\kappa}(z)-p_n(z)q_{n-\kappa+1}(z)=(-1)^{\kappa+1}\left(\frac{\alpha_{n+1,n+1}h_n}{\pi\alpha_{n,n}}\right)B(z). \tag{3.16}$$

Proof. Firstly, let us suppose that $t\in(-1,1)$. Then by (3.8)

$$\begin{aligned}p_{n+1}(t)&q_{n-\kappa}(t)-p_n(t)q_{n-\kappa+1}(t)\\
&=(-1)^{\kappa}b(t)c(t)\,\frac{1}{\pi}\,\text{⨍}_{-1}^{1}\frac{Z(\tau)}{c(\tau)r(\tau)}\cdot\left\{\frac{p_{n+1}(t)p_n(\tau)-p_n(t)p_{n+1}(\tau)}{\tau-t}\right\}d\tau\\
&=(-1)^{\kappa+1}\frac{b(t)c(t)}{\pi}\cdot\frac{\alpha_{n+1,n+1}h_n}{\alpha_{n,n}}\sum_{k=0}^{n}\frac{p_k(t)}{h_k}\,\text{⨍}_{-1}^{1}w_1(\tau)p_k(\tau)\,d\tau,\\
&\qquad\qquad\text{by the Christoffel–Darboux formula (see Szegö [7, §3.2])},\\
&=(-1)^{\kappa+1}\frac{b(t)c(t)}{\pi}\cdot\frac{\alpha_{n+1,n+1}h_n}{\alpha_{n,n}}\end{aligned}$$

by the orthogonality of the polynomials p_n with respect to w_1. This result has been proved for all $t\in(-1,1)$. Since all quantities occurring in (3.16) are polynomials, the region of validity of the equation can be extended into the complex plane so that it is true for all z. $\square$

Equation (3.16) provides some information on the relationship between the zeros of p_n and $q_{n-\kappa}$. It is well known (see [7]) that all the zeros of each polynomial are real, distinct and lie in $(-1,1)$. Suppose $\tau_{j,n}, j=1(1)n$, are the zeros of p_n and $t_{i,n-\kappa}, i=1(1)(n-\kappa)$, are those of $q_{n-\kappa}$. From (3.16) with $z=\tau_{j,n}$ we find

$$p_{n+1}(\tau_{j,n})q_{n-\kappa}(\tau_{j,n})=(-1)^{\kappa+1}\left(\frac{\alpha_{n+1,n+1}h_n}{\pi\alpha_{n,n}}\right)B(\tau_{j,n}).$$

Now $p_{n+1}(\tau_{j,n}) \neq 0$ so that $\tau_{j,n}$ can only be a zero of $q_{n-\kappa}$ if it is a zero of B. Similarly $t_{i,n-\kappa}$ can only be a zero of p_n if it is a zero of B. Thus, in general, the zeros of p_n and $q_{n-\kappa}$ are distinct unless either polynomial has a zero which coincides with a zero of B.

4. Expansions about the point at infinity. The canonical function X and its inverse X^{-1} are of fundamental importance in the analysis of singular integral equations. Previously (see §3) we have introduced the expansions of these functions about the point at infinity, and we shall now give computable forms for the coefficients a_j^* and a_j in these expansions.

THEOREM 4.1. *Suppose*

$$X(z) = \sum_{j=-\infty}^{-\kappa} a_j^* z^j \quad \text{and} \quad X^{-1}(z) = \sum_{j=-\infty}^{\kappa} a_j z^j,$$

where $a_{-\kappa}^* = a_\kappa = (-1)^\kappa$. *Then*

$$a_j^* = \frac{1}{\pi} \int_{-1}^{1} w_1(\tau) B(\tau) \tau^{-j-1} \, d\tau \quad \text{for} \quad j \leq \min(-\kappa, -1) \tag{4.1}$$

and

$$a_j = -\frac{1}{\pi} \int_{-1}^{1} w_2(\tau) B(\tau) \tau^{-j-1} \, d\tau \quad \text{for} \quad j \leq \min(\kappa, -1). \tag{4.2}$$

Proof. From (2.10), on choosing $P = z^{k-1}(z-t)$ for $k = 1, 2, 3, \cdots$, we find

$$\begin{aligned} \frac{1}{\pi} \int_{-1}^{1} w_1(\tau) B(\tau) \tau^{k-1} \, d\tau &= \text{p.p.}\{z^{k-1}(z-t) X(z); t\} \\ &= a_{-k}^* \quad \text{provided} \quad k \geq \kappa, \end{aligned}$$

using (2.8). Equation (4.1) now follows. The proof of (4.2) follows similarly on using (2.11). □

In numerical work the integrals in (4.1) and (4.2) may be evaluated by quadrature thereby giving approximate values for these coefficients. We can also use techniques similar to those in the above proof to obtain some further relations between the coefficients in the polynomials p_n and q_m. Since (see (3.2)) we have $q_n = \text{p.p.}((-1)^\kappa X p_{n+\kappa})$ and since (see (3.3)) we have written $q_n(z) = \sum_{j=0}^{n} \beta_{j,n} z^j$, let us write

$$(-1)^\kappa p_{n+\kappa}(z) X(z) = \sum_{j=-\infty}^{n} \beta_{j,n} z^j. \tag{4.3}$$

We now have the following expressions for the coefficients $\beta_{j,n}$.

THEOREM 4.2. *For* $n + \kappa \geq 0$ *and* $j = -1, -2, -3, \cdots$,

$$\beta_{j,n} = \frac{(-1)^\kappa}{\pi} \int_{-1}^{1} w_1(\tau) B(\tau) \tau^{-j-1} p_{n+\kappa}(\tau) \, d\tau. \tag{4.4}$$

Proof. On choosing $P = (-1)^\kappa p_{n+\kappa} z^{j-1}(z-t)$ in (2.10) and using (2.8), the result follows at once. □

Since B is of degree μ, then because of the orthogonality of the polynomials p_n with respect to w_1 on $(-1, 1)$, it is an immediate consequence of (4.4) that for $\mu - (n+\kappa) \leq k \leq -1$, the coefficients $\beta_{k,n}$ are zero.

We can proceed similarly starting from (3.6). If we write, for $n - \kappa \geq 0$,

$$(-1)^\kappa q_{n-\kappa}(z) X^{-1}(z) = \sum_{j=-\infty}^{n} \alpha_{j,n} z^j, \tag{4.5}$$

then, arguing as above, we find that, for $j=-1,-2,-3,\cdots$,

$$\alpha_{j,n}=\frac{(-1)^{\kappa+1}}{\pi}\int_{-1}^{1} w_2(\tau)B(\tau)\tau^{-j-1}q_{n-\kappa}(\tau)\,d\tau. \tag{4.6}$$

The identification of these integrals is of some interest in numerical work (see Dow and Elliott [1]).

5. Some comments on approximate methods for solving (1.1). Numerical methods for solving the dominant equation (1.1) and the so-called complete equation, where a term $\int_{-1}^{1} k(t,\tau)\phi(\tau)\,d\tau$ is added to the left-hand side of (1.1), have been extensively discussed in the literature for the case of the prototype equation, see, for example, Linz [5]. In these algorithms extensive use is made of the properties of the Chebyshev polynomials. It is the purpose of this section to consider similar methods for our more general equation using the polynomials p_n and q_n. Elsewhere (see [2]) we shall consider in detail the method of classical collocation, and in this section, we shall consider a Galerkin type method for the dominant equation only.

If we take (1.1), multiply it through by the function c and then write $\phi=Z\psi/cr$ so that ψ is the new dependent variable, we obtain an equation which we write as

$$A\psi=F, \qquad F=cf, \tag{5.1}$$

where the linear operator A is defined by

$$A\psi=\frac{aZ}{r}\psi+B\Im\left(\frac{Z}{cr}\psi\right). \tag{5.2}$$

In terms of the linear operator A, we observe that the first of equations (3.13) can be written as

$$Ap_j=(-1)^{\kappa}q_{j-\kappa}, \qquad j\geq\max(\mu,\kappa). \tag{5.3}$$

This suggests how we might proceed to obtain an approximate solution to (5.1). Suppose we write $\psi_n=\Sigma_{j=0}^{n}\alpha_j p_j$; then the residual r_n is given by

$$r_n=\sum_{j=0}^{n}\alpha_j Ap_j-F=\sum_{j=0}^{\max(\mu,\kappa)-1}\alpha_j Ap_j+(-1)^{\kappa}\sum_{j=\max(\mu,\kappa)}^{n}\alpha_j q_{j-\kappa}-F, \tag{5.4}$$

using (5.3). We obtain a system of $(n-\kappa+1)$ linear algebraic equations for the $(n+1)$ unknown coefficients α_j by supposing that

$$\int_{-1}^{1} w_2(\tau)r_n(\tau)q_j(\tau)\,d\tau=0 \quad \text{for } j=0(1)(n-\kappa). \tag{5.5}$$

(Note that in the case when $\kappa>0$ we need a further κ condition to give a system of $(n+1)$ linear algebraic equations for the $(n+1)$ unknown coefficients α_j. When $\kappa<0$ we can "remove" $-\kappa$ equations since F has to satisfy $-\kappa$ consistency conditions (see [6])). In terms of the operator A, we have from (2.10) that

$$Ap_j=\text{p.p.}(p_jX;t)-\frac{1}{\pi}\int_{-1}^{1} w_1(\tau)\left(\frac{B(\tau)-B(t)}{\tau-t}\right)p_j(\tau)\,d\tau$$

so that Ap_j is a polynomial of degree $\leq\max(j-\kappa,\mu-1)$. From (5.4) and (5.5), if $\beta_j=(1/h_{j+\kappa})\int_{-1}^{1} w_2(\tau)F(\tau)q_j(\tau)\,d\tau$, then we find simply that

$$\alpha_j=(-1)^{\kappa}\beta_{j-\kappa} \quad \text{for } \max(\mu,\kappa)\leq j\leq n. \tag{5.6}$$

For $j\leq\max(\mu,\kappa)-1$ we have slightly more complicated equations for α_j which we shall consider no further here. From (5.6) we obtain the following result.

THEOREM 5.1. *Suppose* $\psi \sim \Sigma_{j=0}^{\infty} \alpha_j p_j$, $F \sim \Sigma_{j=0}^{\infty} \beta_j q_j$, *where*

$$\alpha_j = \int_{-1}^{1} w_1 \psi p_j \, d\tau \Big/ \int_{-1}^{1} w_1 p_j^2 \, d\tau, \qquad \beta_j = \int_{-1}^{1} w_2 F q_j \, d\tau \Big/ \int_{-1}^{1} w_2 q_j^2 \, d\tau.$$

Then, for all $n \geq \max(\mu, \kappa)$,

$$\int_{-1}^{1} w_1(\tau)(\psi - \psi_n)^2 d\tau = \int_{-1}^{1} w_2(\tau)(F - F_{n-\kappa})^2 d\tau, \tag{5.7}$$

where $\psi_n = \Sigma_{j=0}^{n} \alpha_j p_j$ *and* $F_{n-\kappa} = \Sigma_{j=0}^{n-\kappa} \beta_j q_j$.

Proof. Now, for $n \geq \max(\mu, \kappa)$,

$$\begin{aligned}
\int_{-1}^{1} w_2(\tau)(F - F_{n-\kappa})^2 d\tau &= \sum_{j=n-\kappa+1}^{\infty} \beta_j^2 \int_{-1}^{1} w_2(\tau) q_j^2(\tau) \, d\tau \\
&= \sum_{j=n-\kappa+1}^{\infty} \alpha_{j+\kappa}^2 \int_{-1}^{1} w_1(\tau) p_{j+\kappa}^2(\tau) \, d\tau, \\
&\qquad \text{and using both (5.6) and (3.9),} \\
&= \sum_{k=n+1}^{\infty} \alpha_k^2 \int_{-1}^{1} w_1(\tau) p_k^2(\tau) \, d\tau \\
&= \int_{-1}^{1} w_1(\tau)(\psi - \psi_n)^2 d\tau,
\end{aligned}$$

as required. □

From (5.7) we have that $\lim_{n \to \infty} \int_{-1}^{1} w_1(\tau)(\psi - \psi_n)^2 d\tau = 0$, the rate of this mean convergence being the same as that of $F_{n-\kappa}$ to F.

We shall conclude this section by considering A as an operator from the space $L^2_{w_1}$ into $L^2_{w_2}$, and obtaining a bound for $\|A\|$. The space $L^2_{w_1}$ is the space of functions square integrable on $(-1, 1)$ with respect to the weight function w_1 so that the norm $\|\cdot\|_{w_1}$ is defined by

$$\|g\|_{w_1} = \left(\int_{-1}^{1} w_1(\tau) g^2(\tau) \, d\tau \right)^{1/2}. \tag{5.8}$$

Similarly for the space $L^2_{w_2}$, where its norm $\|\cdot\|_{w_2}$ is defined by

$$\|g\|_{w_2} = \left(\int_{-1}^{1} w_2(\tau) g^2(\tau) \, d\tau \right)^{1/2}. \tag{5.9}$$

We now define the norm of A by

$$\|A\| = \sup_{g \neq 0} \|Ag\|_{w_2} / \|g\|_{w_1}. \tag{5.10}$$

THEOREM 5.2. *Suppose* $f = Ag$, *then*

$$\|f\|_{w_2}^2 = \|g\|_{w_1}^2 + 2 \int_{-1}^{1} w_1 g \, \text{p.p.}(BX^{-1}) \Im(w_1 g) \, d\tau \tag{5.11}$$

and

$$\|A\| \leq \left\{ 1 + \left(\frac{1}{\pi} \int_{-1}^{1} w_1(\tau) \left(\frac{1}{\pi} \int_{-1}^{1} w_1(\xi) R^2(\xi, \tau) \, d\xi \right) d\tau \right)^{1/2} \right\}^{1/2}, \tag{5.12}$$

where $R(\xi, \tau) = (\text{p.p.}(BX^{-1}; \xi) - \text{p.p}(BX^{-1}; \tau))/(\xi - \tau)$.

Proof.

$$\int_{-1}^{1} w_1 f^2 d\tau = \int_{-1}^{1} \frac{f}{cZr}\left(\frac{aZg}{r} + bc\Im(w_1 g)\right) d\tau$$

$$= \int_{-1}^{1}\left\{\frac{afg}{cr^2} - \frac{Zg}{cr}\Im\left(\frac{fb}{Zr}\right)\right\} d\tau \quad \text{by (2.13)}$$

$$= \int_{-1}^{1}\left\{\frac{a^2 w_1 g}{r^2} + \frac{abg}{r^2}\Im(w_1 g)\right\} d\tau - \int_{-1}^{1} w_1 g \Im\left\{\frac{abg}{r^2} + \frac{cb^2}{Zr}\Im(w_1 g)\right\} d\tau$$

since $f = Ag$. On using (2.13) again, we find

$$\|f\|_{w_2}^2 = \int_{-1}^{1}\left\{\frac{a^2 w_1 g^2}{r^2} + \frac{2abg}{r^2}\Im(w_1 g)\right\} d\tau - \int_{-1}^{1} w_1 g \Im\left(\frac{cb^2}{Zr}\Im(w_1 g)\right) d\tau. \tag{5.13}$$

By the Poincaré–Bertrand formula (see (2.12)), we have

$$\Im\left(\frac{cb^2}{Zr}\Im(w_1 g)\right) = \Im\left(\frac{cb^2}{Zr}\right)\Im(w_1 g) - \frac{b^2 g}{r^2} - \Im\left(w_1 g \Im\left(\frac{cb^2}{Zr}\right)\right). \tag{5.14}$$

But, from (2.11), since $cb = B$ and B is a polynomial,

$$\Im\left(\frac{cb^2}{Zr}\right) = \Im\left(\frac{bB}{Zr}\right) = \frac{aB}{rZ} - \text{p.p.}(BX^{-1}).$$

Substituting this into (5.14) gives

$$\Im\left(\frac{cb^2}{Zr}\Im(w_1 g)\right) = \left\{\frac{aB}{rZ} - \text{p.p.}(BX^{-1})\right\}\Im(w_1 g) - \frac{b^2 g}{r^2} - \Im\left\{\frac{abg}{r^2} - w_1 g\,\text{p.p.}(BX^{-1})\right\}.$$

On substituting this expression into (5.13) and using (2.13) again, we obtain (5.11).

Now $\text{p.p.}(BX^{-1})$ is a polynomial of degree $\leq \kappa + \mu$ and is independent of g. By (2.13),

$$\int_{-1}^{1} w_1 g\,\text{p.p}(BX^{-1})\Im(w_1 g)\, d\tau = -\int_{-1}^{1} w_1 g \Im\left(w_1 g\,\text{p.p.}(BX^{-1})\right) d\tau,$$

but

$$\Im\left(w_1 g\,\text{p.p.}(BX^{-1})\right) - \text{p.p.}(BX^{-1})\cdot\Im(w_1 g) = \frac{1}{\pi}\int_{-1}^{1} w_1 g R\, d\tau,$$

from definition of R. Consequently,

$$2\int_{-1}^{1} w_1 g\,\text{p.p.}(BX^{-1})\Im(w_1 g)\, d\tau = -\frac{1}{\pi}\int_{-1}^{1} w_1(\tau) g(\tau)\left(\int_{-1}^{1} w_1(\xi) g(\xi) R(\xi,\tau)\, d\xi\right) d\tau. \tag{5.15}$$

On repeated use of the Cauchy–Schwarz inequality, we find

$$\left|2\int_{-1}^{1} w_1 g\,\text{p.p.}(BX^{-1})\Im(w_1 g)\, d\tau\right| \leq \|g\|_{w_2}^2\left\{\frac{1}{\pi}\int_{-1}^{1} w_1(\tau)\left(\frac{1}{\pi}\int_{-1}^{1} w_1(\xi) R^2(\xi,\tau)\, d\xi\right) d\tau\right\}^{1/2}$$

from which (5.12) follows. □

We see from (5.12) that A is a bounded linear operator from $L^2_{w_1}$ into $L^2_{w_2}$. In conclusion, we might note that on choosing $g = p_n$, with $n \geq \max(\mu, \kappa)$, $f = (-1)^{\kappa} q_{n-\kappa}$ (see (5.3)). The second term on the right-hand side of (5.11), which from (5.15) can be

rewritten as

$$-\frac{1}{\pi}\int_{-1}^{1} w_1(\tau)p_n(\tau)\left(\int_{-1}^{1} w_1(\xi)p_n(\xi)R(\xi,\tau)\,d\xi\right)d\tau,$$

vanishes if either $\mu+\kappa\leq 0$ or, by orthogonality, if $n\geq\mu+\kappa$. Thus, we recover (3.9) in the particular case when $j=k$. This particular case shows that we must also have $\|A\|\geq 1$.

6. Conclusion. In this paper we have identified two sequences of orthogonal polynomials which are induced by the singular integral equation (1.1). As indicated already, these sets of polynomials are of considerable importance in constructing approximate methods for the solution of both (1.1) and the so-called *complete* equation in which a term of the form $\int_{-1}^{1} k(t,\tau)\phi(\tau)\,d\tau$ is added to the left-hand side of (1.1). An approximate method of solution of the complete equation, based on collocation at the zeros of the polynomials q_n, has been described by the author in [2]. There it is noted that much algorithmic investigation remains to be done, and in particular, we might mention that of computing the zeros and Christoffel numbers of the polynomials p_n and q_n given only the weight functions w_1 and w_2. For some singular integral equations we can identify the polynomials p_n and q_n immediately.

When a,b are constants such that $b\neq 0$ and $a^2+b^2=1$, we find (see [1, Ex. 9.1]) that $\kappa=1$ and the fundamental function Z is given by $Z(t)=(1-t)^{-1-\theta}(1+t)^{\theta}$, where $\theta=(1/\pi)\arctan(b/a)+N$ with $N=0$ if $b/a<0$ or $N=-1$ if $b/a>0$. In this case we choose $c\equiv 1$ so that $w_1(t)=(1-t)^{-1-\theta}(1+t)^{\theta}$ and $w_2(t)=(1-t)^{1+\theta}(1+t)^{-\theta}$. The polynomials p_n, q_n can be chosen as the Jacobi polynomials $P_n^{(-1-\theta,\theta)}$ and $P_n^{(1+\theta,-\theta)}$ respectively.

For an equation with nonconstant coefficients suppose in (1.1) that

$$a(t)=T_{|\kappa|}(t), \qquad b(t)=\operatorname{sgn}(\kappa)(1-t^2)^{1/2}U_{|\kappa|-1}(t). \tag{6.1}$$

Here $T_{|\kappa|}$ and $U_{|\kappa|-1}$ denote respectively the Chebyshev polynomials of the first kind of degree $|\kappa|$ and of the second kind of degree $|\kappa|-1$. The index of this equation, as evaluated in [1], is κ. We exclude the case of $\kappa=0$ since this gives $b=0$; otherwise, κ can take any integer value. From (6.1) we see that $r(t)=1$, and since b is not a polynomial let us choose $c(t)=(1-t^2)^{-1/2}$. Then $B=cb$ is a polynomial of degree $\mu=|\kappa|-1$. With this a,b we find that $X(z)=(-1)^{\kappa}2^{\kappa}(z+(z^2-1)^{1/2})^{-\kappa}$ so that the fundamental function Z is given by $Z(t)=2^{\kappa}$, for all $t\in(-1,1)$. Consequently, $w_1=2^{\kappa}(1-t^2)^{1/2}$ and $w_2=2^{-\kappa}(1-t^2)^{1/2}$ so that the polynomials p_n, q_n can both be chosen in this case to be U_n. For another example with nonconstant coefficients see [2, §7].

REFERENCES

[1] M. L. Dow and David Elliott, *The numerical solution of singular integral equations over* (−1, 1), SIAM J. Numer. Anal., 16 (1979), pp. 115–134.

[2] David Elliott, *The classical collocation method for singular integral equations*, SIAM J. Numer. Anal., 19 (1982), pp. 816–832.

[3] F. D. Gakhov, *Boundary Value Problems*, Pergamon, Oxford, 1966.

[4] L. N. Karpenko, *Approximate solution of a singular integral equation by means of Jacobi polynomials*, J. Appl. Math. Mech., 30 (1966), pp. 668–675.

[5] Peter Linz, *An analysis of a method for solving singular integral equations*, BIT, 17 (1977), pp. 329–337.

[6] N. I. Muskhelishvili, *Singular Integral Equations*, Noordhoff, Gröningen, 1953.

[7] G. Szegö, *Orthogonal Polynomials*, AMS Colloquium Publications 23 (1939), American Mathematical Society, Providence, RI (3rd edition, 1967).

[8] F. G. Tricomi, *On the finite Hilbert transformation*, Quart. J. Math. (2), 2 (1951), pp. 199–211.

[9] _____, *Integral Equations*, Interscience, New York, 1957.

SIAM J. MATH. ANAL.
Vol. 13, No. 6, November 1982

0036-1410/82/1306-0013 $01.00/0

ON A FORMULA OF INVERSION*

D. NAYLOR† AND P. H. CHANG‡

Abstract. This paper develops a formula of inversion for an integral transform of a type similar to that associated with the names of Kontorovich and Lebedev except that the kernel involves the Neumann function $Y_u(kr)$ and the variable r varies over the truncated infinite interval $a \leq r < \infty$ where $a > 0$. The transform is useful in the investigation of functions that satisfy the Helmholtz equation and a condition of radiation at infinity.

1. Introduction. In previous papers [3], [4] one of the authors considered the integral transform defined by the equation

$$G(u)=\int_a^\infty \left[J_u(kr)H_u^{(1)}(ka)-J_u(ka)H_u^{(1)}(kr)\right] f(r)\,\frac{dr}{r},$$

where $a>0$, $k>0$. This transform is useful in the solution of certain boundary problems that are associated with the Helmholtz equation and which involve the radiation condition

$$\lim_{r\to\infty} r^{1/2}\left[f'(r)-ikf(r)\right]=0.$$

A formula of inversion for the above transform was devised in [4] where it was assumed that the function $f(r)$ satisfied the integrability condition

$$r^{-1/2}\left[rf''(r)+f'(r)+k^2rf(r)\right]\in L(a,\infty)$$

as well as the radiation condition.

In this paper the authors consider the transform defined by the equation

$$F(u)=\int_a^\infty Y_u(kr)f(r)\,\frac{dr}{r}, \tag{1}$$

where the kernel is a Neumann type Bessel function, the notation being that of Watson [9]. The transform (1), which was introduced in the earlier paper [5], has a simpler form than that of the transform $G(u)$, yet is applicable to the same class of boundary problems. A formula of inversion for transform (1) was constructed in [5]. This formula involved an integral term as well as a series expansion and was established when the function $f(r)$ to be expanded satisfied the radiation condition as well as the stated integrability condition. In this paper an alternative inversion formula for (1) is developed without assuming that $f(r)$ satisfies the above integrability condition. It is not assumed either that $f(r)$ satisfies the radiation condition; the essential assumption is only that $r^{-1}f(r)\in L(a,\infty)$. However, although not a necessary condition for the validity of the inversion formula itself, in practice the radiation condition must be enforced to permit the application of the transform (1) to the Helmholtz equation.

The method followed to obtain the alternative formula of inversion for the transform (1) is basically similar to that adopted in [1]. This entails an integration in the complex plane together with an appeal, in the final stage of the proof, to the Mellin inversion theorem.

*Received by the editors May 28, 1980, and in revised form October 16, 1981.
†Department of Applied Mathematics, University of Western Ontario, London, Ontario, Canada 96A5B9.
‡Department of Mathematics and Computer Science, University of Nebraska, Omaha, Nebraska 68182.

The inversion formula established in this paper contains an integral term as well as a series expansion. The series involves the functions $Y_{u_n}(kr)$ where $u_1, u_2, \cdots$ are the zeros of the function $Y_u(ka)$ regarded as function of the order u of the Bessel function. These functions will not be described as eigenfunctions since they do not form an orthogonal set on the interval (a, ∞).

The method followed in [5] was based on a Green's function technique which yielded an expansion involving the functions $Y_{u_n}(kr)$ associated with those zeros u_n positioned in the half plane $\mathrm{Re}(u)<0$. All of these zeros are real. The series that appears in the inversion formula proved in the present paper involves the functions $Y_{u_n}(kr)$ corresponding to zeros u_n situated in the half plane $\mathrm{Re}(u)>0$, of which all but a finite number are complex.

The zeros of the function $Y_u(x)$ regarded as a function of u for given positive x were discussed in [5] where it was proved that there exist three infinite sets of zeros as follows:

(i) an infinite set of real zeros u_n such that $u_n \to -\infty$ as $n \to \infty$. For large n the zero u_n is given by the asymptotic formula

$$u_n \approx -\left(n+\frac{1}{2}\right)+[ex/(2n+1)]^{2n+1}; \tag{2}$$

(ii) two infinite sets of complex zeros $u_n' = R_n e^{\pm i\theta_n}$ located in the first and fourth quadrants of the complex u-plane and given for large n by the asymptotic formulas

$$\theta_n \approx \frac{\pi}{2}\left[1-\frac{1}{\log(2R_n/ex)}\right], \tag{3}$$

$$R_n \log(2R_n/ex) \approx \left(n-\frac{1}{4}\right)\pi. \tag{4}$$

It is also known that the zeros u_n are all simple and that there are no purely imaginary zeros.

The formula proved in this paper is stated in the following theorem which is proved in §3.

THEOREM 1. *Suppose that $f(r)$ is continuous for $r \geq a > 0$ and that $r^{-1}f(r) \in L(a, \infty)$. Let the transform $F(u)$ be defined for positive values of k by means of* (1). *Then, if $r > a$,*

$$f(r) = \frac{1}{2i}\int_L \frac{[Y_u(kr)J_u(ka) - Y_u(ka)J_u(kr)]F(u)u\,du}{Y_u(ka)} + \lim_{c\to 0}\sum_{u_n}\frac{uY_u(kr)J_u(ka)F(u)e^{cu^2}}{(\partial/\partial u)Y_u(ka)}, \tag{5}$$

where L denotes the imaginary axis in the complex u-plane and the summation is extended over all the zeros u_n of the function $Y_u(ka)$, regarded as a function of the order u, that are located in the half plane $\mathrm{Re}(u)>0$.

The exponential function occurring in the series in the formula (5) is a summability factor, the parameter c tending to zero through positive values.

In order to establish the above theorem it is first necessary to prove a preliminary result which appears as Theorem 2 in §2 of the paper.

2. The integral theorem.

THEOREM 2. *Let the functions $f(r)$, $F(u)$ be defined as in Theorem* 1. *Then if $r>a$,*

$$f(r)=-\frac{1}{2i}\lim_{c\to 0}\int_L uJ_u(kr)F(u)e^{cu^2}du, \tag{6}$$

where L denotes the imaginary axis of the complex u-plane and the parameter c tends to zero through positive values.

To obtain the above result the expression (1) for $F(u)$ is inserted into the integral appearing on the right-hand side of (6) and the order of integration changed in the resulting repeated integral. This leads to the equation

$$\int_L uJ_u(kr)F(u)e^{cu^2}du=\int_a^\infty f(t)\frac{dt}{t}\int_L uJ_u(kr)Y_u(kt)e^{cu^2}du. \tag{7}$$

The inversion of the order of integration can be justified with the aid of the following bounds, derived in [4], [5], which show that the repeated integral is absolutely convergent for any $c>0$.

$$\begin{aligned}|J_{is}(kr)|&\le\left(\frac{2}{\pi kr}\right)^{1/2}\cosh(s\pi/2),\\ |Y_{is}(kt)|&\le\left(\frac{2}{\pi kt}\right)^{1/2}\cosh(s\pi/2).\end{aligned} \tag{8}$$

These inequalities apply on the imaginary axis where $u=is$ and s is real.

It will now be shown that the L-integral appearing in (7) is a symmetric function of the variables r,t in the sense that

$$\int_L uJ_u(kr)Y_u(kt)e^{cu^2}du=\int_L uJ_u(kt)Y_u(kr)e^{cu^2}du. \tag{9}$$

To obtain this result we appeal to the identity

$$Y_u(x)=J_u(x)\cot u\pi-J_{-u}(x)\operatorname{cosec}u\pi \tag{10}$$

from which it is seen that

$$\begin{aligned}&\int_L u\left[J_u(kr)Y_u(kt)-J_u(kt)Y_u(kr)\right]e^{cu^2}du\\ &\quad=-\int_L u\left[J_u(kr)J_{-u}(kt)-J_u(kt)J_{-u}(kr)\right]e^{cu^2}\operatorname{cosec}u\pi\,du=0,\end{aligned} \tag{11}$$

since the integrand is an odd function of the variable u. Therefore (9) follows from (11).

The domain of the t-integration in (7) is now decomposed into the parts (a,r) and (r,∞). This leads to the equation

$$\begin{aligned}\int_L uJ_u(kr)F(u)e^{cu^2}du&=\int_a^r f(t)\frac{dt}{t}\int_L uJ_u(kt)Y_u(kr)e^{cu^2}du\\ &\quad+\int_r^\infty f(t)\frac{dt}{t}\int_L uJ_u(kr)Y_u(kt)e^{cu^2}du,\end{aligned} \tag{12}$$

the variables (r,t) having been interchanged in the first integral on the right-hand side by virtue of (9).

It is convenient to introduce the function $h(u,r,t)$ defined by the equation

$$h(u,r,t)=uJ_u(kr)Y_u(kt)+\frac{1}{\pi}(r/t)^u. \tag{13}$$

The function h is, like the Bessel functions themselves, an entire function of the complex variable u. If the Bessel functions present on the right-hand side of (12) are eliminated with the aid of (13) we find the equation

$$\begin{aligned}\int_L uJ_u(kr)F(u)e^{cu^2}du &= -\frac{1}{\pi}\int_a^r f(t)\frac{dt}{t}\int_L (t/r)^u e^{cu^2}du \\ &\quad -\frac{1}{\pi}\int_r^\infty f(t)\frac{dt}{t}\int_L (r/t)^u e^{cu^2}du + I_1 + I_2,\end{aligned} \tag{14}$$

where

$$I_1 = \int_a^r f(t)\frac{dt}{t}\int_L h(u,t,r)e^{cu^2}du, \tag{15}$$

$$I_2 = \int_r^\infty f(t)\frac{dt}{t}\int_L h(u,r,t)e^{cu^2}du. \tag{16}$$

The first integral on the right-hand side of (14) can be combined with the second integral appearing therein after changing the sign of the variable u in the latter. This procedure yields the equation

$$\begin{aligned}\int_L uJ_u(kr)F(u)e^{cu^2}du &= -\frac{1}{\pi}\int_a^\infty f(t)\frac{dt}{t}\int_L \left(\frac{t}{r}\right)^u e^{cu^2}du + I_1 + I_2 \\ &= -\frac{1}{\pi}\int_L r^{-u}e^{cu^2}du\int_a^\infty t^{u-1}f(t)\,dt + I_1 + I_2,\end{aligned} \tag{17}$$

the inversion of the order of integration in the repeated integral being again justified for any $c>0$ by absolute convergence. The L-integral on the right-hand side of (17) is [8, p. 26] uniformly convergent for $c\geq 0$ and it reduces when $c=0$ to $i\pi f(r)$ by the Mellin inversion theorem [7, p.46]. It will next be proved that I_1 and I_2 tend to zero as $c\to 0$ so that the desired result (6) will follow from (17) on taking the limit of each side as $c\to 0$.

To investigate the integrals I_1 and I_2 it is necessary to determine the behaviour of the function h when u is large and to this end we make use of the formula

$$J_u(x) = \frac{(x/2)^u}{\Gamma(u+1)}\left[1-\frac{x^2}{4(1+u)}+O(u^{-2})\right]. \tag{18}$$

Equation (18) applies whenever u is large compared with x and bounded away from the negative integers. Upon inserting (18) and (10) and making use of the identity $\Gamma(u)\Gamma(1-u)=\pi\operatorname{cosec}u\pi$, it is found that the function h defined by (13) possesses the asymptotic form defined by the equation

$$h(u,r,t) = \left[\frac{k^2(r^2-t^2)}{4u\pi}(r/t)^u + (k^2rt/4)^u\frac{\cot u\pi}{u\Gamma(u)\Gamma(u)}\right]\left[1+O(u^{-1})\right] \tag{19}$$

for large values of u bounded away from the integers. The Γ-functions appearing in this formula may be estimated for large values of u by means of Stirling's formula

$$\Gamma(u) = (2\pi/u)^{1/2}\exp(u\log u - u)\left[1+O(u^{-1})\right], \tag{20}$$

which holds as $u\to\infty$ in $|\arg u|\leq\pi-\varepsilon$. It is then seen that

$$h(u,r,t)=\frac{1}{4u\pi}k^2(r^2-t^2)(r/t)^u\left[1+O(u^{-1})\right]$$
$$+O\left[\left(\frac{k^2rt}{4}\right)^u\cot u\pi\exp\left(-2u\log u+2u\right)\right] \tag{21}$$

for large u bounded away from integer values, $|\arg u|\leq\pi-\varepsilon$ and uniformly in any bounded domain of values of t.

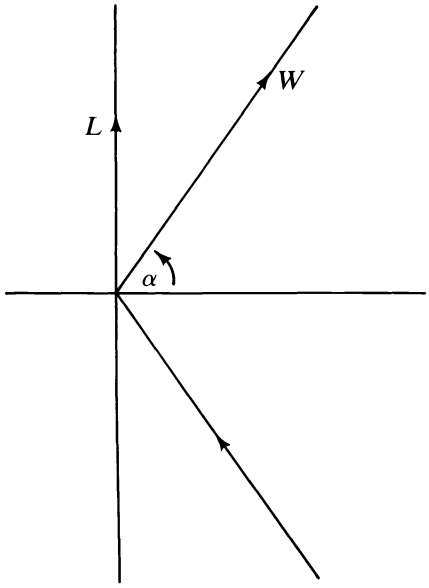

FIG. 1

Let W (Fig. 1) denote the wedge $|\arg u|=\alpha$ where $\frac{\pi}{4}<\alpha<\frac{\pi}{2}$. An inspection of the asymptotic formula (21) reveals that the dominant term in the integrand of the L integral in equation (15) is the summability factor $\exp(cu^2)$. If we set $u=Re^{i\theta}$ this factor has magnitude $\exp(cR^2\cos 2\theta)$ which provides convergence as $R\to\infty$ in the sector $\alpha\leq\theta\leq\frac{\pi}{2}$ since $\cos 2\theta<0$ there. The path L may then be deformed onto W so that

$$\int_L h(u,t,r)e^{cu^2}du=\int_W h(u,t,r)e^{cu^2}du. \tag{22}$$

Now $|\exp(cu^2)|<1$ on W, therefore, by (21),

$$|h(u,t,r)e^{cu^2}|\leq|h(u,t,r)|=O\left[u^{-1}\left(\frac{t}{r}\right)^u\right],$$

which is absolutely integrable on W if $a\leq t\leq r-\varepsilon$ where $\varepsilon>0$. Therefore, if $t\leq r-\varepsilon$,

$$\lim_{c\to 0}\int_L h(u,t,r)e^{cu^2}du=\int_W h(u,t,r)\,du=0$$

after closing the contour W on the right-hand side by means of a sequence of arcs of radii $|u|=n+\frac{1}{2}$, where $n\to\infty$ and applying Cauchy's theorem. Since the convergence as

$c \to 0$ is uniform for $a \le t \le r-\varepsilon$ it follows that

$$\lim_{c\to 0} \int_a^{r-\varepsilon} f(t)\,\frac{dt}{t} \int_L h(u,t,r) e^{cu^2}\,du = 0. \tag{23}$$

It will now be shown that, uniformly for $c>0$,

$$\int_{r-\varepsilon}^{r} f(t)\,\frac{dt}{t} \int_L h(u,t,r) e^{cu^2}\,du = O(\varepsilon). \tag{24}$$

To verify this result the path L is again deformed onto W which is then decomposed into the parts W_1 on which $|u|<R_1$ together with the remainder W_2 on which $|u| \ge R_1$. The quantity R_1 is chosen large enough to ensure that the asymptotic formula (21) applies on W_2. Let $h_1(u,r,t)$ denote the dominant (leading) term on the right-hand side of (21); then

$$\begin{aligned}
\left| \int_{W_2} h_1(u,t,r) e^{cu^2}\,du \right| &\le \frac{1}{2\pi} k^2 (r^2-t^2) \int_{R_1}^{\infty} R^{-1} \left(\frac{t}{r}\right)^{R\cos\alpha} e^{cR^2\cos\alpha}\,dR \\
&\le \frac{1}{2\pi} k^2 (r^2-t^2) R_1^{-1} \int_{R_1}^{\infty} \left(\frac{t}{r}\right)^{R\cos\alpha} e^{cRR_1\cos 2\alpha}\,dR \\
&\le \frac{k^2(r^2-t^2)}{2\pi R_1 [\cos\alpha \log(r/t) - cR_1 \cos 2\alpha]} \\
&\le \frac{k^2(r^2-t^2)}{2\pi R_1 \cos\alpha \log(r/t)} \le \frac{2r^2}{2\pi R_1 \cos\alpha},
\end{aligned}$$

since $\log(r/t) \ge (r-t)/r$ for $a \le t \le r$. Therefore the integral of h_1 along W_2 is bounded uniformly for $c>0$ and $a \le t \le r$. Similarly, the contributions from the remaining terms in (21) as well as that corresponding to the finite part W_1 are bounded. Therefore (24) follows and upon combining (23), (24) it is seen that $I_1 \to 0$ as $c \to 0$.

The investigation of I_2 is carried out in a similar manner by deforming L onto W and decomposing the domain of the t-integration in (16) into three parts $(r, r+\varepsilon)$, $(r+\varepsilon, r_0)$ and (r_0, ∞), where r_0 is chosen large enough to ensure that

$$\int_{r_0}^{\infty} |f(t)|\,\frac{dt}{t} \int_W |h(u,r,t)\,du| < \varepsilon.$$

To justify this inequality it is by (13) sufficient to verify that the double integrals

$$\int_W u J_u(kr)\,du \int_{r_0}^{\infty} Y_u(kt) f(t)\,\frac{dt}{t}, \qquad \int_{r_0}^{\infty} f(t)\,\frac{dt}{t} \int_W (r/t)^u\,du \tag{25}$$

are absolutely convergent. The second double integral in (25) is evidently absolutely convergent since $t \ge r_0 > r+\varepsilon$ therein and $t^{-1} f(t) \in L(a,\infty)$. To discuss the first double integral in (25) we note that it equals

$$\int_W u J_u(kr) F_0(u)\,du, \tag{26}$$

where

$$F_0(u) = \int_{r_0}^{\infty} Y_u(kt) f(t)\,\frac{dt}{t}. \tag{27}$$

The function $F_0(u)$ can be estimated for large values of u by means of equation (35), developed in the following section of this paper, with a replaced by r_0 therein. This shows that

$$F_0(u)=O\left[u^{-1/2}(2/kr_0)^u\Gamma(u)\right]. \tag{28}$$

Since from (18),

$$J_u(kr)=O\left[(kr/2)^u/\Gamma(u+1)\right], \tag{29}$$

it follows from (28), (29) that the integrand in (26) is $O[u^{-1/2}(r/r_0)^u]$ and this is absolutely integrable on W since $r_0>r+\varepsilon$.

The contributions to I_2 of the double integrals over the remaining intervals $(r,r+\varepsilon)$ and $(r+\varepsilon,r_0)$ are treated in a fashion similar to that used in discussing I_1 which shows that the first such interval contributes a term that is $O(\varepsilon)$ uniformly for all $c>0$ while the contribution from the second interval tends to zero as $c\to 0$. Thus $I_2\to 0$ as $c\to 0$.

3. The complex residue series. The formula (6) just established may now be used to construct the expansion formula stated in Theorem 1. The first step is to prove that

$$\int_{W_0}\frac{uY_u(kr)J_u(ka)F(u)\,du}{Y_u(ka)}=0, \tag{30}$$

where W_0 is the contour, illustrated in Fig. 2, whose asymptotes are the lines $\arg u=\pm\alpha$ where $\frac{\pi}{4}<\alpha<\frac{\pi}{2}$ and which is positioned to lie to the right of all of the zeros u_n of the function $Y_u(ka)$. This choice of W_0 is possible, since by (3) the complex zeros are such that $\arg u_n\to\pm\frac{\pi}{2}$ as $|u_n|\to\infty$ while there is at most a finite number of positive zeros.

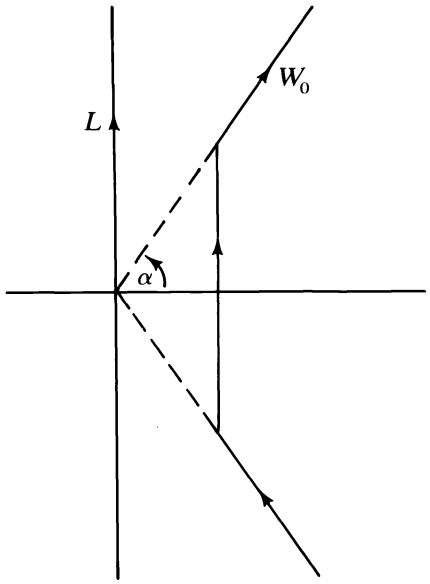

FIG. 2

It is first noted that the function $F(u)$ is an entire function of the complex variable u. This follows from the fact that the function $Y_u(kr)$ is itself an entire function of u and is $O(r^{-1/2})$ as $r\to\infty$ uniformly in any bounded domain of values of u. Since $r^{-1}f(r)\in L(a,\infty)$ the integral (1) is absolutely and uniformly convergent in any such

domain so that $F(u)$ is entire. The contour W_0 appearing in (30) will now be closed on the right-hand side and Cauchy's theorem applied. To carry out this procedure it is necessary to determine the asymptotic behaviour of the integrand. Upon substituting (18) into (10) it is found that

$$Y_u(kr) = -\frac{1}{\pi}\Gamma(u)\left(\frac{2}{kr}\right)^u\left[1 - \frac{\pi(kr/2)^{2u}\cot u\pi}{\Gamma(u)\Gamma(u+1)}\right]\left[1+O(u^{-1})\right] \tag{31}$$

for large u bounded away from integer values. As $u\to\infty$ on W_0, $\cot u\pi \to \pm i$ so, by Stirling's formula (20), equation (31) simplifies to the equation

$$Y_u(kr) = -\frac{1}{\pi}\Gamma(u)\left(\frac{2}{kr}\right)^u\left[1+O(u^{-1})\right] \tag{32}$$

as $u\to\infty$ on W_0. The contour W_0 will be closed by adding the part of the line γ, $\mathrm{Re}(u) = n+\frac{1}{2}$ cut off inside W_0. Here n is a positive integer which is allowed to tend to infinity. The line γ passes midway between the poles of $\cot u\pi$ and on this line we set $u = n+\frac{1}{2}+is$ where s is real and find that $|\cot u\pi| = |\tanh s\pi| \le 1$. Thus (32) is valid on γ as well as on W_0 itself so that

$$\frac{Y_u(kr)}{Y_u(ka)} = \left(\frac{a}{r}\right)^u\left[1+O(u^{-1})\right] \tag{33}$$

as $u\to\infty$ on γ and W_0.

An asymptotic bound on the function $F(u)$ appearing in (30) can be obtained from (1) by applying the Schwarz inequality which shows that $|F(u)| \le \|f(r)\|\cdot\|Y_u(kr)\|$, where

$$\|f(r)\| = \left\{\int_a^\infty |f(r)|^2 r^{-1}\,dr\right\}^{1/2}.$$

The factor $\|Y_u(kr)\|$ can be obtained from the formula [5, eq. (10)]

$$\int_a^\infty r^{-1}|Y_u(kr)|^2\,dr = \frac{\sinh(\pi R\sin\theta) + ka\,\mathrm{Im}\,Y_u(ka)Y_{\bar u}'(ka)}{\pi R^2\sin 2\theta}, \tag{34}$$

where $u = Re^{i\theta}$, $\bar u = Re^{-i\theta}$. Upon using (32) to estimate the Bessel functions occurring in (34) it is found that

$$\int_a^\infty r^{-1}|Y_u(kr)|^2\,dr = \frac{\sinh(\pi R\sin\theta)}{\pi R^2\sin 2\theta} + \frac{(2/ka)^{2R\cos\theta}|\Gamma(u)|^2}{2\pi^2 R\cos\theta}\left[1+O(u^{-1})\right].$$

For $|\arg u| \le \alpha$ the dominant term on the right-hand side of the preceding equation is that involving the Γ-function, therefore

$$F(u) = O\left[R^{-1/2}\left(\frac{2}{ka}\right)^{R\cos\theta}\Gamma(u)\right] \tag{35}$$

as $u\to\infty$ on γ and W_0. Upon combining (18), (33) and (35) it is found that the integrand appearing in (30) is

$$O\left[R^{-1/2}\left(\frac{a}{r}\right)^{R\cos\theta}\right].$$

This tends to zero sufficiently rapidly as $R\to\infty$ to permit the contour to be closed in the manner described. Since the integrand has no poles inside W_0 it follows that the integral in (30) is equal to zero, by Cauchy's theorem.

The next step is to deform the path W_0 in (30) onto the imaginary axis. The justification of this procedure requires the insertion of the summability factor $\exp(cu^2)$ which enables (30) to be written as the limit

$$\lim_{c\to 0}\int_{W_0}\frac{uY_u(kr)J_u(ka)F(u)e^{cu^2}du}{Y_u(ka)}=0. \tag{36}$$

The introduction of the summability factor into (30) is permissible since the integral in question is absolutely convergent and $|\exp(cu^2)|=\exp(cR^2\cos 2\alpha)\leq 1$ on W_0 for all $c\geq 0$ since $\frac{\pi}{4}<\alpha<\frac{\pi}{2}$.

Upon deforming W_0 onto L and taking into account the residues at the poles it is found that

$$\lim_{c\to 0}\int_{L}\frac{uY_u(kr)J_u(ka)F(u)e^{cu^2}du}{2i\pi Y_u(ka)}=-\lim_{c\to 0}\sum\frac{uY_u(kr)J_u(ka)F(u)e^{cu^2}}{(\partial/\partial u)Y_u(ka)}, \tag{37}$$

the summation extending over all of the zeros of the function $Y_u(ka)$ that are located in the half plane $\mathrm{Re}(u)\geq 0$.

To establish the truth of (37) estimates of the various functions appearing in (36) are needed in the region traversed. The path W_0 in (36) will be connected to the imaginary axis L by a sequence of paths C_n which will be defined shortly and which avoid the zeros of the function $Y_u(ka)$. On the paths C_n, $|\mathrm{Im}\,u|\to\infty$ and $\cot u\pi$ tends to $-i$ in the upper half plane or to $+i$ in the lower half plane. Introducing this change into (31) and estimating the Γ-functions by means of Stirling's formula (20) yields the equation

$$Y_u(ka)=-2(\pi u)^{-1/2}e^{-i\pi/4}\sinh\left[u\log\left(\frac{2u}{kae}\right)+\log\sqrt{2}+\frac{i\pi}{4}\right]\left[1+O(u^{-1})\right], \tag{38}$$

which applies as $u\to\infty$ in $0<\delta\leq\arg u\leq\frac{\pi}{2}$. Upon setting $u=Re^{i\theta}$ equation (38) can be written as the equation

$$Y_u(ka)=-2(\pi u)^{-1/2}e^{-i\pi/4}\sinh(A+iB)\left[1+O(R^{-1})\right], \tag{39}$$

where

$$\begin{aligned}A&=R\cos\theta\log\left(\frac{2R}{kae}\right)-R\theta\sin\theta+\log\sqrt{2},\\ B&=R\sin\theta\log\left(\frac{2R}{kae}\right)+R\theta\cos\theta+\frac{\pi}{4}.\end{aligned} \tag{40}$$

The complex zeros $u=Re^{\pm i\theta}$ lie in the half plane $\mathrm{Re}(u)\geq 0$ and those of large modulus are given by the equations $A=0$, $B=n\pi$, where n is a large (positive) integer. The solution of these equations gives the values $R=R_n$, $\theta=\pm\theta_n$ quoted in §1, (3) and (4) which show that $R_n\to\infty$ and $\theta_n\to\frac{\pi}{2}$ as $n\to\infty$. Therefore if $\delta>0$ the complex zeros of sufficiently large modulus all lie in the sectors $\frac{\pi}{2}-\delta\leq|\theta|\leq\frac{\pi}{2}$.

As the curve C_n we may select that whose polar equation in the sectors $\frac{\pi}{2}-\delta<|\theta|\leq\frac{\pi}{2}$ is $B=(n+\frac{1}{2})\pi$ and which is continued beyond these sectors by means of two circular arcs $|u|=R$ of suitable radius to meet the wedge W_0. On the parts of C_n lying

inside the stated sectors, $\sinh(A+iB)$ reduces to $\pm i\cosh A$ so that

$$\begin{aligned}\left|\frac{Y_u(kr)}{Y_u(ka)}\right| &\sim \left|\frac{\sinh\left[A+iB+u\log(a/r)\right]}{\sinh(A+iB)}\right| \\ &= \left|\frac{\cosh\left[A+u\log(a/r)\right]}{\cosh A}\right| \\ &= \left|\cosh\left[u\log\left(\frac{a}{r}\right)\right]+\tanh A\sinh\left[u\log\left(\frac{a}{r}\right)\right]\right| \\ &\le \cosh\left[R\cos\theta\log\left(\frac{a}{r}\right)\right]+\sinh\left|R\cos\theta\log\left(\frac{a}{r}\right)\right| \\ &= \exp\left[R\cos\theta\log\left(\frac{r}{a}\right)\right]\end{aligned} \tag{41}$$

as $R\to\infty$ in $\frac{\pi}{2}-\delta\le|\theta|\le\frac{\pi}{2}$. The behaviour of the function $F(u)$ in this sector can as before be found by means of the Schwarz inequality in conjunction with equation (34) in which the function $Y_u(ka)$ and its derivatives are calculated from (39). This gives the equation

$$\begin{aligned}F(u) = {} & O\left[\left|R^2\sin 2\theta\right|^{-1/2}\exp\left(\frac{\pi}{2}R\sin\theta\right)\right] \\ & + O\left[\left|R^2\sin 2\theta\right|^{-1/2}\exp\left|R\cos\theta\log\left(\frac{2R}{kae}\right)-R\theta\sin\theta+\log\sqrt{2}\right|\right]\end{aligned} \tag{42}$$

as $R\to\infty$. The preceding bound breaks down on the imaginary axis $\theta=\pm\frac{\pi}{2}$ where, however, it is permissible to use the following alternative bound:

$$F(u)=O\left[\exp\left|\frac{\pi}{2}R\sin\theta\right|\right]. \tag{43}$$

This bound, which is derived in Appendix 1 to this paper, applies as $|R\sin\theta|\to\infty$ in the strip $|R\cos|\le\frac{1}{3}$.

In the remaining sectors $\alpha\le|\theta|\le\frac{\pi}{2}-\delta$, $|\cot u\pi|\to 1$ while $\Gamma(u)\to\infty$ as $u\to\infty$ so that the leading term on the right-hand side of (31) is the dominant one. Therefore (32), (33), (35) also apply in this sector.

The behaviour of the Bessel function $J_u(ka)$ occurring in (36) can be obtained immediately by combining (18) and (20), which leads to the asymptotic equation

$$|J_u(ka)|\sim R^{-1/2}\exp\left[-R\cos\theta\log\left(\frac{2R}{kae}\right)+R\theta\sin\theta\right]. \tag{44}$$

An inspection of the equations (33), (35), (41), (42), (43), (44) reveals that on the curve C_n the expression

$$\frac{uJ_u(ka)Y_u(kr)F(u)}{Y_u(ka)}$$

tends to zero as $n\to\infty$ except in the vicinity of the imaginary axis where it is $O[R^{-1/2}\exp(R\log R)]$ at most. However the presence of the summability factor, which has magnitude $\exp(cR^2\cos 2\theta)$, where $\cos 2\theta<0$, ensures that when $c>0$ the actual integrand in (36) tends to zero sufficiently rapidly as $n\to\infty$ to permit the path W_0 to be deformed onto L. This verifies the validity of equation (37), which when combined with

(6) yields the following formula:

$$f(r)=-\frac{1}{2i}\lim_{c\to 0}\int_L \frac{u[J_u(kr)Y_u(ka)-J_u(ka)Y_u(kr)]F(u)e^{cu^2}du}{Y_u(ka)} + \lim_{c\to 0}\sum \frac{uY_u(kr)J_u(ka)F(u)e^{cu^2}}{(\partial/\partial u)Y_u(ka)}. \tag{45}$$

The expansion formula stated in Theorem 1 follows from (45) on setting $c=0$ in the integrand, a step that will be justified by demonstrating that the integral in (45) is uniformly convergent for $c\geq 0$. To establish this result, appeal is made to the following theorem, similar to that proved in [8, p. 26] concerning the integrals

$$\int_{-\infty}^{\infty} k(s)\,ds, \qquad \int_{-\infty}^{\infty} e^{-cs^2}k(s)\,ds.$$

If the first integral above exists, the second is uniformly convergent for $c\geq 0$. The integral appearing in (45) is of the above type since $u=is$ on L where s is real. This integral will be uniformly convergent for $c\geq 0$ provided that the integral

$$\int_{-\infty}^{\infty} \frac{s[J_{is}(kr)Y_{is}(ka)-J_{is}(ka)Y_{is}(kr)]F(is)\,ds}{Y_{is}(ka)} \tag{46}$$

exists.

To discuss the convergence of (46) it is first noted that by (10) the cross-product of Bessel functions appearing in (46) is equal to

$$\begin{aligned} &J_{is}(kr)Y_{is}(ka)-J_{is}(ka)Y_{is}(kr) \\ &\quad =i[J_{is}(kr)J_{-is}(ka)-J_{is}(ka)J_{-is}(kr)]\operatorname{cosech}(s\pi) \\ &\quad =-\frac{2}{s\pi}\sin bs-\frac{k^2(r^2-a^2)}{2\pi s^2}\cos bs+O(s^{-3}) \end{aligned} \tag{47}$$

after estimating the Bessel functions by means of (18), using the identity $\Gamma(1+is)\Gamma(1-is)=s\pi\operatorname{cosech}(s\pi)$ and setting $b=\log(r/a)$.

The behaviour of the function $Y_{is}(kx)$ for large values of s will be obtained from the formula

$$Y_{is}(kx)=-(2\pi\rho)^{-1/2}\exp\left(\frac{1}{2}s\pi+i\varnothing+\frac{1}{4}i\pi\right)\left[1-\frac{i}{8\rho}+\frac{5is^2}{24\rho^3}+O(s^{-2})\right], \tag{48}$$

where

$$\varnothing=\rho-s\log\left[\frac{s+\rho}{kx}\right], \tag{49}$$

$$\rho=(s^2+k^2x^2)^{1/2} \tag{50}$$

and (48) applies as $s\to+\infty$, uniformly for $0\leq x<\infty$. The formula (48) follows upon substituting a similar formula for $J_{is}(kx)$ into the relation (10), with u set equal to is therein. The asymptotic formula for $J_{is}(kx)$ which is quoted in [2, p. 140] can be obtained from an integral representation of the Bessel function by applying the method of steepest descents, which requires the ratio s/x to remain fixed as $s, x\to\infty$. This is stated in [2]. (The authors are indebted to Professor Olver for pointing out that the expansion in question can also be obtained from the more general theory developed in

[6, pp. 374–382], where uniform asymptotic expansions for the modified Bessel functions are obtained from the differential equation satisfied by these functions, after applying the transformations of variables appropriate to the discussion of the case in which both s and x may be large.) The formula (48) can be obtained from the expression quoted [6, p. 382] with the aid of the identity

$$-\pi Y_{is}(x)=e^{\pi s/2}K_{is}(xe^{-i\pi/2})+e^{-\pi s/2}K_{is}(xe^{i\pi/2}).$$

In the present notation Olver's expansion for $K_{is}(sx)$ applies as $s\to+\infty$, uniformly for $0<x<\infty$.

If, however, attention is confined to positive real values of s and x and if only the leading terms of the expansion are required, these can be obtained by following the initial transformations outlined in [6] but without appealing to the general theory, which is designed to provide the full expansion as well as allow for general complex values of the variables. The simplified procedure suggested is carried out in Appendix 2 of this paper. An examination of the method followed there confirms that the expansion (48) is valid as $s\to\infty$, uniformly for $0\le x<\infty$.

If $s\to+\infty$ while x remains bounded the formula (48) simplifies to give the result

(51)
$$Y_{is}(kx)=-(2\pi s)^{-1/2}\exp\left[\frac{1}{2}\pi s-is\log\left(\frac{2s}{kxe}\right)+\frac{1}{4}\pi i\right]\left[1+\frac{i}{12s}+\frac{ik^2x^2}{4s}+O(s^{-2})\right].$$

This formula applies in particular for $x=a$. On dividing (48) by the expression for $Y_{is}(ka)$ obtained from (51) we find the formula

(52)
$$\frac{Y_{is}(kx)}{Y_{is}(ka)}=\left(\frac{s}{\rho}\right)^{1/2}\exp[i\Psi(s)][1+g(s)],$$

where

(53)
$$\Psi(s)=\rho-s\log\left(\frac{s+\rho}{kx}\right)+s\log\left(\frac{2s}{kae}\right),$$

(54)
$$g(s)=\frac{1}{8i\rho}+\frac{1}{12is}+\frac{k^2a^2}{4is}+\frac{5is^2}{24\rho^3}+O(s^{-2})$$

Equation (52) holds as $s\to+\infty$ uniformly for $a\le x<\infty$. If $0\le x\le as^{\alpha}$ where $0\le\alpha\le\frac{1}{4}$ the formula (52) simplifies, after expanding the various terms appearing there, to yield the equation

(55)
$$\frac{Y_{is}(kx)}{Y_{is}(ka)}=\left[1+\frac{ik^2(x^2-a^2)}{4s}+H(x,s)\right]e^{is\log(x/a)},$$

where $|H(x,s)|\le Cs^{4\alpha-2}$, C being a constant.

The equation (52) shows that $|Y_{is}(kx)/Y_{is}(ka)|\le C'$, some constant, whenever s is large enough, so that the substitution of the $O(s^{-3})$ term in (47) into (46) leads to a term of $O(s^{-2})$ which is absolutely integrable.

We consider now the first two terms appearing in the expression (47). Let s_1, s_2 be large and positive and $s_2>s_1$. On expressing the trigonometric functions appearing in (47) in terms of exponentials and substituting the resulting expression into (46) it is seen necessary to consider the integrals

(56)
$$I=\int_{s_1}^{s_2}\frac{e^{\pm ibs}\,ds}{Y_{is}(ka)}\int_a^{\infty}f(x)Y_{is}(kx)\frac{dx}{x},$$

$$J=\int_{s_1}^{s_2}\frac{e^{\pm ibs}\,ds}{sY_{is}(ka)}\int_a^\infty f(x)Y_{is}(kx)\,\frac{dx}{x}. \tag{57}$$

The integration with respect to x in (56) is now decomposed into the parts (a, as_1^α) and (as_1^α, ∞). The contribution of the first such interval to I is

$$I_1=\int_{s_1}^{s_2}\frac{e^{\pm ibs}\,ds}{Y_{is}(ka)}\int_a^{as_1^\alpha}f(x)Y_{is}(kx)\,\frac{dx}{x}. \tag{58}$$

On substituting the expression (55) we find that

$$\begin{aligned} I_1=&\int_{s_1}^{s_2}e^{\pm ibs}\,ds\int_a^{as_1^\alpha}e^{is\log(x/a)}f(x)\,\frac{dx}{x}\\ &+\frac{ik^2}{4}\int_{s_1}^{s_2}s^{-1}e^{\pm ibs}\,ds\int_a^{as_1^\alpha}(x^2-a^2)e^{is\log(x/a)}f(x)\,\frac{dx}{x}+I_1', \end{aligned} \tag{59}$$

where the term I_1' which arises from the $H(x,s)$ term in (55) does not exceed the quantity

$$C\int_{s_1}^{s_2}s^{4\alpha-2}\,ds\int_a^{as_1^\alpha}|f(x)|\,\frac{dx}{x}\le C\int_{s_1}^{s_2}s^{4\alpha-2}\,ds\int_a^\infty|f(x)|\,\frac{dx}{x}=O\big(s_1^{4\alpha-1}\big).$$

This tends to zero as $s_1\to\infty$ since $\alpha<\frac{1}{4}$.

Consider now the first term on the right-hand side of (59). To show that this term tends to zero as $s_1, s_2\to\infty$ we write it as the difference of two integrals, in the form

$$\int_{s_1}^{s_2}e^{\pm ibs}\,ds\int_a^\infty e^{is\log(x/a)}f(x)\,\frac{dx}{x}-\int_{s_1}^{s_2}e^{\pm ibs}\,ds\int_{as_1^\alpha}^\infty e^{is\log(x/a)}f(x)\,\frac{dx}{x}. \tag{60}$$

The first integral occurring here tends to zero as $s_1, s_2\to\infty$ by the theory of Fourier integrals, since if we set $y=\log(x/a)$ it becomes equal to the integral

$$\int_{s_1}^{s_2}e^{\pm ibs}\,ds\int_0^\infty e^{isy}f(ae^y)\,dy.$$

The second repeated integral in (60) is absolutely convergent, since $x^{-1}f(x)\subset L(a,\infty)$. On changing the order of integration and carrying out the resulting integration with respect to s we find that its modulus does not exceed the quantity

$$2\int_{as_1^\alpha}^\infty\frac{|f(x)|\,dx}{x[\log(x/a)\pm b]}\le\frac{2}{\alpha\log s_1-b}\int_{as_1^\alpha}^\infty|f(x)|\,\frac{dx}{x}.$$

This tends to zero as $s_1\to+\infty$. Therefore both terms in the expression (60) tend to zero as $s_1 s_2\to\infty$ so that the same is true of the first term on the right-hand side of (59).

To discuss the second term on the right-hand side of (59) it is written as the sum of two integrals as follows:

$$\begin{aligned} &\int_{s_1}^{s_2}s^{-1}e^{\pm ibs}\,ds\int_a^{2r}x^{-1}(x^2-a^2)f(x)e^{is\log(x/a)}\,dx\\ &\quad+\int_{s_1}^{s_2}s^{-1}e^{\pm ibs}\,ds\int_{2r}^{as_1^\alpha}x^{-1}(x^2-a^2)f(x)e^{is\log(x/a)}\,dx. \end{aligned} \tag{61}$$

If, as before, we set $y=\log(x/a)$ the first integral occurring in the preceding expression becomes equal to the integral

$$\int_{s_1}^{s_2}s^{-1}e^{\pm ibs}\,ds\int_0^{\log(2r/a)}(e^{2y}-1)f(ae^y)e^{isy}\,dy. \tag{62}$$

Now the theory of Fourier integrals applied to a function of compact support shows that the integral

$$\int_{s_1}^{s_2} e^{\pm ibs}\,ds \int_0^{\log(2r/a)} (e^{2y}-1) f(ae^y) e^{isy}\,dy$$

is convergent as $s_2 \to \infty$. The same is true, a fortiori, of the integral (62) which therefore tends to zero as $s_1, s_2 \to \infty$.

The second term in (61) can be shown to tend to zero as $s_1, s_2 \to \infty$ by changing the order of integration and applying the second mean value theorem to the resulting s-integrals. If we write $\lambda = \log(x/a) \pm b$ and recall the definition $b = \log(r/a)$ it is seen that the possible values of λ are positive throughout the relevant interval $2r \le x \le as_1^\alpha$. In fact $\lambda \ge \log(x/a) - \log(r/a) = \log(x/r) \ge \log 2$, therefore, if $s_1 \le s_2' \le s_2$,

$$\int_{s_1}^{s_2} s^{-1} \cos\lambda s\,ds = s_1^{-1} \int_{s_1}^{s_2'} \cos\lambda s\,ds = (\lambda s_1)^{-1} (\sin\lambda s_2' - \sin\lambda s_1)$$

and similarly for the corresponding integral involving $\sin\lambda s$. Since $\lambda \ge \log 2$ it follows that

$$\left| \int_{s_1}^{s_2} s^{-1} e^{i\lambda s}\,ds \right| \le \frac{4}{s_1 \log 2}.$$

Thus the modulus of the second repeated integral in (61) does not exceed the quantity

$$\frac{4}{s_1 \log 2} \int_{2r}^{as_1^\alpha} x^{-1}(x^2 - a^2) |f(x)|\,dx \le \frac{4a^2 s_1^{2\alpha-1}}{\log 2} \int_{2r}^{as_1^\alpha} |f(x)|\,\frac{dx}{x},$$

since $x^2 \le a^2 s_1^{2\alpha}$ throughout the interval of integration. Since $\alpha < \frac{1}{4}$ the last expression tends to zero as $s_1, s_2 \to \infty$. Thus both terms appearing in (61) tend to zero as $s_1, s_2 \to \infty$ so that the same is true of the expression (59) for I_1.

It is now necessary to consider the contribution of the interval (as_1^α, ∞) to I. This contribution is defined by the equation

$$I_2 = \int_{s_1}^{s_2} \frac{e^{\pm ibs}\,ds}{Y_{is}(ka)} \int_{as_1^\alpha}^{\infty} f(x) Y_{is}(kx)\,\frac{dx}{x}.$$

This integral is absolutely convergent, since, as has already been noted, $|Y_{is}(kx)/Y_{is}(ka)| \le C$ for all relevant x, s. Therefore, on changing the order of integration, we arrive at the equation

$$I_2 = \int_{as_1^\alpha}^{\infty} f(x)\,\frac{dx}{x} \int_{s_1}^{s_2} \frac{e^{\pm ibs} Y_{is}(kx)\,ds}{Y_{is}(ka)}. \tag{63}$$

To proceed further it is necessary to appeal to the following result, which is developed in Appendix 1 to this paper,

$$\int_{s_1}^{s_2} \frac{e^{\pm ibs} Y_{is}(kx)\,ds}{Y_{is}(ka)} = i\left(\frac{s_1}{\rho_1}\right)^{1/2} \frac{e^{i\psi(s_1)}}{\psi'(s_1)} - i\left(\frac{s_2}{\rho_2}\right)^{1/2} \frac{e^{i\psi(s_2)}}{\psi'(s_2)} + O(s_1^{-1}), \tag{64}$$

where $\psi(s) = \Psi(s) \pm bs$ and

$$\rho_1 = (s_1^2 + k^2x^2)^{1/2}, \qquad \rho_2 = (s_2^2 + k^2x^2)^{1/2}. \tag{65}$$

Equation (64) applies for s_1, s_2 arbitrarily large, $s_2 \ge s_1$ and uniformly for $x \ge 3r$. It is proved in Appendix 1 that $\psi'(s) > \log 2$ for all relevant values of s and x and it follows

from (65) that $\rho_1 \geq s_1$ and $\rho_2 \geq s_2$ therefore (64) shows that

$$\left| \int_{s_1}^{s_2} \frac{e^{\pm ibs} Y_{is}(kx)\, ds}{Y_{is}(ka)} \right| \leq \frac{2}{\log 2} + \frac{B}{s_1},$$

where B is a constant. If the preceding bound is inserted into (63) we find that

$$|I_2| \leq \left(\frac{2}{\log 2} + \frac{B}{s_1} \right) \int_{as_1^{\alpha}}^{\infty} |f(x)|\, \frac{dx}{x},$$

which tends to zero as $s_1 \to \infty$ since $x^{-1} f(x) \in L(a, \infty)$. On collecting these results it is seen that the integral $I = I_1 + I_2$ is convergent as $s_2 \to \infty$. The integral J defined by (57) will also converge, a fortiori, as $s_2 \to \infty$ so that the integral (46) itself is convergent.

Appendix 1. The formula (64) is derived by substituting the expression (52) into the integral. This gives the equation

$$\int_{s_1}^{s_2} \frac{e^{\pm ibs} Y_{is}(kx)\, ds}{Y_{is}(ka)} = \int_{s_1}^{s_2} \left(\frac{s}{\rho} \right)^{1/2} e^{i\psi}\, ds + \int_{s_1}^{s_2} \left(\frac{s}{\rho} \right)^{1/2} e^{i\psi} g(s)\, ds. \tag{A.1}$$

It will be shown that each of the five terms appearing on the right-hand side of (54) for $g(s)$ contributes a term of magnitude $O(s_1^{-1})$ to the last integral in (A.1). First it is shown that $\psi'(s) \geq \log 2$ for $s \geq 3kr$ and $x \geq 3r$. From (53) by successive differentiation we find the formulas

$$\psi'(s) = \pm b - \log a + \log \left[\frac{2sx}{s+\rho} \right],$$

$$\psi''(s) = \frac{\rho - s}{s\rho}, \qquad \psi'''(s) = \frac{s^3 - \rho^3}{s^2 \rho^3}.$$

Since $\rho > s$ then $\psi'' > 0$ and ψ' is an increasing function of s. Upon recalling that $b = \log(r/a)$ it is seen that ψ' cannot be less than

$$\log \left[\frac{2sx}{r(s+\rho)} \right] \geq \log 2$$

since $\rho \leq s + kx$, so that

$$\frac{r(s+\rho)}{2sx} \leq \frac{r(2s+kx)}{2sx} = \frac{r}{x} + \frac{kr}{2s} \leq \frac{1}{2}$$

as required.

The s^{-1} terms present in (54) lead upon substitution into (A.1) to an integral of the form

$$\int_{s_1}^{s_2} \frac{e^{i\psi}\, ds}{(s\rho)^{1/2}}. \tag{A.2}$$

We write $e^{i\psi} = \cos\psi + i \sin\psi$ and consider the integral

$$\begin{aligned} \int_{s_1}^{s_2} \frac{\cos\psi\, ds}{(s\rho)^{1/2}} = \int_{s_1}^{s_2} \frac{d(\sin\psi)}{(s\rho)^{1/2} \psi'(s)} &= \frac{1}{s_1^{1/2} \rho_1^{1/2} \psi'(s_1)} \int_{s_1}^{s_2'} d(\sin\psi) \\ &= \frac{\sin\psi(s_2') - \sin\psi(s_1)}{s_1^{1/2} \rho_1^{1/2} \psi'(s_1)}, \qquad s_1 < s_2' < s_2, \end{aligned} \tag{A.3}$$

after applying the mean value theorem, the functions ρ and $\psi'(s)$ being positive increasing functions of s. The quantity ρ_1 is defined by equation (65). Thus the expression (A.3) is $O(s_1^{-1})$ uniformly for all relevant x, that is $x \geq 3r$. The corresponding integral involving $\sin\psi$ can be treated in the same way so that it too is $O(s_1^{-1})$. Consider next the ρ^{-1} term in (54) which when inserted into (A.1) leads to the integral

$$\int_{s_1}^{s_2} \left(\frac{s}{\rho}\right)^{1/2} \frac{e^{i\psi}\, ds}{\rho}.$$

The function $s/\rho = (1+k^2x^2/s^2)^{-1/2}$ is an increasing function of s so that application of the mean value theorem twice shows that

$$\text{(A.4)} \qquad \int_{s_1}^{s_2} \left(\frac{s}{\rho}\right)^{1/2} \frac{\cos\psi\, ds}{\rho} = \left(\frac{s_2}{\rho_2}\right)^{1/2} \frac{[\sin\psi(s_2') - \sin\psi(s_1')]}{\rho_1'\psi'(s_1')}$$

where $s_1 < s_1' < s_2' < s_2$ and $\rho_1' = [(s_1')^2 + k^2x^2]^{1/2}$. Since $\rho_2 > s_2$ and $\rho_1 > s_1$ the expression (A.4) is $O(s_1^{-1})$, uniformly for $x \geq 3r$, and similarly for the corresponding integral involving $\sin\psi$. The s^2/ρ^3 term present in (54) can be treated in a similar fashion, with two applications of the mean value theorem, to show that its contribution to the final term in (A.1) is also $O(s_1^{-1})$. Finally the $O(s^{-2})$ term in (54) leads upon insertion into (A.1) to an absolutely convergent integral which is $O(s_1^{-1})$. Thus the last term on the right-hand side of (A.1) is $O(s_1^{-1})$ uniformly for $x > 3r$. If the remaining term is transformed by an integration by parts we find the equation

(A.5)

$$\int_{s_1}^{s_2} \frac{e^{\pm ibs} Y_{is}(kx)\, ds}{Y_{is}(ka)} = i\left(\frac{s_1}{\rho_1}\right)^{1/2} \frac{e^{i\psi(s_1)}}{\psi'(s_1)} - i\left(\frac{s_2}{\rho_2}\right)^{1/2} \frac{e^{i\psi(s_2)}}{\psi'(s_2)}$$
$$+ i\int_{s_1}^{s_2} \left[\frac{k^2x^2e^{i\psi}}{2s^{1/2}\rho^{5/2}\psi'(s)} - \left(\frac{s}{\rho}\right)^{1/2} \frac{\psi''(s)e^{i\psi}}{[\psi'(s)]^2}\right] ds + O(s_1^{-1}).$$

Now s/ρ is an increasing function of s whereas $\psi''(s)$ and $\psi'(s)^{-1}$ are decreasing functions of s. Therefore,

$$\text{(A.6)} \qquad \int_{s_1}^{s_2} \left(\frac{s}{\rho}\right)^{1/2} \frac{\psi''(s)\cos\psi\, ds}{[\psi'(s)]^2} = \left(\frac{s_2}{\rho_2}\right)^{1/2} \int_{s_1'}^{s_2} \frac{\psi''(s)\cos\psi\, ds}{[\psi'(s)]^2}$$
$$= \left(\frac{s_2}{\rho_2}\right)^{1/2} \frac{\psi''(s_1')}{[\psi'(s_1')]^3} [\sin\psi(s_2') - \sin\psi(s_1')].$$

Since $\rho_2 > s_2$ and $\psi''(s_1') \leq (s_1')^{-1} \leq s_1^{-1}$ the expression (A.6) is again $O(s_1^{-1})$, uniformly for $x \geq 3r$. The integral remaining on the right-hand side of (A.5) can be estimated by means of a single application of the mean value theorem to show that it is $O(s_1^{-1})$ uniformly for $x \geq 3r$. On collecting these results we obtain the formula (64) quoted earlier.

It remains to establish the bound (43) used in §3. This bound can be deduced from the following formula [2, p. 93]

$$J_u(x)J_v(x) + Y_u(x)Y_v(x)$$
$$= \frac{4}{\pi^2} \int_0^\infty K_{u+v}(2x\sinh\theta)[e^{(u-v)\theta}\cos v\pi + e^{-(u-v)\theta}\cos u\pi]\, d\theta,$$

which is valid for $x>0$ and $|\mathrm{Re}(u+v)|<1$. Upon setting $u=t+is$, $v=t-is$ where t,s are real it follows that

$$\begin{aligned}|J_u(x)|^2+|Y_u(x)|^2&=\frac{4}{\pi^2}\int_0^\infty K_{2t}(2x\sinh\theta)(e^{2is\theta}\cos v\pi+e^{-2is\theta}\cos u\pi)\,d\theta\\&\le\frac{8}{\pi^2}\cosh s\pi\int_0^\infty K_{2t}(2x\sinh\theta)\,d\theta\end{aligned}\tag{A.7}$$

since $|\cos u\pi|=|\cos v\pi|\le\cosh s\pi$. Equation (A.7) holds for $x>0$ and $|t|<\frac{1}{2}$. A bound on the Bessel function integral remaining in (A.7) can be obtained by noting that it cannot exceed

$$\int_0^\infty K_{2t}(2x\sinh\theta)\cosh\theta\,d\theta=\frac{1}{2x}\int_0^\infty K_{2t}(y)\,dy=\frac{\pi}{4x}\sec t\pi$$

by Watson [9, p. 388, eq. (8)]. Equation (A.7) now shows that

$$|Y_u(x)|^2\le(2/\pi x)\cosh s\pi\sec t\pi\le(4/\pi x)\cosh s\pi$$

for $|t|<\frac{1}{3}$. Therefore, $|Y_u(x)|\le Cx^{-1/2}\exp|\frac{1}{2}s\pi|$, where C is a constant, and on substituting this in the integral (1) it follows that $F(u)=O(\exp|\frac{1}{2}s\pi|)$ since by hypothesis $r^{-1}f(r)\in L(a,\infty)$ so that $r^{-3/2}f(r)\in L(a,\infty)$, necessarily.

Appendix 2. In this appendix the asymptotic formula for $Y_{is}(x)$ required in §3 is constructed. This function satisfies the differential equation

$$x^2j_{xx}+xj_x+(x^2+s^2)j=0.$$

We now apply the standard transformations [6, p. 375] appropriate to the discussion of the case when both x and s may be large. For the problem at hand the new variables are defined by the equations

$$\begin{aligned}sz&=\int x^{-1}(x^2+s^2)^{1/2}\,dx=(x^2+s^2)^{1/2}-s\log\left[\frac{s+(s^2+x^2)^{1/2}}{x}\right],\\v&=(x^2+s^2)^{1/4}j.\end{aligned}\tag{A.8}$$

The function $v(z)$ introduced in this way satisfies the equation

$$v_{zz}+s^2v=s^2vw(x^2+s^2)^{-1/2},\tag{A.9}$$

where

$$w(x)=\frac{4s^2x-x^3}{4(s^2+x^2)^{5/2}}.\tag{A.10}$$

A particular solution of (A.9) is that defined by the integral equation

$$v(z)=e^{isz}-s\int_z^\infty v(z')w(t)(t^2+s^2)^{-1/2}\sin s(z-z')\,dz',\tag{A.11}$$

where

$$sz'=(t^2+s^2)^{1/2}-s\log\left[\frac{s+(s^2+t^2)^{1/2}}{t}\right]\tag{A.12}$$

If we set $y(x)=v(z)e^{-isz}$, $y(t)=v(z')e^{-isz'}$ in (A.11) and change the variable of integration from z' to t we obtain the equation

$$y(x)=1-\frac{1}{2i}\int_x^\infty y(t)w(t)[1-e^{-2is(z(x)-z(t))}]\,dt, \tag{A.13}$$

so that

$$|y(x)|\le 1+\int_x^\infty |y(t)w(t)|\,dt. \tag{A.14}$$

Now, by (A.10),

$$\int_x^\infty |w(t)|\,dt\le\int_x^\infty \frac{(4s^2t+t^3)\,dt}{4(s^2+t^2)^{5/2}}=\frac{1}{4s}\left[\left(1+\frac{x^2}{s^2}\right)^{-1/2}+\left(1+\frac{x^2}{s^2}\right)^{-3/2}\right]\le\frac{1}{2s}, \tag{A.15}$$

the integrations being elementary. On applying Gronwall's lemma to (A.14) and using (A.15) we find that $|y(x)|\le\exp(1/2s)\le 2$ for $s>1/\log 4$, therefore (A.3) shows that $y(x)=1+K(x)$, where, by (A.15),

$$|K(x)|\le 2\int_x^\infty |w(t)|\,dt\le\frac{1}{s}. \tag{A.16}$$

On substituting $y=1+K$ into the integral term in (A.13) we find that

$$\begin{aligned} y(x)=1&-\frac{1}{2i}\int_x^\infty w(t)\,dt+\frac{1}{2i}\int_x^\infty w(t)e^{2is[z(t)-z(x)]}dt\\ &-\frac{1}{2i}\int_x^\infty K(t)w(t)\{1-e^{2is[z(t)-z(x)]}\}\,dt. \end{aligned} \tag{A.17}$$

Since $|K(x)|\le s^{-1}$ it follows immediately in view of (A.15) that the modulus of the last term on the right-hand side of (A.17) does not exceed s^{-2}. The second integral on the right-hand side of (A.17) is also $O(s^{-2})$ uniformly for $x\ge 0$. This can be verified by integrating by parts, which yields the equation,

$$\begin{aligned} \int_x^\infty w(t)e^{-2isz(t)}\,dt&=\frac{-1}{8i}\int_x^\infty\frac{(4s^2t^2-t^4)}{(s^2+t^2)^3}\frac{d}{dt}[e^{-2isz(t)}]\,dt\\ &=\frac{(4s^2x^2-x^4)e^{-2isz(x)}}{8i(s^2+x^2)^3}+\frac{1}{8i}\int_x^\infty e^{-2isz(t)}\frac{d}{dt}\left[\frac{4s^2t^2-t^4}{(s^2+t^2)^3}\right]dt. \end{aligned} \tag{A.18}$$

The modulus of the first term on the right-hand side of the preceding equation does not exceed

$$\frac{4(s^2+x^2)x^2}{8(s^2+x^2)^3}\le\frac{1}{2s^2}.$$

If we set $t=s\tau$ in the integral present on the right-hand side of (A.18), we see that its modulus is not greater than

$$\frac{1}{8s^2}\int_0^\infty\left|\frac{d}{d\tau}\left\{\frac{(4-\tau^2)\tau^2}{(1+\tau^2)^3}\right\}\right|d\tau.$$

This is $O(s^{-2})$ since the above integral is convergent. Thus both terms on the right-hand side of equation (A.18) are $O(s^{-2})$ so that the same is true of the second integral

appearing in equation (A.17), which therefore reduces to

$$y(x)=1-\frac{1}{2i}\int_x^\infty w(t)\,dt+O(s^{-2})$$

uniformly for $0\leq x<\infty$. Upon substituting the formula (A.10) and evaluating the resulting integral we find that

$$y(x)=1-\frac{i}{8(s^2+x^2)^{1/2}}+\frac{5is^2}{24(s^2+x^2)^{3/2}}+O(s^{-2})$$

so that

$$\text{(A.19)}\quad j(x)=\frac{e^{isz(x)}}{(s^2+x^2)^{1/4}}\left[1-\frac{i}{8(s^2+x^2)^{1/2}}+\frac{5is^2}{24(s^2+x^2)^{3/2}}+O(s^{-2})\right].$$

The asymptotic form of a second solution of Bessel's equation, independent of $j(x)$, can be obtained by taking the complex conjugate $j^*(x)$ and the general solution is a combination of these. Therefore $J_{is}(x)=Aj(x)+Bj^*(x)$ where the values of the coefficients A, B can be obtained from the known limiting forms of $J_{is}(x)$, either as $x\to 0$ or as $x\to\infty$. If we adopt the former procedure and appeal to the result

$$\Gamma(is)=\left(\frac{2\pi}{s}\right)^{1/2}\exp\left(is\log s-is-\frac{1}{2}s\pi-\frac{1}{4}i\pi\right)\left[1+O(s^{-1})\right],$$

which follows from Stirling's formula (20), we find that we must take $B=0$ and

$$A=(2\pi)^{-1/2}\exp\left(\frac{1}{2}s\pi-\frac{1}{4}i\pi\right)$$

so that

$$J_{is}(x)=(2\pi)^{-1/2}j(x)\exp\left(\frac{1}{2}s\pi-\frac{1}{4}i\pi\right),$$

where $j(x)$ is defined by equation (A.19). The formula (48) follows upon using the above formula in conjunction with (10).

Acknowledgment. The authors wish to thank the referees for drawing their attention to an error in the original version of the manuscript and for other comments that led to the clarification of the paper.

REFERENCES

[1] B. L. J. Braaksma, B. Meulenbeld and H. Lemei, *Integral transforms related to a class of second order linear differential equations*, Indagationes, Proc. Kon. Akad. Wetensch, Amsterdam, Series A, 72 (1969), pp. 77–88.

[2] W. Magnus, F. Oberhettinger and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1965.

[3] D. Naylor, *On an integral transform associated with a condition of radiation, Part* 2, Math. Proc. Camb. Phil. Soc., 77 (1975), pp. 189–197.

[4] _____, *On an integral transform occurring in the theory of diffraction*, this Journal, 8 (1977), pp. 402–411.

[5] _____, *On an integral transform*, Glasgow Math. J., 20 (1979), pp. 1–14.

[6] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[7] E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Oxford University Press, London, 1950.

[8] _____, *The Theory of Functions*, second edition, Oxford University Press, London 1959.

[9] G. N. Watson, *Theory of Bessel Functions*, second edition, Cambridge University Press, London, 1958.